A CUSUM Chart for Monitoring a Proportion When Inspecting Continuously

MARION R. REYNOLDS, JR.

Virginia Polytechnic Institute and State University, Blacksburg, VA 24061

ZACHARY G. STOUMBOS

Rutgers, The State University of New Jersey, Newark, NJ 07102-1895

A control chart is considered for the problem of monitoring a process when all items from the process are inspected and classified into one of two categories. The objective is to detect changes in the proportion, p, of items in the first category. The control chart being considered is a cumulative sum (CUSUM) chart based on the Bernoulli observations corresponding to the inspection of the individual items. Bernoulli CUSUM charts can be constructed to detect increases in p, decreases in p, or both increases and decreases in p. The properties of the Bernoulli CUSUM chart are evaluated using exact Markov chain methods and by using a corrected diffusion theory approximation. The corrected diffusion theory approximation provides a relatively simple method of designing the chart for practical applications. It is shown that the Bernoulli CUSUM chart will detect changes in p substantially faster than the traditional approach of grouping items into samples and applying a Shewhart p-chart. The Bernoulli CUSUM chart is also better than grouping items into samples and applying a CUSUM chart to the sample statistics. The Bernoulli CUSUM chart is equivalent to a geometric CUSUM chart which is based on counting the number of items in the second category that occur between items in the first category.

Introduction

Many process monitoring applications concern the proportion of items from a process that are classified into one of two categories, such as defective and nondefective or conforming and nonconforming. It is usually assumed that items from the process are independent, and p is usually used to represent the probability that an item falls into the category of interest. For purposes of exposition in this paper, we will use defective to represent the category of interest, and thus, p will be the probability that an item is defective. If p is the probability of a defective, then it follows that the most desirable value for p is 0, but of course this value may not be attainable in practice. In most cases, there will be an in-control value of p, say, p_0, that is currently attainable by the process in the absence of any special or assignable cause of process variation. Then the purpose of using a control chart to monitor the process is to detect any change in p from the in-control value p_0 that is the result of a special or assignable cause. In most cases, in practice, the primary objective will be to detect any increase in p which would correspond to a deterioration in the quality of the process. In some cases, however, detecting a decrease in p may be of interest; for example, it may be important to determine whether a process improvement program is having a detectable effect. A general review of control charts that can be applied to the problem of monitoring p is given in Woodall (1997).

In many applications it will not be feasible to inspect all items produced by a process; thus, samples will be taken from the process output during specified inspection periods. Each sample taken from the process would normally correspond to one plotted

Dr. Reynolds is a Professor in the Departments of Statistics and Forestry. He is a Member of ASQ. His email address is mrr@vt.edu.

Dr. Stoumbos is an Assistant Professor in the Department of Management Science and Information Systems and a Member of the Rutgers Center for Operations Research (RUTCOR). He is a Member of ASQ.

point on a control chart. For example, samples of size $n = 100$ items might be taken at the start of every shift, and the proportion defective determined for this sample. The traditional chart to use in this application is the Shewhart p-chart, which plots the proportion defective for each sample.

In some applications, however, there may be a continuous stream of inspection data from the process. This could be the case when inspection is done automatically; thus, the inspection results are available for all items that have been produced. A continuous stream of inspection data from the process may also occur with less than 100% inspection. This could occur when the inspection process goes on continuously, but the production rate is higher than the inspection rate. For example, if items are produced at the rate of 10 per minute and the inspection process can inspect only one item per minute, then only 10% of the process output will be inspected. The inspection process running continuously will then generate a stream of one inspected item per minute. In cases in which a continuous stream of inspected items is generated, there may be no natural divisions of the stream into groups of items that would form the samples for the control chart. In these cases, the samples could be formed by artificially designating n consecutive items as a sample. A standard control chart, such as the Shewhart p-chart, could then be applied to these samples of n items.

Although the Shewhart p-chart has the advantage of simplicity, there are several important disadvantages to using this chart. The p-chart is usually set up with 3σ control limits, and this can result in an undesirable false alarm rate because the distribution of the number of defectives in n items is the discrete binomial distribution. In particular, if p_0 is close to zero and n is not very large, then the binomial distribution is not well approximated by a normal distribution. In this case, the false alarm rate can be quite different from what would be expected from using 3σ limits with a normal distribution. Instead of using 3σ limits derived for the normal distribution, it would be better to use the exact binomial distribution to determine the control limits of the p-chart. Using the binomial distribution allows the false alarm rate to be calculated for any choice of the control limits, but there is still the discreteness problem which limits the possible choices for the false alarm rate. Ryan and Schwertman (1997) provide a detailed discussion of the problems associated with determining the control limits for the p-chart.

Another disadvantage of the p-chart is that it is not effective for detecting small changes in p unless n is very large. Runs rules are sometimes used to improve the effectiveness of Shewhart charts, but in the case of the p-chart, the discreteness of the binomial distribution again causes difficulties if standard runs rules based on the normal distribution are used. A better alternative for overcoming the insensitivity of the p-chart to small shifts is to use a cumulative sum (CUSUM) chart. A CUSUM chart can be based on the binomial distribution when samples of n items are available (see, e.g., Gan (1993)).

When there is a continuous stream of inspected items from the process there are disadvantages to any control chart which has to group the items into samples. When items are grouped into samples it would usually be desirable to do the grouping based on administrative convenience and the need to have items in the sample produced under relatively homogeneous conditions. For example, all items produced during one eight-hour shift might be designated as a sample. The disadvantage to this is that a decision about the process can be made only at the end of a shift when a point is plotted on the control chart. If a large process change occurs early in a shift, then this process change would not be detected until the end of the shift. In some cases, it may be possible to partially overcome this problem by designating the samples as the inspected items from a shorter time period, such as an hour. However, if the rate of inspection is not high, the resulting value of n may be too small for the p-chart to effectively detect changes in p.

Another disadvantage to designating the items from a fixed time period as a sample is that there may be variations in the rate of production or the rate of inspection which would result in variation in the sample size from sample to sample. Some methods have been proposed for handling the problem of varying sample size, but it is difficult to determine the properties of the resulting control charts.

If the results of the inspection of an individual item are available immediately after the inspection of this item, then an alternative to using samples of n items is to record the number of items between defective items and use this statistic in a control chart. A point would be plotted on the chart every time a defective occurs. The number of items between defectives will have a geometric distribution; thus, a Shewhart chart can be constructed based on the geometric distribution or, alternately, a CUSUM chart for the geometric distribution can be used (see Bourke (1991)).

The use by practitioners of CUSUM charts for monitoring p has been limited by the difficulty in designing these charts to achieve specified properties. For example, the control limit required for the binomial CUSUM chart to achieve a specified false alarm rate depends on p_0, n, and the specified false alarm rate. The expected time required for the binomial CUSUM chart to detect a shift in p to some $p_1 \neq p_0$ depends on p_0, p_1, n, and the control limit. Although figures or tables have been published for designing CUSUM charts for monitoring p (see, e.g., Gan (1993)), these tables do not cover all of the combinations of p_0, p_1, n, and the control limit that might be needed in applications.

The objective of this paper is to consider a CUSUM control chart which can be applied when the results of the inspection of an individual item are available immediately after the inspection of this item. The individual item inspections can be represented as Bernoulli observations or trials, so this CUSUM chart will be called the *Bernoulli CUSUM chart*. The Bernoulli CUSUM chart can naturally be applied when a continuous stream of inspected items is available. It can also be applied when samples from the process output are inspected during certain inspection periods, but the investigation of this case will be considered in a separate paper. The Bernoulli CUSUM chart plots a point after each inspection, thereby plotting the original data as it is obtained; thus, there is no need to wait until a complete sample has been obtained before a summary statistic can be plotted and a decision made about the state of the process. Some simple approximations to the statistical properties of the Bernoulli CUSUM chart are given here to provide a method for designing this chart to achieve specified properties. The Bernoulli CUSUM chart will be particularly useful in situations in which the in-control value p_0 is low and large samples would be required to effectively use a standard p-chart.

The problem of monitoring p when p_0 is low is similar to the problem of monitoring the defect rate in a sequence of items when the defect rate is low (see, e.g., Lucas (1989)). However, the distribution of defects is usually assumed to be Poisson so that an item can have zero, one, or more than one defect. In the current paper, it is assumed that there are only two categories for an item; therefore the Bernoulli distribution is used.

The Bernoulli CUSUM chart for detecting an increase in p is defined explicitly in the next section, and then exact and approximate methods for evaluating its properties are discussed. A relatively simple method for designing the chart for practical applications is given and illustrated with an example. Comparisons with several competing charts are given next. It is shown that the Bernoulli CUSUM chart will offer faster detection of changes in p than either the Shewhart p-chart or the binomial CUSUM chart. It is also shown that the Bernoulli CUSUM chart is closely related to the geometric CUSUM chart. The problem of detecting a decrease in p is considered later in the paper. Technical details about the methods used in this paper to evaluate the properties of the Bernoulli CUSUM chart are given in the appendices.

The Bernoulli CUSUM Chart

Suppose that there is a continuous stream of items from the process that have been inspected and classified as defective or nondefective. This stream of items can be all items produced by the process (100% inspection) or the items selected for inspection when the production rate is higher than the inspection rate. The results of the inspection of the i^{th} item can be represented as a Bernoulli observation, X_i, which is 1 if the i^{th} item is defective and 0 otherwise. Then, p corresponds to $P(X_i = 1)$.

Before defining the Bernoulli CUSUM chart, it will be instructive to consider some more traditional approaches to constructing a control chart for the monitoring problem under consideration. If the items are grouped into samples of n consecutive items, then the appropriate statistics to consider would be $T_1, T_2, \ldots,$ where $T_j = \sum_{i=n(j-1)+1}^{nj} X_i$ is the total number of defectives in the j^{th} sample. The statistics $T_1, T_2, \ldots$ could be used in a Shewhart p-chart or in a binomial CUSUM chart. In the Shewhart p-chart with 3σ limits, the statistics T_1/n, $T_2/n, \ldots$ would be plotted on the chart with control limits

$$p_0 \pm 3\sqrt{\frac{p_0(1-p_0)}{n}}. \tag{1}$$

Unless n is very large, the calculated lower control limit for the p-chart will be negative for low values of p_0, and thus, there will be no lower control limit. In this case, the p-chart is a one-sided chart for detecting an increase in p. In the binomial CUSUM chart, a cumulative sum involving $T_1, T_2, \ldots$ would be plotted (see, e.g., Gan (1993)). A disadvantage of control charts based on $T_1, T_2, \ldots$ is that a decision about the state of the process would be made only at the end of a sample, when a value of T_j is obtained.

The Bernoulli CUSUM chart is based directly on the individual observations, $X_1, X_2, \ldots$, without using a summary statistic based on grouping the items into samples. The Bernoulli CUSUM chart will be defined here for the problem of detecting an increase in p. The problem of detecting a decrease in p will be considered later in this paper. For detecting an increase in p, the Bernoulli CUSUM control statistic is

$$B_k = \max(0, B_{k-1}) + (X_k - \gamma_B), \quad k = 1, 2, \ldots, \tag{2}$$

where the chart parameter $\gamma_B > 0$ is usually called the *reference value*. The starting value, B_0, for the statistic is frequently taken to be 0 but can be taken to be a positive value if a head start is desired (see Lucas and Crosier (1982)). This CUSUM chart will signal that there has been an increase in p if $B_k \geq h_B$, where h_B is the control limit. Note that after the inspection of item k, the value of B_k is obtained by adding the increment $(X_k - \gamma_B)$ to B_{k-1} if $B_{k-1} \geq 0$; but if $B_{k-1} < 0$, then there is a reset to 0 before $(X_k - \gamma_B)$ is added. Thus, this definition allows B_k to take on negative values. Although this definition of the CUSUM statistic has been used previously (see, e.g., Reynolds, Amin, and Arnold (1990) and Reynolds (1996)), it is slightly different from the traditional CUSUM statistic which immediately resets to 0 so that negative values are not possible. The reason for using our definition is that approximations to the CUSUM statistic depend on knowing the value of B_k when it is below 0.

The reference value γ_B can be chosen using the representation of a CUSUM chart as a sequence of tests where each test is a sequential probability ratio test (SPRT) (see, e.g., Page (1954), Gan (1993), Stoumbos and Reynolds (1997a), or Hawkins and Olwell (1998) for a discussion of the SPRT). To determine the value of γ_B, it is necessary to specify a value $p_1 > p_0$ which represents an out-of-control value of p which should be detected quickly. For a given in-control value p_0 and out-of-control value p_1, define the constants r_1 and r_2 by

$$r_1 = -\log\left(\frac{1-p_1}{1-p_0}\right) \text{ and } r_2 = \log\left(\frac{p_1(1-p_0)}{p_0(1-p_1)}\right). \tag{3}$$

Then, starting from the basic definition of an SPRT, it can be shown that the appropriate choice for γ_B is

$$\gamma_B = \frac{r_1}{r_2}. \tag{4}$$

It will usually be convenient if $\gamma_B = 1/m$, where m is an integer. For example, if $p_0 = 0.01$ and p_1 is chosen to be $p_1 = 0.025$, then this will give $r_1 = 0.0153$, $r_2 = 0.9316$, and $r_1/r_2 = 0.0164 = 1/61.02$. In this case, r_1/r_2 can be increased slightly up to $1/61$ by a slight increase in p_1. The value of p_1 required is approximately 0.02501. Having $r_1/r_2 = 1/61$ means that the possible values of B_k will be integer multiples of $1/61$, which will be convenient for plotting the chart. In general, if p_0 and p_1 are small, then a slight change in p_1 will be sufficient to make $\gamma_B = 1/m$, where m is an integer. In most applications, the precise specification of p_1 will not be critical, so this slight change in p_1 will be of no practical consequence.

Properties of the Bernoulli CUSUM Chart

The performance of a control chart is usually evaluated by looking at the average run length (ARL). The ARL is usually defined as the expected number of samples required for a control chart to signal (see, e.g., Gan (1993)), but it is sometimes used as the expected number of individual observations required to signal (see, e.g., Runger and Willemain (1995)).

This paper is concerned with the application of the Bernoulli CUSUM chart to a continuous stream of inspected items, but comparisons will be made to control charts, such as the p-chart, which are based on samples of n items. Thus, it is desirable to have separate measures for the number of samples and the number of observations required to signal. Define the average number of samples to signal (ANSS) as the expected number of samples (or segments) of n observations required for the chart to signal and define the average number of observations to signal (ANOS) as the expected number of individual observations required for the chart to signal. For the p-chart, the ANSS is computed as the inverse of the probability of a signal for a sample, and the ANOS is the product of n and the ANSS. Note that the Bernoulli CUSUM chart does not require that the inspected items be grouped into samples. Thus, this chart could signal at an item which does not correspond to the end of a sample for the p-chart.

When the time between the individual inspected items is constant, the ANOS can be easily converted to time units if desirable. In this paper, we will frequently refer to the ANOS as a measure of the time until the chart signals. It is desirable to have a large ANOS when $p = p_0$ so that the rate of false alarms is low. When there is a significant shift from p_0, then it is desirable to have a small ANOS at this shifted value of p. Recall that p_1 was defined in the previous section as a value of p which should be detected quickly. Thus, it is desirable to have the ANOS small at $p = p_1$. However, in practice, it would usually be desirable to consider a range of values of p around p_1 and to have a chart with good performance for all of these values of p.

There are several approaches that can be used to evaluate the statistical properties of the Bernoulli CUSUM chart. One approach is based on the standard method of formulating the CUSUM statistic as a Markov chain (see, e.g., Brook and Evans (1972)). Additional discussion of this method is given in Appendix A. This method gives the exact ANOS when r_1/r_2 is a rational number, but the disadvantage is that when p_0 is small, the size of the transition probability matrix for the Markov chain can be quite large, and the computational effort required to work with this matrix can be high.

A second approach to evaluating the ANOS uses the basic Markov chain formulation but explicitly solves the resulting linear equations involving the ANOS (see, e.g., Page (1954) for an example of solving linear equations to find the quantity of interest). Although this second approach avoids the direct manipulation of the transition probability matrix and is relatively easy to program for computer computation, the computational effort here is great enough that computation with a simple pocket calculator is not really feasible. A description of this approach is given in Appendix B.

A third approach to evaluating the ANOS is based on using approximations developed by Wald (1947) and diffusion theory corrections to these approximations that extend the work of Siegmund (1979, 1985). The approximation for the ANOS obtained using this approach will be called the *corrected diffusion* (CD) approximation. The CD approximation will form the basis of a relatively simple method which requires only a pocket calculator to design the Bernoulli CUSUM for practical applications. More details about the CD approximation are given in Appendix C.

A Method for Designing the Bernoulli CUSUM Chart

To design a Bernoulli CUSUM chart for a particular application, it will be necessary to specify p_0 (the in-control value of p) and p_1 (the value of p which the chart is designed to detect). The values of p_0 and p_1 will then determine the reference value γ_B through Equation (4). As discussed previously, when p_0 and p_1 are small, it will usually be convenient to adjust p_1 slightly so that $\gamma_B = 1/m$, where m is an integer.

In designing the chart, it is also necessary to determine the value for the control limit, h_B. The value of h_B will determine the false alarm rate and the speed with which the chart detects increases in p. A reasonable approach to determining h_B is to specify a desired value of the ANOS when $p = p_0$ and then choose h_B to achieve approximately this value of the ANOS. It will usually not be possible to achieve exactly a desired value of the ANOS because the Bernoulli distribution is discrete. However, if p_0 is small, it will usually be possible to obtain an in-control ANOS that is very close to the desired value. Once h_B is chosen to achieve approximately the specified ANOS at $p = p_0$, it will be desirable to look at the ANOS at $p = p_1$ and at other values of p to determine whether detection of shifts in p will be fast enough. In practice, it may be necessary to adjust h_B to achieve a reasonable balance between the desire to have a low false alarm rate (achieved by choosing a large h_B) and fast detection of shifts in p (achieved by choosing a small h_B).

The CD approximation to the ANOS of the Bernoulli CUSUM uses an adjusted value of h_B, which will be called h_B^*, in a relatively simple formula. This adjusted value of h_B is

$$h_B^* = h_B + \epsilon(p_0)\sqrt{p_0 q_0}\,, \tag{5}$$

where $q_0 = 1 - p_0$ and $\epsilon(p)$ is defined in Appendix C. The interpretation of $\epsilon(p_0)\sqrt{p_0 q_0}$ is as the limiting expected excess of the control statistic over h_B. For use in applications, $\epsilon(p)$ can be approximated by

$$\epsilon(p) \approx \begin{cases} 0.410 - 0.0842(\log(p)) & \\ -0.0391(\log(p))^3 & \\ -0.00376(\log(p))^4 & \\ -0.000008(\log(p))^7 & \text{if } 0.01 \le p \le 0.5 \\ & \\ \frac{1}{3}\left(\sqrt{\frac{1-p}{p}} - \sqrt{\frac{p}{1-p}}\right) & \text{if } 0 < p < 0.01\,. \end{cases} \tag{6}$$

A method for obtaining $\epsilon(p)$ when $p > 0.5$ is given in Appendix C. When $p = p_0$, the CD approximation to the ANOS is

$$\text{ANOS}(p_0) \approx \frac{\exp\{h_B^* r_2\} - h_B^* r_2 - 1}{|r_2 p_0 - r_1|}\,. \tag{7}$$

For given values of r_1 and r_2 and a desired value for the in-control ANOS, Equation (7) can be used to find the required value of h_B^*, and then Equations (5) and (6) can be used to find the required value of h_B. Finding h_B^* using Equation (7) can be accomplished by simple trial and error. Appendix D describes a Newton-Raphson procedure that can be used to find

h_B^*, if it is desirable to write a simple program to find h_B using Equation (7).

Table 1 was constructed to provide an additional design aid in determining h_B for the Bernoulli CUSUM chart. This table gives, for several values of p_0, the value of h_B that will give an in-control ANOS closest to several specified values. When p_0 is quite small, say 0.001, it will usually be desirable to choose p_1 to be a relatively large multiple of p_0, such as $3p_0$ or $4p_0$. However, for larger values of p_0, such as 0.10, it will usually be desirable to choose p_1 to be a relatively small multiple of p_0, such as $1.5p_0$ or $2p_0$. All of the values of h_B given in Table 1 correspond to an in-control ANOS that is within 20% of the specified value (in most cases the in-control ANOS is much closer than 20%). In the few cases for which it was not possible to obtain an in-control ANOS within 20%, no entry was made in the corresponding cell of Table 1.

In most applications it will be desirable to determine how fast a shift from p_0 to p_1 will be detected. The CD approximation to the ANOS when $p = p_1$ is

$$\text{ANOS}(p_1) \approx \frac{\exp\{-h_B^* r_2\} + h_B^* r_2 - 1}{|r_2 p_1 - r_1|}, \quad (8)$$

where h_B^* is given by Equation (5). Note that h_B^* uses p_0 even though the ANOS is being approximated at p_1. Approximations to the ANOS for other values of p and a discussion of the accuracy of the CD approximation to the ANOS are given in Appendix C.

An Example of Designing a Bernoulli CUSUM Chart

Suppose that a Shewhart p-chart is currently being used to monitor a process for which the in-control value of p is $p_0 = 0.01$. There is a continuous stream of inspected items which are grouped into samples of $n = 100$ items for purposes of applying this p-chart. This problem could arise from several different inspection situations. One situation is the straightforward case in which 100% of the process output is inspected. As an example, suppose that the production of an item takes one minute and the inspection of an item also takes one minute. The p-chart based on using samples of $n = 100$ would then require 100 minutes to obtain the information for one plotted point.

A second situation would occur when inspection is done continuously, but the production rate is faster than the maximum inspection rate so that it is not possible to inspect all of the output. For example, suppose that an item is produced every 5 seconds, but the maximum inspection rate is one item every minute. If an item is selected every minute for inspection, then 8.33% of the output will be inspected. As in the first situation above, the p-chart would require 100 minutes to obtain the 100 inspected items required to plot a point.

As a third situation, consider a modification of the second situation above in which it is more convenient or desirable to have the inspected items be consecutive items from production. If an item is produced every 5 seconds, then it would take 500 seconds (8.33 minutes) to produce the 100 consecutive items required to make a sample for the p-chart. If the inspection rate is one item per minute, then inspecting these items would take 100 minutes. If inspection is started on another sample of 100 items after finishing with the previous 100, then the inspection process will be going on continuously. In each of the three situations considered above, inspection is done continuously over time, and the differences in the three situations pertain to how the items being inspected relate to the items being produced.

If the p-chart from this problem is set up with the 3σ control limits given by Equation (1), then the lower control limit is negative; thus, there is no effective lower control limit. The upper control limit is equivalent to giving a signal if $T_j \geq 4$, where T_j is the number of defectives in the j^{th} sample. Using this upper control limit gives $P(T_j \geq 4) = 0.01837$ as the probability of a signal for a sample when $p = p_0 = 0.01$; thus, the in-control ANSS is $1/0.01837 = 54.42$. This computation illustrates the fact that 3σ limits applied to the discrete binomial distribution can give a false alarm rate much different than what would occur for a normal distribution. In particular, an upper 3σ control limit for a normal distribution corresponds to an in-control ANSS of $1/0.00135 = 740.8$; thus, the value of 54.4 for the p-chart corresponds to a false alarm rate $740.8/54.4 = 13.6$ times higher than expected.

Suppose that it is desirable to obtain a more reasonable false alarm rate per sample for the p-chart, and so the upper control limit is adjusted so that a signal is given if $T_j \geq 5$. It is not possible to obtain a reasonable lower control limit for the p-chart when $n = 100$ because signaling when $T_j = 0$, the lowest possible nonnegative value, would result in $P(T_j = 0) = 0.3660$ when $p = p_0$. This would result in an extremely high false alarm rate. When $p = p_0$, $P(T_j \geq 5) = 0.00343$; thus, the in-control

TABLE 1. Values of h_B Which Will Approximately Give a Desired In-Control ANOS (Exact In-Control ANOS Values Appear in Parentheses)

Nominal	Adjusted			Desired In-Control ANOS								
p_1	p_0	p_1	m	500	1000	2000	4000	8000	16000	32000	64000	128000
	0.010	0.015027	81	1.765	2.420	3.272	4.309	5.519	6.864	8.321	9.864	11.469
				(503)	(1002)	(1997)	(3996)	(8013)	(15996)	(31921)	(63859)	(128143)
	0.020	0.029363	41	2.415	3.293	4.341	5.585	6.976	8.488	10.098	11.756	13.463
				(497)	(1004)	(1997)	(4012)	(8019)	(16020)	(32107)	(64016)	(128158)
	0.040	0.061450	20	3.150	4.150	5.300	6.550	7.900	9.300	10.750	12.250	13.800
$p_1 = 1.5p_0$				(495)	(1004)	(2029)	(4038)	(8060)	(15955)	(31675)	(63529)	(129339)
	0.060	0.084107	14	3.857	5.071	6.429	8.000	9.643	11.429	13.214	15.071	17.000
				(494)	(999)	(1971)	(3994)	(7908)	(16001)	(31640)	(63385)	(129275)
	0.100	0.153236	8	4.250	5.375	6.625	7.875	9.250	10.625	12.000	13.375	14.750
				(495)	(1005)	(2058)	(4038)	(8233)	(16490)	(32689)	(64420)	(126525)
	0.200	0.304899	4	4.750	6.000	7.000	8.250	9.500	10.500	11.750	13.000	14.250
				(469)	(1060)	(1962)	(4121)	(8508)	(15082)	(30687)	(62224)	(125933)
	0.001	0.002001	693	*	$h_B < 1$	*	1.453	1.970	2.667	3.411	4.237	5.123
				*	(1000)	*	(4000)	(8001)	(15988)	(31994)	(64024)	(128009)
	0.002	0.003990	347	$h_B < 1$	*	1.452	1.968	2.669	3.412	4.239	5.127	6.058
				(500)	*	(1999)	(3995)	(8009)	(15993)	(32012)	(64004)	(127995)
	0.003	0.006000	231	*	1.000	1.805	2.381	3.078	3.874	4.740	5.649	6.593
				*	(1001)	(1998)	(3993)	(7988)	(15993)	(32042)	(63958)	(127838)
	0.004	0.008020	173	*	1.451	1.965	2.659	3.399	4.220	5.098	6.023	6.971
				*	(1002)	(2005)	(4004)	(8008)	(16029)	(32008)	(64124)	(128056)
	0.005	0.009947	139	1.000	1.669	2.187	2.878	3.662	4.511	5.417	6.360	7.324
				(600)	(999)	(1999)	(3999)	(8000)	(15982)	(32026)	(64108)	(128093)
	0.006	0.012091	115	1.000	1.800	2.374	3.061	3.843	4.704	5.600	6.530	7.487
$p_1 = 2p_0$				(502)	(1002)	(2008)	(4004)	(7976)	(16052)	(32000)	(63950)	(128340)
	0.008	0.015870	87	1.448	1.966	2.655	3.402	4.230	5.115	6.046	7.000	7.977
				(500)	(1009)	(1993)	(3999)	(8016)	(16005)	(32027)	(63852)	(127803)
	0.010	0.020142	69	1.652	2.174	2.855	3.623	4.449	5.333	6.261	7.203	8.159
				(496)	(1006)	(2010)	(4011)	(7974)	(15947)	(32081)	(64134)	(128267)
	0.020	0.039221	35	2.171	2.857	3.629	4.486	5.371	6.314	7.286	8.257	9.257
				(501)	(1002)	(1988)	(4008)	(7918)	(15878)	(31929)	(63516)	(128101)
	0.040	0.082392	17	2.765	3.471	4.235	5.059	5.941	6.824	7.706	8.588	9.471
				(512)	(1003)	(1978)	(3954)	(8081)	(16252)	(32384)	(64192)	(126862)

TABLE 1. (Continued) Values of h_B Which Will Approximately Give a Desired In-Control ANOS (Exact In-Control ANOS Values Appear in Parentheses)

Nominal		Adjusted		Desired In-Control ANOS								
p_1	p_0	p_1	m	500	1000	2000	4000	8000	16000	32000	64000	128000
$p_1 = 2p_0$	0.060	0.111466	12	3.250	4.083	5.000	5.917	6.917	7.917	8.917	9.917	10.917
				(490)	(989)	(2015)	(3954)	(8048)	(16133)	(32067)	(63422)	(125088)
	0.100	0.194358	7	3.571	4.429	5.143	6.000	6.857	7.857	8.714	9.571	10.429
				(490)	(1044)	(1903)	(3826)	(7583)	(16671)	(32589)	(63545)	(123731)
$p_1 = 3p_0$	0.001	0.003002	549	*	$h_B < 1$	*	1.273	1.847	2.315	2.849	3.435	4.029
				*	(1000)	*	(4002)	(8003)	(15985)	(32030)	(63971)	(127935)
	0.002	0.005984	275	$h_B < 1$	*	1.273	1.847	2.316	2.847	3.436	4.029	4.640
				(500)	*	(2000)	(4011)	(8005)	(15985)	(32006)	(63874)	(127944)
	0.003	0.009000	183	*	1.000	1.667	2.038	2.628	3.186	3.765	4.372	4.989
				*	(1125)	(1996)	(3998)	(7972)	(16013)	(31921)	(63847)	(128242)
	0.004	0.012034	137	*	1.263	1.839	2.307	2.832	3.416	4.007	4.613	5.226
				*	(997)	(1997)	(4006)	(7969)	(15953)	(31978)	(64039)	(127873)
	0.005	0.014945	110	*	1.518	1.936	2.491	3.018	3.609	4.209	4.818	5.436
				*	(1002)	(1981)	(4004)	(7969)	(16002)	(31942)	(63822)	(127677)
	0.006	0.018151	91	1.000	1.659	2.022	2.615	3.165	3.736	4.341	4.945	5.549
				(565)	(1000)	(1997)	(4008)	(8041)	(16001)	(32194)	(64278)	(127565)
	0.008	0.023744	69	1.261	1.841	2.304	2.841	3.420	4.014	4.623	5.232	5.855
				(498)	(1008)	(1999)	(4033)	(7999)	(16029)	(32111)	(63596)	(127402)
	0.010	0.029844	55	1.509	1.927	2.473	3.000	3.582	4.182	4.782	5.400	6.018
				(501)	(992)	(1992)	(3986)	(7954)	(15964)	(31742)	(63839)	(127774)
	0.020	0.061350	27	1.889	2.407	2.926	3.481	4.074	4.630	5.222	5.815	6.407
				(485)	(983)	(2006)	(3953)	(8112)	(15768)	(31716)	(63541)	(127040)
	0.040	0.114800	14	2.429	2.929	3.500	4.071	4.714	5.286	5.929	6.500	7.071
				(508)	(994)	(1991)	(3916)	(8292)	(16021)	(33429)	(64129)	(122877)
	0.060	0.181420	9	2.556	3.111	3.667	4.111	4.667	5.222	5.778	6.333	6.889
				(483)	(1024)	(2135)	(3760)	(7599)	(15287)	(30631)	(61282)	(122508)
	0.100	0.334514	5	2.600	3.000	3.400	4.000	4.400	4.800	5.400	5.800	6.200
				(501)	(908)	(1718)	(4280)	(7847)	(14417)	(35718)	(65378)	(119628)
	0.001	0.003999	462	*	$h_B < 1$	*	1.123	1.758	2.078	2.595	3.043	3.535
				*	(1000)	*	(4001)	(8013)	(16008)	(32044)	(64020)	(128084)
	0.002	0.007994	231	$h_B < 1$	*	1.121	1.753	2.074	2.589	3.039	3.528	4.017
				(500)	*	(2000)	(3991)	(7993)	(15949)	(32018)	(63920)	(128170)
	0.003	0.011984	154	*	*	1.558	1.929	2.396	2.838	3.325	3.805	4.299
				*	*	(1992)	(3980)	(8015)	(16049)	(31897)	(64177)	(128112)
	0.004	0.016083	115	*	1.113	1.748	2.061	2.574	3.017	3.504	3.991	4.478
				*	(1000)	(2000)	(3994)	(7949)	(15936)	(31806)	(64024)	(127900)
	0.005	0.020093	92	*	1.391	1.859	2.250	2.707	3.185	3.652	4.141	4.630

TABLE 1. (Continued) Values of h_B Which Will Approximately Give a Desired In-Control ANOS (Exact In-Control ANOS Values Appear in Parentheses)

Nominal	Adjusted			Desired In-Control ANOS								
p_1	p_0	p_1	m	500	1000	2000	4000	8000	16000	32000	64000	128000
$p_1 = 4p_0$				*	(1002)	(2024)	(4018)	(9737)	(16113)	(31909)	(63838)	(128264)
	0.006	0.023929	77	*	1.558	1.922	2.390	2.818	3.312	3.792	4.273	4.766
				*	(1004)	(1993)	(4028)	(7927)	(15988)	(32284)	(63491)	(127741)
	0.008	0.031647	58	1.121	1.741	2.069	2.569	3.017	3.500	4.000	4.483	4.983
				(502)	(996)	(2015)	(3962)	(7966)	(15813)	(32328)	(64121)	(129417)
	0.010	0.040072	46	1.370	1.848	2.239	2.696	3.152	3.630	4.109	4.587	5.087
				(495)	(1014)	(2023)	(4022)	(7956)	(16090)	(31888)	(63481)	(129262)
	0.020	0.079680	23	1.826	2.217	2.652	3.087	3.565	4.043	4.522	5.000	5.478
				(509)	(1026)	(1993)	(3879)	(7925)	(15919)	(32091)	(64138)	(128361)
	0.040	0.168670	11	2.091	2.545	2.909	3.364	3.818	4.182	4.636	5.091	5.545
				(514)	(1058)	(1980)	(4080)	(8481)	(15147)	(31217)	(64237)	(131947)

ANSS is $1/0.00343 = 291.35$. Each sample consists of $n = 100$ items which corresponds to an in-control ANOS of 29,135 items. If an inspected item is available every minute, then this in-control ANOS corresponds to a false alarm rate of one per 29,135 minutes or 485.6 hours.

To design a Bernoulli CUSUM chart for this problem, suppose that process engineers decide that it would be desirable to quickly detect any special cause which increases p from 0.01 to 0.025 and that the in-control ANOS should be roughly 29,135 (the value corresponding to the p-chart in current use). From a previous discussion of the case of $p_0 = 0.01$ and $p_1 = 0.025$, it was shown that increasing p_1 slightly from 0.025 to about 0.02501 would give $r_1 = 0.0153$, $r_2 = 0.9320$, $r_1/r_2 = 1/61$, and thus, $m = 61$. An approximate value for h_B can be obtained using Table 1. If $p_1 = 2p_0$, then h_B should be about 6.1; and if $p_1 = 3p_0$, then h_B should be about 4.7. Thus, for $p_1 = 2.5p_0$, the value of h_B should be between 4.7 and 6.1. Using a starting value in this range and using trial and error (or the numerical procedure in Appendix D) to solve Equation (7) to give $\text{ANOS}(p_0) \approx 29{,}135$ results in a value of h_B^* of 5.57 (this value of h_B^* will give an in-control ANOS of 29,120 according to the approximation in Equation (7)). Then, using Equations (5) and (6) to convert to h_B gives $\epsilon(p) = 3.28$, $\epsilon(p_0)\sqrt{p_0 q_0} = 0.33$, and $h_B = 5.24$. As a point of interest, the exact in-control ANOS using $h_B = 5.24$ can be calculated to be 29,249 using the methods given in Appendices A and B. Thus, in this case, the CD approximation gives results which are extremely close to the exact value and certainly good enough for practical applications.

After h_B has been determined, Equation (8) can be used to determine how fast a shift from p_0 to p_1 will be detected. Using $h_B^* = 5.57$ in Equation (8) gives $\text{ANOS}(p_1) \approx 522.5$. As a point of interest, the exact ANOS can be calculated to be 526.0, so the CD approximation is also very good at $p = p_1$. Also, if it is assumed that the shift in p occurs between samples for the p-chart, then the ANOS of the p-chart at p_1 will be 939.7. Thus, the Bernoulli CUSUM chart will provide faster detection than the p-chart. The CD approximation can also be used to find the approximate ANOS at other values of p (see Appendix C). If the detection time at p_1 or at other values of $p > p_0$ is judged to be too high by process engineers, then the detection time can be decreased by decreasing h_B, but this will also increase the false alarm rate.

To illustrate the application of the Bernoulli CUSUM chart that has been designed, suppose that this CUSUM chart is applied to data from the process, and in the next 80 items inspected, the items numbered 3, 69, 72, 74, 77, 78, and 80 are found to be defective, while the other items are nondefective. For this sequence of inspection outcomes, the values of X_k, the increment $(X_k - \gamma_B)$, and the Bernoulli CUSUM statistic B_k are given in Table 2.

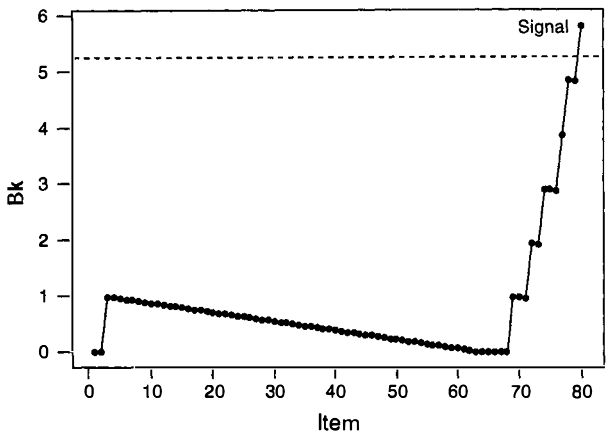

FIGURE 1. A Plot of the Bernoulli CUSUM Chart for the Example Data.

The Bernoulli CUSUM chart with B_k plotted and control limit $h_B = 5.24$ is shown in Figure 1. In the sequence of 80 items, there is one defective item early, followed by a long sequence of nondefectives. Then, starting at item number 69, many defectives begin to appear which would seem to indicate that something has happened to the process at about the time that item number 69 was produced. Note that with $\gamma_B = 1/61$, the possible values for the increment $(X_k - \gamma_B)$ are -1/61 and 60/61; thus, B_k drops down by -1/61 when the current item is nondefective and increases by 60/61 when the current item is defective. However, from the definition given in Equation (2), whenever B_k drops below 0, it resets back to 0 for adding the next increment. For the given data, B_k resets after items 1 and 2 but then increases to 60/61 after observing that item 3 is defective. Following item 3, there is a steady decline in B_k until it drops below 0 at item 64. The next 4 items are nondefective, so B_k continues to reset until a defective occurs at item number 69. After item 69, the frequent occurrences of defectives leads to B_k crossing the control limit and producing a signal at item number 80.

From the example data plotted in Figure 1 or from the definition in Equation (2), it is clear that a signal can be given only at an item which is defective. Thus, it is sufficient to know the value of B_k only at items which are defective. This means that the chart can be examined only at the defective items, and the path of B_k back to the previous defective can be easily reconstructed. For example, after the defective at item number 3, the value of B_k is 60/61; then there are 66 nondefectives before the next defective at item

TABLE 2. Sampling Pattern for the Example Data: $\gamma_B = 1/61$ and $h_B = 5.24$

k	X_k	$X_k - \gamma_B$	B_k	Action
1	0	−1/61	−1/61	Reset
2	0	−1/61	−1/61	Reset
3	1	60/61	60/61	*
4	0	−1/61	59/61	*
5	0	−1/61	58/61	*
⋮	⋮	⋮	⋮	⋮
62	0	−1/61	1/61	*
63	0	−1/61	0	*
64	0	−1/61	−1/61	Reset
65	0	−1/61	−1/61	Reset
66	0	−1/61	−1/61	Reset
67	0	−1/61	−1/61	Reset
68	0	−1/61	−1/61	Reset
69	1	60/61	60/61	*
70	0	−1/61	59/61	*
71	0	−1/61	58/61	*
72	1	60/61	118/61	*
73	0	−1/61	117/61	*
74	1	60/61	177/61	*
75	0	−1/61	176/61	*
76	0	−1/61	175/61	*
77	1	60/61	235/61	*
78	1	60/61	295/61	*
79	0	−1/61	294/61	*
80	1	60/61	354/61	Signal

number 69. If it is known at item number 69 that the previous defective was at item number 3, then it is easy to determine that there would be resets after items numbered 64–68 so that at item number 69, the value of B_k would be 60/61. If plotting is done by computer, then all points could easily be plotted as they are obtained; but for hand computation, some work could be saved by constructing the plot only after a defective is obtained.

It is useful to note that the CUSUM plot in Figure 1 clearly shows where the defectives occurred in the sequence of inspected items. In contrast, a chart based on samples of n, such as the p-chart, plots a summary statistic at the end of the sample and the location of the defectives cannot be determined from the plot. Thus, plotting a summary statistic may make it more difficult to effectively estimate the time that a special cause occurred.

Comparisons with Other Control Charts

The performance of the Bernoulli CUSUM chart will be compared to the performance of other charts by setting up all charts so that ANOS(p_0) is approximately the same for the charts and then comparing the ANOS values of the charts at p_1 and at other values of $p > p_0$. As discussed earlier, it would usually be desirable to consider the performance of the charts for a range of values of p around p_1; the specified p_1 is used mainly as a convenient design device to set up the Bernoulli CUSUM chart. Although the ANOS approximations given previously are highly accurate, the ANOS values used in the comparisons presented below are based on the exact Markov chain methods given in Appendices A and B.

Consider the previous example with $p_0 = 0.01$, $p_1 = 0.025$, and a required in-control ANOS value of approximately 29,135, corresponding to the p-chart in current use. A Bernoulli CUSUM chart with $m = 61$ and $h_B = 320/61 = 5.2459$ has an in-control ANOS of 29,248.6 items. The ANOS of this chart is given in Table 3 for a number of values of p. The sensitivity of the Bernoulli CUSUM to different shifts can be changed by changing p_1, which results in a change in the reference value γ_B. For example, suppose that p_1 is specified to be 0.04 instead of 0.025. This gives $r_2/r_1 = 46.05$, but if p_1 is adjusted slightly to approximately 0.04007, then r_2/r_1 will be the integer 46. Using $h_B = 186/46 = 4.043$ will give an

TABLE 3. Exact ANOS Values with Control Limits for Upper One-Sided Bernoulli CUSUM Charts, Upper One-Sided Shewhart p-Charts, and Upper One-Sided Binomial CUSUM Charts

	Bernoulli CUSUM Chart		Shewhart p-Chart			Binomial CUSUM Chart	
			$n = 51$	$n = 100$	$n = 158$	$n = 51$	$n = 100$
	$p_1 = 0.025$	$p_1 = 0.040$				$p_1 = 0.025$	
p	$h_B = 320/61$	$h_B = 186/46$	Control Limit = 4	Control Limit = 5	Control Limit = 6	$h_S = 275/61$	$h_S = 250/61$
0.010	29248.6	29050.8	29679.1	29134.8	29215.3	29499.0	30278.9
0.015	2847.2	3875.3	7061.0	5651.9	4825.5	2879.0	2897.6
0.020	951.7	1201.2	2688.2	1967.3	1598.8	973.4	986.0
0.025	526.6	587.4	1323.5	941.0	770.6	546.9	561.2
0.030	359.5	366.6	766.3	549.0	467.7	379.6	394.4
0.040	219.2	202.6	348.1	269.4	259.7	240.7	251.9
0.050	157.8	139.0	203.5	177.3	195.9	181.0	188.0
0.060	123.3	105.8	139.1	138.3	172.3	147.3	152.9
0.070	101.2	85.4	105.6	119.5	163.2	124.9	131.8
0.080	85.8	71.6	86.3	109.9	159.8	108.8	118.7
0.090	74.4	61.6	74.5	105.0	158.6	96.7	110.6
0.100	65.7	54.2	66.8	102.4	158.2	87.4	105.8
0.150	41.2	34.0	53.2	100.0	158.0	61.6	100.2
0.200	30.2	25.1	51.2	100.0	158.0	53.2	100.0
0.300	20.0	16.7	51.0	100.0	158.0	51.0	100.0
0.500	12.0	10.0	51.0	100.0	158.0	51.0	100.0
0.750	8.0	6.7	51.0	100.0	158.0	51.0	100.0
1.000	6.0	5.0	51.0	100.0	158.0	51.0	100.0

in-control ANOS of 29,050.8. ANOS values for this CUSUM chart are also given in Table 3. Compared to the CUSUM chart with $p_1 = 0.025$, the CUSUM chart with $p_1 = 0.04$ will offer faster detection of large shifts but slower detection of small shifts.

Table 3 includes ANOS values for the Shewhart p-chart in current use which is based on samples of $n = 100$ observations. The performance of the p-chart is affected by the choice of the number of items grouped into a sample, so two other sample sizes were considered for comparison purposes. The two sample sizes and corresponding control limits were chosen to give an in-control ANOS of approximately 29,135, corresponding to the p-chart with $n = 100$. Using samples of size $n = 51$ and a control limit of 4 for the p-chart will give an in-control ANOS of 29,679.1; using samples of size $n = 158$ with a control limit of 6 will give an in-control ANOS of 29,215.3. ANOS values for other values of p are given in Table 3 for these two p-charts. Comparing the ANOS values for the three p-charts in Table 3 shows that using a small value of n is better for detecting large shifts, while a large value of n is better for detecting small shifts.

Comparing the ANOS values of the two Bernoulli CUSUM charts and the three p-charts in Table 3 shows that the Bernoulli CUSUM chart will detect shifts in p much faster than the p-chart. For example, if p shifts to 0.025, then the p-chart with $n = 100$ requires, on average, 941.0 items to detect this shift, while the Bernoulli CUSUM chart with $p_1 = 0.025$ requires only 526.6 items, on average. If p shifts to 0.10, the p-chart requires an average of 102.4 items to detect this shift, while the Bernoulli CUSUM chart requires an average of 65.7 items.

When items are grouped into samples, an alternative to the p-chart is a CUSUM chart based on the statistics $T_1, T_2, \ldots$ which have a binomial distribution. This binomial CUSUM chart uses the control statistic

$$S_j = \max(0, S_{j-1}) + (T_j - \gamma_S), \quad j = 1, 2, \ldots,$$

and signals at segment j if $S_j \geq h_S$, where S_0 is the starting value and γ_S is the reference value. Gan (1993) discusses Markov chain methods for evaluating the properties of the binomial CUSUM chart. The appropriate choice of γ_S is nr_1/r_2 to detect a shift to p_1. For the case of $n = 100$ and $p_1 = 0.025$ in the previous example, using $\gamma_S = 100/61$ and $h_S = 250/61$ gives an in-control ANOS of 30,278.9 items. For the case of $n = 51$ and $p_1 = 0.04$, using $\gamma_S = 51/61$ and $h_S = 275/61$ gives an in-control ANOS of 29,499.0 items. The ANOS values for these two binomial CUSUM charts are also given in Table 3. The ANOS values in Table 3 show that, for detecting shifts in p, using $n = 51$ is a little better than using $n = 100$ in the binomial CUSUM chart. From Table 3, the binomial CUSUM chart appears to be a little slower in detecting shifts than the Bernoulli CUSUM chart. The binomial CUSUM chart has the worst relative performance when n and p are large. Grouping items into samples will always be a disadvantage in detecting very large shifts in p with the binomial CUSUM chart because no matter how many defectives are suddenly observed this CUSUM chart must wait until a sample is completed before a statistic can be computed and a signal given. Note that when the sample size is $n = 1$, the binomial and Bernoulli CUSUM charts are equivalent.

The Relationship Between the Bernoulli and Geometric CUSUM Charts

The earlier discussion of plotting B_k (in the example section) shows that it is necessary to plot B_k only at items which are defective. This suggests that the essential information needed is the item numbers for the defectives that occur. To explain this more clearly, let D_j represent the number of the item which is the j^{th} defective. For the data in the previous example, $D_1 = 3$, $D_2 = 69$, $D_3 = 72$, ... Knowing the values of the D_j's is equivalent to knowing D_1 and the number of items between the other D_j's. Thus, let $Y_j = D_j - D_{j-1}$ represent the number of items from defective number $j - 1$ to defective number j, where $D_0 = 0$. For the example, $Y_1 = D_1 - D_0 = 3$, $Y_2 = D_2 - D_1 = 66$, $Y_3 = D_3 - D_2 = 3$, ... The statistics $Y_1, Y_2, \ldots$ are independent geometric random variables and can be used in a Shewhart chart or in a CUSUM chart for the geometric distribution. Note that the subscript j in Y_j does not correspond to the j^{th} segment (as does T_j used in the p-chart and binomial CUSUM chart), but rather corresponds to the j^{th} defective which will occur at an unknown item. The fact that knowing the values of the original X_k's is equivalent to knowing the values of the Y_j's suggests that it would be reasonable to base a CUSUM chart directly on the Y_j's. Note that if $p = p_0$, then Y_j would be expected to be large; while if p increases, then Y_j would be expected to be smaller. Thus, observing small values of Y_1, Y_2, ... would tend to indicate that p has increased. A CUSUM chart based on this sequence would use the control statistic

$$G_j = \min(0, G_{j-1}) + (Y_j - \gamma_G), \quad j = 1, 2, \ldots,$$

and would signal at item D_j (defective j) if $G_j \le h_G$, where G_0 is the starting value, $\gamma_G = r_2/r_1 = 1/\gamma_B$ is the reference value, and $h_G < 0$ is the control limit. This CUSUM chart was investigated in detail by Bourke (1991). However, Bourke (1991) counted the number of nondefective items between items D_{j-1} and D_j so that his statistic at item D_j is $Y_j - 1$ rather than Y_j. The two ways of defining the geometric CUSUM are equivalent because the two sequences $\{Y_j - 1\}$ and $\{Y_j\}$ contain exactly the same information about the process parameter p.

The geometric CUSUM chart is equivalent to the Bernoulli CUSUM chart if the Bernoulli CUSUM is used with a head start. In particular, if the starting value in the Bernoulli CUSUM is $B_0 = 1$, then the two charts will be equivalent if $G_0 = 0$, $\gamma_G = r_2/r_1 = 1/\gamma_B$, and $h_G = -mh_B - m + 1$. Under these conditions, it follows that the two charts will have the same ANOS; thus, there is no reason to prefer one to the other in terms of the ability to detect shifts in p.

For practical applications, a possible advantage of the Bernoulli CUSUM chart relative to the geometric CUSUM chart is that the Bernoulli CUSUM statistic B_k may be easier to interpret. The value of B_k can be plotted after each observation, and the occurrence of a defective corresponds to a large jump of size $(m-1)/m$, while for the geometric CUSUM statistic G_j, low values indicate a possible increase in p.

Another advantage of the Bernoulli CUSUM chart is that the relatively simple CD approximation can be used to design the chart for practical applications. However, the Bernoulli CUSUM chart and the geometric CUSUM chart are equivalent; thus, it follows that the approximations for the Bernoulli CUSUM chart can be used to design the geometric CUSUM chart for any applications for which the form of the geometric CUSUM chart seems to be preferable. This design can be accomplished by designing a Bernoulli CUSUM chart to meet the required objectives and then using the relationships discussed above to construct the corresponding geometric CUSUM chart.

The Bernoulli CUSUM Chart for Detecting a Decrease in p

In some situations, it will be desirable to detect any decrease in p below the in-control value p_0. A significant decrease in the observed proportion of defective items could be the result of attempts to improve the process or could indicate that the inspection process has deteriorated and is not correctly identifying all of the defectives. For detecting a decrease in p from p_0 to $p_1 < p_0$, the Bernoulli CUSUM control statistic becomes

$$B_k = \min(0, B_{k-1}) + (X_k - \gamma_B), \quad k = 1, 2, \ldots. \tag{9}$$

The Bernoulli CUSUM chart will signal that there has been a decrease in p if $B_k \le h_B$, where $h_B < 0$ is the control limit. Based on the representation of a CUSUM chart as a sequence of SPRT's, the reference value γ_B should be $\gamma_B = r_1/r_2$ as given in Equation (4), where r_1 and r_2 are defined in Equation (3). When $p_1 < p_0$, both r_1 and r_2 will be negative, and thus, γ_B will be positive. For detecting a decrease in p, we look for nondefectives which correspond to $(X_k - \gamma_B) = -\gamma_B < 0$. Thus, B_k will move down when nondefectives occur, and a signal that p has decreased will be given if B_k drops too far below 0.

As in the case of detecting an increase in p, when detecting a decrease in p it will usually be convenient if $\gamma_B = 1/m$, where m is an integer. As an example, suppose that $p_0 = 0.02$ and it is desirable to detect a decrease to $p_1 = 0.01$. Then, $r_1 = -0.01015$, $r_2 = -0.70330$, and $r_1/r_2 = 0.01444 = 1/69.27$. In this case, if p_1 is increased slightly from 0.01 to 0.01009, then r_1/r_2 will increase slightly to 1/69, and the possible values of B_k will be integer multiples of 1/69.

The CD approximation in Equation (7) for the ANOS at p_0 and in Equation (8) for the ANOS at p_1 still apply in the case of $p_1 < p_0$. To relate the adjusted value h_B^* to h_B, modify Equation (5) to

$$h_B^* = h_B - \epsilon(p_0)\sqrt{p_0 q_0}, \tag{10}$$

where Equation (6) is used to obtain $\epsilon(p)$. For given values of r_1 and r_2 and a desired value for the in-control ANOS, Equation (7) can be used to find the required value of h_B^*; then Equations (6) and (10) can be used to find the required value of h_B.

As an example of designing the Bernoulli CUSUM chart for detecting a decrease in p, consider the previous example of this section in which the in-control value of p is $p_0 = 0.02$ and it is desirable to detect a shift to $p_1 = 0.01$. If a p-chart is used for this problem, the sample size n must be relatively large in order to have a lower control limit with a reasonable false alarm rate. Suppose that inspected items are grouped into samples of $n = 200$ and that the lower control limit is taken to be 0, so that a signal is given that p has decreased if all 200 items are nondefective. For purposes of this example, assume that there is no upper control limit for the p-chart. When $p = p_0 = 0.02$, the probability that this one-sided lower p-chart will signal after any segment is 0.01759; thus, the ANSS is $1/0.01759 = 56.86$. Each sample consists of 200 items, which corresponds to an in-control ANOS of 11,371 items.

Suppose that a Bernoulli CUSUM chart is set up to have approximately the same in-control ANOS as the p-chart. If p_1 is increased slightly from 0.01 to 0.01009, then this will give $r_1 = -0.01006$, $r_2 = -0.6942$, and $\gamma_B = r_1/r_2 = 1/69$. Using trial and error to solve Equation (7) to give $\text{ANOS}(p_0) \approx 11,371$ results in a value of h_B^* of -5.59 (this value of h_B^* will give an in-control ANOS of 11,398, according to the approximation in Equation (7)). Then, using Equations (6) and (10) to convert to h_B gives $\epsilon(p) = 2.31$, $\epsilon(p_0)\sqrt{p_0 q_0} = 0.32$, and $h_B = -5.27$. The exact in-control ANOS using $h_B = -5.27$ can be calculated to be 11,525, so, in this case, the CD approximation is very good.

From Equation (8), the approximate ANOS of the Bernoulli CUSUM chart at $p_1 = 0.01009$ is 949 (the exact value can be calculated to be 948). The ANOS of the p-chart at p_1 is 1,520, so the Bernoulli CUSUM is substantially faster at detecting this decrease in p.

A Two-Sided Bernoulli CUSUM Scheme

In most cases in which it is desirable to detect a decrease in p, it would also be desirable to detect an increase in p, because it is the increase in p that actually corresponds to lower quality. To detect both increases and decreases in p, two Bernoulli CUSUM charts can be run together to give a two-sided control scheme. One of these two charts would use B_k given by Equation (2) for detecting an increase in p, and the other would use B_k given by Equation (9) for detecting a decrease in p. These two charts will be called the upper and lower charts, respectively. Let $\text{ANOS}_U(p)$ and $\text{ANOS}_L(p)$ be, respectively, the ANOS for the upper and lower charts and let $\text{ANOS}_T(p)$ be the ANOS of the resulting two-sided procedure. Under general conditions (see, e.g., Siegmund (1985) or Yashchin (1985)), the ANOS function of the two-sided Bernoulli CUSUM chart satisfies the relation

$$\frac{1}{\text{ANOS}_T(p)} \geq \frac{1}{\text{ANOS}_L(p)} + \frac{1}{\text{ANOS}_U(p)}. \qquad (11)$$

In the case of monitoring the mean of a normal process characteristic, the previous inequality becomes an equality.

Because the CD approximation in Equation (7) is very good and because its derivation is based on approximating the standardized Bernoulli variable, $Z = (X - p)/\sqrt{p(1-p)}$, by the normal distribution of a Brownian motion (see Appendix C), it should be possible to treat the inequality in Equation (11) as an approximate equality. This results in

$$\text{ANOS}_T(p) \approx \frac{\text{ANOS}_L(p) \cdot \text{ANOS}_U(p)}{\text{ANOS}_L(p) + \text{ANOS}_U(p)}. \qquad (12)$$

The approximation in Equation (12) can be applied using the individual CD approximations for $\text{ANOS}_L(p)$ and $\text{ANOS}_U(p)$. For example, suppose that it is desirable to have a two-sided Bernoulli CUSUM chart with an in-control ANOS of approximately 15,000. If the corresponding lower and upper Bernoulli CUSUM charts are each designed to have an in-control ANOS of approximately 30,000, then from Equation (12) it follows that the in-control ANOS of the two-sided CUSUM scheme will be approximately 15,000. The upper and lower Bernoulli CUSUM charts can then be designed using Equation (7).

Conclusions

The Bernoulli CUSUM chart has been developed here for the problem of monitoring a process proportion p. In contrast to the traditional practice of grouping items into samples and applying a Shewhart p-chart, the Bernoulli CUSUM chart makes use of each inspection result as it becomes available. This gives a CUSUM chart which will detect shifts in p much faster than the p-chart.

An obstacle that has hindered the application of CUSUM charts to the problem of monitoring a proportion is that it has been difficult for the practitioner to determine chart properties such as the ANOS. Without knowledge of the properties of the chart, there is no way to design the chart to give a reasonable false alarm rate or to determine whether the detection time for shifts of interest will be fast enough. The properties of the chart for a given value of p depend on values chosen for p_0 and p_1; thus, it is difficult to publish design tables that will cover all cases that might be of interest in applications. The design method presented here for the Bernoulli CUSUM chart relies on the relatively simple CD approximation. This approximation is very accurate and allows the chart to be designed using only a pocket calculator. Analogous CD approximations to the one presented here for the Bernoulli CUSUM chart have also been developed for CUSUM charts based on other discrete distributions, such as the binomial and Poisson distributions (see, e.g., Reynolds and Stoumbos (1998a)). A rigorous development and evaluation of the CD approximations in the context of various discrete distributions is to be presented in a future theoretical paper.

The Bernoulli CUSUM chart has been investigated here for the situation in which a continuous stream of inspected items is available. However, this chart could be used when there are inspection periods in which samples of a fixed size are taken from the process and when there are non-inspection periods when no items are inspected. For example, suppose that samples of $n = 50$ items are inspected twice per eight-hour shift and that inspection takes 30 seconds per item. Then, in 4 hours there would be a 25-minute inspection period and a 215-minute non-inspection period. In an example such as this, the Bernoulli CUSUM chart could be applied by ignoring the non-inspection periods and treating the individual observations in the samples as a sequence of Bernoulli observations. Thus, the design of the Bernoulli CUSUM chart for this situation presents no significant additional difficulties. However, a thorough comparison of the Bernoulli CUSUM chart with other control charts for this situation would require the consideration of the possibility that the shift in p falls within an inspection period. For charts based on samples of n, this would mean that p changes while a sample is being taken. A comprehensive evaluation of this situation will be considered in a future paper.

The example above with the 25-minute inspection period every 4 hours involves a constant rate of sampling which is less than the maximum possible rate. In situations such as this it is possible to increase the efficiency of the control chart by using a variable sampling rate (VSR) control chart which varies the sampling rate as a function of the data obtained from the process. The sampling rate can be varied by varying the sampling interval and/or by varying the sample size. In the example of taking samples of $n = 50$ every 4 hours, if the current sample shows some indication of an increase in p, then the next sample could be taken sooner than in 4 hours, say, within an hour and, in addition, this sample could be larger than 50. On the other hand, if the current sample shows no sign of an increase in p, then the next sample might be a smaller sample taken in, say, 8 hours. For a given in-control average sampling rate and false alarm rate, VSR charts can detect many process parameter changes much faster than traditional charts. Rendtel (1990) considered VSR CUSUM charts for monitoring p. Reynolds and Stoumbos (1998b) give additional references to VSR charts and develop and evaluate one particularly efficient VSR chart for monitoring p. This VSR chart is based on applying an SPRT at fixed sampling points and is called the SPRT chart. If for a particular application it is feasible to have a sampling rate that varies as a function of the data, then it is recommended that the SPRT chart in Reynolds and Stoumbos (1998b) be considered.

Appendix A: The Exact Statistical Properties in Terms of the Full Transition Probability Matrix of the Markov Chain

The ANOS of the Bernoulli CUSUM chart can be found by modeling the chart as a Markov chain (see Zacks (1981) and Hawkins (1992) for related methods for general discrete distributions). First consider the problem of detecting an increase in p. The possible values of the observation X_k are 0 and 1; thus, the possible values of the control statistic B_k are determined by the choice of $\gamma_B = r_1/r_2$. When γ_B is a rational number, there will be a finite number of possible values for B_k, and these values will correspond to the states of the Markov chain. In most quality control applications, the values of p_0 and p_1 will be relatively small, and, in this case, it will be reasonable to make a slight adjustment in p_1 so that r_2/r_1 will be an integer and so that the possible values of B_k will be integer multiples of r_1/r_2. To make the presentation of the Markov chain as simple as possible, the case in which r_2/r_1 is an integer will be presented here. If $r_2/r_1 = m$, where m is an integer, then the control limit h_B can be taken to be an integer multiple of $1/m$. In this case, the possible values of B_k which are below h_B are $-1/m$, 0, $1/m$, $2/m$, $\ldots$, $h_B - 1/m$. When $B_k = -1/m$, there will be a reset to 0 on the next sample, so it is sufficient to start with the value 0 when defining the transient states of the Markov chain. Let state i correspond to the value $(i-1)/m$ for $i = 1, 2, \ldots, mh_B$ and let $t = mh_B$ be the number of transient states. The possible values of $(X_k - 1/m)$ are $-1/m$ and $(m-1)/m$; thus, after each observation the transition will either be down 1 state or up $(m-1)$ states.

If $h_B < 1$, then from any state a signal will be given if a defective is observed. This means that the Bernoulli CUSUM chart reduces in this case to a rule which signals as soon as a defective is observed, so the ANOS is simply $1/p$. Thus, in what follows we assume that $h_B \geq 1$. Let $c = t - m + 1$ represent the highest state from which it is not possible to signal. This means that if the current state is $i \leq c$, then observing a defective will result in a transition to state $i + m - 1 \leq t$, but if the current state is $i > c$, then observing a defective will result in a signal. If $h_B \geq 1$, then this implies that $c \geq 1$. The transition probability matrix for this Markov chain is relatively simple. In particular, if p_{ij} is the transition probability from state i to state j and $q = 1 - p$, then

$$\begin{aligned} p_{11} &= q, \\ p_{1m} &= p, \\ p_{i,i-1} &= q \quad \text{for } i = 2, 3, \ldots, t, \\ p_{i,i+m-1} &= p \quad \text{for } i = 2, 3, \ldots, c, \text{ if } c \geq 2. \end{aligned} \tag{A1}$$

Let N_i be the ANOS when starting in state i and let $\mathbf{N} = (N_1, N_2, \ldots, N_t)'$ be the vector of ANOS values for the t transient states. If $\mathbf{Q}$ is the transition probability matrix from Equation (A1) for these states, then it follows from basic properties of Markov chains that the ANOS vector is

$$\mathbf{N} = \mathbf{M1}, \tag{A2}$$

where $\mathbf{M} = (\mathbf{I} - \mathbf{Q})^{-1}$ is the fundamental matrix of the Markov chain and $\mathbf{1}$ is a column vector of 1's. Unless a headstart is used in the CUSUM, the ANOS of interest would usually be N_1, the first component of N.

Writing a computer program to find the ANOS using Equation (A2) is straightforward, but when the values of p_0 and p_1 are small, the number of states will be quite large, and then it may not be possible to work directly with the transition probability matrix $\mathbf{Q}$. An alternative to using Equation (A2) for finding the ANOS from the Markov chain is given in Appendix B.

For the problem of detecting a decrease in p, the Bernoulli CUSUM control statistic is B_k, given by Equation (9), and a signal is given if $B_k \leq h_B$, where $h_B < 0$. If $\gamma_B = r_1/r_2 = 1/m$, then the possible values of $(X_k - 1/m)$ are $-1/m$ and $(m-1)/m$. Thus after each observation, B_k will either drop by $1/m$ or increase by $(m-1)/m$ (if there is no reset). In this case, h_B can be taken to be an integer multiple of $1/m$, and the possible values of B_k will be the integer multiples of $1/m$, which are between h_B and $(m-1)/m$, inclusive. To model this chart as a Markov chain, the highest value of B_k that needs to be considered in defining the transient states is 0. Thus, let state i correspond to the value $-(i-1)/m$ for $i = 1, 2, \ldots, -mh_B$ and let $t = -mh_B$ be the number of transient states. The transition probabilities can easily be determined, as was done for the case of $p_1 > p_0$ above, and the ANOS vector can be obtained using the expression in Equation (A2) with the appropriate $\mathbf{Q}$.

The ANOS at an off-target value of p is a measure of the time required to detect a shift to this value of p. Computing N_1 for this value of p gives the ANOS under the assumption that the control statistic B_k is 0 at the time that the shift in p occurs. However, in practice, the shift in p may occur at some random time when B_k is not at 0. In the quality control literature, the expected time to detect a shift that occurs at some random time in the future is frequently modeled assuming that the shift occurs after the control statistic has reached a stationary or steady-state distribution. In this situation, the ANOS would usually be called the *steady-state* ANOS. Using the Markov chain described above, it is a relatively simple matter to compute the steady-state ANOS for the Bernoulli CUSUM chart. However, it is more difficult to compute the steady-state ANOS of the p-chart or the binomial CUSUM chart, because the shift in p would be likely to occur somewhere within a sample; thus the number of defectives in this sample would be the sum of two binomial random variables with different values of p. Because the steady-state ANOS has not been developed for all of the charts that are being compared here, the steady-state ANOS was not used in the comparisons.

Appendix B: The Exact ANOS Based on the Explicit Solution of Linear Equations

The transition probability matrix $\mathbf{Q}$ determined by the transition probabilities in Equation (A1) has a simple structure, so it is possible to solve directly for the ANOS, although the resulting expression is rather complicated. The direct solution will be given here for the case of $p_0 < p_1$. Using Equation (A2) or a direct argument shows that $\mathbf{N}$ satisfies the equation

$$\mathbf{N} = \mathbf{1} + \mathbf{QN}. \tag{B1}$$

Consider first the case in which $c \geq m$. Writing out the system of equations in Equation (B1) gives

$$\begin{aligned} N_1 &= 1 + qN_1 + pN_m \\ N_2 &= 1 + qN_1 + pN_{m+1} \\ N_3 &= 1 + qN_2 + pN_{m+2} \\ &\vdots \\ N_c &= 1 + qN_{c-1} + pN_t \\ N_{c+1} &= 1 + qN_c \\ N_{c+2} &= 1 + qN_{c+1} \\ &\vdots \\ N_t &= 1 + qN_{t-1}\,. \end{aligned} \tag{B2}$$

The approach used to solve these equations is to find N_c; then for $i \neq c$, obtain N_i from N_c. For $i > c$, it is relatively easy to show that

$$N_i = \frac{1 - q^{i-c}}{p} + q^{i-c} N_c\,. \tag{B3}$$

Once N_c and N_i for $i > c$ are obtained, N_i for $i < c$ can be obtained by finding $N_{c-1}, N_{c-2}, \ldots, N_2$ successively, using the equation

$$N_i = \frac{1}{q}\left(N_{i+1} - pN_{i+m} - 1\right)\,. \tag{B4}$$

Equation (B4) is obtained from the equation $N_{i+1} = 1 + qN_i + pN_{i+m}$, which, from Equation (B2), holds for $i = 2, 3, \ldots, c$. N_1 is obtained using the first equation in the system of equations in Equation (B2).

The expression for N_c is obtained by working up through the system of equations in Equation (B2), with each N_i being expressed in terms of $N_{i+1}, N_{i+2}, \ldots$. Once N_c is reached in the sequence, Equation (B3) can be used to give an equation with terms involving only N_c. This equation can then be solved explicitly for N_c. The resulting solution can be expressed using a set of constants $T(i', i)$ defined by

$$T(0,i) = \begin{cases} 0 & \text{if } i = -1, 0 \\ 1 & \text{if } i = 1, 2, \ldots, m-1, \end{cases}$$

$$T(1,i) = \begin{cases} b & \text{if } i = -1 \\ b - ia & \text{if } i = 0, 1, \ldots, m-1, \end{cases}$$

$$T(2,i) = \begin{cases} T(1, m-1) & \text{if } i = -1 \\ T(1, m-1) - aT(1,0) & \text{if } i = 0, 1, \\ & \quad \ldots, m-1, \end{cases}$$

and for $i' \geq 2$,

$$T(i', i) = \begin{cases} T(i'-1, m-1) & \text{if } i = -1 \\ T(i'-1, m-1) - aT(i'-1, 0) & \text{if } i = 0 \\ T(i'-1, m-1) - a \sum\limits_{j=0}^{i} T(i'-1, j) & \\ = T(i', i-1) - aT(i'-1, i) & \text{if } i = 1, 2, \\ & \quad \ldots, m-1, \end{cases}$$

where $a = pq^{m-1}$ and $b = 1 - q^{m-1}$. The constants $T(i', i)$ can be evaluated recursively. Now, express c as $c = i'm + k$ for some $i' = 1, 2, \ldots$ and $k = 0, 1, \ldots, m-1$. Then,

$$\begin{aligned} N_c = {} & \frac{1}{T(i'+1, k-1)} \left[\sum_{j=0}^{i'-2} q^{jm+k} \right. \\ & \left. \times \sum_{i=1}^{m} q^{m-i} T(i'-1-j, i-1) + \frac{q^{(i'-1)m+k}}{p} \right] \\ & + \frac{1}{T(i'+1, k-1)} \\ & \times \left[\sum_{i=k+2}^{m} T(i'-1, i-1)(q^{m+k-i} - q^{m-1}) \right. \\ & \left. + \sum_{i=1}^{k} T(i', i-1)(2q^{k-i} - q^{m-1}) \right]. \end{aligned} \tag{B6}$$

This expression will allow the evaluation of the ANOS for $p_0 < p_1$ for any number of states in the Markov chain. However, Equations (B5) and (B6) are sufficiently complicated that a program would usually be required for evaluation. For the case in which $1 \leq c < m$, a derivation similar to the one above gives

$$N_c = \frac{1 + T(1, c-1)}{pT(1, c-1)}\,.$$

Appendix C: A CD Approximation to the ANOS

The expressions for the ANOS in Appendices A and B give the exact ANOS but are computationally complex. The CD approximation to be presented here is suitable for computation using a simple pocket calculator. The technical, theoretical aspects of the CD approximation are beyond the scope of this paper, and thus, will be presented in a future paper.

The Bernoulli CUSUM chart for detecting a shift from p_0 to p_1 is equivalent to a sequence of identical SPRT's for testing the null hypothesis, $H_0 : p = p_0$, versus the alternative hypothesis, $H_1 : p = p_1$. Wald

(1947) originally developed the SPRT and derived simple approximations to its properties. However, in these derivations, Wald ignored the excess (overshoot) of the SPRT's test statistic over its acceptance limit, g, and rejection limit, h, at the termination of the SPRT. Much later, for a certain exponential family of continuous distributions (which includes the normal distribution), Siegmund (1979, 1985) developed CD approximations by deriving the limiting expected excess of the SPRT's test statistic over g and h and correcting for this excess in Wald's approximations. Although Siegmund's corrections substantially improve the accuracy of Wald's approximations in the context of discrete time, they still provide poor approximations for SPRT's with g that are relatively near or equal to zero (Stoumbos and Reynolds (1997b)), as in the case of the Bernoulli CUSUM and all CUSUM charts in general. However, for the ANOS of CUSUM charts based on the same exponential family of continuous distributions, Siegmund (1979, 1985) developed very good CD approximations by replacing the random walk of the control statistic of a CUSUM chart with a continuous Brownian motion process. Woodall and Adams (1993) and Stoumbos and Reynolds (1997b) considered the design of CUSUM schemes and SPRT charts for monitoring the mean of a normal quality characteristic using Siegmund's (1979, 1985) CD approximations.

The Bernoulli distribution does not belong to the exponential family of continuous distributions that Siegmund (1979, 1985) considered in developing his CD approximations to the ANOS of a CUSUM chart. Therefore, these approximations do not apply directly to the Bernoulli CUSUM chart. Here, the work of Siegmund (1979, 1985) is extended to the ANOS of the Bernoulli CUSUM by standardizing a Bernoulli random variable and approximating the standardized Bernoulli variable by a Brownian motion process. The limiting expected excess of the SPRT's test statistic over the acceptance and rejection limits can be derived using a smoothing argument. The derivation of the CD approximation to the ANOS of the Bernoulli CUSUM chart is outlined below.

Let X denote a Bernoulli random variable with $P(X = 1) = p$. Further, suppose that $0 < p < 1$ and standardize X to obtain $Z = (X - p)/\sqrt{p(1-p)}$. Then,

$$E_p\left(Z^3\right) = \left(\frac{1-p}{p}\right)^{1/2} - \left(\frac{1-p}{p}\right)^{-1/2}$$

and

$$E_p(\exp\{i\lambda Z\}) = \exp\left\{i\lambda\left(\frac{1-p}{p}\right)^{1/2}\right\}p + \exp\left\{-i\lambda\left(\frac{1-p}{p}\right)^{-1/2}\right\}(1-p),$$

where $i = \sqrt{-1}$. Further, let

$$\epsilon(p) = \frac{1}{6}E_p\left(Z^3\right) - \frac{1}{\pi}\lim_{s\to 0}\int_0^\infty \frac{1}{\lambda^2}\,\mathrm{Relog}\left\{\frac{2}{\lambda^2}\times\left[1 - E_p(\exp\{i\lambda Z\})\exp\left\{-\frac{(\lambda s)^2}{2}\right\}\right]\right\}d\lambda. \tag{C1}$$

Then, it can be shown that when $p = p_0$, the limiting expected excess of the control statistic of the upper one-sided Bernoulli CUSUM over h_B is $\epsilon(p_0)\sqrt{p_0 q_0}$. It can also be shown that $\epsilon(p)$ satisfies the relationship

$$\begin{aligned}\epsilon(p) &= -\frac{1}{3}E_{1-p}\left(Z^3\right) + \epsilon(1-p)\\ &= \frac{1}{3}E_p\left(Z^3\right) + \epsilon(1-p).\end{aligned} \tag{C2}$$

For $0 < p \leq 0.5$, a very good simple approximation to $\epsilon(p)$ in Equation (C1) is given by Equation (6). For $0.5 < p < 1$, a very good approximation to $\epsilon(p)$ is obtained by evaluating $\epsilon(1-p)$ using the approximation in Equation (6) and then substituting the result in the relation given by Equation (C2). In Equation (6), the approximation for $0.01 \leq p \leq 0.5$ was fit using 40 points from the interval [0.004, 0.5] with $R^2 = 100\%$. The approximation for $0 < p < 0.01$ gives at least one-decimal accuracy. Since $\epsilon(p)$ requires a complicated numerical evaluation, in many practical situations, the approximation in Equation (6) will be very useful.

Results presented here are for the special case $\gamma_B = r_1/r_2 = 1/m$, where m is an integer. In this special case, each SPRT in the Bernoulli CUSUM chart will have acceptance limit $g_B = -r_1/r_2$ and rejection limit h_B. When $p_0 < p_1$, the issue of expected excess below the reset limit, g_B, does not arise because B_k drops by steps of size r_1/r_2 and thus, will always exactly hit g_B. When $p_1 < p_0$, the issue of expected excess below the control limit h_B does not arise because B_k drops by steps of size r_1/r_2 and thus, will always exactly hit h_B. Replacing the random walk of the control statistic of the Bernoulli CUSUM chart (based on the standardized observations $Z_1 = (X_1 - p)/\sqrt{p(1-p)}$,

$Z_2 = (X_2 - p)/\sqrt{p(1-p)}, \ldots$) by a continuous Brownian motion process and using limiting arguments on the representation of the CUSUM as a sequence of SPRT's, a CD approximation to the ANOS of the Bernoulli CUSUM chart can be expressed as

$$\text{ANOS}(p) \approx \begin{cases} \dfrac{\exp\{h_B^* r_2\} - h_B^* r_2 - 1}{|r_2 p_0 - r_1|} & \text{if } p = p_0, \\ \dfrac{\exp\{-h_B^* r_2\} + h_B^* r_2 - 1}{|r_2 p_1 - r_1|} & \text{if } p = p_1, \\ \dfrac{h_B^*\left(h_B^* + \left|-\frac{r_1}{r_2}\right|\right) r_2^2}{r_1(r_2 - r_1)} & \text{if } p = \frac{r_1}{r_2}, \\ \dfrac{\exp\{\xi(p) h_B^* r_2\} - \xi(p) h_B^* r_2 - 1}{|\xi(p)(r_2 p - r_1)|} & \text{otherwise,} \end{cases} \tag{C3}$$

where $h_B^* = h_B + \epsilon(p_0)\sqrt{p_0(1-p_0)}$, where $\xi(p) \neq 0$ denotes the non-zero solution to the equation

$$E_p\left(\exp\{Z^* \xi(p)\}\right) = \left(\frac{p_1}{p_0}\right)^{\xi(p)} p + \left(\frac{1-p_1}{1-p_0}\right)^{\xi(p)} (1-p) = 1, \tag{C4}$$

and where

$$Z^* = \log\left(\frac{f(X;p_1)}{f(X;p_0)}\right) = \log\left(\frac{p_1^X (1-p_1)^{1-X}}{p_0^X (1-p_0)^{1-X}}\right) = \log\left(\frac{1-p_1}{1-p_0}\right) + X \log\left(\frac{p_1(1-p_0)}{p_0(1-p_1)}\right).$$

That is, $f(x;p)$ denotes the Bernoulli density with $P(X=1) = p$. The case $p = r_1/r_2$ in Equation (C3) arises from the theory of SPRT's (see, e.g., Wald (1947), Ghosh (1970), or Siegmund (1985)) as the solution to the equation $E_p(Z^*) = 0$. That is, for $p = r_1/r_2$, $E_{r_1/r_2}(Z^*) = 0$. In all computations presented in this paper, the approximation in Equation (6) will be used for the computation of $\epsilon(p_0)$ in the approximation in Equation (C3).

To approximate the ANOS of the Bernoulli CUSUM for $p \neq p_0, p_1$, or r_1/r_2 using Equation (C3), the value of $\xi(p)$ that satisfies Equation (C4) must be determined. This can be done by trial and error or by using a numerical procedure such as the iterative Newton-Raphson procedure (see, e.g., Cheney and Kincaid (1985) or Burden and Faires (1993)). The Newton-Raphson procedure applied directly to Equation (C4) was found to be quite sensitive to the choice of starting value. This is undesirable in practical situations. Thus, to determine the value of $\xi(p)$ that satisfies Equation (C4), the Newton-Raphson procedure will be applied to the equation

$$p - \frac{1 - \left(\frac{1-p_1}{1-p_0}\right)^{\xi(p)}}{\left(\frac{p_1}{p_0}\right)^{\xi(p)} - \left(\frac{1-p_1}{1-p_0}\right)^{\xi(p)}} = 0, \tag{C5}$$

which is equivalent to Equation (C4) and was found to be insensitive to the choice of starting value.

TABLE C1. Exact ANOS Values Based on the Markov Chain Approach and CD Approximations to these Values for Upper One-Sided Bernoulli CUSUM Charts

	Exact ANOS	Approximations to ANOS			Exact ANOS	Approximations to ANOS		
	$p_1 = 0.025$ $h_B = 320/61$	$p_1 = 0.025$ $h_B = 320/61$			$p_1 = 0.040$ $h_B = 186/46$	$p_1 = 0.040$ $h_B = 186/46$		
p	MC	CD	$\xi(p)$	k_{min}	MC	CD	$\xi(p)$	k_{min}
0.010	29248.6	29173.9	*	*	29050.8	29150.8	*	*
0.015	2847.2	2838.2	0.19	2	3875.3	3867.3	0.50	3
0.020	951.7	947.5	−0.45	3	1201.2	1196.7	0.12	2
0.030	359.5	356.6	−1.49	3	366.6	364.1	−0.49	3
0.040	219.2	216.9	−2.37	3	202.6	200.6	−1.00	3
0.050	157.8	155.8	−3.18	3	139.0	137.3	−1.44	3
0.070	101.2	99.7	−4.69	3	85.4	84.1	−2.25	3
0.100	65.7	64.8	−6.88	3	54.2	53.2	−3.39	3
0.200	30.2	29.9	−14.60	3	25.1	24.0	−7.23	3
0.500	12.0	11.5	−45.26	3	10.0	9.1	−22.41	3

TABLE C2. Exact ANOS Values Based on the Markov Chain Approach and CD Approximations to these Values for Upper One-Sided Bernoulli CUSUM Charts

	Exact ANOS	Approximations to ANOS			Exact ANOS	Approximations to ANOS		
	$p_1 = 0.252$ $h_B = 38/6$	$p_1 = 0.252$ $h_B = 38/6$			$p_1 = 0.458$ $h_B = 16/4$	$p_1 = 0.458$ $h_B = 16/4$		
p	MC	CD	$\xi(p)$	k_{min}	MC	CD	$\xi(p)$	k_{min}
0.100	20985.0	20783.3	*	*	19547.4	19934.8	*	*
0.120	3680.0	3650.9	0.66	3	5931.3	6010.6	0.82	4
0.140	1007.2	1001.2	0.36	3	2209.0	2228.4	0.67	3
0.160	402.7	400.8	0.09	2	969.2	974.6	0.53	3
0.180	213.9	213.0	−0.17	2	487.6	489.3	0.40	3
0.200	137.0	136.5	−0.41	2	275.7	276.2	0.28	2
0.300	45.5	45.2	−1.52	2	51.3	50.7	−0.25	2
0.400	27.1	26.9	−2.55	2	24.2	23.9	−0.73	2
0.500	19.3	19.1	−3.62	2	15.6	15.4	−1.20	2
0.750	11.2	11.2	−7.50	2	8.4	8.2	−2.71	2

Specifically, let

$$\xi_k(p) = \xi_{k-1}(p) - \frac{1 - \theta_q^{\xi_{k-1}(p)} - p\left(\theta_p^{\xi_{k-1}(p)} - \theta_q^{\xi_{k-1}(p)}\right)}{r_1\theta_q^{\xi_{k-1}(p)} - \frac{\left(1-\theta_q^{\xi_{k-1}(p)}\right)\left(\theta_p^{\xi_{k-1}(p)}\log(\theta_p)+r_1\theta_q^{\xi_{k-1}(p)}\right)}{\theta_p^{\xi_{k-1}(p)}-\theta_q^{\xi_{k-1}(p)}}} \tag{C6}$$

for $k = 1, 2, \ldots$, where $\theta_p = p_1/p_0$, $\theta_q = (1 - p_1)/(1 - p_0)$, and $\xi_0(p)$ denotes a specified starting value. As the number of iterations increases, $\xi_k(p)$ rapidly approaches the value that satisfies Equation (C5) and thus, also satisfies the equivalent Equation (C4). After considering extensive computational results, the starting values $\xi_0(p) = \pm 0.1$ were found to work very well in all cases. Using $\xi_0(p) = 0.1$ in all cases considered, 2–4 iterations sufficed to approximate $\xi(p)$ so that $|\xi_k(p) - \xi_{k-1}(p)| \leq 0.009$. Once the value of $\xi(p)$ that satisfies Equation (C4) is determined, the expression in Equation (C3) (together with the approximation in Equation (6)) can be used to obtain very good approximations to the ANOS of the Bernoulli CUSUM chart.

For $p \neq p_0$, p_1, or r_1/r_2, let k_{min} denote the minimum number of iterations required to approximate $\xi(p)$ using Equation (C6) so that Equation (C3) is evaluated with at least two-decimal accuracy. Tables C1 and C2 provide exact ANOS values for one-sided Bernoulli CUSUM charts computed using the equivalent Markov Chain approaches from Appendices A and B, together with CD approximations to these ANOS values based on Equation (C3) (and the approximation in Equation (6)). Tables C1 and C2 also provide values for h_B, $\xi(p)$, and k_{min}. In particular, Table C1 includes Bernoulli CUSUM charts with $p_0 = 0.01$ and $p_1 = 0.025$ and 0.04, while Table C2 includes Bernoulli CUSUM charts with $p_0 = 0.1$ and $p_1 = 0.252$ and 0.458. Tables C1 and C2 demonstrate that the CD approximations to the ANOS of the Bernoulli CUSUM chart are very close to their exact counterparts and that their determination only requires 2–4 iterations of Equation (C6). Thus, since the CD approximation to the ANOS of the Bernoulli CUSUM chart given in Equation (C3) (together with the approximation in Equation (6)) is very good and relatively simple to compute, it will be very useful in many practical situations where the Bernoulli CUSUM will be employed.

Appendix D: A Method for Designing the One-Sided Bernoulli CUSUM Chart Based on the CD Approximation to the ANOS

In the Bernoulli CUSUM chart, it will usually be convenient to have $r_2/r_1 = m$, where m is an integer. This can usually be achieved by slightly adjusting the choice of p_1. If for a p_0 and an initial choice of p_1 the resulting r_2/r_1 is not integer-valued, then one can use the Newton-Raphson procedure to determine the value of p_1 that yields $\langle r_2/r_1 \rangle$, where $\langle r_2/r_1 \rangle$ denotes the nearest whole number to r_2/r_1. Specifically, suppose that $p_{1,0}$ denotes an initial value chosen for p_1, which (as it is likely) yields a non-integer r_2/r_1. We will refer to $p_{1,0}$ as the *nominal value* for p_1. In employing the Newton-Raphson procedure, let $m = \langle r_2/r_1 \rangle$ and

$$\begin{aligned} p_{1,k} &= p_{1,k-1} \\ &- \left[\frac{1}{\log\left(\frac{1-p_{1,k-1}}{1-p_0}\right) + \log\left(\frac{p_{1,k-1}(1-p_0)}{p_0(1-p_{1,k-1})}\right) p_{1,k-1}} \right] \\ &\times \left[\log\left(\frac{p_{1,k-1}(1-p_0)}{p_0(1-p_{1,k-1})}\right) + m \log\left(\frac{1-p_{1,k-1}}{1-p_0}\right) \right] \\ &\times \left[\log\left(\frac{1-p_{1,k-1}}{1-p_0}\right) p_{1,k-1}(1-p_{1,k-1}) \right] \end{aligned} \tag{D1}$$

for $k = 1, 2, \ldots$, where the starting value in Equation (D1) is the nominal value $p_{1,0}$. As the number of iterations increases, $p_{1,k}$ rapidly approaches the *actual value* for p_1, $p_{1,a}$ that gives $r_2/r_1 = m$. In all cases considered, 1–3 iterations sufficed to approximate $p_{1,a}$ with at least five-decimal accuracy.

Once $p_{1,a}$ is determined, the Newton-Raphson procedure can be used in Equation (7) to determine the value of h_B that gives approximately a specified in-control ANOS. In particular, let

$$\begin{aligned} h^*_{B,k} = h^*_{B,k-1} &- \left[\frac{1}{\left(1 - \exp\{h^*_{B,k-1} r_2\}\right) r_2} \right] \\ &\times \Big[\mathrm{ANOS}(p_0) |r_2 p_0 - r_1| \\ &\quad - \exp\{h^*_{B,k-1} r_2\} + h^*_{B,k-1} r_2 + 1 \Big] \end{aligned}$$

and

$$h_{B,k} = \frac{\left\langle \left(h^*_{B,k} - \epsilon(p_0)\sqrt{p_0(1-p_0)}\right) m \right\rangle}{m}$$

for $k = 1, 2, \ldots$, where $h^*_{B,0}$ denotes a specified starting value and where

$$\left\langle \left(h^*_{B,k} - \epsilon(p_0)\sqrt{p_0(1-p_0)}\right) m \right\rangle$$

denotes the nearest whole number to

$$\left(h^*_{B,k} - \epsilon(p_0)\ \sqrt{p_0(1-p_0)}\right) m.$$

Then, as the number of iterations increases, $h_{B,k}$ fairly quickly approaches the value of the control limit h_B that attains the specified in-control ANOS in the approximation in Equation (7). In most cases, the exact in-control ANOS for this value of h_B is the exact ANOS that is the closest to the specified value. After considering extensive computational results, starting values in the range $4 \le h^*_{B,0} \le 8$ were found to work well in the majority of cases.

Acknowledgments

The authors wish to express thanks to Professor Tze Leung Lai for his helpful discussions. Zachary G. Stoumbos's work was supported in part by a 1996 Rutgers Faculty of Management Research Fellowship and a 1996 Rutgers Faculty of Management Research Grant.

References

BOURKE, P. D. (1991). "Detecting a Shift in Fraction Nonconforming Using Run-Length Control Charts with 100% Inspection". *Journal of Quality Technology* 23, pp. 225–238.

BROOK, D. and EVANS, D. A. (1972). "An Approach to the Probability Distribution of CUSUM Run Length". *Biometrika* 59, pp. 539–549.

BURDEN, R. L. and FAIRES, J. D. (1993). *Numerical Analysis,* 5th ed. PWS Publishing, Boston, MA.

CHENEY, W. and KINCAID, D. (1985). *Numerical Mathematics and Computing,* 2nd ed. Brooks/Cole Publishing, Pacific Grove, CA.

GAN, F. F. (1993). "An Optimal Design of CUSUM Control Charts for Binomial Counts". *Journal of Applied Statistics* 20, pp. 445–460.

GHOSH, B. K. (1970). *Sequential Tests of Statistical Hypotheses.* Addison-Wesley/Longman Publishing, Boston, MA.

HAWKINS, D. M. (1992). "Evaluation of Average Run Lengths of Cumulative Sum Charts for an Arbitrary Data Distribution". *Communications in Statistics—Simulation and Computation* 21, pp. 1001–1020.

HAWKINS, D. M. and OLWELL, D. H. (1998). *Cumulative Sum Charts and Charting for Quality Improvement.* Springer-Verlag, New York, NY.

LUCAS, J. M. (1989). "Control Schemes for Low Count Levels". *Journal of Quality Technology* 21, pp. 199–201.

LUCAS, J. M. and CROSIER, R. B. (1982). "Fast Initial Response for CUSUM Quality Control Schemes: Give Your CUSUM a Head Start". *Technometrics* 24, pp. 199–205.

PAGE, E. S. (1954). "Continuous Inspection Schemes". *Biometrika* 41, pp. 100–114.

RENDTEL, U. (1990). "CUSUM-Schemes With Variable Sampling Intervals and Sample Sizes". *Statistical Papers* 31, pp. 103–118.

REYNOLDS, M. R., JR. (1996). "Variable Sampling Interval Control Charts with Sampling at Fixed Times". *IIE Transactions* 28, pp. 497–510.

REYNOLDS, M. R., JR.; AMIN, R. W.; and ARNOLD, J. C. (1990). "CUSUM Charts with Variable Sampling Intervals". *Technometrics* 32, pp. 371–384.

REYNOLDS, M. R., JR. and STOUMBOS, Z. G. (1998a). "Some Recent Developments in Control Charts for Monitoring a

Proportion" in *Statistical Process Monitoring and Optimization* edited by S. H. Park and G. G. Vining. Marcel Dekker, New York, NY (to appear).

REYNOLDS, M. R., JR. and STOUMBOS, Z. G. (1998b). "The SPRT Chart for Monitoring a Proportion". *IIE Transactions on Quality and Reliability* 30, pp. 545–561.

RUNGER, G. C. and WILLEMAIN, T. R. (1995). "Model-Based and Model-Free Control of Autocorrelated Processes". *Journal of Quality Technology* 27, pp. 283–292.

RYAN, T. P. and SCHWERTMAN, N. C. (1997). "Optimal Limits for Attributes Control Charts". *Journal of Quality Technology* 29, pp. 86–98.

SIEGMUND, D. (1979). "Corrected Diffusion Approximations in Certain Random Walk Problems". *Advances in Applied Probability* 11, pp. 701–719.

SIEGMUND, D. (1985). *Sequential Analysis.* Springer-Verlag, New York, NY.

STOUMBOS, Z. G. and REYNOLDS, M. R., JR. (1997a). "Control Charts Applying a Sequential Test at Fixed Sampling Intervals". *Journal of Quality Technology* 29, pp. 21–40.

STOUMBOS, Z. G. and REYNOLDS, M. R., JR. (1997b). "Corrected Diffusion Theory Approximations in Evaluating Properties of SPRT Charts for Monitoring a Process Mean". *Nonlinear Analysis* 30, pp. 3987–3996.

WALD, A. (1947). *Sequential Analysis.* Dover Publications, New York, NY.

WOODALL, W. H. (1997). "Control Charts Based on Attribute Data: Bibliography and Review". *Journal of Quality Technology* 29, pp. 172–183.

WOODALL, W. H. and ADAMS, B. M. (1993). "The Statistical Design of CUSUM Charts". *Quality Engineering* 5, pp. 559–570.

YASHCHIN, E. (1985). "On a Unified Approach to the Analysis of Two-Sided Cumulative Sum Control Schemes with Headstarts". *Advances in Applied Probability* 17, pp. 562–593.

ZACKS, S. (1981). "The Probability Distribution and the Expected Value of a Stopping Variable Associated with One-Sided CUSUM Procedures for Non-Negative Integer Valued Random Variables". *Communications in Statistics—Theory and Methods* 10, pp. 2245–2258.

Key Words: *Attributes Control Charts, Bernoulli, Cumulative Sum, p Charts, Proportion, Statistical Process Control.*