Univariate and Multivariate Surveillance Methods for Detecting Increases in Incidence Rates

Michael D. Joner, Jr.

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Statistics

William H. Woodall, Co-Chair
Marion R. Reynolds, Jr., Co-Chair
Jeffrey B. Birch
Dan J. Spitzner
G. Geoffrey Vining

March 30, 2007
Blacksburg, Virginia

KEYWORDS: Bernoulli Control Charts, Disease Surveillance, Multivariate Control Charts, One-sided Control Charts, Prospective Monitoring, Scan Statistic, Spatial Correlation.

Univariate and Multivariate Surveillance Methods for Detecting Increases in Incidence Rates

Michael D. Joner, Jr.

ABSTRACT

It is often important to detect an increase in the frequency of some event. Particular attention is given to medical events such as mortality or the incidence of a given disease, infection or birth defect. Observations are regularly taken in which either an incidence occurs or one does not. This dissertation contains the result of an investigation of prospective monitoring techniques in two distinct surveillance situations. In the first situation, the observations are assumed to be the results of independent Bernoulli trials. Some have suggested adapting the scan statistic to monitor such rates and detect a rate increase as soon as possible after it occurs. Other methods could be used in prospective surveillance, such as the Bernoulli cumulative sum (CUSUM) technique. Issues involved in selecting parameters for the scan statistic and CUSUM methods are discussed, and a method for computing the expected number of observations needed for the scan statistic method to signal a rate increase is given. A comparison of these methods shows that the Bernoulli CUSUM method tends to be more effective in detecting increases in the rate. In the second situation, the incidence information is available at multiple locations. In this case the individual sites often report a count of incidences on a regularly scheduled basis. It is assumed that the counts are Poisson random variables which are independent over time, but the counts at any given time are possibly correlated between regions. Multivariate techniques have been suggested for this situation, but many of these approaches have shortcomings which have been demonstrated in the quality control literature. In an attempt to remedy some of these shortcomings, a new control chart is recommended based on a multivariate exponentially weighted moving average. The average run-length performance of this chart is compared with that of the existing methods.

Acknowledgments

Dr. William H. Woodall and Dr. Marion R. Reynolds, Jr., were instrumental in helping me learn how to research and carry out research plans. I had many conversations with them as we went through the process of looking for research projects of statistical and practical significance. Their subject-specific expertise was very helpful when preparing publications and presentations. I appreciate their understanding of the various situations I encountered while studying at Virginia Tech. They watched as I made the occasional mistakes in my research and in my studies and then guided me through the steps necessary to fix those errors. I thank them for their encouragement to openly share my thoughts and ideas and for their patience as I worked on this dissertation.

I also owe significant thanks to Dr. Jeffrey B. Birch and Dr. James D. Williams. They were the main reasons I was attracted to Virginia Tech. They took considerable interest in my progress and provided a great deal of encouragement throughout my time here. I could always rely on their optimism to make any day better.

I made extensive use of Dr. Landon H. Sego's computer code to evaluate the performance of the Bernoulli CUSUM chart. This was particularly useful when comparing that chart with the scan statistic method. I also appreciate the friendship of Dr. Williams, Dr. Sego and Dr. Willis A. Jensen. We accomplished a great deal together as a team.

Dr. Ronald D. Fricker provided some spatial surveillance data which was helpful in learning some of the limitations of the spatial method presented here.

Dr. Woodall secured some of the funding for this research in the form of National Science Foundation Grant DMI-0354859. I appreciate this funding, as well as the financial support provided by the university and the department.

Finally, I am especially grateful for the Lord's help in this work. There were many times over the past years that I was inspired to know how to approach a certain problem. His support helped me fill in a lot of the details of this work.

— Michael D. Joner, Jr.

Contents

List of Figures

List of Tables

Glossary of Acronyms

ANOS	Average Number of Observations to Signal	10
ARL	Average Run Length	64
CUSCORE	Cumulative Score	10
CUSUM	Cumulative Sum	3
EWMA	Exponentially Weighted Moving Average	59
MC1	Pignatiello and Runger's Multivariate CUSUM (variant 1)	3
MCUSUM	Crosier's Multivariate CUSUM	59
MEWMA	Multivariate Exponentially Weighted Moving Average	4
SSANOS	Steady-state Average Number of Observations to Signal	12
SSARL	Steady-state Average Run Length	68
SSMNOS	Steady-state Median Number of Observations to Signal	23

Common Notation

Y_i	The outcome of the ith Bernoulli trial.
p	The probability of success for a particular Bernoulli trial.
p_0	The in-control probability of success for a Bernoulli trial.
p_1	The smallest out-of-control probability of success for a Bernoulli trial which is important to detect quickly.
γ	The multiplier of p_0 which gives p_1.
S_i	The scan statistic after Y_i is observed.
m	The number of consecutive opportunities considered by the scan statistic.
k	The number of incidences in the past m opportunities required for the scan statistic chart to signal.
C_i	The CUSUM chart statistic after Y_i is observed.
r	The reference value parameter for the CUSUM or MC1 chart.
h	The alarm limit parameter for the CUSUM or MC1 chart.
X_j	The Shewhart np chart statistic after the jth group of n consecutive Y_i values.
ANOS_0	The ANOS when $p = p_0$.
SSANOS_1	The SSANOS when $p = p_1$.
c	A minimum ANOS_0 value which must be exceeded.
$se(\text{var})$	The standard error of the specified variable.
d	The duration of a transient shift in the incidence rate.
$X_{i,t}$	The count of incidences in region i at time period t.
$\mathbf{X}_t$	The vector of regional counts at time period t.
$\boldsymbol{\mu}_0$	The vector of in-control mean regional counts.
$\boldsymbol{\mu}_1$	A vector of out-of-control mean regional counts which is important to detect quickly.
n	The number of regions.

$\mathbf{\Sigma}$	A covariance matrix of dimension $n \times n$.
$C_{i,t}$	The CUSUM chart statistic for region i after $X_{i,t}$ is observed.
$\mathbf{C}_t$	The vector of CUSUM chart statistics after $\mathbf{X}_t$ is observed.
$MC1_t$	The MC1 chart statistic value after $\mathbf{X}_t$ is observed.
$\mathbf{Z}_t$	The vector of EWMA chart statistics after $\mathbf{X}_t$ is observed.
λ	The smoothing parameter for the EWMA chart.
$\mathbf{\Sigma}_{\mathbf{Z}_t}$	The covariance matrix of $\mathbf{Z}_t$.
MEW_t	The MEWMA chart statistic after $\mathbf{X}_t$ is observed.
q	The alarm limit parameter for the MEWMA chart.
$\boldsymbol{\mu}$	The vector of actual (but unknown) mean regional counts.
μ	The expected regional count for every region, or equivalently, the size of the population multiplied by p.
ρ	The correlation between any two regions that share a common edge when rooks' case adjacency is assumed.

Chapter 1

General Introduction

Researchers and scientists often monitor the frequencies of events. Public health data, for example, may be available soon after each incidence of a disease is observed or they may be available only at the end of some pre-determined reporting period. In the case of the former, the observations are assumed to be the results of independent Bernoulli trials, in that each observation will be an incidence or it will not. In the latter case, it is assumed that the total number of incidences will be reported at regular intervals of time (as opposed to after some number of observations), and the summaries are assumed to be Poisson-distributed and independent over time.

It is possible that data collected in either manner (near-immediate or on a regular schedule) may also specify the spatial location of the event. For example, observations may be taken at several locations (and therefore on different populations). In the case of data taken on humans and animals, the exact location of the official residence or location of the subject might be available. Therefore, the observations could be reported with exact time and location information or could be aggregated to some extent in time and/or space. In some cases the location information might not be used at all.

In this dissertation, particular attention is given to medical surveillance, although the methods presented here could be used in other applications. The following are some examples

of medical events that could be monitored using the methods to be considered. Heffernan *et al.* (2004) considered reports made by emergency rooms throughout New York City on patients arriving with respiratory, fever, diarrhea, and vomiting symptoms. The data were reported to the city health department on a daily basis. Thacker *et al.* (1996) considered monitoring occurrences of birth defects and diagnoses of asthma, both at individual locations and as reported to public health agencies.

An increase in the rates of these types of medical events represents an increased risk to the public. It is therefore important to rapidly detect any increase. When it appears likely that an increase has occurred, epidemiologists and other interested parties could then conduct a more in-depth investigation to determine the cause of the increase and, if necessary, take action to control or remedy the situation. Therefore, the methods studied and proposed in this dissertation are specifically intended to detect rate increases.

Surveillance methods can be used retrospectively or prospectively to detect whether there has been an increase in the rate of interest over time. In retrospective surveillance, a past sequence of data is examined to determine whether there has been an increase in the incidence rate sometime in the past. In prospective surveillance, new information is used as it becomes available to help detect any rate increase as soon as possible after it occurs. In the public health literature, the prospective methods are not as developed as the retrospective methods. Farrington and Beale (1998), Sonesson and Bock (2003), and Woodall (2006) have reviewed some of the surveillance methods currently in use.

In prospective analyses it is relatively easy to detect large increases in rates. It is harder to detect increases of smaller magnitude, however, unless one is willing to increase the probability of indicating an increase when no such increase has occurred (that is, a false alarm). In industrial applications, control charts are frequently used to detect changes in rates. The performance of all commonly used control charts has been thoroughly studied, for both the "in-control" situation when no change has occurred, and for the "out-of-control" situation when some specified change has occurred. Woodall (1997) reviewed the literature for control charts based on count data. Much more work has appeared since.

1.1 Research purpose and outline

It is clear that there are several methods which could be used to detect increases in incidence rates. Appropriate methods will vary depending on the underlying data collection system. The purpose of this dissertation is to evaluate and compare appropriate methods under two of these systems. In Chapter 2 the observations are assumed to be binary and their realizations are recorded as soon as they are available. In Chapter 3 the observations are assumed to be the total number of events over some recording period, reported from several locations at the conclusion of regular reporting cycles. In each case some of the appropriate monitoring techniques will be presented. Also, issues in implementation of these techniques will be discussed before making comparisons of performance.

In journals dedicated to statistical applications in medicine, the scan statistic has been recommended as a prospective monitoring tool (for example, Naus and Wallenstein, 2006; Ismail *et al.*, 2003). Scan statistic methods have seen little use in industrial applications, and little is known about the performance of these methods, especially as compared to control chart methods. In Chapter 2, the (temporal-only) scan statistic is given, issues relevant to its usage in a surveillance system are discussed, and comparisons to the Bernoulli cumulative sum (CUSUM) control chart developed by Reynolds and Stoumbos (1999) are given. The Bernoulli CUSUM method is found to perform better than the scan statistic based method in most circumstances. However, the difference in performance is not large, and thus the use of either method is defensible.

In the medical-statistical literature, control charting methods are discussed and recommended more frequently when a spatial correlation structure is present. Rogerson and Yamada (2004) review and compare two CUSUM-based control charts. One of these, a system of one-sided univariate CUSUM charts, ignores spatial correlation structures. However, since it is one-sided, it is designed to detect only increases in the regional incidence rates being monitored. The other CUSUM-based method is the MC1 chart presented by Pignatiello and Runger (1990). This chart can be used with a spatial correlation structure, but

detects either increases or decreases in the incidence rates. Another control charting method that could be applied when a spatial correlation structure is present is the multivariate exponentially weighted moving average (MEWMA) control chart of Lowry *et al.* (1992). In Chapter 3, the MEWMA chart is modified to enhance the detection increases in one or more spatial regions, regardless of the behavior of the rates in other regions. This new "one-sided" MEWMA chart is then compared against the CUSUM-based control charts and found to be superior, regardless of the spatial correlation structure considered.

Finally, Chapter 4 contains a brief summary and conclusion. Some issues in medical surveillance and some topics for future work are also given.

Chapter 2

Detecting a Rate Increase Using Bernoulli Scan and CUSUM Statistics

2.1 Introduction

The problem of monitoring Bernoulli rates arises frequently in public health and medical applications. Two examples are monitoring the rate of positive diagnoses (e.g. the rate of a congenital malformation), and monitoring the rate of surgical failures (e.g. the mortality rate after a certain surgery). These diagnostics or surgeries are repeatedly performed, and at each repetition (or *opportunity*), an *observation* is taken. The outcome at the ith repetition, Y_i, is said to be an incidence if an event of interest occurs, such as a congenital malformation or a surgical mortality. If the outcome is an incidence, then $Y_i = 1$; otherwise $Y_i = 0$. Let p be the incidence rate (the probability that an incidence occurs), and assume independence of the Bernoulli trials over time.

The rate, p, is assumed to be constant over time while the process is in control with value p_0. The goal of monitoring is to detect an increase in p from p_0 to a significantly larger rate. It is possible to search for the "best" chart if it is known that an increase to a specific rate, $p_1 = \gamma p_0$, where $\gamma > 1$, is most important.

In some situations the in-control probability is not constant from observation to observation. As an example, this probability can be affected in a surgical context by patient characteristics (they can have very different medical histories), methodological considerations (certain surgical techniques may influence the outcome of the surgery), and personnel changes (surgeons have different assistants). In these situations a risk adjustment method is essential. Such risk adjustment methods have been discussed elsewhere (see Grigg and Farewell, 2004; Steiner *et al.*, 2000; Sego *et al.*, 2007*a*).

The Bernoulli scan statistic has been used often in the retrospective surveillance literature. This statistic is the maximum number of incidences occurring within any m consecutive opportunities. All $N - m + 1$ possible subsets of m consecutive opportunities taken from a larger set of $N > m$ consecutive opportunities need to be considered. In this context, the scan statistic is retrospective, as it is assumed that all of the necessary N observations have already been collected before the scan statistic is computed. Numerous applications and examples in a variety of fields, as well as further discussion of the properties of scan statistics (particularly when they are used retrospectively), are given by Glaz *et al.* (2001) and Balakrishnan and Koutras (2002). Recently, Naus and Wallenstein (2006) and Ismail *et al.* (2003) have recommended using the scan statistic for prospective surveillance.

Woodall *et al.* (2007) have discussed some of the issues that arise when the modified scan statistic method of Kulldorff (2001) is used prospectively. One issue is that Kulldorff's method considers multiple scan statistics based on several values of m. This chapter only considers the traditional scan statistic with a single value of m. Using multiple values of m would require investigation of several issues, the most important concerning the ability to control the overall false alarm rate while maintaining the detection power of the chart.

It is well known in the quality control literature (e.g., Hawkins and Olwell, 1998) that a properly designed CUSUM control chart is optimal for detecting a specified sustained change in a parameter of interest under a "worst-case scenario." That is, the CUSUM control chart will detect this change faster on average than any other monitoring method when all methods have the same frequency of false alarms and the control chart statistics for all methods happen

to be at the values which are least favorable for detecting the change. A proof of optimality for this scenario was given by Moustakides (1986). In this dissertation, the time to signal is based on the time required to signal based on the steady-state distribution of the (in-control) chart statistic, rather than using a worst-case scenario. Further, instead of considering only one particular shift for which the charts are optimized, a range of possible shifts in the rate will be studied here.

The objectives of this chapter are to investigate the properties of the Bernoulli scan statistic method when used in prospective surveillance, and to compare the performance of this method to the performance of the Bernoulli form of the CUSUM chart. These methods can be compared exactly by modeling each as a Markov process. In some cases, to be explained in Section 2.3.1, an approximate comparison is performed via simulation.

In Section 2.2, the definitions of the control charts based on the Bernoulli scan statistic and on the Bernoulli CUSUM statistic are given, and some other monitoring schemes are briefly discussed. In Section 2.3, the details of the evaluation and comparison of the performance of these monitoring methods, as well as some examples, are given. Section 2.4 contains a discussion on the conclusions of this study.

2.2 Monitoring Bernoulli rates

2.2.1 Bernoulli scan statistic chart

For a control chart based on the Bernoulli scan statistic one computes, after each trial i, the chart statistic

$$S_i = \sum_{j=\max(1,i-m+1)}^{i} Y_j, \tag{2.1}$$

as given by Glaz *et al.* (2001, p. 44).

Consistent with Ismail *et al.* (2003), a signal or alarm is given if $S_i \geq k$. It is assumed that $m \geq k \geq 2$. The case of $k = 1$ is excluded because the chart with $k = 1$ signals upon

the first incidence, regardless of the value of m. Such a chart does not seem to be useful for applications.

Unfortunately, it is not clear which specific values of k and m should be chosen for particular applications. It is easy to see that increasing k while holding m constant (subject to $k \leq m$) means that incidences will need to happen more frequently in order to produce a signal (alarm). This also implies that it will be harder to produce a false alarm, i.e., indicate an increase in the incidence rate when none has occurred. On the other hand, increasing m while holding k constant means that it will become easier to signal, regardless of whether a rate increase has really occurred. By using these relationships it is possible to design algorithms to help target a specified average number of observations until a false alarm. Such an algorithm is discussed in Section 2.3.2.

2.2.2 Bernoulli cumulative sum chart

The Bernoulli CUSUM control chart developed by Reynolds and Stoumbos (1999) is based on the following statistics:

$$\begin{aligned} C_0 &= 0 \\ C_i &= \max\left(0, C_{i-1} + Y_i - \tfrac{1}{r}\right), \quad i = 1, 2, 3, \cdots, \end{aligned} \tag{2.2}$$

where

$$r = -\frac{\log\left(\frac{p_1(1-p_0)}{p_0(1-p_1)}\right)}{\log\left(\frac{1-p_1}{1-p_0}\right)}. \tag{2.3}$$

For a given value of p_0, the value of r is determined by the choice of p_1 (through a likelihood ratio). Generally, r is rounded to the nearest integer because it simplifies the analysis of the chart's performance. Further, as explained in Section 2.3.3, the value of r may need to be increased slightly to obtain the optimal chart when steady-state performance is considered (see Section 2.3.1).

When using this chart, a signal occurs as soon as $C_i \geq h$. Increasing the value of h will increase the average time until a false alarm. For further discussion of this chart, see

Reynolds and Stoumbos (1999). Lucas (1989) showed that a Bernoulli scan statistic chart with $k = 2$ is equivalent to a Bernoulli CUSUM chart with $h = 1$ and $r = m$.

2.2.3 The Shewhart np chart

The traditional control chart used to monitoring a Bernoulli rate in industrial applications is the Shewhart np chart. To apply the np chart, observations must be grouped together into non-overlapping groups, which are often of constant size n. The chart statistic cannot be computed as each observation becomes available. Instead, when the jth group of n observations has been obtained, one computes

$$X_j = \sum_{i=(j-1)n+1}^{jn} Y_i. \tag{2.4}$$

Traditionally, a chart based on this statistic will signal as soon as any X_j exceeds $np_0 + z\sqrt{np_0(1-p_0)}$, where z is typically taken to be 3. This choice of z gives a limit that is three standard deviations above the in-control expected value of X_j. Increasing the value of z will decrease the false alarm rate. Alternatively, a signal may be given as soon as $X_j \geq c$, where c is determined by using the binomial distribution. In this case, the false alarm rate is determined by $\Pr(X_j \geq c)$, where X_j has a binomial distribution with parameters n and p_0.

The statistical performance of the np chart is adversely affected by the fact that one must wait for n observations to accumulate in order to detect whether there has been an increase in the incidence rate of interest. Although this control chart is currently in widespread use in a variety of applications, Reynolds and Stoumbos (1999, 2000) have shown that this method is almost always slower to detect a shift when compared to a Bernoulli CUSUM chart. For this reason, the np chart is not used in this chapter, and its use is not recommended in surveillance when the Bernoulli trials are observed one at a time.

2.2.4 Other techniques

There are several other methods that could be applied to monitor rates associated with Bernoulli trials. These include the Sets method of Chen (1978), and a variant of it proposed by Sitter *et al.* (1990), as well as the CUSCORE method described in Wolter (1987). Sego *et al.* (2007*b*) compared these methods with the Bernoulli CUSUM chart described above, using methods similar to those used in this chapter. They concluded that the Bernoulli CUSUM chart was almost always superior to these methods. They observed that the CUSCORE method was sometimes better at detecting relatively small increases in the incidence rate when the CUSUM and CUSCORE methods were both designed to detect a large change. Since these other methods have already been evaluated relative to the Bernoulli CUSUM chart, they will not be discussed further in this dissertation.

2.3 Performance

2.3.1 Methodology

Naus and Wallenstein (2006) measured the performance of the scan statistic method by computing the probability of obtaining a signal within 100 observations. Measuring performance in this way is useful only if it is expected that the surveillance technique will be in use for a prespecified amount of time.

The performance of Bernoulli control charts in industrial settings is often evaluated using the average (or expected) number of observations to signal, or ANOS. For prospective surveillance, the ANOS seems more useful since it provides information on how quickly the chart will signal under the assumptions that the monitoring is an ongoing concern and that any increase in the incidence rate is sustained until it is detected.

In some situations a rate increase may last only for a short period of time. In this situation, it may be more useful to consider the distribution of the number of observations

to signal, from which one can obtain the probability that a signal will occur while the rate increase is present (see, for example, Reynolds and Stoumbos, 2004*a*, *b*).

The ANOS of Bernoulli-based control charts can often be evaluated exactly. Formulas for the ANOS of the *np* chart and the Bernoulli CUSUM chart are provided in Reynolds and Stoumbos (1999). They indicated that the ANOS for these charts can also be found via a Markov chain approach. The ANOS for the Bernoulli scan statistic chart can also be found using Markov chains. The details of these computations are presented in Section 2.5.

The in-control ANOS is the ANOS associated with the in-control incidence rate p_0, so this ANOS is a measure of how often false alarms occur. Thus, in designing or comparing control chart schemes it is necessary to specify a desired value, say c, for the in-control ANOS. Ideally, when comparing two competing control charts, both would have in-control ANOS values equal to c. However, the Bernoulli distribution is discrete, so it is usually not possible to obtain in-control ANOS values precisely equal to c. The actual achieved in-control ANOS value for a given chart is often called the ANOS_0 value for that chart. Then, to make fair comparisons between different control charts, the ANOS_0 values should be as close as possible to, but above, c. One common method for selecting the best control chart is to choose the one which has the smallest ANOS at some specified incidence rate $p_1 = \gamma p_0$, among those charts that satisfy $\text{ANOS}_0 \geq c$.

When determining the out-of-control ANOS for a given value of $p > p_0$, many researchers have assumed that this increased rate begins at the same time that the monitoring begins or, equivalently, when the control chart statistic is at its initial value. In the process control literature, this is called an *initial-state* analysis. Initial-state analyses are not as useful when using a monitoring technique that uses past information to decide whether or not the incidence rate has increased. It is more likely that the monitoring began under the in-control incidence rate, p_0, and that any rate increase occurs at some later point in time. Comparisons between methods should be based on the assumption that some observations have been obtained when the incidence rate is p_0 before the increase in p. This is often referred to as a *steady-state* analysis. A discussion of the steady-state analysis and a derivation of the

expressions for the Bernoulli scan statistic chart's steady state ANOS (SSANOS) based on a Markov chain approach are provided in Section 2.5.

Unfortunately, exact values of the ANOS and SSANOS for the scan statistic chart can be difficult to obtain because the Markov chain approach may require a large number of states. The dimension of the Markov transition probability matrix is the number of states, and this matrix must be inverted to obtain the SSANOS (see Section 2.5). It is noted that it was possible to obtain a Markov chain solution for methods requiring as many as 8,475 states. The matrix inversion was not possible (with the department's computing resources) when the Markov chain approach required a very large number of states. In these cases, approximate ANOS and SSANOS values were obtained using simulation.

When there is a sustained increase in p to a specified value p_1, let ANOS_1 and SSANOS_1 be the values of the ANOS and SSANOS, respectively. Note that the value of r given by Equation 2.3 minimizes ANOS_1. However, the value of r usually needs to be increased slightly when the goal is to minimize SSANOS_1. In many cases, however, the value of r given by Equation 2.3 produces a reasonable chart, even though it doesn't minimize SSANOS_1.

2.3.2 Optimizing the scan statistic chart

A list of (k, m) combinations which satisfy the ANOS_0 requirement needs to be obtained. The first such combination can be found by starting with $k = 2$ and determining the largest m (where $m \geq 2$) which satisfies the requirement. Sometimes the $k = 2$, $m = 2$ chart will not have $\text{ANOS}_0 \geq c$. Since increasing the value of m will reduce the value of ANOS_0, it will not be possible to obtain a scan statistic chart with $k = 2$. In this case, k is increased by 1, and another search for an m is undertaken.

Once a valid (k, m) pair has been obtained, another pair can be obtained by increasing k by 1 and increasing m until the largest m which meets the minimum acceptable ANOS_0 value is determined. The largest acceptable value of m will increase as k increases.

The SSANOS_1 value can be computed for each (k, m) combination. Additional (k, m) combinations should be found until it is clear that the SSANOS_1 values are no longer decreasing as k increases further.

For the Bernoulli scan statistic chart, the ANOS is a decreasing function of m for a fixed value of k, because increasing m gives a larger window in which to obtain k incidences, making it easier to signal. If $m = \infty$ then the Bernoulli scan statistic chart signals as soon as k incidences are observed. By using the negative binomial distribution, it follows that the ANOS will be k/p. Thus, for any $m \geq 1$ and given some $k \geq 2$, it follows that a lower bound for the ANOS of the Bernoulli scan statistic chart is

$$\text{ANOS} \geq \frac{k}{p}. \tag{2.5}$$

When $p = p_0$, it will not be possible to obtain a value of ANOS_0 smaller than $2/p_0$ for any $k \geq 2$ and $m \geq k$. Thus, if c is significantly smaller than $2/p_0$, then it will not be possible to find a scan statistic chart with an ANOS_0 value close to c.

2.3.3 Optimizing the Bernoulli CUSUM chart

The value of r given by Equation 2.3 is appropriate for minimizing when the goal is to minimize the initial-state ANOS_1. The value of r must be increased somewhat when searching for the optimal Bernoulli CUSUM chart parameters when minimizing SSANOS_1. To find the optimal r and h in this steady-state case, one can start with the value of r given by Equation 2.3 (with r rounded to the nearest integer), and then search for the value of h which gives the desired ANOS_0. Note that the only values of h that need to be considered are integer multiples of $1/r$. The SSANOS_1 value (for the desired value of γ) is then be computed for this (r, h) pair. Next, the value of r is increased by one unit, and h increased until the desired ANOS_0 is again achieved. The value of SSANOS_1 should then be computed for these parameters. This process should be repeated until the SSANOS_1 value increases as r is increased.

Note that when r is increased that sometimes the ANOS_0 value must also increase in order to satisfy the minimum ANOS_0 requirement. This may result in an increase in the value of SSANOS_1. Therefore, the search for better values of r and h should continue until the SSANOS_1 value increases when a smaller ANOS_0 value is obtained.

However, the r parameter for the CUSUM chart can get too large. When a process is in control, it is expected that the chart statistic will tend to drift toward zero. Assuming that the incidence rate is at its in control value of p_0, Equation 2.2 shows that the chart statistic will increase by $\frac{r-1}{r}$ with probability p_0 and will decrease by $\frac{1}{r}$ (or remain at zero) with probability $1 - p_0$. It is then possible to determine the expected change in the CUSUM chart statistic at any given time. To maintain the expected change at zero or less, it is necessary for

$$\frac{r-1}{r}p_0 - \frac{1}{r}(1-p_0) \leq 0.$$

This is algebraically equivalent to requiring $rp_0 \leq 1$.

2.3.4 Example

One of the examples given by Naus and Wallenstein (2006) concerns the mortality rate on a cardiac surgeon's neonatal operations. In the example, the in-control mortality rate, p_0, is two percent. They suggest using a Bernoulli scan statistic rule with $k = 3$ and $m = 15$. This choice was justified by showing that the probability of signaling an increase in mortality within the next 100 operations (when the rate remains at two percent) is only 0.0463. The authors then show that this chart is very likely to signal if the rate increases by a factor of $\gamma = 6$ to $p_1 = 0.12$. The probability of signaling such an increase in mortality within the next 100 operations is 0.9588.

The Markov chain technique can be used to show that this scan statistic chart has an in-control ANOS (ANOS_0) of 1,931.54. This means that when the mortality rate remains constant at 0.02, there will be an average of 1,931.54 operations until a false alarm. When the mortality rate increases to 0.12, the out-of-control SSANOS (SSANOS_1) is 34.67. This

means that after a mortality rate increase to 0.12, an average of 34.67 operations will be performed before this increase is detected.

In an attempt to find competing charts, assume that an in-control ANOS of at least $c = 1,900$ is desired. Using the Bernoulli CUSUM method of Reynolds and Stoumbos (1999), the optimal parameters are found to be $r = 20$ and $h = 2\frac{9}{20}$. The Markov chain technique indicates that this chart has $\text{ANOS}_0 = 1{,}928.15$ and $\text{SSANOS}_1 = 31.67$, the latter being under a $\gamma = 6$-fold increase. Therefore, the optimal Bernoulli CUSUM chart signals about 2.71 operations faster (on average) than the Bernoulli scan statistic chart recommended by Naus and Wallenstein (2006). This comparison is fair since the two charts signal at about the same time, on average, when no increase has occurred. The CUSUM chart has a slightly higher false alarm rate than the scan statistic chart, so a clear improvement over the scan statistic chart can be shown by using a CUSUM chart with the (suboptimal) parameters $r = 21$ and $h = 2\frac{11}{21}$. According to the Markov chain method, this chart has $\text{ANOS}_0 = 1{,}969.75$ and $\text{SSANOS}_1 = 31.85$. This chart has fewer false alarms on average than the scan statistic method, and also signals a $\gamma = 6$-fold rate increase faster than the scan statistic chart.

It can be shown that the performance of the Bernoulli scan statistic chart can be improved by increasing k to 4 and m to 38. Using the Markov chain technique, the ANOS_0 for this chart is 1,939.89, while the SSANOS_1 is 32.91. However, the performance of this scan statistic chart is still not as good as the performance of the CUSUM chart.

2.3.5 Ability to detect shifts of varying sizes

The goal in the preceding example is to detect a specified increase in p (from an incidence rate of 0.02 to an increased rate of 0.12) as soon as possible while maintaining a specified in-control ANOS. However, it is possible that the rate will increase to some value other than 0.12, since the size of any shift in p that might occur is unknown. Therefore, it is beneficial to compare the statistical performance of the charts under a variety of possible rate increases.

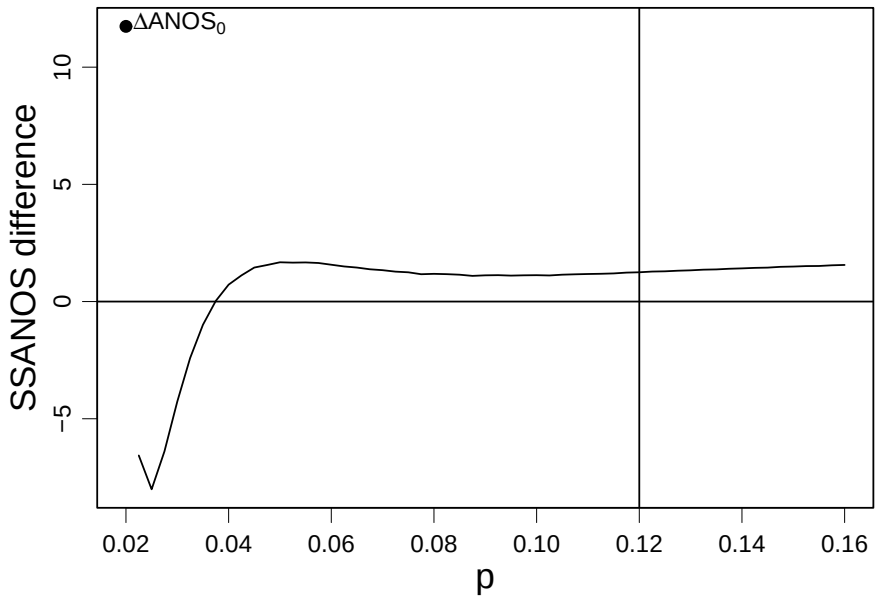


Figure 2.1: SSANOS for Bernoulli scan statistic chart with $k = 4$ and $m = 38$ minus SSANOS for Bernoulli CUSUM chart with $r = 20$ and $h = 49/20$.

Figure 2.1 shows the SSANOS difference when the rate increases to various values of p above 0.02. The SSANOS difference in these figures is computed by subtracting the SSANOS for the Bernoulli CUSUM chart from the SSANOS for the Bernoulli scan statistic chart. Both charts were designed to produce the smallest SSANOS_1 when $p_1 = 0.12$, conditional on $\text{ANOS}_0 \geq 1900$. The CUSUM chart has $r = 20$ and $h = 2\frac{9}{20}$. The scan statistic chart has $k = 4$ and $m = 38$. For a given $p > 0.02$ in this figure, a plotted value that is above zero indicates that the CUSUM chart is better. It can therefore be seen that the scan statistic chart is somewhat superior as long as the rate does not increase above about 0.04, but at $p_1 = 0.12$ (the assumed primary point of interest) the CUSUM chart is better. As an indication of the relative performance of the control charts in the in-control case, a point labeled ΔANOS_0 is plotted at $p = 0.02$. This point is the ANOS for the scan statistic chart minus the ANOS for the CUSUM chart. The positive value plotted here indicates that the CUSUM chart has a slightly lower value of ANOS_0.

An increase from $p_0 = 0.02$ to $p_1 = 0.12$ is quite large. There will be many monitoring situations in which it will be necessary to rapidly detect a smaller increase. The value of p_1 that is selected when designing a chart will affect the SSANOS values at all values of p. A comparison of the SSANOS for two Bernoulli CUSUM charts in Figure 2.2 is provided as an example. One of the CUSUM charts is the one from Figure 2.1 that was designed to produce the smallest SSANOS_1 when $p_1 = 0.12$. It has $r = 20$ and $h = 2\frac{9}{20}$. The other CUSUM chart is designed to produce the smallest SSANOS_1 when $p_1 = 0.05$ ($\gamma = 2.5$). It has $r = 37$ and $h = 3\frac{29}{37}$. Both of these charts have $\text{ANOS}_0 > 1900$. For this figure, the SSANOS difference is the subtraction of the SSANOS for the CUSUM chart optimized for $p_1 = 0.12$ from the SSANOS for the CUSUM chart optimized for $p_1 = 0.05$.

In Figure 2.2 a plotted value above zero indicates that the CUSUM chart optimized for $\gamma = 2.5$ is superior for the given $p > 0.02$. This figure shows that when the rate increases to 0.12, the CUSUM optimized for $\gamma = 6$ is best; its SSANOS value is about 1.5 observations lower. However, when the rate increases to 0.05, the SSANOS value is about 30 observations lower when using the chart designed for that $\gamma = 2.5$-fold increase. This example shows that when γ is relatively large, the optimal SSANOS_1 is often only marginally smaller than it is if the chart is designed for a smaller increase. When the chart is designed for a large increase, however, the SSANOS for smaller increases is usually much larger than when the chart is designed for a small increase. Similar results occur when comparing the optimal scan statistic charts for $\gamma = 6$ and $\gamma = 2.5$. Thus, when designing control charts it is generally better to set p_1 to the smallest sustained increase for which rapid detection is needed.

The optimal Bernoulli scan statistic chart for detecting a $\gamma = 2.5$-fold increase when $\text{ANOS}_0 \geq 1900$ has $k = 7$ and $m = 142$. Its SSANOS performance is compared with the SSANOS performance of the above-mentioned Bernoulli CUSUM chart optimized for a $\gamma = 2.5$-fold shift. Figure 2.3 shows the differences in SSANOS values for the Bernoulli CUSUM chart and the Bernoulli scan statistic chart for increases to different values of p. The CUSUM chart is superior to the scan statistic method for all $p > 0.02$. The CUSUM chart has slightly worse in-control performance.

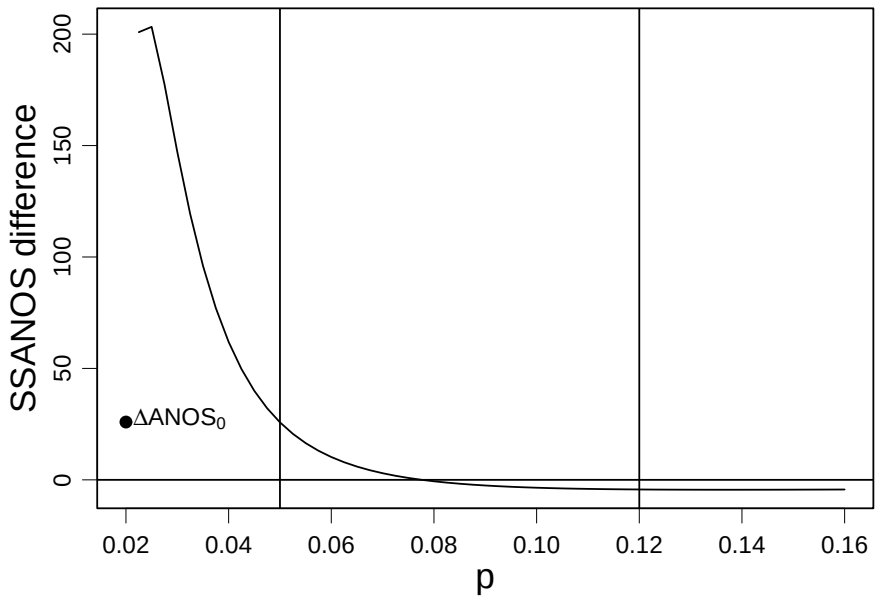


Figure 2.2: SSANOS for Bernoulli CUSUM chart with $r = 37$ and $m = 140/37$ minus SSANOS for Bernoulli CUSUM chart with $r = 20$ and $h = 49/20$.

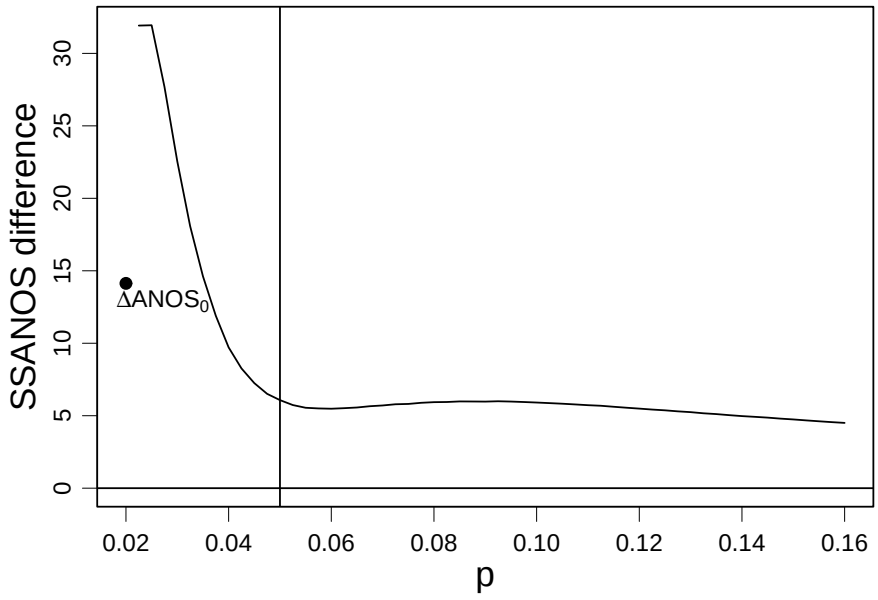


Figure 2.3: SSANOS for Bernoulli scan statistic chart with $k = 7$ and $m = 142$ minus SSANOS for Bernoulli CUSUM chart with $r = 37$ and $h = 140/37$.

Table 2.1: Approximate range of values of p_1 for which the specified scan statistic chart is optimal, subject to $\text{ANOS}_0 > 1900$.

p_1 range		k	m	ANOS_0	$se(\text{ANOS}_0)$
0.0350	0.0375	10	274	1904.39	0.45
0.0375	0.0400	9	228	1903.69	0.46
0.0400	0.0450	8	184	1903.21	0.46
0.0450	0.0525	7	142	1916.34	0.47
0.0525	0.0625	6	103	1932.73	0.48
0.0625	0.0850	5	68	1938.63	0.49
0.0850	0.1425	4	38	1939.88	$*$
0.1425	1.0000	3	15	1931.54	$*$

$*$The ANOS_0 value was calculated exactly.

Figure 2.1 suggests that there can be rate increases to values somewhat smaller than p_1 (the value at which the charts are optimized) where the scan statistic is superior to the CUSUM. However, it happens that a scan statistic chart with a particular k and m is optimal over a range of values of p_1. Further, a CUSUM chart that is designed for the smallest value of p_1 for which the selected scan statistic chart is optimal gives better overall performance than this scan statistic method.

Table 2.1 presents the optimal k and m values for a scan statistic chart designed to detect an increase from $p_0 = 0.02$ to values of p_1 between 0.035 and 1, subject to $\text{ANOS}_0 \geq 1900$ and $k \leq 10$. The cutoff values of p_1 given in this table are approximate. This is due to simulation error and the selection of multiples of 0.0025 as candidate values for p_1.

Table 2.1 shows that the $k = 4,\ m = 38$ scan statistic chart is optimal for a value of p_1 as small as 0.085. The CUSUM chart that is optimal for detecting a shift to 0.085 ($\gamma = 4.25$) has $r = 27$ and $h = 2\frac{8}{9}$. This chart has a lower ANOS_0 than the competing scan statistic chart. Figure 2.4(a) is a plot of the scan statistic SSANOS minus the CUSUM chart SSANOS. This figure shows that for all values of p the CUSUM chart signals faster on average than the optimal scan statistic chart.

In Figures 2.1, 2.3, and 2.4(a), the scan statistic chart always has a higher ANOS_0 than the CUSUM chart. This can be seen as giving a slight advantage to the CUSUM chart when

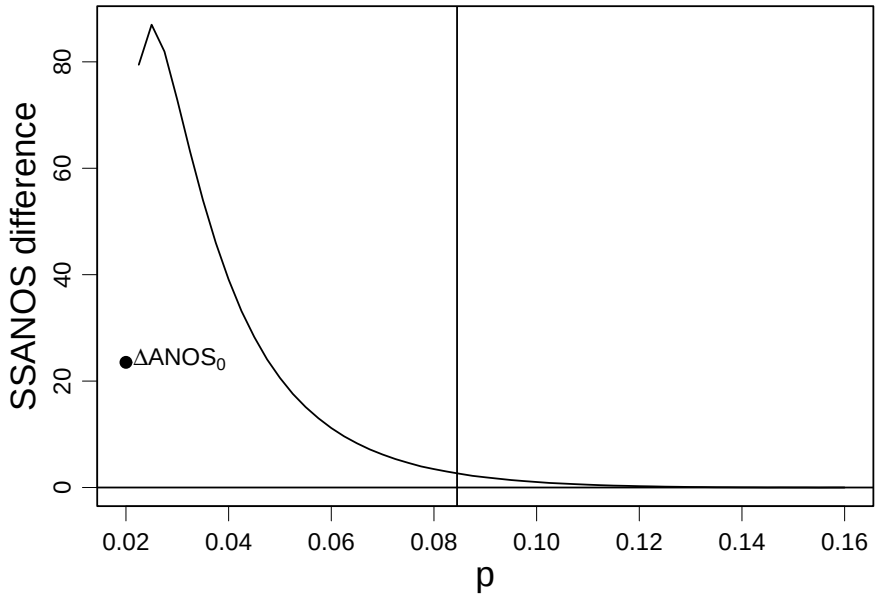
comparing the out-of-control SSANOS values of the two charts. Therefore, the same scan statistic chart is compared to a CUSUM chart with the parameters $r = 26$ and $h = 2\frac{11}{13}$. By using these suboptimal parameters, the scan statistic chart will have a lower ANOS_0 value than the CUSUM chart. The relative performance of these charts is given in Figure 2.4(b). Figure 2.4(b) shows that if the CUSUM chart has the larger value of ANOS_0, its performance still overtakes that of the scan method quickly as p increases. If the scan method has the higher value of ANOS_0, as shown in Figure 2.4(a), then its performance never overtakes that of the CUSUM chart.

According to Table 2.1, the $k = 7,\ m = 142$ scan chart is optimal for a shift to a value of p_1 as small as 0.045. An optimal CUSUM chart for this $\gamma = 2.25$ shift has $r = 38$ and $h = 3\frac{17}{19}$. These charts, which both have $\text{ANOS}_0 \geq 1900$, are compared in Figure 2.5(a). Once again, the scan statistic method has a higher ANOS_0, and the CUSUM chart signals faster than the scan statistic chart for all values of $p > .02$ considered. Figure 2.5(b) shows the comparison between a CUSUM chart with $r = 39$ and $h = 4\frac{1}{39}$ and the optimal scan statistic chart. In this comparison, the ANOS_0 value for the scan statistic chart is smaller than the ANOS_0 value for the CUSUM chart, and the CUSUM chart detects increases in p faster except for values of p very close to p_0.

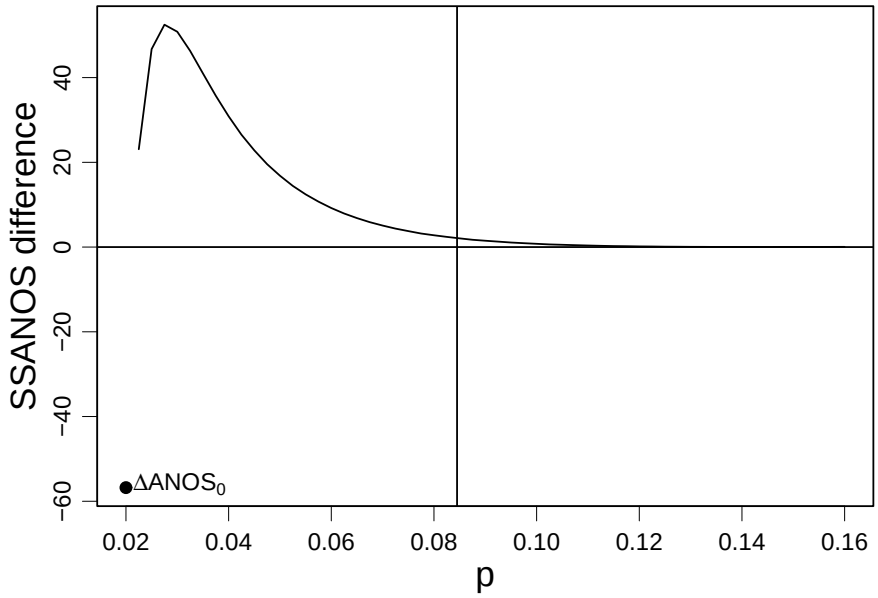
These comparisons are complicated by the fact that it is not possible to obtain exactly the same in-control ANOS_0 values for the charts being compared. However, in all of these examples, the CUSUM charts have better overall performance than the scan methods.

2.3.6 Other performance measures

The CUSUM chart also has better performance than the scan statistic chart with respect to another performance measure, the minimum number of observations which can cause the chart to signal. Suppose each chart statistic is at its worst-case value when a very large increase in the rate occurs. For the scan chart, this would imply that there were no incidences in the last $m - k + 1$ opportunities. The worst-case situation for the CUSUM

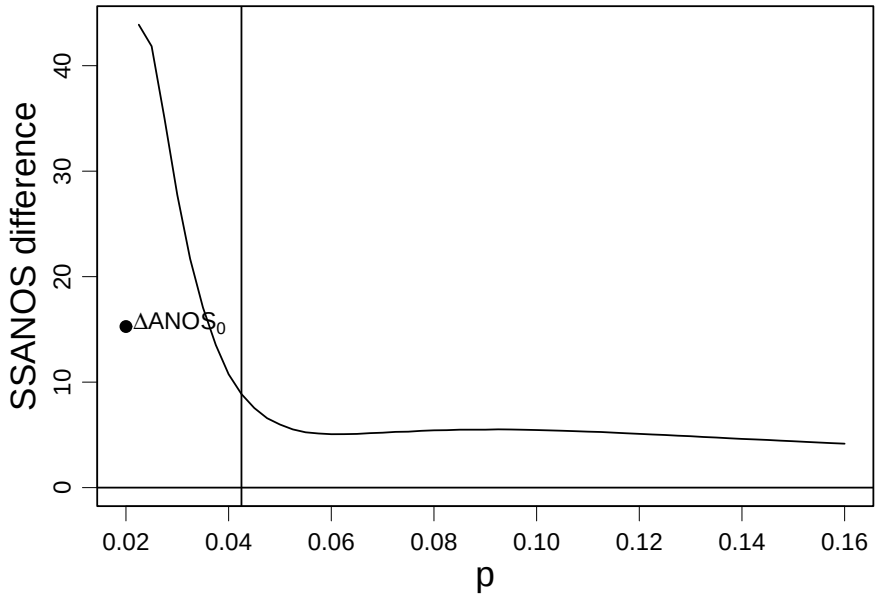


(a)

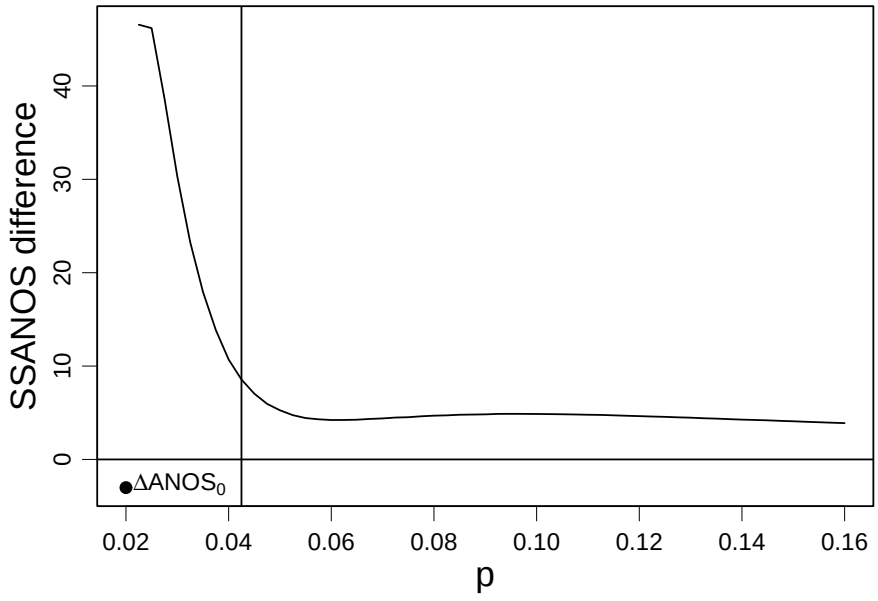


(b)

Figure 2.4: SSANOS for Bernoulli scan statistic chart with $k = 4$ and $m = 38$ minus SSANOS for Bernoulli CUSUM charts (a) with $r = 27$ and $h = 26/9$ and (b) with $r = 26$ and $h = 37/13$.



(a)



(b)

Figure 2.5: SSANOS for Bernoulli scan statistic chart with $k = 7$ and $m = 142$ minus SSANOS for Bernoulli CUSUM chart (a) with $r = 38$ and $h = 74/19$ and (b) with $r = 39$ and $h = 157/39$.

chart is a CUSUM value of zero. Under a worst-case condition, it will take the scan statistic chart at least k consecutive incidences to produce a signal. However, the minimum number of consecutive incidences which will cause the CUSUM chart to signal is $h\frac{r}{r-1}$ rounded up to the next higher integer. In all cases considered in this chapter, the CUSUM chart is superior with respect to this minimum number of consecutive observations that must be taken before an out-of-control signal can be given. For example, the charts optimal for an increase by a factor of $\gamma = 4.25$ (Figure 2.4) have $h = 2\frac{8}{9}$, $r = 27$, and $k = 4$. It is therefore possible for the CUSUM chart to signal the increase in as few as three observations, whereas the scan statistic chart can only signal after four observations. This worst-case performance analysis is related to that of Woodall and Mahmoud (2005) who primarily considered charts that could signal immediately based on a single observation.

An alternative to the SSANOS_1 is the steady-state median number of observations required to obtain a signal (SSMNOS). For the $k = 4$, $m = 38$ scan statistic chart used in Section 3.2, the estimate of this value is 28. For the CUSUM with $r = 20$ and $h = 2\frac{9}{20}$ given in the same section, the estimate of the SSMNOS is 26. Therefore, the CUSUM chart is superior with respect to the SSMNOS after a shift to $p_1 = 0.12$.

There are cases in which the scan statistic has an advantage with respect to the SSMNOS. Consider the comparison in Figure 2.4(a), where the charts are designed for a rate increase from 0.02 to 0.085 while meeting a target ANOS_0 of 1900. The optimal Bernoulli CUSUM chart has $\text{SSANOS}_1 = 53.15$, while the SSANOS_1 for the scan statistic chart is 55.73. However, the estimated SSMNOS for the CUSUM chart is 44, but is 42 for the scan statistic chart.

A comparison of the distributions of the number of observations to signal in the steady-state case for the Bernoulli CUSUM chart and the scan statistic chart helps explain this discrepancy between the comparisons of the average and median observations to signal. The distributions are plotted in Figure 2.6 and are computed using Markov chain theory. The x-axis indicates the number of observations taken since the sustained rate increase occurred.

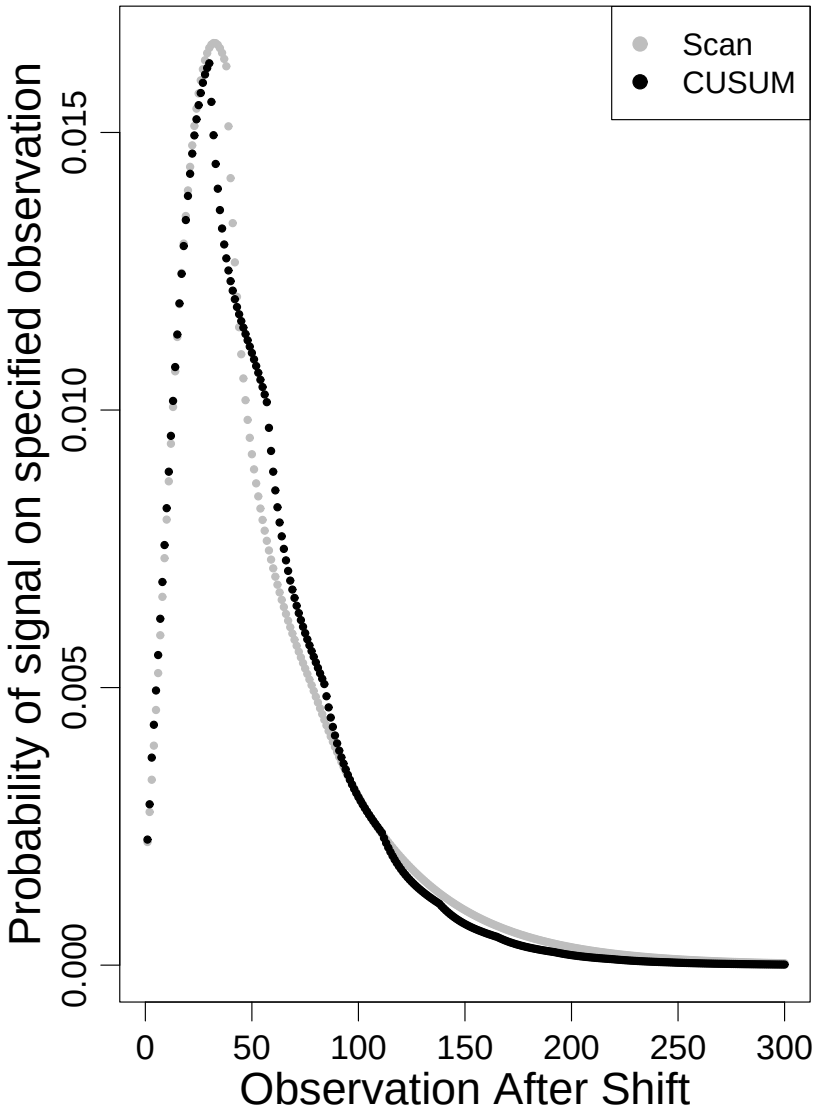


Figure 2.6: Distributions of number of observations to signal for the Bernoulli scan statistic chart with $k = 4$ and $m = 38$ and the Bernoulli CUSUM chart with $r = 27$ and $h = 26/9$.

Points are plotted in the figure (rather than histogram bars) because of the overlap in the distributions. The numbers of observations to signal for the scan method are plotted with gray circles, while the CUSUM chart's values are plotted with black circles. The plot demonstrates the approximate equivalence between the two methods until observation 30, at which point the scan statistic method gains a considerably higher signal probability for several observations. This gives the scan method the lower SSMNOS. Examination of the upper tails in the plot helps explain the scan method's higher SSANOS_1, as there is greater probability in that upper tail for the scan method. That is, the scan statistic chart is more likely than the CUSUM chart to require a very long detection time.

It is interesting to note that the distribution of signal times for the CUSUM is not as smooth as the distribution of signal times for the scan statistic chart. The sudden changes in slope in the distribution of signal times for the CUSUM chart are spaced at intervals of r apart. This property is present in the signal time distributions of all of the CUSUM charts we have considered. One possible explanation is that if an incidence does not occur in r consecutive trials, then the reference value will have caused the chart statistic to decrease by one full unit. In this case, an additional incidence will be required in order to obtain a signal.

2.3.7 Additional comparisons

The preceding examples suggest that, as a general rule, the Bernoulli CUSUM chart can be designed to have better overall statistical performance than the Bernoulli scan statistic chart. To investigate the generality of this result, the relative performance of the charts is examined while varying the three factors c, p_0, and γ (or equivalently p_1).

Two levels are considered for the minimum in-control ANOS_0, $c = 500$ and 10000. For the in-control incidence rate (p_0) the values 0.001 and 0.02 are considered, and for the shift magnitude of interest (γ) the values 1.5, 2.5, and 6.0 are studied. These additional levels have been chosen under the assumption that it is important to detect small increases in

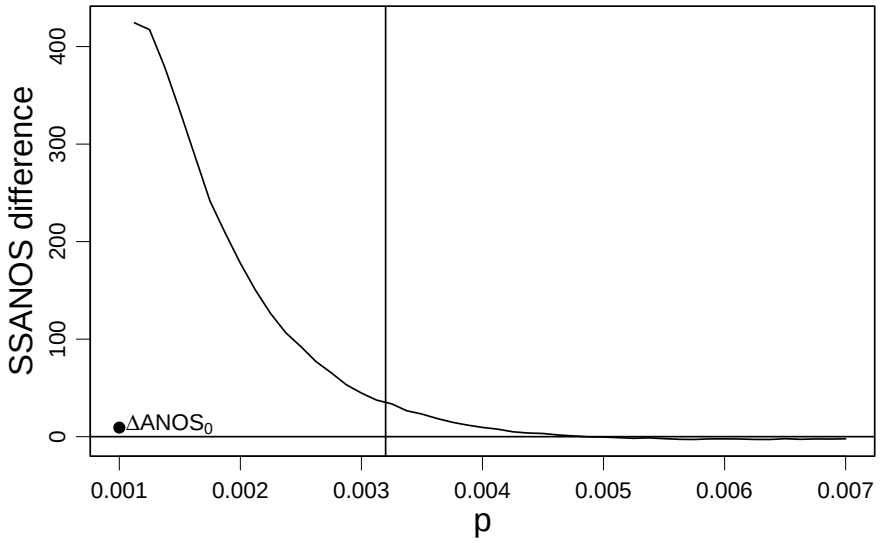
the incidence rate. At each combination of these factor levels, the chart of each type with the best performance is found. It is also possible to determine which chart has superior performance over a wide range of shifts.

When $p_0 = 0.001$, $c = 10000$, and $\gamma = 6$, the best scan statistic chart has $k = 3$ and $m = 692$. This scan statistic chart is the best one for shifts as small as $\gamma = 3.2$. The optimal Bernoulli CUSUM chart for this γ has $r = 812$ and $h = 2\frac{355}{812}$. One CUSUM chart which has a higher ANOS_0 value than the scan statistic chart has $r = 812$ and $h = 2\frac{89}{203}$. These CUSUM charts are compared with the $k = 3, m = 692$ scan statistic chart in Figure 2.7.

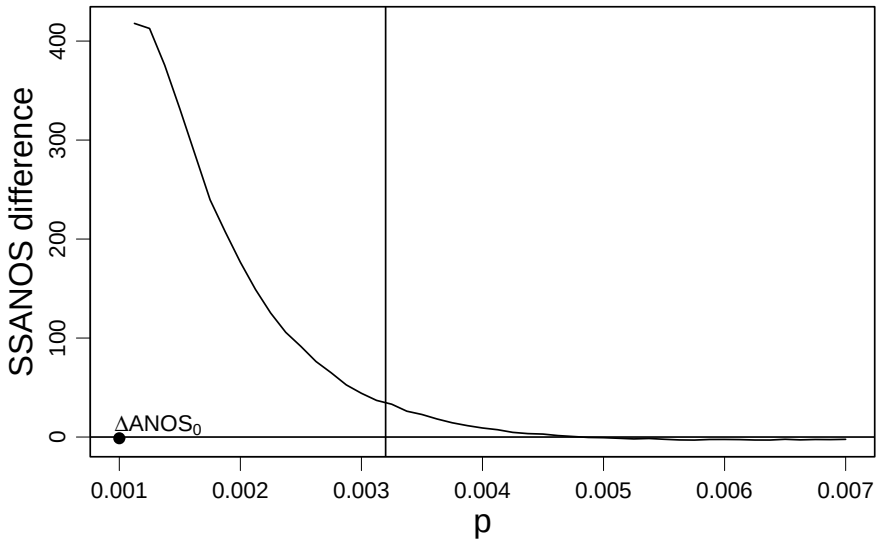
When p_0 is increased to 0.02 (keeping $c = 10000$ and $\gamma = 6$), the best scan statistic chart has $k = 6$ and $m = 62$. These parameters hold for shifts as small as $\gamma = 4.5$. The optimal Bernoulli CUSUM chart for that shift has $r = 23$ and $h = 3\frac{17}{23}$. This chart is unique amongst the optimal CUSUM charts considered throughout this chapter in that it has a higher ANOS_0 value than the corresponding optimal scan statistic chart. A suboptimal CUSUM chart which produces a lower ANOS_0 value than the scan statistic chart has $r = 26$ and $h = 4\frac{3}{26}$. In Figure 2.8, the CUSUM charts are compared with the optimal scan statistic chart.

When $p_0 = 0.02$ and $c = 10000$, but γ is decreased to 2.5, the optimal scan statistic chart has $k = 11$ and $m = 218$. This chart is optimal for shifts as small as $\gamma = 2.4$, but it turns out that the optimal Bernoulli CUSUM chart is the same for both $\gamma = 2.4$ and $\gamma = 2.5$. The optimal CUSUM chart has $r = 34$ and $h = 5\frac{8}{17}$. A suboptimal CUSUM chart which gives a higher ANOS_0 value than the scan statistic chart has $r = 33$ and $h = 5\frac{3}{11}$. These two CUSUM charts are compared with the optimal scan statistic chart in Figure 2.9.

If $\gamma = 1.5$ while still maintaining $p_0 = 0.02$ and $c = 10000$, the optimal scan statistic chart has $k = 27, m = 882$. This chart is optimal for shifts as small as $\gamma = 1.45$. The optimal Bernoulli CUSUM chart for this shift has $r = 45, h = 9\frac{16}{45}$. A suboptimal Bernoulli CUSUM chart which has a higher ANOS_0 value uses $r = 46, h = 10$. Figure 2.10 compares each of these CUSUM charts with the optimal scan statistic chart.

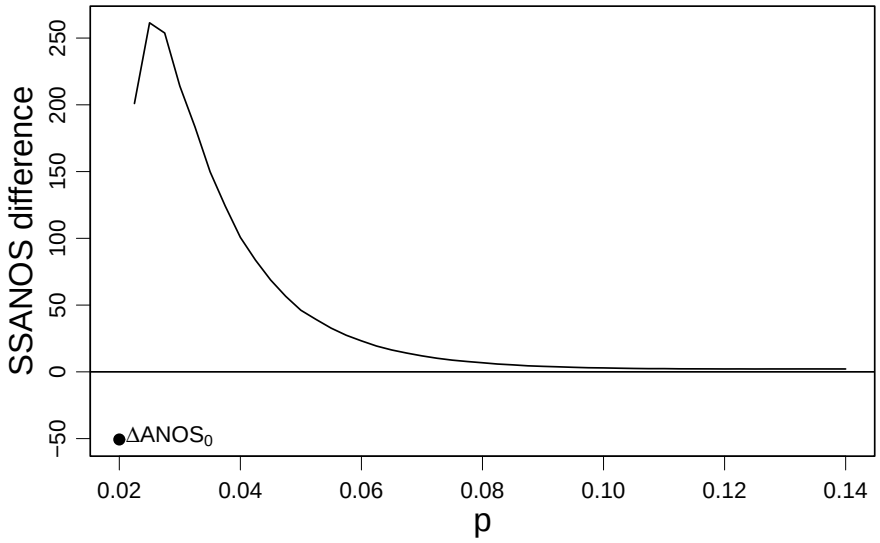


(a)

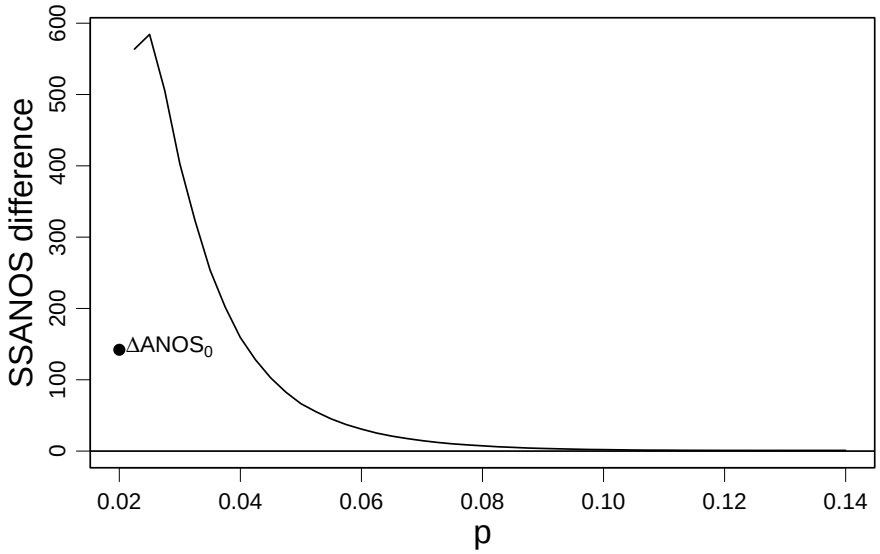


(b)

Figure 2.7: SSANOS for Bernoulli scan statistic chart with $k = 3$ and $m = 692$ minus SSANOS for Bernoulli CUSUM chart (a) with $r = 812$ and $h = 1979/812$ and (b) with $r = 812$ and $h = 495/203$.

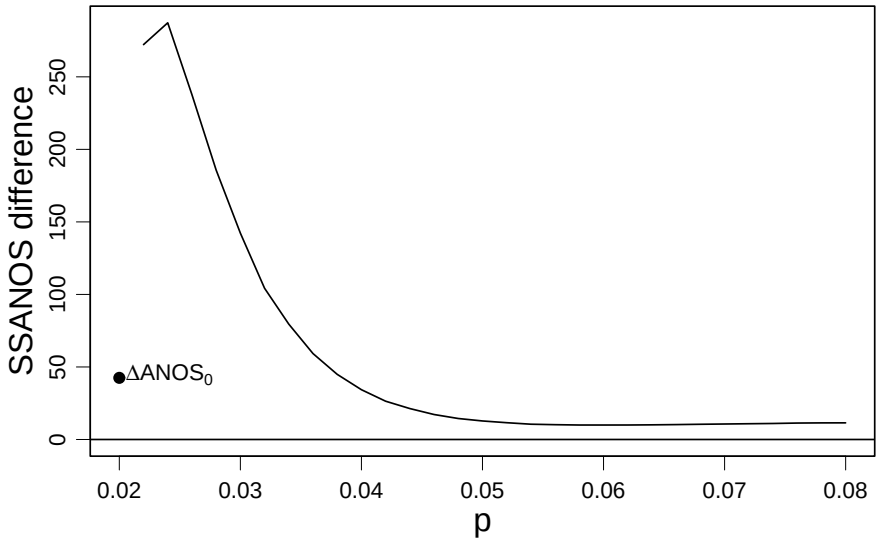


(a)

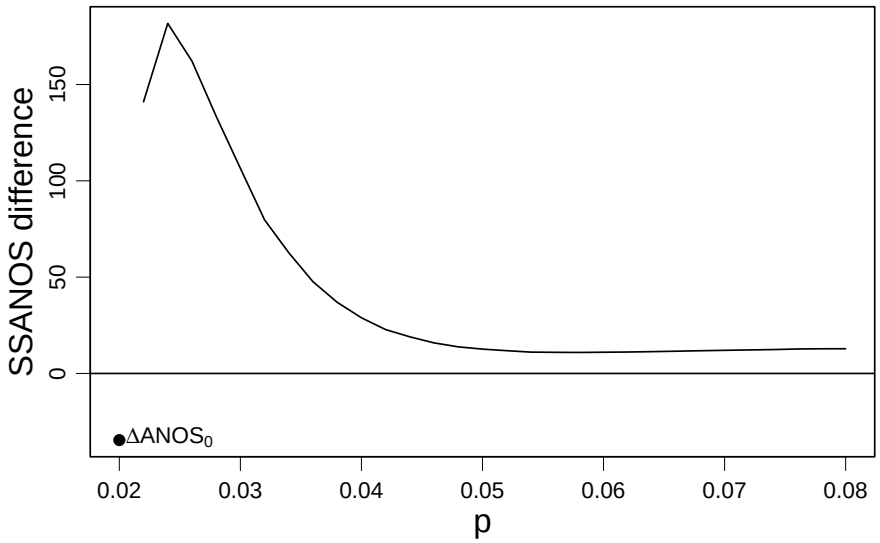


(b)

Figure 2.8: SSANOS for Bernoulli scan statistic chart with $k = 6$ and $m = 62$ minus SSANOS for Bernoulli CUSUM chart (a) with $r = 23$ and $h = 86/23$ and (b) with $r = 26$ and $h = 107/26$.

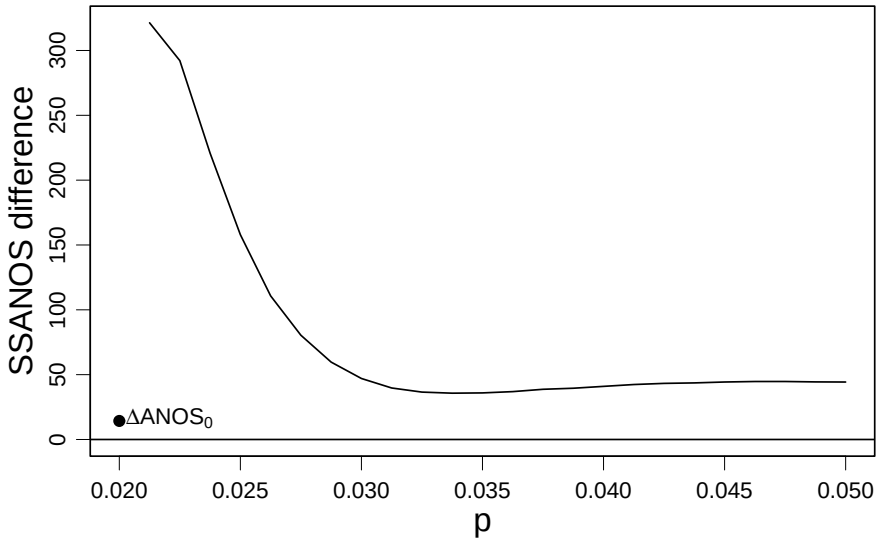


(a)

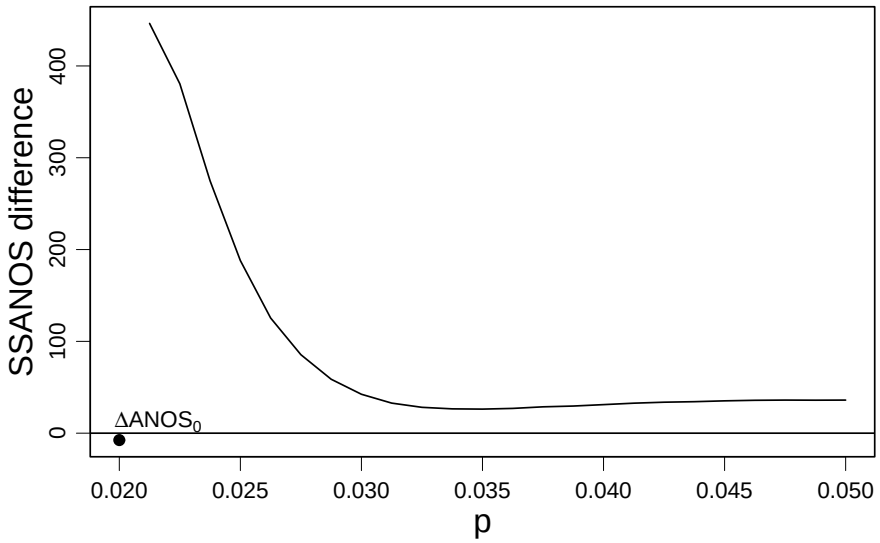


(b)

Figure 2.9: SSANOS for Bernoulli scan statistic chart with $k = 11$ and $m = 218$ minus SSANOS for Bernoulli CUSUM chart (a) with $r = 34$ and $h = 93/17$ and (b) with $r = 33$ and $h = 58/11$.



(a)



(b)

Figure 2.10: SSANOS for Bernoulli scan statistic chart with $k = 27$ and $m = 882$ minus SSANOS for Bernoulli CUSUM chart (a) with $r = 45$ and $h = 421/45$ and (b) with $r = 46$ and $h = 10$.

A final comparison involves $p_0 = 0.02$, $c = 500$, and $\gamma = 6$. The optimal scan statistic chart has $k = 3$ and $m = 35$, and is optimal for shifts as small as $\gamma = 3.25$. The optimal Bernoulli CUSUM chart for the $\gamma = 3.25$ shift has $r = 35$ and $h = 2\frac{6}{35}$. Increasing h to $2\frac{1}{5}$ gives a suboptimal Bernoulli CUSUM chart with a higher ANOS_0 value than that of the optimal scan statistic chart. These CUSUM charts are compared with the scan statistic chart with $k = 3$ and $m = 35$ in Figure 2.11.

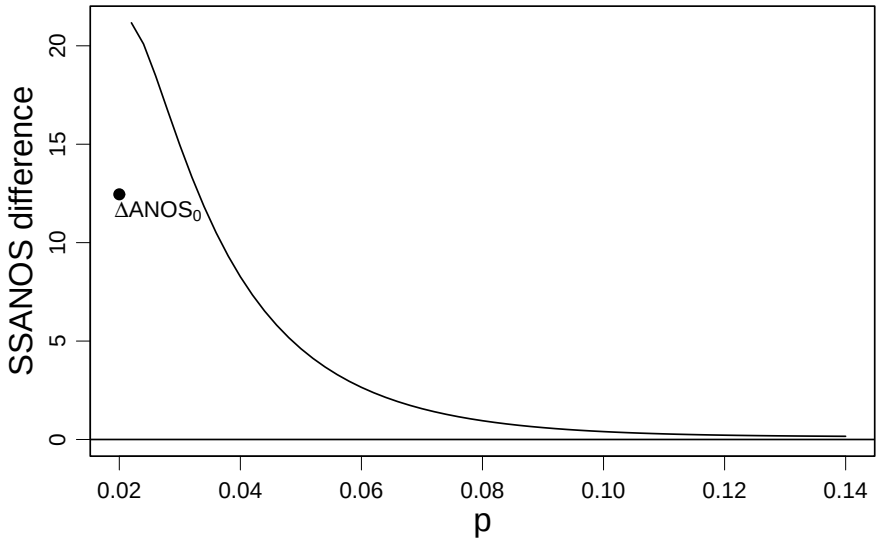
Figures 2.7–2.11 all exhibit results similar to those given earlier in this chapter. Thus it is apparent that, overall, the CUSUM method has superior SSANOS performance.

2.3.8 Detecting shifts of limited duration

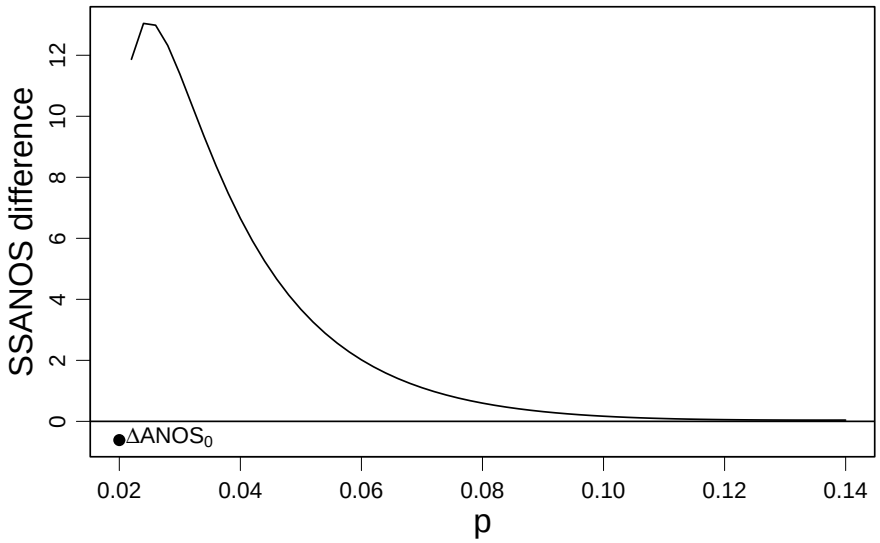
Many increases in rates are caused by transitory events. Here it is assumed that the incidence rate begins at its normal level, p_0 (thus creating the need to draw comparisons using the steady state). After some time, the incidence rate will increase to p_1, but since the event is transitory, the rate will naturally return to p_0, after some amount of time, d. In this case, the SSANOS is not particularly meaningful. It is much more informative to determine the probability of indicating an increase in the incidence rate to p_1 on observation t, $\Pr(T = t)$, for $t \leq d$. These are the probabilities plotted on the y-axis of Figure 2.6.

The information on the probability distribution of T can be used to find $\Pr(T \leq d)$, the probability that the rate increase is detected before the rate returns to p_0. Note that when $d = \infty$, the rate increase is indefinitely sustained. In this case, the probability of eventually detecting the rate increase is 1. Thus, when $d = \infty$ the SSANOS is equal to $E(T)$.

There are two general methods to determine the probability that a monitoring technique generates an alarm while a transitory shift is present. The first method involves multiplying the Markov transition probability matrix (if it can be obtained) by itself d times. The other method involves simulating data with the rate increased (to some $p > p_0$), determining the proportion of simulations that signal at observation t (for several values of t) and then computing the proportion of simulations which generate an alarm within d observations. In



(a)



(b)

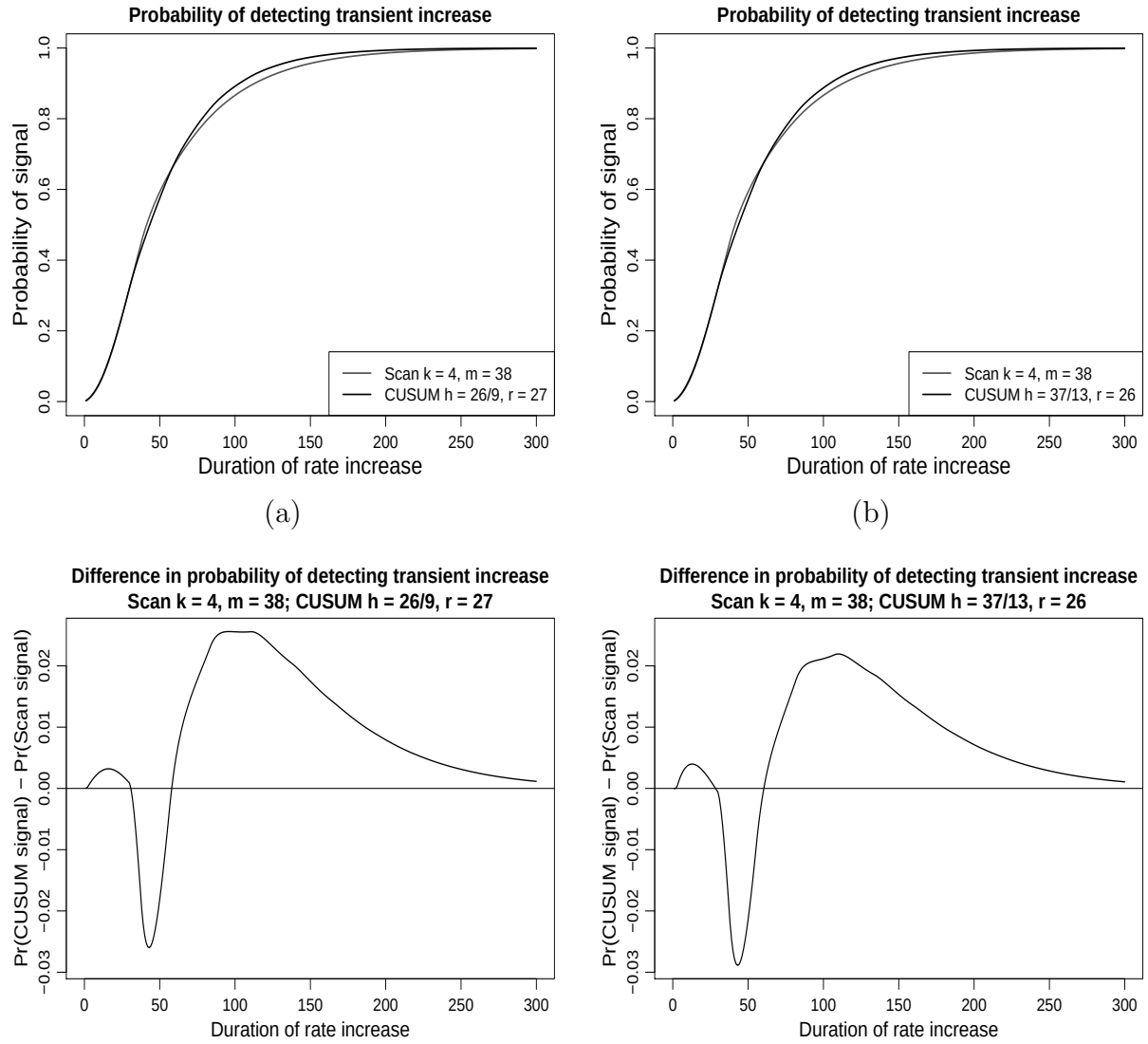
Figure 2.11: SSANOS for Bernoulli scan statistic chart with $k = 3$ and $m = 35$ minus SSANOS for Bernoulli CUSUM chart (a) with $r = 35$ and $h = 76/35$ and (b) with $r = 35$ and $h = 11/5$.

the simulations in this section, 100 million trials are used. Steady-state methodology will be used with the Markov approach and the simulations.

The signal probabilities for detecting a rate increase lasting d observations with the CUSUM and scan statistic methods are shown in Figures 2.12–2.18. Each figure uses one of the combinations of p_0, p_1, and c that were compared earlier in this chapter (under the assumption that $d = \infty$). All of these figures are divided into four parts, with each part using the same range on the x-axis (the length of the rate increase). In the top portions, the y-axis indicates the probability that the shift is detected for both a scan method and a competing CUSUM method. The y-axis in the lower portions indicates the difference in signal probabilities. If the difference is positive, the CUSUM method is more likely to signal than the scan statistic method. The optimal scan statistic method is compared with the optimal CUSUM method in the left hand side of these figures. In the right hand side, the same optimal scan statistic method is compared with a slightly suboptimal CUSUM method, as was done in the SSANOS comparisons presented earlier.

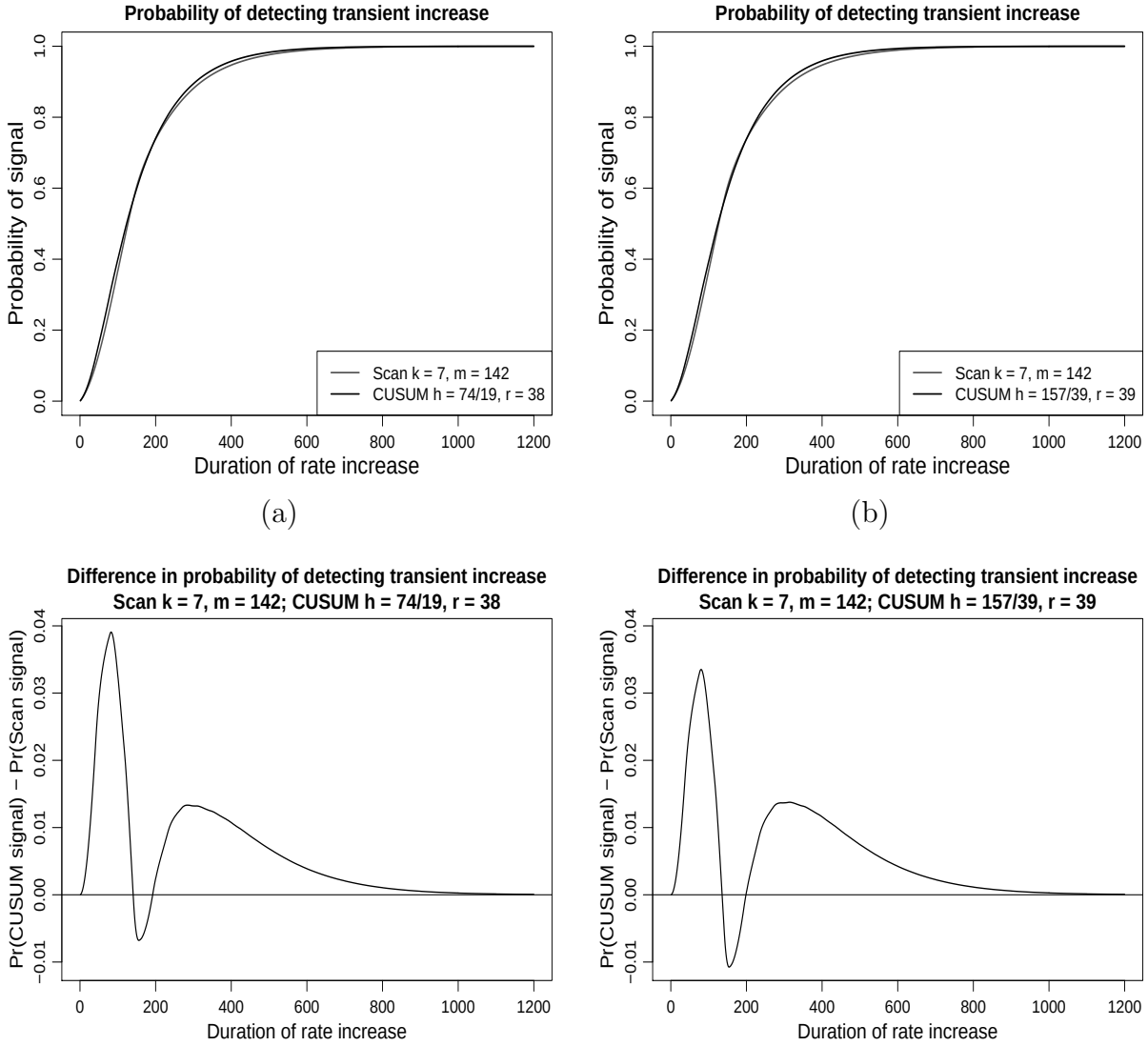
In Figure 2.12, the methods for the $p_0 = 0.02,\ p_1 = 0.085,\ c = 1900$ case (discussed earlier with Figure 2.4) are compared. Figure 2.12 is based on the Markov chain method for both charts, and therefore the signal probabilities are exact. The figure shows that the optimal CUSUM chart has a slightly higher probability of signal than the scan statistic chart for $d \leq 30$. For $30 < d \leq 57$ the scan statistic chart is better; in some cases it is over 2% more likely than the CUSUM chart to detect an increase. However, for $d > 57$ the optimal CUSUM chart is consistently more likely to signal the rate increase. The comparison of the optimal scan statistic chart with the suboptimal CUSUM chart shows similar patterns.

The comparison for the $p_0 = 0.02,\ p_1 = 0.046,\ c = 1900$ case (used earlier in Figure 2.5) is given in Figure 2.13. Here the probabilities for the scan statistic chart are obtained by simulation. Note that the optimal CUSUM chart has a considerable advantage (up to four percentage points) over the optimal scan statistic chart for $d \leq 140$. Indeed, the scan statistic method's advantage is relatively small, and exists over a rather small range of d ($140 < d \leq 192$). For $d > 192$, the optimal CUSUM chart is consistently superior, although



$$p_0 = 0.02, p_1 = 0.085, \mathrm{ANOS}_0 = 1900$$

Figure 2.12: (a, b) Probability of signal and (c, d) difference of probability of signal for shifts of limited duration for the Bernoulli scan statistic chart with $k = 4$ and $m = 38$ and the Bernoulli CUSUM chart (a, c) with $r = 27$ and $h = 26/9$ and (b, d) with $r = 26$ and $h = 37/13$.



$$p_0 = 0.02, p_1 = 0.046, \mathrm{ANOS}_0 = 1900$$

Figure 2.13: (a, b) Probability of signal and (c, d) difference of probability of signal for shifts of limited duration for the Bernoulli scan statistic chart with $k = 7$ and $m = 142$ and the Bernoulli CUSUM chart (a, c) with $r = 38$ and $h = 74/19$ and (b, d) with $r = 39$ and $h = 157/39$.

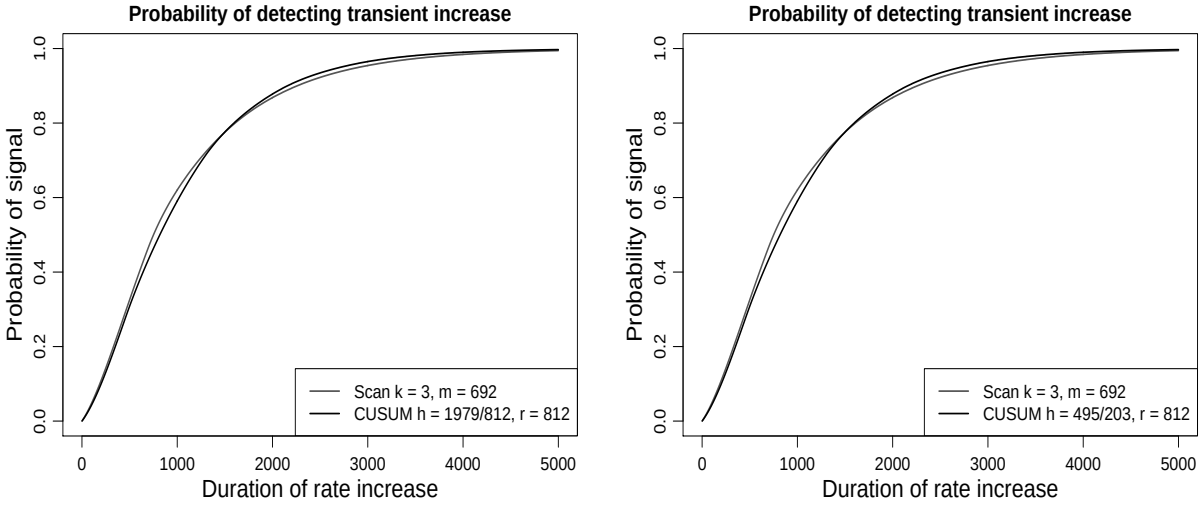
generally by less than a percentage point. There is a small difference in the performance of the suboptimal CUSUM chart against the optimal scan statistic chart, in that the CUSUM chart does not perform quite as well for the smaller values of d.

Figure 2.14 provides a comparison when c is increased to 10000, with $p_0 = 0.001$ and $p_1 = 0.0032$ (the sustained case is considered in Figure 2.7). Again, the probabilities associated with the scan method are determined by simulation. This is the only comparison in which the CUSUM chart is somewhat less likely to detect shifts for small values of d. The reason for this is not known, but may be because p_0 is so small. The optimal CUSUM chart is only superior for shifts of duration $d > 1484$.

When $p_0 = 0.02,\ p_1 = 0.09$, and $c = 10000$ (compared earlier in Figure 2.8), there is a larger difference in the r values chosen for the optimal and suboptimal CUSUM charts than in the other pairs of CUSUM charts. This allows us to see some of the impact of the r value. The comparisons of these charts against the optimal scan statistic chart are shown in Figure 2.15. The comparisons in this figure are based on the exact signal probabilities for the CUSUM charts and the simulated probabilities of signal for the scan statistic chart. The optimal CUSUM chart is substantially more likely to signal than the optimal scan statistic chart (by as many as four percentage points) for shifts of duration $1 < d \leq 57$. The optimal CUSUM chart is also more likely to signal than the optimal scan statistic chart when $d > 87$.

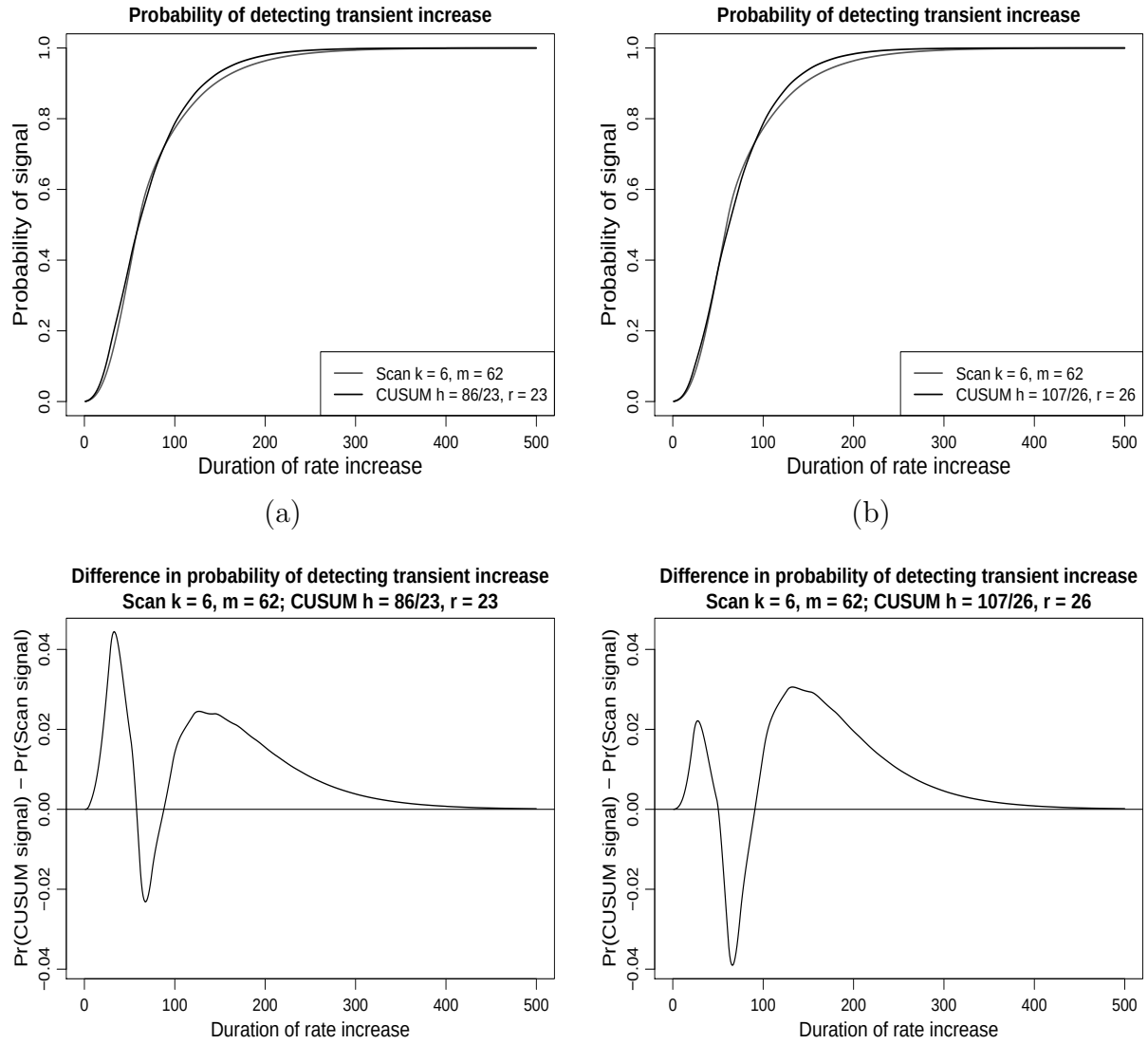
In Figure 2.15, the suboptimal CUSUM chart does not perform quite as well against the optimal scan statistic chart for short shifts ($d \leq 50$) and performs considerably worse for midrange values of d ($50 < d \leq 90$). However, the suboptimal CUSUM chart "makes up some ground" for larger values of d. The suboptimal CUSUM chart attains attains a difference in probability of signal of 0.03, whereas the optimal CUSUM chart does not.

The optimal and suboptimal CUSUM charts for the case of $p_0 = 0.02,\ p_1 = 0.048$, and $c = 10000$ (first presented in Figure 2.9) are much more similar in performance with respect to probability of signal. These charts are compared against the optimal scan statistic chart (for which the signal probabilities were determined by simulation) in Figure 2.16. The



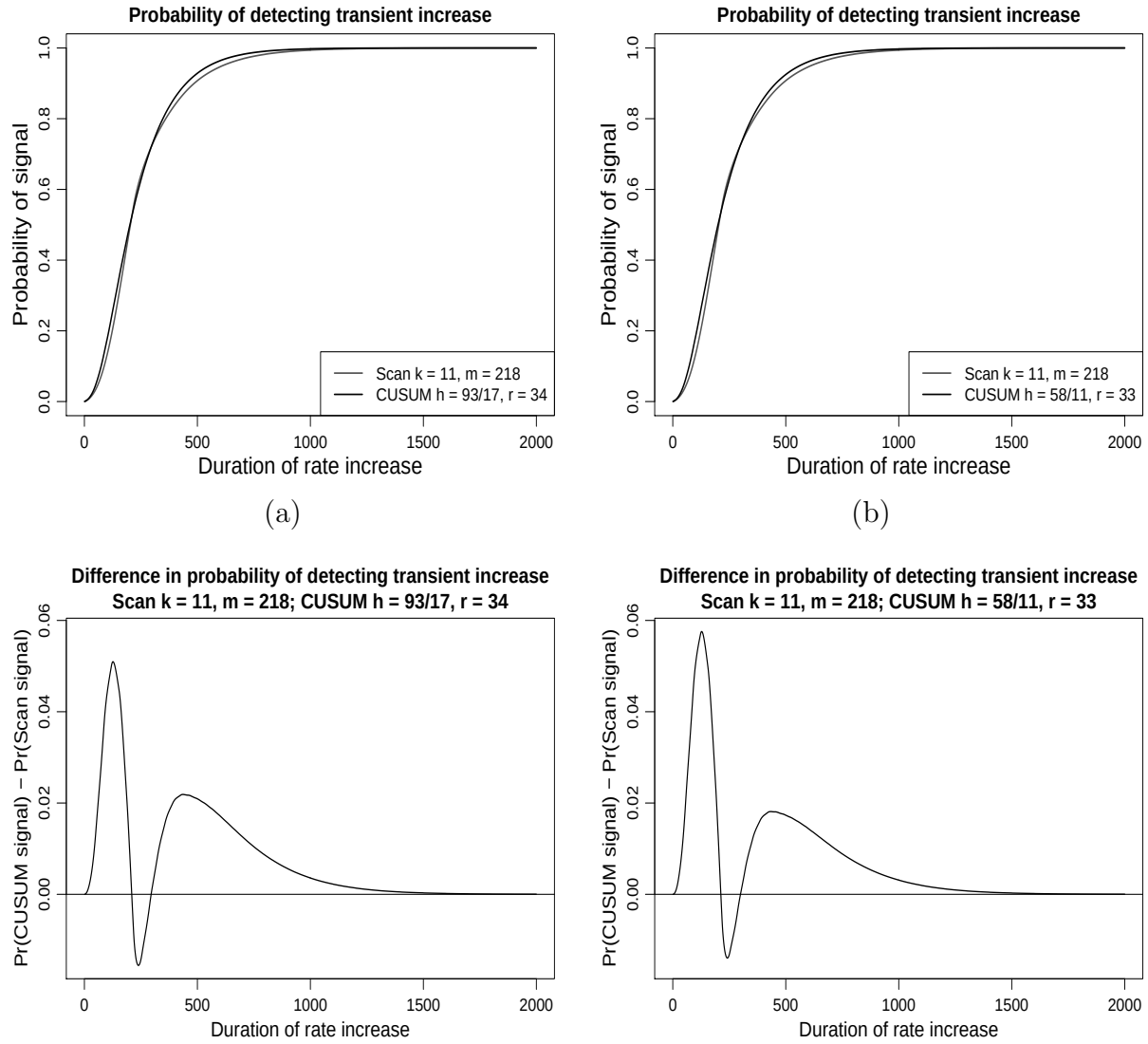
$$p_0 = 0.001, p_1 = 0.0032, \mathrm{ANOS}_0 = 10000$$

Figure 2.14: (a, b) Probability of signal and (c, d) difference of probability of signal for shifts of limited duration for the Bernoulli scan statistic chart with $k = 3$ and $m = 692$ and the Bernoulli CUSUM chart (a, c) with $r = 812$ and $h = 1979/812$ and (b, d) with $r = 812$ and $h = 495/203$.



$$p_0 = 0.02, p_1 = 0.09, \mathrm{ANOS}_0 = 10000$$

Figure 2.15: (a, b) Probability of signal and (c, d) difference of probability of signal for shifts of limited duration for the Bernoulli scan statistic chart with $k = 6$ and $m = 62$ and the Bernoulli CUSUM chart (a, c) with $r = 23$ and $h = 86/23$ and (b, d) with $r = 26$ and $h = 107/26$.



$$p_0 = 0.02, p_1 = 0.048, \mathrm{ANOS}_0 = 10000$$

Figure 2.16: (a, b) Probability of signal and (c, d) difference of probability of signal for shifts of limited duration for the Bernoulli scan statistic chart with $k = 11$ and $m = 218$ and the Bernoulli CUSUM chart (a, c) with $r = 34$ and $h = 93/17$ and (b, d) with $r = 33$ and $h = 58/11$.

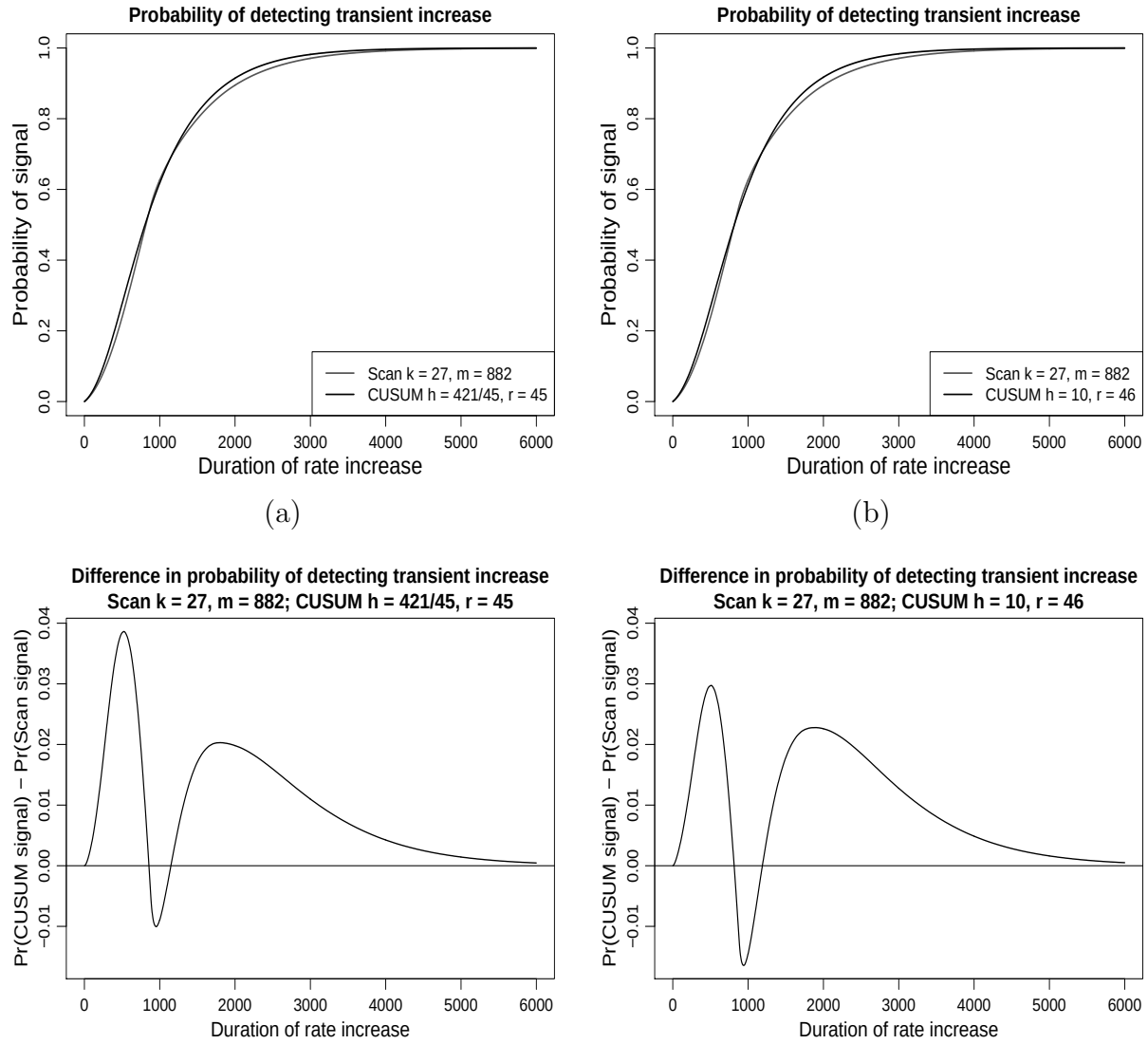
optimal CUSUM chart has a greater probability of signal than the optimal scan statistic chart for $d \leq 209$ and $d > 295$. The suboptimal CUSUM chart has a greater signal probability than the optimal scan statistic chart when $1 < d \leq 212$ and $d > 299$.

Figure 2.17 compares the monitoring methods for $p_1 = 0.029$ (retaining $p_0 = 0.02$ and $c = 10000$, as used in Figure 2.10). The probabilities associated with the scan statistic method are obtained by simulation. The optimal CUSUM chart has a larger probability of signal than the optimal scan statistic chart for $d \leq 857$ and $d > 1148$, whereas the suboptimal CUSUM chart has a greater probability of signal than the optimal scan statistic chart when $d \leq 814$ and $d > 1190$. In both of these comparisons it is clear that the scan statistic chart is less likely to signal for large values of d.

The final comparison is for the case where c is decreased to 500 and $p_0 = 0.02$, $p_1 = 0.065$ (this combination was evaluated earlier in Figure 2.11.) The signal probabilities for both methods were determined by applying Markov chain theory, and are therefore exact. The comparison for shifts of limited duration is made in Figure 2.18. Here the optimal CUSUM chart has a larger signal probability than the optimal scan statistic chart for $d \leq 34$ and $d > 52$, and the suboptimal CUSUM chart has a larger signal probability than the optimal scan statistic chart for $d \leq 31$ and $d > 58$. Indeed, the plot of differences in signal probabilities, although somewhat erratic, indicate fairly similar performance (less than one percentage point difference) for shifts lasting for less than roughly 60 observations.

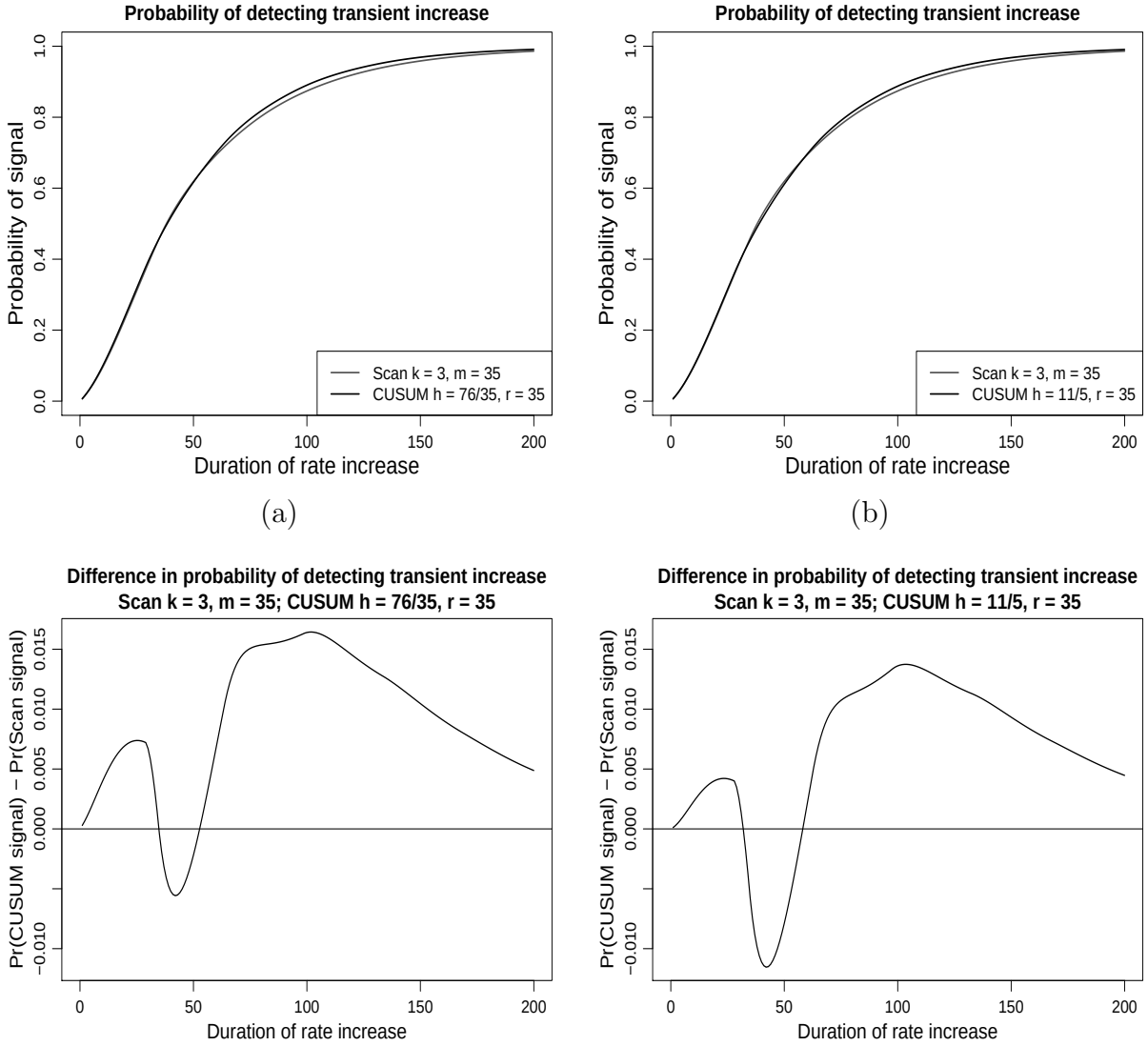
In all of these comparisons, the CUSUM charts have greater probability of indicating a rate increase during the increase if d is large. The CUSUM charts also tend to have greater signal probability when the duration is very short, although the probability of obtaining a signal from either method in such cases is very small. On the other hand, in these comparisons the scan statistic method is more likely to detect an increase in the incidence rate when d is in the neighborhood of m.

There is some additional information to be considered in the steady-state median number of observations to signal (SSMNOS). This value is defined as the smallest t such that



$$p_0 = 0.02, p_1 = 0.029, \mathrm{ANOS}_0 = 10000$$

Figure 2.17: (a, b) Probability of signal and (c, d) difference of probability of signal for shifts of limited duration for the Bernoulli scan statistic chart with $k = 27$ and $m = 882$ and the Bernoulli CUSUM chart (a, c) with $r = 45$ and $h = 421/45$ and (b, d) with $r = 46$ and $h = 10$.



$$p_0 = 0.02, p_1 = 0.065, \mathrm{ANOS}_0 = 500$$

Figure 2.18: (a, b) Probability of signal and (c, d) difference of probability of signal for shifts of limited duration for the Bernoulli scan statistic chart with $k = 3$ and $m = 35$ and the Bernoulli CUSUM chart (a, c) with $r = 35$ and $h = 76/35$ and (b, d) with $r = 35$ and $h = 11/5$.

the probability that a rate increase is detected by observation t is at least 0.5. This value is more relevant to the sustained shift case. However, a comparison of the SSMNOS and the SSANOS does reveal information about the skewness of the distribution of t. Indeed, there are several cases for which the use of SSANOS values may imply that, say, the Bernoulli CUSUM chart is better, but the use of SSMNOS values suggests the scan statistic method is better. In the cases considered here, the detection probability of the CUSUM chart is no more than 1.5% below the detection probability of the scan statistic chart for a transitory increase lasting for some duration d whenever the CUSUM chart has better SSMNOS performance. The SSMNOS and SSANOS values for all of the comparisons are provided in Table 2.2 as both a reference and a summary.

2.4 Discussion

In a monitoring context, public health officials need to know about increases in incidence rates of undesirable events as soon as possible after an increase occurs. This allows for appropriate action to be taken. An exact method for obtaining the SSANOS values of the scan method was developed in this chapter. Further, Section 2.5 introduced a technique for simulating the SSANOS when the exact solution is computationally difficult. The results of the study indicate that the Bernoulli CUSUM chart tends to be superior to the Bernoulli scan statistic chart when the chart parameters are chosen so that the charts are optimal in detecting a specified sustained shift in the parameter. However, the scan statistic chart would be easier to use in practice once the chart parameters have been obtained because less computation is involved. Neither chart has a particular advantage concerning the ease of determining parameters when the goal is to acheive a specified value of the in-control ANOS and the minimize the SSANOS_1. Therefore the choice of chart becomes a tradeoff between simplicity (in the scan method) and a slight reduction in SSANOS_1 values (for the CUSUM).

Furthermore, when the increase in rate is temporary, there can be no clear recommendation concerning which monitoring method is to be used. This is because the method that

Table 2.2: Chart parameters, SSANOS values, and SSMNOS values for specified p_0, p_1, and c.

			Scan		Opt. CUSUM	Sub. CUSUM
				k, m	h, r	h, r
			SSANOS	se	SSANOS[ae]	SSANOS[ae]
p_0	p_1	c	SSMNOS	se	SSMNOS[ae]	SSMNOS[ae]
0.0200	0.0850	1900		4, 38	26/9, 27	37/13, 26
			55.73	[e]	53.15	53.68
			42	[e]	44	44
0.0200	0.0460	1900		7, 142	74/19, 38	157/39, 39
			159.98	0.13	152.97	153.53
			127	[na]	124	125
0.0010	0.0032	10000		3, 692	1979/812, 812	495/203, 812
			1039.57	0.21	1037.43	1037.90
			751	[na]	818	818
0.0200	0.0900	10000		6, 62	86/23, 23	107/26, 26
			75.83	0.05	71.73	72.17
			60	[na]	61	63
0.0200	0.0480	10000		11, 218	93/17, 34	58/11, 33
			253.61	0.18	239.15	239.80
			204	[na]	202	201
0.0200	0.0290	10000		27, 882	421/45, 45	10, 46
			1027.87	0.39	971.19	973.00
			817	[na]	805	818
0.0200	0.0650	500		3, 35	76/35, 35	11/5, 35
			52.25	[e]	50.22	50.76
			39	[e]	39	39

[e]This value is exact.
[ae]All values in this column are exact.
[na]The error estimate was not computed.

is most likely to signal the rate increase depends on the duration of the rate increase, which in some cases may be unknown.

One question that arises when the rate increases for a limited duration is the appropriateness of choosing chart parameters that minimize the $SSANOS_1$. It may be more appropriate to minimize the out-of-control steady-state median number of observations to signal, or to maximize the signal probability given a specific duration for the rate increase. Such analyses would require additional simulation and are not pursued here.

In some applications, the results of Bernoulli trials may not come in the same order that the trials are performed. For example, a surgical procedure may be evaluated based on mortality rates and the order of deaths of patients might not match the order in which their operations occurred. A similar result can occur when diagnosing disease if laboratories do not test samples in the order they are received. This would be problematic for any existing method because monitoring tools make the assumption that the results are reported in the order that the tests are performed. Monitoring methods (including the ones discussed in this chapter) are not designed to work in the situation for which the time required to obtain results widely varies.

Furthermore, results of the trials should be considered as they become available. The trials should never be grouped into batches unless this is unavoidable. Ismail *et al.* (2003) indicated that monitoring methods are more sensitive to changes in an incidence rate when observations are considered individually rather than in batches. Reynolds and Stoumbos (1999, 2000) demonstrated this for several methods designed to monitor incidence rates. It is important to note that the methods used in this chapter are not designed to be applied to grouped data. Each observation should be considered individually as it occurs.

2.5 Appendix: Exact Bernoulli scan statistic SSANOS computation

It is possible to represent the Bernoulli scan statistic chart using a Markov chain. To model any method with a Markov chain, one must define a set of *states*. In the case of the Bernoulli CUSUM chart, these states are directly associated with the values of C_i such that $0 \leq C_i < h$. (The possible values of C_i are finite if an integer value is used for r.)

An analogous, but more complicated, approach is needed to model the Bernoulli scan statistic chart with a Markov chain.

A scan statistic chart which has not yet signaled must have $k-1$ or fewer incidences in the past m opportunities. (Thus the possible values of S_i are the integers between zero and $k-1$.) At any time i, the change in S_i from the previous value S_{i-1} will result in one of four outcomes:

1. S_i increases by 1 to k (resulting in a signal) if $S_{i-1} = k-1$ and Y_i is an incidence, regardless of the value of Y_{i-m+1}.

2. S_i increases by 1 to a value less than k if $S_{i-1} < k-1$, Y_i is an incidence, and Y_{i-m+1} is either a non-incidence or an undefined value (Y_{i-m+1} is undefined if $i-m<0$).

3. S_i is unchanged if both Y_i and Y_{i-m+1} are incidences, or both Y_i and Y_{i-m+1} are non-incidences.

4. S_i decreases by 1 if Y_i is a non-incidence and Y_{i-m+1} is defined and is an incidence.

Because of items 2 through 4, the values of Y for time indices $i-k$ through $i-1$ must be known in order to monitor the behavior of S_i using Markov chains. Thus the definition of the Markov states must involve not the values of S_i but rather the $(m-1)$-vectors associated with the possible outcomes of the most recent $m-1$ opportunities. The outcomes of each of these opportunities must be known because each will eventually be the critical Y_{i-m+1} value

(since additional data will be collected). After any given observation when using a Markov chain based on the scan statistic, one must move from a particular state to one of two states depending on whether or not the new observation is an incidence. (The two states that can be reached must be consistent with the outcomes of the preceding $m-1$ observations.) Again, the states for modeling the scan statistic are associated with vectors of length $m-1$. Each of the states yields a particular value of S_i. Those states which are associated with values of $S_i < k$ are called *transient states*.

The upper bound on the total number of transient states required is

$$N = \sum_{i=0}^{k-1} \left(\begin{array}{c} m-1 \\ i \end{array} \right).$$

One of these states corresponds to the absence of any incidences in the past $m-1$ opportunities. Here N is an upper bound because there are some states with incidences that cannot contribute to a signal. Incidences that cannot contribute to a signal can be considered equivalent to non-incidences. This will result in some states being collapsed together.

Associating states with the past $m-1$ values of Y is very similar to the assignments used in the Markov chain model used by Champ and Woodall (1987) to evaluate the performance of control charts with runs rules. As an example, Table 2.3 enumerates all possible $(m-1)$-vectors for a $k=3$, $m=5$ Bernoulli scan statistic chart and assigns a state number to each of the vectors. Note that only the vectors representing transient states are given.

To find the exact SSANOS, a Markov transition matrix, $\mathbf{Q}$, must be created. Each row and each column of $\mathbf{Q}$ represents one of the N possible states. The element in the ath row and bth column of the transition matrix is the probability of moving from state a to state b. For any of the transient states, there are two possible states that can be reached after the next data point is collected. For a given time i, these states are determined by appending the result of the next trial, Y_i, (either 0 or 1) to the right side of the $(m-1)$-dimensional vector representing the present state. If an incidence occurs, then a check should be made to see if there are now k incidences (this is the distinction between items 1 and 2 in the list given earlier). If there are, the chart will signal, and there will be no need to determine the future

Table 2.3: Possible transient states for a Bernoulli scan statistic chart with $k = 3$, $m = 5$.

State	Last observations*
1	0 0 0 0 *or* 1 0 0 0
2	0 1 0 0
3	0 0 1 0
4	0 0 0 1
5	1 1 0 0
6	1 0 1 0
7	1 0 0 1
8	0 1 1 0
9	0 1 0 1
10	0 0 1 1

*A 1 indicates that an incidence occurred. These indicator variables are ordered with the most recent observation appearing to the right.

state of the chart. Otherwise, the result of the oldest opportunity (the first element of the m-vector) is removed to get a new $(m-1)$-vector. One then determines the state number which identifies the listing of incidences for both of the possibilities: when an incidence occurs and when one does not. The probabilities of moving to these states are assigned to the appropriate elements in $\mathbf{Q}$. Table 2.4 contains the probabilities of moving from any state a to any state b for the $k = 3$, $m = 5$ example.

It is well-known that the average number of observations required to signal from a state i can be found by taking the ith element of $\boldsymbol{\mu} = (\mathbf{I} - \mathbf{Q})^{-1}\mathbf{1}$, where $\mathbf{I}$ is an $N \times N$ identity matrix and $\mathbf{1}$ is a N-vector of ones. The ANOS referred to throughout this chapter is the average number of observations required to signal assuming the rate increase occurs when effectively no incidences have occurred in the past $m-1$ opportunities. As mentioned earlier, this is not a realistic assumption. Instead, a steady-state analysis should be done. The first step is to determine the average number of observations required to signal a rate increase for all of the transient states. The second step is to determine the approximate probability of being in each of these states when the rate increase occurs. As described by Reynolds and Stoumbos (2000), a vector of these probabilities can be obtained by determining the

Table 2.4: Markov transition probabilities for a Bernoulli scan statistic chart with $k = 3$, $m = 5$.

	To state:									
From:	1	2	3	4	5	6	7	8	9	10
1	$1-p$	0	0	p	0	0	0	0	0	0
2	$1-p$	0	0	0	0	0	p	0	0	0
3	0	$1-p$	0	0	0	0	0	0	p	0
4	0	0	$1-p$	0	0	0	0	0	0	p
5	$1-p$	0	0	0	0	0	0	0	0	0
6	0	$1-p$	0	0	0	0	0	0	0	0
7	0	0	$1-p$	0	0	0	0	0	0	0
8	0	0	0	0	$1-p$	0	0	0	0	0
9	0	0	0	0	0	$1-p$	0	0	0	0
10	0	0	0	0	0	0	0	$1-p$	0	0

eigenvector associated with the largest eigenvalue of $\mathbf{Q}_0$, which is $\mathbf{Q}$ as determined when assuming the in-control incidence rate p_0. This dominant eigenvector is normalized and then multiplied by $\boldsymbol{\mu}$ to yield the SSANOS.

The matrix inversion was not possible (with the department's computing resources) when the Markov chain approach required a very large number of states. In these cases, approximate ANOS and SSANOS values were obtained using simulation. The simulation can be made significantly more efficient by using the geometric distribution and using the relationship between the geometric and binomial distributions (as discussed by Reynolds and Stoumbos (1999)). The geometric distribution can be used to model the number of independent Bernoulli trials that must occur before a "success" occurs (see, for example, Section 3.2 of Casella and Berger, 2002). Therefore, rather than repeatedly simulating the Bernoulli distribution to determine the outcome of each opportunity, it is easier to determine the observation numbers that indicate incidences by generating independent geometric random variables. When finding the SSANOS when an increase occurs, a rate increase is simulated by increasing p to some larger value after first simulating a specified number of observations from an in-control period. In the context of the Bernoulli scan statistic chart, one can compare the observation numbers of the most recent and the kth most recent inci-

dences. When the range of these two values is strictly less than m, then the scan statistic chart will signal a rate increase. The observation number for the incidence that results in the signal is recorded. The simulation is repeated many times and the average of these numbers of observations to signal is used to obtain an estimate of the SSANOS.

Chapter 3

Multivariate Control Charts and Spatial Surveillance

3.1 Introduction

In surveillance applications, data are frequently collected in multiple regions at regular time intervals. Often, an outbreak will be somewhat localized, but an increase might be spread across several regions. Monitoring each region separately (perhaps by using several separate control charts) does not allow for the possibility that some of the regions may be similar to each other in geographic location or other characteristics. In some surveillance applications, it is possible to obtain the exact spatial coordinates of each reported case. In other cases it is more practical to define regional boundaries and aggregate the data within these boundaries. The regional boundaries may be based on geographic features, political boundaries, postal codes, census tracts, or other criteria. The data to be analyzed may instead come from a variety of sources, such as hospitals. In this chapter, it is assumed that the data are spatially aggregated.

Elliott and Wartenberg (2004) discussed some of the developments in spatial methodology, including spatial surveillance techniques. Some methods for prospective spatial surveil-

lance have been presented or summarized by Raubertas (1989), Rogerson (1997, 2001), and Kulldorff (2001), among others. A good review of spatial surveillance methods is given by Shmueli and Fienberg (2006). Many of the present problems and challenges in surveillance are discussed in Shmueli (2007). Spatial surveillance has also received attention in several recent books, including Wilson *et al.* (2006), Lawson and Kleinman (2005), Brookmeyer and Stroup (2004), and Lawson (2001).

It is possible to combine the information from each of several regions into a single control chart. Such charts are called multivariate control charts. Rogerson and Yamada (2004) have compared the use of individual univariate charts with the use of a multivariate chart for monitoring spatial disease patterns. They have also discussed some of the issues involved in this monitoring problem.

It is assumed that there is an expected rate of incidence for each region based on demographic and possibly other data. This rate can be determined based on some appropriate existing methodology. During each data collection period, a certain number of cases are observed in each region. These observed counts can be scaled by the population size of the respective regions to produce an observed incidence rate. If the observed incidence rate exceeds the expectation, then some evidence has been provided that an outbreak has occurred. If such a result occurs repeatedly, or if the incidence rate is much larger than the expectation, then the result is significant and the proper conclusion is that the incidence rate has increased. It is important to be able to detect an outbreak as quickly as possible.

It is also necessary to make some assumptions about the nature of the regions and how the incidence count in one region is related to the incidence count in another region. It is important that the data be collected at a regular interval. In some applications, data may be available daily or weekly. More often, data will be collected on a monthly or quarterly basis. In this chapter, data will be required from all regions at each time period.

The length of the reporting time period and the size of the regions being monitored affect the ability to detect an outbreak condition in a timely manner. Haining (2003, p. 61) and

Shmueli and Fienberg (2006) discussed this issue further. They explained that if the data are gathered too frequently and/or the data are excessively localized, the regional counts in each sampling interval may yield low counts. In this situation, it is difficult to maintain an adequate false-alarm rate while still maintaining power to detect an increase when one occurs. A researcher might attempt to resolve this problem by increasing the time between reporting periods or by increasing the size of the regions. In this case, a localized rate increase (in time or space) could be harder to detect if there is not an increase in some parts of a region or in some times in the reporting cycle.

In Section 3.2, methods for examining the spatial structure of the data are discussed. Then, in Section 3.3, the most common multivariate control charts are introduced. A discussion of some issues concerning their use in spatial surveillance follows. Next, a modification of the MEWMA control chart of Lowry *et al.* (1992) is proposed. In Section 3.4, the performance of this modified chart is compared with the charts suggested by Rogerson and Yamada (2004). In Section 3.5, issues related to spatial surveillance when using multivariate control charts are discussed.

3.2 Spatial Structure

There are several types of spatial processes. These processes are best understood when the locations of the incidences are thought of in the continuous sense. These locations can later be aggregated by establishing particular boundaries within the *study area*. In many cases, this aggregation is necessitated by privacy concerns or limited information on location. The choice of the boundaries of these regions can greatly influence the ability to detect the increases. This has been referred to as the modifiable areal unit problem (for a more information on this problem, see Section 4.5 of Waller and Gotway, 2004).

Sometimes the counts taken from the regions can be interrelated. In this case it will be necessary to estimate the spatial correlation structure.

In this section, some of the types of spatial processes are described. Then some methods for estimating the spatial covariance structure will be given.

3.2.1 Types of spatial processes

The homogeneous spatial Poisson process assumes that 1) in any study area the number of incidences occurring within a given time frame follows a Poisson distribution, and 2) the exact locations of these incidences represent an independent random sample of locations within the study area. This definition implies that there is a constant baseline incidence rate at all locations in the study area.

However, it is rarely true that the incidence rate is the same at all locations in the study area because the population density generally varies within the study area. This is remedied by defining the heterogeneous spatial Poisson process. This process requires knowledge of a function that defines the expected number of incidences at each location in the study region. The exact locations of the incidences are then independently allocated in proportion to this function. When incidences are aggregated spatially, it is important to note that this definition simply allows the expected number of incidences within the regions to vary. This definition does not induce an underlying correlation between regions.

The Cox process is a variation on the heterogeneous Poisson process, which uses a random function to represent the expected number of incidences over the locations in the study region. Variations of the Cox process can result in spatial correlation.

Waller and Gotway (2004, p. 202) mention that some analysts avoid use of the Poisson distribution because it assigns a nonzero probability to observing more cases than there are people in each region. However, this should rarely present a problem unless the incidence rate is fairly large.

The methods in this chapter apply to both the homogeneous and heterogeneous Poisson processes. The methods also allow for spatial dependencies among the regions.

3.2.2 Estimating spatial relationships

As mentioned, the homogeneous and heterogeneous spatial Poisson processes are based on the assumption that the realizations of the regions are independent of each other. If the incidence rate increases in one or more of the regions, the increase would be reflected in both the mean and in the apparent correlation structure in the study area. This has been referred to as "first-order clustering" (Kulldorff *et al.*, 2003; Waller and Gotway, 2004, p. 259). In many cases, however, the counts of certain regions may depend on other (usually nearby) regions. The counts obtained from a study area containing such spatially dependent regions may inherently contain clusters, and this has been called "second-order clustering" (Kulldorff *et al.*, 2003; Waller and Gotway, 2004, p. 259). In this case, one would expect values at adjacent locations to be similar. Haining (2003) states that this is "the usual expectation." If such similarities (second-order clustering) are believed to exist, then it is important to decide how to estimate the nature of the spatial association.

The multivariate control charts that are considered later in this chapter can account for relationships between the regional counts within the study area. This is accounted for through the $n \times n$ covariance matrix $\mathbf{\Sigma}$, in which the (i, j) element represents the covariance between the counts in regions i and j, and the (i, i) element represents the variance of the count in region i. If the counts are standardized, then $\mathbf{\Sigma}$ becomes a correlation matrix with a diagonal of unity and correlations on the off-diagonal elements.

These correlations can be difficult to estimate. Therefore, it is common to use systematic methods to model the spatial proximity of any pair of regions. There are parameters for each of these methods for estimating spatial association. It is not at all clear how these parameters ought to be chosen. Also, there does not seem to be any indication as to the scenarios in which any one of these methods ought to be used. These are not questions which will be addressed in this dissertation. However, the various techniques will be listed here.

One approach is taken by Rogerson and Yamada (2004). They employ a method called rooks' case adjacency. The $\mathbf{\Sigma}$ matrix is defined such that the ith and jth regional counts

have correlation equal to ρ^c, where c is the smallest number of boundary lines that must be crossed to travel between the ith and jth regions. The "rooks' case" (or "rooks' move") name is used to represent the movements of the rook in chess, for which only horizontal or vertical moves are allowed. The rooks' move idea is addressed by Haining (2003, p. 78), although he only discusses the determination of c.

Another option is the binary connectivity matrix, in which the (i, j) element is defined as a positive value if regions i and j share a boundary, and zero otherwise. Haining (2003) points out that if the binary connectivity matrix is multiplied by itself, the resulting matrix will indicate all of the regions that could be reached in two steps; multiplying this matrix by the original binary connectivity matrix yet again indicates all of the regions that can be reached in three steps, etc.

In another method, the (i, j) element of the matrix is defined as a positive value if the centers of regions i and j are within some specified distance of each other; otherwise, the (i, j) element is set to zero. This positive value could be a constant, or it could be a function of the distance between the centers of regions i and j. Usually such a function consists of raising the distance between the regions to a negative power, although other functions are possible, as discussed by Haining (2003).

A variant of this approach sets the (i, j) element to a positive value if the center of region j is one of the c closest to the center of region i. However, this approach does not necessarily yield a symmetric matrix. This approach also behaves differently for a region that is on the edge of the study area than it does for a region in the interior of the study area.

Another approach (which also produces a non-symmetric matrix) involves setting the (i, j) elements of the matrix proportional to the percentage of the perimeter of region i that is shared with region j. The percentage could also be raised to a positive power. Any portion of the perimeter of region i that is on the border of the study area is usually excluded from this calculation.

The binary connectivity matrix method, the centroid-based matrix method, and the perimeter proportion method were all suggested by Waller and Gotway (2004) and Haining (2003), who used them as weights for methods that compute an overall ("global") index of spatial association for a single realization of values from the regions of the study area (not in a surveillance or monitoring system). Haining (2003) pointed out that the use of such systematic methods is most useful when there is reason to suspect relationships among nearby areas but no external information is available which could be used to determine the nature of such spatial association.

Finally, it is not clear whether to assume second-order clustering. Perhaps the easiest way to address this question is to obtain some data from a time period in which no cluster of disease is believed to exist, and then estimate the correlation structure from those data.

3.3 Multivariate Control Charts and Their Use

3.3.1 Common multivariate control charts

Assume that disease count data are collected at discrete time periods $t = 1, 2, \cdots$ from regions (or other sources) labeled $i = 1, \cdots, n$. The count for region i at time period t is denoted by $X_{i,t}$ and assumed to follow a Poisson distribution. At each time period, the $X_{i,t}$ values can be combined into a vector $\mathbf{X}_t = (X_{1t}, X_{2t}, \cdots, X_{nt})$. The expected count of each region when there has been no outbreak, $\mu_{i,0}$, must be specified. These can be aggregated into the vector $\boldsymbol{\mu}_0 = (\mu_{10}, \mu_{20}, \cdots, \mu_{n0})$. We assume that each of these Poisson means are sufficiently large to permit the use of normal approximations to the Poisson distributions. This same assumption is implied by Rogerson and Yamada (2004). Because the $X_{i,t}$ follow Poisson distributions, the regional counts have different variances (unless all regions have the same expected counts), and the observations from the regions may be correlated to some extent with each other. This information is incorporated into a $n \times n$ covariance matrix, $\boldsymbol{\Sigma}$.

Rogerson and Yamada (2004) have evaluated the use of multiple control charts, with one for each region. They used CUSUM charts. The chart for the ith region is based on

$$C_{i,t} = \begin{cases} \max(C_{i,t-1} + X_{i,t} - r, 0) & \text{if } t > 0 \\ 0 & \text{if } t = 0, \end{cases} \tag{3.1}$$

where r is a parameter called the reference value. (Unlike the discussion in Chapter 2, r need not be an integer. Note also that r itself appears in this equation, rather than its reciprocal.) The CUSUM chart for the ith region signals if $C_{i,t}$ exceeds an alarm limit, h, which is chosen to produce a specified average time between false alarms. To detect an increase in any of the regions, one must maintain n of these charts. An increase is reported if any of these charts signals. This approach was evaluated in the quality control literature by Woodall and Ncube (1985) for observations drawn from the multivariate normal distribution.

Rogerson and Yamada (2004) suggested the use of the MC1 chart presented by Pignatiello and Runger (1990). The MC1 chart is a multivariate CUSUM chart. The construction of this chart is based on the assumption that the difference between the observed and expected values, $(\mathbf{X}_t - \boldsymbol{\mu}_0)$, is normally distributed. To compute the chart statistic, one must first compute

$$\mathbf{C}_t = \sum_{i=t-n_t+1}^{t} (\mathbf{X}_i - \boldsymbol{\mu}_0), \tag{3.2}$$

where

$$n_t = \begin{cases} n_{t-1} + 1 & \text{if } MC1_{t-1} > 0 \\ 1 & \text{otherwise.} \end{cases} \tag{3.3}$$

The chart statistic is

$$MC1_t = \begin{cases} \max\left\{\sqrt{\mathbf{C}_t' \boldsymbol{\Sigma}^{-1} \mathbf{C}_t} - r n_t, 0\right\} & \text{if } t > 0 \\ 0 & \text{if } t = 0. \end{cases} \tag{3.4}$$

The value r again acts as the reference value parameter. The MC1 chart signals if $MC1_t$ exceeds an alarm limit, h. The choice of r and h affects the chart's statistical performance.

When a univariate CUSUM chart is used to monitor the mean of a normal distribution, the optimal choice of r has been shown to be equal to one-half of the shift that the chart is intended to detect quickly. Because of this, a common choice for r for the MC1 chart is

$$r = 0.5((\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)), \tag{3.5}$$

where $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$ represents the shift of interest. Pignatiello and Runger (1990), however, indicated that this value is not necessarily optimal for the MC1 chart. Furthermore, the regional counts are assumed to be Poisson random variables. Therefore, this "optimal" choice for r is not fully justified, since the normal distributions are only approximate distributions for the regional counts.

There are CUSUM-based alternatives to the MC1 chart. For example, another multivariate CUSUM chart was presented by Crosier (1988). This MCUSUM chart has been shown to be similar to, but slightly less effective than, the MC1 chart. Therefore details on this chart are not presented here.

Stoto *et al.* (2006) modified the MCUSUM chart so that it would only respond to increases in regional counts. They then compared the performance of this modified MCUSUM chart with multiple univariate CUSUM charts. They found that both of these methods were preferred to more traditional Shewhart-based multivariate control charts. They also found that in certain circumstances the MCUSUM could be superior in performance, while in others the multiple univariate CUSUM charts could be superior in performance.

A competing chart to the CUSUM chart is the exponentially weighted moving average (EWMA) chart. The performance of the EWMA and CUSUM charts have been shown to be approximately equivalent in the univariate case (e.g., Lucas and Saccucci, 1990; Reynolds and Stoumbos, 2004*a*).

Lowry *et al.* (1992) suggested a multivariate extension of the EWMA chart. First, one computes

$$\mathbf{Z}_t = \begin{cases} \lambda(\mathbf{X}_t - \boldsymbol{\mu}_0) + (1-\lambda)\mathbf{Z}_{t-1} & \text{if } t > 0 \\ \mathbf{0} & \text{if } t = 0, \end{cases} \tag{3.6}$$

where $\mathbf{Z}_0 = \mathbf{0}$. The parameter $\lambda > 0$ is often called the smoothing parameter. It is known that $\mathbf{Z}_t$ has a covariance matrix equal to

$$\boldsymbol{\Sigma}_{\mathbf{Z}_t} = \frac{\lambda[1-(1-\lambda)^{2t}]}{2-\lambda}\boldsymbol{\Sigma}, \tag{3.7}$$

or, as $t \to \infty$,

$$\boldsymbol{\Sigma}_{\mathbf{Z}_\infty} = \frac{\lambda}{2-\lambda}\boldsymbol{\Sigma}. \tag{3.8}$$

Because of the principles of steady-state analysis, which are explained in Section 3.4.1, Equation 3.8 will be used. The MEWMA chart statistic is

$$MEW_t = \mathbf{Z}_t' \boldsymbol{\Sigma}_{\mathbf{Z}_\infty}^{-1} \mathbf{Z}_t \tag{3.9}$$

and a signal is generated if MEW_t exceeds an alarm limit, q. As with the MC1 chart, the choice of the parameters λ and q affects the statistical performance of the chart. It is commonly recommended that $\lambda = 0.2$ be used.

3.3.2 One-sided versus two-sided charts

It is well known that a one-sided statistical hypothesis test is preferable to a two-sided hypothesis test if the researcher is only concerned with an effect in one direction. An analogous property holds true with control charts as well. Rogerson and Yamada (2004) used one-sided control charts when they used individual (univariate) control charts, with a separate chart for each region. This is sensible because a decrease in the incidence rate would often not be of as much interest as an increase.

However, the fundamental multivariate control charts previously discussed can signal an alarm for many types of changes. Such changes may include only increases or only decreases in some regions, and may also include increases in some regions accompanied by decreases in others. A multivariate control chart is called "directionally invariant" if the statistical performance of the chart depends on $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ solely through the non-centrality parameter,

$$(\boldsymbol{\mu} - \boldsymbol{\mu}_0)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0), \tag{3.10}$$

where $\boldsymbol{\mu} - \boldsymbol{\mu}_0$ represents the change in the mean vector, and $\boldsymbol{\Sigma}$ is the applicable covariance matrix, which is usually assumed to remain constant. It follows that directionally invariant charts can signal quickly for shifts in any direction from the target vector. This property can be useful in certain settings, but is not desirable for monitoring disease rates, as one should (generally) not be alarmed by a decrease in a disease rate. Some (e.g., Pignatiello and Runger, 1990) have discussed this principle of directional invariance in greater detail.

The principle of directional invariance is useful because it is easier to evaluate the statistical performance of a chart that has this property. However, as discussed in Section 3.4.2, the covariance matrix changes after a shift in the mean vector when Poisson distributed counts are being monitored. Furthermore, it is not desirable to detect shifts in all directions in a surveillance program intended to detect only increases in disease incidence.

Since it is not of primary importance here to detect decreases in incidence rates, the ideal approach would involve a multivariate control chart that is analogous to a one-sided univariate control chart. One of the approaches of Fassò (1999) would generate an alarm for an increased rate in at least one region, but only under the condition that none of the parameters being monitored have decreased. Specifically, he suggests the use of the hypotheses:

$$\begin{aligned} H_0 &: \boldsymbol{\mu} = \mathbf{0} \qquad \text{(without loss of generality)} \\ H_1 &: \boldsymbol{\mu} = \begin{cases} \mathbf{0} & \text{until process shifts at some unknown time} \\ \boldsymbol{\mu} \in \boldsymbol{\Omega}^+ & \text{after process shifts,} \end{cases} \end{aligned} \tag{3.11}$$

where $\boldsymbol{\Omega}^+$ is the "only-increase" set of the parameter space, defined by

$$\boldsymbol{\Omega}^+ = \{\boldsymbol{\mu} = (\mu_1, \cdots, \mu_n)' : \mu_j \geq 0, \boldsymbol{\mu}'\boldsymbol{\mu} > 0\}. \tag{3.12}$$

The problem with this approach is that the resulting modified MEWMA chart is designed to signal an increase in the incidence rate under the assumption that the rate did not decrease in any of the regions. This is overly restrictive, as the disease rate may increase in some regions while decreasing elsewhere. It is important to detect an increase in the count of any region, regardless of any decrease in the counts of other regions. Fassò (1998) has

also introduced and discussed a hypothesis testing approach which will detect any increase, but this approach has not yet been implemented in the context of control charts.

Testik and Runger (2006) followed the "only-increase" approach. They also provided a variant in which some of the parameters (here these represent regional counts) are permitted to increase or decrease, but the remaining parameters, however, are still subject to the "only-increase" approach used by Fassò (1999). (Essentially, this means that some of the parameters are treated as "two-sided" while others are treated as "one-sided.") Testik and Runger (2006) also mentioned another approach, in which the statistic

$$P = \frac{\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{X}_t}{\sqrt{\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1}}} \tag{3.13}$$

is used. A signal is given for large, positive values of P. This method is designed to signal that there has been an increase in the incidence rate of at least one region as long as the counts recorded in $\mathbf{X}_t$ are such that $\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{X}_t$ is positive.

Sonesson and Frisén (2005) have suggested an alternative to the MEWMA chart introduced by Lowry *et al.* (1992). Sonesson and Frisén pointed out that the MEWMA chart statistic MEW_t is a quadratic form of a vector of individual statistics and an appropriate covariance matrix. They proposed using individual upper CUSUM statistics in a similar quadratic form. The choice of these statistics results in a chart that would detect increases in the rates of some regions without being adversely affected by the presence of decreases in the rates of other regions.

Stoto *et al.* (2006) also used only the upper CUSUM statistics as inputs to the MCUSUM chart, so that decreases in rates would not contribute to producing an alarm. They stated that there would be greater power in detecting increases in rates.

3.3.3 The proposed one-sided MEWMA chart

The following proposal is similar to that of Sonesson and Frisén (2005), who suggested the use of upper CUSUM statistics, and Stoto *et al.* (2006), who modified the MCUSUM chart.

In this proposed method, a "barrier" is placed on the EWMA statistics (as computed in Equation 3.6),

$$\mathbf{Z}_t = \begin{cases} \max\{\lambda(\mathbf{X}_t - \boldsymbol{\mu}_0) + (1-\lambda)\mathbf{Z}_{t-1}, \mathbf{0}\} & \text{if } t > 0 \\ \mathbf{0} & \text{if } t = 0, \end{cases} \tag{3.14}$$

where the maximum operator refers to an element-wise comparison of the two vectors. This should be used in conjunction with Equations 3.8 and 3.9. Since no element of $\mathbf{Z}_t$ will ever be less than zero, the resulting chart will tend to signal only as a result of increases in the observed incidence rate of one or more regions.

This modification is relatively simple. Note that $\boldsymbol{\Sigma}_{\mathbf{Z}_t}$, as defined in Equation 3.8, becomes an approximation under this procedure.

3.3.4 Buildup of credit

The control charts presented in this chapter combine information over time to more efficiently detect small increases in the disease incidence rate. When one of these charts is used, the chart statistic may accumulate information which can result in a delay in the time required to signal that the disease incidence rate has increased. This property is referred to as the building up of "credit" in literature applying control charts to health care (e.g., Grigg and Farewell, 2004). In the quality control literature, this is called the "inertia problem."

Woodall and Mahmoud (2005) thoroughly investigated the inertia properties of several different control charts. They showed that "the MC1 chart can build up an exceedingly large amount of inertia." Runger and Testik (2004) also pointed out that the MC1 chart has considerable inertia. Because of the very high amount of inertia that can be generated by the MC1 chart, its use is not recommended in monitoring applications. The MCUSUM chart of Crosier (1988) does not suffer as significantly from the inertia problem.

There are at least two factors which contribute to a buildup of credit. First, a monitoring technique may be slow to indicate a sudden increase in disease rates if there has been

a gradual buildup of information that indicates a decrease in disease rates. Secondly, a technique may simply not give sufficient weight to the most recent data points. Runger and Testik (2004) also note that many forms of the MEWMA chart suffer from the inertia problem because of this issue. If recent data points are given a low enough weight (small λ in EWMA-based charts), a significant amount of data will be required in favor of the outbreak hypothesis, particularly if historic data have been strongly opposed to that hypothesis. The proposed one-sided MEWMA chart does not permit a buildup of as much of this type of credit.

3.4 Performance

3.4.1 Comparing different control charts

Control chart performance is often evaluated using the average run length, or ARL. The ARL is similar to the ANOS. Recall that the ANOS measures the average number of individual (Bernoulli) observations required to obtain a signal. For example, the ANOS could measure the average number of births until a certain congenital malformation is observed which causes the monitoring scheme to signal. These births need not be equally spaced in time, and the frequency of births may not necessarily be constant. The ANOS, therefore, does not suggest an average number of days, months, or any other time period until a signal is obtained. However, the ARL used here measures the average number of reporting periods (in this case, Poisson observations) until a signal is obtained.

To evaluate in-control ARL performance using simulation, values for $(\mathbf{X}_t - \boldsymbol{\mu}_0)$ are generated using a multivariate normal distribution whose mean vector is the zero vector. Computer simulations generate data until the chart signals, at which point the number of samples generated is recorded. Each simulation reported in Table 3.1 was replicated 50000 times; in the remaining tables, 100000 simulations were used in order to gain greater accuracy. The average of the number of samples required is then estimated. This is the

estimate of the ARL for an in-control condition. By trial-and-error, control limits are chosen so as to produce an in-control ARL equal to 100. Because of simulation error, however, it is not possible to guarantee that the selected chart parameters are the exact parameters that produce an in-control ARL equal to 100.

When the incidence rate increases in any particular region, a control chart should signal that increase faster (on average) than it would falsely signal in the in-control case. Such cases are simulated by assigning a positive value to at least one element of $\boldsymbol{\mu}$. If two competing charts are designed with equal in-control ARLs, the chart with the smaller out-of-control ARL is usually considered to be the superior chart. This corresponds to better sensitivity for the same false alarm rate. Often there are several sets of control chart parameters which yield out-of-control ARL results which are within simulation error. Therefore, our "optimal" results are actually the best estimates available for the specified number of simulations used.

Sonesson and Bock (2003) pointed out that most researchers have made this comparison by starting the chart at $t = 0$ (this means that the control chart statistic is at its starting value) and assuming the process shifts immediately to the out-of-control case. That is, the researchers are performing an initial-state analysis. However, as explained in Chapter 2, a steady-state analysis provides a more realistic comparison, because an out-of-control condition may not occur when the chart is at its initial state.

In this chapter, a steady-state simulation is achieved by assuming that disease rates are at their expected levels for some specified amount of time before increasing those rates. In these simulations, the process is kept in-control for fifty time periods. The out-of-control ARL is still defined as the expected number of samples required to detect the increased rate, beginning the count at the time that the rates increase. If a simulated chart signals that the rates have increased before they actually have (before the fifty periods have passed), that particular run was discarded and a new run replaced it.

3.4.2 A change in incidence rate affects the variance

It is assumed that the number of incidences in each region follows the Poisson distribution. The Poisson distribution has a mean and variance equivalent to the size of the population multiplied by the true incidence rate for each person (if the rate is constant for each person). The Poisson parameter for the ith region is denoted by μ_i. Note that there are methods to determine the rate when individuals do not have the same risk or incidence rate. However, in this chapter it is assumed that a single regional incidence rate can be obtained. If that rate, μ_i, has increased (and assuming the population size is constant), then both the mean and the variance of the count being monitored are increased. In the multivariate framework, the covariance between region i and each of the other regions is also affected. The simulation technique used here will account for the change in variance and covariance. This change is reflected in $\mathbf{\Sigma}$ by multiplying the i, i element by k and all other elements in the ith row and ith column by $\sqrt{k}$.

3.4.3 Simulation environment

Rogerson and Yamada (2004) tested the performance of several different MC1 charts. For purposes of comparing the performance of different control charts, this chapter considers the system of ten regions which they used. They arranged the ten regions in a three-by-three grid, with the tenth region placed in the first column, below the third row. Suppose these regions are numbered with regions 1 through 3 in the first row, 4 through 6 in the second row, 7 through 9 in the third row, and region 10 in the fourth row. It is assumed that these regions have a correlation structure which follows the rooks' case adjacency described in Section 3.2. Therefore, regions 1 and 5 would have correlation equal to ρ^2 because two boundary lines need to be crossed (either between regions 1 and 2 and between regions 2 and 5, or between regions 1 and 4 and between regions 4 and 5).

In the system just described, the regions share boundaries. However, the control chart methods discussed in this chapter will work with any spatial configuration as long as the

correlation structure between regions (or sites) is known. For example, the data may come from situations in which the region boundaries may be irregularly shaped. In some cases, the regions may not share common boundaries (for example, data may only be collected from regions 1, 3, 9, and 10 of the grid configuration). It is also possible that several sources may be used, such as data from hospitals, clinics, and laboratories, which may not correspond to "regions" at all.

Rogerson and Yamada's (2004) choice of 0, 0.2, 0.5, and 0.7 for correlations (values of ρ) will be continued here. For the purpose of comparison, the case in which expected regional counts increase by one standard deviation is considered first. This is, for example, equivalent to a 10% increase when $\mu = 100$.

Note that when $\rho > 0$, an incidence of disease in one location suggests that people in or near that location are more likely to also be diagnosed with the disease. For many types of diseases (particularly those that are not airborne or spread by contact) there should be little or no spatial correlation. When no spatial correlation is assumed, fewer assumptions about the regional structure are necessary.

This chapter makes the simplifying assumption that all regions have the same baseline value of μ. In order to use the normal approximation to the Poisson distribution, it is recommended that μ be at least 10. This study will consider in-control values of μ equal to 10, 50, and 100. If, for example, each region has a population of 10000, then these values of μ could correspond to incidence rates of 0.001, 0.005, and 0.01, respectively, for each individual.

3.4.4 Impact of shifted variance and steady-state analysis

Table 3.1 presents the difference between the traditional initial-state, constant variance analysis and the more relevant steady-state, increasing variance analysis. Section a) of the table makes this comparison when the rate of one region (region 1 is used) increases. The comparison when the rate increases in three adjacent regions (i.e., regions 1, 2, and 4) is made

in section b), while section c) gives the comparison when the rate increases in three disjoint regions (i.e., regions 1, 6, and 10).

Table 3.1 uses the values of h and r considered by Rogerson and Yamada (2004). Except for the choice of $r = 4.82$, the values of r appear to have been arbitrarily chosen and the values of h were determined by use of an approximation cited by Rogerson and Yamada (2004). They determined $r = 4.82$ by simulation when h was set to zero. None of these sets of parameters appear to produce optimal charts. Further efforts to optimize chart performance are reported in Section 3.4.6.

In the columns of Table 3.1 labeled "RY", the out-of-control ARLs are given, following the methods used to generate Figure 1 of Rogerson and Yamada (2004). Therefore, the fact that the variance changes when the rate increases is ignored. Also, these results are evaluating the relative performance under the initial-state case, in which the rates have are increased when the monitoring begins. The standard errors of each ARL estimate given in this column are no more than 0.45% of the associated ARL.

In the columns of Table 3.1 labeled "SSCOV", the out-of-control steady-state ARLs (or SSARLs) are given. Here the covariance matrix is changed with the shift in the mean and steady-state theory is used. The covariance matrix changes depend on the in-control mean vector $\boldsymbol{\mu}$. Here it is assumed that all 10 regions each have an in-control average count of 100. The standard errors of each ARL estimate given in this column are no more than 0.45% of the associated ARL.

A comparison of the two sets of columns in Table 3.1 shows that, except for the Shewhart case where $h = 0$, the MC1 chart has poorer statistical performance when one accounts for the both the shift in variance resulting from the assumption that we have a Poisson process and the use of a steady-state analysis. However, these adjustments provide a more realistic representation of the performance of the chart.

Table 3.1: Estimated out-of-control ARLs (no covariance matrix shift) and SSARLs (with covariance matrix shift) for the MC1 chart. 50000 simulations are used, yielding $se(\widehat{\text{ARL}}) < 0.0045\widehat{\text{ARL}}$. In-control ARL = 100.

a) region 1 shifts by 1 standard deviation.

		$\rho = 0.2$		$\rho = 0.5$		$\rho = 0.7$	
h	r	RY	SSCOV	RY	SSCOV	RY	SSCOV
9.8	0.40	10.56	15.61	8.16	11.92	5.61	8.07
8.5	0.50	10.10	14.79	7.57	11.04	5.09	7.32
7.4	0.60	9.82	14.21	7.09	10.43	4.64	6.74
0.0	4.82	48.02	43.51	32.73	29.55	13.78	12.51

b) three adjacent regions (1, 2, and 4) each shift by 1 standard deviation.

		$\rho = 0.2$		$\rho = 0.5$		$\rho = 0.7$	
h	r	RY	SSCOV	RY	SSCOV	RY	SSCOV
9.8	0.40	6.83	9.84	5.81	8.35	3.94	5.50
8.5	0.50	6.24	9.04	5.28	7.60	3.54	4.95
7.4	0.60	5.74	8.40	4.83	6.96	3.20	4.51
0.0	4.82	22.69	18.97	15.37	13.02	4.36	4.06

c) three disjoint regions (1, 6, and 10) each shift by 1 standard deviation.

		$\rho = 0.2$		$\rho = 0.5$		$\rho = 0.7$	
h	r	RY	SSCOV	RY	SSCOV	RY	SSCOV
9.8	0.40	6.83	8.72	5.81	6.83	3.94	4.72
8.5	0.50	6.24	7.93	5.28	6.15	3.54	4.24
7.4	0.60	5.74	7.36	4.83	5.64	3.20	3.84
0.0	4.82	22.69	14.58	15.37	7.75	4.36	2.55

RY: Initial-state, no covariance shift. See Rogerson and Yamada (2004).
SSCOV: Steady-state, with covariance shift.

3.4.5 Size of shifts

The Poisson assumption means that the variance of an observed count is the same value as the mean. It may be difficult to interpret the size of the shift when an out-of-control condition corresponds to an increase of some number of standard deviations. If, for example, 10 cases of a disease are expected to occur in a population of 10000, then a one standard deviation increase corresponds to an increased rate of about 3.16 cases, or a 31.6% increase. However, if the disease is expected to occur, say, 50 times on average in a population of 10000, then a one standard deviation increase corresponds to an increased rate of about 7.07 cases, or a 14.1% increase.

Because the "size" of a standard deviation is dependent on μ, it is more practical to be concerned with increases in terms of percentage change as opposed to increases in units of the standard deviation. This type of increase has been considered in the quality literature for Poisson variates (e.g., Lucas (1985)).

3.4.6 Comparison between the proposed and existing charts

There are many ways to compare the one-sided MEWMA chart with competing charts such as the MC1 method or the system of multiple one-sided univariate CUSUM charts. The comparisons are established by choosing the optimal parameters for each chart such that the out-of-control SSARL is minimized for a specific out-of-control condition, when the in-control ARL is 100. Simulation is used to approximate the optimal chart parameters (λ, reference value, and alarm limits) for each of these three charts. Three main cases are examined: detecting a 20% shift with $\mu = 10$, a 20% shift with $\mu = 50$, and a 10% shift with $\mu = 100$. This last case represents a one standard deviation increase, which makes it similar to the case simulated by Rogerson and Yamada (2004). The other two cases deal with the possibility of lower regional counts, but assume that the rate increase will be somewhat larger, since small shifts are generally difficult to detect.

Each of these three cases is considered under three different types of regional shifts. These shifts involve a shift in the rate of a single region (region 1), a shift in the rates of three adjacent regions (1, 2, and 4), and a shift in the rates of three non-adjacent regions (1, 6, and 10). Table 3.2 presents the estimated out-of-control SSARLs of the optimized charts in these cases, assuming $\rho = 0.5$.

Note that, in the cases in Table 3.2, detection of the specified increase by a MC1 chart requires at least 14% more time periods on average than the corresponding one-sided MEWMA chart. The one-sided MEWMA chart also detects increases in these cases faster than the system of multiple one-sided univariate CUSUM charts. However, it is important to note that the system of multiple one-sided univariate CUSUM charts is sometimes superior to the MC1 method. This happens when either a single region experiences an increased rate or all regions experience increased rates. In these cases, the one-sided nature of the univariate CUSUM system is more valuable than the spatial information. In the cases where the rates of three regions are shifted, however, the MC1 chart makes better use of the spatial information, and this leads to greater sensitivity to the increases than the one-sided univariate CUSUM charts, which ignore the spatial information.

When $\rho = 0$, the regions are spatially independent of each other. Table 3.3 provides the estimated ARL performance of the three control chart methods in this independent case. Note that the same combinations of μ and percent shift are presented for the out-of-control conditions in which the mean 1) of any one region increases, 2) increases simultaneously in any three regions, and 3) increases simultaneously in all of the regions. Since the regions are spatially independent of each other, there is no need to indicate which of the regions are increasing in the three region case.

When a rate increase occurs in just one region, the results in Table 3.3 indicate that the system of univariate CUSUM charts detects the increase faster (on average) than the competing charts. However, if the increase occurs simultaneously in three regions, the one-sided MEWMA control chart tends to detect the increase faster. In the case of a one-region increase, each univariate control chart is considering only the information from the associated

Table 3.2: Comparison of estimated optimal out-of-control SSARLs ($\rho = 0.5$) using 100000 simulations. In-control ARL = 100.

Regions	%		One-sided MEWMA				MC1				Univariate CUSUM			
shifted	shift	μ	limit	λ	$se(\widehat{\text{ARL}})$	$\widehat{\text{ARL}}$	h	r	$se(\widehat{\text{ARL}})$	$\widehat{\text{ARL}}$	h	r	$se(\widehat{\text{ARL}})$	ARL
1	20	10	12.325	0.05	0.031	15.01	7.875	0.55	0.040	18.65	25.35	0.75	0.033	16.18
1	20	50	15.960	0.19	0.009	5.10	4.270	1.20	0.011	5.95	29.35	4.25	0.009	5.38
1	10	100	14.695	0.11	0.015	8.34	5.680	0.85	0.019	9.91	57.10	4.00	0.016	8.91
1, 2, 4	20	10	14.695	0.11	0.019	9.09	5.960	0.80	0.024	11.38	17.20	1.35	0.022	10.58
1, 2, 4	20	50	16.970	0.37	0.005	3.00	3.310	1.60	0.006	3.50	19.90	6.50	0.006	3.61
1, 2, 4	10	100	16.255	0.22	0.009	4.97	4.430	1.15	0.011	5.90	44.70	5.50	0.010	5.97
1, 6, 10	20	10	14.430	0.10	0.013	7.08	5.430	0.90	0.016	8.46	18.70	1.20	0.017	9.37
1, 6, 10	20	50	17.120	0.44	0.004	2.28	2.635	2.05	0.004	2.60	19.90	6.50	0.005	3.21
1, 6, 10	10	100	16.510	0.26	0.007	3.80	3.730	1.40	0.008	4.37	44.70	5.50	0.008	5.38
All	10	100	16.890	0.34	0.006	3.15	4.270	1.20	0.012	6.35	30.40	8.50	0.006	3.64

Table 3.3: Comparison of estimated optimal out-of-control SSARLs (spatially independent regions) using 100000 simulations. In-control ARL = 100.

Number of	%		One-sided MEWMA				MC1				Univariate CUSUM			
regions	shift	μ	limit	λ	$se(\widehat{\text{ARL}})$	$\widehat{\text{ARL}}$	h	r	$se(\widehat{\text{ARL}})$	$\widehat{\text{ARL}}$	h	r	$se(\widehat{\text{ARL}})$	ARL
1	20	10	15.250	0.04	0.036	17.02	9.06	0.45	0.062	26.74	23.25	0.95	0.035	16.49
1	20	50	18.210	0.14	0.011	6.30	5.68	0.85	0.017	9.10	26.53	4.95	0.010	5.46
1	10	100	17.480	0.09	0.019	9.99	6.97	0.65	0.031	14.89	50.15	5.00	0.017	9.06
3	20	10	18.130	0.13	0.015	7.54	6.59	0.70	0.026	12.75	17.46	1.40	0.017	9.25
3	20	50	18.480	0.44	0.004	2.58	3.61	1.45	0.007	3.97	19.24	6.90	0.005	3.16
3	10	100	18.695	0.25	0.007	4.20	4.78	1.05	0.012	6.69	40.08	6.50	0.009	5.26
10	10	100	18.020	0.60	0.002	1.46	2.76	1.95	0.005	2.68	25.25	10.50	0.005	2.97

data stream (and not looking for changes in covariance), so it makes sense that the univariate method would have the fastest detection. The one-sided MEWMA method is best for an increase in three regions in part because the one-sided MEWMA chart statistic is affected by changes in any of the streams (and not just a single data stream). The MC1 method is inferior in most of these cases. This is likely due to the fact that the MC1 method can signal for decreases in rates, and therefore the control limits need to be increased somewhat to account for the number of signals associated with decreases.

Other combinations of shift size, μ, and ρ were considered, although no attempt was made to optimize the control chart performance for these combinations. However, in that limited exploration, the one-sided MEWMA chart has been consistently superior to the MC1 chart and to the system of one-sided univariate CUSUM charts when either $\rho > 0$ or more than one region is affected.

3.4.7 Ability to detect shifts of varying sizes

The optimal charts used in 3.2 were designed to detect specific percentage increases as rapidly as possible. However, it is possible that larger or smaller shifts will occur, since in practice the size of the shift is not known. Therefore, it is beneficial to compare the statistical performance of the multiple univariate CUSUM chart system, the MC1 chart, and the one-sided MEWMA chart under a variety of shift sizes. The optimal chart parameters (alarm limits and either the reference value or smoothing parameter) that were found for the charts in Table 3.2 are retained. Each set of parameters is tested under shift sizes ranging from 5% to 30%, since this range is in the neighborhood of the intended shift size of each chart.

Comparisons for rate shifts involving region 1 only, regions 1, 2, and 4, and regions 1, 6, and 10 are presented in Figures 3.1, 3.2, and 3.3 respectively. Each of these comparisons is based on $\rho = 0.5$. Each line represents the out-of-control SSARL for a different chart. Each of these charts shows that the best choice is the one-sided MEWMA chart. This is particularly clear in the event of a shift that is smaller than expected.

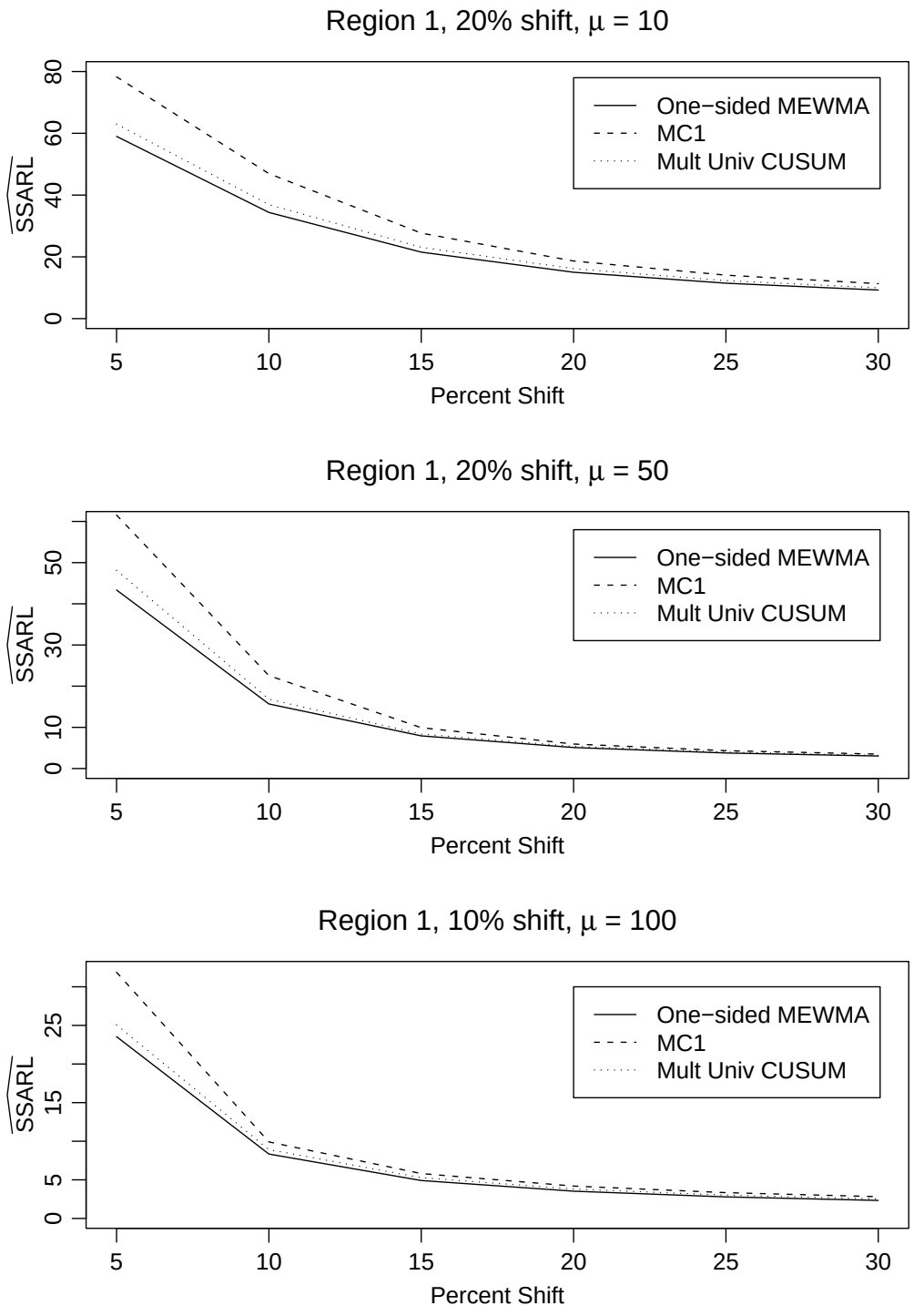


Figure 3.1: Estimated SSARL performance for shifts of different sizes in region 1. In-control ARL = 100.

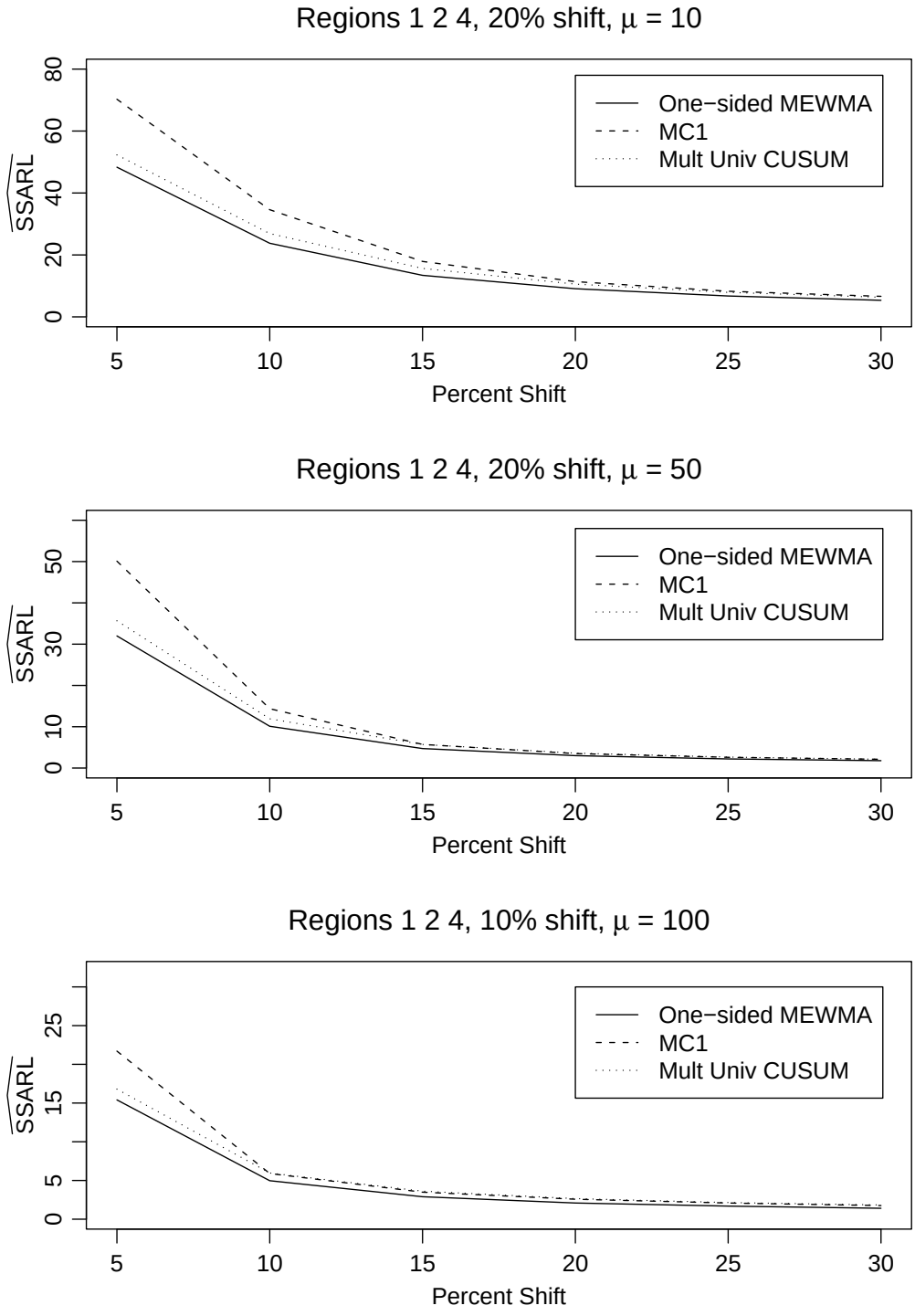


Figure 3.2: Estimated SSARL performance for shifts of different sizes in regions 1, 2, and 4. In-control ARL = 100.

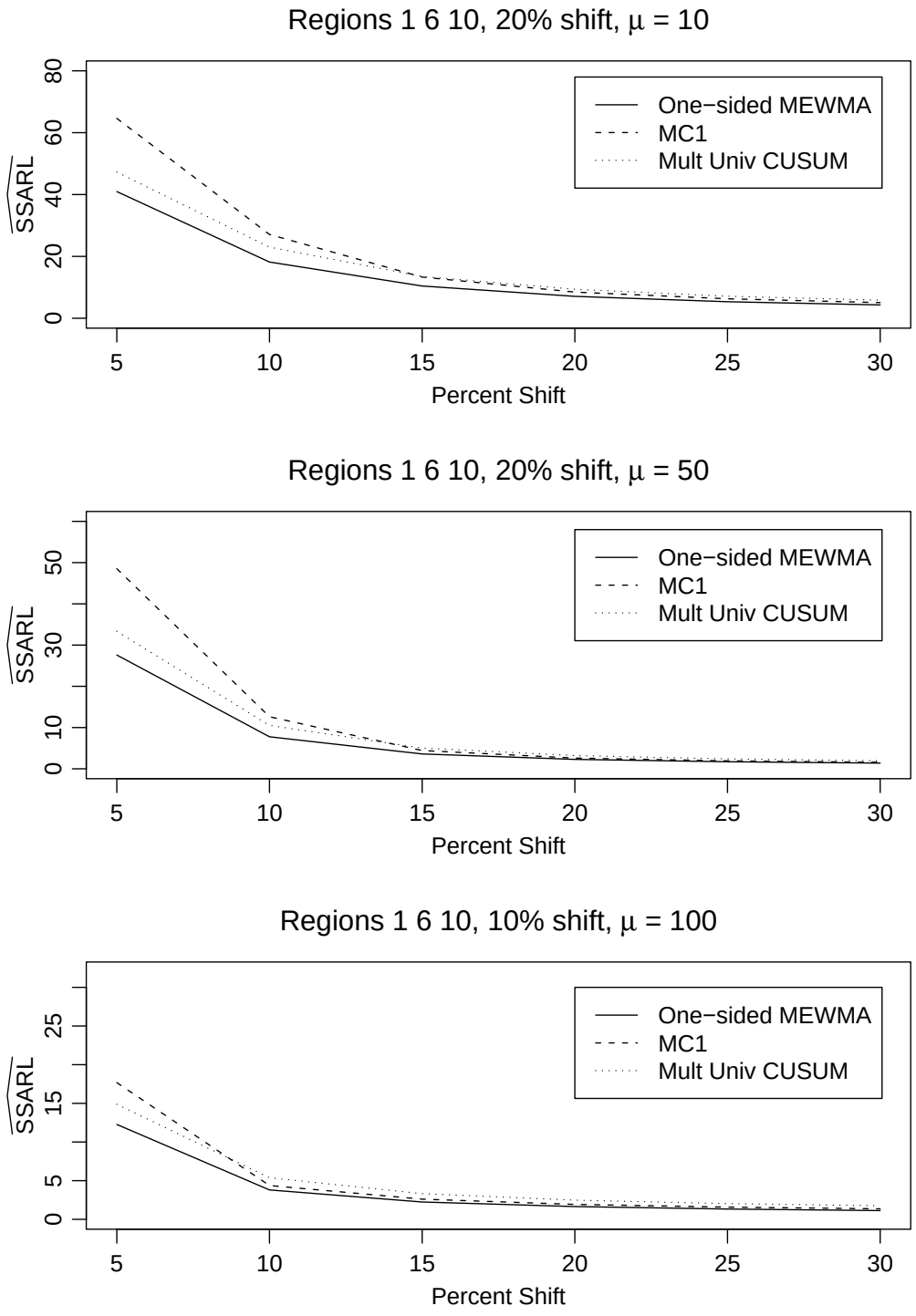


Figure 3.3: Estimated SSARL performance for shifts of different sizes in regions 1, 6, and 10. In-control ARL = 100.

3.5 Discussion

In a monitoring context, it is often essential to detect increases in rates based on regional counts of disease as soon as possible after an increase occurs in the population. Knowing about the increase sooner may allow officials to act more quickly to prevent further spread of the disease. It follows that any improvement in the ARL performance of a monitoring technique should be pursued. The techniques studied in this chapter can be used in detecting increases in rates that are monitored in multiple locations at the same points in time. Some of the issues involved include steady-state analyses, building of "credit" in multivariate control charts, and properties of the Poisson distribution. A simulation-based method for evaluating the ARL performance of a control chart has also been presented.

It has been shown in this chapter that the proposed one-sided MEWMA chart is superior to the other monitoring techniques that have been considered, namely the MC1 chart and the system of multiple univariate CUSUM charts. The one-sided MEWMA chart is, on average, the fastest to signal even when the increase is different from the increase the chart was designed to detect. The MC1 chart, on the other hand, has inferior performance in surveillance applications. Again, because of the credit that can build up when the MC1 chart is used, its use in surveillance is not recommended.

The results in this chapter indicate that the ARL performance for a shift in three regions with a high assumed correlation between them is worse than the ARL performance for a shift of equal size in three regions where the assumed correlation between them is relatively smaller, for all of the charts considered here. When the correlation is high, then a high count in one region implies that the neighboring regions are more likely to have high counts. This means that if three regions each experience rate increases, and those regions are assumed to be highly correlated with each other, then the fact that the rates of these three regions all increase is actually expected. On the other hand, if three regions each experience rate increases, but those regions are not assumed to be highly correlated with each other, then the fact that the three regions all experienced rate increases does not agree with the assumed

correlation structure. Therefore, high spatial correlation can make it harder to detect true outbreaks in some cases. This is an important realization, and somewhat disconcerting, as it implies that the ability of these control charts to detect rate increases depends on the underlying correlation matrix.

Finally, there is a need to establish some set of procedures for responding to the signals given by multivariate control charts. In particular, it is important to know which regions experienced increased rates, when the increase began, and the magnitude of the increase. A reasonable initial action would be to investigate the control charts for the individual regions.

Chapter 4

Conclusions and Future Work

4.1 Summary and Conclusions

In the current statistical research climate, medical surveillance is a topic of great interest. In particular, health specialists are searching for statistical methodology which will quickly indicate an increase in a rate of interest, whether it is monitored locally or at a higher level.

Several methods for detecting increases in incidence rates have been presented in this dissertation. Many of these methods are in use to some degree in practice. The usage of these methods and the details for efficient use of them have been presented in this dissertation, and comparisons have been made. When the data are binary and available as the observations are recorded, the Bernoulli CUSUM chart appears to be best at detecting increases in incidence rates, although the scan statistic method usually offers similar performance. When the observations are aggregated over time and a count of incidences is taken on a regular basis, a one-sided MEWMA chart appears to be best at detecting increases in multiple regions.

The ability of any chart to detect rate increases depends heavily on the probability of observing an incidence under normal conditions, the magnitude (and location) of the increase in the incidence rate which must be rapidly detected, the spatial correlation structure (if

any), and the false alarm rate. When choosing control chart parameters, researchers must decide which type or types of rate increases are most important to detect. Researchers must also choose a false alarm rate. The decisions made relative to these items must be carefully considered. This is discussed briefly in Section 4.2.

Further research in the field of medical surveillance is necessary. This research should be centered in finding or adapting methods to produce rapid detection of rate increases. The quality of health care and the ability to protect the public from disease breakouts relies on the development of efficient monitoring techniques and a proper application of them. Some directions for future research are given in Section 4.3.

4.2 Some issues in medical surveillance

Some researchers in medical surveillance appear to neglect the implications of a false alarm in an attempt to quickly detect a rate increase. For example, Ismail *et al.* (2003) stated that "in the hospital the cost of [a] false alarm is unlikely to be great." This generalization is dangerous. It implies that it is better to use a small in-control ANOS or ARL value because a rate increase would be indicated sooner. The monitoring methods would then frequently issue false alarms, and this could lead health specialists to choose to ignore the charts when they signal an alarm condition. Therefore, increasing the false alarm rate to an unacceptably high level in order to rapidly detect a rate increase is often an inappropriate solution.

Also, in certain applications, a signal from a control chart may trigger immediate precautions until the cause of the signal can be determined. These may include giving patients certain medications or treatments which, under a false alarm condition, are unnecessary and thus result in unnecessary expense. In certain cases, such a response may expose patients to other risks, such as undesirable side effects from these unnecessary treatments. Obviously, there are also arguments for using a relatively small in-control ANOS or ARL. If the ANOS or ARL are too large, it will take too long to detect true rate increases. Health professionals must consider the tradeoffs when selecting these values.

However, another view on the issue of controlling the false alarm rate is given by Rolka (2006). He mentions that at least some of this debate is caused by the current surveillance hypotheses, which equate a false indication of an increased rate to a Type I error, and a failure to respond to a real increase in the rate to a Type II error. He suggests that failing to respond to an actual rate increase is the more severe error. Therefore, it would seem that this error rate should be controlled. However, he acknowledges that this would allow more false signals, which leads a "fatigue" when a response is really needed.

4.3 Additional topics for future work

The following are several additional topics which could be considered as part of future research efforts.

First, the methods discussed in this dissertation have focused on the problem of detecting rate increases. In some cases, it may be important to look for decreases in the frequency of medical events. For example, a decrease in the incidence of a given disease may signal a pharmaceutical company to decrease production of a given drug. The methods studied in this dissertation could be adapted to detect decreases. However, when an event is rare, it may prove difficult to detect a decrease in the rate.

The scan statistic and CUSUM methods of Chapter 2 can be adapted for Poisson counts. It would be interesting to compare their performance, and also compare the performance relative to the performance of the Bernoulli charts.

In the epidemiological literature, significant attention is being given to the methods which can include spatial information. The discussion in this area has focused on the spatio-temporal version of the scan statistic. In particular, Kulldorff (2001) attempts to simultaneously scan over spatial areas of varying sizes and temporal periods of varying lengths to search for clusters of increased rates. This approach is appealing because it is expected to provide a specific location in both space and time for transient rate increases, but also indi-

cate a sustained increase and/or spread of a rate increase. However, at present this method is at best a diagnostic tool. Its statistical properties under in-control and out-of-control conditions are not known. The results of research in this area could be very revealing on a method that appears to be in widespread use. Some of this research is in progress and is expected to become part of Shannon Fraker's dissertation, under the direction of Dr. William H. Woodall and with the assistance of Dr. Howard S. Burkom of The John Hopkins University Applied Physics Laboratory.

As mentioned in Chapter 2, additional research could focus on the different chart parameters that would likely result if the parameters were selected according to other criterion. One such criteron would be to maximize the probability of detecting a specific increased rate within a specified number of observations subject to a relatively low signal probability if the rate does not increase. Another would be to minimize the steady-state out-of-control median number of observations to signal, subject to a specific in-control median number of observations to signal.

When data are collected in several regions and aggregated over time, as assumed in Chapter 3, most, if not all, methods require that the count data must be available from all regions before the control chart statistic can be recalculated. However, particularly in complicated data collection schemes, it will be difficult to have all of the data available before some appointed deadline. Although Strat (2005) briefly discussed this issue, there have not been any studies on the performance of various surveillance methods when the data are incomplete. It is possible that many control-chart based methods will not be usable, and if any are, no one procedure is likely to be optimal. On the other hand, data must also be obtained with minimal reporting delay, as such delays must be added to the average run length when determining the time required to discover an increase in the disease rates. Therefore, methods need to be developed which can handle the missing data, so as to reduce reporting-related delays.

Yet another issue from Chapter 3 concerns the use of the multivariate normal approximation with Poisson data. This approximation is known to be adequate in the univariate

case when the Poisson mean is sufficiently large. One way to deal with this problem would be to develop a multivariate monitoring methods that uses the Poisson framework rather than relying on the normal approximation. However, to evaluate properties of this method in the multivariate case, an algorithm is needed which will efficiently generate data from a multivariate Poisson distribution. These algorithms exist only in a limited form. For example, Johnson *et al.* (1997) explain how to generate multivariate Poisson data if $\text{Cov}(X_i, X_j)$ are equal to each other for all $i \neq j$, where X_i and X_j are Poisson counts.

Finally, several researchers have recently indicated that surveillance data can be autocorrelated (see, for example, Shmueli, 2007). Indeed, the approach taken in Chapter 3 is based on the assumption that all of the Poisson counts are time-independent. When these counts are correlated over time, the variability estimates are affected. This, in turn, influences the chart design needed to give a specified false alarm rate. Even with a modified chart design, autocorrelation affects the ability to detect rate increases. Methods need to be developed that can handle multivariate autocorrelated count data as effectively as possible.

Bibliography

Balakrishnan, N. and Koutras, M. V. (2002). *Runs and Scans with Applications*. New York: Wiley.

Brookmeyer, R. and Stroup, D. F. (2004). *Monitoring the Health of Populations*. New York: Oxford University Press.

Casella, G. and Berger, R. L. (2002). *Statistical Inference*, 2nd edition. Pacific Grove, CA: Duxbury.

Champ, C. W. and Woodall, W. H. (1987). Exact results for Shewhart control charts with supplementary runs rules. *Technometrics* **29**(4), 393–399.

Chen, R. (1978). A surveillance system for congenital malformations. *Journal of the American Statistical Association* **73**, 323–327.

Crosier, R. B. (1988). Multivariate generalizations of cumulative sum quality-control schemes. *Technometrics* **30**(3), 291–303.

Elliott, P. and Wartenberg, D. (2004). Spatial epidemiology: current approaches and future challenges. *Environmental Health Perspectives* **112**(9), 998–1006. First published on 15 April 2004, doi:10.1289/ehp.6735.

Farrington, P. and Beale, A. D. (1998). The detection of outbreaks of infectious disease. In Lierl, L., Cliff, A. D., Valleron, A., Farrington, P. and Bull, M. (eds), *Geomed '97*. Stuttgart: BG Teubner.

Fassò, A. (1998). One-sided multivariate testing and environmental monitoring. *Austrian Journal of Statistics* **27**(1–2), 17–37.

Fassò, A. (1999). One-sided MEWMA control charts. *Communications in Statistics: Simulation and Computation* **28**(2), 381–401.

Glaz, J., Naus, J. and Wallenstein, S. (2001). *Scan Statistics*. New York: Springer.

Grigg, O. and Farewell, V. (2004). An overview of risk-adjusted charts. *Journal of the Royal Statistical Society, Series A* **167**(3), 523–539.

Haining, R. P. (2003). *Spatial Data Analysis: Theory and Practice*. Cambridge, United Kingdom: Cambridge University Press.

Hawkins, D. M. and Olwell, D. H. (1998). *Cumulative Sum Control Charts and Charting for Quality Improvement*. New York: Springer.

Heffernan, R., Mostashari, F., Das, D., Karpati, A., Kulldorff, M. and Weiss, D. (2004). Syndromic surveillance in public health practice, New York City. *Emerging Infectious Diseases* **10**(5), 858–864.

Ismail, N. A., Pettitt, A. N. and Webster, R. A. (2003). 'Online' monitoring and retrospective analysis of hospital outcomes based on a scan statistic. *Statistics in Medicine* **22**, 2861–2876. First published on 27 August 2003, doi:10.1002/sim.1532.

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1997). *Discrete Multivariate Distributions*. New York: Wiley, pp. 124–152.

Kulldorff, M. (2001). Prospective time periodic geographical disese surveillance using a scan statistic. *Journal of the Royal Statistical Society, Series A*, **164**(1), 61–72.

Kulldorff, M., Tango, T. and Park, P. J. (2003). Power comparisons for disease clustering tests. *Computational Statistics and Data Analysis*, **42**(4), 665–684.

Lawson, A. (2001). *Statistical Methods in Spatial Epidemiology*. West Sussex: Wiley.

Lawson, A. B. and Kleinman, K. (2005). *Spatial and Syndromic Surveillance*. West Sussex: Wiley.

Lowry, C. A., Woodall, W. H., Champ, C. W. and Rigdon, S. E. (1992). A multivariate exponentially weighted moving average control chart. *Technometrics* **34**(1), 46–53.

Lucas, J. M. (1985). Counted data CUSUM's. *Technometrics* **27**(2), 129–144.

Lucas, J. M. (1989). Control schemes for low count levels. *Journal of Quality Technology* **21**(3), 199–201.

Lucas, J. M. and Saccucci, M. S. (1990). Exponentially weighted moving average control schemes: properties and enhancements. *Technometrics* **32**(1), 1–12.

Moustakides, G. V. (1986). Optimal stopping times for detecting changes in distributions. *The Annals of Statistics* **14**, 1379–1387.

Naus, J. and Wallenstein, S. (2006). Temporal surveillance using scan statistics. *Statistics in Medicine* **25**, 311–324. First published on 12 December 2005, doi:10.1002/sim.2209.

Pignatiello, J. J., Jr. and Runger, G. C. (1990). Comparisons of multivariate CUSUM charts. *Journal of Quality Technology* **22**(3), 173–186.

Raubertas, R. F. (1989). An analysis of disease surveillance data that uses the geographic locations of the reporting units. *Statistics in Medicine* **8**, 267–271.

Reynolds, M. R., Jr. and Stoumbos, Z. G. (1999). A CUSUM chart for monitoring a proportion when inspecting continuously. *Journal of Quality Technology* **31**(1), 87–108.

Reynolds, M. R., Jr. and Stoumbos, Z. G. (2000). A general approach to modeling CUSUM charts for a proportion. *IIE Transactions* **32**, 515–535.

Reynolds, M. R., Jr. and Stoumbos, Z. G. (2004*a*). Control charts and the efficient allocation of sampling resources. *Technometrics* **46**(2), 200–214. doi:10.1198/004017004000000257.

Reynolds, M. R., Jr. and Stoumbos, Z. G. (2004*b*). Should observations be grouped for effective process monitoring? *Journal of Quality Technology* **36**(4), 343–366.

Rogerson, P. (2001). Monitoring point patterns for the development of space-time clusters. *Journal of the Royal Statistical Society, Series A* **164**(1), 87–96.

Rogerson, P. A. (1997). Surveillance systems for monitoring the development of spatial patterns. *Statistics in Medicine* **16**, 2081–2093.

Rogerson, P. A. and Yamada, I. (2004). Monitoring change in spatial patterns of disease: comparing univariate and multivariate cumulative sum approaches. *Statistics in Medicine* **23**, 2195–2214. First published on 29 June 2004, doi:10.1002/sim.1806.

Rolka, H. (2006). Data analysis research issues and emerging public health biosurveillance directions. In Wilson, A. G., Wilson, G. D. and Olwell, D. H. (eds), *Statistical Methods in Counterterrorism*. New York: Springer, pp. 101–107.

Runger, G. C. and Testik, M. C. (2004). Multivariate extensions to cumulative sum control charts. *Quality and Reliability Engineering International* **20**, 587–606. First published on 10 June 2004, doi:10.1002/qre.571.

Sego, L. H., Reynolds, M. R., Jr. and Woodall, W. H. (2007*a*). Risk-adjusted monitoring of survival times. *Statistics in Medicine*. Under revision for resubmission.

Sego, L. H., Woodall, W. H. and Reynolds, M. R., Jr. (2007*b*). A comparison of surveillance methods for small incidence rates. *Statistics in Medicine*. Conditionally accepted.

Shmueli, G. (2007). Statistical challenges in modern biosurveillance. *Technometrics*. Invited paper, submitted.

Shmueli, G. and Fienberg, S. E. (2006). Current and potential statistical methods for monitoring multiple data streams for biosurveillance. In Wilson, A. G., Wilson, G. D.

and Olwell, D. H. (eds), *Statistical Methods in Counterterrorism*. New York: Springer, pp. 109–140.

Sitter, R. R., Hanrahan, L. P., DeMets, D. and Anderson, H. A. (1990). A monitoring system to detect increased rates of cancer incidence. *American Journal of Epidemiology* **132**, 123–130.

Sonesson, C. and Bock, D. (2003). A review and discussion of prospective statistical surveillance in public health. *Journal of the Royal Statistical Society, Series A* **166**(1), 5–21.

Sonesson, C. and Frisén, M. (2005). Multivariate surveillance. In Lawson, A. B. and Kleinman, K. (eds), *Spatial and Syndromic Surveillance*. West Sussex: Wiley, pp. 153–166.

Steiner, S. H., Cook, R. J., Farewell, V. T. and Treasure, T. (2000). Monitoring surgical performance using risk-adjusted cumulative sum charts. *Biostatistics* **1**(4), 441–452.

Stoto, M. A., Fricker, R. D., Jr., Jain, A., Diamond, A., Davies-Cole, J. O., Glymph, C., Kidane, G., Lum, G., Jones, L., Dehan, K. and Yuan, C. (2006). Evaluating statistical methods for syndromic surveillance. In Wilson, A. G., Wilson, G. D. and Olwell, D. H. (eds), *Statistical Methods in Counterterrorism*. New York: Springer, pp. 141–172.

Strat, Y. L. (2005). Overview of temporal surveillance. In Lawson, A. B. and Kleinman, K. (eds), *Spatial and Syndromic Surveillance*. West Sussex: Wiley, pp. 13–30.

Testik, M. C. and Runger, G. C. (2006). Multivariate one-sided control charts. *IIE Transactions* **38**, 635–645. doi:10.1080/07408170600692176.

Thacker, S. B., Stroup, D. F., Parrish, R. G. and Anderson, H. A. (1996). Surveillance in environmental public health: issues, systems, and sources. *American Journal of Public Health* **86**(5), 633–638.

Waller, L. A. and Gotway, C. A. (2004). *Applied Spatial Statistics for Public Health Data*. Hoboken, New Jersey: Wiley.

Wilson, A. G., Wilson, G. D. and Olwell, D. H. (2006). *Statistical Methods in Counterterrorism*. New York: Springer.

Wolter, C. (1987). Monitoring intervals between rare events: a cumulative score procedure compared with Rina Chen's sets technique. *Methods of Information in Medicine* **26**, 215–219.

Woodall, W. H. (1997). Control charts based on attribute data: bibliography and review. *Journal of Quality Technology* **29**(2), 172–183.

Woodall, W. H. (2006). The use of control charts in health care monitoring and public health surveillance (with discussion). *Journal of Quality Technology* **38**(2), 88–104.

Woodall, W. H. and Mahmoud, M. A. (2005). The inertial properties of quality control charts. *Technometrics* **47**(4), 425–436. doi:10.1198/004017005000000256.

Woodall, W. H., Marshall, J. B., Joner, M. D., Jr., Fraker, S. E. and Abdel-Salam, A.-S. G. (2007). On the use of scan methods in prospective health surveillance. *Journal of the Royal Statistical Society, Series A*. Accepted for publication.

Woodall, W. H. and Ncube, M. M. (1985). Multivariate cusum quality-control procedures. *Technometrics* **27**(3), 285–292.