The Annals of Statistics
1994, Vol. 22, No. 2, 805–823

DYNAMIC SAMPLING PLAN IN SHIRYAYEV–ROBERTS PROCEDURE FOR DETECTING A CHANGE IN THE DRIFT OF BROWNIAN MOTION

BY M. S. SRIVASTAVA AND YANHONG WU

University of Toronto and Stanford University

In this paper, a dynamic sampling plan in the Shiryayev–Roberts procedure is considered. It is shown that a two-rate dynamic sampling plan is optimal in the sense that it minimizes the stationary average delay time (SADT). Analytical results as well as numerical comparisons show that it is significantly superior to the fixed sampling plan. The comparison also shows that it is as powerful as the dynamic sampling procedure of Assaf and Ritov. The generalizations to the fast initial response and to the CUSUM procedure are also briefly discussed.

1. Introduction. Quick detection of changes in distribution is an important problem in quality control. Similar problems arise in many areas, such as in the surveillance of birth records for a possible increment in the frequency of genetic malformation considered by Weatherall and Haskey (1976). Several competing procedures, such as the CUSUM, EWMA and Shiryayev–Roberts procedures, have been discussed extensively in the literature [cf. Pollak and Siegmund (1985), Roberts (1966) and Srivastava and Wu (1993)]. However, usually a fixed sampling plan is carried out to monitor the process. Clearly, a procedure which takes fewer samples when no change is expected but takes more samples when a change is expected should be more efficient than the corresponding fixed sampling plan. Such a sampling scheme is called a dynamic sampling plan.

The idea of dynamic sampling has been used by Girshick and Rubin (1952) from a Bayesian decision point of view. Recently, Assaf (1988) considered a Bayesian dynamic sampling procedure which reinitiated interest. Motivated from batch-type production control, Assaf and Ritov (1988) considered a double sequential sampling plan. Assaf and Ritov (1989) further considered another dynamic sampling plan based on the CUSUM procedure [see also Assaf, Pollak and Ritov (1992) for a more general discussion]. A two-interval sampling CUSUM procedure has also been considered by Reynolds, Amin and Arnold (1990) in the discrete-time case and by Wu and Srivastava (1993) in the continuous-time case.

The main purpose of this paper is to study a dynamic sampling plan in the Shiryayev–Roberts procedure which is formally discussed by Pollak and Siegmund (1985). The proposed procedure can be used as a benchmark for

Received September 1991; revised April 1993.
AMS 1991 *subject classifications*. Primary 62L10; secondary 62N10.
Key words and phrases. Average run length, CUSUM procedure, diffusion process, two-rate sampling plan.

comparison as it has certain optimality and, more important, it can be easily implemented in practical situations. Consider the following two situations. First, in a batch-type production process, instead of sampling a fixed percentage of items in all the batches, one may change the sampling percentage (or number) [cf. Assaf (1988)]: Second, in a one-at-a-time production process, instead of sampling at equally spaced intervals, one may change the sampling interval [cf. Reynolds, Amin and Arnold (1990)]. A restriction in both situations is that the maximum sampling rate is bounded from above due to the finite batch size and the production time between two units, respectively.

A unified continuous-time model for the two situations can be formulated as follows. We consider the change point problem in the drift of a Brownian motion. Under the fixed sampling plan with sampling rate 1, the observation process is

$$dW_t = \delta I_{[t>\theta]}dt + dB_t,$$

where B_t is a standard Brownian motion, θ is the change point and δ is the amount of change in the drift, assumed known. The Shiryayev–Roberts procedure is defined to give an alarm at

$$\tau = \inf\{t > 0: R_t > T\},$$

for a specified control limit T, where

$$R_t = \int_0^t \exp\left[\delta(W_t - W_s) - \frac{\delta^2}{2}(t-s)\right] ds.$$

It can be observed that the integrand in R_t is the likelihood ratio of the observations up to time t with change at time s with respect to no change. From Pollak and Siegmund (1985), the process R_t can be written in differential form as

$$dR_t = dt + \delta R_t \, dW_t \quad \text{with } R_0 = 0.$$

Suppose the sampling rate is $a(y)$ when $R_t = y$. Roughly speaking, this means that the next sampling interval should be taken as $1/a(y)$, or the sampling number in the infinitesimal time interval $[t, t+\Delta t)$ is $a(y)\,\Delta t$, approximately. We shall assume that $a(y)$ is bounded such that

$$0 \le a_1 \le a(y) \le a_2 < \infty,$$

where a_1 and a_2 are the lowest and highest sampling rates. For notational convenience, we shall use the same R_t and τ to denote the detecting process and the alarming time under the dynamic sampling plan. We first introduce some standard notation.

Let $\{\tau_1, \dots, \tau_\nu\}$ denote the consecutive alarming intervals until the change in the process is detected. The delay time will be $\sum_1^\nu \tau_i - \theta$. Let $E_\theta(\cdot)$ denote the mean taken with the change at θ. We shall write

$$\mathrm{ARL}_0 = E_\infty \tau \quad \text{and} \quad \mathrm{ARL}_1 = E_0 \tau$$

for the average in-control and out-of-control run length, respectively. The stationary average delay time defined by

$$\mathrm{SADT} = \lim_{\theta\to\infty} E_\theta\left(\sum_1^{\nu}\tau_i - \theta\right)$$

is particularly important in our discussion. The average in-control sampling rate in the long run will be written as

$$\mathrm{ASR}_0 = E_\infty \int_0^{\tau} \frac{a(R_t)\,dt}{E_\infty \tau}.$$

Without loss of generality, we shall always assume that $\mathrm{ASR}_0 = 1$.

To design a dynamic sampling plan, we shall fix $\mathrm{ARL}_0 = T$ and $\mathrm{ASR}_0 = 1$ This is only reasonable in a long run time, that is, with the change occurring far away from the beginning. In Section 6, we shall show that the optimal sampling plan which minimizes SADT is of the two-rate type:

$$a(y) = a_1 I_{[y<S]} + a_2 I_{[S\le y<T]},$$

where S and T are the switching limit and control limit, respectively. The design of S and T will be discussed in Section 2. Their properties are studied in several interesting cases. The asymptotic behaviors of ARL_1 and SADT are discussed in Section 3. The head-start technique is also investigated in order to reduce the difference between ARL_1 and SADT. In Section 4, we first compare the proposed plan with the fixed sampling plan. Then we compare it with the procedure of Assaf and Ritov (1989) (A-R). It will be shown that although the proposed plan has slightly smaller SADT than the A-R procedure, it has much larger ARL_1. However, with the head start under the proposed plan, the two procedures give the same results in an extreme case. In Section 5, we also give the corresponding results for the two-rate sampling CUSUM procedure. The optimality of the two-rate sampling plan is shown in Section 6. Detailed proofs for the main results are given in the Appendix. Technical tools used here are results from diffusion theory. A convenient reference is Karlin and Taylor (1981); the same notation will be used.

2. Design of the two-rate sampling plan. To design a two-rate sampling plan, we shall fix ARL_0 and ASR_0. The first lemma gives the differential form of R_t under any sampling rate. The derivation will be delayed until Section 6.

LEMMA 1. *Suppose the sampling rate is $a(y)$. Then R_t satisfies the following differential form:*

$$dR_t = dt + \delta R_t\, dW_t,$$

where W_t is the observation process satisfying

$$dW_t = \delta a(R_t) I_{[t>\theta]}\, dt + \sqrt{a(R_t)}\, dB_t,$$

where B_t is the standard Brownian motion.

From Lemma 1, it is easy to see that R_t is always a martingale before the change no matter what the sampling rate is. Thus,

$$\mathrm{ARL}_0 = T$$

if $R_0 = 0$. This implies that the control limit T is always taken as ARL_0. It remains to determine the switching limit S, which should satisfy the condition $\mathrm{ASR}_0 = 1$.

Let $L_{[a,b)}$ denote the total time spent in the interval $[a, b)$ for R_t for $0 \le t \le \tau$. Then the two conditions on ARL_0 and ASR_0 are equivalent to

$$T = E_\infty L_{[0,S)} + E_\infty L_{[S,T)}$$

and

$$T = a_1 E_\infty L_{[0,S)} + a_2 E_\infty L_{[S,T)}.$$

That means S should satisfy

$$E_\infty L_{[S,T)} = \frac{1-a_1}{a_2-a_1} T. \tag{1}$$

To find $E_\infty L_{[S,T)}$, it is convenient to use the results from diffusion theory [cf. Karlin and Taylor (1981), Chapter 15]. Before the change, the drift parameter and diffusion parameter of R_t are given by

$$\mu(x) = 1, \qquad \sigma(x) = \delta x\left[\sqrt{a_1} I_{[x<S]} + \sqrt{a_2} I_{[x>S]}\right].$$

By writing

$$\begin{aligned} s(x) &= \exp\left(-\int \frac{2\mu(x)}{\sigma^2(x)}\,dx\right) \\ &= \exp\left(\frac{2}{\delta^2 a_1 x}\right) I_{[x<S]} + \exp\left(\frac{2}{\delta^2 S}\left(\frac{1}{a_1} - \frac{1}{a_2}\right) + \frac{2}{\delta^2 a_2 x}\right) I_{[x\ge S]}, \end{aligned}$$

and $S(x) = \int s(x)\,dx$, it can be verified that, for $a < S$ and $a < x < S$,

$$\lim_{a\to 0} \frac{S(x) - S(a)}{S(T) - S(a)} = 1 \quad \text{for } x < S.$$

The Green's function cannot be written down directly, as 0 is an entrance boundary. Instead, we first assume that the initial state $R_0 = x$ and the exiting boundaries are $[a, T]$ with $a < x < T$. Then we first let $a \to 0$, then let $x \to 0$. Direct calculations give that

$$
\begin{aligned}
&G(0,z)\\
&= \lim_{x\to 0} G(x,z) = 2\frac{S(T)-S(z)}{\sigma^2(z)s(z)}\\
&= \lim_{x\to 0}\lim_{a\to 0} \frac{2\Big(S(T)-S\big(\max(z,x)\big)\Big)\Big(S\big(\min(z,x)\big)-S(a)\Big)}{\sigma^2(z)s(z)\big(S(T)-S(a)\big)}\\
&= 2\frac{\int_z^S \exp\left(2/\delta^2 a_1 u\right)du + \int_S^T \exp\left(2/\delta^2 a_2 u\right)du \exp\Big[\left(2/\delta^2 S\right)\left(1/a_1 - 1/a_2\right)\Big]}{\delta^2 a_1 z^2 \exp\left(2/\delta^2 a_1 z\right)}\\
&\quad \times I_{[z<S]} + 2\frac{\int_z^T \exp\left(2/\delta^2 a_2 u\right)du}{\delta^2 a_2 z^2 \exp\left(2/\delta^2 a_2 z\right)} I_{[S<z<T]}.
\end{aligned}
\tag{2}
$$

Thus,

$$
\begin{aligned}
E_\infty L_{[S,T)} &= \int_S^T G(0,z)\,dz = 2\int_S^T \frac{\exp\left(-2/\delta^2 a_2 z\right)}{\delta^2 a_2 z^2}\int_z^T \exp\left(2/\delta^2 a_2 u\right)du\,dz\\
&= 2\int_S^T \exp\left(2/\delta^2 a_2 u\right)\int_S^u \frac{\exp\left(-2/\delta^2 a_2 z\right)}{\delta^2 a_2 z^2}\,dz\,du\\
&= \int_S^T \left(1-\exp\left[\frac{-2}{\delta^2 a_2}\left(\frac{1}{S}-\frac{1}{u}\right)\right]\right)du.
\end{aligned}
$$

Therefore, we obtain the following lemma.

LEMMA 2. *For a two-rate sampling plan* (a_1, a_2), *given* $\mathrm{ARL}_0 = T$ *and* $\mathrm{ASR}_0 = 1$, *the control limit is* T *and the switching limit* S *satisfies*

$$
\frac{(1-a_1)T}{a_2-a_1} = \int_S^T \left(1-\exp\left[-\frac{2}{\delta^2 a_2}\left(\frac{1}{S}-\frac{1}{u}\right)\right]\right)du. \tag{3}
$$

The equation can be solved numerically, for example, by the Mathematica language.

Several interesting properties for S and T are summarized as follows.

1. As $T\to\infty$,

$$
S \to \frac{2}{\delta^2 a_2}\left(\ln\frac{a_2-a_1}{a_2-1}\right)^{-1}.
$$

This can be easily proved from Lemma 2. Thus, as the control limit T goes to infinity, the switching limit S goes to a finite number rather than infinity. This is due to the recurrent property of R_t [cf. Pollak and Siegmund (1985)]. From this property, a convenient method to design the switching limit S is to assume $T=\infty$.

TABLE 1
Switching limit S with T = 100

(a_1,a_2)	δ 0.01	0.05	0.1	0.2	0.5	1.0
(0.5, 2)	66.22	56.37	41.22	23.41	7.10	2.86
(0.5, 5)	86.92	67.29	49.26	28.85	9.24	2.94
(0.5, 10)	90.92	69.64	51.29	30.37	9.90	3.17
(0.5, 20)	92.25	70.67	52.23	31.09	10.21	3.29
(0.5, 50)	92.88	71.24	52.77	31.51	10.40	3.36
(0.5, ∞)	93.25	71.61	53.12	31.78	10.53	3.40
(0, 2)	49.75	43.73	31.41	16.62	4.54	1.33
(0, 5)	78.40	57.62	38.84	20.25	5.55	1.64
(0, 10)	86.14	60.57	40.73	21.26	5.85	1.73
(0, 20)	88.79	61.85	41.59	21.74	6.00	1.77
(0, 50)	89.97	62.57	42.09	20.14	6.08	1.81
(0, ∞)	90.63	63.03	42.42	22.20	6.14	1.82

2. If $a_2 \to \infty$, then from (3), S will approximately satisfy

$$\frac{(1-a_1)\delta^2}{2}T = \frac{T-S}{S} - \ln\frac{T}{S}. \tag{4}$$

3. If both T and a_2 go to ∞, then

$$S \to \frac{2}{\delta^2(1-a_1)}.$$

In particular, if $a_1 = 0$,

$$S \to \frac{2}{\delta^2}. \tag{5}$$

4. If $\delta \to 0$, then

$$S \to \frac{a_2-1}{a_2-a_1}T,$$

which is linear in T.

Table 1 gives some numerical values of S with $T = 100$. For several values of δ, we choose several combinations of (a_1, a_2). When $a_1 = 0$, it means no sampling, and when $a_2 \doteq \infty$, the sampling rate is taken as large as possible. It can be seen that, for small δ, the effect of a_1 is not very large and, for a_2 larger than 10, S does not change significantly.

3. Asymptotic properties of ARL_1 and SADT and head start technique. In this section, we study the asymptotic behavior of the ARL_1 and SADT under the dynamic sampling plan. We also investigate the head start technique. The following lemma lists the results for ARL_1 and SADT under the fixed sampling plan; its proof can be found in Shiryayev (1963) or Pollak and Siegmund (1985).

LEMMA 3. *Under the fixed sampling plan,*

$$\mathrm{ARL}_1 = \frac{2}{\delta^2}\exp\left(\frac{2}{\delta^2 T}\right)\int_{2/\delta^2 T}^{\infty}\frac{\exp(-x)}{x}\,dx, \tag{6}$$

$$\mathrm{SADT} = \frac{2}{\delta^2}\left(\exp\left(\frac{2}{\delta^2 T}\right)\int_{2/\delta^2 T}^{\infty}\frac{\exp(-z)}{z}\,dz - 1 + \frac{2}{\delta^2 T}\int_0^{\infty}\exp\left(\frac{-2z}{\delta^2 T}\right)\frac{\ln(1+z)}{z}\,dz\right). \tag{7}$$

When T is large,

$$\mathrm{ARL}_1 \approx \mathrm{SADT} = \frac{2}{\delta^2}\left[\ln\left(\frac{\delta^2 T}{2}\right) + O(1)\right].$$

The exact formulas of ARL_1 and SADT under the dynamic sampling plan can be obtained by using the diffusion theory. For simplicity we only give ARL_1. From Lemma 1, R_t follows the following differential form after the change

$$dR_t = \left[1 + \delta^2 R_t\left(a_1 I_{[R_t < S]} + a_2 I_{[R_t > S]}\right)\right]dt + \delta R_t\left(\sqrt{a_1} I_{[R_t < S]} + \sqrt{a_2} I_{[R_t > S]}\right)dB_t,$$

with the same diffusion parameter as before the change and the drift parameter becomes

$$\mu(x) = 1 + \delta^2 x\left(a_1 I_{[x<S]} + a_2 I_{[x>S]}\right).$$

Following the same lines as in the proof of Lemma 2, we can obtain that

$$\begin{aligned}
\mathrm{ARL}_1 &= \int_0^T G(0,z)\,dz \\
&= \frac{2}{\delta^2 a_1 S}\int_{1/S}^{\infty}\frac{1}{z}\exp\left(\frac{-2z}{\delta^2 a_1}\right)dz \\
&\quad + \frac{2}{\delta^2 a_2}\exp\left(\frac{2}{\delta^2 a_2 T}\right)\int_{1/T}^{1/S}\frac{1}{z}\exp\left(-\frac{2z}{\delta^2 a_2}\right)dz \\
&\quad + \left(1 - \exp\left[-\frac{2}{\delta^2 a_2}\left(\frac{1}{S} - \frac{1}{T}\right)\right]\right) \\
&\quad \times \left(\frac{a_2}{a_1}\exp\left(\frac{2}{\delta^2 S a_1}\right)\int_{1/S}^{\infty}\frac{1}{z^2}\exp\left(-\frac{2z}{\delta^2 a_1}\right)dz - S\right).
\end{aligned}$$

As $T \to \infty$, to first order,

$$\mathrm{ARL}_1 \approx \frac{2}{\delta^2 a_2} \exp\left(\frac{2}{\delta^2 a_2 T}\right) \int_{1/T}^{1/S} \frac{1}{z} \exp\left(\frac{-2z}{\delta^2 a_2}\right) dz,$$

which is approximately equal to ARL_1 under the fixed sampling plan with the shift amount $\delta\sqrt{a_2}$. From Lemma 3, we know that

$$\mathrm{ARL}_1 \approx \frac{2}{\delta^2 a_2} \ln \frac{T\delta^2 a_2}{2} \quad \text{as } T \to \infty.$$

Thus, the reduction on the average delay time is mainly dependent on the higher sampling rate a_2. In the following, we consider the most interesting case with $a_1 = 0$, that is, no sampling when R_t is below the switching limit.

Denote by $\mathrm{ARL}_1(y) = E_0[\tau \mid R_0 = y]$ the average out-of-control run length when the initial state $R_0 = y$, and by $\alpha(y)$ the stationary distribution for the controlled process (instantaneous return process) R_t when there is no change with switching limit S and control limit T. $\mathrm{ARL}_1(y)$ can be obtained similar to ARL_0.

Lemma 4 follows from renewal theory and the Markov property of R_t.

LEMMA 4.

$$\alpha(y) = \frac{1}{T} \int_0^y G(0,z)\, dz$$

and

$$\mathrm{SADT} = \int_0^T \mathrm{ARL}_1(y)\, d\alpha(y),$$

where $G(0,z)$ *is given in* (2).

The following two lemmas give the asymptotic behavior of $\alpha(y)$ and $\mathrm{ARL}_1(y)$ as $a_2 \to \infty$; their exact forms are given by (16) and (17) in the Appendix.

LEMMA 5. *As* $a_2 \to \infty$,

$$\alpha(y) = \frac{y}{T} \quad \text{for } y < S$$

and

$$\alpha(T) - \alpha(S) = O\left(\frac{1}{a_2}\right).$$

Thus, $\alpha(y)$ is uniformly distributed in $(0, S)$ and has mass $1 - S/T$ at $y = S$ as $a_2 \to \infty$.

To find an approximation for $\mathrm{ARL}_1(y)$, we first note that if the initial state is $y < S$, the process will linearly increase to S since no sampling is taken. Thus,

$$\mathrm{ARL}_1(y) = S - y + \mathrm{ARL}_1(S) \quad \text{for } y < S.$$

On the other hand, for $R_0 = y \geq S$, we have the following lemma.

LEMMA 6. *As $a_2 \to \infty, \mathrm{ARL}_1(y) = O(1/a_2)$ for $y > S$, and*

$$\mathrm{ARL}_1(S) = S\left(1 - \frac{S}{T}\right) + \frac{2}{\delta^2 a_2}\left[\ln\frac{T}{S} - \frac{1}{T}(T - S)\right] + O\left(\frac{1}{a_2}\right).$$

The exact form for $\mathrm{ARL}_1(S)$ can be obtained from (18) in the Appendix. Combining the above three lemmas, we get the following theorem.

THEOREM 1. *Suppose $a_1 = 0$, then as $a_2 \to \infty$,*

$$\mathrm{ARL}_1 = 2S\left(1 - \frac{S}{2T}\right) + \frac{2}{\delta^2 a_2}\left[\ln\frac{T}{S} - \frac{1}{T}(T - S)\right] + o\left(\frac{1}{a_2}\right),$$

$$\mathrm{SADT} = S\left(1 - \frac{S}{2T}\right) + \frac{2}{\delta^2 a_2}\left[\ln\frac{T}{S} - \frac{1}{T}(T - S)\right] + o\left(\frac{1}{a_2}\right).$$

From Theorem 1, we can see that although ARL_1 and SADT do not differ much under the fixed sampling plan, they are quite different under the dynamic sampling plan. As $a_2 \to \infty$, ARL_1 is approximately twice the SADT. Also, as $T \to \infty$, $S \to 2/\delta^2$. Thus, we have the following corollary.

COROLLARY 1. *If $T \to \infty$ in addition to the conditions of Theorem 1, then*

(8)
$$\mathrm{ARL}_1 = 2S\left(1 - \frac{S}{2T}\right) + \frac{2}{a_2\delta^2}\left[\ln\frac{\delta^2 T}{2} - 1 + O\left(\frac{1}{T}\right)\right] + o\left(\frac{1}{a_2}\right),$$
$$\mathrm{SADT} = S\left(1 - \frac{S}{2T}\right) + \frac{2}{a_2\delta^2}\left[\ln\frac{\delta^2 T}{2} - 1 + O\left(\frac{1}{T}\right)\right] + o\left(\frac{1}{a_2}\right).$$

In the special case with $a_1 = 0$ and $a_2 = \infty$, the switching limit S satisfies

(9)
$$\frac{\delta^2}{2}T = \frac{T - S}{S} - \ln\frac{T}{S},$$

and the corresponding SADT and ARL_1 are given by

(10)
$$\mathrm{SADT} = S\left(1 - \frac{S}{2T}\right) = \frac{1}{2}\mathrm{ARL}_1.$$

From a theoretical point of view, SADT is the only reasonable measure for comparison since ARL_0 and ASR_0 are meaningful in a long run time. However, a disadvantage under the proposed plan is that ARL_1 is too large compared to its SADT. One method to overcome this is to use the head start or the so-called fast initial response technique [cf. Lucas and Crosier (1982)]. More specifically, instead of starting at $R_0 = 0$, we let $R_0 = S$, that is, taking a sample at the beginning. This has almost no effect on ARL_0, but reduces ARL_1 significantly.

Consider the case wth $a_1 = 0$. Suppose the switching limit is S^* and the control limit is T^* under the head start $R_0 = S^*$. From the martingale property of R_t, $\text{ARL}_0 = T^* - S^*$. On the other hand, from the proof of Lemma 2, the total sample size in a cycle before the change is

$$a_2 \int_{S^*}^{T^*} \left(1 - \exp\left[\frac{-2}{\delta^2 a_2}\left(\frac{1}{S^*} - \frac{1}{u}\right)\right]\right) du = \frac{2}{\delta^2}\left(\frac{T^* - S^*}{S^*} - \ln \frac{T^*}{S^*}\right) + O\left(\frac{1}{a_2}\right),$$

as $a_2 \to \infty$. Thus, S^* and T^* can be chosen from the two conditions $\text{ARL}_0 = T$ and $\text{ASR}_0 = 1$. The corresponding ARL_1 and SADT can be obtained as in the case $R_0 = 0$, and the details will not be given here. For example, ARL_1 is actually the $\text{ARL}_1(S)$ given in Lemma 6 with S and T replaced by S^* and T^*.

As an example, we consider the special case $a_2 = \infty$. In this case, S^* and T^* satisfy

$$T^* = T + S^*, \tag{11}$$

$$\frac{2}{\delta^2}\left(\frac{T^* - S^*}{S^*} - \ln \frac{T^*}{S^*}\right) = T. \tag{12}$$

As $T \to \infty$, $S^* \to 2/\delta^2$, as in the case without a head start. Under the head start, it is obvious that

$$\text{ARL}_1^* = \text{SADT}^* = S^*\left(1 - \frac{S^*}{T^*}\right). \tag{13}$$

4. Comparison with fixed sampling plan and the Assaf–Ritov procedure. We first consider the maximum reduction on the average delay time under the $(0, \infty)$ plan compared with the fixed sampling plan. Table 2 gives some numerical values of ARL_1 and SADT for $T = 100$ and 500 without the head start. For the dynamic sampling plan, the switching limit S, ARL_1 and SADT are calculated from (9) and (10). All numerical evaluations are carried out by using the Mathematica language. We see that the reduction is significant.

Table 3 gives the corresponding value with the head start based on (11)–(13). Only S^* is given as $T^* = T + S^*$. Comparing Table 2 with Table 3, we see that the head start has very little effect on the behavior of SADT except for small δ, but it reduces ARL_1 significantly.

In the following, we compare the proposed plan with the dynamic sampling CUSUM procedure of Assaf and Ritov (1989). Their procedure can be briefly explained as follows. Suppose the samples are taken at the time points $\{\varepsilon, 2\varepsilon, \ldots\}$.

TABLE 2
Comparison of dynamic $(0, \infty)$ and fixed sampling plans

	$T = 100$			$T = 500$		
δ	Dynamic S	SADT($\frac{1}{2}$ARL$_1$)	Fixed SADT(ARL$_1$)	Dynamic S	SADT($\frac{1}{2}$ARL$_1$)	Fixed SADT(ARL$_1$)
0.1	42.42	33.42	39.61(72.37)	97.35	87.86	128.45(209.57)
0.2	22.20	19.74	27.81(46.15)	36.74	35.39	68.60(100.73)
0.5	6.14	5.95	12.15(17.57)	7.38	7.32	22.17(29.05)
1.0	1.82	1.81	5.16(6.85)	1.95	1.94	8.05(9.94)
1.5	0.85	0.84	2.92(3.73)	0.88	0.88	4.27(5.13)
2.0	0.49	0.48	1.91(2.38)	0.50	0.50	2.68(3.17)
2.5	0.31	0.31	1.36(1.66)	0.32	0.32	1.86(2.17)

At each sampling point, we take samples sequentially. If we let W_{it} denote the information process at the ith sampling point, then the sampling will be stopped at

$$\tau_i = \inf\{t > 0\colon W_{it} < -C\varepsilon \text{ or } > A\},$$

where C and A are two prespecified constants. If $W_{it} < -C\varepsilon$, we move to the next sampling point; otherwise, an alarm will be made. It is also assumed that the entire batch consists of either defectives or nondefectives, that is, W_{it} has fixed drift parameter 0 or δ. The parameters C and A are chosen such that $\mathrm{ARL}_0 = T$ and $\mathrm{ASR}_0 = 1$ given the sampling interval length ε.

In general, the two procedures are not comparable, as the assumptions are different. However, we shall show that when $\varepsilon \to 0$, the Assaf–Ritov procedure gives the same results as the proposed $(0, \infty)$ plan. It is clear that in the Assaf–Ritov procedure, when $\varepsilon = 0$, $\mathrm{ARL}_1 = \mathrm{SADT}$ and both achieve their minimum values. Notationally, their δ is our ε here; T_{fa} is our ARL_0; τ_0 is our ASR_0; T_d is our ARL_1; and μ_0 is our δ. From Assaf and Ritov [(1989), Theorem 1] it is not difficult to verify that C and A satisfy

$$\frac{\exp(\delta A) - 1 - A\delta}{\delta^2/2} = T \quad \text{and} \quad C = \frac{\exp(\delta A) - 1}{\delta T}, \tag{14}$$

TABLE 3
Average delay times with head start and under the Assaf–Ritov (A-R) procedure

	$T = 100$			$T = 500$		
δ	(A, C) (A-R)	SADT(ARL$_1$)	S^*	(A, C) (A-R)	SADT(ARL$_1$)	S^*
0.1	(8.58,014)	42.40	73.61	(16.36,6.08)	97.38	120.93
0.2	(7.53,0.18)	22.18	28.50	(13.06,0.13)	36.69	36.60
0.5	(5.58,0.31)	6.14	6.54	(8.43,0.27)	7.39	7.50
1.0	(4.01,0.54)	1.81	1.85	(5.55,0.51)	1.94	1.95
1.5	(3.18,0.78)	0.85	0.85	(4.23,0.76)	0.88	0.88
2.0	(2.67,1.04)	0.48	0.48	(3.46,1.01)	0.50	0.50

and the corresponding SADT and ARL_1 are given by

$$\text{SADT} = \mathrm{ARL}_1 = \frac{1 - \exp(-\delta A)}{\delta C}. \tag{15}$$

Carefully comparing (11)–(13) with (14)–(15), we can see that the two procedures give the same results if we write $T^*/S^* = \exp(\delta A)$. This means that the maximum reductions under the two procedures are the same.

Table 3 also gives the corresponding values for (A, C). Again, the Mathematica language is used for the numerical evaluation. It is clear from this comparison that without the head start, the proposed plan has much longer ARL_1 than the A-R procedure. With the head start in the proposed plan, the two become equally powerful.

REMARK 1. As the referee pointed out, even without the head start under the proposed plan, one can modify the A-R procedure such that the two procedures give the same result in the above extreme case. More specifically, in the A-R procedure with $\varepsilon = 0$, we wait time W following every alarm without sampling. Then (A, C) should satisfy

$$W + T_{fa} = T \quad \text{and} \quad \tau_0 T_{fa} = T.$$

The corresponding SADT is given by

$$\text{SADT} = \frac{W}{2}\frac{W}{T} + T_d,$$

as the stationary distribution is uniformly lying in $(0, W)$ with total probability W/T. Now one can choose the optimal W such that SADT achieves its minimum value. It can be shown that this modified A–R procedure does give the same results as the proposed $(0, \infty)$ plan without the head start.

It may be noted that Assaf and Ritov (1989) also compared the average sampling number during the delay detection time (SADN). In their procedure, when $\varepsilon = 0$, it is the same as ARL_1 under the fixed sampling plan. In our case,

$$\text{SADN} = \lim_{\theta \to \infty} E \int_{\theta}^{\sum_1^{\nu} \tau_i} a(R_t)\, dt,$$

where the τ_i are defined in the introduction.

The following lemma gives the result for SADN.

LEMMA 7. *Under the* $(0, \infty)$ *sampling plan,*

$$\text{SADN} = \frac{2}{\delta^2}\left[\ln \frac{T}{S} - \frac{T - S}{T}\right].$$

The proof will be given in the Appendix. For example, as $T \to \infty$,

$$\text{SADN} = \frac{2}{\delta^2} \ln \frac{\delta^2 t}{2}\big(1 + o(1)\big),$$

which is approximately equal to ARL_1 under the fixed sampling plan from Lemma 3. Thus, the implementation of a dynamic sampling plan reduces the average delay time, but not the average sample number in delay time.

5. Two-rate sampling CUSUM procedure. A natural problem is to consider the two-rate sampling plan in the CUSUM procedure as in Reynolds, Amin and Arnold (1990). Wu and Srivastava (1993) have studied this procedure, following the same lines as in this paper. Suppose the switching limit and control limit are denoted by c and d, respectively. The main results are given in the following theorem.

THEOREM 2. *Given* $\mathrm{ARL}_0 = T$ *and* $\mathrm{ASR}_0 = 1$, c *and* d *satisfy*

$$\frac{\delta^2 T}{2}\frac{a_2(1-a_1)}{(a_2-a_1)} = \exp\big[(d-c)\delta\big] - 1 - \delta(d-c),$$

$$\frac{\delta^2 T}{2} = \exp(\delta d) - 1 - \delta d.$$

(i) *As* $T \to \infty$,

$$\mathrm{ARL}_1 = \frac{2}{\delta^2 a_2}\left(\ln\frac{\delta^2 T}{2} - 1\right) - \frac{2}{\delta^2}\left(\frac{1}{a_1} - \frac{1}{a_2}\right)\ln\frac{1-a_1}{1-a_1/a_2} + O\left(\frac{\ln T}{T}\right).$$

(ii) *If in addition* $a_1 \to 0$,

$$\mathrm{ARL}_1 = \frac{2}{\delta^2 a_2}\left(\ln\frac{\delta^2 T}{2} - 2\right) + \frac{2}{\delta^2} + O\left(\frac{\ln T}{T}\right).$$

(iii) *As* $a_1 \to 0$ *and* $a_2 \to \infty$,

$$\lim(\mathrm{ARL}_1) = \lim(\mathrm{SADT}) = T\frac{1-\exp(-\delta d)}{\exp(\delta d) - 1}.$$

Again we note that, under the $(0, \infty)$ sampling plan, it gives the same result as the other two procedures.

It would be interesting to compare the three dynamic sampling plans mentioned in this paper in the discrete-time case, especially in the one-at-a-time production case. The Assaf–Ritov procedure seems more natural to be adapted to this situation after some modifications since it is periodic. The proposed plan is also quite convenient as the sampling procedure consists of a sequence of sampling and no-sampling intervals and, more important, the no-sampling interval is predictable [cf. Girshick and Rubin (1952)]. It seems that the two-rate sampling CUSUM procedure has certain disadvantages compared to the other two. For example, the lower sampling rate cannot be taken as zero, and also the switching time from lower to higher sampling rate is random [cf. Reynolds, Amin and Arnold (1990)]. We shall present the results in a future communication.

6. Optimality of the two-rate sampling plan. Now we give a brief proof for the optimality of the two-rate sampling plan in the sense that it minimizes the SADT.

The idea of the proof is based on the results of Shiryayev (1963), where it was shown that under the fixed sampling plan the Shiryayev–Roberts procedure is optimal in the sense that it minimizes SADT for fixed ARL_0.

We shall show that the two-rate sampling plan is the limit of a sequence of two-rate Bayesian sampling procedures which are known to be optimal in a certain sense.

Suppose the change point θ has an exponential prior distribution with parameter λ, that is,

$$P(\theta > t) = \exp(-\lambda t).$$

Let H_t denote the history of the observation process up to time t. Write

$$\pi_t = P\big(\theta < t \,|\, H_t\big)$$

as the posterior probability of $\{\theta < t\}$ up to time t, and write

$$R_t^{(\lambda)} = \frac{\pi_t}{\lambda(1-\pi_t)}.$$

Assume that the sampling rate is dependent on π_t. Since $R_t^{(\lambda)}$ is a deterministic monotone increasing function of π_t, we may assume that the sampling rate is $a(R_t^{(\lambda)})$ at time t. Thus, the observation process dW_t has the same probability law as

$$a\big(R_t^{(\lambda)}\big)\delta I_{[\theta > t]}\,dt + \sqrt{a\big(R_t^{(\lambda)}\big)}\,dB_t.$$

Since the dynamic sampling procedure is adaptive, we have

$$\pi_t = \frac{\lambda \int_0^t \exp(-\lambda s)\exp\Big[-\int_s^t \big[a\big(R_u^{(\lambda)}\big)\big]^{-1}\big(\delta a(R_u^{(\lambda)})\,dW_u - \frac{1}{2}\delta^2 a^2(R_u^{(\lambda)})\,du\big)\Big]\,ds}{\exp(-\lambda t) + \lambda\int_0^t \exp(-\lambda s)\exp\Big[-\int_s^t \big[a(R_u^{(\lambda)})\big]^{-1}\big(\delta a(R_u^{(\lambda)})\,dW_u - \frac{1}{2}\delta^2 a^2(R_u^{(\lambda)})\,du\big)\Big]\,ds}$$

and

$$R_t^{(\lambda)} = \int_0^t \exp\big[\lambda(t-s)\big]\exp\bigg[-\int_s^t \Big[a\big(R_u^{(\lambda)}\big)\Big]^{-1}\Big(\delta a\big(R_u^{(\lambda)}\big)\,dW_u - \tfrac{1}{2}\delta^2 a^2\big(R_u^{(\lambda)}\big)\,du\Big)\bigg]\,ds.$$

Now, from Itô's formula, we find that $R_t^{(\lambda)}$ satisfies the following differential form:

$$dR_t^{(\lambda)} = \big(1 + \lambda R_t^{(\lambda)}\big)\,dt + \delta R_t^{(\lambda)}\,dW_t,$$

with $R_0^{(\lambda)} = 0$.

Suppose τ_λ is a stopping time adapted to $R_t^{(\lambda)}$. We consider the minimization of $E(\tau_\lambda - \theta \mid \tau_\lambda > \theta)$ under the following two conditions:

$$P(\tau_\lambda < \theta) = \alpha$$

and

$$E\int_0^{\tau_\lambda} a\big(R_t^{(\lambda)}\big)\,dt = E\tau_\lambda.$$

The first condition is the constraint on the false alarm probability. The second condition is the constraint on the average sampling rate, which is equal to 1 in our case.

To satisfy the first condition, the optimal stopping time is

$$\begin{aligned}\tau_\lambda &= \inf\{t > 0\colon \pi_t \geq 1-\alpha\}\\ &= \inf\left\{t > 0\colon R_t^{(\lambda)} \geq \frac{1-\alpha}{\lambda\alpha}\right\},\end{aligned}$$

by the definition of $R_t^{(\lambda)}$ for any sampling rate $a(y)$. Assaf (1988) has considered the optimal dynamic Bayesian sampling procedure and showed that, when the sampling rate is unbounded, the $(0, \infty)$ sampling plan is optimal under the two constraints in the sense that it minimizes the average delay time. Using the same argument here, we can easily show that the two-rate (a_1, a_2) sampling plan is optimal under the two constraints if the sampling rate is bounded. This means that there exists a switching limit such that the sampling rate will be switched from a_1 to a_2 as long as $R_t^{(\lambda)}$ crosses the limit from below. Obviously, (a_1, a_2) should depend on the parameter λ as well as on the parameter α.

Now let $\lambda \to 0$, that is, the change point occurs far from the beginning. Then $R_t = R_t^{(0)}$ satisfies

$$dR_t = dt + \delta R_t\,dW_t,$$

as given in Lemma 1. Let $\alpha \to 1$ in such a way that the ratio $(1-\alpha)/\lambda \to T$, a specified number. Then it is easy to see that

$$\tau_\lambda \to \tau = \inf\{t > 0\colon R_t \geq T\}.$$

Now the first constraint becomes

$$\lim_{\lambda\to 0}\frac{P(\tau_\lambda > \theta)}{\lambda} = T.$$

However, since θ has an exponential distribution, it can easily be shown that

$$\lim_{\lambda\to 0}\frac{P(\tau_\lambda > \theta)}{\lambda} = \lim_{\lambda\to 0}\frac{1 - E\exp(-\lambda\tau_\lambda)}{\lambda} = \lim_{\lambda\to 0} E\tau_\lambda.$$

Thus, the first condition becomes

$$E_\infty\tau = T.$$

The second condition becomes

$$\lim_{\lambda\to 0} E\int_0^{\tau_\lambda} a\big(R_t^{(\lambda)}\big)\,dt = E_\infty \int_0^{\tau} a(R_t)\,dt = \int_0^T a(y)G(0,y)\,dy = T,$$

which is the condition on the average sampling rate before the change.

From the memoryless property of the exponential distribution, it is easy to show that $E[\tau_\lambda - \theta \mid \tau_\lambda > \theta]$ is the same as the unconditional average delay time [cf. Shiryayev (1963) and Assaf (1988)]. In addition, the limit of the switching limit is only dependent on the constant T. Thus, we have proved that the two-rate sampling plan in the Shiryayev–Roberts procedure is optimal in the sense that it minimizes the stationary average delay time.

APPENDIX

Proofs of main results. In this appendix, we prove Lemmas 5–7 and Theorem 1.

PROOF OF LEMMA 5. For fixed a_1 and a_2, from renewal theory we see that for $y < S$ the stationary distribution is equal to

$$\begin{aligned}
\alpha(y) &= \frac{1}{T}\int_0^y G(0,x)\,dx \\
&= \frac{2}{T}\int_0^y \frac{\exp(-2/\delta^2 a_1 u)}{\delta^2 a_1 z^2}\int_z^S \exp\left(\frac{2}{\delta^2 a_1 u}\right) du\,dz \\
&\quad + \frac{2}{T}\int_0^y \frac{\exp(-2/\delta^2 a_2 u)}{\delta^2 a_1 z^2}\,dz \int_S^T \exp\left(\frac{2}{\delta^2 a_2 u}\right) du \exp\left[\frac{2}{\delta^2 S}\left(\frac{1}{a_1} - \frac{1}{a_2}\right)\right] \qquad (16) \\
&= \frac{y}{T} + \frac{1}{T}\exp\left(-\frac{2}{\delta^2 a_1 y}\right)\int_y^S \exp\left(\frac{2}{\delta^2 a_1 u}\right) du \\
&\quad + \frac{1}{T}\exp\left(-\frac{2}{\delta^2 a_1 y}\right)\exp\left[\frac{2}{\delta^2 S}\left(\frac{1}{a_1} - \frac{1}{a_2}\right)\right]\int_S^T \exp\left(\frac{2}{\delta^2 a_2 u}\right) du.
\end{aligned}$$

As $a_1 \to 0$, S as a function of a_1 is decreasing to a constant. For notational convenience, we still use the same S to denote this limit. Thus, for $y < S$,

$$\alpha(y) \to \frac{y}{T} \quad \text{as } a_1 \to 0.$$

That means $\alpha(y)$ is uniformly distributed in $(0,S)$. Similarly, from (2) we can show that, for $y > S$,

$$\begin{aligned}
\alpha(T) - \alpha(y) &= \frac{1}{T}\int_y^T G(0,z)\,dz \\
&= \frac{1}{T}(T-y) - \frac{1}{T}\int_y^T \exp\left[\frac{2}{\delta^2 a_2}\left(\frac{1}{u} - \frac{1}{y}\right)\right] du,
\end{aligned} \qquad (17)$$

which is free of a_1 functionally. Thus as $a_2 \to \infty$, if we still use the same S to denote the limit of the switching limit, then, for $y > S$,

$$\begin{aligned}\alpha(T) - \alpha(y) &= \frac{1}{T}\int_y^T \frac{2}{\delta^2 a_2}\left(\frac{1}{u} - \frac{1}{y}\right) du + o\left(\frac{1}{a_2}\right)\\ &= O\left(\frac{1}{a_2}\right),\end{aligned}$$

which proves Lemma 5. □

PROOF OF LEMMA 6.

$$(18)\quad G(s,z) = \begin{cases} 2\dfrac{S(T) - S(z)}{\sigma^2(z)s(z)} \\ \quad = \dfrac{2}{\delta^2 a_2}\displaystyle\int_z^T \frac{1}{x^2}\exp\left(\frac{2}{\delta^2 a_2 x}\right) dx \exp\left(-\frac{2}{\delta^2 a_2 z}\right), & \text{for } z \geq S,\\ 2\dfrac{S(T) - S(S)}{\sigma^2(z)s(z)} \\ \quad = \dfrac{2}{a_1\delta^2}\exp\left(-\dfrac{2}{\delta^2 a_1 z}\right)\displaystyle\int_S^T \frac{1}{x^2}\exp\left(\frac{2}{\delta^2 a_2 x}\right) dx \\ \quad \times \exp\left[\dfrac{2}{\delta^2 S}\left(\dfrac{1}{a_1} - \dfrac{1}{a_2}\right)\right], & \text{for } z < S.\end{cases}$$

First, we note that, for $z > S, G(S,z)$ is free of a_1. Letting $a_2 \to \infty$, we find that

$$G(S,z) = \frac{2}{\delta^2 a_2}\left(\frac{1}{z} - \frac{1}{T}\right) + o\left(\frac{1}{a_2}\right) \quad \text{for } S < z < T.$$

On the other hand, as $a_1 \to 0$,

$$\begin{aligned}&\frac{2}{a_1\delta^2}\int_0^S \exp\left(-\frac{2}{\delta^2 a_1 z}\right) dz \exp\left(\frac{2}{\delta^2 S a_1}\right)\\ &\quad = \left(\frac{2}{\delta^2 a_1}\right)^2 \int_{2/a_1\delta^2 S}^{\infty} \frac{1}{z^2}\exp(-z)\, dz \exp\left(\frac{2}{\delta^2 S a_1}\right) \quad \to \quad S^2.\end{aligned}$$

Similarly, as $a_2 \to \infty$,

$$\begin{aligned}&\int_S^T \frac{1}{x^2}\exp\left(\frac{2}{\delta^2 a_2 x}\right) dx \exp\left(-\frac{2}{\delta^2 S a_2}\right)\\ &\quad = \left(\frac{1}{S} - \frac{1}{T}\right) + O\left(\frac{1}{a_2}\right).\end{aligned}$$

Thus, for $a_1 = 0$, we have, as $a_2 \to \infty$,

$$\begin{aligned}\mathrm{ARL}_1(S) &= \int_0^S G(S,z)\,dz + \int_S^T G(S,z)\,dz \\ &= S\left(1 - \frac{S}{T}\right) + \frac{2}{\delta^2 a_2}\left[\ln\frac{T}{S} - \frac{1}{T}(T-S)\right] + O\left(\frac{1}{a_2}\right).\end{aligned}$$

Obviously, for $y > S$, $\mathrm{ARL}_1(y) \to 0$ as $a_2 \to \infty$. □

The proof of Theorem 1 can thus be obtained by Lemmas 5 and 6.

PROOF OF LEMMA 7. The proof of Lemma 7 can be obtained from the proof of Lemma 6. We note that for any a_1, as $a_2 \to \infty$,

$$a_2 G(S,z) = \frac{2}{\delta^2}\left(\frac{1}{z} - \frac{1}{T}\right) \quad \text{for } S < z < T.$$

On the other hand, we have shown that for $a_1 = 0$, as $a_2 \to \infty$,

$$\int_0^S G(S,z)\,dz \to S\left(1 - \frac{S}{T}\right).$$

Thus, under the $(0, \infty)$ sampling plan, the total sample number after the change is

$$\int_S^T \frac{2}{\delta^2}\left(\frac{1}{z} - \frac{1}{T}\right)dz = \frac{2}{\delta^2}\left[\ln\frac{T}{S} - \frac{T-S}{T}\right].$$

Acknowledgments. We wish to thank the referee and an Associate Editor for many helpful suggestions which improved the presentation of Section 4 and also for pointing out an error in Lemma 6 in the original version. Thanks are also due to Ashim Mallik for several valuable discussions.

REFERENCES

ASSAF, D. (1988). A dynamic sampling procedure for detecting a change in distribution. *Ann. Statist.* **16** 236–253.

ASSAF, D. and RITOV, Y. (1988). A double sequential procedure for detecting a change in distribution. *Biometrika* **75** 715–722.

ASSAF, D. and RITOV, Y. (1989). Dynamic sampling procedures for detecting a change in the drift of Brownian motion: a non-Bayesian model. *Ann. Statist.* **17** 793–800.

ASSAF, D., POLLAK, M. and RITOV, Y. (1992). A new look at warning and action lines of surveillance schemes. *J. Amer. Statist. Assoc.* **87** 889–895.

GIRSHICK, M. A. and RUBIN, H. M. (1952). A Bayes approach to a quality control model *Ann. Math. Statist.* **23** 114–125.

KARLIN, S. and TAYLOR, H. M. (1981). *A Second Course in Stochastic Processes*. Academic, New York.

LUCAS, J. M. and CROSIER, P. R. (1982). Fast initial response (FIR) for the cumulative sum quality control schemes. *Technometrics* **24** 199–205.

POLLAK, M. and SIEGMUND, D. (1985). A diffusion process and its application to detecting a change in the drift of a Brownian motion process. *Biometrika* **72** 267–280.

REYNOLDS, M. R., JR., AMIN, R.M. and ARNOLD, J. C. (1990). CUSUM charts with variable sampling intervals. *Technometrics* **32** 371–396.

ROBERTS, S. M. (1966). A comparison of some control chart procedures. *Technometrics* **8** 411–430.

SHIRYAYEV, A. N. (1963). On optimum methods in quickest detection problems. *Theory Probab. Appl.* **13** 22–46.

SIEGMUND, D. (1985). *Sequential Analysis: Tests and Confidence Intervals*. Springer, Berlin.

SRIVASTAVA, M. S. and WU, Y. (1993). Comparison of EWMA, CUSUM and Shiryayev–Roberts procedures for detecting a shift in the mean. *Ann. Statist.* **21** 645–670.

WU, Y. and SRIVASTAVA, M. S. (1993). Dynamic sampling plans in on-line control charts. *Canad. J. Statist.* **21** 409–419.

WEATHERALL, J. A. C. and HASKEY, J. C. (1976). Surveillance of malformations. *British Medical Bulletin* **32** 39–44.

DEPARTMENT OF STATISTICS
UNIVERSITY OF TORONTO
TORONTO, ONTARIO
M5S 1A1, CANADA

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305