BIBLIOGRAPHY

Alt, F. B. (1976). "Small Sample Probability Limits for the Mean of a Multivariate Normal Process". *ASQC Technical Conference Transactions*, pp. 170-176.

Alt, F. B. (1985). "Multivariate Quality Control" in *Encyclopedia of Statistical Sciences* 6, edited by S. Kotz and N. L. Johnson, John Wiley & Sons, New York, NY.

IMSL STAT/LIBRARY (1987). IMSL STAT/LIBRARY User's Manual, Vol. 1-3, IMSL Inc., Houston, TX.

Montgomery, D. C. (1996). *Introduction to Statistical Quality Control*. John Wiley & Sons, New York, NY.

Nedumaran, G. and Pignatiello, J. J., Jr. (1999). "On Constructing T^2 Control Charts for On-line Process Monitoring". *IIE Transactions* 31, pp. 529-536.

Seo, T. and Siotani, S. (1993). "Approximations to the Upper Percentiles of T^2_{max}-type Statistics" in *Statistical Science & Data Analysis*, edited by K. Matsusita et al., VSP, Tokyo.

Siotani, M. (1959). "The Exact Value of the Generalized Distances of the Individual Points in the Multivariate Normal Sample". *Annals of the Institute of Statistical Mathematics* 10, pp. 183-203.

Received May, 1997; Revised January, 2000.

COMMUN. STATIST.—SIMULA., 29(2), 633–646 (2000)

SOME POWER ASPECTS OF METHODS FOR DETECTING DIFFERENT SHIFTS IN THE MEAN

E. Järpe P. Wessman

Department of Statistics, Göteborg University,
SE 405 30 Göteborg, Sweden

Key words: Surveillance, change point, false alarm, expected delay, predictive value.

ABSTRACT

We study, by means of simulations, the performance of the Shewhart method, the Cusum method, the Shiryaev-Roberts method and the likelihood ratio method in the case when the true shift differs from the shift for which the methods are optimal. The methods are compared for a fixed expected time until false alarm. The comparisons are made with respect to some measures associated with power such as probability of alarm when the change occurs immediately, expected delay of true alarm and predictive value of an alarm.

1 INTRODUCTION

Statistical surveillance is needed to detect changes in random processes in e.g. environmetrics, manufacturing industry, econometry, biometry, etc. Suppose that a random process is observed at discrete time points. Each random variable is normally distributed with zero mean and unit variance. At a random time point, τ, the mean shifts to a new level, μ, but there is no shift

Copyright © 2000 by Marcel Dekker, Inc. www.dekker.com

in variance. We are interested in situations where the methods for detecting the shift are optimal for one specified size of shift. We determine the power aspects for different shift sizes of the process.

Srivastava and Wu (1993) made a similar kind of comparison study of several well known surveillance methods, the Exponentially weighted moving average (EWMA), the Cusum and the Shiryaev-Roberts methods. They considered the case when the size of the shift of the stochastic process is unknown while the shift for which the methods are specified is fixed. However, there are several differences between their study and our study. We treat the case of active surveillance, i.e. when the surveillance stops as soon as the first alarm is signalled. Srivastava and Wu studied passive surveillance, when a number of false alarms have been signalled before the change. Furthermore, they considered continuous time while we consider discrete time (i.e. time is present only through a countable set of observations made at discrete time points). In our study the EWMA method is not considered. Because of some debate about the proper choice of the parameter λ in the EWMA method (see Frisén, 1999), it needs a special treatment. But the likelihood ratio (LR) method, discussed by Frisén and de Maré (1991), is included.

The measures of performance we use here are *expected time until alarm given that a change happens immediately* (ARL^1): $\mathsf{E}[t_A \mid \tau = 1]$, *the distribution of the time of alarm given that a change happens immediately*: $\mathsf{P}[t_A = t \mid \tau = 1]$, *the expected delay of true alarm when a change occurs at time t* (CED): $\mathsf{E}[t_A - \tau \mid t_A > \tau = t]$ and *the total expected delay of true alarm when τ is geometrically distributed* (ED): $\mathsf{E}_\nu[t_A - \tau \mid t_A > \tau]$. Srivastava and Wu used *the stationary average delay time* (SADT): $\lim_{t\to\infty} \mathsf{E}[t_1 + \ldots + t_{n+1} - \tau \mid \tau = t]$ where $t_1, \ldots, t_n$ are time intervals between n false alarms before τ, and t_{n+1} is the interval between the last false alarm and the first alarm after change. The main difference between SADT, CED and ED is that SADT is the limit expected delay when the time of change $\tau = t$ tends to infinity while CED and ED are measures of expected delay without this limit.

Frisén and Wessman (1999) made a study of differences between and robustness of some methods for surveillance based on the likelihood ratio when there is a shift in the mean of a sequence of normally distributed random variables. They made a simulation study of a unit shift in mean when the methods were specified for a 0.5, 1 and 2 shift in mean having fixed the ARL^0 so that ARL^0 was the same for all methods. The comparisons were made with respect to the same measures of performance as in this paper.

The main differences were shown to be between (the Shewhart method, the Cusum method) and (the Shiryaev-Roberts method, the LR method) for $\mathrm{ARL}^0 = 11$. For $\mathrm{ARL}^0 = 100$ the properties of the Cusum method were more similar to those of the LR method and the Shiryaev-Roberts method. They concluded that increasing specified size of shift makes the properties of all considered surveillance methods tend to the Shewhart method. This means that an increased weight is put on the latest observation. The properties of the LR method considered are well approximated by those of the Shiryaev-Roberts method for most cases studied here.

This paper is organized as follows. In Section 2 the statistical model and the surveillance methods are described. In Section 3 the properties of distribution of true alarm, average run length, expected delay and predictive value are examined for some surveillance methods, by means of simulations. An example of the use of the results for spatial surveillance is given in Section 4 and in Section 5 the results are discussed.

2 Methods

We consider the problem of shift in the mean of a normal distribution. Suppose that we have a sequence of random variables $\{X(t) : t \in \mathbb{N}\}$ and a random time of change τ. We also assume that, given the change-point τ, all variables, $X(1), X(2), X(3), \ldots$, are conditionally independent. We assume that

$$X(t) \overset{\mathcal{D}}{=} \begin{cases} \mathrm{N}(0,1) & \text{if } t < \tau \\ \mathrm{N}(\mu,1) & \text{if } t \geq \tau \end{cases}.$$

In this paper we consider the shift sizes $\mu \in \{0.5, 1, 2\}$.

Some measures of performance include specification of the distribution of τ. In these cases τ is geometrically distributed with an intensity parameter $\nu(t) = \nu$. In general τ has a frequency function, $\pi_t = \mathsf{P}[\tau = t]$, with intensity $\nu(t) = \mathsf{P}[\tau = t | \tau \geq t]$.

A surveillance method is a stopping rule. In this paper the stopping rules considered may be represented as

$$t_A = \min\{s : p(X_s) > K\}$$

where X_s denotes the history of X up to time s, $\{X(t) : t \leq s\}$, $p(\cdot)$ is called an alarm function and K is a threshold (sometimes called critical limit). See e.g. Lai (1995) for a more thorough presentation of surveillance methods.

Let $\mathrm{lr}(t)$ denote the likelihood ratio $f(X(t) \mid \tau \leq t)/f(X(t) \mid \tau > t)$. Thus, for a shift of size M we have

$$\mathrm{lr}(t) \;=\; \exp\Big(M(X(t) - M/2)\Big).$$

The surveillance methods considered here can all be defined by using $\mathrm{lr}(t)$. The **Shewhart method** (Shewhart, 1931) is defined by writing the alarm function as

$$p(X_s) = \mathrm{lr}(s) \;\text{ for all } s = 1, 2, 3, \ldots$$

The **Cusum method** is the surveillance method suggested by Page (1954) and presented by Lorden (1971) with an alarm function

$$p(X_s) = \begin{cases} \log \mathrm{lr}(1) & \text{when } s = 1 \\ \max(0\,,\, \log \mathrm{lr}(s) + p(X_{s-1})) & \text{when } s = 2, 3, 4, \ldots \end{cases}$$

The **Shiryaev-Roberts method**, derived by Shiryaev (1963), has an alarm function

$$p(X_s) = \begin{cases} \mathrm{lr}(1) & \text{when } s = 1 \\ \mathrm{lr}(s)\,(p(X_{s-1}) + 1) & \text{when } s = 2, 3, 4, \ldots \end{cases}$$

The **(full) likelihood ratio (LR) method** is the surveillance method presented by Frisén and de Maré (1991)

$$t_A = \min\{s : p(X_s) > k_s\}$$

where $k_s = K\mathsf{P}[\tau > s]/\mathsf{P}[\tau \leq s]$ and

$$p(X_s) = \sum_{u=1}^{s} \frac{\pi_u}{\mathsf{P}[\tau \leq s]} \prod_{w=u}^{s} \mathrm{lr}(w) \quad \text{for all } s = 1, 2, 3, \ldots$$

This is equivalent to

$$t_A = \min\{s : q(X_s) > K\}$$

where K is a constant with respect to s and

$$q(X_s) = \sum_{u=1}^{s} \frac{\pi_u}{1 - \sum_{w=1}^{s} \pi_w} \prod_{w=u}^{s} \mathrm{lr}(w) \quad \text{for all } s = 1, 2, 3, \ldots$$

When not otherwise stated, the LR method used is optimized for $\nu = 0.1$.

Throughout the paper all surveillance methods considered are optimized for the shift in the mean of size $M = 1$ and no shift in variance which is unity. The values of μ of the processes studied are $\mu \in \{0.5, 1, 2\}$ and $\sigma = 1$.

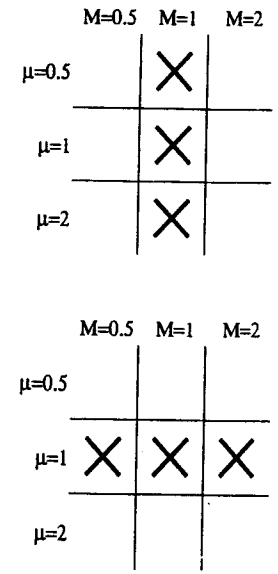

Figure 1: *To the left: the cases of this study. To the right: the cases of the study by Frisén and Wessman (1999).*

Choosing a low value for the threshold K makes the alarm function $p(X_s)$ more likely to exceed K and thus the stopping rule more likely to signal alarm soon after a change has occurred but also more prone to give false alarms. Choosing a high value will give fewer false alarms but also a longer delay of true alarm.

This study is intended as a complement to the study by Frisén and Wessman (1999). There $\mu = 1$ and the shift size for which the surveillance methods were optimized $M \in \{0.5, 1, 2\}$ while in our study the surveillance methods are optimized for $M = 1$. This is illustrated in Figure 1.

3 Results

For the Shewhart method, the run length distribution $\mathsf{P}[t_A = t \mid \tau = t]$ was analytically calculated while, for the other surveillance methods, simulations

were used to approximate the run length distribution. 10^7 replicates of the run length were simulated for $\tau = 1, \ldots, 150$. In these simulations standard normal random numbers were generated by the NAG subroutine g05ddf. To make the plots comparable, the thresholds K were chosen so that ARL^0=100 for all surveillance methods. These thresholds were determined with a level of accuracy which made the deviation between the intended and estimated ARL^0 less than 0.1% of the intended value. The large number of replicates made it possible to neglect the sampling error from this determination.

3.1 Probability of false alarm

Even though the surveillance methods are calibrated so that $\mathrm{ARL}^0 = 100$, the false alarm probability, $\mathsf{P}[t_A < \tau]$, differs between the methods. It can be thought of as a characteristic for surveillance corresponding to the level of significance for hypothesis testing. The false alarm probability is not a function of the shift size μ. Thus, the results by Frisén and Wessman (1999) apply also here.

3.2 Delay of true alarm

In the same way that the false alarm probabilities correspond to the level of significance, characteristics involving events concerned with alarm when the change has occurred, such as the conditional probability that alarm is signalled at time t given that $\tau \leq t$ and the expected delay of true alarm, can be said to correspond to the power of a test. Figure 2 shows the run length distribution $\mathsf{P}[t_A = t \mid \tau = 1]$ for shift sizes $\mu = 0.5, 1, 2$.

In Figure 3 we see the conditional expected time of delay of an alarm given that the change has already happened at time t, i.e. $\mathsf{E}[t_A - \tau \mid t_A > \tau = t]$.

The expected time until alarm when the change has already occurred when surveillance starts, $\mathsf{E}[t_A \mid \tau = 1]$, is usually denoted by ARL^1. This has to do with the *delay of true alarm* in a similar way as ARL^0 has to do with *false alarm*. The limits of $1 + \mathsf{E}_\nu[t_A - \tau \mid t_A > \tau]$, as the intensity ν tends to 1, are the values of ARL^1. The expected delay, $\mathsf{E}_\nu[t_A - \tau \mid t_A > \tau]$, is shown in Figure 4 as a function of ν.

As can be seen in Figure 3 and Figure 4, the Shewhart method has a larger conditional expected delay (CED) and expected delay (ED) than the other methods when $(\mu, M) = (0.5, 1)$ and when $(\mu, M) = (1, 1)$. For $(\mu, M) = (2, 1)$ the Shewhart method performs better and is comparable to the others with respect to CED and ED.

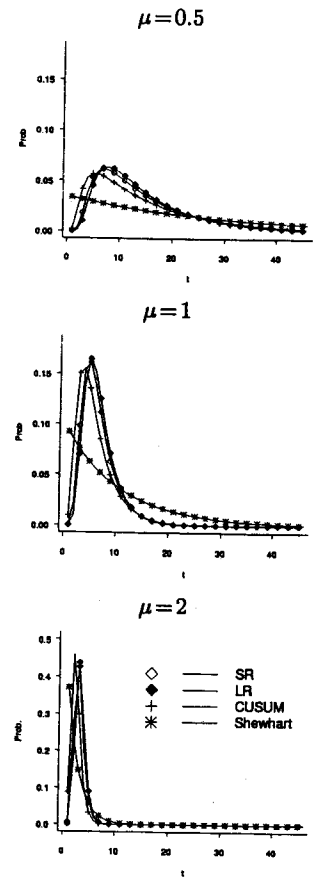

Figure 2: *Probabilities of alarm at time t, given that* $\tau = 1$.

In Table 1 the numerical values of expected delay have been tabulated for $\nu = 0.1, 0.25, 0.5, 0.75, 0.9$. For the Shewhart method the expected delay is $\mathrm{ARL}^1 - 1$, regardless of ν. The differences in ED between all except the Shewhart method are small. The LR method and the Shiryaev-Roberts method perform marginally better for $\mu = 0.5$ and Cusum for $\mu = 1$ and $\mu = 2$.

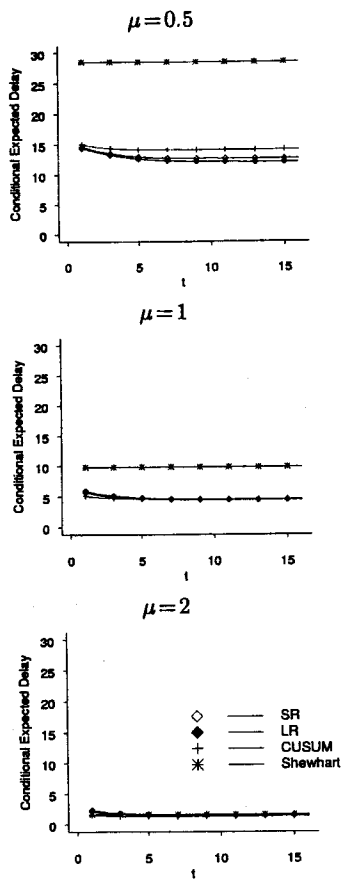

Figure 3: *Conditional expected delay given that the change occurs at time t.*

3.3 Predictive value

Another power aspect on surveillance methods is the predictive value, $PV(t) = \mathsf{P}_\nu[\tau \le t \,|\, t_A = t]$, which measures the credibility of an alarm at time t. In Figure 5 the predictive value is plotted having specified τ to have geometrical distribution with intensity $\nu = 0.1$. From Figures 2 and 5, one

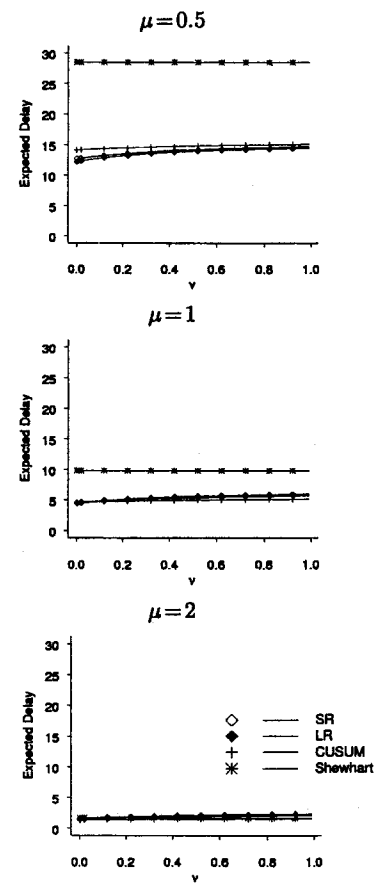

Figure 4: *Expected delay as a function of the intensity ν.*

may conclude that high probability of fast detection is penalized by low credibility of alarms at an early stage.

4 Spatial surveillance

A new area, where the results presented in the previous section can be used, is spatial surveillance. For example, one may want to detect a change of

Table 1: *Expected delay when the intensity is* $\nu=0.1, 0.25, 0.5, 0.75, 0.9$.

Method	$\mu \backslash \nu$	0.10	0.25	0.50	0.75	0.90
Shewhart	0.5	28.50	28.50	28.50	28.50	28.50
Cusum		14.33	14.55	14.81	14.99	15.07
LR		12.76	13.36	13.94	14.27	14.40
SR		13.17	13.67	14.19	14.49	14.62
Shewhart	1	9.83	9.83	9.83	9.83	9.83
Cusum		4.68	4.79	4.93	5.03	5.08
LR		4.83	5.19	5.56	5.78	5.87
SR		4.71	5.02	5.34	5.55	5.64
Shewhart	2	1.69	1.69	1.69	1.69	1.69
Cusum		1.39	1.43	1.49	1.54	1.56
LR		1.74	1.93	2.14	2.27	2.33
SR		1.58	1.75	1.93	2.05	2.11

levels of gamma radiation in a specified geographical region. This is one situation when spatial dependencies might need to be taken into account. Järpe (1999) made a study of a surveillance situation in the Ising model. Figure 6 shows two simulated Ising patterns.

It was shown that, in this case, the spatial surveillance problem considered could be reduced to an ordinary univariate surveillance problem. In the Ising model, a certain amount of shift in the interaction parameter ϕ in an $n \times n$-lattice corresponds to a shift in mean of a statistic sufficient for this interaction parameter. A Shewhart chart based on the sufficient statistic for the changing parameter, ϕ, was used. In the Ising model $\phi=0$ corresponds to no interaction, $\phi<0$ to attraction and $\phi>0$ to repulsion. The performance was evaluated by the expected delay. For example a shift from $\mu=0$ to $\mu=0.5$ corresponds to a larger shift of the interaction parameter ϕ in a 5×5-lattice than in a 6×6-lattice. Supposing that ϕ changes from $\phi_0=0$ to ϕ_1 in an $n\times n$-lattice, Table 2 gives the values of CED which correspond to a change from $\mu=0$ to $\mu=0.5, 1, 2$ for lattice sizes $n=5, 7, 10$ when the LR method is specified for a shift in mean of size 1.

5 Discussion

Since Shiryaev-Roberts method is the limit method of the likelihood ratio

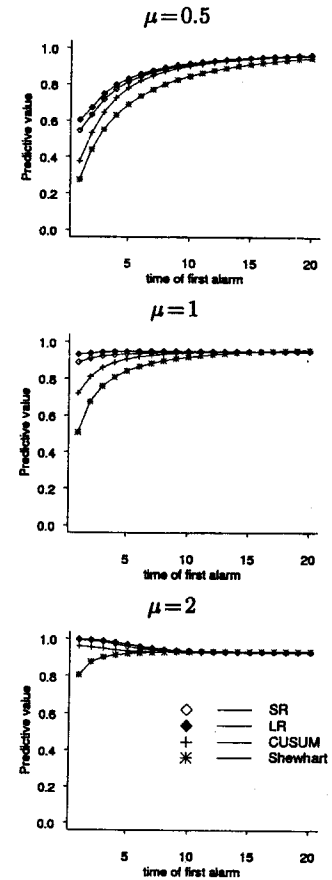

Figure 5: $\mathrm{PV}(t) = \mathsf{P}[\tau \le t \mid t_A = t]$ *when* τ *is geometrically distributed with intensity* $\nu=0.1$.

method when the ν, for which it is optimized, tends to 0, these methods are quite similar as can be expected from the study by Frisén and Wessman (1999).

Srivastava and Wu (1993) considered the problem of a Brownian motion with drift $\mu \in \{0.5, 1, 1.5, 2\}$, diffusion $\sigma=1$ and methods optimized for drift

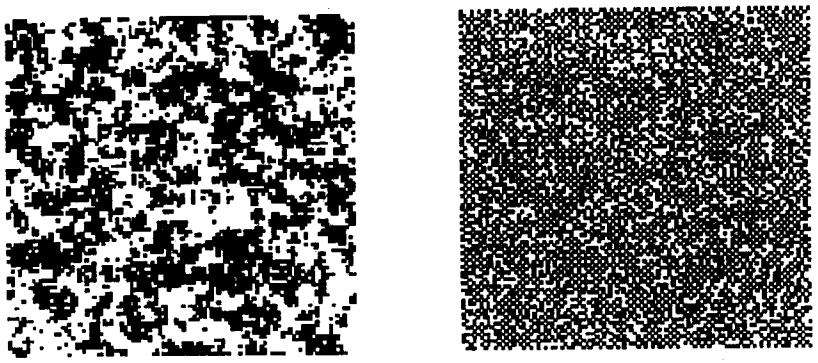

Figure 6: *To the left: a simulation of an attractive Ising pattern with* $\phi=-0.8$. *To the right: a simulation of a repulsive Ising pattern with* $\phi=0.8$.

Table 2: *Examples of conditional expected delay of the LR method for the Ising model.*

n	ϕ_1	μ	$\mathsf{E}_{\phi_1}[t_A-\tau \mid t_A>\tau=t]$ $t=1$	$t=15$
5	−0.142	0.5	14.470	12.077
	−0.274	1	5.925	4.441
	−0.274	2	2.360	1.535
7	−0.101	0.5	14.470	12.077
	−0.199	1	5.925	4.441
	−0.380	2	2.360	1.535
10	−0.071	0.5	14.470	12.077
	−0.142	1	5.925	4.441
	−0.274	2	2.360	1.535

1 and diffusion 1 where ARL^0 was fixed 100 and 500. The comparisons were made with respect to the stationary average delay time (SADT). The closest correspondence to SADT in this paper would be the limit expected delay as ν tends to zero

$$\lim_{\nu\to 0} \mathsf{E}_\nu[t_A-\tau \mid t_A>\tau]$$

(where τ is assumed geometrical with intensity ν) and the limit conditional expected delay given that change occurs at time t as t tends to infinity

$$\lim_{t\to\infty} \mathsf{E}[t_A-\tau \mid t_A>\tau=t].$$

In the study by Srivastava and Wu $\mathrm{SADT}_{\mathrm{Cusum}} > \mathrm{SADT}_{\mathrm{SR}}$ when the shift size is 0.5 and $\mathrm{ARL}^0=100$ which coincides with our result for both expected delay and conditional expected delay. In the Srivastava and Wu study, the Shiryaev-Roberts method is the best method with respect to SADT for all cases $\mu=0.5, 1, 1.5, 2$ and $\nu=0.1, 0.25, 0.5, 0.75, 0.9$. In our study, however, the Cusum method is better with respect to expected delay for $\mu=1,2$. (In the conditional expected delay respect no such conclusion is perceivable for $\mu=1,2$.) The differences are small and may depend on the fact that expected delay, conditional expected delay and SADT are different measures of performance.

As expected, the power performance of all methods is worse when a smaller shift than specified occurs. The loss of performance can be large as seen in Figures 3 – 5. For example the differences of CED and ED between $\mu=0.5$ and $\mu=1$ more than doubles regardless of τ and t respectively. However, in absolute time the deterioration in performance in this respect is more pronounced for the Shewhart method. Also the $\mathrm{PV}(t)$ for early alarm decreases as one moves from $(\mu, M)=(0.5, 1)$ to $(\mu, M)=(1, 1)$. For $(\mu, M)=(2, 1)$ when the shift is larger than specified the methods perform better than for the case they are optimized for. This is especially true for the Shewhart method; only for $(\mu, M)=(2, 1)$ it has CED and ED values comparable with the other methods considered.

ACKNOWLEDGEMENTS

We would very much like to thank Professor Marianne Frisén for introducing us to surveillance and for essential discussions. We also thank Associate Professor Aila Särkkä for many helpful comments on an earlier version of this paper.

This work was supported by the Swedish Radiation Protection Institute and the Swedish Council for Research in the Humanities and Social Sciences.

REFERENCES

Frisén, M. (1999), Characterisation of Methods for Surveillance by Optimality, *Research Report 1999:13*, Göteborg University, Göteborg.

Frisén, M. and de Maré, J. (1991), Optimal Surveillance, *Biometrika*, vol. 78, pp 271–280.

Frisén, M. and Wessman, P. (1999), Evaluations of Likelihood Ratio Methods for Surveillance, Differences and Robustness, *Communications in Statistics, Computation and simulation*, vol. 28., nr. 3, pp 597–622.

Järpe, E. (1999), Surveillance of Spatial Patterns, Change of interaction in the Ising Model, to appear in *Communications in Statistics, Theory and Methods*, vol. 28., nr. 12.

Lai, T.L. (1995), Sequential Changepoint Detection in Quality Control and Dynamical Systems, *Journal of the Royal Statistical Society*, series B, vol. 57, pp 613–658.

Lorden, G. (1971), Procedures for Reacting to a Change in Distribution, *Annals of Mathematical Statistics*, vol. 42, pp 1897–1908.

Page, E.S. (1954), Continuous Inspection Schemes, *Biometrika*, vol. 41, pp 100–115.

Shewhart, W.A. (1931), Economic Control of Quality Control, *Reinhold Company*, Princeton N.J.

Shiryaev, A.N. (1963), On Optimum Methods in Quickest Detection Problems, *Theory of Probability and its Applications*, vol. 8, pp 28–46.

Srivastava, M.S. and Wu, Y. (1993), Comparison of some EWMA, CUSUM and Shiryaev-Roberts Procedures for Detecting a Shift in the Mean, *Annals of Statistics*, vol. 21, pp 645–670.

Received April, 1999; Revised December, 1999.

COMMUN. STATIST.—SIMULA., 29(2), 647–665 (2000)

CONFIDENCE INTERVALS FOR THE KAPPA PARAMETER, WITH APPLICATION TO THE SEMICONDUCTOR INDUSTRY

Rachelle C. Wilkinson, G. Bruce Schaalje and Bruce Jay Collings

Brigham Young University
Department of Statistics
230 TMCB
Provo, UT 84602

Key Words and Phrases: kappa statistic; measures of agreement for categorical data; bootstrap; bias-corrected bootstrap; square root transformation

ABSTRACT

In the semiconductor industry, test tapes are used to classify microprocessors into one of several functional categories before packaging. Test tapes must be changed frequently, and so it is necessary make decisions about whether the new tape performs similarly enough to the old tape. The kappa statistic ($\hat{\kappa}$), a chance-corrected measure of agreement for categorical data, would be useful for this purpose if a reliable confidence interval procedure were available. Unfortunately, confidence intervals for κ have not been investigated for applications in which the true value is expected to be close to 1, more than two categories are involved, and marginal distributions are not uniform. Simulation was used to study properties of 5 confidence interval procedures for κ under these conditions. Confidence intervals were compared based on average confidence interval length and coverage rate. Although no confidence interval method consistently performed better than the rest, the bias-corrected bootstrap method generally performed well.

Copyright © 2000 by Marcel Dekker, Inc. www.dekker.com