UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Chapter 1: Introduction

1.1. Notation

SPC	Statistical process control
NSPC	Nonparametric statistical process control
pmf	Probability mass function
cdf	Cumulative distribution function
pgf	Probability generating function
mgf	Moment generating function
cgf	Cumulant generating function
n	Sample size
$X_1, X_2, \ldots X_n$	Random variables in a sample
$x_1, x_2, \ldots x_n$	Observations in a sample
θ_0	Target value / Known or specified in-control location parameter[1]
CUSUM	Cumulative sum
EWMA	Exponentially weighted moving average
ARL	Average run length
ARL_0	In-control average run length
ARL_δ	Out-of-control average run length
$SDRL$	Standard deviation of the run length
MRL	Median run length
UCL	Upper control limit
CL	Center line
LCL	Lower control limit
FAR	False alarm rate
FAP	False alarm probability
VSI	Variable sampling interval
FSI	Fixed sampling interval
a_U	Upper action limit / Upper control limit
w_U	Upper warning limit
w_L	Lower warning limit
a_L	Lower action limit / Lower control limit
TPM	Transition probability matrix
A	Absorbent
NA	Non-absorbent

[1] The location parameter could be the mean, median or some percentile of the distribution. When the underlying distribution is known to be highly skewed, the median or some percentile is preferred to the mean.

1.2. Distribution of chance causes

One of the main goals of statistical process control (SPC) is to distinguish between two sources of variability, namely common cause (chance cause) variability and assignable cause (special cause) variability. Common cause variability is an inherent or natural (random) variability that is present in any repetitive process, whereas assignable cause variability is a result of factors that are not solely random. In SPC, the pattern of chance causes is usually assumed to follow some parametric distribution (such as the normal). The charting statistic and the control limits depend on this assumption and as such the properties of these control charts are "exact" only if this assumption is satisfied. However, the chance distribution is either unknown or far from being normal in many applications and consequently the performance of standard control charts is highly affected in such situations. Thus there is a need for some easy to use, flexible and robust control charts that do not require normality or any other specific parametric model assumption about the underlying chance distribution. Distribution-free or nonparametric control charts can serve this broader purpose. On this point see for example, Woodall and Montgomery (1999) and Woodall (2000). These researchers and others provide more than enough reasons for the development of nonparametric control charts.

1.3. Nonparametric or distribution-free

The term nonparametric is not intended to imply that there are no parameters involved, in fact, quite the contrary. While the term distribution-free seems to be a better description of what we expect from these charts, that is, they remain valid for a large class of distributions, nonparametric is perhaps the term more often used. In the statistics literature there is now a rather vast collection of nonparametric tests and confidence intervals and these methods have been shown to perform well compared to their normal theory counterparts. Remarkably, even when the underlying distribution is normal, the efficiency of some nonparametric methods relative to the corresponding (optimal) normal theory methods can be as high as 0.955 (see, e.g., Gibbons and Chakraborti, 2003). In fact, for some heavy-tailed distributions like the double exponential, nonparametric tests can be more efficient. It may be argued that nonparametric methods will be "less efficient" than their parametric counterparts when one has a complete knowledge of the process distribution for which that parametric method was specifically designed. However, the reality is that such information is seldom, if ever, available in practice. Thus it seems natural to develop and use nonparametric methods in SPC and the quality practitioners will be well advised to have these techniques in their toolkits.

We only discuss univariate nonparametric control charts designed to track the location of a continuous process since very few charts are available for monitoring the scale and simultaneously monitoring the location and scale of a process. The field of multivariate control charts is interesting and the body of literature on nonparametric multivariate control charts is growing. However, in our opinion, it hasn't yet reached a critical mass and a discussion on this topic is better postponed for the future.

1.4. Nonparametric control charts

Chakraborti, Van der Laan and Bakir (2001) (hereafter CVB) provided a systematic and thorough account of the nonparametric control chart literature. A nonparametric control chart is defined in terms of its in-control run length distribution. If the in-control run length distribution of a control chart is the same for every continuous distribution, the chart is called nonparametric or distribution-free. CVB summarized the advantages of nonparametric control charts as follows: (i) simplicity, (ii) no need to assume a particular parametric distribution for the underlying process, (iii) the in-control run length distribution is the same for all continuous distributions, (iv) more robust and outlier resistant, (v) more efficiency in detecting changes when the true distribution is markedly non-normal, particularly with heavier tails, and (vi) no need to estimate the variance to set up charts for the location parameter. It is emphasized that from a technical point of view most nonparametric procedures require the population to be continuous in order to be distribution-free and thus in a SPC context we consider the so-called "variables control charts." Some disadvantages of nonparametric control charts are as follows: (i) they will be "less efficient" than their parametric counterparts when one has a complete knowledge of the process distribution for which that parametric method was specifically designed, (ii) one usually requires special tables when the sample sizes are small, and (iii) nonparametric methods are not well-known amongst all researchers and quality practitioners.

1.5. Terminology and formulation

Two important problems in usual SPC are monitoring the process mean and/or the process standard deviation. In the nonparametric setting, we consider, more generally, monitoring the center or the location (or a shift) parameter and/or a scale parameter of a process. The location parameter represents a typical value and could be the mean or the median or some percentile of the distribution; the latter two are especially attractive when the underlying distribution is expected to be skewed. Also in the nonparametric setting, the processes are implicitly assumed to follow (i) a location model, with a cdf $F(x-\theta)$, where θ is the location parameter or (ii) a scale model, with a cdf $F\left(\frac{x}{\tau}\right)$, where $\tau(>0)$ is the scale parameter. Even more generally, one might consider (iii) the location-scale model with cdf $F\left(\frac{x-\theta}{\tau}\right)$, where θ and τ are the location and the scale parameter, respectively. Under these model assumptions, the problem is to track θ and τ (or both), based on random samples or subgroups taken (usually) at equally spaced time points. In the usual (parametric) control charting problems F is assumed to be the cdf Φ of the standard normal distribution whereas in the nonparametric setting, for variables data, F is some unknown continuous cdf. Although the location-scale model seems to be a natural model to consider paralleling the normal theory case with mean and variance both unknown, most of what is currently available in the nonparametric statistical process control (NSPC) literature deals mainly with the location model.

As we noted earlier, a comprehensive survey of the literature until about 2000 can be found in CVB. Here, we mention some of the key contributions and ideas and a few of the more recent developments in the area; the literature on nonparametric methods continues to grow at a rapid pace. In fact, Woodall and Montgomery (1999) stated: 'There would appear to be an increasing role for nonparametric methods, particularly as data availability increases'. Most nonparametric charts, however, have been developed for Phase II applications. There are generally two phases in SPC. In Phase I (also called the retrospective phase), typically, preliminary analysis is done which includes planning, administration, data collection, data management, exploratory work including graphical and numerical analysis, goodness-of-fit analysis etc. to ensure that the process is in-control. This means that the process is managed to operate at or near some acceptable target value along with some natural variation and no special causes of concern are expected to be present. Once this is ascertained, SPC moves to the next phase, Phase II, (or the prospective phase), where the control limits and/or the estimators obtained in Phase I are used for process monitoring based on new samples of data. When the underlying parameters of the process distribution are known or specified, this is referred to as the "standard(s) known" case and is denoted case K. In contrast, if the parameters are unknown and need to be estimated, it is typically done in Phase I, with in-control data. This situation is referred to as the "standard(s) unknown" case and is denoted case U. In this text we are going to consider decision problems under both Phase I and Phase II. One of the main differences between the two phases is the fact that the *FAR* (or in-control average run length ARL_0) is typically used to construct and evaluate Phase II control charts, whereas the false alarm probability (*FAP*) is used to construct and evaluate Phase I control charts. The *FAP* is the probability of at least one false alarm out of many comparisons, whereas the *FAR* is the probability of a single false alarm involving only a single comparison. Various authors have studied the Phase I problem; see for example King (1954), Chou and Champ (1995), Sullivan and Woodall (1996), Jones and Champ (2002), Champ and Chou (2003), Champ and Jones (2004), Koning (2006) and Human, Chakraborti and Smit (2007). Since not much is typically known or can be assumed about the underlying process distribution in a Phase I setting, nonparametric Phase I control charts are of great use.

There are three main classes of control charts: the Shewhart chart, the cumulative sum (CUSUM) chart and the exponentially weighted moving average (EWMA) chart and their refinements. Relative advantages and disadvantages of these charts are well documented in the literature (see, e.g., Montgomery, 2001). Analogs of these charts have been considered in the nonparametric setting. We describe some of the charts under each of the three classes.

1.6. Shewhart-type charts

Shewhart-type charts are the most popular charts in practice because of their simplicity, ease of application, and the fact that these versatile charts are quite efficient in detecting moderate to large shifts. Both one-sided and two-sided charts are considered. The one-sided charts are more useful when only a directional shift (higher or lower) in the median is of interest. The two-sided charts, on the other hand, are typically used to detect a shift or change in the median in any direction.

1.7. CUSUM-type charts

While the Shewhart-type charts are widely known and most often used in practice because of their simplicity and global performance, other classes of charts, such as the CUSUM charts are useful and sometimes more naturally appropriate in the process control environment in view of the sequential nature of data collection. These charts, typically based on the cumulative totals of a plotting statistic, obtained as data accumulate, are known to be more efficient for detecting certain types of shifts in the process. The normal theory CUSUM chart for the mean is typically based on the cumulative sum of the deviations of the individual observations (or the subgroup means) from the specified target mean. It seems natural to consider analogs of these charts using the nonparametric plotting statistics discussed earlier. These lead to nonparametric CUSUM (NPCUSUM) charts.

1.8. EWMA-type charts

Another popular class of control charts is the exponentially weighted moving average (EWMA) charts. The EWMA charts also take advantage of the sequentially (time ordered) accumulating nature of the data arising in a typical SPC environment and are known to be efficient in detecting smaller shifts but are easier to implement than the CUSUM charts.

Section A: Monitoring the location of a process when the target location is specified (Case K)

Chapter 2: Sign control charts

2.1. The Shewhart-type control chart

2.1.1. Introduction

The sign test is one of the simplest and broadly applicable nonparametric tests (see, e.g., Gibbons and Chakraborti, 2003) that can be used to test hypotheses (or find confidence intervals) for the median (or any specified percentile) of a continuous distribution. In this thesis, we will only consider the 50th percentile, i.e. the median. The fact that the sign test is applicable for any continuous population is an advantage to quality practitioners. Suppose that the median of a continuous process needs to be maintained at a specified value θ_0. Amin et al. (1995) presented Shewhart-type nonparametric charts for this problem using what are called "within group sign" statistics. This is called a sign chart (also referred to as the SN chart).

2.1.2. Definition of the sign test statistic

Let $X_{i1}, X_{i2},..., X_{in}$ denote the i^{th} $(i = 1,2,...)$ sample or subgroup of independent observations of size $n > 1$ from a process with an unknown continuous distribution function F. Let θ_0 denote the known or specified value of the median when the process is in-control, then θ_0 is called the target value. Compare each x_{ij} $(j = 1,2,..., n)$ with θ_0. Record the difference between θ_0 and each x_{ij} by subtracting θ_0 from x_{ij}. There will be n such differences, $x_{ij} - \theta_0$ $(j = 1,2,..., n)$, in the i^{th} sample. Let n^+ denote the number of observations with values greater than θ_0 in the i^{th} sample. Let n^- denote the number of observations with values less than θ_0 in the i^{th} sample. Provided there are no ties we have that $n^+ + n^- = n$.

Define

$$SN_i = \sum_{j=1}^{n} sign(x_{ij} - \theta_0) \tag{2.1}$$

where $sign(x)$ = -1, 0, 1 if $x < 0$, $= 0$, > 0.

Then SN_i is the difference between n^+ and n^- in the i^{th} sample, i.e. SN_i is the difference between the number of observations with values greater than θ_0 and the number of observations with values less than θ_0 in the i^{th} sample.

Define

$$T_i = \frac{SN_i + n}{2}, \tag{2.2}$$

assuming there are no ties within a subgroup. The random variable T_i is the number of sample observations greater than or equal to θ_0 in the i^{th} sample. In (2.2) the statistic T_i is expressed in terms of the sign test statistic SN_i. Using the relationship in (2.2), the sign test statistic SN_i can be expressed in terms of the statistic T_i (if there are no ties within a subgroup) and we obtain

$$SN_i = 2T_i - n. \tag{2.3}$$

This relationship is evident from the fact that

$$SN_i = \sum_{j=1}^{n} sign(x_{ij} - \theta_0) = \sum_{j=1}^{n} \left(2\psi(x_{ij} - \theta_0) - 1\right) = 2T_i - n$$

where $\psi(x) = 0$, 1 if $x \leq 0$, > 0.

In the literature the statistic T_i is also well-known under the name *sign test statistic* (see, for example, Gibbons and Chakraborti (2003)). For the purpose of this study, SN_i will be referred to as the sign test statistic.

Zero differences

For a continuous random variable, X, the probability of any particular value is zero; thus, $P(X = a) = 0$ for any a. Since the distribution of the observations is assumed to be continuous, $P(X_{ij} - \theta_0 = 0) = 0$. Theoretically, the case where $sign(x_{ij} - \theta_0) = 0$ should occur with zero probability, but in practice zero differences do occur as a result of, for example, truncation or rounding of the observed values. A common practice in such cases is to discard all the observations leading to zero differences and to redefine n as the number of nonzero differences.

2.1.3. Plotting statistic

Sign control charts are based on the well-known sign test. A control chart is a graph consisting of values of a plotting (or charting) statistic and the associated control limits. The plotting statistic for the sign chart is $SN_i = \sum_{j=1}^{n} sign(x_{ij} - \theta_0)$ for $i = 1,2,3,...$.

Distributional properties of the charting statistic

The random variable T_i has a binomial distribution with parameters n and $p = P(X_{ij} \geq \theta_0)$, i.e. $T_i \sim BIN(n, p)$. Hence, we can find the distribution of SN_i via the relationship given in (2.3).

Table 2.1. Moments and the probability mass function of the T_i and SN_i statistics, respectively.

	T_i	SN_i
Expected value	$E(T_i) = np$	$E(SN_i)$ $= E(2T_i - n)$ $= n(2p-1)$
Variance	$\text{var}(T_i) = np(1-p)$	$\text{var}(SN_i)$ $= \text{var}(2T_i - n)$ $= 4np(1-p)$
Standard deviation	$stdev(T_i) = \sqrt{np(1-p)}$	$stdev(SN_i)$ $= 2\sqrt{np(1-p)}$
Probability mass function (pmf)	$f(t) = P(T_i = t) = \binom{n}{t} p^t (1-p)^{n-t}$	$f(s)$ $= P(SN_i = s)$ $= P(2T_i - n = s)$ $= P\left(T_i = \frac{n+s}{2}\right)$

The probability distributions of T_i and SN_i are both **symmetric**[*] as long as the median remains at θ_0. In this case:

- the probability distributions are referred to as the *in-control* probability distributions;
- $p = P(X_{ij} \geq \theta_0) = 0.5$; and
- since the in-control distribution of the plotting statistic SN_i is symmetric, the control limits will be equal distances away from 0.

Figure 2.1 illustrates for $n = 10$ that the in-control probability distributions of T_i and SN_i are symmetric about their means, that is, T_i is symmetric around its mean given by $n \times p = 10 \times 0.5 = 5$ and SN_i is symmetric around its mean given by $n(2p-1) = 10(2 \times 0.5 - 1) = 0$.

[*] T_i and SN_i are symmetric about np and zero, respectively, as long as the median remains at θ_0.

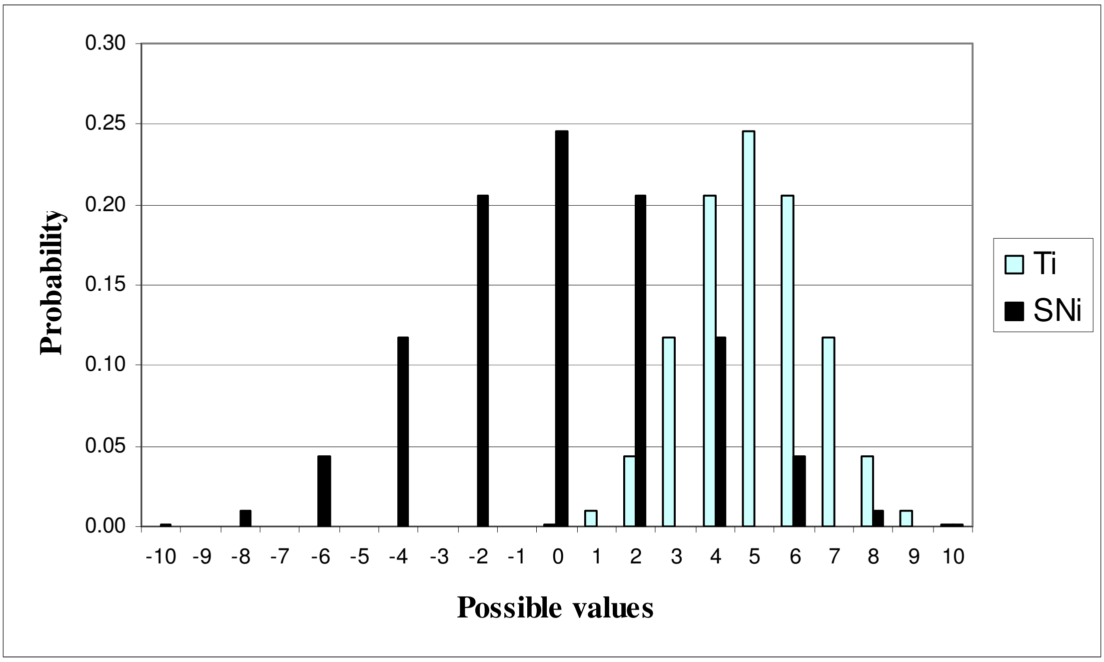

Figure 2.1. The in-control probability distribution of T_i and SN_i for $n = 10$.

2.1.4. Determination of control limits

In order to find the control limits and study chart performance, the distribution of SN_i is necessary; this can be most easily obtained using the relationship $SN_i = 2T_i - n$. Since the in-control distribution of T_i is binomial with parameters n and 0.5, it follows that the in-control distribution of SN_i is symmetric about 0 and hence the control limits and the center line of the two-sided nonparametric Shewhart-type sign chart (for the median) are given by

$$\begin{aligned} UCL &= c \\ CL &= 0 \\ LCL &= -c \end{aligned} \tag{2.4}$$

where $c \in \{1,2,...,n\}$.

If the plotting statistic SN_i falls between the control limits, that is, $-c < SN_i < c$, the process is declared to be in-control, whereas if the plotting statistic SN_i falls on or outside one of the control limits, that is, if $SN_i \le -c$ or $SN_i \ge c$, the process is declared to be out-of-control. In the latter case corrective action and a search for assignable causes is necessary.

Take note that T_i can also be calculated and plotted against the control limits. This is done by assuming that the LCL is equal to some constant a and that the UCL is equal to some constant b, i.e. the control limits are given by: $UCL = b$ and $LCL = a$. Since the in-control probability distribution of T_i is symmetric when working with the median, that is, $P(T_i = a) = P(T_i = n - a)$, a sensible choice for b is therefore $n - a$.

The control limits and the center line of the nonparametric Shewhart chart (for the median) using T_i as the plotting statistic are given by

$$\begin{aligned} UCL &= n - a \\ CL &= np \\ LCL &= a \end{aligned}$$

where a denotes a positive integer which is selected such that $LCL < UCL$.

Although both T_i and SN_i can be calculated for each sample and be compared to the control limits, the statistic SN_i has the advantage of keeping the control limits symmetric around zero. Therefore, the plotting statistic SN_i is calculated and used as the plotting statistic. The terms ‘plotting statistic’ and ‘charting statistic’ will be used interchangeably throughout this text.

The question arises: When using SN_i as the plotting statistic, what should the values of the control limits be set equal to? In other words, what is the value of the charting constant c? Specifying control limits is one of the critical decisions that must be made in designing a control chart. By moving the control limits farther away from the center line, we decrease the risk of a type I error – that is, the risk of a point falling beyond the control limits, indicating an out-of-control condition when no assignable cause is present. However, widening the control limits will also increase the risk of a type II error – that is, the risk of a point falling between the control limits when the process is really out-of-control. If we move the control limits closer to the center line, the opposite effect is obtained: The risk of type I error is increased, while the risk of type II error is decreased. Consequently, the control limits are chosen such that if the process is in-control, nearly all of the sample points will fall between them. In other words, the charting constant c is typically obtained for a specified in-control average run length, which, in case K, is equal to the reciprocal of the nominal FAR, α. Thus, using the symmetry of the binomial distribution, c is the smallest integer such that $P_{\theta_0}(SN_i \geq c) \leq \frac{\alpha}{2}$. For example, using Table G of Gibbons and Chakraborti (2003) we give some $P_{\theta_0}(T \geq t)$ values in Table 2.2 that may be considered "small" in a SPC context. The charting constant c is obtained using $c = 2t - n$ (recall that the sign test statistic SN_i is expressed in terms of the statistic T_i by the relationship $SN_i = 2T_i - n$). The false alarm rate is obtained by adding the probability in the left tail, $P_{\theta_0}(T \leq n-t)$, and the probability in the right tail, $P_{\theta_0}(T \geq t)$, i.e. $FAR = P_{\theta_0}(T \leq n-t) + P_{\theta_0}(T \geq t)$. Since the probability distribution of T_i is symmetric (as long as the median remains at θ_0), the FAR is also obtained using $FAR = 2P_{\theta_0}(T \geq t)$. For example, for $n = 5$ we get $t = 5$ and thus $c = 5$ for a FAR of $2(0.0312) = 0.0624$ and this is the lowest FAR achievable. However for $n = 10$ the FAR drops to 0.0020 if $c = 10$. It should be noted that the lowest attainable FAR is always obtained when $n = t$.

Table 2.2. FAR and ARL_0 of a sign control chart for various values of $n = t$.

n	5	6	7	8	9	10
$P_{\theta_0}(T \geq t)$	0.0312	0.0156	0.0078	0.0039	0.0020	0.0010
FAR (α)	0.0624	0.0312	0.0156	0.0078	0.0040	0.0020
ARL_0	16.00	32.00	64.00	128.00	256.00	512.00

Looking at the attainable FAR and ARL_0 values shown in Table 2.2, we see that unless the sample size is at least 10, the sign chart would be somewhat unattractive (from an operational point of view) in SPC applications, where one often stipulates a large in-control average run length, as large as 370 or 500, and a small FAR, as small as 0.0027. If, for example, the FAR is too 'large', which is the case for 'small' sample sizes, many false alarms will be expected by this chart leading to a possible loss of time and resources. Then again, the sign chart is the simplest of nonparametric charts that works under minimal assumptions. In fact, from the hypothesis testing literature, it is known that the sign test (and so the chart) is more robust and efficient when the chance distribution is symmetric like the normal but with heavier tails such as the double exponential.

Example 2.1

A Shewhart-type sign chart for the Montgomery (2001) piston ring data

We illustrate the Shewhart-type sign chart using a set of data from Montgomery (2001; Tables 5.1 and 5.2) on the inside diameters of piston rings manufactured by a forging process. A part of this data, fifteen prospective samples (Table 5.2) each of five observations, is used here. The rest of the data (Table 5.1) will be used later. We assume that the underlying distribution is symmetric with a known median $\theta_0 = 74\,mm$. From Table G (see Gibbons and Chakraborti (2003)) we obtain $t = 5$ (when $n = 5$) for an achieved false alarm rate of $2(0.0312) = 0.0624$. Therefore, $c = 2 \times 5 - 5 = 5$ and the control limits and the center line of the nonparametric Shewhart sign chart are given by $UCL = 5$, $CL = 0$ and $LCL = -5$.

Panel *a* of Table 2.3 displays the sample number. The two rows of each cell in panel *b* shows the individual observations and $sign(x_{ij} - \theta_0)$ values, respectively. The SN_i and T_i values are shown in panel *c* and panel *d*, respectively.

As an example, the calculation of SN_1 (found in Table 2.3) is given.

$$\begin{aligned} SN_1 &= sign(x_{11} - \theta_0) + sign(x_{12} - \theta_0) + sign(x_{13} - \theta_0) + sign(x_{14} - \theta_0) + sign(x_{15} - \theta_0) \\ &= sign(74.012 - 74) + sign(74.015 - 74) + sign(74.030 - 74) + sign(73.986 - 74) + sign(74 - 74) \\ &= sign(0.012) + sign(0.015) + sign(0.03) + sign(-0.014) + sign(0) \\ &= 1 + 1 + 1 - 1 + 0 \\ &= 2. \end{aligned}$$

Table 2.3. Data and calculations for the Shewhart sign chart[*].

Panel *a*	Panel *b*					Panel *c*	Panel *d*
Sample number	**Individual observations** $sign(x_{ij}-\theta_0)$					SN_i	T_i
1	74.012[†] 1	74.015 1	74.030 1	73.986 -1	74.000 0	2	3
2	73.995 -1	74.010 1	73.990 -1	74.015 1	74.001 1	1	3
3	73.987 -1	73.999 -1	73.985 -1	74.000 0	73.990 -1	-4	0
4	74.008 1	74.010 1	74.003 1	73.991 -1	74.006 1	3	4
5	74.003 1	74.000 0	74.001 1	73.986 -1	73.997 -1	0	2
6	73.994 -1	74.003 1	74.015 1	74.020 1	74.004 1	3	4
7	74.008 1	74.002 1	74.018 1	73.995 -1	74.005 1	3	4
8	74.001 1	74.004 1	73.990 -1	73.996 -1	73.998 -1	-1	2
9	74.015 1	74.000 0	74.016 1	74.025 1	74.000 0	3	3
10	74.030 1	74.005 1	74.000 0	74.016 1	74.012 1	4	4
11	74.001 1	73.990 -1	73.995 -1	74.010 1	74.024 1	1	3
12	74.015 1	74.020 1	74.024 1	74.005 1	74.019 1	5	5
13	74.035 1	74.010 1	74.012 1	74.015 1	74.026 1	5	5
14	74.017 1	74.013 1	74.036 1	74.025 1	74.026 1	5	5
15	74.010 1	74.005 1	74.029 1	74.000 0	74.020 1	4	4

The sign chart is shown in Figure 2.2 with $UCL = 5$, $CL = 0$ and $LCL = -5$.

[*] See SAS Program 1 in Appendix B for the calculation of the values in Table 2.3.
[†] The two rows of each cell in panel *b* shows the x_{ij} and $sign(x_{ij}-\theta_0)$ values, respectively, for example,

$x_{11} = 74.012$ $sign(x_{11}-\theta_0) = 1$	is presented as	74.012 1

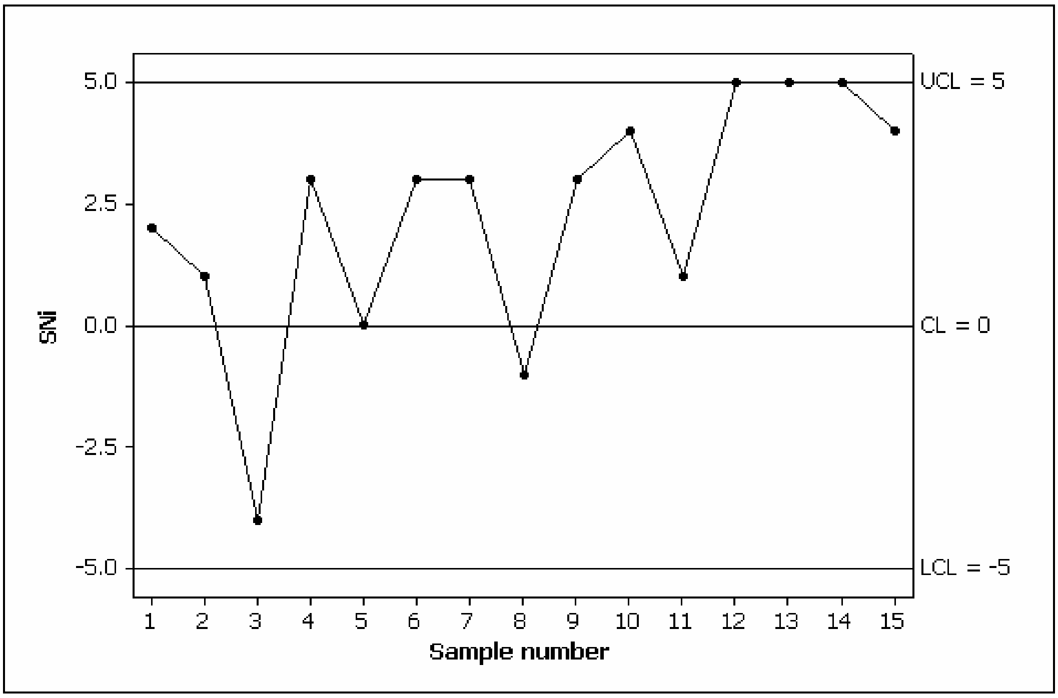

Figure 2.2. Shewhart-type sign control chart for Montgomery (2001) piston ring data.

Observations 12, 13 and 14 lie on the upper control limit which indicates that the process is out-of-control starting at sample 12. It appears most likely that the process median has shifted upwards from the target value of 74 *mm*. Corrective action and a search for assignable causes is necessary.

Control charts are often compared on the basis of various characteristics of the run length distribution, such as the ARL. One prefers a "large" in-control average run length (denoted ARL_0) and a "small" out-of-control ARL (denoted ARL_δ) under a shift. Amin et al. (1995) compared the ARL of the classical Shewhart $\overline{X}$ chart and the Shewhart-type sign chart for various shift sizes and underlying distributions. One practical advantage of sign charts, and of all nonparametric charts (if, of course, their assumptions are satisfied), is that the FAR (and the ARL_0) remains the same (eg. $FAR = 0.0624$ and $ARL_0 = 16$ for $n = 5$) for all continuous distributions. This is so because the in-control run length distribution is the same for every continuous distribution, for nonparametric charts, by definition. This does not hold for parametric charts (except for EWMA charts), and, as a result, parametric charts (again, with the EWMA chart being the exception) do not enjoy the same kind of robustness properties as nonparametric charts do. It should be noted that the EWMA control chart can be designed so that it is robust to the normality assumption. On this point, Borror, Montgomery and Runger (1999) showed that the ARL_0 of the EWMA chart is reasonably close to the normal-theory value for both skewed and heavy-tailed symmetric non-normal distributions.

2.1.5. Run length distribution

The number of subgroups or samples that need to be collected (or, equivalently, the number of plotting statistics that must be plotted) before the next out-of-control signal is given by a chart is called the run length. The run length is a random variable denoted by N. A popular measure of chart performance is the 'expected value' or the 'mean' of the run length distribution, called the average run length (ARL). Various researchers, see for example, Barnard (1959) and Chakraborti (2007), have suggested using other characteristics for assessment of chart performance, for example, the standard deviation of the run length distribution ($SDRL$), the median run length (MRL) and/or other percentiles of the run length distribution. This recommendation is warranted seeing as (i) the run-length can only take on positive integer values by definition, (ii) the shape of this distribution is significantly right-skewed and (iii) it's known that in a right-skewed distribution the mean is greater than the median and thus is usually not a fair representation of a typical observation or the center.

Since the observations plotted on the control chart are assumed to be independent, the number of points that must be plotted until the first plotted point plots on or exceeds a control limit is a geometric random variable with parameter p, where p denotes the probability of a success (or, equivalently, the probability of a signal). Therefore, $N \sim GEO(P(\text{Signal}))$ where $P(\text{Signal}) = p$. The well-known properties of the geometric distribution are given in panel *a* of Table 2.4 and we use the fact that if q denotes the probability of no signal then $P(\text{Signal}) + P(\text{No Signal}) = p + q = 1$, i.e. $q = 1 - p$. The properties of the run length N are derived using the well-known properties of the geometric distribution and they are displayed in panel *b* of Table 2.4.

Table 2.4. The properties of the geometric and run length distribution.

	a	***b***
	$X \sim GEO(p)$	$N \sim GEO(P(Signal))$
Expected value	$E(X)=\frac{1}{p}$	$E(N)=ARL=\frac{1}{P(\text{Signal})}$
Variance	$\text{var}(X)=\frac{q}{p^2}$	$\text{var}(N)=\frac{1-P(\text{Signal})}{(P(\text{Signal}))^2}$
Standard deviation	$stdev(X)=\frac{\sqrt{q}}{p}$	$SDRL=\frac{\sqrt{1-P(\text{Signal})}}{P(\text{Signal})}$
Probability mass function (pmf)	$f(x)=P(X=x)=p(1-p)^{x-1}$ for $x=1,2,3,...$	$P(N=a)=P(\text{Signal})(1-P(\text{Signal}))^{a-1}$ for $a=1,2,3,...$
Cumulative distribution function (cdf)	$F(x)=P(X\le x)=1-(1-p)^x$ for $x=1,2,3,...$	$P(N\le a)=1-(1-P(\text{Signal}))^a$ for $a=1,2,3,...$

The $100\rho^{th}$ $(0<\rho<1)$ percentile is defined as the smallest l such that the cdf, given by $P(N\le l)=1-(1-P(\text{Signal})^l$ for $l=1,2,.....$, at the integer l is at least $(100\times\rho)\%$, that is,

$$l=\min\{j:1-(1-P(\text{Signal}))^j\ge\rho\} \text{ for } j=1,2,... \tag{2.5}$$

which reduces to finding the smallest positive integer l such that

$$l\ge\frac{\ln(1-\rho)}{\ln(1-P(\text{Signal}))}. \tag{2.6}$$

The run length distribution can be described via percentiles, for example, using the 5^{th}, 25^{th} (the first quartile, Q_1), 50^{th} (the median run length, MRL), 75^{th} (the third quartile, Q_3) and the 95^{th} percentiles by substituting ρ in expression (2.6) by 0.05, 0.25, 0.50, 0.75 and 0.95, respectively.

2.1.6. One-sided control charts

A lower one-sided chart will have a LCL equal to some constant value with no UCL, whereas an upper one-sided chart will have an UCL equal to some constant value with no LCL. One-sided control charts are particularly useful in situations where only an upward (or only a downward) shift in a particular process parameter is of interest. For example, we might be monitoring the breaking strength of material used to make parachutes. If the breaking strength of the material decreases it might tear at a critical time, whereas if the breaking strength of the material increases it is beneficial to the user, since the material would, most likely, not tear while being used. In such a scenario a lower one-sided chart will be sufficient, since we are only interested in detecting a downward shift in a process parameter.

For the sign control chart, if we are only interested in detecting a downward shift we will use a lower one-sided sign control chart with $LCL = -c$ and no upper control limit. Consequently, if the plotting statistic SN_i falls on or below the LCL the process is declared to be out-of-control. On the other hand, if we are only interested in detecting an upward shift we will use an upper one-sided sign control chart with $UCL = c$ and no lower control limit. Consequently, if the plotting statistic SN_i falls on or above the UCL the process is declared to be out-of-control.

2.1.6.1. Lower one-sided control charts

Result 2.1: Probability of a signal

The probability that the control chart signals, that is, the probability that the plotting statistic SN_i is smaller than or equal to the lower control limit, can be expressed in terms of $p = P(X_{ij} \geq \theta_0)$, the sample size n and the constant c. Let $P^L(\text{Signal})$ denote the probability of a signal, where superscript L refers to the lower one-sided chart. The probability of a signal is then given by

$$P^L(\text{Signal}) = P^L(SN_i \le LCL) = P^L(SN_i \le -c) = P^L(2T_i - n \le -c) = P^L\left(T_i \le \frac{n-c}{2}\right). \quad (2.7)$$

Note that (2.7) can be solved by using the cdf of a Binomial distribution.

The probabilities, $P^L(\text{Signal})$'s, were computed using Mathcad (see Mathcad Program 2 in Appendix B). In doing so, we kept in mind that we ultimately wanted to use the T_i statistic in the calculation of $P^L(\text{Signal})$, because then we could use the cdf of a Binomial distribution to find $P^L(\text{Signal})$. Therefore, the probability of a signal for the lower one-sided sign chart was computed using

$$P^L(\text{Signal}) = P^L(T_i \le LCL) = P^L(T_i \le a) = \sum_{i=0}^{a} \binom{n}{i} p^i (1-p)^{n-i}. \quad (2.8)$$

The results are given in Tables 2.5, 2.6 and 2.7 for $n = 5$, $n = 10$ and $n = 15$, respectively, for $p = 0.1(0.1)0.9$ and $a = 0(1)n$. The shaded column ($p = 0.5$) contains the value of the *in-control* average run length (ARL_0) and the false alarm rate (FAR), whereas the rest of the columns ($p \ne 0.5$) contain the values of the *out-of-control* average run length (ARL_δ) and the probability of a signal (when the process is considered to be out-of-control).

Result 2.2: Average run length

Since the run length has a geometric distribution (recall that $N \sim GEO(P^L(\text{Signal}))$ the expected value of this specific geometric distribution will be equal to $\frac{1}{P^L(\text{Signal})}$. The ARL is the mean of the run length distribution. Therefore, we have that

$$ARL^L = E(N) = \frac{1}{P^L(\text{Signal})}. \quad (2.9)$$

Result 2.3: Standard deviation of the run length

Since the run length has a geometric distribution (see Result 2.2) the standard deviation will be equal to $\frac{\sqrt{1-P^L(\text{Signal})}}{P^L(\text{Signal})}$. Therefore, we have that

$$SDRL^L(N) = \frac{\sqrt{1-P^L(\text{Signal})}}{P^L(\text{Signal})}. \qquad (2.10)$$

Example 2.2

For a sample size of 10 ($n = 10$), $p = 0.5$ and $a = 2$, we can calculate the probability of a signal and the average run length using (2.8) and (2.9), respectively, and we obtain $P^L(\text{Signal}) =$ $\sum_{i=0}^{2}\binom{10}{i}(0.5)^i(1-0.5)^{10-i} = 0.055$ and $ARL^L = \frac{1}{0.055} = 18.29$.

2.1.6.2. Upper one-sided control charts

Result 2.4: Probability of a signal

The probability that the control chart signals, that is, the probability that the plotting statistic SN_i is greater than or equal to the upper control limit, can be expressed in terms of $p = P(X_{ij} \geq \theta_0)$, the sample size n and the constant c. Let $P^U(\text{Signal})$ denote the probability of a signal, where superscript U refers to the upper one-sided chart. The probability of a signal is then given by

$$P^U(\text{Signal}) = P^U(SN_i \geq UCL) = P^U(SN_i \geq c) = P^U\left(T_i \geq \frac{n+c}{2}\right) = 1 - P^U\left(T_i \leq \frac{n+c}{2} - 1\right). \qquad (2.11)$$

Note that (2.11) can be solved by using the cdf of a Binomial distribution.

The probabilities, $P^U(\text{Signal})$'s, were computed using Mathcad (see Mathcad Program 1 in Appendix B). In doing so, we kept in mind that we ultimately wanted to use the T_i statistic in the calculation of $P^U(\text{Signal})$, because then we could use the cdf of a Binomial distribution to find $P^U(\text{Signal})$. Therefore, the probability of a signal for the upper one-sided sign chart was computed using

$$P^U(\text{Signal}) = P^U(T_i \geq UCL) = P^U(T_i \geq n-a) = \sum_{i=n-a}^{n} \binom{n}{i} p^i (1-p)^{n-i}. \qquad (2.12)$$

The results are given in Tables 2.5, 2.6 and 2.7 for $n=5$, $n=10$ and $n=15$, respectively, for $p = 0.1(0.1)0.9$ and $a = 0(1)n$. The shaded column ($p = 0.5$) contains the value of the *in-control* average run length (ARL_0) and the false alarm rate (FAR), whereas the rest of the columns ($p \neq 0.5$) contain the values of the *out-of-control* average run length (ARL_δ) and the probability of a signal (when the process is considered to be out-of-control).

Result 2.5: Average run length

Since the run length has a geometric distribution (recall that $N \sim GEO(P^U(\text{Signal}))$ the expected value of this specific geometric distribution will be equal to $\dfrac{1}{P^U(\text{Signal})}$. The ARL is the mean of the run length distribution. Therefore, we have that

$$ARL^U = E(N) = \frac{1}{P^U(\text{Signal})}. \qquad (2.13)$$

Result 2.6: Standard deviation of the run length

Since the run length has a geometric distribution (see Result 2.5) the standard deviation will be equal to $\dfrac{\sqrt{1-P^U(\text{Signal})}}{P^U(\text{Signal})}$. Therefore, we have that

$$SDRL^U(N) = \frac{\sqrt{1-P^U(\text{Signal})}}{P^U(\text{Signal})}. \qquad (2.14)$$

Example 2.3

For a sample size of 10 ($n = 10$), $p = 0.5$ and $a = 2$, we can calculate the probability of a signal and the average run length using (2.12) and (2.13), respectively, and we obtain $P^U\,(\text{Signal}) = \sum_{i=10-2}^{10} \binom{10}{i} (0.5)^i (1-0.5)^{10-i} = 0.055$ and $ARL^U = \frac{1}{0.055} = 18.29$.

Application

The average run length values for the lower and upper one-sided Shewhart sign charts are calculated by evaluating expressions (2.8) and (2.9) for the lower one-sided chart and expressions (2.12) and (2.13) for the upper one-sided chart using $n = 5, 10$ and 15, respectively. These values are shown in Table 2.5, Table 2.6 and Table 2.7, respectively.[*] As mentioned previously, the shaded column ($p = 0.5$) contains the value of the *in-control* average run length (ARL_0) and the false alarm rate (FAR), whereas the rest of the columns ($p \neq 0.5$) contain the values of the *out-of-control* average run length (ARL_δ) and the probability of a signal (when the process is considered to be out-of-control).

[*] Table 2.5, Table 2.6 and Table 2.7 should preferably be studied together.

Table 2.5. The average run length for the one-sided Shewhart sign chart with $n = 5$.[**]

ARL **P(Signal)**	p^U	**0.9**	**0.8**	**0.7**	**0.6**	**0.5**	**0.4**	**0.3**	**0.2**	**0.1**
	p^L	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
a	0	1.69 0.590	3.05 0.328	5.95 0.168	12.86 0.078	32.00 0.031	97.66 0.010	411.52 0.002	3125.00 0.000	100000.00 0.000
	1	1.09 0.919	1.36 0.737	1.89 0.528	2.97 0.337	5.33 0.188	11.49 0.087	32.49 0.031	148.81 0.007	2173.91 0.000
	2	1.01 0.991	1.06 0.942	1.19 0.837	1.47 0.683	2.00 0.500	3.15 0.317	6.13 0.163	17.27 0.058	116.82 0.009
	3	1.00 1.000	1.01 0.993	1.03 0.969	1.10 0.913	1.23 0.813	1.51 0.663	2.12 0.472	3.81 0.263	12.28 0.081
	4	1.00 1.000	1.00 1.000	1.00 0.998	1.01 0.990	1.03 0.969	1.08 0.922	1.20 0.832	1.49 0.672	2.44 0.410
	5	1.00 1.000	1.00 1.000	1.00 1.000	1.00 1.000	1.00 1.000	1.00 1.000	1.00 1.000	1.00 1.000	1.00 1.000

[**] See Mathcad Programs 1 and 2 in Appendix B for the calculation of the values in Table 2.5.

Table 2.6. The average run length for the one-sided Shewhart sign chart with $n = 10$.[††]

ARL **P(Signal)**	p^U	**0.9**	**0.8**	**0.7**	**0.6**	**0.5**	**0.4**	**0.3**	**0.2**	**0.1**
	p^L	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
	0	2.87 0.349	9.31 0.107	35.40 0.028	165.38 0.006	1024.00 0.001	9536.74 0.000	169350.88 0.000	9765625.00 0.000	10000000000.00 0.000
	1	1.36 0.736	2.66 0.376	6.70 0.149	21.57 0.046	93.09 0.011	596.05 0.002	6959.63 0.000	238185.98 0.000	109890109.89 0.000
	2	1.08 0.930	1.48 0.678	2.61 0.383	5.98 0.167	18.29 0.055	81.34 0.012	628.78 0.002	12832.62 0.000	2676659.53 0.000
	3	1.00 0.987	1.14 0.879	1.54 0.650	2.62 0.382	5.82 0.172	18.26 0.055	94.41 0.011	1156.93 0.001	109629.89 0.000
	4	1.00 0.998	1.03 0.967	1.18 0.850	1.58 0.633	2.65 0.377	6.02 0.166	21.12 0.047	157.00 0.006	6807.23 0.000
a	5	1.00 1.000	1.01 0.994	1.05 0.953	1.20 0.834	1.61 0.623	2.73 0.367	6.65 0.150	30.49 0.033	611.64 0.002
	6	1.00 1.000	1.00 0.999	1.01 0.989	1.06 0.945	1.21 0.828	1.62 0.618	2.85 0.350	8.27 0.121	78.15 0.013
	7	1.00 1.000	1.00 1.000	1.00 0.998	1.01 0.988	1.06 0.945	1.20 0.833	1.62 0.617	3.10 0.322	14.25 0.070
	8	1.00 1.000	1.00 1.000	1.00 1.000	1.00 0.998	1.01 0.989	1.05 0.954	1.18 0.851	1.60 0.624	3.79 0.264
	9	1.00 1.000	1.00 1.000	1.00 1.000	1.00 1.000	1.00 0.999	1.01 0.994	1.03 0.972	1.12 0.893	1.54 0.651
	10	1.00 1.000	1.00 1.000	1.00 1.000	1.00 1.000	1.00 1.000	1.00 1.000	1.00 1.000	1.00 1.000	1.00 1.000

[††] See Mathcad Programs 1 and 2 in Appendix B for the calculation of the values in Table 2.6.

Table 2.7. The average run length for the one-sided Shewhart sign chart with $n = 15$.[‡‡]

ARL **P(Signal)**	p^U	**0.9**	**0.8**	**0.7**	**0.6**	**0.5**	**0.4**	**0.3**	**0.2**	**0.1**
	p^L	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
a	0	4.86 0.206	28.42 0.035	210.63 0.005	2126.82 0.000	32768.00 0.000	931322.57 0.000	69691719.38 0.000	30517578125.00 0.000	999999999999987.00 0.000
	1	1.82 0.549	5.98 0.167	28.35 0.035	193.35 0.005	2048.00 0.000	39630.75 0.000	1935881.09 0.000	500288165.98 0.000	7352941176470.50 0.000
	2	1.23 0.816	2.51 0.398	7.88 0.127	36.88 0.027	270.81 0.004	3585.46 0.000	114687.42 0.000	17528764.00 0.000	115727346371.95 0.000
	3	1.06 0.944	1.54 0.648	3.37 0.297	11.05 0.091	56.89 0.018	518.73 0.002	10910.04 0.000	988871.98 0.000	2938272765.74 0.000
	4	1.01 0.987	1.20 0.836	1.94 0.515	4.60 0.217	16.88 0.059	106.98 0.009	1487.58 0.001	80245.85 0.000	107571980.98 0.000
	5	1.00 0.998	1.07 0.939	1.39 0.722	2.48 0.403	6.63 0.151	29.56 0.034	273.78 0.004	8831.92 0.000	5358475.36 0.000
	6	1.00 1.000	1.02 0.982	1.15 0.869	1.64 0.610	3.29 0.304	10.52 0.095	65.61 0.015	1273.91 0.001	351310.79 0.000
	7	1.00 1.000	1.00 0.996	1.05 0.950	1.27 0.787	2.00 0.500	4.69 0.213	19.99 0.050	235.86 0.004	29739.88 0.000
	8	1.00 1.000	1.00 0.999	1.02 0.985	1.11 0.905	1.44 0.696	2.56 0.390	7.63 0.131	55.37 0.018	3219.26 0.000
	9	1.00 1.000	1.00 1.000	1.00 0.996	1.04 0.966	1.18 0.849	1.68 0.597	3.59 0.278	16.38 0.061	444.51 0.002
	10	1.00 1.000	1.00 1.000	1.00 0.999	1.01 0.991	1.06 0.941	1.28 0.783	2.06 0.485	6.09 0.164	78.61 0.013
	11	1.00 1.000	1.00 1.000	1.00 1.000	1.00 0.998	1.02 0.982	1.10 0.909	1.42 0.703	2.84 0.352	18.00 0.056
	12	1.00 1.000	1.00 1.000	1.00 1.000	1.00 1.000	1.00 0.996	1.03 0.973	1.15 0.873	1.66 0.602	5.43 0.184
	13	1.00 1.000	1.00 1.000	1.00 1.000	1.00 1.000	1.00 1.000	1.01 0.995	1.04 0.965	1.20 0.833	2.22 0.451
	14	1.00 1.000	1.00 1.000	1.00 1.000	1.00 1.000	1.00 1.000	1.00 1.000	1.00 0.995	1.04 0.965	1.26 0.794
	15	1.00 1.000	1.00 1.000	1.00 1.000	1.00 1.000	1.00 1.000	1.00 1.000	1.00 1.000	1.00 1.000	1.00 1.000

[‡‡] See Mathcad Programs 1 and 2 in Appendix B for the calculation of the values in Table 2.7.

2.1.7. Two-sided control charts

Result 2.7: Probability of a signal

The probability that the control chart signals, that is, the probability that the plotting statistic SN_i is greater than or equal to the UCL, or smaller than or equal to the LCL, can be expressed in terms of $p = P(X_{ij} \geq \theta_0)$, the sample size n and the constant c. The probability of a signal is then given by

$$P(\text{Signal}) = 1 - P(-c < 2T_i - n < c) = 1 - P\left(T_i \leq \frac{n+c}{2} - 1\right) + P\left(T_i \leq \frac{n-c}{2}\right). \quad (2.15)$$

Note that (2.15) can be solved by using the cdf of a binomial distribution.

Result 2.8: Average run length

Since the run length has a geometric distribution we have that

$$ARL = E(N) = \frac{1}{P(\text{Signal})}. \quad (2.16)$$

Compare expression (2.16) to expressions (2.9) and (2.13).

Result 2.9: Standard deviation of the run length

Since the run length has a geometric distribution we have that

$$SDRL(N) = \frac{\sqrt{1 - P(\text{Signal})}}{P(\text{Signal})}. \quad (2.17)$$

Compare expression (2.17) to expressions (2.10) and (2.14).

2.1.8. Summary

In Section 2.1 we have described and evaluated the nonparametric Shewhart-type sign control chart. Generally speaking, when the underlying process distribution is either asymmetric or symmetric with heavy tails, sign charts are more efficient while the reverse is true for normal and normal-like distributions with light tails. One practical advantage of the nonparametric Shewhart-type sign control chart is that there is no need to assume a particular parametric distribution for the underlying process (see Section 1.4 for other advantages of nonparametric charts).

2.2. The Shewhart-type control chart with warning limits

2.2.1. Introduction

It is known that standard Shewhart charts are efficient in detecting large process shifts quickly, but are insensitive to small shifts (see, for example, Montgomery (2005)). Additional supplementary rules have been suggested to increase the sensitivity of standard Shewhart charts to small process shifts. Shewhart (1941) gave the first proposal in making the standard Shewhart chart more sensitive to small process shifts by proposing that additional sensitizing tests should be incorporated into the standard Shewhart chart. Various rules or 'tests for special causes' have been considered in the literature for parametric control charts; see for example, the rules associated with the Shewhart control chart in Nelson (1984) and in the Western Electric handbook (1956). See also the discussion in Montgomery (2001).

Runs rules can be used to increase the sensitivity of standard Shewhart charts. Denote each runs rule by $R(r,n,k,l)$ where a signal is given if r out of the last n points fall in the interval (k,l), where $r \leq n$ are integers and $k < l$. The well-known standardized Shewhart $\overline{X}$ control chart is denoted by $\{R(1,1,-\infty,-3) \cup R(1,1,3,\infty)\}$, since the standardized Shewhart $\overline{X}$ control chart signals if any charting statistic (1 out of 1 point) falls in the interval $(-\infty,-3)$ or if any charting statistic (1 out of 1 point) falls in the interval $(3,\infty)$.

Page (1955) considered a Markov-chain approach for simple combinations of runs rules. Amin et al. (1995) considered Shewhart-type sign charts with warning limits and runs rules. Page (1962), Weindling, Littauer and Oliveira (1979) and Champ and Woodall (1987) studied the properties of $\overline{X}$ charts with warning limits.

Incorporating the runs rules $\{R(1,1,-\infty,a_L) \cup R(1,1,a_U,\infty)\}$ into the Shewhart sign chart is similar to using *action limits* where action will be taken if any 1 point falls outside the action limits. Incorporating the two runs rules $\{R(r,r,w_U,a_U) \cup R(r,r,a_L,w_L)\}$ into the Shewhart sign chart is similar to using *warning limits* where action will be taken if r successive points fall between the warning and action limits, that is, action will be taken if r successive points fall between w_U and a_U *or* action will be taken if r successive points fall between a_L and w_L. Hence, rule A follows: Action will be taken if r successive points fall between w_U and a_U, *or* if r successive points fall between a_L and w_L, *or* if any point falls outside the action limits. Let L denote the ARL of rule A. Assume that the upper action and upper warning limits are equal to some constants represented by a and w, respectively, that is, $a_U = a$ and $w_U = w$. In the case of the Shewhart-type sign control chart with warning limits, sensible choices for the lower action and lower warning limits are $-a$ and $-w$, respectively, that is, $a_L = -a$ and $w_L = -w$. The latter choices are sensible, since the in-control distribution of SN_i is symmetric about zero (see Section 2.1.3).

In Section 2.2.3.1 two runs rules are incorporated into the upper one-sided Shewhart sign chart. Similarly, in Section 2.2.3.2 two runs rules are incorporated into the lower one-sided Shewhart sign chart. The average run lengths are computed for the upper and lower charts, respectively. Finally, in Section 2.2.4 two runs rules are incorporated into the two-sided Shewhart sign chart.

2.2.2. Markov chain representation

A Markov chain representation of a Shewhart chart supplemented with runs rules is used for calculating the probability that any subset of runs rules will give an out-of-control signal. In this section some basic concepts of matrices and transition probabilities are given and explained. An illustrative example follows in the next section, i.e. Section 2.2.3.1.

Let p_{ij} represent the probability that the process will, when in state i, next make a transition to state j. Since probabilities are non-negative, $p_{ij} \geq 0$. Let TPM denote the matrix of one-step transition probabilities. The abbreviation TPM will be used throughout the text for *transition probability matrix* which is given by

$$TPM_{(n+1)\times(n+1)} = [p_{ij}] = \begin{pmatrix} p_{00} & p_{01} & \cdots & p_{0n} \\ p_{10} & p_{11} & \cdots & p_{1n} \\ \vdots & \vdots & \vdots & \vdots \\ p_{i0} & p_{i1} & \cdots & p_{in} \\ \vdots & \vdots & \vdots & \vdots \\ p_{n0} & p_{n1} & \cdots & p_{nn} \end{pmatrix} \text{ for } i = 0,1,...,n \text{ and } j = 0,1,...,n \text{ with } n \geq 2.$$

The i^{th} row, $(p_{i0}, p_{i1},..., p_{in})$, contains all the transition probabilities to go from state i to one of the states in Ω, where Ω denotes the state space, i.e. $\Omega = \{0,1,...,n\}$. We have that

$$\sum_{j\in\Omega} p_{ij} = 1 \quad \forall i \tag{2.18}$$

since it's certain that starting in state i the process will go to one of the states in one step.

2.2.3. One-sided control charts

2.2.3.1. Upper one-sided control charts

The upper one-sided Shewhart sign chart, described previously, is efficient in detecting large process shifts quickly. Since it is known to be inefficient in detecting small process shifts, an upper warning limit is drawn below the upper action limit to increase its sensitivity for detecting small shifts.

Define rule A_U as: 'Action will be taken if r successive points fall between w_U and a_U (denoted by $R(r,r,w_U,a_U)$) *or* if any point falls above a_U (denoted by $R(1,1,a_U,\infty)$)'. Clearly, rule A_U is created to detect upward shifts. Let L^U denote the *ARL* of rule A_U. L^U can be calculated by enumerating the possible combinations of the positions of the plotted points and treating them as the states of a discrete Markov process. The following set of rules is used: $\{ R(r,r,w_u,a_u) \cup R(1,1,a_u,\infty) \}$. The 3 mutually exclusive intervals (also referred to as zones) which are considered are given by:

Zone Z_0 = the interval $(-\infty, w_U)$

Zone Z_1 = the interval $[w_U, a_U)$

Zone Z_2 = the interval $[a_U, \infty)$

These zones are graphically represented in Figure 2.3.

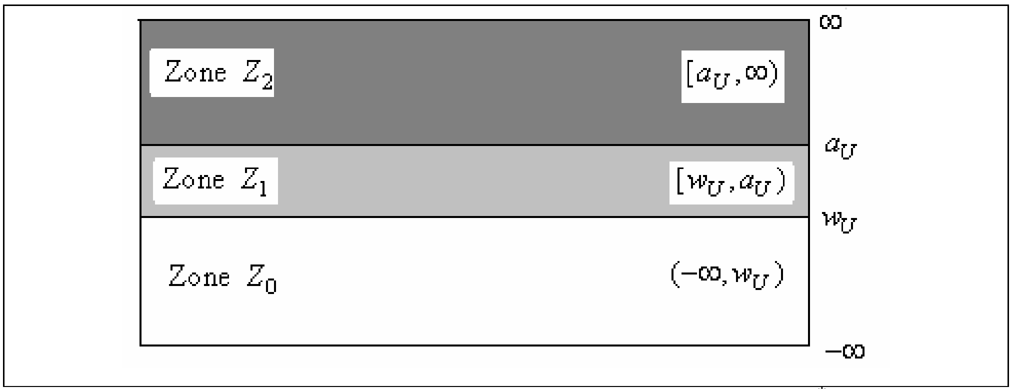

Figure 2.3. A control chart partitioned into 3 zones*.

$R(r, r, w_u, a_u)$: The chart will signal if any r successive points fall in Zone Z_1; or

$R(1,1, a_u, \infty)$: The chart will signal if any 1 point falls in Zone Z_2.

Classification of states

If a state is entered once and can't be left, the state is said to be absorbent. As a result, the probability of going from an absorbent state to the same absorbent state is equal to one. The transient (non-absorbent) states are the remaining states of which the time of return or the number of steps before return is uncertain.

Table 2.8. Classifications and descriptions of states.

State number	Description of state	Absorbent (A)/ Non-absorbent (NA)
0	1 point plots in Zone Z_0	NA
1	1 point plots in Zone Z_1	NA
2	2 successive points plot in Zone Z_1	NA
3	3 successive points plot in Zone Z_1	NA
$\vdots$	$\vdots$	$\vdots$
$r-1$	$r-1$ successive points plot in Zone Z_1	NA
r	r successive points plot in Zone Z_1 or 1 point plots in Zone Z_2	A

* Any point plotting on a line is to be taken as plotting into the adjacent more extreme zone of the chart.

Let p_i denote the probability of plotting in Zone Z_i for $i = 0,1,2$. Therefore:

> p_0 is the probability of plotting in Zone Z_0; $p_0 = P(SN_i < w_U)$;
>
> p_1 is the probability of plotting in Zone Z_1; $p_1 = P(w_U \le SN_i < a_U)$; and
>
> p_2 is the probability of plotting in Zone Z_2; $p_2 = P(SN_i \ge a_U)$.

Clearly, $\sum_{i=0}^{2} p_i = p_0 + p_1 + p_2 = 1$, since the statistic **must** plot in one of the 3 zones. The transition probability matrix, $TPM = [p_{ij}]$, for $i = 0,1,2,\ldots,r$ and $j = 0,1,2,\ldots,r$ is given by

$$TPM_{(r+1)\times(r+1)} = \begin{pmatrix} p_{00} & p_{01} & p_{02} & \cdots & p_{0(r-1)} & p_{0r} \\ p_{10} & p_{11} & p_{12} & \cdots & p_{1(r-1)} & p_{1r} \\ p_{20} & p_{21} & p_{22} & \cdots & p_{2(r-1)} & p_{2r} \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ p_{(r-1)0} & p_{(r-1)1} & p_{(r-1)2} & \cdots & p_{(r-1)(r-1)} & p_{(r-1)r} \\ p_{r0} & p_{r1} & p_{r2} & \cdots & p_{r(r-1)} & p_{rr} \end{pmatrix} = \begin{pmatrix} p_0 & p_1 & 0 & \cdots & 0 & p_2 \\ p_0 & 0 & p_1 & \cdots & 0 & p_2 \\ p_0 & 0 & 0 & \cdots & 0 & p_2 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ p_0 & 0 & 0 & \cdots & 0 & p_1 + p_2 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}.$$

From expression (2.18) we have $\sum_{j \in \Omega} p_{ij} = 1 \quad \forall i$. This is easily proven for $i = 0,1,2,\ldots,r$. For example, for $i = 0$ we have that $\sum_{j=0}^{r} p_{0j} = p_0 + p_1 + 0 + \cdots + 0 + p_2 = 1$. The rest of the calculations follow similarly. Table 2.9 illustrates that the TPM can be partitioned into 4 sections.

Table 2.9. Transition probabilities for a Markov chain with one absorbing state.

	States at time $t+1$					
States at time t	0 (NA)	1 (NA)	2 (NA)	…	r - 1 (NA)	r (A)
0 (NA)	p_0	p_1	0	…	0	p_2
1 (NA)	p_0	0	p_1	…	0	p_2
2 (NA)	p_0	0	0	…	0	p_2
⋮	⋮	⋮	⋮	…	⋮	⋮
r - 1 (NA)	p_0	0	0	…	0	$p_1 + p_2$
r (A)	0	0	0	…	0	1

=

States at time $t+1$					
0 (NA)	1 (NA)	2 (NA)	…	r - 1 (NA)	r (A)
$Q_{r\times r}$					$\underline{p}_{r\times 1}$
$\underline{0}'_{1\times r}$					$1_{1\times 1}$

where the sub-matrix $Q_{r\times r} = \begin{pmatrix} p_0 & p_1 & 0 & \cdots & 0 \\ p_0 & & p_1 & \cdots & 0 \\ p_0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ p_0 & 0 & 0 & \cdots & 0 \end{pmatrix}$ is called the essential transition probability sub-matrix and it contains all the transition probabilities of going from a non-absorbent (transient) state to a non-absorbent state, $Q:(NA \rightarrow NA)$. $\underline{p}_{r\times 1} = \begin{pmatrix} p_2 & p_2 & p_2 & \cdots & p_1 + p_2 \end{pmatrix}'$ contains all the transition probabilities of going from each non-absorbent state to the absorbent states, $\underline{p}:(NA \rightarrow A)$. $\underline{0}'_{1\times r} = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \end{pmatrix}$ contains all the transition probabilities of going from each absorbent state to the non-absorbent states, $\underline{0}' :(A \rightarrow NA)$. $\underline{0}'$ is a row vector with all its elements equal to zero, since it is impossible to go from an absorbent state to a non-absorbent state, because once an absorbent state is entered, it is never left. $1_{1\times 1}$ represents the scalar value one which is the probability of going from an absorbent state to an absorbent state, $1:(A \rightarrow A)$. Therefore,

$$TPM_{(r+1)\times(r+1)} = \begin{pmatrix} Q_{r\times r} & | & \underline{p}_{r\times 1} \\ - & - & - \\ \underline{0}'_{1\times r} & | & 1_{1\times 1} \end{pmatrix}. \tag{2.19}$$

Let L_i^U denote the run length of the upper one-sided chart with initial state i for $i = 0,1,2,\ldots,r-1$. To calculate the probability mass function, define the $r\times 1$ vector $\underline{L}_h^U$ by

$$\underline{L}_h^U = \begin{pmatrix} P(L_0^U = h) \\ P(L_1^U = h) \\ \vdots \\ P(L_{r-1}^U = h) \end{pmatrix} = \begin{pmatrix} \text{Probability that the run length of the chart with initial state 0 is } h \\ \text{Probability that the run length of the chart with initial state 1 is } h \\ \vdots \\ \text{Probability that the run length of the chart with initial state } (r-1) \text{ is } h \end{pmatrix}.$$

Brook and Evans (1972) showed that these vectors can be calculated recursively using

$$\underline{L}_1^U = (I - Q)\underline{1}$$

and (2.20)

$$\underline{L}_h^U = Q\underline{L}_{h-1}^U \text{ for } h = 2,3,\ldots$$

where $\underline{1}$ is a $r \times 1$ column vector of $1's$, I is the $r \times r$ identity matrix and Q is the $r \times r$ essential transition probability sub-matrix obtained from the partitioned TPM.

Multiplying out the matrices in (2.20) we get expressions for L_i^U for $i = 1,2,...r-1$.

$$L_i^U = 1 + p_0 L_0^U + p_1 L_{i+1}^U \quad \text{for} \quad (i = 0,1,2,...,r-2)$$

and (2.21)

$$L_{r-1}^U = 1 + p_0 L_0^U .$$

These equations may be solved recursively for $L_1^U ,..., L_{r-1}^U$ in terms of L_0^U :

$$L_0^U = 1 + p_0 L_0^U + p_1 L_1^U$$
$$L_0^U = 1 + p_0 L_0^U + p_1 (1 + p_0 L_0^U + p_1 L_2^U)$$
$$L_0^U = 1 + p_0 L_0^U + p_1 + p_1 p_0 L_0^U + p_1^{\ 2} L_2^U$$
$$L_0^U = 1 + p_0 L_0^U + p_1 + p_1 p_0 L_0^U + p_1^{\ 2} (1 + p_0 L_0^U + p_1 L_3^U)$$
$$L_0^U = 1 + p_0 L_0^U + p_1 + p_1 p_0 L_0^U + p_1^{\ 2} + p_1^{\ 2} p_0 L_0^U + p_1^{\ 3} L_3^U$$
$$\vdots$$
$$L_0^U = 1 + p_0 L_0^U + p_1 + p_1 p_0 L_0^U + p_1^{\ 2} + p_1^{\ 2} p_0 L_0^U + p_1^{\ 3} + ... + p_1^{\ r-1} L_{r-1}^U$$
$$L_0^U = 1 + p_0 L_0^U + p_1 + p_1 p_0 L_0^U + p_1^{\ 2} + p_1^{\ 2} p_0 L_0^U + p_1^{\ 3} + ... + p_1^{\ r-1} (1 + p_0 L_0^U)$$
$$L_0^U = 1 + p_0 L_0^U + p_1 + p_1 p_0 L_0^U + p_1^{\ 2} + p_1^{\ 2} p_0 L_0^U + p_1^{\ 3} + ... + p_1^{\ r-1} + p_1^{\ r-1} p_0 L_0^U$$

$$L_0^U - p_0 L_0^U - p_1 p_0 L_0^U - p_1^{\ 2} p_0 L_0^U - ... - p_1^{\ r-1} p_0 L_0^U = 1 + p_1 + p_1^{\ 2} + p_1^{\ 3} + ... + p_1^{\ r-1} \quad (2.22)$$

It can be proven by induction on r that $\sum_{i=0}^{r-1} p_1^{\ i} = \frac{1 - p_1^{\ r}}{1 - p_1}$ for $p_1 \neq 1$. Making use of this, expression (2.22) can be simplified to

$$L_0^U - p_0 L_0^U \left(\frac{1 - p_1^{\ r}}{1 - p_1} \right) = \frac{1 - p_1^{\ r}}{1 - p_1}$$

$$L_0^U \left[\frac{1 - p_1 - p_0 (1 - p_1^{\ r})}{1 - p_1} \right] = \frac{1 - p_1^{\ r}}{1 - p_1}$$

$$L_0^U = \frac{1 - p_1^{\ r}}{1 - p_1 - p_0 (1 - p_1^{\ r})}. \quad (2.23)$$

Expression (2.23) of this thesis is given in Amin et al. (1995) and determined in Page (1962). Expression (2.23) is a closed form expression of the in-control average run length of a one-sided chart with warning and action limits in the positive direction only.

Therefore, the in-control average run length of the one-sided (upper or positive direction) chart with warning limit w_U and control limit (action limit) at a_U is given by (2.23) where

$$p_0 = P(SN_i < w_U) = \sum_{i=0}^{\frac{w+n-2}{2}} \binom{n}{i} p^i (1-p)^{n-i} \qquad (2.24)$$

and

$$p_1 = P(w_U \le SN_i < a_U) = \sum_{i=0}^{\frac{a+n-2}{2}} \binom{n}{i} p^i (1-p)^{n-i} - \sum_{i=0}^{\frac{w+n-2}{2}} \binom{n}{i} p^i (1-p)^{n-i} . \qquad (2.25)$$

The derivations of (2.24) and (2.25) are given below.

Derivation of expression (2.24):

p_0

$= P(SN_i < w_U)$

by using the relationship between SN_i and T_i (recall that $SN_i = 2T_i - n$) we obtain

$= P(2T_i - n < w_U)$

$= P\left(T_i < \frac{w_U + n}{2}\right)$

$= P\left(T_i \le \frac{w_U + n - 2}{2}\right)$

given that T_i is binomially distributed with parameters n and $p = P(X_{ij} \ge \theta_0)$ we obtain

$= \sum_{i=0}^{\frac{w_U+n-2}{2}} \binom{n}{i} p^i (1-p)^{n-i}$

since the upper warning limit is equal to some constant w, i.e. $w_U = w$ (see Section 2.2.1) we obtain

$= \sum_{i=0}^{\frac{w+n-2}{2}} \binom{n}{i} p^i (1-p)^{n-i}$.

Derivation of expression (2.25):

p_1

$= P(w_U \le SN_i < a_U)$

by using the relationship between SN_i and T_i (recall that $SN_i = 2T_i - n$) we obtain

$$= P(w_U \le 2T_i - n < a_U)$$

$$= P\left(\frac{w_U + n}{2} \le T_i < \frac{a_U + n}{2}\right)$$

$$= P\left(T_i \le \frac{a_U + n - 2}{2}\right) - P\left(T_i \le \frac{w_U + n - 2}{2}\right)$$

given that T_i is binomially distributed with parameters n and $p = P(X_{ij} \ge \theta_0)$ we obtain

$$= \sum_{i=0}^{\frac{a_U + n - 2}{2}} \binom{n}{i} p^i (1-p)^{n-i} - \sum_{i=0}^{\frac{w_U + n - 2}{2}} \binom{n}{i} p^i (1-p)^{n-i}$$

since the upper action and warning limits are equal to constant values represented by a and w, respectively, i.e. $a_U = a$ and $w_U = w$ (see Section 2.2.1), we obtain

$$= \sum_{i=0}^{\frac{a + n - 2}{2}} \binom{n}{i} p^i (1-p)^{n-i} - \sum_{i=0}^{\frac{w + n - 2}{2}} \binom{n}{i} p^i (1-p)^{n-i} .$$

The in-control average run length of the one-sided (lower or negative direction) chart with warning limit at w_L and control limit (action limit) at a_L can be found by replacing p_0 and p_1 by q_0 and q_1 where $q_0 = P(SN_i > w_L)$ and $q_1 = P(a_L < SN_i \le w_L)$. The in-control average run length for the two-sided chart, denoted L_0, can then be obtained using a result in Roberts (1958), $\frac{1}{L_0} = \frac{1}{L_0^U} + \frac{1}{L_0^L}$. The lower one-sided and two-sided charts with warning limits are discussed in detail in Sections 2.2.3.2 and 2.2.4 respectively.

The in-control average run length for the upper one-sided control chart with both warning and action limits is calculated for a specific example ($n = 10$, $p = 0.5$) by evaluating expressions (2.23), (2.24) and (2.25). These values are shown in Table 2.10. Amin, Reynolds and Bakir

(1995) studied Shewhart charts with both warning and action limits. They constructed a table containing the values of L_0^U for Shewhart charts using the sign statistic when $n=10$ and $p=0.5$. The values of L_0^U were noted for $a_U=8$ and 10, $w_U=0(2)8$ and $r=2,3,...,7$. Table 2.10 is similar to the table constructed by Amin, Reynolds and Bakir (1995). Note that the values of L_0^U can also be constructed for other values of n, a_U, w_U and r.

Table 2.10. Values of L_0^U for Shewhart charts with both warning and action limits when $n=10$ and $p=0.5$.*

	$a_U=8$				$a_U=10$			
w_U / r	**0**	**2**	**4**	**6**	**2**	**4**	**6**	**8**
2	4.1	9.2	30.2	79.4	9.6	38.6	269.2	933.7
3	7.9	23.0	70.1	92.4	27.8	194.7	890.3	1023.0
4	13.5	44.7	88.4	93.1	73.0	593.7	1015.8	1024.0
5	21.2	66.9	92.3	93.1	175.4	911.2	1023.6	1024.0
6	31.0	81.5	93.0	93.1	364.4	1002.8	1024.0	1024.0
7	42.2	88.5	93.1	93.1	609.8	1020.3	1024.0	1024.0

Studying Table 2.10 we observe the following. For values of w_U close to a_U and r reasonably large, the introduction of warning lines will have little effect on L_0^U. The reason being that if w_U is close to a_U, the probability of having r successive points plot in this small interval $[w_u, a_U)$ is small. As an example, the calculation of the in-control average run length for $n=10$, $p=0.5$, $a_U=8$, $w_U=2$ and $r=6$ will be given. By substituting these values into equations (2.23), (2.24) and (2.25) we obtain

$$p_0 = \sum_{i=0}^{5}\binom{10}{i}(0.5)^{10} = 0.6230,$$

$$p_1 = \sum_{i=0}^{8}\binom{10}{i}(0.5)^{10} - \sum_{i=0}^{5}\binom{10}{i}(0.5)^{10} = 0.3662 \text{ and}$$

$$L_0^U = \frac{1-p_1^{\,r}}{1-p_1-p_0(1-p_1^{\,r})} = \frac{1-(0.3662)^6}{1-0.3662-0.6230\left(1-(0.3662)^6\right)} = 81.4689 \approx 81.5.$$

* See Mathcad Program 3 in Appendix B for the calculation of the values in Table 2.10. This table also appears in Amin, Reynolds and Bakir (1995), page 1606, Table 2.

2.2.3.2. Lower one-sided control charts

The lower one-sided Shewhart sign chart, described previously, is efficient in detecting large process shifts quickly. Since it is known to be inefficient in detecting small process shifts, a lower warning limit is drawn above the lower action limit to increase its sensitivity for detecting small shifts.

Define rule A_L as: 'Action will be taken if r successive points fall between a_L and w_L (denoted by $R(r,r,a_L,w_L)$) *or* if any point falls below a_L (denoted by $R(1,1,-\infty,a_L)$)'. Clearly, rule A_L is created to detect downward shifts. Let L^L denote the ARL of rule A_L. L_0^L can be computed similarly as L_0^U (see equation (2.23)) with p_0 and p_1 being replaced by q_0 and q_1, where q_0 denotes the probability that a given sample point falls above w_L and q_1 denotes the probability that a given sample point falls between a_L and w_L. Therefore, we have that

$$L_0^L = \frac{1-q_1{}^r}{1-q_1-q_0(1-q_1{}^r)} \tag{2.26}$$

where

$$q_0 = P(SN_i > w_L) = 1 - \sum_{i=0}^{\frac{n-w}{2}} \binom{n}{i} p^i (1-p)^{n-i} \tag{2.27}$$

and

$$q_1 = P(a_L < SN_i \le w_L) = \sum_{i=0}^{\frac{n-w}{2}} \binom{n}{i} p^i (1-p)^{n-i} - \sum_{i=0}^{\frac{n-a}{2}} \binom{n}{i} p^i (1-p)^{n-i} . \tag{2.28}$$

Compare expressions (2.27) and (2.28) to (2.24) and (2.25).

The in-control average run length for the lower one-sided control chart with both warning and action limits is calculated for a specific example ($n=10$, $p=0.5$) by evaluating expressions (2.26), (2.27) and (2.28). These values are shown in Table 2.11.

Table 2.11. Values of L_0^L for Shewhart charts with both warning and action limits when $n=10$ and $p=0.5$.*

w_L / r	$a_L=8$				$a_L=10$			
	0	**2**	**4**	**6**	**2**	**4**	**6**	**8**
2	4.1	9.2	30.2	79.4	9.6	38.6	269.2	933.7
3	7.9	23.0	70.1	92.4	27.8	194.7	890.3	1023.0
4	13.5	44.7	88.4	93.1	73.0	593.7	1015.8	1024.0
5	21.2	66.9	92.3	93.1	175.4	911.2	1023.6	1024.0
6	31.0	81.5	93.0	93.1	364.4	1002.8	1024.0	1024.0
7	42.2	88.5	93.1	93.1	609.8	1020.3	1024.0	1024.0

Studying Table 2.11 we observe the following. For values of w_L close to a_L and r reasonably large, the introduction of warning lines will have little effect on L_0^L. The reason being that if w_L is close to a_L, the probability of having r successive points plot in this small interval $(a_L, w_L]$ is small. As stated earlier, due to the symmetry of the Binomial distribution we have that if $a_U = a$ then let $a_L = -a$ and if $w_U = w$ then let $w_L = -w$. As a result the values of L_0^L and the values of L_0^U are equal.

As an example, the calculation of the in-control average run length for $n=10$, $p=0.5$, $a_L=8$, $w_L=2$ and $r=6$ will be given. By substituting these values into equations (2.26), (2.27) and (2.28) we obtain

$$q_0 = 1-\sum_{i=0}^{4}\binom{10}{i}(0.5)^{10} = 0.6230,$$

$$q_1 = \sum_{i=0}^{4}\binom{10}{i}(0.5)^{10} - \sum_{i=0}^{1}\binom{10}{i}(0.5)^{10} = 0.3662 \text{ and}$$

$$L_0^L = \frac{1-q_1^{\,r}}{1-q_1-q_0(1-q_1^{\,r})} = \frac{1-(0.3662)^6}{1-0.3662-0.6230\left(1-(0.3662)^6\right)} = 81.4689 \approx 81.5.$$

* See Mathcad Program 3 in Appendix B for the calculation of the values in Table 2.11.

2.2.4. Two-sided control charts

Roberts (1958) provided a method of approximating the *ARL* of the two-sided Shewhart chart with both warning and action limits. The *ARL* for each separate one-sided Shewhart chart was calculated and then combined by applying equation (2.29)

$$\frac{1}{L_0} = \frac{1}{L_0^U} + \frac{1}{L_0^L}. \tag{2.29}$$

(See Appendix A Theorem 1 for a step-by-step derivation of equation (2.29)). Equation (2.29) can be re-written as

$$L_0 = \frac{L_0^L L_0^U}{L_0^L + L_0^U} \tag{2.30}$$

where L_0 denotes the *ARL* of a two-sided chart. In practice some observations can be tied with the specified median. If the number of such cases, within a sample, is small (relative to *n*) one can drop the tied cases and reduce *n* accordingly. On the other hand, if the number of ties is large, more sophisticated analysis might be necessary.

The in-control average run length for the two-sided control chart with both warning and action limits is calculated by evaluating expression (2.30). These values are shown in Table 2.12 for $n = 10$, $p = 0.5$, $a = 8$ and 10, $w = 0(2)8$ and $r = 2,3,...,7$.

Table 2.12. Values of L_0 for Shewhart charts with both warning and action limits when $n = 10$ and $p = 0.5$.*

	$a = 8$				$a = 10$			
w / ***r***	**0**	**2**	**4**	**6**	**2**	**4**	**6**	**8**
2	2.1	4.6	15.1	39.7	4.8	19.3	134.6	466.9
3	4.0	11.5	35.0	46.2	13.9	97.4	445.2	511.5
4	6.7	22.4	44.2	46.5	36.5	296.8	507.9	512.0
5	10.6	33.5	46.2	46.5	87.7	455.6	511.8	512.0
6	15.5	40.7	46.5	46.5	182.2	501.4	512.0	512.0
7	21.1	44.2	46.5	46.5	304.9	510.2	512.0	512.0

* See Mathcad Program 3 in Appendix B for the calculation of the values in Table 2.12.

Studying Table 2.12 we observe the following: For values of w close to a and r reasonably large, the introduction of warning limits will have little effect on L_0. The reason being that if w is close to a (and, consequently, $-w$ is close to $-a$), the probability of having r successive points plot in the small interval $[w,a)$ or $(-a,-w]$ is small. These procedures can't be meaningfully illustrated using the data of Montgomery (2001) because the sample size n = 5 used there is too small. It may be noted that the highest possible L_0 values for the basic (without warning limits) two-sided sign chart can be seen to be 2^{n-1} (see Amin, Reynolds and Bakir (1995) pp. 1609-1610 and their Appendix on pp. 1620-1621 for a detailed discussion on and a proof that $max\, L_0 = 2^{n-1}$). Thus, achievable values of L_0 are too small for practical use, unless n is about 10. In Table 2.13, the charting constants, i.e. the warning and action limits, are shown, along with the achieved *ARL* values, for the in-control and one out-of-control case. The *ARL* values for the two-sided sign chart, without the warning limits, are shown in each case, within parentheses, for reference.

Table 2.13. In-control *ARL* values for the two-sided sign chart with and without warning limits for $n = 10$*.

	***r* = 2**	***r* = 3**	***r* = 6**
		***w* = 7 and *a* = 10**	
***p* = 0.5 (in-control)**	208.97 (512.00)	476.03 (512.00)	511.99 (512.00)
***p* = 0.6 (out-of-control)**	35.03 (162.60)	103.28 (162.60)	162.17 (162.60)
		***w* = 7 and *a* = 8**	
***p* = 0.5 (in-control)**	42.86 (46.55)	46.37 (46.55)	46.55 (46.55)
***p* = 0.6 (out-of-control)**	16.37 (20.82)	20.16 (20.82)	20.82 (20.82)

It is seen that adding warning limits to a control chart decreases its average run length. For example, adding a warning limit at 7 to the basic sign chart with an action limit at 10 decreases the ARL_0 approximately 59% (from 512 to 208.97), when $r = 2$, 7% (512 to 476.03) when $r = 3$ and 0.002% (512 to 511.99) when $r = 6$, respectively. The out-of-control average

* See Mathcad Program 3 in Appendix B for the calculation of the values in Table 2.13.

run length is decreased by approximately 79% (from 162.6 to 35.03) when $r = 2$, 36% (from 162.6 to 103.28) when $r = 3$ and 0.26% (from 162.6 to 162.17) when $r = 6$, respectively. Note that although the out-of-control average run length is reduced significantly (which means a quicker detection of shift) by the addition of warning limits, the ARL_0 is also reduced significantly. This poses a dilemma in practice, since it is desirable to have a high ARL_0 and a low *FAR*, so one would need to strike a balance. One possibility is to use warning limits closer to the action limits. For example, from the second panel of Table 2.13, we see that adding a warning limit at 7 to the sign chart with an action limit of 8, decreases the ARL_0 by only 8% (from 46.55 to 42.86) when $r = 2$ and has little effect on ARL_0 when r is reasonably large. Amin et al. (1995) concluded that for the upper one-sided Shewhart-type sign chart, introduction of warning limits will have little effect on the in-control average run length, but can significantly reduce the out-of-control average run length for small shifts when the warning limits are chosen close to the action limits and r is reasonably large. Similar conclusions are expected to hold for two-sided charts.

Up to this point we have discussed methods to increase the sensitivity of standard Shewhart control charts to small process shifts. Another method is to extend the existing charts by incorporating various signaling rules involving runs of the plotting statistic. The signaling rules considered include the following: A process is declared to be out-of-control when (a) a single point (charting statistic) plots outside the control limit(s) (*1-of-1* rule) (b) k consecutive points (charting statistics) plot outside the control limit(s) (*k-of-k* rule) or (c) exactly k of the last w points (charting statistics) plot outside the control limit(s) (*k-of-w* rule). We can consider these signaling rules where both k and w are positive integers with $1 \le k \le w$ and $w \ge 2$. Rule (a) is the simplest and is the most frequently used in the literature. Thus, the *1-of-1* rule corresponds to the usual control chart, where a signal is given when a plotting statistic falls outside the control limit(s). Rules (a) and (b) are special cases of rule (c); rules (b) and (c) have been used in the context of supplementing the Shewhart charts with warning limits and zones.

Example 2.4

A two-sided Shewhart chart incorporating the *2-of-2* rule with *one* absorbing state that corresponds to the out-of-control signal

In this example, a control chart is viewed as consisting of the zones shown in Figure 2.4.

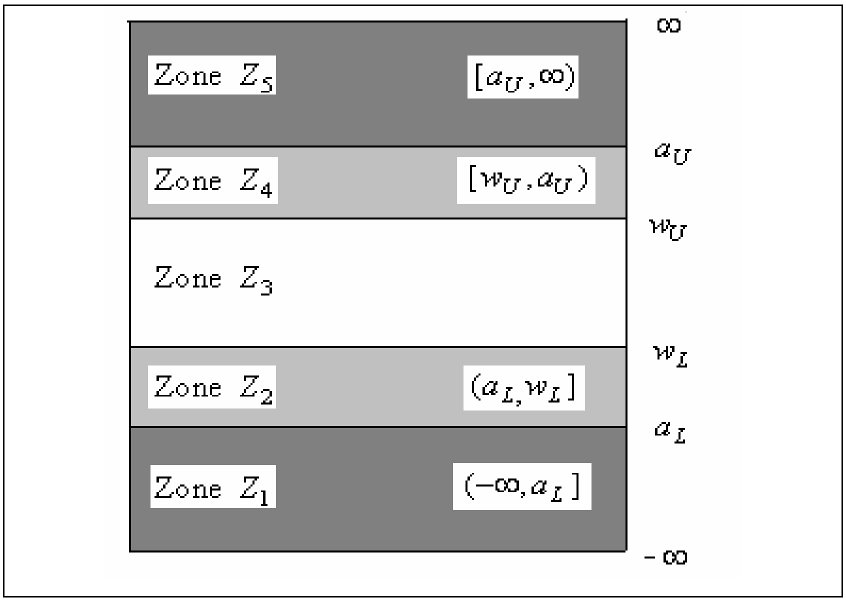

Figure 2.4. A control chart partitioned into 5 zones.

Let p_i denote the probability of plotting in Zone Z_i for $i = 1,2,3,4,5$. To illustrate the calculation of signal probabilities, the following set of rules is used:

$\{R(1,1,-\infty,a_L) \cup R(2,2,a_L,w_L) \cup R(2,2,w_U,a_U) \cup R(1,1,a_U,\infty)\}$.

$R(1,1,-\infty,a_L)$: The chart will signal if any 1 point falls in Zone Z_1 (below LCL).

$R(2,2,a_L,w_L)$: The chart will signal if any 2 successive points fall in Zone Z_2.

$R(2,2,w_U,a_U)$: The chart will signal if any 2 successive points fall in Zone Z_4.

$R(1,1,a_U,\infty)$: The chart will signal if any 1 point falls in Zone Z_5 (above UCL).

Table 2.14. Classifications and descriptions of states.

State number	Description of state	Absorbent (A)/ Non-absorbent (NA)
0	No points beyond any of the control limits. Point plots in Zone Z_3	NA
1	Point plots in Zone Z_2	NA
2	Point plots in Zone Z_4	NA
3	Point plots below a_L or above a_U or 2 successive points fall between w_U and a_U or 2 successive points fall between w_L and a_L.	A

Clearly, $\sum_{i=1}^{5} p_i = p_1 + p_2 + p_3 + p_4 + p_5 = 1$, since the statistic **must** plot in one of the 5 zones. The transition probabilities are given in the transition probability matrix, $TPM = [p_{ij}]$, for $i = 0,1,2,3$ and $j = 0,1,2,3$.

$$TPM_{4\times4} = \begin{pmatrix} p_{00} & p_{01} & p_{02} & p_{03} \\ p_{10} & p_{11} & p_{12} & p_{13} \\ p_{20} & p_{21} & p_{22} & p_{23} \\ p_{30} & p_{31} & p_{32} & p_{33} \end{pmatrix} = \begin{pmatrix} p_3 & p_2 & p_4 & p_1 + p_5 \\ p_3 & 0 & p_4 & p_1 + p_5 + p_2 \\ p_3 & p_2 & 0 & p_1 + p_5 + p_4 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

From (2.18) we have $\sum_{j\in\Omega} p_{ij} = 1 \quad \forall i$. This is easily shown for $i = 0,1,2,3$. For example, for $i = 0$ we have that $\sum_{j=0}^{3} p_{0j} = p_3 + p_2 + p_4 + p_1 + p_5 = 1$. The rest of the calculations follow similarly. Table 2.15 illustrates that the TPM can be partitioned into 4 sections.

Table 2.15. Transition probabilities of the *2-of-2* rule for a Markov chain with *one* absorbing state.

	States at time $t+1$					States at time $t+1$			
States at time t	0 (NA)	1 (NA)	2 (NA)	3 (A)		0 (NA)	1 (NA)	2 (NA)	3 (A)
0 (NA)	p_3	p_2	p_4	$p_1 + p_5$	=				
1 (NA)	p_3	0	p_4	$p_1 + p_5 + p_2$			$Q_{3\times3}$		$\underline{p}_{3\times1}$
2 (NA)	p_3	p_2	0	$p_1 + p_5 + p_4$					
3 (A)	0	0	0	1			$\underline{0}'_{1\times3}$		$1_{1\times1}$

Brook and Evans (1972) showed that the *ARL* for initial state i can be calculated by adding the elements in the i^{th} row of $(I_{3\times3} - Q_{3\times3})^{-1}$. Making use of $(I - Q)^{-1}$ is typically done in stochastic processes where one works with recurrence and first passage times (see, for example, Bartlett (1953)).

$$I_{3\times3} - Q_{3\times3} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} p_3 & p_2 & p_4 \\ p_3 & 0 & p_4 \\ p_3 & p_2 & 0 \end{pmatrix} = \begin{pmatrix} 1-p_3 & -p_2 & -p_4 \\ -p_3 & 1 & -p_4 \\ -p_3 & -p_2 & 1 \end{pmatrix}$$

$$(I_{3\times3} - Q_{3\times3})^{-1} = \begin{pmatrix} 1-p_3 & -p_2 & -p_4 \\ -p_3 & 1 & -p_4 \\ -p_3 & -p_2 & 1 \end{pmatrix}^{-1} =$$

$$\frac{-1+p_4p_2}{(-1+p_4p_2+p_3+p_3p_4p_2+p_3p_2+p_3p_4)} \begin{pmatrix} -1+p_4p_2 & -p_2(1+p_4) & -p_4(1+p_2) \\ -p_3(1+p_4) & -1+p_3+p_3p_4 & -p_4 \\ -p_3(1+p_2) & -p_2 & -1+p_3+p_3p_2 \end{pmatrix}.$$

The *ARL* for initial state i can be calculated by adding the elements in the i^{th} row of $(I_{3\times3} - Q_{3\times3})^{-1}$ for $i = 1,2,3$.

Example 2.5

A two-sided Shewhart chart incorporating the *2-of-2* rule with *more than one* absorbing state that corresponds to an out-of-control signal.

This example is similar to the previous example in having three transient states, but differs from the previous example by having *more than one* absorbing state. By changing the classification of the states, the generalization to more than one absorbing state is considered. We need to introduce a rule number and this is done by adding a subscript to each rule, i.e. in general we have that $R(r,n,k,l)$ which now becomes $R_j(r,n,k,l)$ where *j* denotes the rule number. This modification allows for a separate absorbing state, A_j, that is associated with each of the runs rules, $R_j(r,n,k,l)$. This modification of Champ and Woodall's (1987) method was done by Champ and Woodall (1997). As a result, we have the following rules with the corresponding absorbing states (see Figure 2.4 for the partitioning of the control chart into 5 zones):

- $R_1(1,1,-\infty,a_L)$ associated with absorbing state A_1: The chart will signal if any 1 point falls in Zone Z_1 (below LCL).
- $R_2(2,2,a_L,w_L)$ associated with absorbing state A_2: The chart will signal if any 2 successive points fall in Zone Z_2.
- $R_3(2,2,w_U,a_U)$ associated with absorbing state A_3: The chart will signal if any 2 successive points fall in Zone Z_4.
- $R_4(1,1,a_U,\infty)$ associated with absorbing state A_4: The chart will signal if any 1 point falls in Zone Z_5 (above UCL).

Table 2.16. States and next-state transitions by zones.

		Zones					
State number	**State vector**	**Z1** $(-\infty, a_L]$	**Z2** $(a_L, w_L]$	**Z3** (w_L, w_U)	**Z4** $[w_U, a_U)$	**Z5** $[a_U, \infty)$	**Absorbent (A)/ Non-absorbent (NA)**
0	(0,0)	A_1	(1,0)	(0,0)	(0,1)	A_4	NA
1	(1,0)	A_1	A_2	(0,0)	(0,1)	A_4	NA
2	(0,1)	A_1	(1,0)	(0,0)	A_3	A_4	NA
3	A_1	A_1	A_1	A_1	A_1	A_1	A
4	A_2	A_2	A_2	A_2	A_2	A_2	A
5	A_3	A_3	A_3	A_3	A_3	A_3	A
6	A_4	A_4	A_4	A_4	A_4	A_4	A

Each non-absorbing state in Table 2.16 is represented by a vector of 0's and 1's. The vector indicates by the 1's only those observations that may contribute to an out-of-control signal. Let p_i denote the probability of plotting in Zone Z_i for $i = 1,2,3,4,5$. Clearly, $\sum_{i=1}^{5} p_i = p_1 + p_2 + p_3 + p_4 + p_5 = 1$, since the statistic **must** plot in one of the 5 zones. The transition probabilities are given in the transition probability matrix, $TPM = [p_{ij}]$, for $i = 0,1,...,6$ and $j = 0,1,...,6$.

$$TPM_{7\times7} = \begin{pmatrix} p_{00} & p_{01} & p_{02} & p_{03} & p_{04} & p_{05} & p_{06} \\ p_{10} & p_{11} & p_{12} & p_{13} & p_{14} & p_{15} & p_{16} \\ p_{20} & p_{21} & p_{22} & p_{23} & p_{24} & p_{25} & p_{26} \\ p_{30} & p_{31} & p_{32} & p_{33} & p_{34} & p_{35} & p_{36} \\ p_{40} & p_{41} & p_{42} & p_{43} & p_{44} & p_{45} & p_{46} \\ p_{50} & p_{51} & p_{52} & p_{53} & p_{54} & p_{55} & p_{56} \\ p_{60} & p_{61} & p_{62} & p_{63} & p_{64} & p_{65} & p_{66} \end{pmatrix} = \begin{pmatrix} p_3 & p_2 & p_4 & p_1 & 0 & 0 & p_5 \\ p_3 & 0 & p_4 & p_1 & p_2 & 0 & p_5 \\ p_3 & p_2 & 0 & p_1 & 0 & p_4 & p_5 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

From (2.18) we have $\sum_{j \in \Omega} p_{ij} = 1 \quad \forall i$. This is easily shown for $i = 0,1,...,6$. For example, for $i = 0$ we have that $\sum_{j=0}^{6} p_{0j} = p_3 + p_2 + p_4 + p_1 + 0 + 0 + p_5 = 1$. The rest of the calculations follow similarly. Table 2.17 illustrates that the TPM can be partitioned into 4 sections.

Table 2.17. Transition probabilities of the *2-of-2* rule for a Markov chain with more than one absorbing state.

States at time t	States at time $t+1$: 0 (NA)	1 (NA)	2 (NA)	3 (A)	4 (A)	5 (A)	6 (A)
0 (NA)	p_3	p_2	p_4	p_1	0	0	p_5
1 (NA)	p_3	0	p_4	p_1	p_2	0	p_5
2 (NA)	p_3	p_2	0	p_1	0	p_4	p_5
3 (A)	0	0	0	1	0	0	0
4 (A)	0	0	0	0	1	0	0
5 (A)	0	0	0	0	0	1	0
6 (A)	0	0	0	0	0	0	1

=

	States at time $t+1$: 0 (NA)	1 (NA)	2 (NA)	3 (A)	4 (A)	5 (A)	6 (A)
NA	$Q_{3\times3}$			$C_{3\times4}$			
A	$Z_{4\times3}$			$I_{4\times4}$			

where the essential transition probability sub-matrix $Q_{3\times3}=\begin{pmatrix} p_3 & p_2 & p_4 \\ p_3 & 0 & p_4 \\ p_3 & p_2 & 0 \end{pmatrix}$ contains all the transition probabilities of going from a non-absorbent state to a non-absorbent state, $Q:(NA\to NA)$. $C_{3\times4}=\begin{pmatrix} p_1 & 0 & 0 & p_5 \\ p_1 & p_2 & 0 & p_5 \\ p_1 & 0 & p_4 & p_5 \end{pmatrix}$ contains all the transition probabilities of going from each non-absorbent state to the absorbent states, $C:(NA\to A)$. $Z_{4\times3}=\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$ contains all the transition probabilities of going from each absorbent state to the non-absorbent states, $Z:(A\to NA)$. The Z matrix is the zero matrix, since it is impossible to go from an absorbent state to a non-absorbent state, because once an absorbent state is entered, it is never left. $I_{4\times4}=\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$ contains all the transition probabilities of going from an absorbent state to an absorbent state, $I:(A\to A)$. A square matrix of this form is called the identity matrix.

$TPM = TPM^1$ represents the probability that the process will, when in state i, next make a transition to state j, in **one** step. Consider the matrix, TPM^n (read: "TPM to the power of n"), the non-absorbing state i and the absorbing state j. The j^{th} component of TPM^n is the probability that a signal will be caused by the j^{th} set of runs rules that can cause a signal on or before the n^{th} sampling stage given the chart begins in non-absorbing state i. For that reason, an equation for TPM^n is desired. In addition, we will show that $\lim_{n\to\infty} TPM^n = \begin{pmatrix} Z & B \\ Z & I \end{pmatrix}$ where the elements of the matrix B are the probabilities that the chart will go from a non-absorbent state (where no signal is given) to an absorbent state (where a signal is given) in n transitions. We are interested in the matrix B, because the $(i,j)^{th}$ element of B is the long run proportion of times the j^{th} set of runs rules causes the chart to signal given the chart starts in a non-absorbing state i.

Probabilities on the n^{th} sampling stage

$$TPM_{7\times7} = TPM^1_{7\times7} = \begin{pmatrix} Q_{3\times3} & C_{3\times4} \\ Z_{4\times3} & I_{4\times4} \end{pmatrix}$$

$$\begin{aligned}
&TPM^2_{7\times7} \\
&= TPM^1_{7\times7} \cdot TPM^1_{7\times7} \\
&= \begin{pmatrix} Q_{3\times3} & C_{3\times4} \\ Z_{4\times3} & I_{4\times4} \end{pmatrix}\begin{pmatrix} Q_{3\times3} & C_{3\times4} \\ Z_{4\times3} & I_{4\times4} \end{pmatrix} \\
&= \begin{pmatrix} Q^2_{3\times3} + C_{3\times4}Z_{4\times3} & Q_{3\times3}C_{3\times4} + C_{3\times4}I_{4\times4} \\ Z_{4\times3}Q_{3\times3} + I_{4\times4}Z_{4\times3} & Z_{4\times3}C_{3\times4} + I^2_{4\times4} \end{pmatrix} = \begin{pmatrix} Q^2_{3\times3} & Q_{3\times3}C_{3\times4} + C_{3\times4} \\ Z_{4\times3} & I_{4\times4} \end{pmatrix}
\end{aligned}$$

$$
\begin{aligned}
&TPM^{3}_{7\times7}\\
&= TPM^{2}_{7\times7}\cdot TPM^{1}_{7\times7}\\
&= \begin{pmatrix} Q^{2}_{3\times3} & Q_{3\times3}C_{3\times4}+C_{3\times4} \\ Z_{4\times3} & I_{4\times4} \end{pmatrix}\begin{pmatrix} Q_{3\times3} & C_{3\times4} \\ Z_{4\times3} & I_{4\times4} \end{pmatrix}\\
&= \begin{pmatrix} Q^{3}_{3\times3}+Q_{3\times3}C_{3\times4}Z_{4\times3}+C_{3\times4}Z_{4\times3} & Q^{2}_{3\times3}C_{3\times4}+Q_{3\times3}C_{3\times4}I_{4\times4}+C_{3\times4}I_{4\times4} \\ Z_{4\times3}Q_{3\times3}+I_{4\times4}Z_{4\times3} & Z_{4\times3}C_{3\times4}+I^{2}_{4\times4} \end{pmatrix}\\
&= \begin{pmatrix} Q^{3}_{3\times3} & Q^{2}_{3\times3}C_{3\times4}+Q_{3\times3}C_{3\times4}+C_{3\times4} \\ Z_{4\times3} & I_{4\times4} \end{pmatrix}
\end{aligned}
$$

continuing in this way, we obtain

$$
TPM^{n}_{7\times7} = \begin{pmatrix} Q^{n}_{3\times3} & Q^{n-1}_{3\times3}C_{3\times4}+Q^{n-2}_{3\times3}C_{3\times4}+\ldots+Q_{3\times3}C_{3\times4}+C_{3\times4} \\ Z_{4\times3} & I_{4\times4} \end{pmatrix} = \begin{pmatrix} Q^{n}_{3\times3} & (Q^{n-1}_{3\times3}+Q^{n-2}_{3\times3}+\ldots+Q_{3\times3}+I_{3\times3})C_{3\times4} \\ Z_{4\times3} & I_{4\times4} \end{pmatrix}.
$$

This expression can be simplified by applying the following Corollary.

Corollary 2.1

$$(Q^{n-1}+Q^{n-2}+\ldots Q+I)C=(I-Q)^{-1}(I-Q^{n})C$$

Proof:

$$
\begin{aligned}
&(I-Q)(I+Q+Q^{2}+Q^{3}+\ldots+Q^{n-1})\\
&= I+Q+Q^{2}+Q^{3}+\ldots+Q^{n-1}-(Q+Q^{2}+Q^{3}+\ldots+Q^{n-1}+Q^{n})\\
&= I+Q+Q^{2}+Q^{3}+\ldots+Q^{n-1}-Q-Q^{2}-Q^{3}-\ldots-Q^{n-1}-Q^{n}\\
&= I-Q^{n}
\end{aligned}
$$

$$
TPM^{n}_{7\times7} = \begin{pmatrix} Q^{n}_{3\times3} & (Q^{n-1}_{3\times3}+Q^{n-2}_{3\times3}+\ldots+Q_{3\times3}+I_{3\times3})C_{3\times4} \\ Z_{4\times3} & I_{4\times4} \end{pmatrix} = \begin{pmatrix} Q^{n}_{3\times3} & (I_{3\times3}-Q_{3\times3})^{-1}(I_{3\times3}-Q^{n}_{3\times3})C_{3\times4} \\ Z_{4\times3} & I_{4\times4} \end{pmatrix}
$$

$$
\lim_{n\to\infty} TPM^{n}_{7\times7} = \begin{pmatrix} \lim_{n\to\infty} Q^{n}_{3\times3} & \lim_{n\to\infty}(I_{3\times3}-Q_{3\times3})^{-1}(I_{3\times3}-Q^{n}_{3\times3})C_{3\times4} \\ \lim_{n\to\infty} Z_{4\times3} & \lim_{n\to\infty} I_{4\times4} \end{pmatrix} = \begin{pmatrix} Z_{3\times3} & (I_{3\times3}-Q_{3\times3})^{-1}C_{3\times4} \\ Z_{4\times3} & I_{4\times4} \end{pmatrix} = \begin{pmatrix} Z_{3\times3} & B_{3\times4} \\ Z_{4\times3} & I_{4\times4} \end{pmatrix}
$$

where $B_{3\times4}=(I_{3\times3}-Q_{3\times3})^{-1}C_{3\times4}$.

$\lim_{n\to\infty} Q_{3\times3}^n = Z_{3\times3}$, because the elements of $Q_{3\times3}^n$ are the transition probabilities that the chart will go from a non-absorbent state to a non-absorbent state in n transitions. These probabilities will tend to 0 as n tends to infinity, because once the system has moved from a non-absorbent state to an absorbent state, that absorbent state can't be left, i.e. the system will not be able to move back to a non-absorbent state.

Recall that we are interested in the matrix B, because the $(i, j)^{th}$ element of B is the long run proportion of times the j^{th} set of runs rules causes the chart to signal given the chart starts in a non-absorbing state i.

$$B_{3\times4} = (I_{3\times3} - Q_{3\times3})^{-1} C_{3\times4}$$

$$= \left[\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} p_3 & p_2 & p_4 \\ p_3 & 0 & p_4 \\ p_3 & p_2 & 0 \end{pmatrix}\right]^{-1} \begin{pmatrix} p_1 & 0 & 0 & p_5 \\ p_1 & p_2 & 0 & p_5 \\ p_1 & 0 & p_4 & p_5 \end{pmatrix}$$

$$= \begin{pmatrix} 1-p_3 & -p_2 & -p_4 \\ -p_3 & 1 & -p_4 \\ -p_3 & -p_2 & 1 \end{pmatrix}^{-1} \begin{pmatrix} p_1 & 0 & 0 & p_5 \\ p_1 & p_2 & 0 & p_5 \\ p_1 & 0 & p_4 & p_5 \end{pmatrix}$$

$$= \begin{pmatrix} b_{11} & b_{12} & b_{13} & b_{14} \\ b_{21} & b_{22} & b_{23} & b_{24} \\ b_{31} & b_{32} & b_{33} & b_{34} \end{pmatrix}.$$

The effort in inverting $(I - Q)$ could be substantial and therefore some type of statistical software package is desirable. Using Mathcad's **Symbolics** $\rightarrow$ **Evaluate** $\rightarrow$ **Symbolically** we can easily calculate the inverse of $(I - Q)$ and multiply the two matrices, $(I - Q)^{-1}$ and C, to get an expression for the matrix B. The long run signal probabilities are given by b_{11}, b_{12}, b_{13}, b_{14}, b_{21}, b_{22}, b_{23}, b_{24}, b_{31}, b_{32}, b_{33} and b_{34}. Since these are all very long expressions, only one will be given and explained: $b_{11} = \dfrac{p_1(1+p_2)(1+p_4)}{1 - p_2p_4 - p_3 - p_2p_3p_4 - p_2p_3 - p_3p_4}$ is the long run proportion of times that the runs rule R_1 causes the chart to signal when the chart starts in state 1. In general: b_{ij} is the long run proportion of times that the runs rule R_j causes the chart to signal when the chart starts in state i.

2.2.5. Summary

The necessary steps for calculating the probability that any subset of runs rules will give an out-of -control signal:

STEP 1: Classification of states:

- State number
- Description of state
- Absorbent (A) / Non-absorbent (NA)

STEP 2: Setting up the transition probability matrix $TPM = [p_{ij}]$

STEP 3: Partitioning of the transition probability matrix into 4 sections $TPM = [p_{ij}] = \begin{pmatrix} Q & C \\ Z & I \end{pmatrix}$

- $Q : (NA \rightarrow NA)$
- $C : (NA \rightarrow A)$
- $Z : (A \rightarrow NA)$
- $I : (A \rightarrow A)$

STEP 4: Obtain $(I - Q)^{-1}$

STEP 5: Calculate $B = (I - Q)^{-1} C$

STEP 6: Interpret B. b_{ij} is the long run proportion of times that the runs rule R_j causes the chart to signal when the chart starts in state i.

2.3. The tabular CUSUM control chart

2.3.1. Introduction

Cumulative sum (or CUSUM) control charts were first introduced by Page (1954) (although not in its present form) and have been studied by many authors, for example, Barnard (1959), Ewan and Kemp (1960), Johnson (1961), Goldsmith and Whitfield (1961), Page (1961), Ewan (1963), Van Dobben de Bruyn (1968), Woodall and Adams (1993) and Hawkins and Olwell (1998). Montgomery (2005) related CUSUM ideas to other SPC methodologies.

The statistical design of CUSUM charts

While the Shewhart-type charts are widely known and most often used in practice because of their simplicity and global performance, other classes of charts, such as the CUSUM charts are useful and sometimes more naturally appropriate in the process control environment in view of the sequential nature of data collection. The CUSUM chart incorporates all the information in the sequence of sample values by plotting a function of the cumulative sums of the deviations of the sample values from a target value. For example, suppose that samples of size $n=1$ are collected and let x_j denote the j^{th} observation. The case of individual observations occurs very often in practice, so that situation will be treated first. Later we will see how to modify these results for subgroups. Then if θ_0 is the target value, the CUSUM chart is formed by plotting C_i where

$$C_i = \sum_{j=1}^{i}(x_j - \theta_0) = (x_i - \theta_0) + \sum_{j=1}^{i-1}(x_j - \theta_0) = (x_i - \theta_0) + C_{i-1}.$$

The upper one-sided CUSUM works by accumulating deviations from $\theta_0 + K$ that are above target. For the upper one-sided CUSUM chart we use

$$C_i^+ = \max[0, C_{i-1}^+ + (x_i - \theta_0) - K] \quad \text{for } i = 1,2,3,... \tag{2.31}$$

to detect positive deviations from θ_0. A signaling event occurs for the first i such that $C_i^+ \geq H$.

The lower one-sided CUSUM works by accumulating deviations from $\theta_0 - K$ that are below target. For the lower one-sided CUSUM chart we use

$$C_i^- = \min[0, C_{i-1}^- + (x_i - \theta_0) + K] \quad \text{for } i = 1,2,3,... \tag{2.32}$$

or

$$C_i^{-*} = \max[0, C_{i-1}^{-*} - (x_i - \theta_0) - K] \quad \text{for } i = 1,2,3,... \tag{2.33}$$

to detect negative deviations from θ_0. A signaling event occurs for the first i such that $C_i^- \leq -H$ (if expression (2.32) is used) or $C_i^{-*} \geq H$ (if expression (2.33) is used). For a visually appealing chart, expression (2.32) will be used to construct the lower one-sided CUSUM.

The two-sided CUSUM chart signals for the first i at which either one of the two inequalities is satisfied, that is, either $C_i^+ \geq H$ or $C_i^- \leq -H$. Both K and H are non-negative integers and they are needed in order to implement the CUSUM chart. Details regarding how to choose these constants are given in Section 2.3.1 in the sub-section called *Recommendations for the design of the CUSUM control chart*.

Note that both C_i^+ and C_i^- accumulate deviations from the target value θ_0 that are greater than K. Originally, Page (1954) set the starting values equal to zero, that is, $C_0^+ = 0$ and $C_0^- = 0$. Later on, Lucas and Crosier (1982) recommended setting the starting values equal to some nonzero value to improve the sensitivity of the CUSUM at process start-up. This is referred to as the fast initial response (FIR) or head start feature.

The standardized CUSUM

The variable x_i can be standardized by subtracting its mean and dividing by its standard deviation, that is,

$$y_i = \frac{(x_i - \theta_0)}{\sigma}. \tag{2.34}$$

The resulting standardized upper one-sided CUSUM is given by

$$S_i^+ = \max[0, S_{i-1}^+ + y_i - k] \quad \text{for } i = 1,2,3,... \tag{2.35}$$

while the resulting standardized lower one-sided CUSUM is given by

$$S_i^- = \min[0, S_{i-1}^- + y_i + k] \quad \text{for } i = 1,2,3,... \tag{2.36}$$

or

$$S_i^{-*} = \max[0, S_{i-1}^{-*} - y_i - k] \quad \text{for } i = 1,2,3,... \tag{2.37}$$

The two-sided standardized CUSUM is constructed by running the upper and lower one-sided standardized CUSUM charts simultaneously and signals at the first i such that $S_i^+ \geq H_s$ or $S_i^- \leq -H_s$*. Both k and H_s are non-negative integers and they are needed in order to implement the standardized CUSUM chart. As mentioned previously, details regarding how to choose these constants are given in Section 2.3.1 in the sub-section called *Recommendations for the design of the CUSUM control chart*.

The unstandardized CUSUM C_i and the standardized CUSUM S_i contains the same information. The question arises: Should unstandardized or standardized data be used? Unstandardized data has the advantage that the units of the vertical axis are in their original measurements which makes interpretation easier. Standardized data has the advantage that different CUSUM charts can be compared.

The CUSUM for monitoring the process mean and other sample statistics

A CUSUM chart for monitoring the process mean can be obtained by replacing x_i in expression (2.34) with the sample average $\bar{x}_i$ and by replacing σ by $\sigma/\sqrt{n}$. It is also possible to develop CUSUM charts for other sample statistics, for example, standard deviations and defects. These CUSUM charts for other sample statistics have been studied by many authors, for example, Lucas (1985), Gan (1993) and White, Keats and Stanley (1997).

Recommendations for the design of the CUSUM control chart

Phase II CUSUM charts should be designed on the basis of *ARL* performance. The parameters K and H are obtained for a specified in-control average run length. Both

---

* The vertical axis of the standardized CUSUM will be measured in multiples of the standard deviation (σ) of the data, whereas the vertical axis of the unstandardized data will be measured in the same units of *X*, for example, in meters, millimeters, ect. To avoid confusion, H and H_s will be used to denote the decision intervals for the unstandardized and standardized CUSUM charts, respectively.

parameters are non-negative integers. Let σ denote the standard deviation of the sample variable used in forming the cumulative sum. Parametric CUSUM charts (see Page (1954)) are used for detecting shifts in a normal mean based on the cumulative sum of differences from target. Let $H = h\sigma$ and $K = k\sigma$ where h is usually taken to be equal to 4 or 5 and k is usually taken to equal 0.5 (see Montgomery (2005) page 395). By choosing $h = 4$ or $h = 5$ and $k = 0.5$ (the values most commonly used in practice) we generally get a good average run length performance for parametric CUSUM charts. In the next section we will show that choosing $h = 4$ or $h = 5$ and $k = 0.5$ is not recommended for nonparametric CUSUM charts, since it usually gives a poor in-control average run length performance. Since we will not be using $H = 4\sigma$ or 5σ and $K = 0.5\sigma$ for nonparametric control charts, we will denote the decision interval and reference value by h and k, respectively, from this point forward.

The proposed nonparametric CUSUM chart

Amin, Reynolds and Bakir (1995) proposed a nonparametric CUSUM chart for the median (or any other percentile) of any continuous population based on sign statistics. Recall that for the i^{th} random sample the plotting statistic in the Shewhart-type chart was $SN_i = \sum_{j=1}^{n} sign(x_{ij} - \theta_0)$. The chart proposed by Amin et al. (1995) instead uses the cumulative sum of the statistic SN_i with a stopping rule. They also calculated the *ARL* of the chart using a Markov chain approach where the transition probabilities are calculated via the distribution of the sign statistic, which is of course binomial. The procedure is distribution-free since the in-control distribution of SN_i does not depend on the underlying distribution for all continuous distributions. A CUSUM sign chart can be obtained by replacing y_i in expressions (2.35), (2.36) and (2.37) with SN_i. In other words, for the upper one-sided CUSUM sign chart we use

$$S_i^+ = \max[0, S_{i-1}^+ + SN_i - k] \quad \text{for } i = 1,2,3,... \qquad (2.38)$$

to detect positive deviations from the known target value θ_0. A signaling event occurs for the first i such that $S_i^+ \geq h$.

For a lower one-sided CUSUM sign chart we use

$$S_i^- = \min[0, S_{i-1}^- + SN_i + k] \quad \text{for } i = 1,2,3,... \qquad (2.39)$$

or

$$S_i^{-*} = \max[0, S_{i-1}^{-*} - SN_i - k] \quad \text{for } i = 1,2,3,... \tag{2.40}$$

to detect negative deviations from the known target value θ_0. A signaling event occurs for the first i such that $S_i^- \leq -h$ (if expression (2.39) is used) or $S_i^{-*} \geq h$ (if expression (2.40) is used).

The corresponding two-sided CUSUM chart signals for the first i at which either one of the two inequalities is satisfied, that is, either $S_i^+ \geq h$ or $S_i^- \leq -h$. Starting values are typically chosen to equal zero, that is, $S_0^+ = S_0^- = 0$.

The constants k and h are obtained for a specified in-control average run length. In-control average run length (ARL_0), standard deviation of the run length ($SDRL_0$), 5^{th}, 25^{th} (the first quartile, Q_1), 50^{th} (the median run length, MRL_0), 75^{th} (the third quartile, Q_3) and 95^{th} percentile values will be computed and tabulated for various values of h and k later on.

2.3.2. One-sided control charts

2.3.2.1. Upper one-sided control charts

Various expressions for the exact run length distribution and its parameters have been given for the normal theory one-sided CUSUM procedure by, for example, Ewan and Kemp (1960), Brook and Evans (1972), Woodall (1983) and Hawkins and Olwell (1998). Many authors have presented various approximations for the run length distribution and its parameters for the one-sided CUSUM procedure. A Markov chain representation of the one-sided CUSUM procedure based on integer-valued cumulative sums is presented in this section. The number of states included in the Markov chain is minimized in order to make the methods as efficient as possible.

Markov chain approach

Brook and Evans (1972) and Amin et al. (1995) considered a method for evaluating the exact average run length and its moments for the upper one-sided CUSUM chart by treating the cumulative sum as a Markov chain with the state space a subset of $\{0,1,2,...,h\}$. Markov techniques have a great advantage as they are adjustable to many runs related problems and they often simplify the solutions to the specific problems they are applied on. Fu, Spiring and Xie (2002) presented three results that must be satisfied before implementing the finite-state Markov chain approach.

Let S_t^+ be a finite-state homogenous Markov chain on the state space Ω^+ with a TPM such that (i) $\Omega^+ = \{\varsigma_0, \varsigma_1, ..., \varsigma_{r+s-1}\}$ where $0 = \varsigma_0 < \varsigma_1 < ... < \varsigma_{r+s-1} = h$ and ς_{r+s-1} is an absorbent state; (ii) the TPM is given by $TPM = [p_{ij}]$ for $i = 0,1,...,r+s-1$ and $j = 0,1,...,r+s-1$ where r denotes the number of non-absorbent states and s the number of absorbent states, respectively, and (iii) the starting value should be in the "dummy" state with probability one, that is, $P(S_0^+ = \varsigma_0) = 1$, to ensure the process starts in-control . Assume that the Markov chain S_t^+ satisfies conditions (i), (ii) and (iii), then from Fu, Spiring and Xie (2002) and Fu and Lou (2003) we have

$$P(N = n \mid S_0^+ = 0) = \underline{\xi} Q^{n-1}(I-Q)\underline{1} \tag{2.41}$$

$$E(N) = \underline{\xi}(I-Q)^{-1}\underline{1} \tag{2.42}$$

$$E(N^2) = \underline{\xi}(I+Q)(I-Q)^{-2}\underline{1} \tag{2.43}$$

$$\text{var}(N) = E(N^2) - (E(N))^2 = \underline{\xi}(I+Q)(I-Q)^{-2}\underline{1} - \left(\underline{\xi}(I-Q)^{-1}\underline{1}\right)^2 \tag{2.44}$$

$$SDRL = \sqrt{\text{var}(N)} = \sqrt{\underline{\xi}(I+Q)(I-Q)^{-2}\underline{1} - \left(\underline{\xi}(I-Q)^{-1}\underline{1}\right)^2} \tag{2.45}$$

where the essential transition probability sub-matrix Q is the $r \times r$ matrix that contains all the transition probabilities of going from a non-absorbent state to a non-absorbent state, I is the $r \times r$ identity matrix, $\underline{\xi}$ is a $1 \times r$ row vector with 1 at the 1^{st} element and zero elsewhere and $\underline{1}$ is an $r \times 1$ column vector with all elements equal to unity. See Theorem 2 in Appendix A for the derivations done by Fu, Spiring and Xie (2002) and Fu and Lou (2003).

The time that the procedure signals is the first time such that the finite-state Markov chain S_t^+ enters one of the absorbent states where the state space is given by $\Omega^+ = \{\varsigma_0, \varsigma_1, ..., \varsigma_{r+s-1}\}$, $S_0^+ = 0$ and

$$S_t^+ = \min\{h, \max\{0, S_{t-1}^+ + SN_t - k\}\}. \tag{2.46}$$

The state corresponding to a signal by the CUSUM chart is called an absorbent state. Clearly, there is only one absorbent state, since the chart signals when S_t^+ falls on or above *h*, i.e. $s = 1$.

The distribution of SN_t can easily be obtained from the binomial distribution (recall that $SN_i = 2T_i - n \quad \forall i$, where T_i is binomially distributed with parameters n and $p = P(X_{ij} \geq \theta_0))$. The binomial probabilities are given in Table G of Gibbons and Chakraborti (2003) and can easily be calculated using some type of statistical software package, for example, Excel or SAS.

Example 2.6

An upper one-sided CUSUM sign chart where the sample size is odd ($n = 5$)

The statistical properties of an upper one-sided CUSUM sign chart with a decision interval of 4 ($h = 4$), a reference value of 1 ($k = 1$) and a sample size of 5 ($n = 5$) is examined. For n odd, the reference value is taken to be odd, because this leads to the sum $\sum(SN_i - k)$ being equal to even values which reduces the size of the state space for the Markov chain. This will halve the size of the matrices of transition probabilities. For $h = 4$ we have that the state space is $\Omega^+ = \{\varsigma_0, \varsigma_1, \varsigma_2\} = \{0,2,4\}$ with $0 = \varsigma_0 < \varsigma_1 < \varsigma_2 = h$. The state space is calculated using equation (2.46) and the calculations are shown in Table 2.18.

Table 2.18. Calculation of the state space when $h = 4$, $k = 1$ and $n = 5$.

SN_t	$S_{t-1}^{+} + SN_t - k$	$\max\{0, S_{t-1}^{+} + SN_t - k\}$	$S_t^{+} = \min\{h, \max\{0, S_{t-1}^{+} + SN_t - k\}\}$
-5	-6*	0	0
-3	-4	0	0
-1	-2	0	0
1	0	0	0
3	2	2	2
5	4	4	4

Table 2.19. Classifications and descriptions of the states.

State number	Description of the state	Absorbent (A)/ Non-absorbent (NA)
0	$S_t^{+} = 0$	NA
1	$S_t^{+} = 2$	NA
2	$S_t^{+} = 4$	A

From Table 2.19 we see that there are two non-absorbent states, i.e. $r = 2$, and one absorbent state, i.e. $s = 1$. Therefore, the corresponding TPM will be a $(r+s)\times(r+s) = 3\times 3$ matrix. It can be shown (see Table 2.20) that the TPM is given by

$$TPM_{3\times 3} = \begin{pmatrix} p_{00} & p_{01} & p_{02} \\ p_{10} & p_{11} & p_{12} \\ p_{20} & p_{21} & p_{22} \end{pmatrix} = \begin{pmatrix} 26/32 & 5/32 & | & 1/32 \\ 16/32 & 10/32 & | & 6/32 \\ - & - & - & - \\ 0 & 0 & | & 1 \end{pmatrix} = \begin{pmatrix} Q_{2\times 2} & | & \underline{p}_{2\times 1} \\ - & - & - \\ \underline{0}'_{1\times 2} & | & 1_{1\times 1} \end{pmatrix}$$

where the essential transition probability sub-matrix $Q_{2\times 2} : (NA \rightarrow NA)$ is an $r \times r = 2\times 2$ matrix, $\underline{p}_{2\times 1} : (NA \rightarrow A)$ is an $(r+s-1)\times 1 = 2\times 1$ column vector, $\underline{0}'_{1\times 2} : (A \rightarrow NA)$ is a $1\times(r+s-1) = 1\times 2$ row vector and $1_{1\times 1} : (A \rightarrow A)$ represents the scalar value one.

The one-step transition probabilities are calculated by substituting SN_t in expression (2.46) by $2T - n$ and substituting in values for h, k, S_t^{+} and S_{t-1}^{+}. The calculation of the one-step transition probabilities are given in Table 2.20 for illustration.

The probabilities in the last column of the TPM can also be calculated using the fact that $\sum_{j\in\Omega} p_{ij} = 1 \;\; \forall i$ (see equation (2.18)). Therefore,

---

* Note: Since only the state space needs to be described, S_{t-1}^{+} can be any value from Ω^{+} and we therefore take, without loss of generality, $S_{t-1}^{+} = 0$. Any other possible value for S_{t-1}^{+} would lead to the same Ω^{+}.

$p_{02} = 1 - (p_{00} + p_{01}) = 1 - (26/32 + 5/32) = 1/32$;

$p_{12} = 1 - (p_{10} + p_{11}) = 1 - (16/32 + 10/32) = 6/32$; and

$p_{22} = 1 - (p_{20} + p_{21}) = 1 - (0 + 0) = 1$.

Since it is easier to calculate the probabilities in the last column of the TPM using the latter approach, it will be used throughout the text from this point forward.

Table 2.20. The calculation of the transition probabilities when $h = 4$, $k = 1$ and $n = 5$.

p_{00} $= P(S_t = 0 \mid S_{t-1} = 0)$ $= P(\min\{4, \max\{0, 0 + SN_t - 1\}\} = 0)$ $= P(\max\{0, 0 + SN_t - 1\} = 0)$ $= P(SN_t - 1 \le 0)$ $= P(SN_t \le 1)$ $= P(2T - 5 \le 1)$ $= P(T \le 3)$ $= 26/32$	p_{01} $= P(S_t = 2 \mid S_{t-1} = 0)$ $= P(\min\{4, \max\{0, 0 + SN_t - 1\}\} = 2)$ $= P(\max\{0, SN_t - 1\} = 2)$ $= P(SN_t - 1 = 2)$ $= P(SN_t = 3)$ $= P(2T - 5 = 3)$ $= P(T = 4)$ $= 5/32$	p_{02} $= P(S_t = 4 \mid S_{t-1} = 0)$ $= P(\min\{4, \max\{0, 0 + SN_t - 1\}\} = 4)$ $= P(\max\{0, SN_t - 1\} \ge 4)$ $= P(SN_t - 1 \ge 4)$ $= P(SN_t \ge 5)$ $= P(2T - 5 \ge 5)$ $= 1 - P(T \le 4)$ $= 1/32$
p_{10} $= P(S_t = 0 \mid S_{t-1} = 2)$ $= P(\min\{4, \max\{0, 2 + SN_t - 1\}\} = 0)$ $= P(\max\{0, SN_t + 1\} = 0)$ $= P(SN_t \le -1)$ $= P(2T - 5 \le -1)$ $= P(T \le 2)$ $= 16/32$	p_{11} $= P(S_t = 2 \mid S_{t-1} = 2)$ $= P(\min\{4, \max\{0, 2 + SN_t - 1\}\} = 2)$ $= P(\max\{0, SN_t + 1\} = 2)$ $= P(SN_t = 1)$ $= P(2T - 5 = 1)$ $= P(T = 3)$ $= 10/32$	p_{12} $= P(S_t = 4 \mid S_{t-1} = 2)$ $= P(\min\{4, \max\{0, 2 + SN_t - 1\}\} = 4)$ $= P(\max\{0, SN_t + 1\} \ge 4)$ $= P(SN_t \ge 3)$ $= P(2T - 5 \ge 3)$ $= 1 - P(T \le 3)$ $= 6/32$
p_{20} $= P(S_t = 0 \mid S_{t-1} = 4)$ $= 0$ *	p_{21} $= P(S_t = 2 \mid S_{t-1} = 4)$ $= 0$	p_{22} $= P(S_t = 4 \mid S_{t-1} = 4)$ $= 1$ †

Using the TPM the ARL can be calculated using $ARL = \underline{\xi}(I - Q)^{-1}\underline{1}$. A well-known concern is that important information about the performance of a control chart can be missed when only examining the ARL (this is especially true when the process distribution is skewed). Various authors, see for example, Radson and Boyd (2005) and Chakraborti (2007), have suggested that one should examine a number of percentiles, including the median, to get the complete information about the performance of a control chart. Therefore, we now also consider percentiles. The $100\rho^{th}$ percentile is defined as the smallest integer l such that the

* The probability equals zero, because it is impossible to go from an absorbent state to a non-absorbent state.
† The probability equals one, since the probability of going from an absorbent state to an absorbent state is equal to one (once an absorbent state is entered, it is never left).

cdf is at least $(100\times\rho)\%$. Thus, the $100\rho^{th}$ percentile l is found from $P(N\leq l)\geq\rho$. The median (50^{th} percentile) will be considered, since it is a more representative performance measure than the *ARL* (see the discussion in Section 2.1.5). The first and third quartiles (25^{th} and 75^{th} percentiles) will also be considered, since it contains the middle half of the distribution. The 'tails' of the distribution should also be examined and therefore the 5^{th} and 95^{th} percentiles are calculated. The calculation of these percentiles is shown in Table 2.21 for illustration purposes. The first column of Table 2.21 contains the values that the run length variable (N) can take on.

Table 2.21. Calculation of the percentiles when $h=4$, $k=1$ and $n=5$[*].

N	$P(N\leq l)$	The 5^{th}, 25^{th}, 50^{th}, 75^{th} and 95^{th} percentiles
1	0.0313	
2	0.0859	$\rho_{0.05}=2$ (smallest integer such that the cdf is at least 0.05)
3	0.1420	
4	0.1954	
5	0.2456	
6	0.2928	$\rho_{0.25}=6$ (smallest integer such that the cdf is at least 0.25)
7	0.3370	
8	0.3784	
9	0.4173	
10	0.4537	
11	0.4878	
12	0.5198	$\rho_{0.5}=12$ (smallest integer such that the cdf is at least 0.5)
13	0.5499	
14	0.5780	
15	0.6044	
16	0.6291	
17	0.6523	
18	0.6740	
19	0.6944	
20	0.7135	
21	0.7314	
22	0.7482	
23	0.7639	$\rho_{0.75}=23$ (smallest integer such that the cdf is at least 0.75)
24	0.7787	
25	0.7925	
26	0.8055	
27	0.8176	
28	0.8290	
29	0.8397	
30	0.8497	
⋮	⋮	
48	0.9530	$\rho_{0.95}=48$ (smallest integer such that the cdf is at least 0.95)
49[†]	0.9559	

[*] See SAS Program 2 in Appendix B for the calculation of the values in Table 2.21.
[†] The value of the run length variable is only shown for some values up to *N*=49 for illustration purposes.

The formulas of the moments and some characteristics of the run length distribution have been studied by Fu, Spiring and Xie (2002) and Fu and Lou (2003) – see equations (2.41) to (2.45). By substituting $\underline{\xi}_{1\times 2} = (1 \quad 0)$, $Q_{2\times 2} = \frac{1}{32}\begin{pmatrix} 26 & 5 \\ 16 & 10 \end{pmatrix}$ and $\underline{1}_{2\times 1} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ into these equations, we obtain the following:

$$ARL = E(N) = \underline{\xi}(I - Q)^{-1}\underline{1} = 16.62$$

$$E(N^2) = \underline{\xi}(I + Q)(I - Q)^{-2}\underline{1} = 516.59$$

$$SDRL = \sqrt{Var(N)} = \sqrt{E(N^2) - (E(N))^2} = 15.51$$

$$5^{th} \text{ percentile} = \rho_{0.05} = 2$$

$$25^{th} \text{ percentile} = \rho_{0.25} = 6$$

$$\text{Median} = 50^{th} \text{ percentile} = \rho_{0.5} = 12$$

$$75^{th} \text{ percentile} = \rho_{0.75} = 23$$

$$95^{th} \text{ percentile} = \rho_{0.95} = 48$$

Other values of *h*, *k* and *n* were also considered and the results are given in Table 2.22.

Table 2.22. The in-control average run length (ARL_0^+), standard deviation of the run length ($SDRL$), $5^{th}, 25^{th}, 50^{th}, 75^{th}$ and 95^{th} percentile* values for the upper one-sided CUSUM sign chart when $n = 5$†.

k	*h*	
	2	**3 or 4**
1	5.33 4.81 (1, 2, 4, 7, 15)	16.62 15.51 (2, 6, 12, 23, 48)
3	32.00 31.50 (2, 10, 22, 44, 95)	‡

* The three rows of each cell shows the ARL_0^+, the $SDRL$, and the percentiles ($\rho_5, \rho_{25}, \rho_{50}, \rho_{75}, \rho_{95}$), respectively.

† See SAS Program 2 in Appendix B for the calculation of the values in Table 2.22.

‡ Since the decision interval is taken to satisfy $h \le n - k$ there are open cells in Table 2.22.

Note that the summary measures for odd values of h will be equal to the summary measures of the subsequent even integer. More on this later (refer to example (2.8)). Values of k and h are restricted to be integers so that the Markov chain approach could be employed to obtain expressions for the exact run length distribution and its parameters. In order to allow for the possibility of stopping after one sample, i.e. issuing a signal, the values of h is taken to satisfy $h \leq n - k$.

The five percentiles (given in Table 2.22) are displayed in boxplot-like graphs for various h and k values in Figure 2.5. It should be noted that these boxplot-like graphs differ from standard box plots. In the latter case the whiskers are drawn from the ends of the box to the smallest and largest values inside specified limits, whereas, in the case of the boxplot-like graphs, the whiskers are drawn from the ends of the box to the 5th and 95th percentiles, respectively. In this thesis "boxplot" will refer to a boxplot-like graph from this point forward.

Figure 2.5 clearly shows the effects of h and k on the run length distribution and it portrays the run length distribution when the process is in-control. We would prefer a "boxplot" with a high valued (large) in-control average run length and a small spread. Applying this criterion, we see that the "boxplot" corresponding to the $(h,k) = (2,3)$ combination has the largest in-control average run length, which is favorable, but it also has the largest spread which is unattractive. The "boxplot" furthest to the right is exactly opposite from the "boxplot" furthest to the left. The latter has the smallest spread, which is favorable, but it also has the smallest in-control average run length, which is unattractive. In conclusion, no "boxplot" is optimal relative to the others.

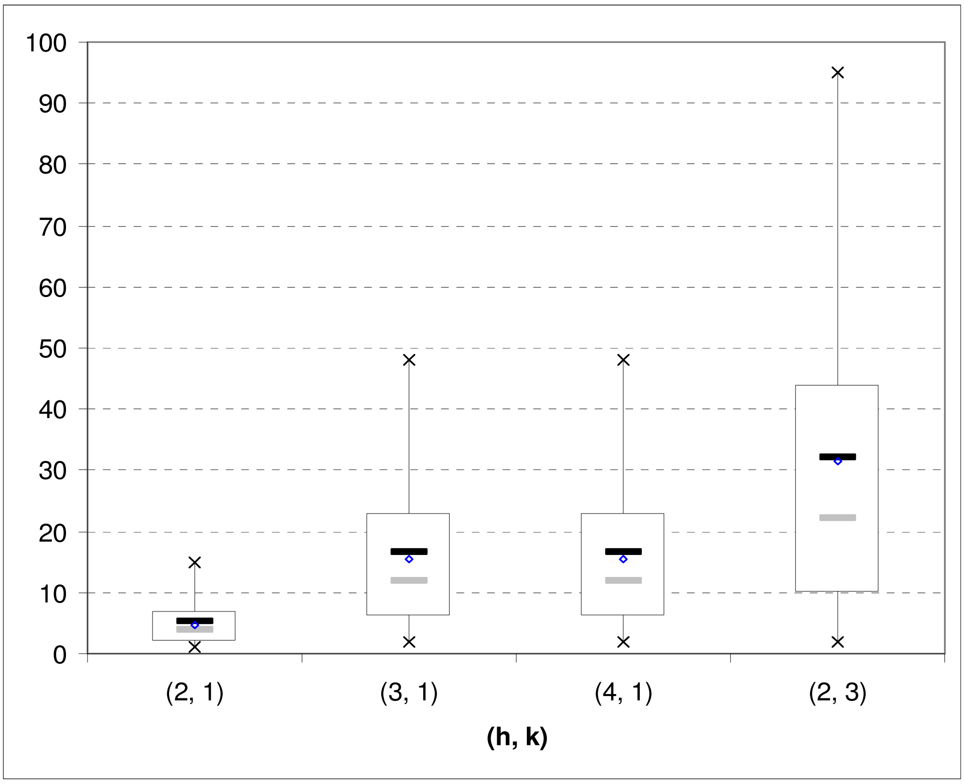

Figure 2.5. Boxplot-like graphs for the in-control run length distribution of various upper one-sided CUSUM sign charts when $n = 5$. The whiskers extend to the 5th and the 95th percentiles. The symbols “▬”, “◇” and “▬” denote the *ARL*, *SDRL** and *MRL*, respectively.

Example 2.7

An upper one-sided CUSUM sign chart where the sample size is even (*n*=6)

The statistical properties of an upper one-sided CUSUM sign chart with a decision interval of 4 ($h = 4$), a reference value of 2 ($k = 2$) and a sample size of 6 ($n = 6$) is examined. For n even, the reference value is taken to be even, because this leads to the sum $\sum(SN_i - k)$ being equal to even values which reduces the size of the state space for the Markov chain. For $h = 4$ we have that the state space is $\Omega^+ = \{\varsigma_0, \varsigma_1, \varsigma_2\} = \{0,2,4\}$ with $0 = \varsigma_0 < \varsigma_1 < \varsigma_2 = h$. The state space is calculated using equation (2.46) and the calculations are shown in Table 2.23.

* For ease of interpretation, the standard deviation (as measure of spread) is included in the (location) measures of percentiles.

Table 2.23. Calculation of the state space when $h=4$, $k=2$ and $n=6$.

SN_t	$S_{t-1}^{+}+SN_t-k$	$\max\{0,S_{t-1}^{+}+SN_t-k\}$	$S_t^{+}=\min\{h,\max\{0,S_{t-1}^{+}+SN_t-k\}\}$
-6	-8*	0	0
-4	-6	0	0
-2	-4	0	0
0	-2	0	0
2	0	0	0
4	2	2	2
6	4	4	4

Table 2.24. Classifications and descriptions of the states.

State number	Description of the state	Absorbent (A)/ Non-absorbent (NA)
0	$S_t^{+}=0$	NA
1	$S_t^{+}=2$	NA
2	$S_t^{+}=4$	A

From Table 2.24 we see that there are two non-absorbent states, i.e. $r=2$, and one absorbent state, i.e. $s=1$. Therefore, the corresponding TPM will be a $(r+s)\times(r+s)=3\times3$ matrix. It can be shown (see Table 2.25) that the TPM is given by

$$TPM_{3\times3}=\begin{pmatrix} p_{00} & p_{01} & p_{02} \\ p_{10} & p_{11} & p_{12} \\ p_{20} & p_{21} & p_{22} \end{pmatrix}=\begin{pmatrix} 57/64 & 6/64 & | & 1/64 \\ 42/64 & 15/64 & | & 7/64 \\ - & - & - & - \\ 0 & 0 & | & 1 \end{pmatrix}=\begin{pmatrix} Q_{2\times2} & | & \underline{p}_{2\times1} \\ - & - & - \\ \underline{0}'_{1\times2} & | & 1_{1\times1} \end{pmatrix}$$

where the essential transition probability sub-matrix $Q_{2\times2}:(NA\to NA)$ is an $r\times r=2\times2$ matrix, $\underline{p}_{2\times1}:(NA\to A)$ is an $(r+s-1)\times1=2\times1$ column vector, $\underline{0}'_{1\times2}:(A\to NA)$ is a $1\times(r+s-1)=1\times2$ row vector and $1_{1\times1}:(A\to A)$ represents the scalar value one. The one-step transition probabilities are calculated by substituting SN_t in expression (2.46) by $2T-n$ and substituting in values for h, k, S_t^{+} and S_{t-1}^{+}. The calculation of the one-step transition probabilities are given in Table 2.25 for illustration.

---

* Note: Since only the state space needs to be described, S_{t-1}^{+} can be any value from Ω^{+} and we therefore take, without loss of generality, $S_{t-1}^{+}=0$. Any other possible value for S_{t-1}^{+} would lead to the same Ω^{+}.

Table 2.25. The calculation of the transition probabilities when $h=4$, $k=2$ and $n=6$.

p_{00} $=P(S_t=0 \mid S_{t-1}=0)$ $=P(\min\{4,\max\{0,0+SN_t-2\}\}=0)$ $=P(\max\{0,0+SN_t-2\}=0)$ $=P(SN_t-2\leq 0)$ $=P(SN_t\leq 2)$ $=P(T\leq 4)$ $=\frac{57}{64}$	p_{01} $=P(S_t=2 \mid S_{t-1}=0)$ $=P(\min\{4,\max\{0,0+SN_t-2\}\}=2)$ $=P(\max\{0,SN_t-2\}=2)$ $=P(SN_t-2=2)$ $=P(SN_t=4)$ $=P(T=5)$ $=\frac{6}{64}$	p_{02} $=1-(p_{00}+p_{01})$ $=1-(\frac{57}{64}+\frac{6}{64})$ $=\frac{1}{64}$
p_{10} $=P(S_t=0 \mid S_{t-1}=2)$ $=P(\min\{4,\max\{0,2+SN_t-2\}\}=0)$ $=P(\max\{0,SN_t\}=0)$ $=P(SN_t\leq 0)$ $=P(T\leq 3)$ $=\frac{42}{64}$	p_{11} $=P(S_t=2 \mid S_{t-1}=2)$ $=P(\min\{4,\max\{0,2+SN_t-2\}\}=2)$ $=P(\max\{0,SN_t\}=2)$ $=P(SN_t=2)$ $=P(T=4)$ $=\frac{15}{64}$	p_{12} $=1-(p_{10}+p_{11})$ $=1-(\frac{42}{64}+\frac{15}{64})$ $=\frac{7}{64}$
p_{20} $=P(S_t=0 \mid S_{t-1}=4)$ $=0$ *	p_{21} $=P(S_t=2 \mid S_{t-1}=4)$ $=0$	p_{22} $=1-(p_{20}+p_{21})$ $=1$

The formulas of the moments and some characteristics of the run length distribution have been studied by Fu, Spiring and Xie (2002) and Fu and Lou (2003) – see equations (2.41) to (2.45). By substituting $\underline{\xi}_{1\times 2}=(1 \quad 0)$, $Q_{2\times 2}=\frac{1}{64}\begin{pmatrix}57 & 6\\ 42 & 15\end{pmatrix}$ and $\underline{1}_{2\times 1}=\begin{pmatrix}1\\1\end{pmatrix}$ into these equations, we obtain the following:

$$ARL=E(N)=\underline{\xi}(I-Q)^{-1}\underline{1}=38.68$$

$$E(N^2)=\underline{\xi}(I+Q)(I-Q)^{-2}\underline{1}=2918.19$$

$$SDRL=\sqrt{Var(N)}=\sqrt{E(N^2)-(E(N))^2}=37.71$$

$$5^{th}\text{ percentile}=\rho_{0.05}=3$$

$$25^{th}\text{ percentile}=\rho_{0.25}=12$$

$$\text{Median}=50^{th}\text{ percentile}=\rho_{0.5}=27$$

$$75^{th}\text{ percentile}=\rho_{0.75}=53$$

$$95^{th}\text{ percentile}=\rho_{0.95}=114$$

* The probability equals zero, because it is impossible to go from an absorbent state to a non-absorbent state.

Other values of h, k and n were also considered and the results are given in Table 2.26.

Table 2.26. The in-control average run length (ARL_0^+), standard deviation of the run length ($SDRL$), $5^{th}, 25^{th}, 50^{th}, 75^{th}$ and 95^{th} percentile* values for the upper one-sided CUSUM sign chart when $n = 6$†.

k	**h**		
	2	**3 or 4**	**5 or 6**
0	2.91 2.36 (1, 1, 2, 4, 8)	5.92 4.96 (1, 2, 4, 8, 16)	10.66 8.80 (2, 4, 8, 14, 28)
2	9.14 8.63 (1, 3, 6, 12, 26)	38.68 37.71 (3, 12, 27, 53, 114)	‡
4	64.00 63.50 (4, 19, 45, 89, 191)		

The five percentiles (given in Table 2.26) are displayed in boxplot-like graphs for various h and k values in Figure 2.6. Recall that we would prefer a "boxplot" with a high valued (large) in-control average run length and a small spread. Applying this criterion, we see that the "boxplot" corresponding to the $(h,k) = (2,4)$ combination has the largest in-control average run length, which is favorable, but it also has the largest spread which is unattractive. The "boxplot" furthest to the right is exactly opposite from the "boxplot" furthest to the left. The latter has the smallest spread, which is favorable, but it also has the smallest in-control average run length, which is unattractive. In conclusion, no "boxplot" is optimal relative to the others.

* The three rows of each cell shows the ARL_0^+, the $SDRL$, and the percentiles ($\rho_5, \rho_{25}, \rho_{50}, \rho_{75}, \rho_{95}$), respectively.
† See SAS Program 2 in Appendix B for the calculation of the values in Table 2.26.
‡ Since the decision interval is taken to satisfy $h \le n - k$ there are open cells in Table 2.26.

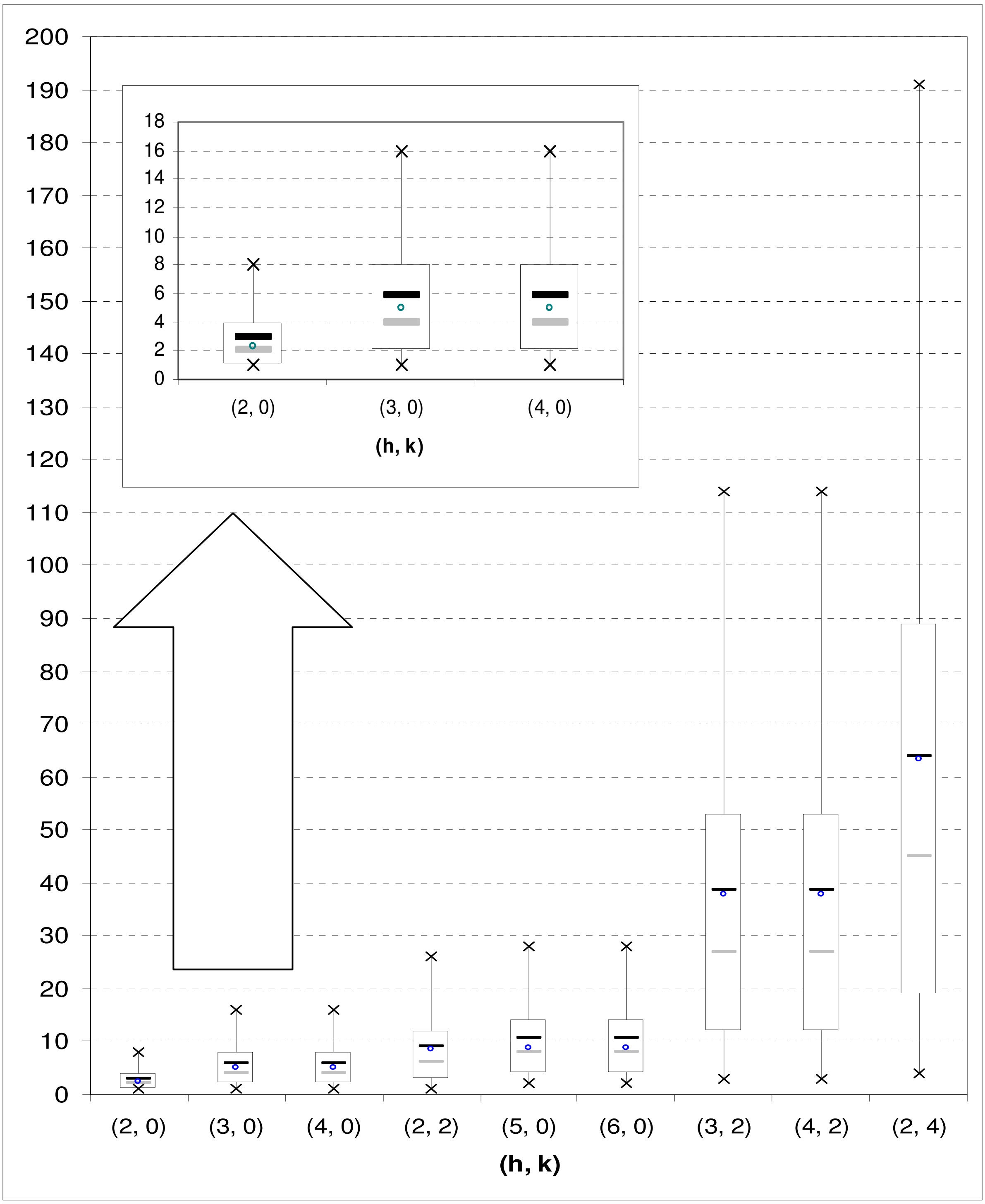

Figure 2.6. Boxplot-like graphs for the in-control run length distribution of various upper one-sided CUSUM sign charts when $n = 6$. The whiskers extend to the 5th and the 95th percentiles. The symbols "▬", "○" and "▬" denote the *ARL*, *SDRL** and *MRL*, respectively.

* For ease of interpretation, the standard deviation (as measure of spread) is included in the (location) measures of percentiles.

On the performance side, note that the largest in-control average run length that the upper one-sided CUSUM sign chart can obtain is 2^n. Therefore, for a sample size of 6 the largest ARL_0^+ equals $2^6 = 64$ (this is obtained when $h = 2$ and $k = 4$). For this case we find the $TPM = \begin{pmatrix} 63/64 & 1/64 \\ 0 & 1 \end{pmatrix}$ and as a result the in-control average run length equals $ARL_0^+ = \underline{\xi}(I - Q)^{-1}\underline{1} = 1 \times (1 - 63/64)^{-1} \times 1 = 64$. Since the largest ARL_0^+ is only 64 for $n = 6$, many false alarms will be expected by this chart leading to a possible loss of time and resources. Larger sample sizes should therefore preferably be taken when implementing the upper one-sided CUSUM sign chart.

....

Example 2.8

An upper one-sided CUSUM sign chart with a decision interval of 4 (*h*=4), a reference value of 1 (*k*=1) and a sample size of 5 (*n*=5)

In the previous two examples it can be seen that summary measures for odd values of h will be equal to the summary measures of the subsequent even integer. This will be illustrated by the use of an example.

For the upper one-sided CUSUM sign chart with a decision interval of 4 ($h = 4$), a reference value of 1 ($k = 1$) and a sample size of 5 ($n = 5$) the TPM was given by $TPM = \begin{pmatrix} 26/32 & 5/32 & 1/32 \\ 16/32 & 10/32 & 6/32 \\ 0 & 0 & 1 \end{pmatrix}$ (see example (2.6). By keeping the reference value and the sample size fixed and changing h to an odd integer ($h = 3$) we obtain the same TPM and therefore we obtain the same summary measures. Stated differently, the summary measures of h odd ($h = 3$) will be equal to the summary measures of the subsequent even integer ($h = 4$).

....

We've considered sample sizes of $n = 5$ and 6 and established that larger sample sizes should preferably be taken when implementing the upper one-sided CUSUM sign chart. Therefore, a larger sample size ($n = 10$) is considered and the results are given in Table 2.27.

Table 2.27. The in-control average run length (ARL_0^+), standard deviation of the run length ($SDRL$), $5^{th}, 25^{th}, 50^{th}, 75^{th}$ and 95^{th} percentile* values for the upper one-sided CUSUM sign chart when $n = 10$ †.

k	***h***		
	3 or 4	**5 or 6**	**7 or 8**
2	14.34 13.58 (1, 5, 10, 20, 41)	36.81 35.48 (3, 12, 26, 51, 108)	91.59 89.45 (7, 28, 64, 126, 270)
4	77.97 77.29 (5, 23, 54, 108, 232)	464.86 463.68 (25, 135, 323, 644, 1390)	‡
6	929.97 929.37 (48, 268, 645, 1289, 2785)		

Table 2.27 gives values of ARL_0^+ for various values of h and k when the sample size is equal to 10. Amin, Bakir and Reynolds (1995) provided a similar Table (see Table 5 on page 1613) containing the in-control run length summary values for the upper one-sided CUSUM sign chart (ARL_0^+) for a range of k and h values when $n = 10$.

The five percentiles (given in Table 2.27) are displayed in boxplot-like graphs for various h and k values in Figure 2.7. Recall that we would prefer a "boxplot" with a high valued (large) in-control average run length and a small spread. Applying this criterion, we see that the "boxplot" corresponding to the $(h,k) = (3,6)$ or $(h,k) = (4,6)$ combination has the largest in-control average run length, which is favorable, but it also has the largest spread which is unattractive. The "boxplot" furthest to the right is exactly opposite from the "boxplot" furthest to the left. The latter has the smallest spread, which is favorable, but it also has the smallest in-control average run length, which is unattractive. In conclusion, no "boxplot" is optimal relative to the others.

* The three rows of each cell shows the ARL_0^+, the $SDRL$, and the percentiles ($\rho_5, \rho_{25}, \rho_{50}, \rho_{75}, \rho_{95}$), respectively.

† See SAS Program 2 in Appendix B for the calculation of the values in Table 2.27.

‡ Since the decision interval is taken to satisfy $h \le n - k$ there are open cells in Table 2.27.

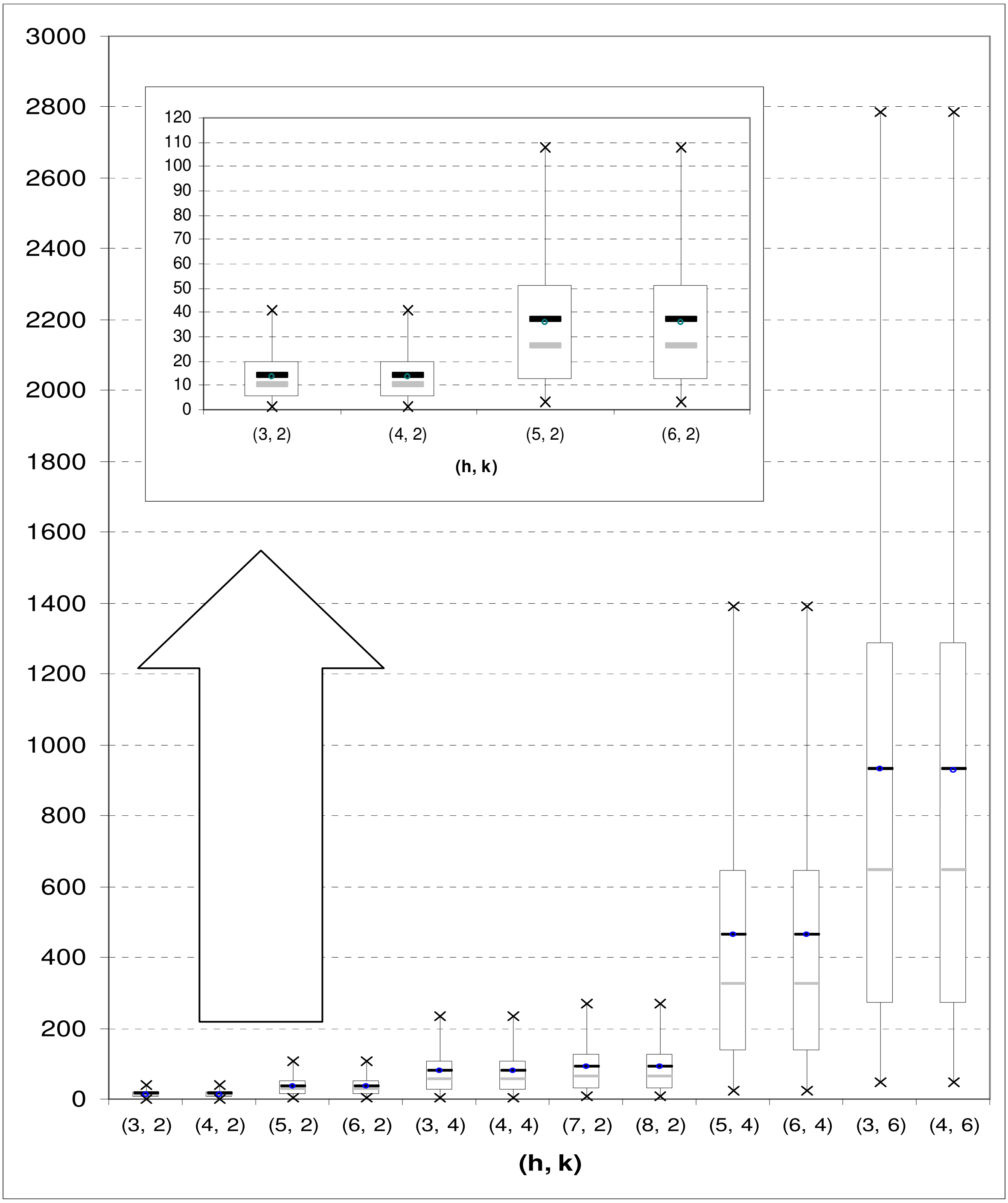

Figure 2.7. Boxplot-like graphs for the in-control run length distribution of various upper one-sided CUSUM sign charts when $n = 10$. The whiskers extend to the 5th and the 95th percentiles. The symbols "▬", "○" and "▬" denote the *ARL*, *SDRL*[*] and *MRL*, respectively.

[*] For ease of interpretation, the standard deviation (as measure of spread) is included in the (location) measures of percentiles.

Example 2.9

An upper one-sided CUSUM sign chart for the Montgomery (2001) piston ring data

We conclude this sub-section by illustrating the upper one-sided CUSUM sign chart using a set of data from Montgomery (2001; Table 5.2) on the inside diameters of piston rings manufactured by a forging process. The dataset contains 15 samples (each of size 5). We assume that the underlying distribution is symmetric with a known target value of $\theta_0 = 74\,mm$.

Let $k = 3$. Once k is selected, the constant h should be chosen to give the desired in-control average run length performance. By choosing $h = 2$ we obtain an in-control average run length of 32 which is the highest in-control average run length attainable when $n = 5$ (see Table 2.22).

The plotting statistics for the Shewhart sign chart (SN_i for $i = 1,2,...,15$) are given in the second row of Table 2.28. The upper one-sided CUSUM plotting statistics (S_i^+ for $i = 1,2,...,15$) are given in the third row of Table 2.28.

Table 2.28. SN_i and S_i^+ values for the piston ring data in Montgomery (2001)*.

Sample No:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
$\boldsymbol{SN_i}$	2	1	-4	3	0	3	3	-1	3	4	1	5	5	5	4
$\boldsymbol{S_i^+}$	0	0	0	0	0	0	0	0	0	1	0	2	4	6	7

To illustrate the calculations, consider sample number 1. The equation for the plotting statistic S_i^+ is $S_1^+ = \max[0, S_0^+ + SN_1 - k] = \max[0, 0 + 2 - 3] = \max[0, -1] = 0$ where a signaling event occurs for the first i such that $S_i^+ \geq h$, that is, $S_i^+ \geq 2$. The graphical display of the upper one-sided CUSUM sign chart is shown in Figure 2.8.

* See SAS Program 3 in Appendix B for the calculation of the values in Table 2.28.

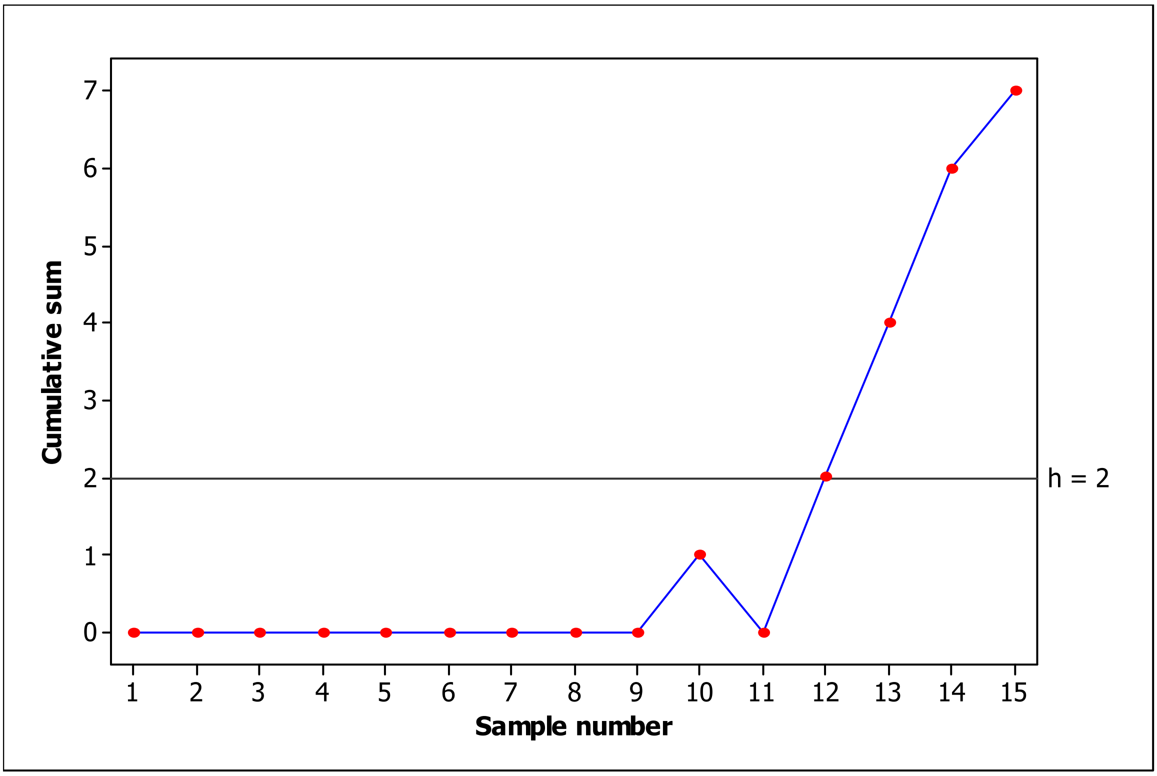

Figure 2.8. The upper one-sided CUSUM sign chart for the Montgomery (2001) piston ring data.

The upper one-sided CUSUM sign chart signals at sample 12, indicating a most likely upward shift from the known target value θ_0. The action taken following an out-of-control signal on a CUSUM chart is identical to that with any control chart. A search for assignable causes should be done, corrective action should be taken (if required) and, following this, the CUSUM is reset to zero. Different control charts are compared by designing the control charts to have the same ARL_0 and then evaluating the ARL_δ. The control chart with the lower ARL_δ is the preferred chart. These procedures can not be meaningfully illustrated using the data from Montgomery (2001) because the sample size $n = 5$ used here is too small. It may be noted that the highest ARL_0^+ is 32 for $n = 5$. Thus, achievable values of ARL_0^+ are too small for practical use, unless n is 'large'.

2.3.2.2. Lower one-sided control charts

Analogous to the previous section, a Markov chain representation of the one-sided CUSUM procedure based on integer-valued cumulative sums is presented in this section. The number of states included in the Markov chain is minimized in order to make the methods as efficient as possible. The time that the procedure signals is the first time such that the finite-state Markov chain S_t^- enters the state ς_0 where the state space is given by $\Omega^- = \{\varsigma_0, \varsigma_1, ..., \varsigma_{r+s-1}\}$ with $-h = \varsigma_0 < ... < \varsigma_{r+s-1} = 0$, $S_0^- = 0$ and

$$S_t^- = \max\{-h, \min\{0, S_{t-1}^- + SN_t + k\}\}. \tag{2.47}$$

Clearly, there is only one absorbent state, since the chart signals when S_t^- falls on or below $-h$, i.e. $s = 1$.

The distribution of SN_t can easily be obtained from the binomial distribution (recall that $SN_i = 2T_i - n \quad \forall i$, where T_i is binomially distributed with parameters n and $p = P(X_{ij} \geq \theta_0)$) . The binomial probabilities are given in Table G of Gibbons and Chakraborti (2003) and can easily be calculated using some type of statistical software package, for example, Excel or SAS.

Example 2.10

A lower one-sided CUSUM sign chart where the sample size is odd ($n = 5$)

The statistical properties of a lower one-sided CUSUM sign chart with a decision interval of 4 ($h = 4$), a reference value of 1 ($k = 1$) and a sample size of 5 ($n = 5$) is examined. For n odd, the reference value is taken to be odd, because this leads to the sum $\sum (SN_i - k)$ being equal to even values which reduces the size of the state space for the Markov chain. For $h = 4$ we have Ω^- = $\{\varsigma_0, \varsigma_1, \varsigma_2\} = \{-4, -2, 0\}$ with $-h = \varsigma_0 < \varsigma_1 < \varsigma_2 = 0$. The state space is calculated using equation (2.47) and the calculations are shown in Table 2.29.

Table 2.29. Calculation of the state space when $h = 4$, $k = 1$ and $n = 5$.

SN_t	$S^-_{t-1} + SN_t + k$	$\min\{0, S^-_{t-1} + SN_t + k\}$	$S^-_t = \max\{-h, \min\{0, S^-_{t-1} + SN_t + k\}\}$
-5	-4[*]	-4	-4
-3	-2	-2	-2
-1	0	0	0
1	2	0	0
3	4	0	0
5	6	0	0

Table 2.30. Classifications and descriptions of the states.

State number	Description of the state	Absorbent (A)/ Non-absorbent (NA)
0	$S^-_t = 0$	NA
1	$S^-_t = -2$	NA
2	$S^-_t = -4$	A

From Table 2.30 we see that there are two non-absorbent states, i.e. $r = 2$, and one absorbent state, i.e. $s = 1$. Therefore, the corresponding TPM will be a $(r+s)\times(r+s) = 3\times 3$ matrix. It can be shown (see Table 2.31) that the TPM is given by

$$TPM_{3\times3} = \begin{pmatrix} p_{00} & p_{0(-2)} & p_{0(-4)} \\ p_{(-2)0} & p_{(-2)(-2)} & p_{(-2)(-4)} \\ p_{(-4)0} & p_{(-4)(-2)} & p_{(-4)(-4)} \end{pmatrix} = \begin{pmatrix} 26/32 & 5/32 & | & 1/32 \\ 16/32 & 10/32 & | & 6/32 \\ - & - & - & - \\ 0 & 0 & | & 1 \end{pmatrix} = \begin{pmatrix} Q_{2\times2} & | & \underline{p}_{2\times1} \\ - & - & - \\ \underline{0}'_{1\times2} & | & 1_{1\times1} \end{pmatrix}$$

where the essential transition probability sub-matrix $Q_{2\times2} : (NA \to NA)$ is an $r \times r = 2\times 2$ matrix, $\underline{p}_{2\times1} : (NA \to A)$ is an $(r+s-1)\times 1 = 2\times 1$ column vector, $\underline{0}'_{1\times2} : (A \to NA)$ is a $1\times(r+s-1) = 1\times 2$ row vector and $1_{1\times1} : (A \to A)$ represents the scalar value one.

The one-step transition probabilities are calculated by substituting SN_t in expression (2.47) by $2T - n$ and substituting in values for h, k, S^-_t and S^-_{t-1}. The calculation of the one-step transition probabilities are given for illustration in Table 2.31.

[*] Note: Since only the state space needs to be described, S^-_{t-1} can be any value from Ω^- and we therefore take, without loss of generality, $S^-_{t-1} = 0$. Any other possible value for S^-_{t-1} would lead to the same Ω^-.

Table 2.31. The calculation of the transition probabilities when $h=4$, $k=1$ and $n=5$.

p_{00} $=P(S_t=0\mid S_{t-1}=0)$ $=P(\max\{-4,\min\{0,0+SN_t+1\}\}=0)$ $=P(\min\{0,0+SN_t+1\}=0)$ $=P(SN_t+1\geq 0)$ $=P(SN_t\geq -1)$ $=P(2T-5\geq -1)$ $=P(T\geq 2)$ $=1-P(T\leq 1)$ $=26/32$	$p_{0(-2)}$ $=P(S_t=-2\mid S_{t-1}=0)$ $=P(\max\{-4,\min\{0,0+SN_t+1\}\}=-2)$ $=P(\min\{0,SN_t+1\}=-2)$ $=P(SN_t+1=-2)$ $=P(SN_t=-3)$ $=P(2T-5=-3)$ $=P(T=1)$ $=5/32$	$p_{0(-4)}$ $=1-(p_{00}+p_{0(-2)})$ $=1-(26/32+5/32)$ $=1/32$
$p_{(-2)0}$ $=P(S_t=0\mid S_{t-1}=-2)$ $=P(\max\{-4,\min\{0,-2+SN_t+1\}\}=0)$ $=P(\min\{0,-2+SN_t+1\}=0)$ $=P(SN_t-1\geq 0)$ $=P(SN_t\geq 1)$ $=P(2T-5\geq 1)$ $=P(T\geq 3)$ $=1-P(T\leq 2)$ $=16/32$	$p_{(-2)(-2)}$ $=P(S_t=-2\mid S_{t-1}=-2)$ $=P(\max\{-4,\min\{0,-2+SN_t+1\}\}=-2)$ $=P(\min\{0,-2+SN_t+1\}=-2)$ $=P(SN_t-1=-2)$ $=P(SN_t=-1)$ $=P(2T-5=-1)$ $=P(T=2)$ $=10/32$	$p_{(-2)(-4)}$ $=1-(p_{(-2)0}+p_{(-2)(-2)})$ $=1-(16/32+10/32)$ $=6/32$
$p_{(-4)0}$ $=P(S_t=0\mid S_{t-1}=-4)$ $=0$*	$p_{(-4)(-2)}$ $=P(S_t=-2\mid S_{t-1}=-4)$ $=0$	$p_{(-4)(-4)}$ $=1-(p_{(-4)0}+p_{(-4)(-2)})$ $=1-(0+0)$ $=1$

The formulas of the moments and some characteristics of the run length distribution have been studied by Fu, Spiring and Xie (2002) and Fu and Lou (2003) – see equations (2.41) to (2.45). By substituting $\underline{\xi}_{1\times 2}=(1\quad 0)$, $Q_{2\times 2}=\begin{pmatrix} 26/32 & 5/32 \\ 16/32 & 10/32 \end{pmatrix}$ and $\underline{1}_{2\times 1}=\begin{pmatrix}1\\1\end{pmatrix}$ into these equations, we obtain the following:

$$ARL=E(N)=\underline{\xi}(I-Q)^{-1}\underline{1}=16.62$$

$$E(N^2)=\underline{\xi}(I+Q)(I-Q)^{-2}\underline{1}=516.59$$

$$SDRL=\sqrt{Var(N)}=\sqrt{E(N^2)-(E(N))^2}=15.51$$

$$5^{th}\text{ percentile}=\rho_5=2$$

$$25^{th}\text{ percentile}=\rho_{25}=6$$

* The probability equals zero, because it is impossible to go from an absorbent state to a non-absorbent state.

Median $= 50^{th}$ percentile $= \rho_{50} = 12$

75^{th} percentile $= \rho_{75} = 23$

95^{th} percentile $= \rho_{95} = 48$

The in-control average run length values, standard deviation of the run length values and percentiles for the lower one-sided CUSUM sign chart are exactly the same as for the upper one-sided CUSUM sign chart, since the one-step transition probabilities matrices are the same (compare the transition probabilities matrices of examples 2.6 and 2.10). Therefore, we obtain Result 2.10:

Result 2.10:

The in-control average run length (ARL_0^+), standard deviation of the run length ($SDRL$), 5^{th}, 25^{th}, 50^{th}, 75^{th} and 95^{th} percentile values tabulated for the upper one-sided CUSUM sign chart will also hold for the lower one-sided CUSUM sign chart.

Example 2.11

A lower one-sided CUSUM sign chart for the Montgomery (2001) piston ring data

We conclude this sub-section by illustrating the lower one-sided CUSUM sign chart using a set of data from Montgomery (2001; Table 5.2) on the inside diameters of piston rings manufactured by a forging process. The dataset contains 15 samples (each of size 5). We assume that the underlying distribution is symmetric with a known target value of $\theta_0 = 74\,mm$.

From Table 2.22 it can be seen that the in-control average run length equals 32 when $h = 2$ and $k = 3$ (recall that this is the largest possible in-control average run length value that the chart can obtain, since $2^5 = 32$). The plotting statistics for the Shewhart sign chart (SN_i for $i = 1,2,...,15$) are given in the second row of Table 2.32. The lower one-sided CUSUM plotting statistics (S_i^- for $i = 1,2,...,15$) are given in the third row of Table 2.32.

Table 2.32. SN_i and S_i^- values for the piston ring data in Montgomery (2001)*.

Sample No:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
SN_i	2	1	-4	3	0	3	3	-1	3	4	1	5	5	5	4
S_i^-	0	0	-1	0	0	0	0	0	0	0	0	0	0	0	0

To illustrate the calculations, consider sample number 1. The equation for the plotting statistic S_1^- is

$$S_1^{-*} = \max[0, S_0^{-*} - SN_1 - k] = \max[0, 0 - 2 - 3] = \max[0, -5] = 0 \qquad (2.48)$$

or

$$S_1^- = \min[0, S_0^- + SN_1 + k] = \min[0, 0 + 2 + 3] = \min[0, 5] = 0\,. \qquad (2.49)$$

A signaling event occurs for the first i such that $S_i^{-*} \geq h$, that is, $S_i^{-*} \geq 2$ if expression (2.48) is used or $S_i^- \leq -h$, that is, $S_i^- \leq -2$ if expression (2.49) is used.

The graphical display of the lower one-sided CUSUM sign chart (using expression (2.49)) is shown in Figure 2.9 and does not signal.

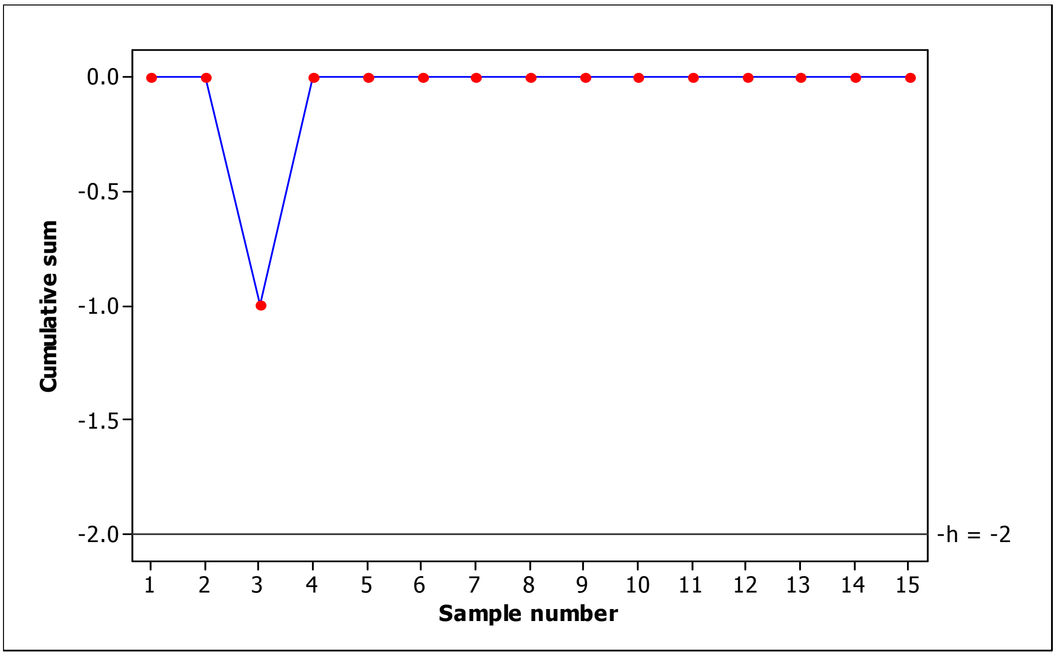

Figure 2.9. The lower one-sided CUSUM sign chart for the Montgomery (2001) piston ring data.

* See SAS Program 3 in Appendix B for the calculation of the values in Table 2.32.

2.3.3. Two-sided control charts

Various authors have studied the two-sided CUSUM scheme, for example, Kemp (1971) gives the average run length of the two-sided CUSUM scheme in terms of the average run lengths of the two corresponding one-sided schemes. Lucas and Crosier (1982) used a Markov chain representation of the two-sided CUSUM scheme to determine the run length distribution and the average run length. In this thesis, the approach taken by Brook and Evans (1972) for the one-sided CUSUM scheme is extended to the two-sided CUSUM scheme. A Markov chain representation of the two-sided CUSUM scheme based on integer-valued random variables will be presented. This is done since the nonparametric statistics that are the building blocks of the CUSUM scheme are discrete random variables. The number of states included in the Markov chain is minimized by taking the reference value k so that the state space of the Markov chain is a set of even numbers. This reduces the size of the TPM and thus eliminates unnecessary calculations in order to make the methods as efficient as possible.

Recall that for the upper one-sided CUSUM sign chart we use

$$S_t^+ = \min\{h, \max\{0, SN_t - k + S_{t-1}^+\}\} \text{ for } n = 1,2,... \quad (2.50)$$

For a lower one-sided CUSUM sign chart we use

$$S_t^- = \max\{-h, \min\{0, SN_t + k + S_{t-1}^-\}\} \text{ for } n = 1,2,... \quad (2.51)$$

For the two-sided scheme the two one-sided schemes are performed simultaneously. The corresponding two-sided CUSUM chart signals for the first t at which either one of the two inequalities is satisfied, that is, either $S_t^+ \geq h$ or $S_t^- \leq -h$. Starting values are typically chosen to equal zero, that is, $S_0^+ = S_0^- = 0$. The two-sided scheme signals at $N = \min_t\{t : S_t^+ \geq h \quad \text{or} \quad S_t^- \leq -h\}$ where h is a positive integer.

The two-sided CUSUM scheme can be represented by a Markov chain with states corresponding to the possible combinations of values of S_t^+ and S_t^-. The states corresponding to values for which a signal is given by the CUSUM scheme are called absorbent states. Clearly, there are two absorbent states since the chart signals when S_t^+ falls on or above h or when S_t^- falls on or below $-h$, i.e. $s = 2$. Recall that r denotes the number of non-absorbent states and, consequently, the corresponding TPM is an $(r+s)\times(r+s)$, i.e. an $(r+2)\times(r+2)$ matrix.

The time that the procedure signals is the first time such that the finite-state Markov chain enters the state ς_0 or ς_{r+s-1} where the state space is given by $\Omega = \Omega^+ \cup \Omega^- = \{\varsigma_0, \varsigma_1, \ldots, \varsigma_{r+s-1}\}$ with $-h = \varsigma_0 < \ldots < \varsigma_{r+s-1} = h$.

Example 2.12

A two-sided CUSUM sign chart where the sample size is odd ($n = 5$)

The statistical properties of a two-sided CUSUM sign chart with a decision interval of 4 ($h = 4$), a reference value of 1 ($k = 1$) and a sample size of 5 ($n = 5$) is examined. For n odd, the reference value k is taken to be odd, because this leads to the sum $\sum(SN_i - k)$ being equal to even values which reduces the size of the state space for the Markov chain. This reduces the size of the TPM and thus eliminates unnecessary calculations in order to make the methods as efficient as possible. Let Ω denote the state space for the two-sided chart. Ω is the union of the state space for the upper one-sided chart $\Omega^+ = \{0,2,4\}$ and the state space for the lower one-sided chart $\Omega^- = \{-4,-2,0\}$. Therefore, $\Omega = \Omega^+ \cup \Omega^- = \{-4,-2,0\} \cup \{0,2,4\} = \{-4,-2,0,2,4\} = \{\varsigma_0, \varsigma_1, \varsigma_2, \varsigma_3, \varsigma_4\}$ with $-h = \varsigma_0 < \varsigma_1 < \varsigma_2 < \varsigma_3 < \varsigma_4 = h$.

Table 2.33. Classifications and descriptions of the states.

State number	Values of the CUSUM statistic(s)	Absorbent (A) Non-absorbent (NA)
0	$S_t^- = 0$ and $S_t^+ = 0$	NA
1	$S_t^- = 2$ or $S_t^+ = 2$ *	NA
2	$S_t^- = -2$ or $S_t^+ = -2$ †	NA
3	$S_t^- = 4$ or $S_t^+ = 4$ ‡	A
4	$S_t^- = -4$ or $S_t^+ = -4$ §	A

From Table 2.33 we see that there are three non-absorbent states, i.e. $r = 3$, and two absorbent states, i.e. $s = 2$. Therefore, the corresponding TPM will be a $(r+s)\times(r+s) = 5\times 5$ matrix. The layout of the TPM is as follows. There are three transient states and two absorbent states. By convention we first list the non-absorbent states and then we list the absorbent states. In column one we compute the probability of moving from state i to state 0, for all i. Note that the process reaches state 0 when both the upper and the lower cumulative sums equal zero. In columns two and three, we compute the probabilities of moving from state i to the remaining non-absorbent states, for all i. Finally, in the remaining two columns we compute the probabilities of moving from state i to the absorbent states, for all i. Thus, the TPM can be conveniently partitioned into 4 sections given by

$$TPM_{5\times 5} = \begin{pmatrix} p_{00} & p_{01} & p_{02} & p_{03} & p_{04} \\ p_{10} & p_{11} & p_{12} & p_{13} & p_{14} \\ p_{20} & p_{21} & p_{22} & p_{23} & p_{24} \\ p_{30} & p_{31} & p_{32} & p_{33} & p_{34} \\ p_{40} & p_{41} & p_{42} & p_{43} & p_{44} \end{pmatrix} = \begin{pmatrix} 20/32 & 5/32 & 5/32 & | & 1/32 & 1/32 \\ 10/32 & 10/32 & 5/32 & | & 6/32 & 1/32 \\ 10/32 & 5/32 & 10/32 & | & 1/32 & 6/32 \\ - & - & - & - & - & - \\ 0 & 0 & 0 & | & 1 & 0 \\ 0 & 0 & 0 & | & 0 & 1 \end{pmatrix} = \begin{pmatrix} Q_{3\times 3} & | & C_{3\times 2} \\ - & - & - \\ Z_{2\times 3} & | & I_{2\times 2} \end{pmatrix}$$

* Moving from state 0 to state 1 can happen when either the upper cumulative sum or the lower cumulative sum equals 2. But the lower cumulative sum can not equal 2 since by definition the lower cumulative sum can only take on integer values smaller than or equal to zero. Therefore, we only use the probability that the upper cumulative sum equals 2 in the calculation of the probabilities in the TPM.

† Moving from state 0 to state 2 can happen when either the upper cumulative sum or the lower cumulative sum equals -2. But the upper cumulative sum can not equal -2 since by definition the upper cumulative sum can only take on integer values greater than or equal to zero. Therefore, we only use the probability that the lower cumulative sum equals -2 in the calculation of the probabilities in the TPM.

‡ A similar argument to the argument in the first footnote on this page holds. Therefore, we only use the probability that the upper cumulative sum equals 4 in the calculation of the probabilities in the TPM.

§ A similar argument to the argument in the second footnote on this page holds. Therefore, we only use the probability that the lower cumulative sum equals -4 in the calculation of the probabilities in the TPM.

where the essential transition probability sub-matrix $Q_{3\times 3}:(NA \to NA)$ is an $r \times r = 3\times 3$ matrix, $C_{3\times 2}:(NA \to A)$ is an $r \times s = 3\times 2$ matrix, $Z_{2\times 3}:(A \to NA)$ is an $s \times r = 2\times 3$ matrix and $I_{2\times 2}:(A \to A)$ is a $s \times s = 2\times 2$ matrix.

The calculation of the elements of the TPM is illustrated next. Note that this essentially involves the calculation of the matrices Q and C. First consider the transient states. Note that the process moves to state 0, i.e., $j = 0$, when both the upper *and* the lower cumulative sums equal 0. Thus the required probability of moving to 0, from any other transient state, is the probability of an *intersection* of two sets involving values of the upper and the lower CUSUM statistics, respectively. On the other hand, the probability of moving to any state $j \neq 0$, from any other state, is the probability of a *union* of two sets involving values of the upper and the lower CUSUM statistics, respectively. However, one of these two sets is empty so that the required probability is the probability of only the non-empty set.

The calculation of the entry in the first row and the first column of the matrix Q, p_{00}, will be discussed in detail. This is the probability of moving from state 0 to state 0 in one step. As we just described, this can happen only when the upper *and* the lower cumulative sums both equal 0 at time t. For the upper one-sided CUSUM p_{00} is the probability that the upper CUSUM equals 0 at time t, given that the upper CUSUM equaled 0 at time $t-1$, that is, $P(S_t^+ = 0 \mid S_{t-1}^+ = 0)$. For the lower one-sided procedure p_{00} is the probability that the lower CUSUM equals 0 at time t, given that the lower CUSUM equaled 0 at time $t-1$, that is, $P(S_t^- = 0 \mid S_{t-1}^- = 0)$. For the two-sided procedure the two one-sided procedures are performed simultaneously. Therefore we have that $p_{00} = P\left(\left\{S_t^+ = 0 \mid S_{t-1}^+ = 0\right\} \cap \left\{S_t^- = 0 \mid S_{t-1}^- = 0\right\}\right)$.

We have that

$$p_{00}$$
$$= P\left(\left\{S_t^+ = 0 \mid S_{t-1}^+ = 0\right\} \cap \left\{S_t^- = 0 \mid S_{t-1}^- = 0\right\}\right)$$

this is computed by substituting in values for h, k, S_t^+, S_{t-1}^+, S_t^- and S_{t-1}^- into equations (2.50) and (2.51)

$$= P\left(\left\{\min\left\{4, \max\{0, SN_t - 1 + 0\}\right\} = 0\right\} \cap \left\{\max\left\{-4, \min\{0, SN_t + 1 + 0\}\right\} = 0\right\}\right)$$
$$= P\left(\left\{\max\{0, SN_t - 1 + 0\} = 0\right\} \cap \left\{\min\{0, SN_t + 1 + 0\} = 0\right\}\right)$$
$$= P\left((SN_t - 1 \le 0) \cap (SN_t + 1 + 0 \ge 0)\right)$$
$$= P\left((SN_t \le 1) \cap (SN_t \ge -1)\right)$$

recall that $SN_t = 2T - n$ where T is binomially distributed with parameters n and $p = P(X_{ij} \ge \theta_0)$

$$= P\left((2T - 5 \le 1) \cap (2T - 5 \ge -1)\right)$$
$$= P\left((T \le 3) \cap (T \ge 2)\right)$$
$$= P(T = 2) + P(T = 3)$$
$$= 2P(T = 2)$$
$$= 2 \times \left(10/32\right)$$
$$= 20/32$$

since T is binomially distributed with parameters $n = 5$ and $p = 0.5$.

The remaining entries in the first column of the matrix Q can be calculated similarly and we find that $p_{10} = 10/32$ and $p_{20} = 10/32$.

Next we discuss the calculation of the entry in the first row and the second column of the matrix Q, p_{01}, in detail. This is the probability of moving from state 0 to state 1 in one step. This can happen when either the upper cumulative sum or the lower cumulative sum equals 2. But the lower cumulative sum can not equal 2 since by definition the lower cumulative sum can only take on integer values smaller than or equal to zero. Therefore although the required probability is the probability of the *union* of two sets involving values of the upper and the lower CUSUM statistics, one of these sets is empty so that the required probability is the probability of only the non-empty set. Hence, in this case, we will only have to calculate the upper one-sided probability. For the upper one-sided CUSUM, p_{01} is the probability that the upper cumulative sum equals 2 at time t, given that the upper cumulative sum equaled 0 at time $t-1$, that is, $P(S_t^+ = 2 \mid S_{t-1}^+ = 0)$. We have that

p_{01}
$= P(S_t^+ = 2 \mid S_{t-1}^+ = 0)$
$= P(\min\{4, \max\{0, SN_t - 1 + 0\}\} = 2)$
$= P(\max\{0, SN_t - 1 + 0\} = 2)$
$= P(SN_t - 1 = 2)$
$= P(SN_t = 3)$
$= P(2T - 5 = 3)$
$= P(T = 4)$
$= 5/32$.

The remaining entries in the second column of the matrix Q can be calculated similarly and we find that $p_{11} = 10/32$ and $p_{21} = 5/32$.

Next we discuss the calculation of the entry in the first row and the third column of the matrix Q, p_{02}, in detail. This is the probability of moving from state 0 to state 2 in one step. This happens when *only* the lower cumulative sum moves to -2, since the upper cumulative sum can not move to -2. Recall that the upper cumulative sum can only take on integer values greater than or equal to zero. As in the case of p_{01}, this probability is also the probability of the *union* of two sets, involving values of the CUSUM statistics, one of which is empty, so that the required probability is the probability of only the non-empty set. Hence, in this case since the lower CUSUM is involved, we will only have to calculate the probability associated with the lower one-sided procedure. Now, for the lower one-sided procedure p_{02} is the probability that the lower cumulative sum equals -2 at time t, given that the lower cumulative sum equaled 0 at time $t-1$, that is, $P(S_t^- = -2 \mid S_{t-1}^- = 0)$. We have that

p_{02}
$= P(S_t^- = -2 \mid S_{t-1}^- = 0)$
$= P(\max\{-4, \min\{0, SN_t + 1 + 0\}\} = -2)$
$= P(\min\{0, SN_t + 1 + 0\} = -2)$
$= P(SN_t + 1 = -2)$
$= P(SN_t = -3)$
$= P(2T - 5 = -3)$
$= P(T = 1)$
$= 5/32$.

The remaining entries in the third column of the matrix Q can be calculated similarly and we find that $p_{12} = 5/32$ and $p_{22} = 10/32$.

Next we discuss the calculation of the entry in the first row and the first column of the matrix C, p_{03}, in detail. This is the probability of moving from state 0 to an absorbent state, state 3, in one step. Again, this can happen when *only* the upper cumulative sum moves to 4, since the lower cumulative sum can not move to 4. Recall that the lower cumulative sum can only take on integer values smaller than or equal to zero. Therefore, once again the required probability is the probability of the *union* of two sets involving values of the CUSUM statistics, one of which is empty so that the probability is the probability of only the non-empty set. Therefore we will only have to calculate the upper one-sided probability in this case. For the upper one-sided procedure p_{03} is the probability that the upper cumulative sum equals 4 at time t, given that the upper cumulative sum equaled 0 at time $t-1$, that is, $P(S_t^+ = 4 \mid S_{t-1}^+ = 0)$. We have that

$$
\begin{aligned}
& p_{03} \\
&= P(S_t^+ = 4 \mid S_{t-1}^+ = 0) \\
&= P\left(\min\{4, \max\{0, SN_t - 1 + 0\}\} = 4\right) \\
&= P\left(SN_t - 1 \geq 4\right) \\
&= P\left(SN_t \geq 5\right) \\
&= P\left(2T - 5 \geq 5\right) \\
&= P\left(T \geq 5\right) \\
&= 1/32 .
\end{aligned}
$$

The remaining entries in the first column of the matrix C can be calculated similarly and we find that $p_{13} = 6/32$ and $p_{23} = 1/32$.

Next we discuss the calculation of the entry in the first row and the last column of the matrix C, p_{04}, in detail. This is the probability of moving from state 0 to state 4 in one step. This can happen when *only* the lower cumulative sum moves to -4, since the upper cumulative sum can not move to -4. Recall that the upper cumulative sum can only take on integer values greater than or equal to zero. Therefore although the required probability is the probability of the *union* of two sets involving values of the upper and the lower CUSUM statistics, one of these sets is empty so that the required probability is the probability of only the non-empty set. Therefore we will only have to calculate the lower one-sided probability. For the lower one-sided procedure p_{04} is the probability that the lower cumulative sum equals -4 at time t, given that the lower cumulative sum equaled 0 at time $t-1$, that is, $P(S_t^- = -4 \mid S_{t-1}^- = 0)$. We have that

$$\begin{aligned}
&p_{04} \\
&= P(S_t^- = -4 \mid S_{t-1}^- = 0) \\
&= P\left(\max\{-4, \min\{0, SN_t + 1 + 0\}\} = -4\right) \\
&= P(SN_t + 1 \le -4) \\
&= P(SN_t \le -5) \\
&= P(2T - 5 \le -5) \\
&= P(T \le 0) \\
&= \tfrac{1}{32}.
\end{aligned}$$

The remaining entries in the last column of the matrix C can be calculated similarly and we find that $p_{14} = \tfrac{1}{32}$ and $p_{24} = \tfrac{6}{32}$.

The run length distribution and its parameters

The run length distribution and its parameters are calculated using the matrix Q. The ARL is given by $\underline{\xi}(I-Q)^{-1}\underline{1}$ where $\underline{\xi}_{1\times 3} = (1 \quad 0 \quad 0)$, $Q_{3\times 3} = \frac{1}{32}\begin{pmatrix} 20 & 5 & 5 \\ 10 & 10 & 5 \\ 10 & 5 & 10 \end{pmatrix}$ and $\underline{1}_{3\times 1} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$.

As a result, $ARL = E(N) = \underline{\xi}(I-Q)^{-1}\underline{1} = 8.31$.

Let ARL^+ and ARL^- denote the average run lengths of the upper and lower one-sided charts, respectively. The ARL of the two-sided chart can be expressed as a function of the average run lengths of the one-sided charts through the expression

$$ARL = \frac{(ARL^+)(ARL^-)}{\left(ARL^+ + ARL^-\right)} \tag{2.52}$$

(see Theorem 1 in Appendix A for the proof of result (2.52)).

From the lower- and upper CUSUM sign sections we have that $ARL^{+} = 16.62$ and $ARL^{-} = 16.62$. Using equation (2.52) we have that $ARL = \frac{(16.62)(16.62)}{(16.62+16.62)} = 8.31$.

Table 2.34. The in-control average run length (ARL_0), standard deviation of the run length ($SDRL$), 5^{th}, 25^{th}, 50^{th}, 75^{th} and 95^{th} percentile* values for the two-sided CUSUM sign chart when $n = 5$†.

k	***h***	
	2	**3 or 4**
1	2.67 2.11 (1, 1, 2, 3, 7)	8.31 7.16 (1, 3, 6, 11, 23)
3	16.00 15.50 (1, 5, 11, 22, 47)	‡

Analogous to what was done for the upper one-sided chart, the five percentiles (given in Table 2.34) are displayed in boxplot-like graphs for various h and k values in Figure 2.10. Recall that we would prefer a "boxplot" with a high valued (large) in-control average run length and a small spread. Applying this criterion, we see that the "boxplot" corresponding to the $(h,k) = (2,3)$ combination has the largest in-control average run length, which is favorable, but it also has the largest spread which is unattractive. The "boxplot" furthest to the right is exactly opposite from the "boxplot" furthest to the left. The latter has the smallest spread, which is favorable, but it also has the smallest in-control average run length, which is unattractive. In conclusion, no "box plot" is optimal relative to the others.

* The three rows of each cell shows the ARL_0, the $SDRL$, and the percentiles (ρ_5, ρ_{25}, ρ_{50}, ρ_{75}, ρ_{95}), respectively.
† See SAS Program 2 in Appendix B for the calculation of the values in Table 2.34.
‡ Since the decision interval is taken to satisfy $h \le n-k$ there are open cells in Table 2.34.

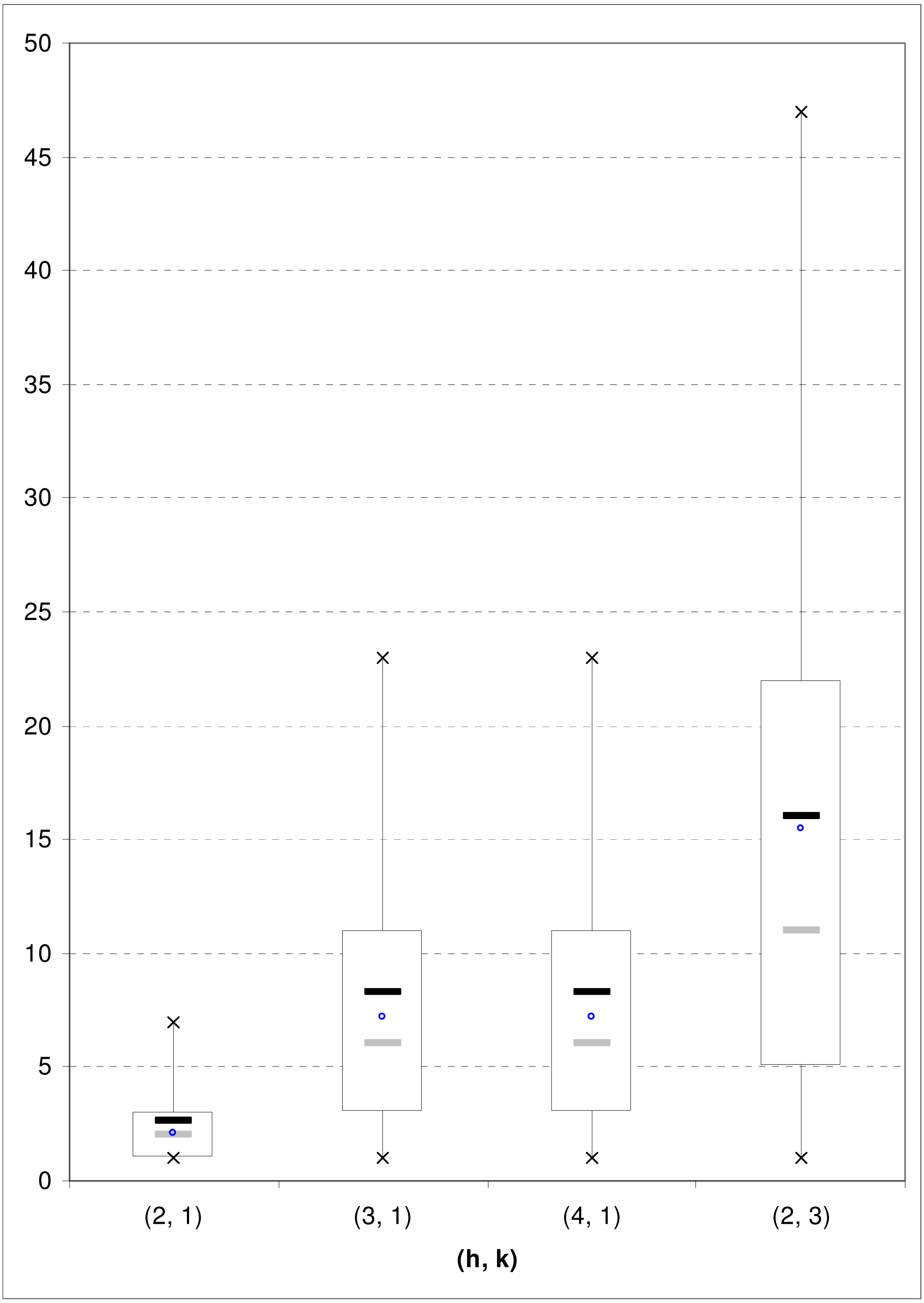

Figure 2.10. Boxplot-like graphs for the in-control run length distribution of various two-sided CUSUM sign charts when $n = 5$. The whiskers extend to the 5th and the 95th percentiles. The symbols "▬", "○" and "▬" denote the *ARL*, *SDRL** and *MRL*, respectively.

* For ease of interpretation, the standard deviation (as measure of spread) is included in the (location) measures of percentiles.

Other values of *h*, *k* and *n* were also considered and the results are given in Table 2.35.

Table 2.35. The in-control average run length (ARL_0), standard deviation of the run length ($SDRL$), $5^{th}, 25^{th}, 50^{th}, 75^{th}$ and 95^{th} percentile* values for the two-sided CUSUM sign chart when $n = 6$†.

k	***h***		
	2	**3 or 4**	**5 or 6**
0	1.45 0.81 (1, 1, 1, 2, 3)	2.96 1.88 (1, 2, 3, 4, 7)	5.33 4.22 (1, 2, 4, 7, 14)
2	4.57 4.04 (1, 2, 3, 6, 13)	19.34 18.36 (2, 6, 14, 26, 56)	‡
4	32.00 31.50 (2, 10, 22, 44, 95)		

** The three rows of each cell shows the ARL_0, the $SDRL$, and the percentiles ($\rho_5, \rho_{25}, \rho_{50}, \rho_{75}, \rho_{95}$), respectively.

† See SAS Program 2 in Appendix B for the calculation of the values in Table 2.35.

‡ Since the decision interval is taken to satisfy $h \le n - k$ there are open cells in Table 2.35.

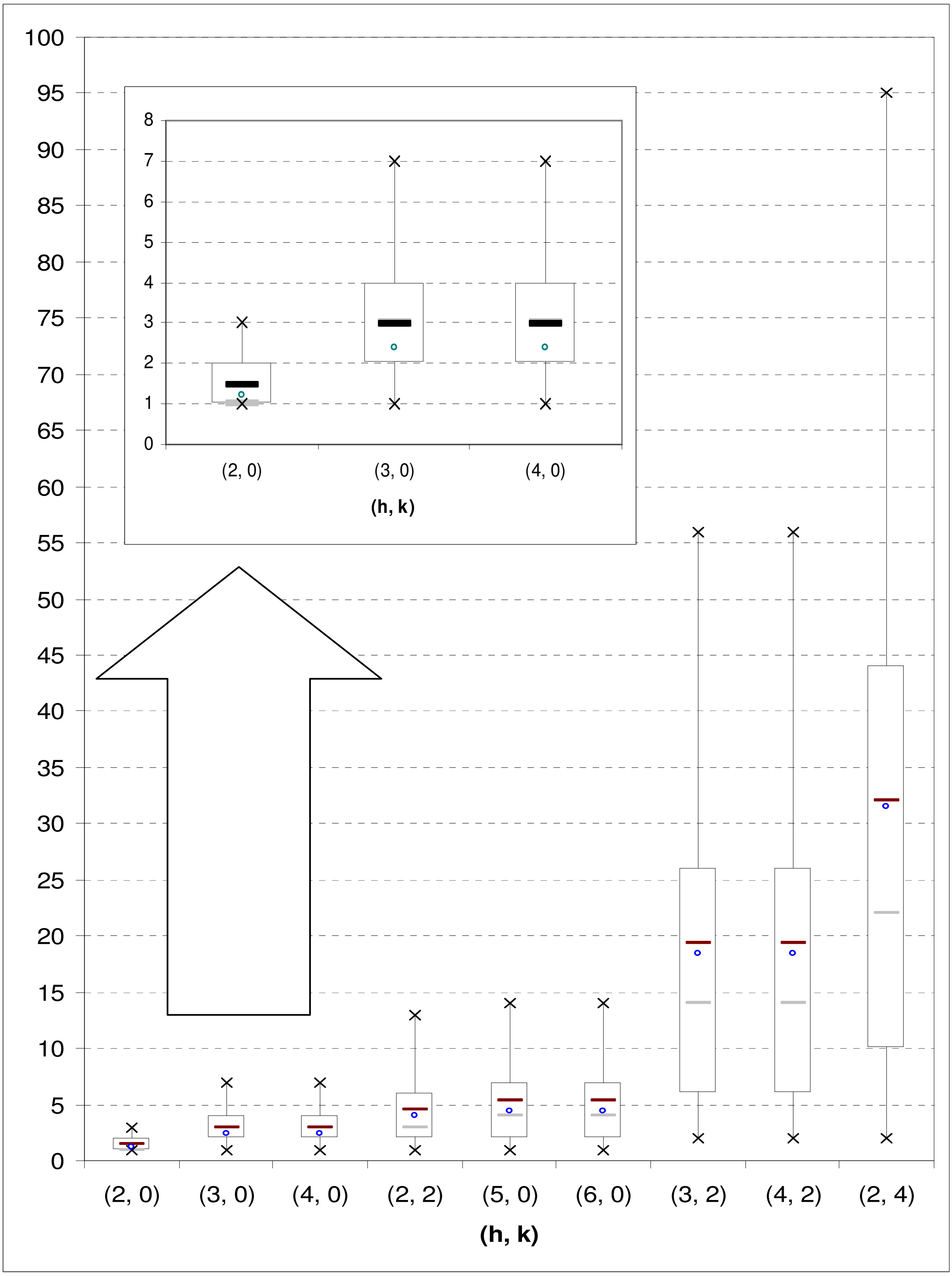

Figure 2.11. Boxplot-like graphs for the in-control run length distribution of various two-sided CUSUM sign charts when $n = 6$. The whiskers extend to the 5th and the 95th percentiles. The symbols "▬", "○" and "▬" denote the *ARL*, *SDRL*[*] and *MRL*, respectively.

[*] For ease of interpretation, the standard deviation (as measure of spread) is included in the (location) measures of percentiles.

Table 2.36. The in-control average run length (ARL_0), standard deviation of the run length ($SDRL$), 5^{th} , 25^{th} , 50^{th} , 75^{th} and 95^{th} percentile* values for the two-sided CUSUM sign chart when $n = 10$ †.

k	***h***		
	3 or 4	**5 or 6**	**7 or 8**
2	7.17 6.39 (1, 3, 5, 10, 20)	18.41 17.05 (2, 6, 13, 25, 52)	45.80 43.63 (4, 15, 32, 63, 133)
4	38.98 38.30 (3, 12, 27, 54, 115)	232.43 231.26 (13, 68, 161, 322, 694)	‡
6	464.98 464.39 (24, 134, 322, 644, 1392)		

* The three rows of each cell shows the ARL_0, the $SDRL$, and the percentiles (ρ_5 , ρ_{25} , ρ_{50} , ρ_{75} , ρ_{95}), respectively.

† See SAS Program 2 in Appendix B for the calculation of the values in Table 2.36.

‡ Since the decision interval is taken to satisfy $h \le n - k$ there are open cells in Table 2.36.

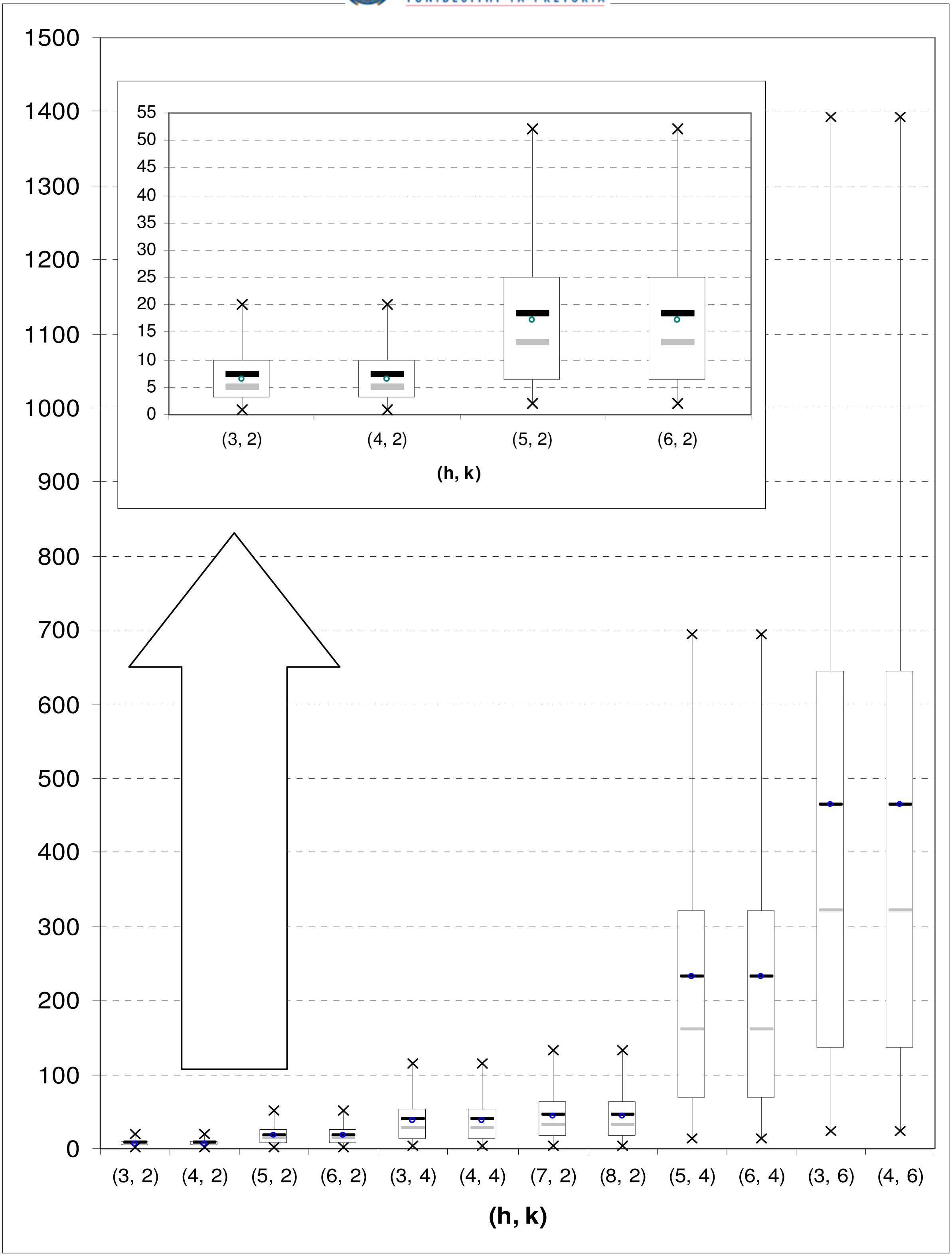

Figure 2.12. Boxplot-like graphs for the in-control run length distribution of various two-sided CUSUM sign charts when $n = 10$. The whiskers extend to the 5th and the 95th percentiles. The symbols "▬", "○" and "▬" denote the *ARL*, *SDRL*[*] and *MRL*, respectively.

[*] For ease of interpretation, the standard deviation (as measure of spread) is included in the (location) measures of percentiles.

Example 2.13

A two-sided CUSUM sign chart for the Montgomery (2001) piston ring data

We conclude this sub-section by illustrating the two-sided CUSUM sign chart using the piston ring data set from Montgomery (2001). We assume that the underlying distribution is symmetric with a known target value of $\theta_0 = 74\,mm$. Let $k = 3$. Once k is selected, the constant h should be chosen to give the desired in-control average run length performance. By choosing $h = 2$ we obtain an in-control average run length of 16 which is the highest in-control average run length attainable when $n = 5$ (see Table 2.34). Table 2.37 shows the upper and lower sign CUSUM statistics, respectively.

Table 2.37. One-sided sign (S_i^+ and S_i^-) statistics*.

Sample number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
S_i^+	0	0	0	0	0	0	0	0	0	1	0	2	4	6	7
S_i^-	0	0	-1	0	0	0	0	0	0	0	0	0	0	0	0

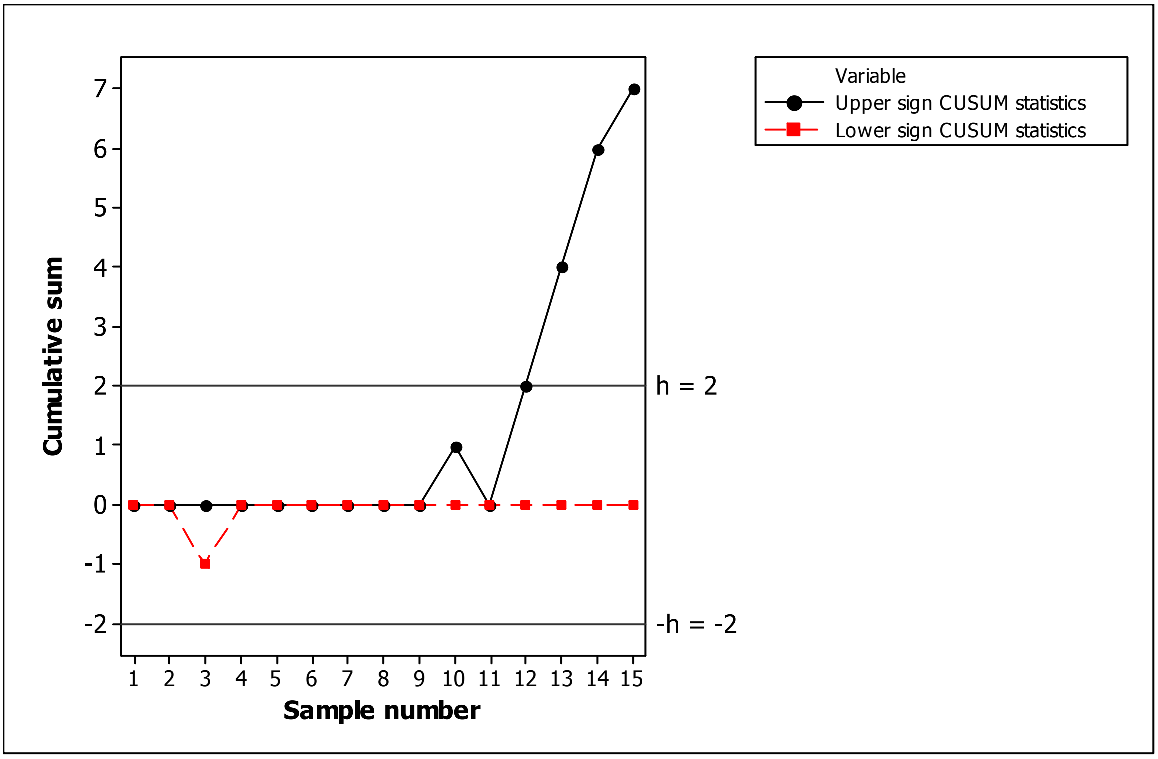

Figure 2.13. The two-sided CUSUM sign chart for the Montgomery (2001) piston ring data.

* See SAS Program 3 in Appendix B for the calculation of the values in Table 2.37.

The two-sided CUSUM sign chart signals at sample number 12, indicating a most like upward shift in the process median. The action taken following an out-of-control signal on a CUSUM chart is identical to that with any control chart. A search for assignable causes should be done, corrective action should be taken (if required) and, following this, the CUSUM is reset to zero.

2.3.4. Summary

While the Shewhart-type charts are widely known and most often used in practice because of their simplicity and global performance, other classes of charts, such as the CUSUM charts are useful and sometimes more naturally appropriate in the process control environment in view of the sequential nature of data collection. In this section we have described the properties of the CUSUM sign chart and given tables for its implementation. Detailed calculations have been given to help the reader to understand the subject more thoroughly.

2.4. The EWMA control chart

2.4.1. Introduction

The exponentially weighted moving average (EWMA) scheme was first introduced by Roberts (1959). In a subsequent article, Roberts (1966) compared the performance of EWMA charts to Shewhart and CUSUM charts. Various authors have studied EWMA charts (see for example Robinson and Ho (1978) and Crowder (1987)). EWMA charts have become very popular over the last few years. It is one of several charting methods aimed at correcting a deficiency of the Shewhart chart – insensitivity to small process shifts.

An EWMA control chart scheme accumulates statistics $X_1, X_2, X_3, ...$ with the plotting statistics defined as

$$Z_i = \lambda X_i + (1-\lambda)Z_{i-1} \qquad (2.53)$$

where $0 < \lambda \leq 1$ is a constant called the weighting constant. The starting value Z_0 is often taken to be the process target value, i.e. $Z_0 = \theta_0$.

The EWMA chart is constructed by plotting Z_i against the sample number i (or time). If the plotting statistic Z_i falls between the two control limits, that is, $LCL < Z_i < UCL$, the process is considered to be in-control. If the plotting statistic Z_i falls on or outside one of the control limits, that is $Z_i \le LCL$ or $Z_i \ge UCL$, the process is considered to be out-of-control.

To illustrate that the plotting statistic Z_i is a weighted average of all the previous statistics, Z_{i-1} may be substituted by $Z_{i-1} = \lambda X_{i-1} + (1-\lambda)Z_{i-2}$ in equation (2.53) to obtain

$$\begin{aligned}
Z_i &= \lambda X_i + (1-\lambda)\left(\lambda X_{i-1} + (1-\lambda)Z_{i-2}\right) \\
&= \lambda X_i + \lambda(1-\lambda)X_{i-1} + (1-\lambda)^2 Z_{i-2} \\
&= \lambda X_i + \lambda(1-\lambda)X_{i-1} + (1-\lambda)^2\left(\lambda X_{i-2} + (1-\lambda)Z_{i-3}\right) \\
&= \lambda X_i + \lambda(1-\lambda)X_{i-1} + \lambda(1-\lambda)^2 X_{i-2} + (1-\lambda)^3 Z_{i-3}.
\end{aligned}$$

This method of substitution is called *recursive substitution.* By continuing the process of recursive substitution for Z_{i-p}, $p = 2,3,...,t$, we obtain

$$Z_i = \lambda \sum_{p=0}^{i-1} (1-\lambda)^p X_{i-p} + (1-\lambda)^i Z_0. \tag{2.54}$$

We can see from expression (2.54) that Z_i can be written as a moving average of the current and past observations which has geometrically decreasing weights $\lambda(1-\lambda)^p$ associated with increasingly aged observations X_{i-p} ($p = 1,2,...$). Therefore, the EWMA has been referred to as a geometric moving average (see, for example, Montgomery (2005)).

If the observations $\{X_i, i = 1,2,...\}$ are independent identically distributed variables with mean μ and variance σ^2, then the mean and the variance of the plotting statistic Z_i are given by

$$\mu_{Z_i} = E(Z_i) = \mu \quad \text{for} \quad i = 1,2,...$$

and

$$\sigma_{Z_i}^2 = \sigma^2 \left(\frac{\lambda}{2-\lambda}\right)\left(1-(1-\lambda)^{2i}\right) \quad \text{for} \quad i = 1,2,... \; .$$

The *exact* control limits and the center line of the EWMA control chart are given by

$$\begin{aligned} UCL &= \theta_0 + L\sigma\sqrt{\left(\frac{\lambda}{2-\lambda}\right)\left(1-(1-\lambda)^{2i}\right)} \\ CL &= \theta_0 \\ LCL &= \theta_0 - L\sigma\sqrt{\left(\frac{\lambda}{2-\lambda}\right)\left(1-(1-\lambda)^{2i}\right)} \end{aligned} \quad . \tag{2.55}$$

From (2.55) we see that we have two design parameters of interest, namely the multiplier L, ($L > 0$) and the smoothing constant λ. We also see that $\left(1-(1-\lambda)^{2i}\right) \to 1$ as i increases. Therefore, as i increases the control limits will approach steady-state values given by

$$\begin{aligned} UCL &= \theta_0 + L\sigma\sqrt{\left(\frac{\lambda}{2-\lambda}\right)} \\ LCL &= \theta_0 - L\sigma\sqrt{\left(\frac{\lambda}{2-\lambda}\right)} \end{aligned} \quad . \tag{2.56}$$

The above-mentioned control limits are called *steady-state* control limits.

Various authors recommend choosing the EWMA constants L and λ by minimizing the average run length at a specified shift for a desired in-control average run length. In general, values of λ in the interval $0.05 \le \lambda \le 0.25$ work well in the normal theory case with $\lambda = 0.05$, $\lambda = 0.1$ and $\lambda = 0.2$ being popular choices. The ARL_0, standard deviation of the run length ($SDRL$), 5^{th}, 25^{th} (the first quartile, Q_1), 50^{th} (the median run length, MRL), 75^{th} (the third quartile, Q_3) and 95^{th} percentile values can be computed and tabulated for various values of L and λ.

Lucas and Saccucci (1990) have investigated some properties of the EWMA chart under the assumption of independent normally distributed observations. Lucas and Saccucci's most important contribution is the use of a Markov-chain approach to evaluate the run-length properties of the EWMA chart. It is important to note that the successive observations are assumed to be independent over time in their evaluation. Lucas and Saccucci (1990) used a procedure similar to that described by Brook and Evans (1972) to approximate the properties of an EWMA scheme. They evaluate the properties of the *continuous state* Markov chain by *discretizing* the infinite state transition probability matrix (TPM). This procedure entails dividing the interval between the upper control limit and the lower control limit into N subintervals of width 2δ. Then the plotting statistic, Z_i, is said to be in the non-absorbing state j at time i if

$$S_j - \delta < Z_i \le S_j + \delta \quad \text{for} \quad j = 1,2,...,N-1$$

and

$$S_j - \delta < Z_i < S_j + \delta \quad \text{for} \quad j = N$$

where S_j denotes the midpoint of the j^{th} interval. Let r denote the number of non-absorbing states. Z_i is said to be in the absorbing state if Z_i falls on or outside one of the control limits, that is, $Z_i \le LCL$ or $Z_i \ge UCL$. Clearly, there are $r+1$ states, since there are r non-absorbing states and one absorbing state. Lucas and Saccucci (1990) have done a thorough job of evaluating the run length properties of the EWMA chart and provided helpful tables for the design of the EWMA chart. Additional tables are provided in the technical report by Lucas and Saccucci (1987). In their 1990 paper they concentrate on the average run length characteristics of various charting combinations. The authors conclude that EWMA procedures have average run length properties similar to those for CUSUM procedures. This point has also been made by various authors, for example, Ewan (1963), Roberts (1966) and Montgomery, Gardiner and Pizzano (1987). In this thesis, the approach taken by Lucas and Saccucci (1990) is extended to the use of the sign statistic resulting in an EWMA sign chart that accumulates the statistics $SN_1, SN_2, SN_3,...$.

2.4.2. The proposed EWMA sign chart

A nonparametric EWMA-type of control chart based on the sign statistic can be obtained by replacing X_i in expression (2.53) with SN_i (recall that $SN_i = \sum_{j=1}^{n} sign(x_{ij} - \theta_0)$).

The EWMA sign chart accumulates the statistics $SN_1, SN_2, SN_3,...$ with the plotting statistics defined as

$$Z_i = \lambda SN_i + (1-\lambda)Z_{i-1} \quad (2.57)$$

where $0 < \lambda \le 1$ and the starting value Z_0 is usually taken to equal zero, i.e. $Z_0 = 0$.

The expected value, variance and standard deviation of SN_i are found from the fact that the distribution of SN_i can easily be obtained from the distribution of the binomial distribution (recall that $SN_i = 2T_i - n$ if there are no ties within a subgroup, where T_i has a binomial distribution with parameters n and $p = P(X_{ij} \geq \theta_0)$). The formulas for the expected value, variance and standard deviation of SN_i was derived in Section 2.1 and we obtained $E(SN_i) = n(2p-1)$, $\text{var}(SN_i) = 4np(1-p)$ and $stdev(SN_i) = \sigma_{SN_i} = 2\sqrt{np(1-p)}$, respectively. The starting value Z_0 can also be taken to be the expected value of SN_i, therefore $Z_0 = E(SN_i) = n(2p-1)$ and in the in-control case where $p = 0.5$ we have $Z_0 = n(2\times 0.5 - 1) = 0$ for all n.

From the similarity between the definitions of the normal EWMA scheme and the sign EWMA scheme, it follows that the exact control limits and the center line of the EWMA sign control chart can be obtained by replacing σ in (2.55) with σ_{SN_i} which yields

$$\begin{aligned} UCL &= 0 + L\sigma_{SN_i}\sqrt{\left(\frac{\lambda}{2-\lambda}\right)\left(1-(1-\lambda)^{2i}\right)} \\ CL &= 0 \\ LCL &= 0 - L\sigma_{SN_i}\sqrt{\left(\frac{\lambda}{2-\lambda}\right)\left(1-(1-\lambda)^{2i}\right)} \end{aligned} . \qquad (2.58)$$

It is important to note θ_0 in (2.55) is replaced by 0 in (2.58). This is because the EWMA sign chart is designed for the sign test statistic and not for the observations (the X_i's).

Similarly, the *steady-state* control limits can be obtained by replacing σ in (2.56) with σ_{SN_i} and θ_0 by zero which yields

$$\begin{aligned} UCL &= L\sigma_{SN_i}\sqrt{\left(\frac{\lambda}{2-\lambda}\right)} \\ LCL &= -L\sigma_{SN_i}\sqrt{\left(\frac{\lambda}{2-\lambda}\right)} \end{aligned} . \qquad (2.59)$$

2.4.3. Markov-chain approach

Lucas and Saccucci (1990) evaluated the properties of the *continuous state* Markov chain by *discretizing* the infinite state TPM. This procedure entails dividing the interval between the UCL and the LCL into N subintervals of width 2δ. The width of each subinterval can be obtained by setting $\delta = \frac{1}{2} \times \frac{(UCL - LCL)}{N}$ and we get that the width of an interval equals $2\delta = \frac{(UCL - LCL)}{N}$. Thus, the endpoints of the subintervals will be given by

$$LCL, \quad LCL + \frac{(UCL - LCL)}{N}, \quad LCL + 2\frac{(UCL - LCL)}{N}, \ldots, \quad LCL + (N-1)\frac{(UCL - LCL)}{N}, \quad UCL,$$

respectively (see Figure 2.14). In general, the endpoints of the j^{th} interval will be given by

$$(LCL_j, UCL_j) = \left(LCL + (j-1) \times \frac{(UCL - LCL)}{N}, LCL + j \times \frac{(UCL - LCL)}{N} \right).$$

The midpoint of the j^{th} interval, S_j, is easily obtained by taking the sum of the two endpoints of the j^{th} interval and dividing it by 2. Thus, we obtain

$$\begin{aligned} &S_j \\ &= \frac{LCL_j + UCL_j}{2} \\ &= \frac{\left(LCL + \frac{(j-1)(UCL - LCL)}{N} \right) + \left(LCL + \frac{j(UCL - LCL)}{N} \right)}{2} \\ &= \frac{(2j-1)(UCL - LCL)}{2N}. \end{aligned}$$

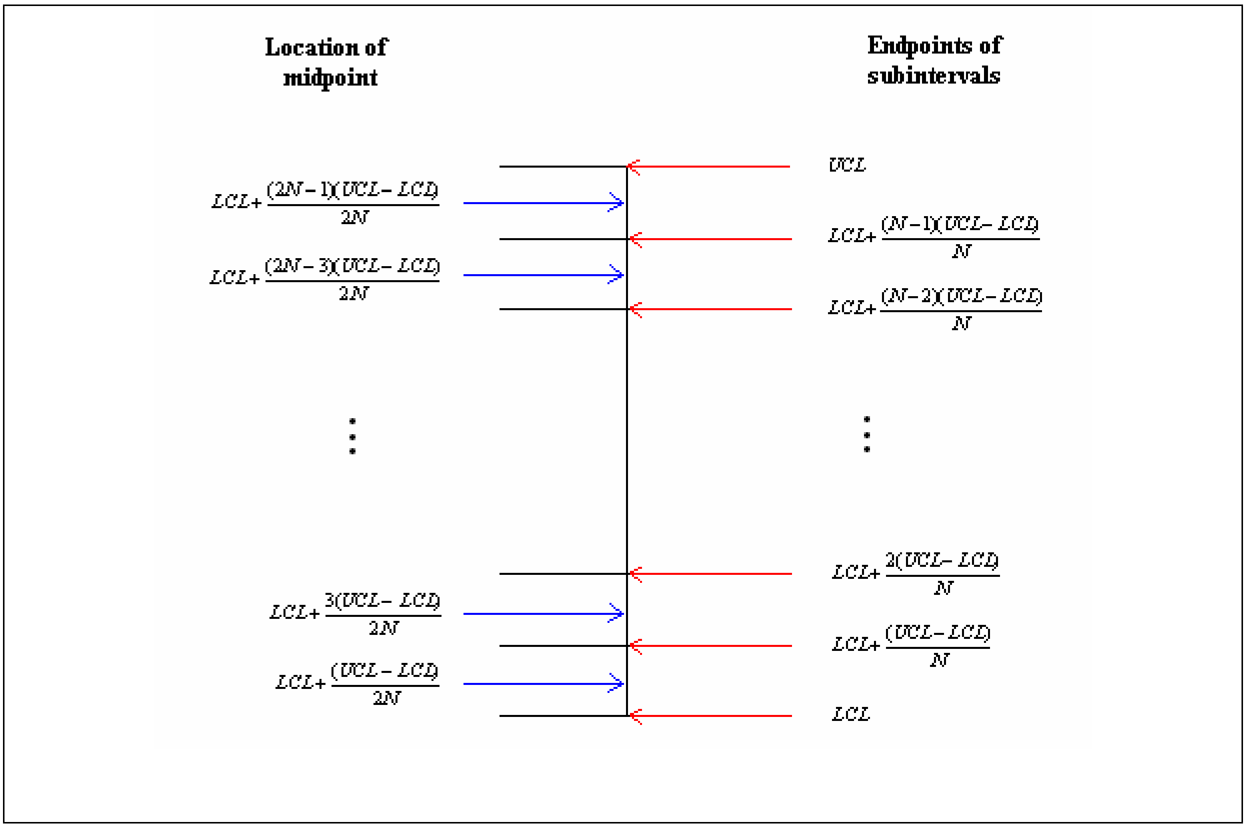

Figure 2.14. Partitioning of the interval between the UCL and the LCL into N subintervals.

Then the plotting statistic, Z_i, is said to be in the non-absorbing state j at time i if

$$S_j - \delta < Z_i \le S_j + \delta \quad \text{for} \quad j = 1,2,...,N-1 \tag{2.60}$$

and

$$S_j - \delta < Z_i < S_j + \delta \quad \text{for} \quad j = N\,. \tag{2.61}$$

Z_i is said to be in the absorbing state if Z_i falls on or outside one of the control limits, that is, $Z_i \le LCL$ or $Z_i \ge UCL$.

Let p_{ij} denote the probability of moving from state i to state j in one step. We have that $p_{ij} = P(\text{Moving to state } j \mid \text{in state } i)$. To calculate this probability we assume that the plotting statistic is equal to S_i whenever it is in state i. For all j non-absorbing we obtain $p_{ij} = P(S_j - \delta < Z_k \le S_j + \delta \mid Z_{k-1} = S_i)$. This is the probability that Z_k is within state j, conditioned on Z_{k-1} being equal to the midpoint of state i. By using the definition of the plotting statistic given in expression (2.57) this transition probability can be written as

$$
\begin{aligned}
p_{ij} &= P\left(S_j - \delta < \lambda SN_k + (1-\lambda)Z_{k-1} \le S_j + \delta \mid Z_{k-1} = S_i\right) \\
&= P\left(S_j - \delta < \lambda SN_k + (1-\lambda)S_i \le S_j + \delta\right) \\
&= P\left(\frac{(S_j - \delta) - (1-\lambda)S_i}{\lambda} < SN_k \le \frac{(S_j + \delta) - (1-\lambda)S_i}{\lambda}\right).
\end{aligned}
$$

Recall that $SN_k = 2T_k - n$ where T_k is binomially distributed with parameters n and $p = P(X_{ij} \ge \theta_0)$

$$
\begin{aligned}
p_{ij} &= P\left(\frac{(S_j - \delta) - (1-\lambda)S_i}{\lambda} < 2T_k - n \le \frac{(S_j + \delta) - (1-\lambda)S_i}{\lambda}\right) \\
&= P\left(\frac{\left(\frac{(S_j - \delta) - (1-\lambda)S_i}{\lambda} + n\right)}{2} < T_k \le \frac{\left(\frac{(S_j + \delta) - (1-\lambda)S_i}{\lambda} + n\right)}{2}\right).
\end{aligned} \tag{2.62}
$$

The probability of transition to the out-of-control state can be determined similarly. For all j absorbing we obtain

$$
\begin{aligned}
p_{ij} &= P\left(Z_k \le LCL \mid Z_{k-1} = S_i\right) + P\left(Z_k \ge UCL \mid Z_{k-1} = S_i\right) \\
&= P\left(\lambda SN_k + (1-\lambda)Z_{k-1} \le LCL \mid Z_{k-1} = S_i\right) + P\left(\lambda SN_k + (1-\lambda)Z_{k-1} \ge UCL \mid Z_{k-1} = S_i\right) \\
&= P\left(\lambda SN_k + (1-\lambda)S_i \le LCL\right) + P\left(\lambda SN_k + (1-\lambda)S_i \ge UCL\right) \\
&= P\left(SN_k \le \frac{LCL - (1-\lambda)S_i}{\lambda}\right) + P\left(SN_k \ge \frac{UCL - (1-\lambda)S_i}{\lambda}\right) \\
&= P\left(2T_k - n \le \frac{LCL - (1-\lambda)S_i}{\lambda}\right) + P\left(2T_k - n \ge \frac{UCL - (1-\lambda)S_i}{\lambda}\right) \\
&= P\left(T_k \le \frac{\left(\frac{LCL - (1-\lambda)S_i}{\lambda} + n\right)}{2}\right) + P\left(T_k \ge \frac{\left(\frac{UCL - (1-\lambda)S_i}{\lambda} + n\right)}{2}\right).
\end{aligned} \tag{2.63}
$$

Since the values δ, λ, n, S_i and S_j are known constants the binomial probabilities in expressions (2.62) and (2.63) can easily be calculated using some type of statistical software package, for example, Excel or SAS.

Once the one-step transition probabilities are calculated, the TPM can be constructed and is given by $TPM = \begin{pmatrix} Q & | & \underline{p} \\ - & - & - \\ \underline{0}' & | & 1 \end{pmatrix}$ (written in partitioned form) where Q is the matrix that contains all the transition probabilities of going from a non-absorbing state to a non-absorbing state. In other words, Q is the transition matrix among the in-control states, $Q:(NA \rightarrow NA)$. $\underline{p}$ contains all the transition probabilities of going from each non-absorbing state to the absorbing states, $\underline{p}:(NA \rightarrow A)$. $\underline{0}' = (0 \quad 0 \quad 0 \quad \cdots \quad 0)$ contains all the transition probabilities of going from the absorbing state to each non-absorbing state, $\underline{0}' : (A \rightarrow NA)$. 1 represents the scalar value one which is the probability of going from the absorbing state to the absorbing state, $1:(A \rightarrow A)$.

Lucas and Saccucci (1990) have investigated some properties of the EWMA chart under the assumptions of independent normally distributed observations. From the similarity between the definitions of the normal EWMA scheme and the sign EWMA scheme, it follows that the formulas derived by Lucas and Saccucci (1990) can be extended to the use of the sign EWMA scheme. The formulas derived by Lucas and Saccucci (1990) have been studied by other authors, for example, Fu, Spiring and Xie (2002) and Fu and Lou (2003). The latter two used the moment generating function and the probability generating function, respectively, to derive expressions for the first and second moments of the run length variable N. See Theorem 2 in Appendix A for the derivations done by Fu, Spiring and Xie (2002) and Fu and Lou (2003). For the formulas refer to equations (2.41) to (2.45) of this thesis.

Example 2.14

The EWMA sign chart where the sample size is even (*n* = 6)

We consider the EWMA sign chart with a smoothing constant of 0.1 ($\lambda = 0.1$) and a multiplier of 3 ($L = 3$). The interval between the UCL and the LCL is divided into 4 subintervals ($N = 4$). For a sample size of 6, the sign statistic SN_i can take on the values $\{-6,-4,-2,0,2,4,6\}$ and the statistic T_i takes on the values $\{0,1,2,3,4,5,6\}$.

The *steady-state* control limits are given in (2.59) by

$$UCL = L\sigma_{SN_i}\sqrt{\left(\frac{\lambda}{2-\lambda}\right)}$$

$$LCL = -L\sigma_{SN_i}\sqrt{\left(\frac{\lambda}{2-\lambda}\right)}$$

where $L = 3$, $\lambda = 0.1$, and $\sigma_{SN_i} = 2.449$, since $\sigma_{SN_i} = 2\sqrt{np(1-p)} = 2\sqrt{6(0.5)(1-0.5)} = 2.449$.

Clearly, we only have to calculate the UCL since $LCL = -UCL$. We have that $UCL = 3\times 2.449\sqrt{\left(\frac{0.1}{2-0.1}\right)} = 1.686$. Therefore, $LCL = -1.686$.

This Markov-chain procedure entails dividing the interval between the UCL and the LCL into subintervals of width 2δ. Figure 2.15 illustrates the partitioning of the interval between the UCL and the LCL into subintervals.

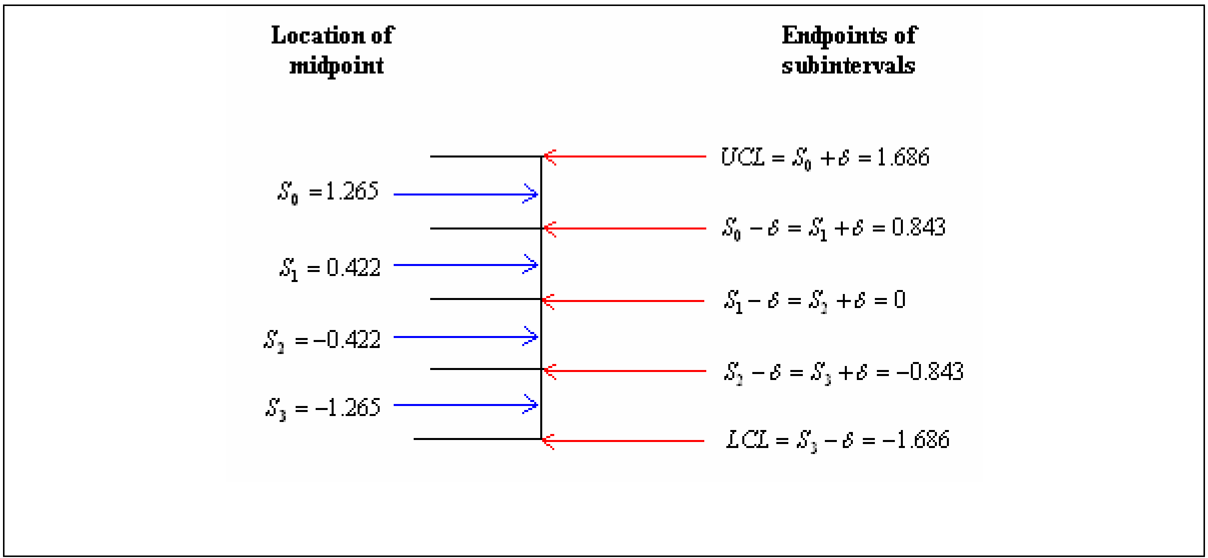

Figure 2.15. Partitioning of the interval between the UCL and the LCL into 4 subintervals.

From Figure 2.15 we see that there are 4 non-absorbing states, i.e. $r = 4$. The TPM is given by

$$TPM_{5\times5} = \begin{pmatrix} p_{00} & p_{01} & p_{02} & p_{03} & p_{04} \\ p_{10} & p_{11} & p_{12} & p_{13} & p_{14} \\ p_{20} & p_{21} & p_{22} & p_{23} & p_{24} \\ p_{30} & p_{31} & p_{32} & p_{33} & p_{34} \\ p_{40} & p_{41} & p_{42} & p_{43} & p_{44} \end{pmatrix} = \begin{pmatrix} Q_{4\times4} & | & \underline{p}_{4\times1} \\ - & - & - \\ \underline{0}'_{1\times4} & | & 1_{1\times1} \end{pmatrix}.$$

The plotting statistic, Z_i, is said to be in the non-absorbing state j at time i if $S_j - \delta < Z_i \le S_j + \delta$ for $j = 0,1,2,3$ where S_j denotes the midpoint of the j^{th} interval. Each sub-interval has a width of $2\delta = 0.843$, therefore $\delta = 0.4215$.

Table 2.38. Calculation of the one-step probabilities in the first row of the TPM.

$p_{00} = P(\textit{Moving to state } 0 \mid \textit{in state } 0)$ $= P(S_0 - \delta < Z_k \leq S_0 + \delta \mid Z_{k-1} = S_0)$ using expression (2.62) we obtain $= P\left(\frac{\left(\frac{(S_0 - \delta) - (1-\lambda)S_0}{\lambda} + n\right)}{2} < T_k \leq \frac{\left(\frac{(S_0 + \delta) - (1-\lambda)S_0}{\lambda} + n\right)}{2}\right)$ with $\delta = 0.4215$, $\lambda = 0.1$, $L = 3$ and $S_0 = 1.265$ $= P(1.525 < T_k \leq 5.740)$ $= P(T_k = 2) + P(T_k = 3) + P(T_k = 4) + P(T_k = 5)$ $= \frac{56}{64}$
$p_{01} = P(\textit{Moving to state } 1 \mid \textit{in state } 0)$ $= P(S_1 - \delta < Z_k \leq S_1 + \delta \mid Z_{k-1} = S_0)$ using expression (2.62) we obtain $= P\left(\frac{\left(\frac{(S_1 - \delta) - (1-\lambda)S_0}{\lambda} + n\right)}{2} < T_k \leq \frac{\left(\frac{(S_1 + \delta) - (1-\lambda)S_0}{\lambda} + n\right)}{2}\right)$ $= P(-2.690 < T_k \leq 1.525)$ $= P(T_k \leq 1)$ $= \frac{7}{64}$

$$p_{02} = P(\text{Moving to state } 2 \mid \text{in state } 0)$$

$$= P(S_2 - \delta < Z_k \le S_2 + \delta \mid Z_{k-1} = S_0)$$

using expression (2.62) we obtain

$$= P\left(\frac{\left(\frac{(S_2 - \delta) - (1-\lambda)S_0}{\lambda} + n \right)}{2} < T_k \le \frac{\left(\frac{(S_2 + \delta) - (1-\lambda)S_0}{\lambda} + n \right)}{2} \right)$$

$$= P(-6.904 < T_k \le -2.690)$$

$$= 0$$

$$p_{03} = P(\text{Moving to state } 3 \mid \text{in state } 0)$$

$$= P(S_3 - \delta < Z_k \le S_3 + \delta \mid Z_{k-1} = S_0)$$

using expression (2.62) we obtain

$$= P\left(\frac{\left(\frac{(S_3 - \delta) - (1-\lambda)S_0}{\lambda} + n \right)}{2} < T_k \le \frac{\left(\frac{(S_3 + \delta) - (1-\lambda)S_0}{\lambda} + n \right)}{2} \right)$$

$$= P(-11.119 < T_k \le -6.904)$$

$$= 0$$

$$p_{04} = P(\text{Moving to state } 4 \mid \text{in state } 0)$$

$$= P(Z_k \le LCL \mid Z_{k-1} = S_0) + P(Z_k \ge UCL \mid Z_{k-1} = S_0)$$

using expression (2.63) we obtain

$$= P\left(T_k \le \frac{\frac{LCL - (1-\lambda)S_0}{\lambda} + n}{2} \right) + P\left(T_k \ge \frac{\frac{UCL - (1-\lambda)S_0}{\lambda} + n}{2} \right)$$

$$= P(T_k \le -11.119) + P(T_k \ge 5.739)$$

$$= 0 + P(T_k = 6)$$

$$= \frac{1}{64}$$

The calculations of the other transition probabilities can be done similarly. Therefore the TPM is given by

$$TPM = \begin{pmatrix} 56/64 & 7/64 & 0 & 0 & 1/64 \\ 1/64 & 56/64 & 7/64 & 0 & 0 \\ 0 & 7/64 & 56/64 & 1/64 & 0 \\ 0 & 0 & 7/64 & 56/64 & 1/64 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} Q_{4\times 4} & | & \underline{p}_{4\times 1} \\ - & - & - \\ \underline{0}'_{1\times 4} & | & 1_{1\times 1} \end{pmatrix}.$$

Other values of the multiplier (L) and the smoothing constant (λ) were also considered and the results are given in Tables 2.39 and 2.40[*].

Table 2.39. The in-control average run length (ARL_0), standard deviation of the run length ($SDRL$), $5^{th}, 25^{th}, 50^{th}, 75^{th}$ and 95^{th} percentile values[†] for the EWMA sign chart when $n = 6$ and $N = 5$, i.e. there are 5 subintervals between the lower and upper control limit[‡].

	$L=1$	$L=2$	$L=3$
$\lambda = 0.05$	6.79 8.56 (1, 1, 3, 9, 24)	22.86 30.29 (1, 3, 10, 31, 85)	736.00 827.24 (4, 134, 472, 1051, 2393)
$\lambda = 0.1$	4.84 5.36 (1, 1, 3, 7, 16)	83.69 104.13 (1, 6, 47, 121, 294)	736.00 819.78 (4, 142, 477, 1049, 2377)
$\lambda = 0.2$	4.73 5.08 (1, 1, 3, 6, 15)	34.12 39.63 (1, 5, 21, 49, 114)	585.80 608.31 (9, 152, 398, 820, 1800)

Similar tables can be constructed by changing the sample size (n), the number of subintervals between the lower and upper control limit (N), the multiplier (L) and the smoothing constant (λ) in the SAS program for the EWMA sign chart given in Appendix B.

[*] These results were calculated through the formulas given in equations (2.41) to (2.45).
[†] The three rows of each cell shows the ARL_0, the $SDRL$, and the percentiles ($\rho_5, \rho_{25}, \rho_{50}, \rho_{75}, \rho_{95}$), respectively.
[‡] See SAS Program 4 in Appendix B for the calculation of the values in Table 2.39.

Table 2.40. The in-control average run length (ARL_0), standard deviation of the run length ($SDRL$), $5^{th}, 25^{th}, 50^{th}, 75^{th}$ and 95^{th} percentile values* for the EWMA sign chart when $n = 10$ and $N = 5$, i.e. there are 5 subintervals between the lower and upper control limit†.

	$L = 1$	$L = 2$	$L = 3$
$\lambda = 0.05$	25.47 31.96 (1, 2, 13, 37, 90)	166.06 207.27 (1, 11, 93, 241, 585)	1773.34 2089.12 (6, 228, 1087, 2557, 5970)
$\lambda = 0.1$	10.35 11.02 (1, 2, 7, 14, 33)	75.61 81.91 (1, 16, 50, 107, 204)	845.42 890.53 (9, 208, 570, 1188, 2624)
$\lambda = 0.2$	3.67 3.68 (1, 1, 2, 5, 11)	25.47 31.96 (1, 2, 13, 37, 90)	272.79 305.97 (1, 51, 176, 389, 886)

From Tables 2.39 and 2.40 we see that the ARL_0, $SDRL$ and percentiles increase as the value of the multiplier (L) increases. In contrast, the ARL_0, $SDRL$ and percentiles decrease as the value of the smoothing constant (λ) increases. From Table 2.40 we find an in-control average run length of 272.79 for $n = 10$ when the multiplier is taken to equal 3 ($L = 3$) and the smoothing constant 0.2 ($\lambda = 0.2$). The chart performance is good, since the attained in-control average run length of 272.79 is in the region of the desired in-control average run length which is generally taken to be 370 or 500.

2.4.4. Summary

EWMA charts are popular control charts; they take advantage of the sequentially (time ordered) accumulating nature of the data arising in a typical SPC environment and are known to be efficient in detecting smaller shifts but are easier to implement than the CUSUM charts. We have described the properties of the EWMA sign chart and given tables for its implementation. Although a lot has been done over the past few years concerning EWMA-type charts, more work is necessary on the practical implementation of the charts as well as on adaptations in case U.

* The three rows of each cell shows the ARL_0, the $SDRL$, and the percentiles ($\rho_5, \rho_{25}, \rho_{50}, \rho_{75}, \rho_{95}$), respectively.

† See SAS Program 4 in Appendix B for the calculation of the values in Table 2.40.