RC 11880 Revised (#53510) 5/13/86 Revised 4/14/87
Mathematics 93 pages

Package CONTRD for Design, Analysis and Running of Cusum-Shewhart Control Schemes.

Emmanuel Yashchin

IBM Corporation
Thomas J. Watson Research Center
Department of Mathematical Sciences
Yorktown Heights, NY 10598.

Abstract

In recent years, Cumulative Sum (Cusum) control schemes (charts) have become increasingly popular in industrial quality control as means for monitoring the quality of manufactured products. This popularity is based on the fact that performance of Cusum schemes is proven to be statistically superior to their classical counterparts - Shewhart schemes ($\bar{x}$ -charts, p-charts, etc.) in the sense that with the same degree of protection against false alarms, they have a much better sensitivity with respect to out-of-control situations. One of the most attractive properties of a Cusum control scheme is its "designability". In other words, once the "good" and "bad" levels of the process as well as corresponding sensitivity requirements are specified, one can come up with a Cusum scheme (and determine the relevant sampling intensity) to meet these requirements. This property of Cusum schemes is especially important in situations where data is collected and/or processed automatically and in situations where several parameters are controlled simultaneously. In the present work we discuss some simple methods for design of one-sided and two-sided Cusum-Shewhart schemes. We introduce the package CONTRD for design, analysis and running of Cusum-Shewhart schemes and give examples of its application.

Contents

List of abbreviations

d.f.	distribution function
iid	independent and identically distributed
r.v.	random variable
s.d.	standard deviation
Cusum	Cumulative Sum
RL	Run Length
ARL	Average Run Length
SDRL	Standard Deviation of the Run Length
P(UP)	probability that the signal is triggered by the upper scheme
TS	Time elapsed before an out-of-control Signal is triggered
SI	Sampling Intensity
ATS	Average Time elapsed before an out-of-control Signal is triggered
$T[i]$	the i-th component of vector T

1. Introduction. Control schemes and characterization of their performance

Let x_1, x_2, ... be a sequence of observations related to a certain process. The observation x_i may represent, for example

- sample percentage of defective chips in the i^{th} produced lot;
- total number of defects found in the i^{th} produced wafer;
- sample mean of 4 diameters of ball bearings chosen at random during the i^{th} production period;
- sample standard deviation of 10 simultaneous measurements (corresponding to various locations) of polyethylene film thickness taken during the i^{th} sampling period;
- waiting time of the i^{th} customer in the queue;
- discrepancy between the actual amount of product shipped in the i^{th} month and that predicted by a given model,

and so on - for purposes of our discussion the nature of incoming observations is immaterial. In most practical situations we would like our observations to behave in a certain way, ex. to fall as close as possible to some target value, to stay below some prescribed limit, etc. Failure of the observations to comply with this desired behavior is considered as an out-of-control situation; we would like to detect such behavior as early as possible.

In order to monitor sequences of observations we use control schemes. A control scheme is a set of criteria in order to test, at any given moment of time whether the process generating the observations is under control. Clearly, many different control schemes can be associated with the same sequence of observations; some of the better known include Shewhart schemes, Moving Average schemes of various types, etc. In order to compare different types of schemes we need to introduce some criterion of performance of a control scheme. The most important one is represented by the Run Length (RL) of a scheme. If the input observations correspond to on-target situation, we would like the RL to be as long as possible; otherwise, it should be as short as possible. Since the RL is a random variable, the

actual comparison between control schemes is usually based on some of its characteristics, such as Average Run Length (ARL), Median or some other quantile of the Run Length, etc.

For example, let us assume that the observations are independent, identically distributed (iid) and normal with mean μ and s.d. $\sigma = 1$. The target level of the process is $\mu = 0$. Let us draw an ARL curve as a function of the process level μ for a Shewhart scheme (signal is triggered if a single observation falls above 3) and for an (unspecified) Cusum scheme (Fig. 1.1).

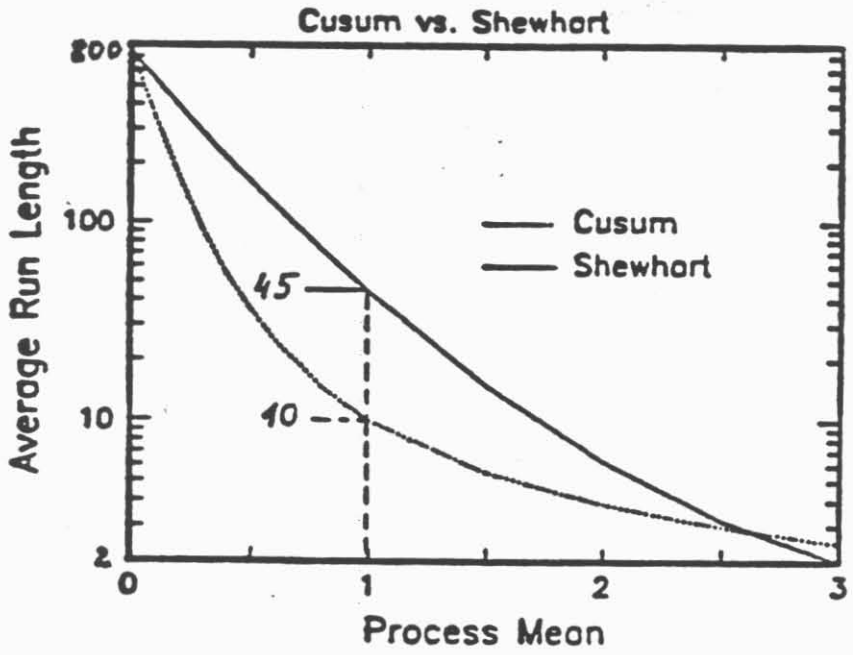

Fig. 1.1. Comparison between Cusum and Shewhart schemes.

Thus, if the process is on target, both schemes have roughly the same degree of protection against false alarms (ARL $\approx$ 740). However, as the process level shifts, the Cusum scheme becomes much more sensitive. For example, for $\mu = 1$ we have ARL(Cusum) $\approx$ 10 while ARL(Shewhart)$\approx$ 45. Another interesting question is as follows: if we took n observations at a time and applied a Shewhart scheme to their sample means, how large should n be to assure the same sensitivity at $\mu = 1$ as our Cusum scheme? One can show that to achieve that we need to take $n = 3$; a direct conclusion is that in some situations by using a Cusum scheme instead of a Shewhart one can reduce the sampling in-

tensity by a factor of 3 and still keep the same "resolution" between "good" and "bad" levels of the process.[1]

In the present work we consider a Cusum - Shewhart class of control schemes. These schemes assure about the best possible sensitivity for a given level of protection against false alarms, and, in addition, possess certain desirable features listed, for example, in Yashchin (1985a, p. 378). First of all, Cusum - Shewhart schemes are "analyzable". In other words, it is possible to examine, by analytic means, the RL behavior of a scheme for any given stochastic pattern of incoming (iid) observations; approximate results for some non-iid cases are also available (ex. see Bagshaw and Johnson (1974, 1975)). Another important quality is "designability". Indeed, once the "good" and "bad" levels of the process as well as corresponding sensitivity requirements are specified, one can, in a relatively straightforward way, design a Cusum - Shewhart scheme and determine the relevant sampling intensity to meet these requirements (see Woodall (1985 a, b), Yashchin (1985a)). Since analysis of this type of schemes is associated with an extensive computational effort, including matrix analysis, both problems of analysis and design are hardly treatable unless an appropriate software package is available. In the present work we introduce such package (CONTRD, previously called DARCS) and give several examples of its application. A separate package CONTRP for plotting of Cusum-Shewhart schemes is presently under testing and will be described in a forthcoming report.

Analysis of the RL and careful design of control schemes are especially important in situations where measurements are taken and processed automatically and/or where several parameters are controlled simultaneously. In such situations frequent out-of-control signals associated with *practically* non-important changes in process parameters may cause frequent unjustified corrective actions and/or eventually ruin the discipline of the operator; on the other hand, failure to detect a truly out-of-control situation rapidly may result in a substantial amount of poor-quality product.

---

[1] Formally, one can define the "resolution" of a control scheme, for example, as ratio between the ARL's corresponding to acceptable ("good") and unacceptable ("bad") levels of the process (these levels are determined on the basis of practical and/or economical considerations). In situations where the sampling interval is not a fixed number, it is natural to characterize the performance of a scheme in terms of the Time to Signal (TS) instead of the Run Length; in such cases one can define the resolution as a ratio between the ATS's (Average Time to Signal) corresponding to "good" and "bad" levels of the process.

For example, consider the following situation related to the production of surface-mounted printed circuit boards. Assume that a board has 400 pads each containing a certain amount of solder paste deposited by squeezing it through a mask. Before mounting the components onto the board and re-flowing the solder, the volumes of solder paste on each pad are measured by an optical scanner. If the measurements corresponding to some pad show an erratic behavior (which may be caused, for example, by a partially clogged slot in the mask), an out-of-control signal is triggered. It is clear that use of a 3-sigma Shewhart scheme to control the subsequent volumes on a pad would result, on the average, in one false out-of-control signal per board! (Indeed, Fig. 1.1 implies that the on-target ARL for a two-sided Shewhart scheme is approximately 740/2=370). So, if we wanted the probability of a false alarm within an 8-hour shift not to exceed 5%, we should have undergone the appropriate design and analysis procedure. The final control scheme would probably represent some kind of a compromise between the desired sensitivity and degree of protection against false alarms.

This example makes it clear that one cannot blindly apply standard control schemes considered in some Quality Control textbooks to situations involving simultaneous control of several parameters. Yet, such situations are rather common in modern industry, and it is not unusual to see thousands of sequences monitored simultaneously. To summarize, *any control scheme associated with automatic data processing and/or simultaneous control of several parameters should be thoroughly analyzed before it can be recommended for use*. The analysis should involve identification of various possible joint distributions of observations and investigation of the corresponding run length distributions. Its ultimate aim is to assure that the run length of the scheme under consideration is sufficiently long if the changes in process parameters are not *practically* important and sufficiently short if they are.

In the context of modern process control, another property of a control scheme becomes crucial, namely, its capability to incorporate new information immediately upon its arrival, and update itself accordingly. This criterion corresponds to one of the weakest points of Shewhart control schemes, which are typically associated with first subgrouping observations into samples and only then updating the scheme. Clearly, in situations where observations (measurements) are not "naturally" grouped, but rather arrive one at a time, such artificial subgrouping leads to waste of time and loss

of resolution power of the scheme; it is not inherently tied to the problem of control itself. One of the main reasons for creating artificial samples when running Shewhart schemes is related to concern that individual observations may have other than normal distribution; by using sample averages one could bring the scheme characteristics closer to those predicted by the normal model.[2]

As the reader will see from the next section, in the case of Cusum - Shewhart schemes the process of cumulative summation itself brings us (by virtue of the Central Limit Theory) into the normal domain, eliminating any necessity for artificial grouping. In general, every scheme considered in the present work is based on the principle of *immediate utilization of incoming information;* the term "sample size" will typically refer to a group of observations (measurements) arriving into the control system at the same moment of time.

In practical applications, it is not always clear what actions should be taken as a result of an out-of-control signal. The strictest one calls for an immediate stopping of the production process until the situation is clarified and the problems (if any) dealt with. Another possibility would be to increase the sampling intensity and/or switch to a tighter mode of operation which, in turn, could lead to either more drastic actions or return to the normal operating mode, depending on subsequent behaviour of the process. In situations related to automated control (ex. robot control) one could try to estimate the current level of the process of observations and introduce a correction by means of a feedback loop. Clearly, many other actions could be suggested; the actual choice will always depend primarily on the specific nature of the situation. Unfortunately, the scope of the present work does not enable us to discuss in detail the questions related to actions following an out-of-control signal as well as many other important aspects of the Cusum technique; the reader will undoubtedly find useful the monographs by van Dobben de Bruyn (1968) and Woodward, R. and Goldsmith (1964), Bissell (1969) and guide by the British Standards Institution (1980-1983).

We believe that because of their excellent statistical properties, "designability", easy visual interpretation and other important features, Cusum-Shewhart control schemes will become a dominant tool

---

[2] Other reasons for not updating the schemes immediately may be related to cost of information processing of even some statistical considerations (see Example B.6, Appendix B).

for on-line process control in the coming years. Our hope is that engineers working in the area of quality control will find the package CONTRD a helpful and easy to use tool for design, analysis and running of this type of control schemes.

2. Cusum and Cusum – Shewhart control schemes. Page's and V-mask graphical representations of a Cusum – Shewhart scheme

In this section we give a short reminder on application of some typical Cusum-Shewhart schemes to our sequence of observations $x_1, x_2, \ldots$. As we shall see, these schemes can be used in one of two modes: Page's mode and a V-mask mode. We start by introducing the upper Page's scheme.

(i) Upper Page's scheme

Let us suppose that we are primarily concerned about the possibility that the process might shift up towards an unacceptable level (typical example - monitoring sample proportions of defectives in successive lots). Upper Page's schemes represent a type of Cusum control schemes that can be used to detect the presence of such conditions. The scheme is defined in terms of three parameters: $h^+ \geq 0$ (signal level), k^+ (reference value) and $0 \leq s_0^+ \leq h^+$ (headstart). It is applied as follows:

a) Start from s_0^+ and compute the sequence of cumulative sums:

$$s_i^+ = \max \{s_{i-1}^+ + (x_i - k^+), 0\}, \quad i = 1, 2, \ldots \tag{2.1}$$

b) If N^+ is the first index i for which $s_i^+ > h^+$, trigger an out-of-control signal at time N^+.

Note that N^+ represents the RL of the scheme. If an additional signal criterion is introduced, namely

c) If a single observation x_i satisfies $x_i > c^+$, trigger the out-of-control signal at the moment i,

the procedure is called an upper Page's scheme with parameters (h^+, k^+, s_0^+) supplemented by Shewhart's limit c^+.[3] Here and in what follows we refer to such (supplemented) Page's schemes as Cusum-Shewhart control schemes.

---

[3] It is clear that in order to affect the performance of the control scheme the Shewhart's limit must satisfy $c^+ < h^+ + k^+$. Also, if $c^+ \leq k^+$, then an out of control signal can be triggered only if the Shewhart's limit has been violated i.e. we obtain a pure Shewhart scheme with upper control limit c^+.

Let us clarify the roles of the parameters in a Cusum-Shewhart scheme. The reference value k^+ is usually chosen to be close to the midpoint between the acceptable and unacceptable levels of the process, as shown in Fig. 2.1. Thus, it acts as an "anchor" keeping the scheme from drifting in on-target situations. On the other hand, if the process level is unacceptable, the successive differences $(x_i - k^+)$ become typically positive, they accumulate in (2.1) causing the scheme to eventually "float up" and signal.

The signal level h^+ characterizes the degree of accumulation of information allowed in the control scheme. If $h^+ = 0$, we do not allow any accumulation of evidence against the on-target hypothesis and are prepared to signal on the basis of a single observation - in other words our Cusum scheme turns into a pure Shewhart scheme with upper control limit k^+.

The headstart s_0^+ implements the Fast Initial Response feature, i.e. it provides an instrument for detecting *initially present* out-of-control conditions earlier than similar conditions occurring later. The rationale for using a headstart is as follows: if the process is on target, the Page's scheme will be (most likely) brought to zero by the reference value, so that in this case the expected effect of the headstart is minimal; otherwise, however, the out-of-control will be triggered much sooner (ex. see Lucas and Crossier (1982)). Finally, supplementing the scheme by a Shewhart's limit improves the sensitivity of the scheme with respect to substantial increases in the process level - in other words, it removes some of the "inertia" of a Cusum scheme when facing a sharp change of the process (ex. see Lucas (1982)). There are also cases in which Shewhart's limits are introduced because of some special features of the associated production process or other considerations.

Note that schemes based on only two parameters, signal level and reference value, are frequently found quite satisfactory for practical purposes.

(ii) **Lower and two-sided Page's schemes**

Now assume that we are primarily concerned about the possibility that the process might shift down to some unacceptable lower level (typical example - monitoring successive inter-failure times of an electronic device). A natural way to monitor such sequences is to apply an *upper* Page's scheme with

parameters $h^- \geq 0$, k^-, $0 \leq s_0^- \leq h^-$ and c^- to the sequence of "reflected" observations, $-x_1$, $-x_2$, ... Such procedure defines a *lower* Page's scheme:

$$s_i^- = \max \{s_{i-1}^- + (-x_i - k^-), 0\}, \quad i = 1, 2, \ldots, \tag{2.2}$$

signal if $s_i^- > h^-$. In accordance with our recommendations regarding choice of the reference value, $(-k^-)$ should be chosen close to the "midway" between the acceptable and (lower) unacceptable process level (Fig. 2.1). Analogously, if the lower scheme is supplemented by a Shewhart limit, an immediate out-of-control signal should be triggered if $(-x_i) > c^-$. Note that if the target level of our sequence is 0, the reference values (and Shewhart limits, if present) of both upper and lower schemes will be positive.

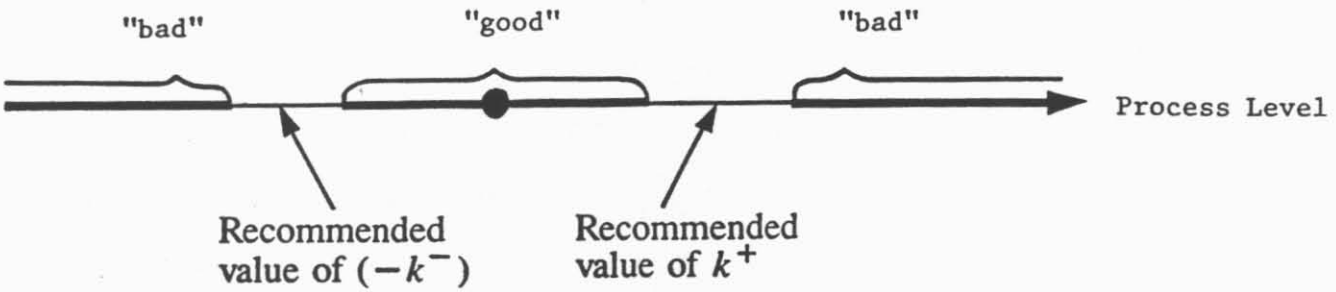

Fig. 2.1. Choice of the reference values.

In situations where we would like to detect rapidly both types of shift of process from its target level, it makes sense to run both schemes simultaneously and to trigger an out-of-control signal as soon as one of the one-sided schemes signals. This procedure will be called a two-sided Page's scheme with parameters $(h^+, k^+, s_0^+, h^-, k^-, s_0^-)$, possibly supplemented by Shewhart's limits (c^-, c^+). It is clear that we must always have $(-c^-) < c^+$. Moreover, Fig. 2.1. indicates that for all "reasonable" two-sided schemes $(-k^-) < k^+$; in what follows, we shall always assume that this condition is satisfied.

To illustrate the use of Page's schemes, let us consider the following example.

Example 2.1 In the oxidation process of silicon wafers, we are interested in keeping the difference between the actual mean thickness of the grown SiO_2 layer and its target value (we denote this difference by Δ) as close to zero as possible. In order to achieve that, we take n measurements of film thickness per lot and test for presence of significant systematic deviation between the actual mean thickness and the target value. Let the measurements corresponding to the i^{th} lot be $y_{i1}, y_{i2}, \ldots y_{in}$ (denote their average by $\bar{y}_{i.}$) and let us base our control scheme on the sequence $x_1, x_2, \ldots$, where x_i is the difference between $\bar{y}_{i.}$ and the target value.

The consequences of systematic deviations between the actual mean thickness and its target level depend not only on the magnitude but also on the sign of the deviation. So, we would like to guard ourselves against situations in which Δ is more than 6Å or less than (-4Å). Let us apply the two-sided Page's scheme with parameters ($h^+ = 9,\ k^+ = 3,\ s_0^+ = 2,\ h^- = 5,\ k^- = 2,\ s_0^- = 1$) to the sequence of observations $x_1, x_2, \ldots$ presented in Table 2.1. The resulting chart is given by Fig. 2.3, a), and the values of the one-sided schemes correspond to columns 3 and 4 of the mentioned table. The out-of-control signal is triggered by the upper scheme at time $i = 40$.

In the situation described in the above example, one would usually try to control not only the mean (level) of the sequence, but also σ, the variability within each sample. This can be done by means of a Cumulative $\hat{\sigma}$- chart, where $\hat{\sigma}$ is the sample standard deviation:

$$\hat{\sigma}_i = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2}\ ,\quad i = 1,\ 2,\ \ldots . \tag{2.3}$$

We do not have a particular "target" value for the sequence $\hat{\sigma}_1, \hat{\sigma}_2, \ldots$; instead, we have a "target" region, namely, we want the underlying "true" standard deviation σ to lie within the interval $0 \le \sigma \le 2$Å. On the other hand, we would like to detect as quickly as possible the situation in which $\sigma \ge 4$Å. Let us apply the scheme $\{ h = 5$Å $k = 3$Å, $s_0 = 1$Å $\}$ to our sequence $\hat{\sigma}_1, \hat{\sigma}_2, \ldots$. In the last two columns of Table 2.1 we give the observed realization of this sequence and the corresponding values of the Page's scheme, $s_1, s_2, \ldots$.[4]

---

[4] Note that for both considered schemes the reference values were chosen about the midway between "good" and "bad" levels of the process.

Finally, in situations of this type there usually are other sources of variability of interest, for example wafer-to-wafer variability within a lot, lot-to-lot variability, etc. These components of variability should be controlled separately by using appropriate sequences of estimators. It is natural to choose the corresponding target regions so as to maintain the total variability (which is a primary factor that determines the process escape rate) sufficiently small. In general, it is also a good practice to maintain a separate chart for controlling the total variability.

To illustrate the above point, let us assume that the underlying mean of the population corresponding to the i-th sample is itself a random variable with mean 0 and standard deviation σ_b. If the within-sample variability σ is small compared to σ_b and generally behaves in a stable way, an appropriate procedure for controlling the lot-to-lot variability could be based on the sequence $d_i = 0.5\sqrt{\pi}\ |\ \bar{x}_{i+1} - \bar{x}_i\ |\ ,\ \ i = 1, 2, \ldots$. Indeed, it is well known that the expected level of this sequence is

$$E(d_i) = \sqrt{\sigma_b^2 + \frac{\sigma^2}{n}} = \sigma_b\{1 + \frac{1}{2n}(\sigma/\sigma_b)^2 - \frac{1}{8n^2}(\sigma/\sigma_b)^4 + \cdots\}; \qquad (2.4)$$

thus, for situations of interest its relative bias is rather small. For example, if $n=5$ and $\sigma/\sigma_b \leq 0.5$, it does not exceed $0.5^2/(2 \times 5) = 0.025$. In situations where σ is not sufficiently small and, consequently, plays a significant role in determining the level of our sequence, an alternative approach is needed; detailed analysis of such situations, however, falls beyond the scope of the present work.

(iii) **An alternative approach: V-mask scheme**

In this subsection we consider an alternative way of applying a Page's schemes to a sequence of observations - V-mask schemes. For purposes of control, both types of schemes are completely equivalent; the difference is only in the graphical representation.

Suppose, as previously, that we observe the process $x_1, x_2, \ldots$. Let us define the cumulative sum process $c_0, c_1, \ldots$ by

$$c_0 = 0, \quad c_i = \sum_{j=1}^{i} (x_j - t_0), \quad i = 1, 2, \ \ldots, \tag{2.5}$$

where t_0 is some "convenient" constant (ex. target value; as we shall see later, this constant does not play any role in the control procedure itself, but rather serves for convenience of graphical interpretation only). Further, let us plot the resulting values c_i against i and construct a mask with parameters (h^+, k^+, h^-, k^-) as shown in Fig. 2.2. Note that if the horizontal line is marked as having slope t_0, then k^+ and $-k^-$ represent the *slopes* of the lower and upper arms of the mask, respectively. If our scheme is supplemented by Shewhart limits (c^-, c^+), the mask will be slightly "parabolized" near the origin. Now let us apply the V-mask to the cumulative sum as shown in Fig. 2.2, and trigger an out-of-control signal at the first moment the Cusum path fails to fall within the arms of the V-mask; for one-sided control we apply the appropriate half of the V-mask only. To implement the Fast Initial Response feature, we put two artificial observations $(0, -s_0^+)$ and $(0, s_0^-)$ onto the chart and trigger a signal also if one of these observations falls outside the mask.

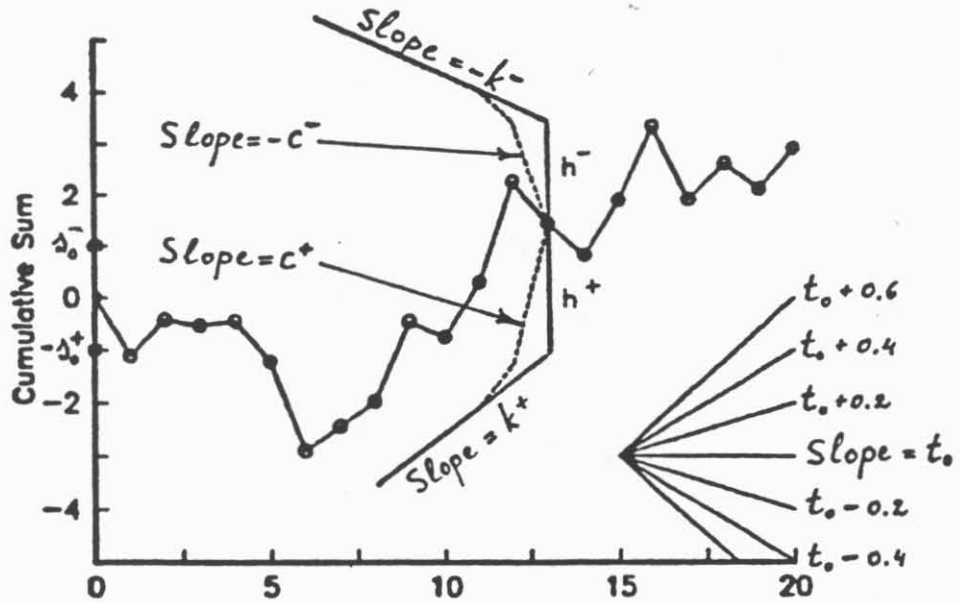

Fig. 2.2. The V-mask scheme.

The described procedure is called a V-mask scheme. As we mentioned earlier, it is completely equivalent to a Page's scheme in the sense that one of the schemes signals at some moment of time if and only if the other one does (ex. see Duncan (1974, p.469)). As an example, let us apply a V-mask scheme to the chart (i, c_i) associated with observations x_1, x_2, ... from Table 2.1 (see Fig. 2.3, b)). As one can see, both schemes signal at the same time $i = 40$.[5]

The main benefits of the method of Cusum plotting used in the V-mask scheme are related to convenience for purposes of graphical data analysis. Indeed, the cumulative sum trajectory represents a natural instrument for smoothing the data without loss of information in the sense that it enables immediate restoration of individual observations. Its slope at any moment of time corresponds to the current level of the process of observations; it can be easily estimated by means of a protractor (slope guide). The latter can be used either as shown in the bottom right part of Fig. 2.2 or as in Fig. 2.3,b, where slopes corresponding to rays of the protractor are displayed instead of the values of cumulative sum. Clearly, these values can be easily restored, since the cusum path always starts at the origin and the protractor size is known. This type of display is especially convenient for automated data processing, as the protractor is always located in the same place. In addition, Cusum plot enables immediate evaluation of the average of observations within any given interval of time by connecting the ends of the Cusum path by the ruler and matching its slope with an appropriate ray of the protractor.

One of the main drawbacks is related to the fact that the Cusum path is not limited to a horizontal strip of paper (screen) and it can run out of the prescribed margins. There are several ways to overcome this difficulty (ex. by re-initiating the chart once it runs out of prescribed margins), but all of them come at the expense of convenience of visual evaluation. An additional drawback is related to necessity to specify the value of t_0 in order to ensure approximate horizontality of the "on-target" Cusum path. In addition, special scaling is required in order to have "reasonable" angles of the V-mask. In some cases, this requirement represents a nuisance, especially when a standard-grid graph

---

[5] Readers familiar with this topic will notice that in statistical literature V-masks are typically defined in terms of so called "leading distances" and angles of the mask arms (ex. see Duncan (1974, p.470)). Our definition has several important advantages. First, our parameters are invariant with respect to scaling of the axes. In addition, our parameters are *the same* for both types of Cusum schemes.

i	x_i	s_i^+	s_i^-	c_i	$\hat{\sigma}_i$	s_i
1	-4.0	0	3	-4.0	2.0	0
2	-1.0	0	2	-5.0	2.5	0
3	3.0	0	0	-2.0	3.5	0.5
4	-2.0	0	0	-4.0	4.0	1.5
5	-2.5	0	0.5	-6.5	4.5	3.0
6	-0.5	0	0	-7.0	2.5	2.5
7	1.5	0	0	-5.5	3.0	2.5
8	-3.0	0	1	-8.5	4.5	4.0
9	4.0	1.0	0	-4.5	2.5	3.5
10	3.5	1.5	0	-1.0	1.5	2.0
11	-2.5	0	0.5	-3.5	2.0	1.0
12	-3.0	0	1.5	-6.5	2.5	0.5
13	-3.0	0	2.5	-9.5	2.0	0
14	-0.5	0	1.0	-10.0	2.5	0
15	-2.5	0	1.5	-12.5	2.5	0
16	1.0	0	0	-11.5	3.5	0.5
17	-1.0	0	0	-12.5	4.5	2.0
18	-3.0	0	1	-15.5	4.5	3.5
19	1.0	0	0	-14.5	2.0	2.5
20	4.5	1.5	0	-10.0	1.5	1.0
21	-3.5	0	1.5	-13.5	2.5	0.5
22	-3.0	0	2.5	-16.5	4.5	2.0
23	-1.0	0	1.5	-17.5	4.5	3.5
24	4.0	1.0	0	-13.5	3.5	4.0
25	-0.5	0	0	-14.0	3.5	4.5
26	-2.5	0	0.5	-16.5	2.5	4.0
27	4.0	1.0	0	-12.5	2.0	3.0
28	-2.0	0	0	-14.5	2.0	2.0
29	-3.0	0	1.0	-17.5	2.5	1.5
30	-1.5	0	0.5	-19.0	3.0	1.5
31	4.0	1.0	0	-15.0	3.5	2.0
32	2.5	0.5	0	-12.5	4.0	3.0
33	-0.5	0	0	-13.0	1.5	1.5
34	7.0	4.0	0	-6.0	2.0	0.5
35	5.0	6.0	0	-1.0	2.0	0
36	4.0	7.0	0	3.0	3.0	0
37	4.5	8.5	0	7.5	2.5	0
38	2.5	8.0	0	10.0	3.5	0.5
39	2.5	7.5	0	12.5	2.0	0
40	5.0	9.5	0	17.5	4.0	1.0

Table 2.1 The observed values of sample means, $\{x_i\}$, corresponding values $\{s_i^+, s_i^-\}$ of the scheme $\{h^+ = 9, k^+ = 3, s_0^+ = 2, h^- = 5, k^- = 2, s_0^- = 1\}$ and $\{c_i\}$ of the "pure" CUSUM (the first five columns); The observed values of the sample standard deviations, $\{\hat{\sigma}_i\}$, and the corresponding values $\{s_i\}$ of the one-sided scheme $\{h = 5, k = 3, s_0 = 1\}$ (the last two columns).

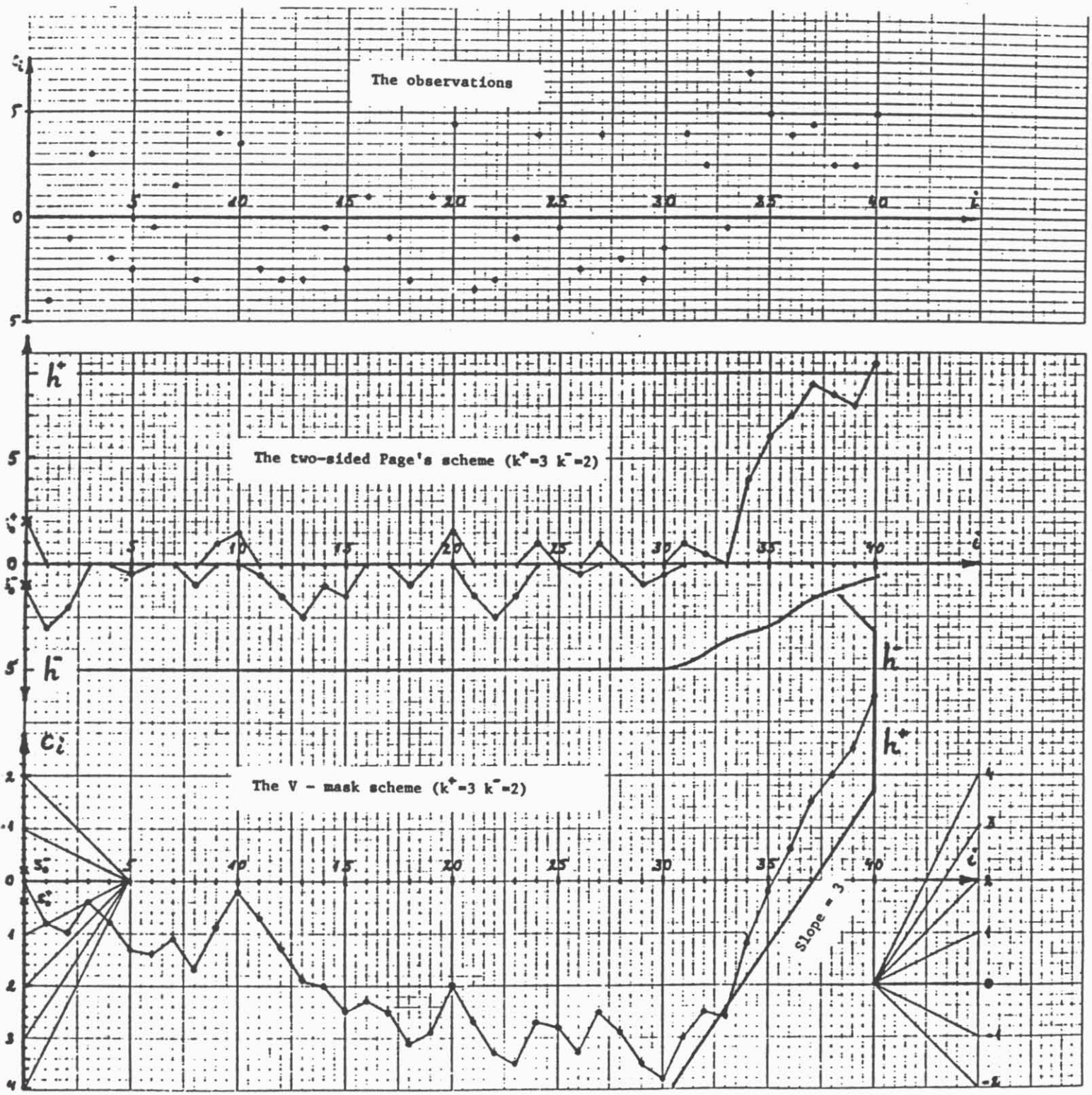

Fig. 2.3. Cusum control schemes. The scale symbols of c_i - axis correspond to slopes. The tic marks on the horizontal ray of the protractor can be used to obtain multiple values of the slopes.

paper is used for plotting. However, our experience shows that none of these drawbacks can be considered as serious in the context of computerized (esp. interactive) plotting. Page's scheme is free of the above drawbacks, but it is less informative since the only information we keep is that required for control purposes. In what follows, we shall work in terms of Page's schemes only; however, the user should bear in mind that all the results can be applied directly to cases in which V-mask is used.

3. The structure of package CONTRD. Typical outputs of the functions for analysis and design

In this section we give a description of APL/APL2 package (workspace) CONTRD for design, analysis and running of Cusum-Shewhart control schemes developed recently in the Department of Mathematical Sciences of Thomas J. Watson Research Center (IBM). It represents a substantially modified and enhanced version of the package DARCS described in Yashchin (1985a). CONTRD can be loaded either by typing)LOAD CONTRD after entering manually the APL/APL2 environment, or by using an exec CONTRD which is supplied with the workspace. To exit from the workspace, one should type)OFF. One of the important features is that there is no need to know APL or to have an APL keyboard in order to use the functions of CONTRD. These functions include:

- Special functions CUSUMC, CUSUMP, CUSUMS, CUSUMT and CUSUMX for design and analysis of Cusum c-charts, p-charts, s-charts, Time-Between-Events (t)-charts and $\bar{x}$ – charts, respectively. This set of functions is sufficient for many users of Cusum schemes;

- Distribution functions of some commonly used random variables; every function whose name starts with letters DF is one of such functions (ex. DFNORM, DFBINOM, etc.). They are used to specify the nature of incoming observations when analyzing properties of the scheme. A full list of provided distribution functions can be found in Appendix A.

- Functions for analysis of upper Page's schemes (ONEAN, ONEVARY, ONEXPLR) and lower schemes (ONEANL, ONEVARYL, ONEXPLRL);

- Functions for design of upper Page's schemes (ONEFIND, ONEXPLRD) and lower schemes (ONEFINDL, ONEXPLRDL);
- Functions for analysis of two-sided Page's schemes (TWOAN, TWOVARY, TWOXPLR);
- Functions for running one- and two-sided Page's schemes (ONERUN, ONERUNL, TWORUN);
- Functions for special purposes (ASSIGN, IDENTIFY, QUIT, RESET, SET, SETI, SETFIND, SETVARY, SETXPLR);
- Other functions (primarily for generating various types of random variables and sequences, and statistical analysis).

The package has a complete internal documentation (functions DESCRIBE and HELP). For example, information related to the function CUSUMC can be obtained by typing HELP 'CUSUMC'. The function HELP can also be used to obtain information on the general structure of the package, list of available distributions, list of abbreviations, etc.

Consider the functions related to design and/or analysis of one-sided schemes. A function for analysis usually results in a table including confirmation of scheme parameters as well as the results of analysis. It typically looks as follows:

```
Analysis of upper Cusum scheme with parameters  H,K = 3 1
The level of discretization is 30
The observations are normal with  SIGMA=1
The changing parameter name is MEU
The scheme is supplemented by the Shewhart limit 3.5

MEU      ARL     SDRL        5       10       20       50
0.0   1507.3   1505.4   .99760   .99430   .98772   .96823
0.5    111.0    108.2   .97515   .93225   .84995   .64403
1.0     17.1     14.1   .84008   .59530   .29071   .03376
1.5      6.3      3.9   .48629   .12590   .00780   .00000
```

As one can see, this output is quite self-explanatory. For example, it says that if the process level is $\mu = 1.5$, the corresponding ARL and SDRL (Average and Standard Deviation of the Run Length) are

6.3 and 3.9, respectively; the probability that the Run Length will be greater than 5 is 0.48629, the probability that it will be greater than 20 is 0.00780, etc.[6]

For purposes of analysis, we assume that the interval $(0, h)$ is subdivided into d groups having the same length δ (except the group containing 0) and, at each step, the values of the Page's scheme are rounded to a center of a corresponding group. We refer to d and δ as the *level* and *interval* of discretization, respectively. The notion of discretization is discussed in detail in Appendix C.

The output table corresponding to two-sided schemes is analogous. In addition to columns shown above, it has a column P(UP) containing the probabilities that the signal is triggered by the upper scheme. Also, some additional information related to the nature of the two-sided scheme (presence and power of intrinsic interaction, etc.) may appear. For details, see Yashchin (1985a, pp. 381-384) or Yashchin (1985b).

The functions for design of one-sided schemes perform a search for a signal level, h, for which the ARL (or specified Quantile of the Run Length) is equal to a specified number. Other parameters of the scheme are fixed; they are either derived automatically (as in special functions CUSUMC, CUSUMP, etc.) or specified by the user. Analogously, the initial approximation h_0 for the search procedure can be chosen automatically (special functions) or should be provided by the user. A typical output of a design function is as follows:

```
Search for the value of H satisfying  ARL=3.9
The length of interval of discretization is 0.53301
The observations are distributed as X-bar with MEU=110 SIGMA=10 SAMPLE=3
H=15.7238  Level of Discr.=30  ARL=3.89715
H=16.2568  Level of Discr.=31  ARL=4.00394
The interpolated value of H is 15.738
```

This output corresponds to Example 5.6 considered later (in this example $h_0 = 12$, $k = 105$). It shows that the procedure of search for h results in an interval. If the design procedure is automatically followed by an analysis of the resulting scheme, the signal level for which the ARL is closer to the target value will be chosen. So, since in our case the ARL=3.897 for $h = 15.738$ is closer to the

---

[6] Note that the last displayed table is always stored in the workspace under the name TABLE. So, to see some additional significant digits of the output table, one should type TABLE.

desired value, 3.9, it will be chosen, by default, for the subsequent automated analysis. Note that this default mode can be changed (Sec. 5.6).

In cases where choice of the initial approximation, h_0 has been grossly unsuccessful, the search procedure may fail. In such cases the last approximation and other related information is displayed; the user should use this information in order to suggest an alternative initial approximation and repeat the relevant function.

The information specific to functions for design and analysis of general Cusum-Shewhart schemes will be provided in Sec. 5.

4. Special functions for design and analysis of commonly used control charts

In this section we consider the special functions for design and analysis of Cusum schemes for several types of "popular" control situations - control of the mean and standard deviation of a normal population, sample proportion of defectives, and so on. These functions have a unified format and are very easy to use. We feel that they may satisfy the needs of a substantial proportion of Cusum users. First, we describe the common features of this group of functions and then provide more specific information and examples. Several sample runs of special function can be found in Appendix B.

4.1. Some common features

Every function has a right argument, vector R, specifying the requirements of the design procedure, and a left argument, vector L, containing such information as sample (subgroup) size, acceptable and non-acceptable levels of the controlled parameter, and standard deviation of a single measurement. The last two components of L are always optional and will be discussed later.

At the first step (design) the function automatically picks an appropriate reference value, k ,[7] and then performs a search for the value of the signal level h for which the *on-target* ARL (or *on-target*

---

[7] The choice of k is usually based on likelihood ratio considerations, which assures about the best possible resolution power as well as certain asymptotic optimality properties (ex. see Lorden (1971)). The reader familiar with the theory of Sequential Probability Ratio Tests (SPRT's) will notice that for situations corresponding to special functions the structure of SPRT is analogous to that of the Page's scheme. So, the value of k used in the special functions (except CUSUMS) can be derived directly from the appropriate SPRT. If the distribution of the observations is similar to

quantile of the Run Length) is equal to a prescribed value; this value should be provided in R [1] . If R [1] represents the quantile, then an additional component, R [2] , is required; this component should provide the order of the quantile. If the right argument, R, contains a third component, then R [1] specifies the desired *off-target* ARL (in this case R [2] must be set to 0) or the desired *off-target* quantile.

The initial point h_0 used in the procedure of search for h is chosen automatically. However,as we mentioned in the previous section, in some cases this choice is too low compared to the sought value of h . In such cases Step 1 fails and a message appears indicating that the upper bound of search has been reached. On the basis of (displayed) last approximation to h , the user should introduce an additional (the first optional) component of L, representing some larger initial point of search, and then repeat the function. It is also recommended to use this optional component if a good approximation to the sought value of h is available - this will save CPU time. The second and last optional component of L represents either the interval of discretization (for functions dealing with counted variables - CUSUMC and CUSUMP) or the level of discretization. The default interval of discretization is 0.1; the default level of discretization is 30.

Once the search procedure for h is completed, Step 2 of a special function (interactive analysis) is initiated by the message:

Enter the values of H, K, HEADST. and CD for further analysis (or 0 to exit):

At this point the user can examine additional properties of the scheme derived at Step 1 (or any other scheme). Only the first component (signal level, H) must be entered. If the third component falls between 0 and H, it will be used as a headstart; otherwise, a steady state analysis will be performed (see Appendix D).[8] The code CD enables the user to control the extent of analysis:

CD=1 will prompt the user to introduce additional values of the controlled parameters to be explored in the analysis;

---

normal, this suggests to choose k in the midway between "good" and "bad" levels of the process, as noted earlier, in Sec. 2. More material on the analogy between SPRT's and Cusum control schemes can be found, for example, in Khan (1984).

[8] In the steady state analysis the on-target distribution is the one corresponding to the 1-st row of the output table. This row usually corresponds to the acceptable level of the controlled parameter.

CD=2 will prompt the user to introduce the values of r for which $Prob.(RL > r)$ will be computed;

CD=3 will prompt the user to introduce a Shewhart limit;

CD=0 cancels all the above conditions and returns to the original extent of analysis.

Other values of CD will prompt the user to introduce part (or all) of the above conditions (ex. CD=12 is equivalent to simultaneous use of CD=1 and CD=2). Some additional possibilities for the choice of CD are available when using the function CUSUMX; see Sec. 4.6.

4.2. Design of a cumulative c – chart for controlling the mean of a Poisson population. The function CUSUMC

The function CUSUMC designs and analyzes a cumulative c-chart. This scheme is used to control the mean λ of a sequence of Poisson random variables (typical example - monitoring numbers of defects found in successively produced units). Thus, the observations in this type of scheme are integers. The format is as follows:

L CUSUMC R

where the right argument R has a general form described in Sec. 4.1. The left argument L contains from two to four components:

L [1] is the acceptable level, λ_0;

L [2] is the unacceptable level, $\lambda_1 > \lambda_0$;

L [3] (optional) is the initial approximation h_0 used in the procedure of search for h. Its default value is 2.95.

L [4] (optional) is the interval of discretization, δ (default 0.1).[9]

The value of k used in the design procedure is

$$k = \frac{\lambda_1 - \lambda_0}{\log \lambda_1 - \log \lambda_0},$$

[9] Note that by specifying the *interval* of discretization we are able to eliminate the roundoff error (see Appendix C). However, by doing that we lose the direct control over the *level* of discretization. So, schemes with signal level, say, above 10 will lead to a level of discretization of order $d \approx 100$ and, consequently, to extensive CPU time requirements. Therefore, if our preliminary run indicates that high values of h may be required, it is recommended to use a longer interval of discretization, say, $\delta = 0.2, 0.5$ or 1.

rounded to the nearest multiple of δ (ex. see Lucas (1985)).

Example 4.2. Let the acceptable and unacceptable levels of λ be $\lambda_0 = 1.5$ and $\lambda_1 = 3.5$. We would like to have a scheme (i.e. to find h ; remember that k is determined automatically) for which the ARL (for $\lambda = 1.5$) is 200. So, we can use

1.5 3.5 CUSUMC 200

If our goal were to find a scheme for which $Prob.(RL > 20 \mid \lambda = 1.5) = 0.9$, and a good initial approximation ($h_0 = 3.7$) were available, we should have used the statement

1.5 3.5 3.7 CUSUMC 20 0.9

If our goal were to find a scheme for which the ARL (for $\lambda = 3.5$) is 2.5, we should have used the statement

1.5 3.5 CUSUMC 2.5 0 1

A case related to use of CUSUMC is considered in Appendix B (Example B.3).

4.3. Design of a cumulative p - chart for controlling the process proportion of defective units. The function CUSUMP

The function CUSUMP designs and analyzes a cumulative p-chart. This scheme is used to control the process proportion of defectives p on the basis of numbers (counts) of defective units found in successive samples of size n. In general, this type of scheme represents an instrument for controlling the parameter p of a binomial population. Clearly, the observations in this type of scheme are integers. The format is as follows:

L CUSUMP R

where the right argument R has a general form described in Sec. 4.1. The left argument L contains from three to five components:

$L\,[1]$ is the sample size, n;

L [2] is the acceptable level, p_0;

L [3] is the unacceptable level, $p_1 > p_0$;

L [4] *(optional)* is the initial approximation h_0 used in the procedure of search for h. Its default value is 2.95.

L [5] *(optional)* is the interval of discretization, δ (default 0.1).[10]

The value of k used in the design procedure is

$$k = \frac{p_1 - p_0}{\log p_1 - \log p_0},$$

rounded to the nearest multiple of δ.

Example 4.3. Let the acceptable and unacceptable levels of p be $p_0 = 0.01$ and $p_1 = 0.04$, and let the sample size be $n = 25$. We would like to have a scheme (i.e. to find h; remember that k is determined automatically) for which the ARL (for $p = 0.01$) is 500. To derive an appropriate scheme, we can use

25 0.01 0.04 CUSUMP 500

If our goal were to find a scheme for which $Prob.(RL > 20 \mid p = 0.01) = 0.95$, and a good initial approximation ($h_0 = 3$) were available, we should have used the statement

25 0.01 0.04 3 CUSUMP 20 0.95

If our goal were to find a scheme for which the ARL (for $p = 0.04$) is 2.5, we should have used the statement

25 0.01 0.04 CUSUMP 2.5 0 1

Note that CUSUMP can also be used to derive a scheme for monitoring p on the basis of sample proportions of defectives (instead of counts). If the sample size n is fixed, one should just divide the derived scheme parameters by n. If the sample size varies, we have no choice but to use the sample

[10] See the footnote on p. 20.

proportions; in this case we could set n to some "expected" sample size, derive an appropriate scheme, and then examine its performance with respect to other, fixed as well as random sample sizes (also see Example 5.7).

4.4. Design of a cumulative s – chart for controlling the standard deviation of a normal population. The function CUSUMS

The function CUSUMS designs and analyzes a cumulative s-chart. This scheme is used to control the standard deviation σ of a normal population on the basis of a sequence of sample standard deviations $\hat{\sigma}_1, \hat{\sigma}_2, \ldots$ (see (2.3)) corresponding to successive samples (subgroups) of size n. Cumulative s-charts represent an alternative for the "classical" s-charts and r-charts (see Duncan(1974, Ch. 21). Note that in most practical situations the subgroup size n does not exceed 5.

The format is as follows:

L CUSUMS R

where the right argument R has a general form described in Sec. 4.1. The left argument L contains from three to five components:

$L[1]$	is the sample size, n;
$L[2]$	is the acceptable level of variability, σ_0;
$L[3]$	is the unacceptable level, $\sigma_1 > \sigma_0$;
$L[4]$ *(optional)*	is the initial approximation h_0 used in the procedure of search for h. Its default value is $3\sigma_0/\sqrt{2n}$.
$L[5]$ *(optional)*	is the level of discretization, d (default 30).

The value of k used in the design procedure is $k = (\sigma_0 + \sigma_1)/2c(n)$, where

$$c(n) = \sqrt{(n-1)/2}\,\Gamma((n-1)/2)/\Gamma(n/2);$$

in particular, for sample size of $n = 2, 3, 4, 5, 6$, and 7 this constant is 1.25, 1.13, 1.09, 1.06, 1.05 and 1.04, respectively.

Example 4.4. Let the acceptable and unacceptable levels of σ be $\sigma_0 = 0.1$ and $\sigma_1 = 0.3$, and let the sample size be $n = 4$. We would like to have a scheme (i.e. to find h; remember that k is determined automatically) for which the ARL (for $\sigma = 0.1$) is 200. To derive an appropriate scheme, we can use

4 0.1 0.3 CUSUMS 200

If our goal were to find a scheme for which $Prob.(RL > 20 \mid \sigma = 0.1) = 0.9$, and a good initial approximation ($h_0 = 0.03$) were available, we should have used the statement

4 0.1 0.3 0.03 CUSUMS 20 0.9

If our goal were to find a scheme for which the ARL (for $\sigma = 0.4$) is 1.5, we should have used the statement

4 0.1 0.3 CUSUMS 1.5 0 1

The form of the reference value given above is related to the fact that sample standard deviation $\hat{\sigma}$ represent a biased (downwards) estimator for σ; to obtain an unbiased estimator, one would need to multiply it by $c(n)$. Therefore, when using the V-mask version of the cumulative s-chart, the current level of the sequence as shown by the protractor should also be multiplied by $c(n)$ to obtain an unbiased assessment of σ. If it is important, for reasons of graphical data presentation, to base the control scheme on the sequence of unbiased estimators, one could simply multiply the sample standard deviations as well as the scheme parameters by $c(n)$. Another way would be, of course, to base the control procedure on the sequence of sample variances. Though this sequence is less appealing from the point of view of graphical presentation and, moreover, its members are relatively highly skewed, it may be preferred in situations of extremely high sampling intensity (ex. control of robots) where speed of computing values of the scheme may become an important factor.[11] Since its use does not typically lead to improvement in scheme performance, this possibility will not be considered in the present work.

---

[11] The value of k recommended for control scheme based on sample variances is $\frac{2 \times (\log \sigma_1 - \log \sigma_0)}{(1/\sigma_0^2) - (1/\sigma_1^2)}$.

One can see that CUSUMS can be used to control internal variability in slightly more general situations. For example, assume that in situation described in Example 2.1 the measurements related to a given lot come from three wafers (taken from the middle and both ends of the lot), each corresponding to a set of five measurements. Then, since one can expect that various parts of the lot are subject to slightly different conditions, it would be natural to estimate the within-wafer standard deviation by taking an average of three standard deviations corresponding to different wafers. Since the number of degrees of freedom for estimating σ is $3 \times (5-1) = 12$, our "pooled" sample standard deviation is distributed in the same way as a sample standard deviation corresponding to a sample of size $12+1 = 13$ taken from a homogenuous population. Thus, a control scheme based on the pooled standard deviation can be designed by using CUSUMS with $n = 13$. For a more general formulation see Yashchin (1984, p.33).

Additional examples illustrating the use of CUSUMS are considered in Examples B.1 and B.7 of Appendix B.

4.5. Design of a cumulative t - chart for controlling the average time between events of a Poisson process. Controlling the process proportion of defectives on the basis of "gaps" between successive defective units. The function CUSUMT

The function CUSUMT designs and analyzes a cumulative Time-Between-Event chart, or t-chart. This type of chart is used to control the mean inter-arrival time θ of a Poisson process (i.e. to control the mean θ of a sequence of iid exponential random variables). They also represent a way of controlling the rate $\lambda = 1/\theta$ of a Poisson process (an alternative way would be to count the number of arrivals in subsequent time intervals of some fixed length and then to use a c-chart). In typical practical situations we are interested in the *lower* control scheme only. Indeed, if our observations represent, say, times between successive breakdowns in some system, or lengths of life corresponding to a sequence of tested devices, or inter-arrival times of customers in a queueing system, we are interested to detect, as soon as possible, situations in which these observations fall systematically below the expected level.

Another interesting application of this type of schemes is based on the relationship between the Bernoulli (0-1) process and the Poisson process. Indeed, let us consider a production process for which every produced item is defective with probability p. If p is small, then the number of units produced until the next defective one is found has a geometric distribution (which, as we know, represent a discrete analogue of an exponential distribution). If the intensity of sampling was constant, then times between consecutive defective units would form a process which can be considered a Poisson process for most practical purposes. Thus, one could control the process proportion of defectives by means of a t-chart by treating the observations (number of units produced between consecutive defectives) as approximately exponential random variables. This approximation works very well for small values of p; interested reader could verify that by applying an appropriate function for analysis of general schemes to a geometric distribution. An alternative way of controlling p would be to form samples of some fixed size, n, and then to use a cumulative p-chart considered earlier. If the units are produced and/or inspected one at a time, this way will cause loss in the resolution power of the scheme - this is one of situations mentioned in the introduction where one should try to use the information as soon as it arrives rather than "create" samples purely for purposes of control.

It is also worth mentioning that, because of the connection between exponential and Weibull distributions, the function CUSUMT can also be used to design a scheme for controlling the mean of a Weibull population (typical application - monitoring the life times of successive devices subjected to an accelerated life testing procedure); see Example B.5 from Appendix B.

The format of CUSUMT is as follows:

L CUSUMT R

where the right argument R has a general form described in Sec. 4.1. The left argument L contains from two to four components:

$L\,[1]$	is the acceptable level of the mean time between events, θ_0;
$L\,[2]$	is the unacceptable level, $\theta_1 < \theta_0$;
$L\,[3]$ *(optional)*	is the initial approximation h_0 used in the procedure of search for h. Its default value is $2\theta_0$;

L [4] (optional) is the level of discretization, d (default 30).

The value of k used in the design procedure is

$$k = -\frac{\theta_0\theta_1(\log\theta_0 - \log\theta_1)}{\theta_0 - \theta_1}.$$

Note that the reference value is negative since t-chart corresponds to a *lower* control scheme. One can see that $(-k)$ is a reciprocal of the reference value used in cumulative c-charts (ex. see Lucas (1985)).

Example 4.5. Let the acceptable and unacceptable levels of θ be $\theta_0 = 1000$ and $\theta_1 = 500$. In the context of controlling the rate of defectives, this means the following. If the average rate of defectives is one per 1000 produced units (i.e. $p = 0.001$), we consider it quite satisfactory and, under these conditions, we would like to avoid false alarms; if, however, the average rate becomes one per 500 units, we would like an out-of-control signal to be triggered as soon as possible.

We are interested in a scheme (i.e. in finding an appropriate value of the signal level h; remember that k is determined automatically) for which the ARL (for $\theta = 1000$) is 200. In other words, if $\theta = 1000$, we would like the Average Time to Signal (ATS) to be $1000 \times 200 = 200000$. To derive an appropriate scheme, we can use

```
1000 500 CUSUMT 200
```

If our goal were to find a scheme for which $Prob.(RL > 20 \mid \theta = 1000) = 0.9$, and a good initial approximation ($h_0 = 2900$) were available, we should have used the statement

```
1000 500 2900 CUSUMT 20 0.9
```

At this point, an important remark is in place. As we saw earlier, the Average Time between Signals can be explicitly derived from the ARL. This is a direct consequence of Wald's identity (ex. see Feller (1971)). Unfortunately, this identity cannot be extended to other characteristics of the Run Length (ex. quantiles). So, $Prob.(RL > 20 \mid \theta = 1000) = 0.9$, does not, in general, imply that

Prob.$(TS > 20 \times 1000 \mid \theta = 1000) = 0.9$. Such relationship is not more than a (rather useful) approximation.

Finally, if our goal were to find a scheme for which the ARL (for $\theta = 500$) is 2 (in other words, if $\theta = 500$, we would like the ATS to be $500 \times 2 = 1000$), we should have used the statement

```
1000 500 CUSUMT 2 0 1
```

4.6. Design of a cumulative $\bar{x}$ – chart for controlling the mean of a normal population. The function CUSUMX

The function CUSUMX designs and analyzes a cumulative $\bar{x}$ -chart. This scheme is used to control the mean μ of a normal population on the basis of sample averages corresponding to successive samples (subgroups of measurements) of size n. Cumulative $\bar{x}$ -charts represent an alternative for the "classical" (Shewhart's) $\bar{x}$ - charts (ex. see Duncan(1974, Ch. 21). In most practical situations the subgroup size n used in this type of scheme does not exceed 5.

The format is as follows:

```
L CUSUMX R
```

where the right argument R has a general form described in Sec. 4.1. The left argument L contains from four to six components:

$L[1]$	is the sample size, n;
$L[2]$	is the standard deviation of a *single* measurement, σ_0;
$L[3]$	is the acceptable level of the mean, $\mu_0 \geq 0$;
$L[4]$	is the unacceptable level, $\mu_1 > \mu_0$;
$L[5]$ *(optional)*	is the initial approximation h_0 used in the procedure of search for h. Its default value is $3\sigma_0/\sqrt{n}$, i.e. three standard deviations of an observation (representing an average of n measurements).
$L[6]$ *(optional)*	is the level of discretization, d (default 30).

The value of k used in the design procedure is $k = (\mu_0 + \mu_1)/2$.

Note that for this function, additional possibilities of analysis are available in its second step. In particular, specifying a four-digit CD will lead to analysis of a symmetric *two-sided* scheme with the given parameters.[12] The last three digits control the extent of analysis as described in Sec. 3.1. For example, CD = 1012 will prompt the user to introduce additional values of μ to be considered as well as values of r for which $Prob.(RL > r)$ are to be computed; CD = 1000 will cancel all special conditions and continue analysis in a *two-sided* mode, etc. To return to the *one-sided* mode of analysis, one should use CD = 0.

Example 4.6. Let the acceptable and unacceptable levels of μ be $\mu_0 = 2.5$ and $\mu_1 = 4.5$, and let the sample (subgroup) size be $n = 4$ and the standard deviation of a single measurement be $\sigma = 1.2$. We would like to design a one-sided scheme (i.e. to find h ; remember that k is determined automatically) for which the ARL (for $\mu = 2.5$) is 200. To derive an appropriate scheme, we can use

4 1.2 2.5 4.5 CUSUMX 200

If our goal were to find a scheme for which $Prob.(RL > 20 \mid \mu = 2.5) = 0.9$, and a good initial approximation ($h_0 = 0.5$) were available, we should have used the statement

4 1.2 2.5 4.5 0.5 CUSUMX 20 0.9

If our goal were to find a scheme for which the ARL (for $\mu = 4.5$) is 1.5, we should have used the statement

4 1.2 2.5 4.5 CUSUMX 1.5 0 1

Additional examples related to the use of CUSUMX can be found in Examples B.2, B.6 and B.7 of Appendix B.

---

[12] This option should be used only when the target level corresponding to the two-sided scheme is 0.

5. Functions for design and analysis of general Cusum – Shewhart schemes

In the previous section our discussion was centered around functions for design and analysis of schemes appropriate to a rather limited class of situations. All of our special functions deal with one-sided schemes the only exception being function CUSUMX; but the only two-sided schemes the latter can handle are symmetric schemes with normally distributed observations. Of course, we could not write special functions for every situation that might become relevant - so, we created functions to handle situations in which a general Cusum-Shewhart control scheme is applied to a general pattern of incoming (iid) observations. These functions can be used not only directly, but also as a toolkit which enables the user to create his own functions for design and analysis of schemes corresponding to specific situations.[13]

In the present section we introduce the class of such functions. We refer to them as general functions as opposed to the special functions considered in the previous section. As we shall see, all of them have a similar format and many other common features. So, we proceed by providing some general information about this group of functions.

5.1. Some general information. Unified format of the general functions. Interactive, EXPLR- and VARY- modes of analysis

To analyze the RL of a general Page's scheme one needs to specify the scheme parameters as well as the nature of incoming observations. So, *every* function for analysis and/or design of a general one-sided Page's scheme has the following form:

Y function DFNAME

where *function* represents the type of design and/or analysis function. The arguments are as follows:

---

[13] In fact, the special functions discussed earlier represent an example of using the general functions as a toolkit. Of course, to use the general functions in this way one should be familiar with APL.

DFNAME represents the *name* of the APL function which returns the value of the d.f. of the observations $F(x)$ for any given value of x. Typically, one will use one of the functions provided with the package (see Appendix A).[14]

Y is a vector specifying the scheme parameters; it consists of two or three components: Y [1] is the signal level, h. Y [2] is the reference value, k. Y [3] is the Shewhart's limit, c (optional).

Each function for design and/or analysis of one-sided schemes comes in two versions. The first one is used for purpose of handling upper Page's schemes. The second plays a similar role when dealing with lower schemes. The name of the latter function differs from that of the previous one by presence of an additional letter, L. For example, ONEAN is used for analysis of upper schemes; its counterpart, ONEANL, performs a similar analysis of lower schemes.

Analogously, every function for analysis of a general two-sided scheme has the following form:

T function DFNAME

where *function* and DFNAME have the same meaning as in the one-sided situations; the vector of scheme parameters T can contain four or six components: T [1] and T [3] are the signal levels, h^+ and h^-, respectively. T [2] and T [4] are the reference values, k^+ and k^-, respectively. T [5] and T [6] are the Shewhart's limits, $(-c^-)$ and c^+, respectively. The last two components are optional; the choice of signs excludes any possibility of confusion. Indeed, if these components are present, they must satisfy the inequality T [5] < T [6] ; an immediate out-of-control signal is triggered if a single observation falls outside the interval (T [5] , T [6]).

Clearly, this setup is sufficient to perform an interactive analysis of a given Cusum-Shewhart scheme with respect to a given pattern of incoming observations (functions ONEAN(L) and TWOAN). In addition to these possibilities, two other modes of analysis are available:

---

[14] If DFNAME is provided by the user, its heading *must* be of type R ← DFNAME X.

EXPLR - mode (ONEXPLR(L), TWOXPLR) enables one to examine the performance of a fixed scheme with respect to a family of distribution functions determined by varying the values of a specified parameter (ex. mean of the normal population. This mode of analysis requires the user to specify the name of the varying parameter of the distribution as well as its values by means of the function SETXPLR.

VARY - mode (ONEVARY(L), TWOVARY) enables one to examine the performance of a family of control schemes determined by varying the values of a specified scheme parameter (ex. the reference value) with respect to a fixed distribution function of incoming observations. This mode of analysis requires the user to specify the varying parameter of the scheme as well as its values by means of the function SETVARY.

The functions for design of general one-sided schemes perform a search for a signal level h for which the ARL (or some specified quantile of the Run Length) is equal to a prescribed number. This number as well as some other conditions which determine the type of the search procedure are specified by the function SETFIND. Other parameters of the scheme (given by Y) are fixed; the first component of Y is used as an initial approximation h_0 in the search procedure. The function ONEFIND(L) performs the search for h only; the function ONEXPLRD(L) first performs the search (Step 1), and then performs an EXPLR - mode analysis of the resulting scheme (Step 2).

5.2. Specifying the optional conditions for analysis (headstarts, level of discretization, etc.). The functions SET, SETI and RESET

Before using any of the functions, the user can specify several special conditions. These conditions correspond to values of selected global variables of the workspace; so, a user familiar with APL could change these default values directly, or localize them if he wants to use the general functions as a toolkit. Another way to change the mentioned conditions is by using the functions SET or SETI. The format of SET is as follows:

CODE SET VAL

where CODE determines which condition is to be specified and VAL is a value assigned to an appropriate global variable. The function SETI has the same format but, before changing the condition, it informs the user of its intentions and provides the possibility of aborting the assignment at the last moment. The meaning of CODE is:

CODE=0 Specifies the headstart(s). If VAL has a single component, it will be used as a headstart for analysis of any one-sided scheme[15] Otherwise, VAL represents a pair of headstarts that will be used for analysis of any two-sided scheme;[16]

CODE=1 specifies the level of discretization, d which should be provided in VAL. This level will be used for design and analysis of one-sided as well as two-sided schemes;[17]

CODE=2 in this case VAL should contain the values of r for which $Prob.(RL > r)$ is to be computed;[18]

CODE=3 is used for the purpose of steady state analysis only. If VAL=1, the next analyzed scheme will be considered as an on-target one and the corresponding steady state

---

[15] The corresponding global variable is HEADSTART its default value is 0. Note that negative value of headstart will lead to a steady state analysis (or an error message, where the latter is inappropriate). This type of analysis is available for one-sided schemes only.

[16] The corresponding global vector is HEADSTWO (its default value is (0, 0)).

[17] The corresponding global variables are DISCRONE and DISCRTWO (the default value for both of them is 0). Note that DISCRTWO will be used as a level of discretization of the scheme with a higher signal level. The level of discretization of the opposite scheme is chosen in such a way that the lengths of discretization intervals of both schemes be as close as possible.

[18] The corresponding global vector is R (in the default mode it is empty).

distribution will be stored. The statement with VAL=0 will cancel this condition.[19]

Some additional possibilities are available for the purpose of running the Cusum-Shewhart schemes. These will be discussed in Sec. 6.1. To restore the default modes of *all* optional conditions simultaneously, one can type RESET.

Next we consider the "general" functions separately.

5.3. Interactive analysis of one-sided schemes. The function ONEAN

This function is used to analyze (in interactive mode) the ARL, SDRL and the run length distribution of a one-sided scheme. Use of ONEAN results in -

1. printout of the basic information about the scheme;
2. printout of the complete set of (discretized) headstarts as well as set of corresponding ARL's and SDRL's;
3. prompting the user to specify values of r for which $Prob.(\mathrm{RL} > r)$ are to be computed;
4. prompting the user to specify the headstart (or bounds of the segment containing the headstarts) for which the above probabilities are to be computed;
5. printout of the table of the computed probabilities; each row of the table corresponds to a single headstart;
6. prompting the user to continue the analysis of the run length distribution.

The function ONEANL performs a similar analysis of a lower scheme.

Example 5.3. Suppose that we would like to analyze the run length of a scheme with parameters $h = 3$, $k = 1$, the observations $\{ x_i \}$ being distributed normally with mean $\mu = 0$ and s.d. $\sigma = 1$. The function returning the values of a normal d.f. exists in our workspace under the name DFNORM. Before using this function, we must set its global variables MEU and SIGMA to be equal to μ and

---

[19] The corresponding global variable is ONTARGET (default 0). The steady state distribution is stored in the global vector STEADY (see Yashchin (1984) for more details).

σ, respectively. This can be done either by using two separate APL assignment statements or by using our function ASSIGN as shown in the example below. Thus, the interactive analysis is initiated by executing the statements

```
'MEU SIGMA' ASSIGN 0 1
(3 1) ONEAN 'DFNORM'
```

5.4. Analyzing a set of one-sided schemes with respect to a fixed distribution of incoming observations. The function ONEVARY

The function ONEVARY performs, in non-iteractive mode, the analysis of a sequence of upper schemes (depending on a single varying parameter) corresponding to a given fixed d.f. of the observations. The function ONEVARYL performs a similar analysis of a sequence of lower schemes.

Execution of ONEVARY results in analysis of the scheme with parameters given by Y and, in addition, of a sequence of schemes corresponding to values of a varying parameter of the scheme. Before using ONEVARY one should specify this parameter and provide its values. This is done by executing the function CODE SETVARY VAL. The left argument, CODE, should be 1, 2 or 3 if the varying parameter of the scheme is h, k or c, respectively. The right argument, VAL, should provide a list of values of the varying parameter. If CODE is 0, the varying parameter is the headstart; in this case VAL should provide the bounds of a segment containing headstarts for which the analysis is to be performed.

The optional conditions (level of discretization, headstart, etc.) are set as described in Sec. 5.1.

Example 5.4. Suppose that we would like to analyze the run length of a scheme $h = 6$, $k = 1$ supplemented by Shewhart's limit $c = 5.5$,and, in addition, of schemes with $h = 6.1, 6.2, 6.5$ with respect to a sequence of iid Weibull observations with shape 1 and scale 2 (i.e. exponential observations with mean 2). We would like the output table to include the probabilities $Prob.(RL > r)$ for $r = 5$, 10 and 20 (optional condition). So, we use the following statements:

```
2 SET 5 10 20
1 SETVARY 6.1 6.2 6.5
'SHAPE SCALE' ASSIGN 1 2
(6 1 5.5) ONEVARY 'DFWEIB2'
```

Note that DFWEIB2 is our APL function computing the values of the distribution function of a two-parametric Weibull random variables; SHAPE and SCALE are its global parameters.[20]

5.5. Analyzing a fixed one-sided scheme with respect to a family of distributions of incoming observations. The function ONEXPLR

The function ONEXPLR performs the analysis of a one-sided Page's scheme with all three (or four, if the scheme is supplemented by Shewhart's limit) parameters fixed for a set of several d.f.'s of the observations corresponding to different values of a specified parameter. This parameter usually corresponds to one of the global variables of the function DFNAME. The function ONEXPLRL performs a similar analysis of a lower scheme.

Before using ONEXPLR, one should specify the name of the changing parameter of the distribution function as well as its values. This is done by executing the statement NAME SETXPLR VAL where NAME is a (character) vector containing the name of the changing parameter of the distribution and VAL is a vector containing its values. In addition one can specify some optional conditions (level of discretization, headstart, etc.) as described in Sec. 5.1.[21]

Example 5.5. Suppose that we would like to analyze the performance of a scheme $h = 29.5$, $k = 9$ and $c = 18.5$ with respect to sequences of (iid) Poisson random variables with means 6.5, 8.5 and 11.5. Let the headstart of the scheme be $s_0 = 10$ (optional condition). To perform the analysis, we execute the following statements:

---

[20] Also note, that one could use the APL statement R ← 5 10 20 instead of our first statement.

[21] If, before using ONEXPLR, one executes the statements 3 SET 1 and 0 SET (-1) (see Sec. 5.2), the steady state analysis will be envoked. In this analysis, the distribution corresponding to the first component of VAL will be treated as on-target distribution of observations.

```
0 SET 10
'LAMBDA' SETXPLR 6.5 8.5 11.5
(29.5 9 18.5) ONEXPLR 'DFPOIS'
```

5.6. Design of a one-sided scheme. The function ONEFIND

The function ONEFIND performs an automated search for the signal level h (upper scheme), satisfying one of the two conditions:

$$\begin{aligned} &a)\ \mathrm{ARL} = q \quad \text{or} \\ &b)\ \mathrm{Prob.}(\mathrm{RL} > q) = \gamma, \end{aligned} \tag{5.1}$$

where q and γ are specified by the user. All the other parameters of the scheme (provided in the left argument, vector Y) remain fixed. The initial point of search, h_0, should be provided in the first component of Y. The function ONEFINDL performs a similar search for a signal level of a lower scheme.

Before using ONEFIND one should specify the conditions needed to perform a search for an appropriate value of the signal level, h. This is done by executing the statement MODE SETFIND FIX, where MODE determines the type of the search procedure. MODE should be 0 if h_0 provided by Y [1] may be a very rough estimate of the sought value of h; it is useful to set MODE to 1 if h_0 is close to the sought value and needs some refining only. The first component of the right argument, FIX, should contain the value of q; if q represents the quantile (i.e. the user wants the condition of type b) to be satisfied), then FIX must have a second component, providing the order of the quantile, γ. If q represents the ARL, no second component is required.

If the search for h was successful, the lower and upper approximations for h will be printed out together with the corresponding values of ARL (if the user wanted the condition a) to be satisfied) or $Prob.(RL > q)$ if he wanted the condition b) to be satisfied; the corresponding levels of discretization will also be printed out. A typical output of ONEFIND is as shown in the second table of Sec. 3.

In the case that search for h fails, only information related to the last approximation examined by the search procedure as well as some diagnostics will be provided. This information can be used to suggest an alternative initial point of search and repeat the function.[22]

Note that ONEFIND will try to preserve the current level of discretization. Therefore, the precision of the search can be controlled by means of the function SET. The default level of discretization, 30, usually leads to quite satisfactory results. When MODE = 1, the function will *always* preserve the length of interval of discretization, computed on the basis of current level of discretization and h_0 (see (C.1), Appendix C). This mode is especially useful for dealing with schemes based on counts, where proper choice of the interval may eliminate roundoff errors usually caused by discretization.

Example 5.6. Let the observations correspond to the sequence of normal means corresponding to subgroups of size $n = 3$. Let the mean and standard deviation of a single measurement be $\mu = 110$ and $\sigma = 10$. Let the reference value be $k = 105$. Under these conditions, if one is interested in finding a value of h for which the ARL is 3.9, he could achieve it by executing the following statements:

```
'MEU SIGMA SAMPLE' ASSIGN 110 10 3
0 SETFIND 3.9
(12 105) ONEFIND 'DFXBAR'
```

(which leads to $h = 15.7$; ex. see Duncan (1974, p.476)). Our initial approximation for the search procedure is $h_0 = 12$. Note that DFXBAR is the function computing the d.f. of sample mean; MEU, SIGMA and SAMPLE are its global parameters.

If, under the stated conditions, we wanted to find h satisfying $Prob.(RL > 3) = 0.95$, the sequence would be the same except the second statement which becomes 0 SETFIND 3 0.95.

---

[22] In some cases failure of ONEFIND is related to non-existense of the value of h having the desired property. This is indicated by an appropriate message.

5.7. Design and analysis of a one-sided scheme. The function ONEXPLRD

The function ONEXPLRD can be viewed as a combination of the functions ONEFIND (Step1) and ONEXPLR (Step2). In its second stage, the function ONEXPLRD performs (as ONEXPLR does) the analysis of a scheme with all three (or four, if the scheme is supplemented by a Shewhart limit) parameters fixed for a set of several distribution functions of the observations corresponding to different values of a specified parameter (which usually corresponds to one of the global variables of the function DFNAME). However, before performing the analysis, the function determines (Step1) the value of the signal level h satisfying one of the two conditions (5.1), where q and γ are specified by the user, exactly as in ONEFIND. Other parameters of the scheme (provided in the left argument, Y) remain fixed. As in ONEFIND, the initial point of search, h_0, should be provided in the first component of Y. The value of h found in the first stage is subsequently used (in Step2) as a signal level of the control scheme.

The function ONEXPLRDL performs a similar analysis for a lower scheme.

Before using ONEXPLRD one should specify the conditions needed to perform a search for h. This is done by executing the statement MODE SETFIND FIX, where MODE and FIX have the same meaning as in ONEFIND. The only exception is that now MODE may have a second (optional) component, MODE [2] , playing the following role. The procedure of search for h (Step 1) results in an interval (see second output table, Sec. 3). The value of MODE [2] determines which of its bounds will be selected as an approximation to h. If MODE [2] = -1, the lower bound will be selected. If MODE [2] = 1, the upper one will be selected. If MODE [2] = 0, (or if this component is absent), the bound for which the value of q is closer to the desired value will be selected. If MODE [2] = 2, the interpolated value will be selected.

One should also specify the name of the changing parameter of the distribution function as well as its values. As in ONEXPLR, this is done by executing the statement NAME SETXPLR VAL where NAME is a (character) vector containing the name of the changing parameter and VAL is a vector containing its values.

In addition, one can specify some optional conditions as described in Sec. 5.1.

Example 5.7. Consider the situation described in Duncan (1974, p.478). We would like to design a cumulative p-chart satisfying the following conditions: ARL($p = 0.04$) = 7.5 and ARL($p = 0.01$) = 500. Let us choose the reference value to be $k = 0.025$ (midway between the good and bad quality). What remains to be determined are the values of the signal level h and the sample size n. First, we specify that we would like the value of h to satisfy ARL(p=0.04) = 7.5. Since for any given sample size this value will be determined automatically (in the first stage of ONEXPLRD), we only need to find, by trial and error, the appropriate sample size, n. We start by trying $n = 10$...

```
'PROB SAMPLE' ASSIGN 0.04 10
0 SETFIND 7.5
'PROB' SETXPLR 0.01 0.02 0.03 0.04
(0.1 0.025) ONEXPLRD 'DFPROPB'
```

This sequence will use 0.1 as an initial approximation for h, determine h so that the ARL=7.5 and apply the resulting scheme to sequences of observations corresponding to $p = 0.01, \ldots, 0.04$. The function DFPROPB computes the d.f. of the sample proportion; SAMPLE and PROB are its global parameters. The reader may verify, that in order to achieve the desired resolution between "good" and "bad" quality, we need to use samples of size $n = 24$.

One can argue that the special function CUSUMP is more appropriate for dealing with situations of this type; in particular, we could initiate the analysis corresponding to $n = 10$ by using the statement:

```
10 0.01 0.04 CUSUMP 7.5 0 1
```

This approach has an advantage that it automatically selects the level of discretization and k so as to eliminate roundoff errors. On the other hand, the latter scheme is based on the sequence of *counts* of defectives, and therefore cannot be recommended in cases where sample size varies from sample to sample; in such cases one should use the sequence of *sample proportions* of defectives. After deriving a scheme appropriate for a certain "most likely" sample size n, one can use the function ONEXPLR to examine its properties with respect to any other (fixed) sample size. It is worth mentioning that the case of random sample size can also be considered. In particular, the function

DFPROP corresponds to distribution of sample proportion under the assumption that the sample size is a Poisson distributed random variable with mean SAMPLE.

5.8. Interactive analysis of two-sided schemes. The function TWOAN

The function TWOAN is used for interactive analysis of the ARL, SDRL and the run length distribution of two-sided Cusum-Shewhart schemes. In addition, it enables one to examine (interactively) the probability that the signal will be triggered by the upper scheme, $P(UP)$.

Use of TWOAN results in -

1. printout of the basic information about the two-sided scheme;
2. printout of the complete set of (discretized) headstarts as well as set of corresponding ARL's and SDRL's for the upper and lower schemes separately;
3. prompting the user to specify the headstarts for which the ARL, SDRL and P(UP) are to be computed. The results for each pair of headstarts are returned immediately.
4. prompting the user to specify the headstarts and the values of r for which $Prob.(\mathrm{RL} > r)$ are to be computed. The results for each pair of headstarts are returned immediately.
5. prompting the user to continue the analysis of the run length distribution.

The user has a possibility to switch from step (4) to (3) and vice versa (or to exit from the function) by specifying a pair of headstarts which falls out of range.

As an example, suppose that we would like to analyze the run length of a scheme with parameters $h^+ = 4,\ k^+ = 1,\ h^- = 4.5,\ k^- = 1.5,$ the observations $\{ x_i \}$ being distributed normally with mean $\mu = 0$ and standard deviation $\sigma = 1$. Let the scheme be supplemented by the Shewhart's limits $c^- = 3,\ c^+ = 3.5,$ i.e. $x_i \leq -3$ or $x_i \geq 3.5$ should trigger an immediate out-of-control signal. Then the statements needed to initiate the interactive analysis are as follows:

```
'MEU SIGMA' ASSIGN 0 1
(4, 1, 4.5, 1.5, (-3), 3.5) TWOAN 'DFNORM'
```

5.9. Analyzing a set of two-sided schemes with respect to a fixed distribution of incoming observations. The function TWOVARY

The function TWOVARY performs the analysis of a sequence of schemes (depending on a single varying parameter or a pair of varying scheme parameters) corresponding to a given fixed d.f. of the observations.

Before using TWOVARY, the user should specify the varying parameter of the scheme and provide its values. As in the function ONEVARY, this is done by executing the function CODE SETVARY VAL. The left argument, CODE, should contain 1, 2, 3, 4, 5 or 6 if the varying parameter of the scheme is h^+, k^+, h^-, k^-, c^- or c^+, respectively.[23] The right argument, VAL, should provide a list of values of the varying parameter. If CODE=0, the varying parameters are headstarts; in this case VAL should provide the pairs of headstarts for which the analysis is to be performed. If CODE is 11, 22 or 33, the varying parameters are pairs (h^+, h^-), (k^+, k^-), or ($-c^-$, c^+), respectively. In this case VAL should provide the relevant pairs of the scheme parameters.

In addition, one can specify some optional conditions as described in Sec. 5.1.

Example 5.9. To analyze the performance of a scheme $h^+ = 3$, $k^+ = 1$, $h^- = 2.7$, $k^- = 0.9$, supplemented by the Shewhart limits $c^- = 3.5$, $c^+ = 3.5$ and, in addition, of schemes with $k^+ = 1.1,\ 1.2,\ 1.5$ with respect to a sequence of iid normal observations with mean 0 and standard deviation 2, one can execute the following statements:

```
'MEU SIGMA' ASSIGN 0 2
2 SETVARY 1.1 1.2 1.5
(3, 1, 2.7, 0.9, (-3.5), 3.5) TWOVARY 'DFNORM'
```

[23] Note that CODE corresponds to sequential order of the component of T being varied.

5.10. Analyzing a fixed two-sided scheme with respect to a family of distributions of incoming observations. The function TWOXPLR

The function TWOXPLR performs, in a non-interactive mode, the analysis of a two-sided scheme with all the parameters fixed for a set of several d.f.'s of the observations corresponding to different values of a specified parameter. It plays the same role as ONEXPLR does in the analysis of one-sided schemes.

Before using TWOXPLR one should specify the name of the changing parameter of the distribution function as well as its values by executing the statement NAME SETXPLR VAL, where, as usual, NAME is a (character) vector containing the name of the changing parameter and VAL is a vector containing its values;

In addition, one can specify some optional conditions as described in Sec. 5.1.

Example 5.10. Suppose that we would like to analyze the performance of a scheme $h^+ = 3$, $k^+ = 1$, $h^- = 3$, $k^- = 1$, supplemented by the Shewhart limits $c^- = 3.5$, $c^+ = 3.5$, with respect to sequences of sample means corresponding to samples of size 4 from the normal population with s.d. $\sigma=2$ and means 0, 0.1, 0.5, 1. In addition to the usual analysis, we would also like to compute $Prob.(RL > r)$ for $r = 5$, 10 and 100 (optional condition). To perform the analysis, we can use the following statements:

```
2 SET 5 10 100
'SIGMA SAMPLE' ASSIGN 2 4
'MEU' SETXPLR 0 0.1 0.5 1
(3, 1, 3, 1, (-3.5), 3.5) TWOXPLR 'DFXBAR'
```

6.Running the Cusum-Shewhart schemes

So far, our discussion was primarily related to design and analysis of Cusum-Shewhart control schemes by using appropriate functions of CONTRD. However, this package could hardly be considered as complete unless it also provided functions for applying Cusum-Shewhart schemes to se-

quences of observations. Such functions could be used not only for real time monitoring of data, but also for purposes of retrospective data analysis, diagnostics, forecasting and graphical data analysis. In addition, they would enable one to study (by simulation) the performance of control schemes with respect to other than iid patterns of incoming data.[24] Thus, we complete this work by introducing the functions ONERUN and TWORUN for running the control schemes. The reader will see that the syntax of these functions is similar to that of functions for analysis we have introduced earlier - the only difference is that the right argument specifies the data rather than the distribution function of the observations.

6.1. The function ONERUN for running one-sided schemes

The function ONERUN is used to apply a one-sided Cusum-Shewhart scheme to a particular set of data. Its format is as follows:

$$S \leftarrow \text{Y ONERUN DATA}$$

where DATA is the name of the vector containing the observations $\{x_i\}$ and Y is the vector containing parameters of the scheme (defined as in Sec. 5.1). The output vector S contains the computed values of the (upper) Page's scheme.[25] The function ONERUNL plays a similar role in running lower Page's schemes.

Use of ONERUN also results in a global matrix OUTCONTR containing information about the detected out-of-control observations.[26] Each row of this matrix corresponds to a single out-of-control observation and its seven elements contain the following information:

1. the sequential order of the out-of-control observation;

---

[24] Clearly, the iid patterns can also be studied by simulation. However, it can be performed much more efficiently by applying a technique based on the use of empirical distributions (see Example B.4, Appendix B).

[25] Instead of the APL assignment operator, ← , one can use the familiar function ASSIGN as follows: 'S' ASSIGN (Y ONERUN DATA)

[26] Here and in what follows this term is used instead of a more precise "observations corresponding to the out-of-control state of the process"

2. the out-of-control code which is 1 if the signal level h has been exceeded, 2 if the Shewhart's limit c has been exceeded and 3 if *both* these criteria have been violated;

3. the corresponding value of the Page's scheme;

4. the corresponding value of the last observation;

5. the number of observations in the last *positive* portion of the Page's scheme;

6. the sample average of these observations;

7. the sample range of these observations.

By using the function LENGTHS one can create a vector containing the successive run lengths. For example, the statement STATIST LENGTHS will perform a statistical analysis of the run lengths corresponding to the first column of OUTCONTR.

Before using ONERUN, one could specify optional conditions by executing the familiar function CODE SET VAL. As usual, CODE=0 is used to specify the headstart of the scheme (see Sec. 5.2). Other possible values of CODE are as follows:

CODE=5 if VAL=1, the scheme will run in a "quiet" mode. If VAL=0, then information about the scheme as well as out-of-control observations (matrix OUTCONTR) will be displayed; in addition, the user will have a possibility to display (interactively) any portion of the data together with associated values of the Page's scheme as well as some basic statistics related to this portion (sample mean, standard deviation, range, etc.).[27]

CODE=6 the values of upper and/or lower schemes will be re-set to zero after each observation the sequential number of which is a component of VAL. If VAL < 0, the

---

[27] The corresponding global variable is INTERACT (its default value is 0).

scheme(s) will be re-set to 0 (if VAL=-1) or to headstart (if VAL=-2) after each detected out-of-control observation. If VAL=0, no re-setting will take place.[28]

We remind that typing RESET will restore the default mode of all optional conditions.

Example 6.1. The sequence of statements

```
0 SET 1.5
6 SET -2
S ← (6.5, (-1), (-4.5)) ONERUNL (1000 SIMWEIB 2 4)
```

will result in values of a lower Cusum-Shewhart scheme with parameters h=6.5, k=-1 (supplemented by the Shewhart limit c=-4.5, i.e. a single observation below 4.5 triggers an immediate signal), corresponding to a simulated sequence of 1000 Weibull observations with Shape 2 and Scale 4. The scheme is automatically re-set to its headstart (1.5) after each out-of-control observation.

6.2. The function TWORUN for running two-sided schemes

This function is used to apply a two-sided Cusum-Shewhart scheme to a given set of data. Its format is as follows:

```
S ← T TWORUN DATA
```

where DATA is the name of the vector containing the observations $\{x_i\}$ and T is the vector containing parameters of the scheme (defined as in Sec. 5.1). The output matrix S contains the computed values of the upper (first row) and lower (second row) Page's schemes.

Use of TWORUN also results in a global matrix OUTCONTR that has the same format as one created by ONERUN except that the out-of-control code can also have values -1, -2 or -3. Negative values of the out-of-control code have the same meaning as their positive counterparts (see 6.1) but are related to signals triggered by the *lower* scheme.

---

[28] The corresponding global vector is RESTORE (its default value is 0).

The optional conditions one can use before applying TWORUN, as well as use of the function LENGTHS are analogous to those described in the previous subsection.

Example 6.2. The sequence of statements

```
5 SET 1
6 SET -1
S←(3 1 3 1) TWORUN (1000 SIMNORM 0.1 1.4)
STATIST LENGTHS
```

will result in values of a symmetric two-sided Cusum-Shewhart scheme with parameters $h^+ = h^- = 3$ and $k^+ = k^- = 1$, corresponding to a simulated sequence of 1000 normal observations with mean 0.1 and standard deviation 1.4. The scheme is automatically re-set to 0 after each out-of-control observation. The first statement leads to automatic mode of execution. The last statement performs a statistical analysis of the resulting run lengths.

7. Other functions

1. The function QUIT. This function is used to "escape" from situations corresponding to an error discovered by APL in the middle of execution. We tried to build CONTRD in such a way that such situations would never occur (so, we hope that QUIT will remain the only function of the package that is never used). However, if they do, please type QUIT to return to the original conditions[29] and notify the author.

2. Functions for simulating random variables. The package includes a set of functions for simulating types of random variables typically encountered in practical applications. Every function whose name starts with letters SIM is one of such functions (ex. SIMBINOM, SIMNORM, etc.). A function of this type has a left argument, L, representing the quantity of variables to be generated and a right argument, R, characterizing the parameters of the distribution. For example 20 SIMNORM 2 0.3 will

---

[29] Warning: if such an error occurs during execution of an EXPLR - type function, the global (varying) parameter of the currently used d.f. may have a different value (after quitting) than that assigned before the execution of the function.

generate 20 normal random variables with mean 2 and standard deviation 0.3. Other possibilities are listed in Appendix E. Note that the function SIMNORM can also generate a matrix of normal variables with a specified "grand" mean, row-to-row variability and within-row variability.

3. The function IDENTIFY. This function is used to identify the names of global parameters of a given distribution function as well as their current values. The syntax is IDENTIFY DFNAME, where DFNAME is the name of APL function which returns the values of the d.f. of observations (see p. 29 and Appendix A). For example, IDENTIFY 'DFNORM' will identify the parameters of the normal distribution.

4. The function QUANTILE. This function computes a set of quantiles corresponding to a given distribution. For example, 0.05 0.5 0.95 QUANTILE 'DFNORM' will compute the quantiles of order 0.05, 0.5 and 0.95 corresponding to a normal distribution (with parameters given by global variables MEU and SIGMA). The function can be applied to distributions listed in Appendix A as well as to those written by the user.

5. The function FITDF. This function fits a specified distribution to a given set of data. For example, if D is a vector containing the data, the statement D FITDF 'DFNORM' will estimate the parameters of the normal distribution and display information related to quality of the fit (Chi-square statistic, Kolmogorov-Smirnov statistic, etc.). At present, the function can be applied to selected members of the list of distributions from Appendix A only.

6. The function STATIST. This function computes and prints out the basic statistics (ex. mean, range, etc.) associated with a given set of data. Its format is STATIST D, where D is either vector or matrix containing the data.

If D is a vector, three esimates of the standard deviation are included: the "usual" sample standard deviation, $\hat{\sigma}$, the standard error of the regression line , s_e (this estimator is invariant with respect to linear trends present in the data), and s_d defined in terms of the successive differences,

$$s_d^2 = \frac{1}{2(r-1)} \sum_{i=1}^{r-1} (x_{i+1} - x_i)^2; \qquad (7.1)$$

clearly, it is roughly invariant with respect to "shifting" portions of the data.

If D is a matrix (having, say, r rows and n columns), then every row is assumed to contain a sample of n measurements taken at the same moment of time. In this case the function will compute the "grand" characteristics of the pooled data (grand mean, median, etc.). In addition, it will estimate the within-row standard deviation,

$$s^2 = \frac{1}{r}\sum_{i=1}^{r} \hat{\sigma}_i^2 \tag{7.2}$$

($\hat{\sigma}_i$ is the sample standard deviation corresponding to the i - th row), and perform a regression analysis. As in the previous case, three estimates of the row-to-row standard deviation are available: the usual variance component estimate s_b,

$$s_b^2 = \left(\frac{1}{r-1}\sum_{i=1}^{r} (\bar{x}_i - \bar{\bar{x}})^2 - \frac{s^2}{n} \right)^+ \tag{7.3}$$

($\bar{\bar{x}}$ is the "grand" mean), the estimate s_{bd} based on successive differences of the sample means and, therefore, invariant with respect to possible "shifts" in the process level,

$$s_{bd}^2 = \left(\frac{1}{2(r-1)}\sum_{i=1}^{r-1} (\bar{x}_{i+1} - \bar{x}_i)^2 - \frac{s^2}{n} \right)^+, \tag{7.4}$$

and s_{br} based on filtering out the linear component, $s_{br}^2 = s_e^2 - s^2$. Clearly, this estimator is invariant with respect to linear trends in the process level.

Example 7.1. Suppose that we take a sample of 4 measurements per lot and using the sample statistics to control the process. Suppose that the lot-to-lot standard deviation is 0.3, the within-lot standard deviation is 0.1, and the "grand" mean (i.e. the process level) is 0. To simulate samples corresponding to 200 lots and then to compute (assuming normality) the basic statistics, one can use the statement

STATIST (200 4 SIMNORM 0 0.3 0.1)

7. The function STATMULT. This function computes and prints out the basic statistics (ex. means, standard deviations, estimates of the correlation matrix, etc.) associated with a given set of multivariate observations. Its format is STATMULT D, where D is a matrix with rows corresponding to values of the multivariate vectors. Note that multivariate normal vectors can be simulated by using the function SIMMULT.

Three estimates of the covariance structure are included: the "usual" method, the estimate based on successive differences of the multivariate vectors (so, it is roughly invariant with respect to "shifts" in levels of some of the variables) and the estimate obtained after removing the linear component from each variable of the multivariate vector (this estimate is invariant with respect to linear trends in some of the variables). The function implicitly produces global (covariance) matrices COVAR, COVARD and COVARR corresponding to these methods as well as the vector MEANS containing the multivariate means (i.e. column averages).

8. The function SELECT. This function selects a certain sequence of statistics from a given set of data for subsequent statistical analysis or application of a Cusum-Shewhart scheme. Its format is $V \leftarrow$ CD SELECT D, where D is the raw data (vector or matrix) and the code CD determines what statistic is to be selected. The result V contains the values of the selected statistic. The list of possible choices of CD corresponds to sequences that seem to be most likely candidates for analysis and/or scheme application; motivated users will find it easy to incorporate additional values of CD, as needed. If CD has several (say, k) components, each one will be used to select an appropriate sequence, and the resulting matrix V will consist of k corresponding columns.

a) D is a vector. If CD=1, then the vector D itself will be selected. If CD=0, the absolute values of successive differences scaled by $0.5\sqrt{\pi} \approx 0.8862$ will be selected (see (2.4)), i.e. V contains the sequence $0.5\sqrt{\pi}\ |\ x_{i+1} - x_i\ |\ ,\ i = 1,\ 2,\ \ldots\ ,$ where x_i are the elements of D (when measurements are taken one at a time, these differences can be used to control the lot-to-lot variability).

b) D is a matrix. In this case every row is assumed to contain a sample of measurements taken at the same moment of time. The selection proceeds as follows: CD=0 takes the sample means of the rows and then selects the scaled absolute values of successive differences (like in the case when D is a

vector); CD=1, 2, 3 or 4 selects the sample means (of the rows), sample standard deviations, skewnesses and curtoses of the rows, respectively; other values of code correspond to sequences which may depend on the the version of the package; the user can find this information in the on-line documentation.

Clearly, when CD is a vector, there is a possibility that some of the selected sequences are shorter than others. In order to be able to combine them into a single matrix, we insert an (artificial) zero leading element into the shorter sequences.

Example 7.2. Suppose that we take a sample of 4 measurements per lot and using the sample means to control the process level. Suppose that the lot-to-lot standard deviation is 0.3, the within-lot standard deviation is 0.1 and the "grand" mean (i.e. the process level) is 0. Let us simulate samples corresponding to 100 lots and apply the symmetric two-sided scheme $h = 0.4$, $k = 0.05$ to the sequence of sample means (we shall assume that the underlying distribution is normal):

$$S \leftarrow 0.4\ 0.05\ 0.4\ 0.05\ \text{TWORUN}\ (1\ \text{SELECT}\ (100\ 4\ \text{SIMNORM}\ 0\ 0.3\ 0.1))$$

9. The function ROUND. This function is useful when applying a Page's scheme to simulated data. It rounds the input vector (or matrix) up to a specified number of digits after the decimal point. For example, the statement $S \leftarrow 3$ ROUND S will round the elements of S up to 3 decimal places and then re-assign the result to S.

Acknowledgements

I would like to thank Betty Flehinger (IBM Research) for introducing me to the topic of Cusum techniques, constructive criticism and consultation, and active participation in ongoing activity in this area at manufacturing sites. I am also grateful to Norman Brenner (IBM Research) for help in the preparation of the software, Marvin Pittler (IBM East Fishkill) and David Withers (IBM Research) for help in identifying promising areas of application of Cusum techniques and establishing the necessary contacts, and numerous IBM-ers whose comments and criticisms were useful in shaping the present form of CONTRD.

Appendix A. List of available distribution functions.

Below is the list of distribution functions currently provided with CONTRD. Every function accepts a right argument (scalar or vector) only; before using it, one should make sure that the distribution parameters are specified correctly by means of the appropriate global variables. To identify the names of these variables and their current values, use the function IDENTIFY. To compute the quantiles of a given distribution, use the function QUANTILE. To fit a distribution of a given type to sets of data, use FITDF. All three functions are described in Sec. 7.

Function Name	Parameters	Comment
DFBETA	ALPHA, BETA	Beta distribution
DFBINOM	PROB, SAMPLE	Binomial with parameters p = PROB and n = SAMPLE
DFEMPIR	LOCATION, SCALE, DATA	Empirical (sample) distribution function based on observations given by DATA, shifted by substracting LOCATION and scaled by dividing the result by SCALE
DFEXP	THETA	Exponential distribution with mean THETA
DFEXTREME	LOCATION, SCALE	Shifted and scaled Least Extreme Value (double exponential) distribution
DFGAM2	ALPHA, BETA	Gamma with parameters ALPHA and BETA
DFGAM3	ALPHA, BETA, CENTER	Gamma with parameters ALPHA and BETA shifted so that the resulting mean is at CENTER
DFGEOM	PROB	Geometric with parameter PROB. The mean of this variable is 1/PROB and its possible values are 1,2,...
DFHYPGEOM	LOTSIZE, SAMPLE, DEFECTS	Hypergeometric with parameters N = LOTSIZE, n = SAMPLE and k = DEFECTS
DFLGST	LOCATION, SCALE	Logistic distribution
DFLOGN2	MEU, SIGMA	Lognormal
DFLOGN3	MEU, SIGMA CENTER	Three-parametric Lognormal with location parameter CENTER

DFNCF	DEGRNUM DEGRDEN NONCNTR	Non-central F with non-centrality parameter NONCNTR ≥ 0, DEGRNUM degrees of freedom of the numerator and DEGRDEN degrees of freedom of the denominator. NONCNTR = 0 corresponds to the central F-distribution
DFNCHI	DEGREES, NONCNTR	Non-central chi square distribution with non-centrality parameter NONCNTR ≥ 0 and DEGREES degrees of freedom. NONCNTR=0 corresponds to the central chi square distribution.
DFNEGB	DEFECTS, PROB	Negative binomial (i.e. distribution of the sum of DEFECTS geometric random variables with mean 1/PROB)
DFNORM	MEU, SIGMA	Normal
DFPOIS	LAMBDA	Poisson
DFPROPA	LOTSIZE, SAMPLE, DEFECTS	Distribution of the sample proportion of defectives without replacement (i.e. of the hypergeometric variable with parameters N = LOTSIZE, n = SAMPLE and k = DEFECTS, divided by the sample size, n)
DFPROPB	PROB, SAMPLE	Distribution of the sample proportion of defectives with replacement (i.e. of the binomial random variable with parameters p = PROB and n = SAMPLE divided by n)
DFPROP	PROB, SAMPLE	Distribution of the sample proportion with replacement when the sample size is a Poisson random variable with mean SAMPLE. The theoretical proportion is p = PROB
DFRANGE	SIGMA, SAMPLE	Distribution of the sample range corresponding to sample of size SAMPLE taken from a normal population with st. deviation SIGMA
DFS	SIGMA, SAMPLE	Distribution of the sample standard deviation (2.3) corresponding to sample of size SAMPLE taken from normal population with st. deviation SIGMA.
DFSTUD	DEGREES	Student's distribution
DFUNIF	LBOUND, UBOUND	Uniform
DFWEIB2	SCALE, SHAPE	Weibull with parameters SCALE, SHAPE
DFWEIB3	SCALE, SHAPE, CENTER	Three-parametric Weibull with parameters SCALE, SHAPE and location parameter CENTER

DFXBAR	MEU, SIGMA SAMPLE	Distribution of the sample mean corresponding to sample of size SAMPLE taken from a normal population with mean MEU and st. deviation SIGMA.

The last two functions compute the d.f. of the Mahalanobis distance between the mulrivariate normal sample mean and some fixed point (ex. centroid of the target region), μ_0. The parameter MAHAL represents the Mahalanobis distance λ (with respect to the covariance matrix) between the population mean and μ_0 (see (B.6)). DEGREES is the dimension of the multivariate observation and SAMPLE is the sample size.

DFMAHAL	MAHAL, DEGREES, SAMPLE	Distribution of the Mahalanobis distance with respect to a "true" covariance matrix.
DFHOTEL	MAHAL, DEGREES, SAMPLE	Distribution of the Mahalanobis distance with respect to an esimated covariance matrix, S. The function F ($\sqrt{x/SAMPLE}$) corresponds to a Hotelling's T^2 - distribution.

Appendix B. Examples

Example B.1 (Cumulative s - chart). Consider the situation described in Example 2.1. Let the observed process of (within sample) standard deviations be $\hat{\sigma}_1$, $\hat{\sigma}_2$, ... (see (2.3)). Suppose that for every lot i, the measurements $\{y_{i1}, y_{i2}, \ldots y_{in}\}$ can be viewed as independent realizations of a normal random variable with certain mean and standard deviation σ. Let us assume that the sample size is fixed ($n = 4$), and that because of the regular equipment maintenance operations the planning "horizon" does not extend beyond 200 samples; in particular, if the process is on-target, the probability of not getting a false signal within 200 samples should be at least 0.99. Under these assumptions, we would like to find a Cusum-Shewhart scheme for controlling σ with the best possible out-of-target performance.

To design a scheme satisfying these criteria we use the special function CUSUMS:

```
      4 2 4 CUSUMS 200 0.99

The observations are distributed as S with  SIGMA=2  SAMPLE=4
Step1: Search for H satisfying Prob.(R.L.>200)=0.99
Step1 complete; H=3.51342, Prob.(R.L.>200)=0.990356
Analysis of upper Cusum scheme with parameters  H,K= 3.51342 2.76395
The level of discretization is 30
The changing parameter name is SIGMA

   SIGMA      ARL     SDRL
     2.0 20411.4 20408.7
     4.0     4.6      2.9

Enter the values of H, K, HEADST. and CD for further analysis (or 0 to exit):
      3.4 2.8 0 12
Enter additional values of SIGMA for which the analysis is to be performed:
      2.5 3 3.5
Enter the values of R for which Prob.(R.L. > R) is to be computed:
      5 10 200

Analysis of upper Cusum scheme with parameters  H,K= 3.4 2.8
The level of discretization is 30
The observations are distributed as S with  SAMPLE=4
The changing parameter name is SIGMA

   SIGMA      ARL     SDRL        5       10      200
     2.0 20529.0 20526.5  .99985  .99961  .99040
     2.5   141.0   137.9  .98160  .94776  .23902
     3.0    18.1    15.4  .83412  .60713  .00000
     3.5     7.4     5.4  .54580  .21099  .00000
     4.0     4.6     2.9  .28578  .04685  .00000

Enter the values of H, K, HEADST. and CD for further analysis (or 0 to exit):
      0
```

As we can see, the ARL corresponding to σ=4 is 4.6. One could roughly predict this value without going into computations by noting that every observation contributes about $E(\hat{\sigma}) - k = \sigma/c(n) - k$ to the upper Page's scheme. Therefore,

$$ARL \approx \frac{h}{\sigma/c(n) - k}, \qquad (B.1)$$

which in our case ($h = 3.4, k = 2.8, c(4) = 1.09, \sigma = 4$) results in 4.5. Clearly, such approximation is appropriate only for off-target situations in which the mean of the observations substantially exceeds the reference value. The formula (B.1) neglects the effect of reflection of the Page's scheme at 0, which tends to overestimate the ARL. On the other hand, in many cases this tendency is more than compensated by the fact that the amount of overshoot of the Page's scheme over h at the moment of signal is not taken into account. In other words, the degree of accuracy of (B.1) is typically unclear; nevertheless, formulas of this type may serve as a yardstick for the purpose of rough assessment of the ARL curve in off-target region, choice of the initial approximation for h to be used in the design procedure, etc.

Finally, let us mention another approximation, related to analysis of *on-target* situations. Suppose that we would like to evaluate the quantile of order γ of the RL distribution, i.e. to find q satisfying Prob.(RL $> q$) = γ. Since the on-target RL usually has a distribution somewhat similar to exponential, one can approximate q by

$$q \approx -ARL \times \log \gamma. \qquad (B.2)$$

This approximation can be rough, especially for high values of γ. For example, for $\gamma = 0.99961$ and ARL=20529, it results in $q \approx 8$, while the above output indicates that the quantile corresponding to σ=2 is q=10. However, it usually produces a good initial "guess" that can be subsequently refined; moreover, it is of use in cases where no analytic analysis of the RL is possible, and one has no choice but to use simulation (see Example B.7).

Example B.2 Consider once more the situation described in Example 2.1. Under the assumption that the sample (subgroup) size is $n = 4$ and the standard deviation of a single measurement is $\sigma = 3$, let us examine the performance of the two-sided Page's scheme with parameters ($h^+ = 9$, $k^+ = 3$, $s_0^+ = 2$, $h^- = 5$, $k^- = 2$, $s_0^- = 1$) considered in this example with respect to the the following values of the process level: $\mu = 6, 3, 1, 0, -2, -4$.

To perform the analysis, we use the function TWOXPLR:

```
'SIGMA SAMPLE'  ASSIGN  3 4
'MEU'  SETXPLR  (6, 3, 1, 0, (-2), (-4))
0  SET  2 1
9 3 5 2  TWOXPLR  'DFXBAR'
```

```
Analysis of the two-sided Cusum scheme with parameters:
H+,K+ = 9 3  and H-,K- = 5 2
The levels of discretization are D+,D- = 30 17
The observations are distributed as X-bar with  SIGMA=3  SAMPLE=4
The changing parameter name is MEU
The values of the headstarts are 2 1

  MEU  P(UP)      ARL     SDRL
    6  1.000      2.9       .9
    3  1.000     47.5     41.8
    1   .102 .484E07 .484E07
    0   .000 39719.5 39722.0
   -2   .000     19.1     16.5
   -4   .000      2.8      1.2
```

Example B.3: Spin dryers are used as one of the steps in the production of integrated circuit chips from semi-conductor wafers. Typically, the process steps are followed by rinses with deionized, filtered water. After the rinsing, the water is removed by placing the wafers into the spin dryer (centrifugal device), that spins the water off the wafers (and accelerates evaporation by using dry filtered gas).

Periodically, test wafers are run through the rinse and drying cycle and the particles on the wafer that are larger than a specified diameter are counted. The recorded counts serve as a basis for the decision to clean and re-test the spin dryer. Under normal conditions, the level of the process $\{ o_1, o_2, \ldots \}$ of the recorded counts does not exceed 6. Levels of the process exceeding 12 are associated with a high rate of defective production -- situations in which the process of counted contaminating particles reaches this level should be detected as soon as possible. On the other hand, since cleaning and re-testing represent an expensive and tedious procedure, we are interested in a cusum control scheme for which the probability of a false signal within 100 tests is not more than 0.01, and, at the same time, the sensitivity with respect to the levels of the process exceeding 12 is as high as possible.

On the basis of theoretical considerations, there is reason to believe that (during a certain initial period of time) the counts $\{ o_i \}$ form a sequence of iid Poisson random variables with parameter λ.

To perform the design and analysis of such cumulative c - chart, we use the function CUSUMC:

```
6 12 CUSUMC 100 0.99

The observations are Poisson with mean  LAMBDA=6
Step1: Search for H satisfying Prob.(R.L.>100)=0.99
Upper bound of the search has been reached. Repeat the search with
a more precise estimate of the signal level. The last approximation:
H=4.55  Level of Discr.=46  Prob.(R.L.>100)=0.47
```

Unfortunately, our first attempt fails. The output indicates that the sought value of h is substantially higher than 4.55. So, let us introduce the initial point of search $h_0 = 10$ and (to save CPU time) use the interval of discretization $\delta = 0.2$.

```
      6 12 10 0.2 CUSUMC 100 0.99
The observations are Poisson with mean  LAMBDA=6
Step1: Search for H satisfying Prob.(R.L.>100)=0.99
Step1 complete; H=10.9, Prob.(R.L.>100)=0.9897
Analysis of upper Cusum scheme with parameters  H,K= 10.9 8.6
The level of discretization is    55
The changing parameter name is LAMBDA

  LAMBDA      ARL     SDRL
       6   9458.8   9455.9
      12      4.0      1.8

Enter the values of H, K, HEADST. and CD for further analysis (or 0 to exit):
      10.9 8.6 0 123
Enter additional values of LAMBDA for which the analysis is to be performed:
      8 10
Enter the values of R for which Prob.(R.L. > R) is to be computed:
      5 10 100
Enter the Shewhart's Limit:
      18.5

Analysis of upper Cusum scheme with parameters  H,K= 10.9 8.6
The level of discretization is    55
The observations are Poisson
The changing parameter name is LAMBDA
The scheme is supplemented by the Shewhart limit 18.5

  LAMBDA      ARL     SDRL        5       10      100
       6   8685.1   8682.4   .99967   .99910   .98880
       8     54.5     50.3   .96458   .88249   .14739
      10      8.1      5.0   .63735   .23663   .00000
      12      4.0      1.9   .17021   .00654   .00000

Enter the values of H, K, HEADST. and CD for further analysis (or 0 to exit):
```

•
•
•

The above output suggests that the scheme $h = 10.9$, $k = 8.6$ assures about the best possible sensitivity (ARL=12) with respect to $\lambda = 12$ particles/wafer. In order to improve sensitivity with respect to very high levels of contamination, we supplemented the scheme by a Shewhart limit, $c = 18.5$. Could these results be independently verified, say, by using simulation? In the off-target case this is not difficult to do:

```
      5 SET 1
      6 SET -1
      S← 10.9 8.6 18.5  ONERUN  (2000 SIMPOIS 12)
      STATIST LENGTHS

Number of observations: 509    Mean: 3.93    Median: 4
Minimum: 1   Maximum: 13    Range: 12
Estimates of Stand. deviation: S=1.93    SD=1.87    SR=1.93
Regression slope: 0.00088   Skewness: 1.38   Kurtosis: 3.27
```

In the on-target situation, however, the simulation may become very expensive. Indeed, one would need to generate about 9000 Poisson variables in order to obtain a *single* out-of-control signal!

Clearly, the number of observations needed in order to get a good estimate of the ARL and other relevant quantities could well run into millions. This example shows that in the problem of analysis of control schemes analytic methods can produce results which cannot be obtained by simulation. There is, however, another way of verifying the results by using simulation, which is outlined in the Example B.4.

Now let us suppose that the process operates for a long time at the level $\lambda = 6$ particles/wafer, and then its level shifts to 12 particles/wafer. What can we say about the distribution of the Residual Run Length? To answer this question, we can invoke the steady state analysis (see Appendix D) right from CUSUMC and continue its run as follows:

```
                                    •
                                    •
                                    •
Enter the values of H, K, HEADST. and CD for further analysis (or 0 to exit):
      10.9, 8.6, (-1)

Analysis of upper Cusum scheme with parameters  H,K = 10.9 8.6
The level of discretization is 55
The observations are Poisson
The changing parameter name is LAMBDA
The headstart is out of range; steady state analysis assumed
The scheme is supplemented by the Shewhart limit 18.5

  LAMBDA      ARL     SDRL        5       10      100
       6   8682.9   8682.3   .99943   .99885  .98855
       8     54.1     50.3   .95950   .87667  .14638
      10      7.9      5.0   .62233   .22985  .00000
      12      3.9      1.8   .16271   .00621  .00000

Enter the values of H, K, HEADST. and CD for further analysis (or 0 to exit):
0
```

As we can see, the outlook of sensitivity is somewhat better in terms of the Residual RL. This is clearly related to the fact that the scheme may have a non-zero value (headstart) at the moment the shift occurs.

Example B.4 (Analysis of a scheme on the basis of an empirical distribution). In our previous discussion we always assumed that our observations come from some known family of distribution functions. Under this assumption, control of the process becomes essentially control of certain "crucial" parameters of the family (ex. normal mean). However, it occurs quite frequently that the only thing we know about the process is data (corresponding to "good" and/or "bad" states of the process), and we would not like to commit ourselves to any particular family of distributions. In such situations one can use the empirical (or sample) distribution function instead of the unknown distribution for purposes of design and analysis.[29]

To illustrate this approach let us first assume that the observations come from a normal family with standard deviation 1 and examine the performance of a scheme $h = 3$, $k = 1$ with respect to the process levels $\mu = 0.5$, 1, 1.5 and 2:

```
 2 SET 10 20 50
'SIGMA' ASSIGN 1
'MEU' SETXPLR 0 0.5 1 1.5 2
 (3 1) ONEXPLR 'DFNORM'

Analysis of upper Cusum scheme with parameters  H,K = 3 1
The level of discretization is 30
The observations are normal with  SIGMA=1
The changing parameter name is MEU

   MEU      ARL     SDRL       10       20       50
     0   1958.0   1955.6  .99589  .99081  .97572
   0.5    117.5    114.4  .93773  .85926  .66098
   1.0     17.4     14.2  .60252  .29662  .03529
   1.5      6.4      3.8  .12799  .00800  .00000
   2.0      3.7      1.7  .00464  .00000  .00000
```

Now let us generate a set of 10000 standard normal observations and assume that this is the data at hand. Suppose that we know that the typical ways of our process going out of control are related to shift and/or scaling of the appropriate on-target distribution. Thus, let us try to shift our data by 0.5, 1, 1.5 and 2 and examine the performance of the above scheme with respect to the corresponding empirical distributions:

---

[29] This type of approach leads to so-called "bootstrap" estimates of the characteristics of the Run Length. For more information about this technique see Efron (1981).

```
'DATA' ASSIGN (10000 SIMNORM 0 1)
'SCALE' ASSIGN 1
'LOCATION' SETXPLR 0 0.5 1 1.5 2
 (3 1) ONEXPLR 'DFEMPIR'

Analysis of upper Cusum scheme with parameters  H,K = 3 1
The level of discretization is 30
The observ. come from empirical d.f. with  SCALE=1
The changing parameter name is LOCATION

LOCATION      ARL     SDRL       10       20       50
       0   1725.1   1722.9   .99523   .98947   .97239
     0.5    113.4    110.3   .93571   .85463   .65106
     1.0     16.9     13.8   .59335   .28559   .03175
     1.5      6.3      3.8   .12173   .00711   .00000
     2.0      3.7      1.7   .00418   .00000   .00000
```

As one could expect, the results are fairly close to those obtained under the normal assumption. By varying the second parameter of the empirical distribution, SCALE, we could examine the effect of varying the standard deviation of the underlying normal process.

Of course, on the basis of our data, we could first estimate the standard deviation, then assume normality and, finally, analyze the scheme under the (fitted) normal model. However, the resulting characteristics of the Run Length would depend strongly on validity of the assumption of normality. The main moral of this example is, of course, that assumptions of this type can be completely unnecessary, especially in situations where substantial amount of data is available - one can simply use the empirical distribution.

There is, however, another important application of this technique. Indeed, the functions of CONTRD enable us to mathematically analyze any distribution of incoming (iid) observations for which we are able to provide an appropriate APL function (DFNORM, DFPOIS, etc.). However, in some cases it may be not easy to write such a function. In such cases analysis can still be performed by using the above technique, provided one can efficiently generate the underlying sequences of observations. Consider, for example, the situation in which one is trying to monitor the wafer-to-wafer variability within successive lots. Suppose that r wafers are selected at random from each lot, then n measurements are taken from each wafer, and a variance component type esimate of the wafer-to-wafer variance σ_b^2 is computed by using (7.3). This estimator is distributed as a non-negative part of

$$V_1[r-1] \times (\sigma_b^2 + \sigma^2/n) \; - \; V_2[r(n-1)] \times \sigma^2/n,$$

where σ is the within-wafer standard deviation and $V[i]$ corresponds to a chi - square random variable with i degrees of freedom divided by i (the variables V_1 and V_2 are independent).

Writing an APL function to efficiently compute the distribution function of the above estimator may prove to be a tedious task. On the other hand, it can be easily simulated; the analysis can then be performed on the basis of the resulting empirical distribution.

Example B.5 (Controlling the mean of a Weibull population). Consider a line for serial production of certain electronic devices. Every 10 minutes a device is picked from the conveyor and subjected to an accelerated life test. The resulting sequence of life times serves as a basis for assessing the quality of the process. Statistical data analysis indicates that the life time of the device is typically distributed as a Weibull r.v. with a relatively stable shape parameter $c = 4$. The variability in life times (and, of course, the process quality) depends primarily on the scale parameter, α.

If the level (mean) of the sequence exceeds 1.5 min, the process quality is satisfactory; under these conditions we are definitely interested in protection against false alarms. If, however, the level falls to 0.5 min, we would like to detect it as soon as possible; in such a case we still can afford to be late by a half an hour or so. Thus, let us try to find an appropriate Cusum control scheme based on the sequence of recorded life times. We start by noting that it makes sense to choose the midpoint $k = -1$ as a reference value of our (lower) control scheme. Since the mean of the Weibull population is $\alpha\,\Gamma(c^{-1}+1)$, the values of α corresponding to the mean life of 1.5, 1 and 0.5 are 1.7, 1.1 and 0.6, respectively. So, let us find a scheme for which the off-target ARL is 3 (given our sampling intensity, this corresponds to ATS ≈ 30 min):

```
'SHAPE SCALE' ASSIGN 4 0.6
'SCALE' SETXPLR 1.7 1.1 0.6
0 SETFIND 3
(3, (-1)) ONEXPLRDL 'DFWEIB2'

The observations are Weibull with  SHAPE=4 SCALE=0.6
Step1: Search for H satisfying ARL=3
Step1 complete; H=1.10625, ARL=2.9836
Analysis of lower Cusum scheme with parameters  H,K = 1.10625 -1
The level of discretization is 30
The changing parameter name is SCALE

  SCALE      ARL     SDRL
    1.7   5373.2   5370.9
    1.1     25.1     20.4
    0.6      3.0       .6
```

Thus, our "brutal" approach results in a scheme with a seemingly reasonable resolution. Can it be substantially improved? To answer this question, we should recall that under the Weibull assumption, sum of the c-th powers of the observations is a sufficient statistics for α and, therefore, one can expect that a more powerful control procedure (when c is known) can be based on the c-th powers of the recorded life times. It is easy to see that c-th power of a Weibull observation is distributed

exponentially with mean α^c. Since $(1.7, 0.9, 0.6)^4 = 8.3521, 0.6561, 0.1296$, we can use the function CUSUMT as follows:

```
              8.3521 0.1296 CUSUMT 3 0 1

The observations are Exponential with mean  THETA= 0.1296
Step1: Search for H satisfying ARL=3
Step1 complete; H=0.998378, ARL=2.96237
Analysis of lower Cusum scheme with parameters  H,K= 0.998378 -0.548399
The level of discretization is 34
The changing parameter name is THETA

   THETA      ARL     SDRL
  8.3521   7345.9   7343.9
  0.1296      3.0       .6

Enter the values of H, K, HEADST. and CD for further analysis (or 0 to exit):
      1 0.55 0 1
Enter additional values of THETA for which the analysis is to be performed:
       0.6561

Analysis of lower Cusum scheme with parameters  H,K= 1 -0.55
The level of discretization is 30
The observations are Exponential
The changing parameter name is THETA

   THETA      ARL     SDRL
  8.3521   7107.2   7105.2
  0.6561     12.2      9.5
  0.1296      3.0       .6

Enter the values of H, K, HEADST. and CD for further analysis (or 0 to exit):
0
```

The above output shows that under our conditions, control scheme based on the 4-th powers of the life times would increase the resolution by about 40%. We must, however, warn the reader not to misinterpret this statement. Indeed, instead of matching the sensitivity, let us try to match the on-target performance of the original scheme, i.e. design a scheme for which the ARL is 5373.2:

```
              8.3521 0.1296 CUSUMT 5373.2

The observations are Exponential with mean  THETA= 8.3521
Step1: Search for H satisfying ARL=5373.2
Step1 complete; H=0.968575, ARL=5338.39
Analysis of lower Cusum scheme with parameters  H,K= 0.968575 -0.548399
The level of discretization is 33
The changing parameter name is THETA

   THETA      ARL     SDRL
  8.3521   5338.4   5336.5
  0.1296      2.9       .6
```

Thus, for a fixed on-target ARL, use of the transformed observations reduces the sensitivity from 3 to 2.9, which could hardly be viewed as a dramatic improvement. This example shows not only that the scheme based on the recorded life times may be not so bad after all, but also that one must be

prepared to sacrifice a lot of protection against false alarms in order to "buy" a relatively small amount of sensitivity. The reason for that is related to the fact that, as h increases, the off-target ARL grows at a roughly *linear* rate, while the on-target ARL grows *exponentially*. One should keep in mind, however, that reduction of the off-target ARL by as little as 0.1 may sometimes significantly reduce the right tail of the RL distribution (i.e. probability of not catching the change within a short period of time). Indeed, ARL=1 means that the change in process level will be definitely detected in the first sampling period; ARL=1.1 means that with probability roughly 10% the change will not be detected immediately.

Example B.6 In the process of galvanic plating, one is interested in monitoring the concentration of copper in the bath. In order do so, he takes periodic measurements from a prescribed place in the bath. The measurements are relative with respect to some target concentration; so, when the process is on target, the level μ of the sequence of recorded measurements must be 0. On the other hand, if the process level settles outside the interval (-2,2), quick corrective action is required; in particular, if $\mu = \overline{+}\ 2$, we can still tolerate it for about an hour, provided there is no deterioration of the process variability. Under normal operating conditions, the relative measurements are normally distributed with mean μ and standard deviation $\sigma = 1.2$, which incorporates both the measurement error and variability in time.

Our main objective is to determine the sampling intensity (SI) necessary for being able to derive a Cusum scheme with satisfactory resolution. We shall characterize the performance of a scheme in terms of its Time to Signal (TS), which corresponds to Run Length divided by the sampling intensity. We shall require that for $\mu = 2$ and -2 the Average TS (ATS) = 1 hour (in other words, the ARL for $\mu = 2$ and -2 must be 5). First, we examine the scheme corresponding to SI = 5 measurements/hour. We start by designing an upper scheme satisfying ATS=1; subsequent combining of it with an analogous lower scheme yields the symmetric two-sided scheme of interest.

```
      1 1.2 0 2  CUSUMX  5 0 1

The observations are distributed as X-bar with
MEU=2  SIGMA=1.2  SAMPLE=1
Step1: Search for H satisfying ARL=5
Step1 complete; H=4.24474, ARL=4.99342
Analysis of upper Cusum scheme with parameters  H,K = 4.24474 1
The level of discretization is 30
The changing parameter name is MEU

     MEU      ARL     SDRL
       0   1755.6   1752.3
       2      5.0      2.4

Enter the values of H, K, HEADST. and CD for further analysis (or 0 to exit):
      4.24 1 0 1012
Enter additional values of MEU for which the analysis is to be performed:
      1
Enter the values of R for which Prob.(R.L. > R) is to be computed:
      5 10 50 100

Analysis of the two-sided Cusum scheme with parameters:
H+,K+ = 4.24 1  and H-,K- = 4.24 1
The levels of discretization are D+,D- = 30 30
The observations are distributed as X-bar with  SIGMA=1.2  SAMPLE=1
The changing parameter name is MEU

     MEU  P(UP)      ARL     SDRL        5       10       50      100
       0    .500    872.0    868.7   .99732   .99176   .94713   .89416
       1   1.000     22.1     18.1   .90845   .70529   .07592   .00466
       2   1.000      5.0      2.4   .32883   .03094   .00000   .00000

Enter the values of H, K, HEADST. and CD for further analysis (or 0 to exit):
0
```

As one can see, combining the one-sided schemes into a two-sided one does not have much effect on the detection capability. Indeed, if the process level settles at say, -2, the upper scheme is practically idle and the two-sided scheme becomes operationally equivalent to a one-sided (lower) scheme. On the other hand, such combining increases the risk of a false alarm; in particular, for our symmetric scheme it reduces the ATS by a factor of 2, to 872.0/5 = 174 hours; the probability of not having a false alarm within 10 hours is 0.94713.

A sample run corresponding to this scheme is given in Fig. B.1. The first 30 observations of this run correspond to the process level μ=0, the next 30 - to μ=2. As we can see, our Cusum scheme signals at the observation 36, i.e. RL=6 (TS=1.2 hour), as could be expected from the design considerations. For comparison, we also plotted the data and applied the Shewhart 3-Sigma control limits. In this particular run, the Shewhart scheme signals later: RL=8 (TS=1.6 hour); the comparison treats the Shewhart scheme more than fairly, since, as one can see from Fig. 1.1., its ATS is (740/2)/5 = 74 hours, which is worse than on-target ATS=174 of the Cusum scheme.

In the process of assessment whether performance of the derived Page's scheme is satisfactory or not, one should take into account, among the other things, the fact that the level of concentration of copper is not likely to be the only parameter to be controlled - the control system will probably process concentrations of other chemicals, variability in time, etc. If, say, the system was to control 15 parameters of the above type, the probability of not having a false alarm within 10 hours becomes $(0.94713)^{15} \approx 0.45$. Clearly, one could improve the performance by increasing the sampling intensity. For example, raising the intensity to 10 measurements/hour leads to a (symmetric) scheme h=9.298, k=1 with on-target ATS of 98000 hours and probability of having no false alarm within 10 hours 0.9999. Another way to improve the resolution would be to take several measurements at a time, and base the control procedure for μ on the sample averages; this could reduce the "within-bath" portion of the total variability. As an exercise, we suggest the reader to examine these possibilities in more detail.

In conclusion, let us give some remarks related to the principle of immediate utilization of the incoming information mentioned in the Introduction. Suppose that one measurement is taken every 0.1

hour, and we have an option to use subgroups of size 1 (i.e. no subgrouping), 2, 5 or 10. For every subgrouping policy, let us design a Cusum scheme having an on-target ATS of 98000 hours, and consider the off-target performance for $\mu = 2$ and $\mu = 3$. The results are as shown in Table B.1.[31]

Sample size	Samples per hour	Signal level	ATS	P(TS > 1)	P(TS > 2)
1	10	9.30	1.00 (0.53)	0.38 (0.0018)	0.013 (0)
2	5	4.38	1.02 (0.55)	0.35 (0.0013)	0.011 (0)
5	2	1.56	1.10 (0.60)	0.28 (0.0006)	0.009 (0)
10	1	0.68	1.20 (1.00)	0.20 (0.0003)	0.006 (0)

Table B.1. Effect of various subgrouping policies on the resolution power of a two-sided scheme. The observations are iid normal with $\sigma = 1.2$, and both reference values are set to 1. The (off-target) ATS and associated probabilities are computed for $\mu = 2$ and 3 (in parentheses). Every scheme has an on-target ATS of 98000 hours.

This table indicates that best resolution power (in terms of the ATS) is achieved when no artificial subgrouping is used, i.e. the scheme is updated 10 times per hour. Such policy is especially of use in situations where large deviations of the process level from its target value are possible. For example, for $\mu = 3$ the off-target ATS is 0.53 hours compared to 1 hour corresponding to the case when a subsample of size 10 is taken once per hour; clearly, this may represent a serious advantage, especially in the environment of conveyor-type manufacturing, where the off-target ATS can be directly translated into amount of substandard product.

On the other hand, the probability of not catching the same value of μ within an hour is smaller when artificial subsampling *is* used! This phenomena is not difficult to explain: when the scheme is updated frequently, shorter values of the TS become possible and the ATS is driven down; however, some of the information can be lost along the way, because of the possible regenerations of the underlying Page's schemes - therefore, some longer runs also become possible.[32] In contrast, schemes based on subgrouping tend to have a longer "memory" (and therefore are somewhat less likely to overlook a long period of persistently poor quality), but are unable to react quickly. The mentioned "memory"

---

[31] To obtain, say, the second entry of this table, note that the on-target ARL of the one-sided scheme must be 98000 × 2 × 5 = 980000; thus, the analysis can be performed by using the statement 2 1.2 0 2 CUSUMX 980000.

[32] The probabilities of such runs can usually be driven down by supplementing a Cusum scheme with Shewhart's limits.

is not always a plus, as it introduces inertia into the control process. Indeed, assume that the process level changed after a 1/2 hour period. Then half of the observations used in the next updating of a scheme based on subsamples of size 10 are "worse than irrelevant".

In summary, we feel that sampling as frequently as possible and updating the scheme(s) as soon as the new information arrives still represents a good policy, since its minor drawbacks are typically more than compensated by our improved ability to react quickly to large deviations in the process level.

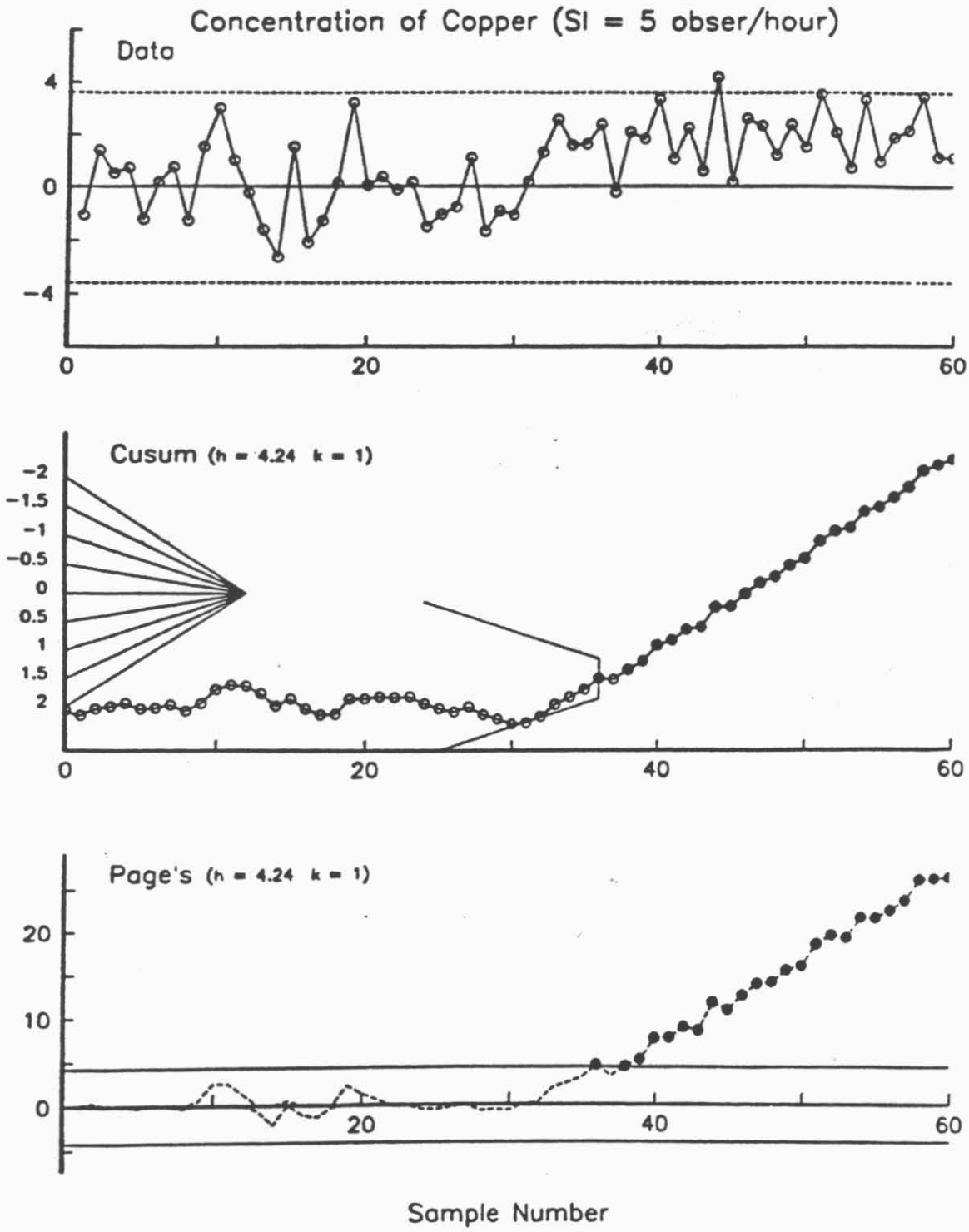

Fig. B.1. A sample run corresponding to Example B.6.

Example B.7 Consider the process of wiring the chips in which we are interested in controlling the height of lines connecting the transistors. The processing is on a lot by lot basis; the control procedure should be based on 12 measurements corresponding to three wafers (4 measurements/wafer) taken from each lot. In accordance with specifications, the height of *every* wire should be between 3 and 5 micron. The primary purpose of the control scheme is to keep the escape rate on this characteristic as low as possible. Extensive data corresponding to various periods of time during which the process operated in a stable and satisfactory mode brought the engineers to the following conclusions:

a)About 50 lots are processed every shift.

b)The process level should be as close as possible to 4 mic.

c)Most of the process variability is attributed to its lot-to-lot component σ_b, which is typically around 0.1 mic; levels exceeding 0.2 were never observed under normal operating conditions and could be considered as "bad". The within-wafer variability is very stable, at the level of $\sigma = 0.05$ mic. The wafer-to-wafer variability within the lot can be neglected.

d)The model for the j-th measurement of the i-th lot

$$y_{ij} = \mu + L_i + \varepsilon_{ij}, \quad j = 1, 2, \ \ldots \ 12, \tag{B.3}$$

where μ is the grand mean (level) of the process, L_i is the effect of the lot (normal with mean 0 and s.d. σ_b), and ε_{ij} is the "noise" (normal with mean 0 and s.d. σ) can be reasonably assumed for the purpose of design and analysis of the control schemes of interest.

The above conclusions imply that under satisfactory operating conditions, the total variability is around $\sqrt{0.1^2 + 0.05^2} = 0.11$ and should not exceed $\sqrt{0.2^2 + 0.05^2} = 0.21$. Therefore, one could consider the interval $(5 - 3 \times 0.21,\ 3 + 3 \times 0.21)$ as a target region for the grand mean; however, let us be conservative and require protection from false alarms if μ belongs to a shorter interval, (3.8, 4.2). Further, if μ settles at the level above $5 - 3 \times 0.11 = 4.67$ or below $3 + 3 \times 0.11 = 3.33$, it becomes impossible to eliminate defective product; to protect against unexpected sources of vari-

ability, we shall declare $\mu \leq 3.4$ and $\mu \geq 4.6$ as regions for which the cumulative $\bar{X}$ - chart should have good detection capability, i.e. "bad" regions (see Fig. B.2).[32]

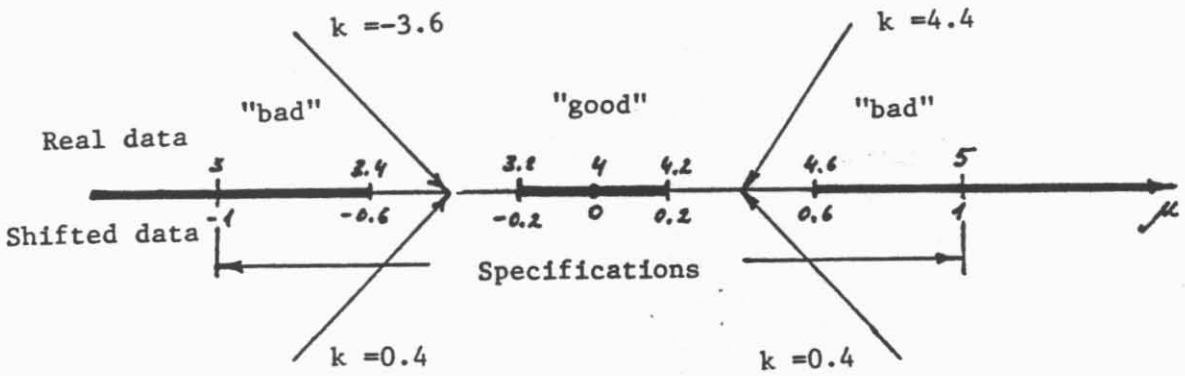

Fig. B.2. "Good" and "bad" domains for the grand mean.

Now let us try to design the mentioned $\bar{X}$ - chart for controlling μ and see what resolution can be obtained under the present SI. Our immediate concern is related to sensitivity: in particular, under normal operating conditions, we would like to assure that the probability of detecting the event $\mu \leq 3.4$ or $\mu \geq 4.6$ within 2 lots be at least 0.95. The s.d. of $\bar{X}$ under normal conditions is $\sqrt{0.1^2 + 0.05^2/12} = 0.101 \approx 0.1$, which clearly indicates that the within-sample variability is not likely to affect the performance of the chart. In order to be able to design a symmetric *two-sided* scheme without exiting from the special function CUSUMX, let us shift the measurements by 4, so that the target level becomes 0. Clearly, the target region for the shifted data becomes (-0.2, 0.2); levels beyond the interval (-0.6, 0.6) correspond to "bad" state of the process (see Fig. B.2). As usual, we start by designing an upper scheme, and then symmetrize it:

---

[32] Note that the analysis of the domain for μ is based primarily on the behavior of total variability and does not take into account the sampling intensity. Indeed, the latter does not determine what is "good" and "bad" behavior of the underlying process, but whether the inflow rate of information enables us to *resolve* between these levels.

```
      1 0.1 0.2 0.6  CUSUMX  2 0.05 1

The observations are distributed as X-bar with
MEU=0.6  SIGMA=0.1  SAMPLE=1
Step1: Search for H satisfying Prob.(R.L.>2)=0.05
Step1 complete; H=0.174907, Prob.(R.L.>2)=0.0492814
Analysis of upper Cusum scheme with parameters  H,K = 0.174907 0.4
The level of discretization is 30
The changing parameter name is MEU

     MEU      ARL     SDRL
     0.2   9403.1   9402.4
     0.6      1.5       .6

Enter the values of H, K, HEADST. and CD for further analysis (or 0 to exit):
      0.175 0.4 0 1012
Enter additional values of MEU for which the analysis is to be performed:
      0 0.4
Enter the values of R for which Prob.(R.L. > R) is to be computed:
      2 5 50 100

Analysis of the two-sided Cusum scheme with parameters:
H+,K+ = 0.175 0.4  and H-,K- = 0.175 0.4
The levels of discretization are D+,D- = 30 30
The observations are distributed as X-bar with  SIGMA=0.1  SAMPLE=1
The changing parameter name is MEU

     MEU  P(UP)      ARL     SDRL        2        5       50      100
       0    .500 .112E09 .112E09 1.00000 1.00000 1.00000 1.00000
     0.2  1.000   9435.8   9435.1   .99981   .99949   .99473   .98948
     0.4  1.000      8.5      7.0   .86364   .57031   .00085   .00000
     0.6  1.000      1.5       .6   .04936   .00004   .00000   .00000

Enter the values of H, K, HEADST. and CD for further analysis (or 0 to exit):
0
```

As one can see, under the given sensitivity requirements, we are also able to achieve a good degree of protection against false alarms; for example, the probability of a false alarm during a shift (50 lots) does not exceed 0.5%. Such good resolution was to be expected, since the width of the intermediate region, 0.4 (=0.6-0.2) is roughly equal to four standard deviations of an observation (which is in our case $\bar{X}$). However, if we lose control over the total variability (esp. its dominant contributor, lot-to-lot variability), the resolution power of the scheme may be severely damaged. For example, consider the situations corresponding to s.d. of of the sample means, $\sqrt{\sigma_b^2 + \sigma^2/12}$, accepting values 0.1, 0.15 and 0.2, when the grand mean μ is at the upper bound of the target region. For demonstration purpose, let us return to our original (non-shifted) data. Clearly, both signal levels remain unaffected, while the reference values become k^+ =0.4+4=4.4 and k^-=0.4-4=-3.6, i.e. as shown in Fig. 2.1.

```
'MEU SAMPLE' ASSIGN 4.2 1
'SIGMA' SETXPLR 0.1 0.15 0.2
2 SET 2 5 50 100
(0.175, 4.4, 0.175, (-3.6)) TWOXPLR 'DFXBAR'

Analysis of the two-sided Cusum scheme with parameters:
H+,K+ = 0.175 4.4  and H-,K- = 0.175 -3.6
The levels of discretization are D+,D- = 30 30
The observations are distributed as X-bar with  MEU=4.2  SAMPLE=1
The changing parameter name is SIGMA

  SIGMA  P(UP)     ARL     SDRL        2        5       50      100
   0.10  1.000  9435.8   9435.1  .99981   .99949   .99473  .98948
   0.15  1.000   129.8    129.1  .98628   .96365   .68004  .46167
   0.20   .999    27.9     27.2  .93504   .83795   .16182  .02603
```

By examining the "bad" levels of μ in a similar way, one can see that increase in s.d. also affects detection capability, though to a much lesser extent.[34] A much more serious is its effect on the escape rate. Indeed, if the total s.d. reaches 0.2 at the time the grand mean is at the level μ =4.6, one could expect 2.5% of the wires to fall out of specifications; taking into account that the number of wires on a chip is very large, one should not expect high yields on the chip level.[35]

Next let us try to design a scheme for controlling the lot-to-lot variability. Since the s.d. of $\bar{X}$ is primarily related to σ_b, the sequence $\{ d_i \}$ as defined in Sec. 2 is appropriate for this purpose. Unfortunately, the successive members of this sequence are correlated (positively), therefore, we have no choice but to use simulation in order to assess the properties of a scheme. Let us choose the reference value k=0.15 (i.e. in the middle between 0.1 and 0.2) and try to find h for which the off-target ARL is about 10, which would, hopefully, lead to a high probability of detecting the presence of excessive lot-to-lot variability within one shift. In accordance with formula of type (B.1), the off-target ARL can be roughly approximated to be $h/(0.2\text{-}0.15)$; therefore, we can start by examining the on-target performance of a scheme corresponding to the signal level h=0.5:

---

[34] This can be easily explained by the fact that under the off-target conditions the RL is determined primarily by the rate of drift of the Page's scheme towards the signal level, i.e. by the excess of the process level over the reference value; on the other hand, under the on-target conditions the RL is strongly affected by the tail properties of the underlying distribution and, consequently, by its s.d. and probabilities of large deviations (also see problems B.1 and B.4).

[35] It is not difficult to derive a formula for percent of defective chips under the model (B.3). As one may expect, the yield depends not just on the total variability, but also on its individual components; clearly, poor yields correspond to situations in which the within-wafer variability is a dominant component.

```
5 SET 1
6 SET -1
S←0.5 0.15 ONERUN (0 SELECT 20000 SIMNORM 0 0.101)
STATIST LENGTHS

Number of observations: 35    Mean: 561    Median: 348
Minimum: 65    Maximum: 1811    Range: 1746
Estimates of Stand. deviation: S=502    SD=498    SR=493
Regression slope: 12.4    Skewness: 1.02    Kurtosis: -0.24
```

Since only 35 out-of-control signals were observed during this run, the study needs to be repeated several times in order to obtain an assessment of better quality.[36] On the basis of the above output, we could expect the ARL to be around 600, and, by (B.2), the probability of no false alarm within a shift to be around $\exp(-50/600) = 0.92$. Analogously, we can examine the off-target performance of our scheme:

```
5 SET 1
6 SET -1
S←0.5 0.15  ONERUN  (0 SELECT 10000 SIMNORM 0 0.2005)
STATIST LENGTHS

Number of observations: 1015    Mean: 9.8    Median: 8
Minimum: 1    Maximum: 67    Range: 66
Estimates of Stand. deviation: S=7.6    SD=7.2    SR=7.6
Regression slope: -0.0007    Skewness: 1.92    Kurtosis: 6.07
```

This simulation study shows that the off-target ARL and SDRL are about 10 and 7.5, respectively. Since relatively many (1015) signals were observed, we can also obtain a good assessment of the RL distribution (one could use the function DFEMPIR for this purpose). In particular, the probabilities that the RL will exceed 20 and 50 were estimated to be 0.09 and 0.001, respectively.

The fact that the resolution of our scheme is much poorer than that of an $\bar{X}$ – scheme considered earlier may or may be not a matter of concern. Indeed, the practical (or economical) consequences of a false alarm produced by an $\bar{X}$ – chart may be completely different from those caused by

---

[36] When the number of out-of-control signals typically observed in a single simulated run is very small, neglecting the last part of the Cusum trajectory (for which no signal was triggered), or disregarding simulations that produced no signals, may lead to serious underestimation of the RL. If no headstarts are used, one could reduce the bias by adding the remainder to the first run length observed in the next simulation.

schemes monitoring lot-to-lot variability, within-lot variability or other parameters - since states of various parameters are frequently associated with different sets of tools, different operators, etc. The same, of course, can be said with respect to sensitivity.

Finally, let us consider the problem of monitoring the within-lot variability. We know that under normal conditions, σ should not exceed 0.05; therefore, any value greater than 0.05, say, 0.0501, could be considered as "bad". However, our sampling intensity does not enable us to resolve between levels so close one to another. Since the s.d. of $\hat{\sigma}$ is

$$s.d.(\hat{\sigma}) = \sigma\sqrt{1 - [c(12)]^{-2}} = \sigma(1 - 1.023^{-2}) = 0.21\sigma, \qquad (B.4)$$

one could expect to be able to detect relatively quickly an increase in σ by about 0.02. Thus, let us declare $\sigma \leq 0.06$ as a target region for σ, and $\sigma \geq 0.08$ as a "bad" region. Now we shall try to design a scheme for which the probability of a false alarm (when σ=0.06) within a shift (50 lots) does not exceed 0.01:

```
      12  0.06 0.08 CUSUMS 50 0.99

The observations are distributed as S with  SIGMA=0.06  SAMPLE=12
Step1: Search for H satisfying Prob.(R.L.>50)=0.99
Step1 complete; H=0.0616966, Prob.(R.L.>50)=0.990405
Analysis of upper Cusum scheme with parameters  H,K = 0.0617 0.0684
The level of discretization is 30
The changing parameter name is SIGMA

   SIGMA      ARL     SDRL
    0.06   4771.0   4766.4
    0.08      7.0      3.8

Enter the values of H, K, HEADST. and CD for further analysis (or 0 to exit):
      0.06 0.07 0 12
Enter additional values of SIGMA for which the analysis is to be performed:
      0.05
Enter the values of R for which Prob.(R.L. > R) is to be computed:
      2 5 50 100

Analysis of upper Cusum scheme with parameters  H,K = 0.06 0.07
The level of discretization is 30
The observations are distributed as S with  SAMPLE=12
The changing parameter name is SIGMA

   SIGMA      ARL     SDRL        2        5       50      100
    0.05 .941E09 .941E09 1.00000 1.00000 1.00000 1.00000
    0.06 10142.0 10138.0  .99999  .99982  .99542  .99052
    0.08      7.8      4.6  .96090  .63028  .00002  .00000

Enter the values of H, K, HEADST. and CD for further analysis (or 0 to exit):
0
```

The above output indicates that with the desired degree of protection against false alarms, we also have a basis for hope that should a significant increase in σ. take place somewhere during a shift, it will be detected by its end. Note that these conclusions depend on the validity of our assumption that wafer-to-wafer variability within a lot can be ignored. If it can't, one should base the control scheme for σ on the average of (three) sample standard deviations corresponding to different wafers of the lot. As mentioned in the end of Sec. 4.4., this is equivalent to reduction of the sample size from 12 to $1+3 \times (4-1)=10$.

A simulated sample run of the process is shown in Fig. B.3. The "true" parameters corresponding to the four successive sets of 20 lots are ($\mu = 4.0, \sigma_b = 0.1, \sigma = 0.05$), ($\mu = 4.0, \sigma_b = 0.2, \sigma = 0.05$), ($\mu = 4.0, \sigma_b = 0.1, \sigma = 0.08$) and ($\mu = 4.6, \sigma_b = 0.1, \sigma = 0.05$), respectively. One of the charts corresponds to the total variability, $\sigma_b^2 + \sigma^2$. It is based on the sequence of unbiased estimators $\{ t_i^2,\ i = 1, 2, \ldots\}$, where

$$t_i^2 = \frac{1}{2}(\bar{x}_{i+1} - \bar{x}_i)^2 + \frac{(n-1)(\hat{\sigma}_i^2 + \hat{\sigma}_{i+1}^2)}{2n}. \qquad (B.5)$$

We used this chart for the purpose of graphical data presentation only (i.e. we do not apply any control scheme to it).

As a final remark, we note that it is not difficult to evaluate the performance of a combined $\bar{X} - s$ - chart, since, under the normal assumptions, the sequences of observations these schemes are based upon are independent. For example, for $\mu = 4.2$ and $\sigma = 0.06$, the probability of no false alarm within 50 lots is $0.99473 \times 0.99542 = 0.99017$. There is no simple way to compute the ARL of the combined scheme exactly. However, in many cases the harmonic mean of the ARL's can serve as a good approximation; in our case it gives ARL $\approx [(1/9435.8) + (1/10142.0)]^{-1} \approx 4900$.

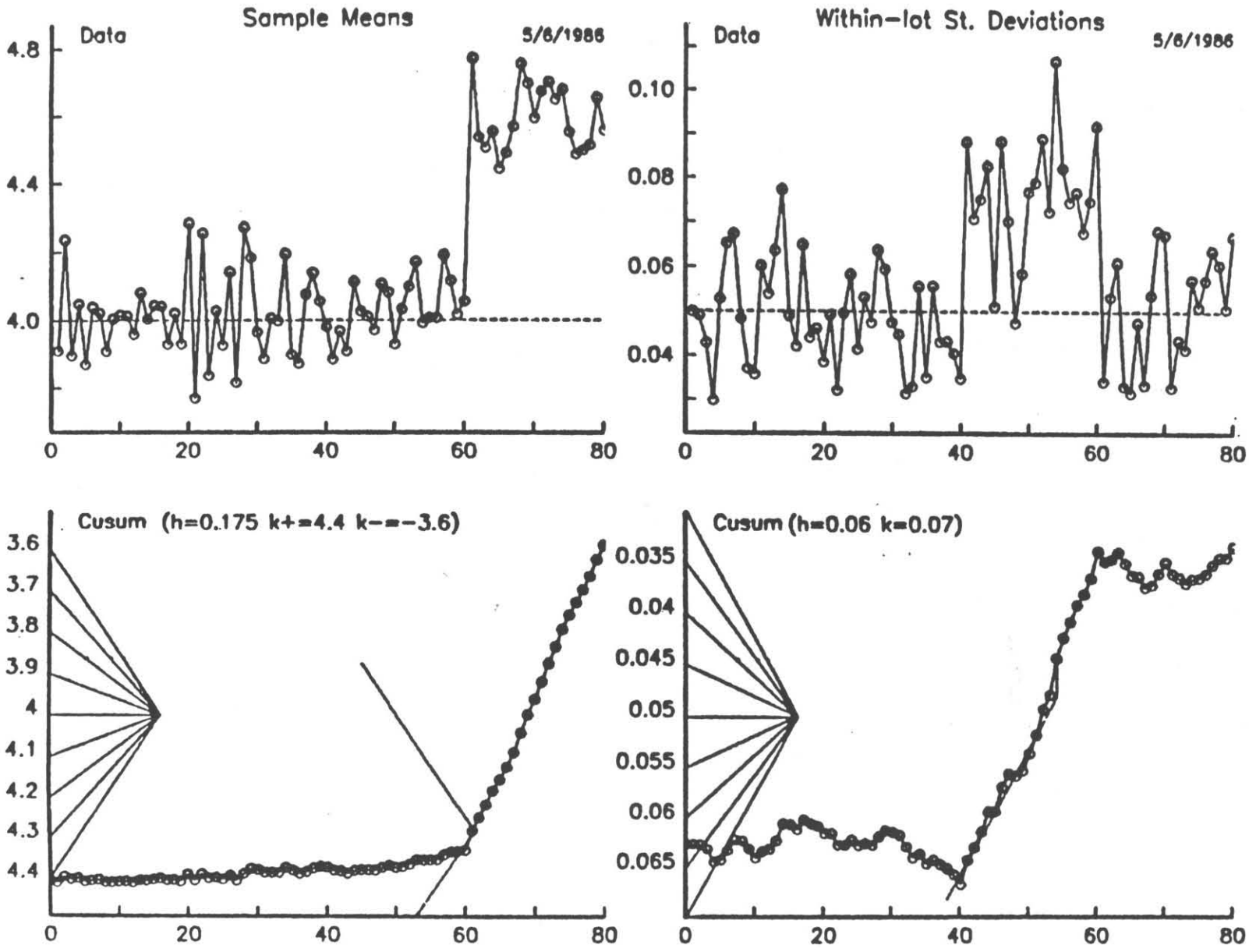

Fig. B.3. A sample run (80 lots) corresponding to Example B.7. Cusum charts in the bottom part of the figure correspond to observations in the top part. Filled plotting symbols correspond to out-of-control state of the process (in terms of the appropriate Cusum scheme); V - mask is shown at the first detected out-of-control observation.

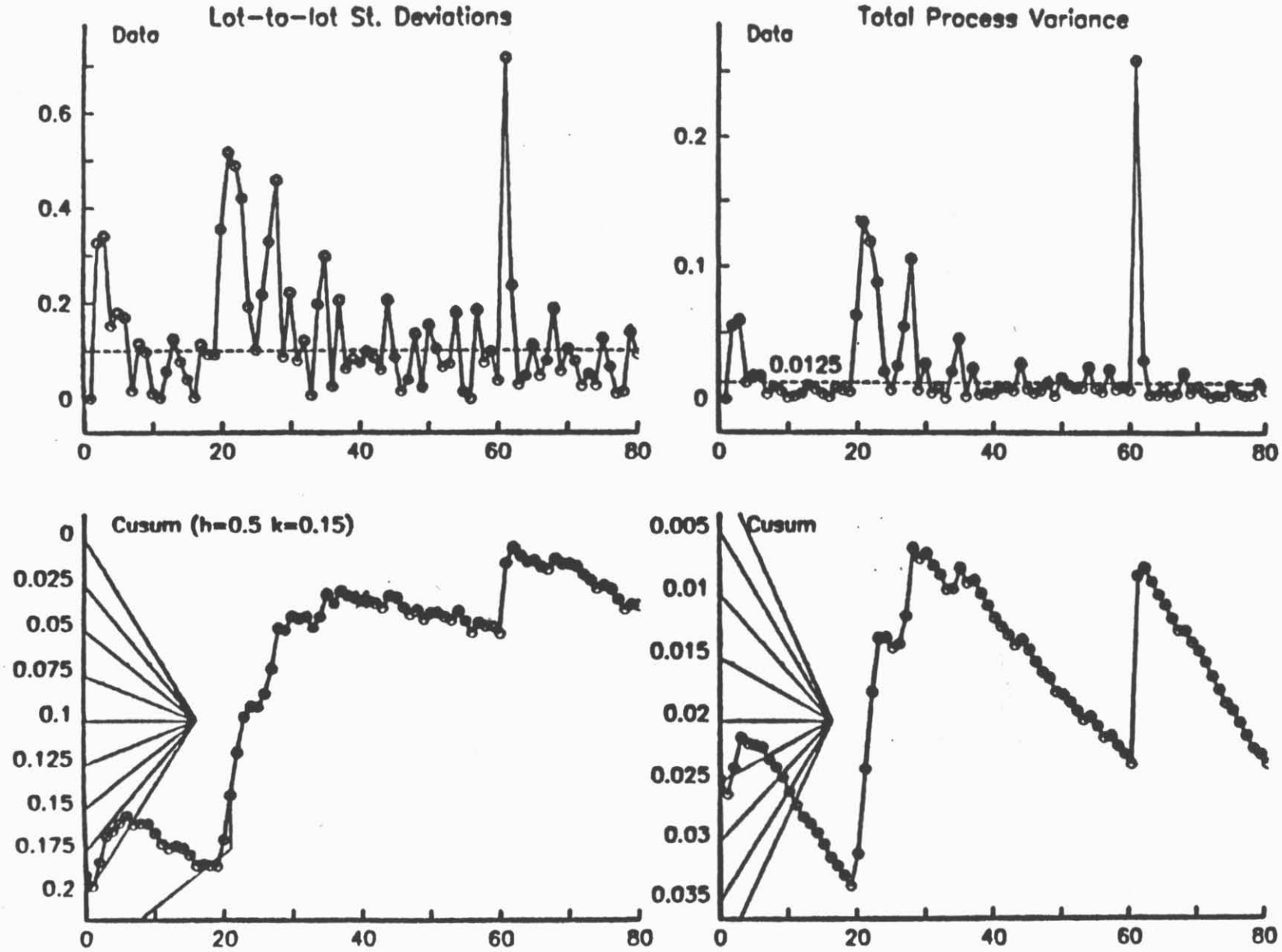

Fig. B.3. (cont) Charts for lot-to-lot variability and total variability. The Cusum scheme corresponding to lot-to-lot variability is re-initiated after the 40 -th lot.

Example B.8 (Control of the multivariate normal mean). Up to this point, we considered the problem of controlling a *single* process parameter. Even in cases where several parameters are controlled simultaneously, we designed and analyzed the individual control schemes as if the sequences of observations the control is based upon were independent. Indeed, there are many situations in which this can be reasonably assumed. For example, it is well known that under the normal assumption, the sequences of sample means and standard deviations are independent. This immediately implies that performance of a cumulative $\bar{X} - s$ - chart can be easily assessed on the basis of the individual components. As an additional example, consider once more the situation related to production of surface-mounted boards mentioned in the Introduction. If a typical out-of-control condition of the process is caused by a random clot or particle clogging one of the slots, the assumption that control charts corresponding to other slots remain unaffected is hardly unreasonable; therefore, assessment of the performance of an ensemble of charts can probably be based on analysis of individual ones and independence assumption.

However, in many cases the sequences different control schemes are based upon cannot be considered as independent. For example, if in the production process of ball bearings one tries to monitor simultaneously the diameters and weights of the balls, the dependency of the associated sequences follows merely from geometric considerations. In situations of this type, the individual control charts are not likely to produce the full resolution power in terms of the vector of parameters, since the underlying correlation structure is not taken into account. Moreover, the individual charts are usually able to detect, relatively quickly, changes of the multivariate mean in the directions of axes, but may have a poor sensitivity with respect to changes of similar magnitude along some other directions.

To improve the power, one can supplement the battery of individual charts by an additional one, intended to control some sort of "distance" between the current multivariate mean of the population, vector μ, and the centroid of the target region, μ_0. One of possible choices is the so-called Mahalanobis distance with respect to some positive definite matrix Σ:

$$\lambda = \sqrt{(\mu - \mu_0)^T \Sigma^{-1} (\mu - \mu_0)} \,. \qquad (B.6)$$

The matrix Σ determines the relative impact of deviation of μ from μ_0 along the different directions on the controlled parameter λ, i.e. the "cost structure". If, for example, we consider the domains $\lambda \leq 1$ and $\lambda \geq 2$ as "good" and "bad", respectively, the corresponding domains in terms of the population mean μ is as shown in Fig. B.4.

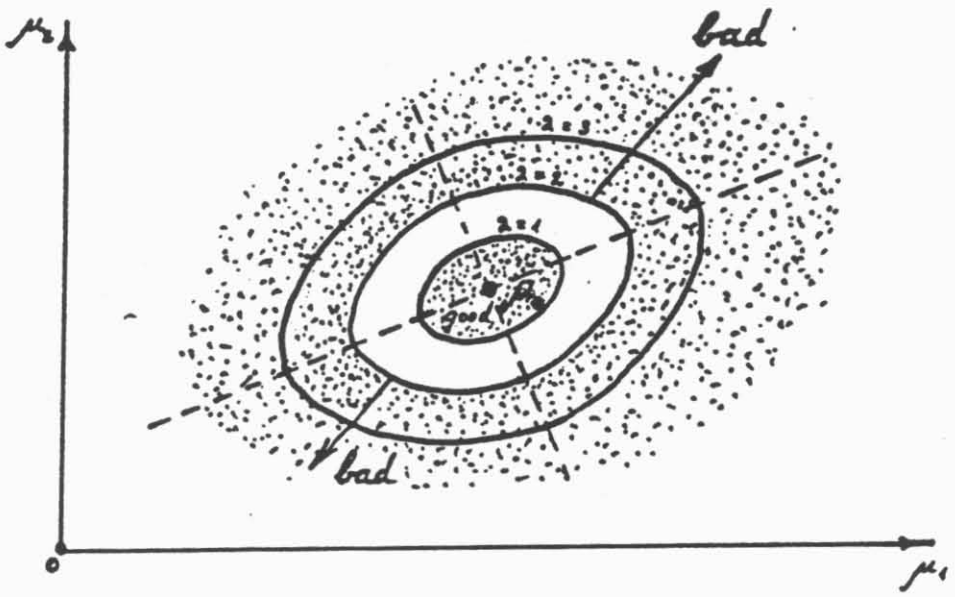

Fig. B.4. "Good" and "bad" domains for the bivariate mean.

It is natural to base the control procedure for λ upon the sequence of sample Mahalanobis distances $\{m_i\}$, defined by means of

$$m_i = \sqrt{(\bar{x}_i - \mu_0)^T \Sigma^{-1} (\bar{x}_i - \mu_0)} \ , \quad i = 1, 2, \ \ldots , \qquad (B.7)$$

where $\bar{x}_i$ is the i-th sample average of a group (sample) of n *multivariate* measurements.[36]

To design an appropriate scheme, we need to be able to compute the distribution of m_i for various multivariate distributions of interest. Unfortunately, this is usually difficult to do; even in cases where the readings follow a multivariate normal distribution, the computations are very involved. In any

[36] It is not difficult to see that when Σ is a diagonal matrix, the Mahalanobis distance turns into a "weighted" Euclidian distance between the point of interest and μ_0.

case, in the present version of the package we cannot suggest an efficient alternative to simulation when dealing with the general case.[38]

There is, however, one relatively simple case which frequently leads to satisfactory results. Namely, assume that the observations come from a p - dimensional normal population, and Σ used in (B.6) is nothing else but the underlying covariance matrix. Such dual use of Σ as both the population parameter and part of the "cost structure" determined by (B.6) may, in some cases, be arguable: indeed, the practical consequences of shifts in μ along different directions need not, in general, be related to the covariance structure. However, in many cases it is anyway very hard (or impossible) to specify Σ to be used in (B.6). Moreover, even if the opposite is true, one could still find that the gains in resolution power resulting from the use of a pre-specified Σ were not worth the more painful design procedure. So, let us consider the problem of controlling λ defined in terms of the underlying covariance matrix in more detail. Two cases will be considered:

a) Σ is known and generally behaves in a stable way.

In this case the distribution of m is closely related to non-central chi-square distribution and is present in the workspace under the name DFMAHAL. Before using this function, one should set its global parameters MAHAL, DEGREES and SAMPLE to λ, p and n, respectively. As an example, let us design a scheme corresponding to a bivariate normal distribution and sample size 5. Suppose that $\lambda \leq 0.5$ and $\lambda \geq 1.5$ are "good" and "bad" regions, respectively. To find an appropriate reference value, we must have some estimate of the central tendency (preferably mean or median) of the sample Mahalanobis distance, (B.7), corresponding to $\lambda = 0.5$ and $\lambda = 1.5$. Once these estimates are available, the reference value can be chosen somewhere in the midway. The simplest way to obtain them is by using simulation; in particular, the statement STATIST (1000 SIMMAHAL 0.5 2 5) estimates the median corresponding to $\lambda=0.5$ to be about 0.7. Similarly, the median corresponding to $\lambda=1.5$ can be estimated to be about 1.5.[39] Therefore, one may expect that a scheme with the ref-

---

[38] For normal populations simulation studies could be performed by using the function SIMMULT.

[39] As an alternative, one could use the function QUANTILE to evaluate the median. However, simulation provides some additional information of interest (ex. the mean).

erence value $k = (0.7+1.5)/2 = 1.1$ will provide about the best possible resolution. Now let us try to design a scheme for which the ARL corresponding to $\lambda=0.5$ is 1000:

```
'MAHAL' SETXPLR 0 0.5 1 1.5
0 SETFIND 1000
'MAHAL DEGREES SAMPLE' ASSIGN 0.5 2 5
2 SET 5 10
2 1.1 ONEXPLRD 'DFMAHAL'
```

```
The observ. represent Mahal. distance (Sigma known) with
MAHAL=0.5  DEGREES=2  SAMPLE=5
Step1: Search for H satisfying ARL=1000
Step1 complete; H=1.08741, ARL=942.8
Analysis of upper Cusum scheme with parameters  H,K = 1.08741 1.1
The level of discretization is 30
The changing parameter name is MAHAL

   MAHAL      ARL     SDRL        5       10
       0 59051.0 59049.7  .99993  .99984
     0.5   942.8   940.8  .99625  .99098
     1.0    13.6    11.1  .76718  .49094
     1.5     3.1     1.5  .06413  .00108
```

Now let us assume that our bivariate distribution corresponds to a pair of independent normal variables with $\sigma=1$, and the centroid of the target region is at the origin. Then, once the Euclidian distance between the process mean and the origin becomes 1.5 (in any direction), the above chart will trigger an out-of-control signal after about 3.1 samples. In particular, consider change of this magnitude in the north-east direction (the corresponding point is $\mu = (1.06, 1.06)$), and examine what would happen if we tried to control the bivariate mean by means of two univariate two-sided $\bar{X}$ - charts. For each chart the "good" region is $-0.5 \leq \mu \leq 0.5$, and the "bad" region is $|\mu| \geq 1.5$, suggesting the reference value of 1. Further, let us design a scheme for which the on-target ARL is 1000:

```
'MEU SIGMA SAMPLE' ASSIGN 0.5 1 5
'MEU' SETXPLR 0 0.5 1.06 1.5
0 SETFIND  1000
3 1 ONEXPLRD 'DFXBAR'
```

```
The observations are distributed as X-bar with
MEU=0.5  SIGMA=1  SAMPLE=5
Step1: Search for H satisfying ARL=1000
Step1 complete; H=1.05125, ARL=949.7
Analysis of upper Cusum scheme with parameters  H,K = 1.05125 1
The level of discretization is 30
The changing parameter name is MEU

     MEU      ARL     SDRL        5       10
       0 .345E06 .345E06  .99999  .99997
     0.5   949.7   948.0  .99596  .99073
    1.06     9.3     7.1  .63545  .31269
     1.5     2.8     1.4  .04432  .00053
```

(clearly, this also reflects the performance of a two-sided scheme, except the latter has a two times shorter ARL when $\mu = 0$). One can see that when the change in mean is along one of the axes, the combination of the univariate charts has a somewhat better sensitivity than a chart based on the Mahalanobis (in our case Euclidian) distance from the origin; however, it is much slower in detecting a change of similar magnitude along the north-east direction. Indeed, the probability that such a change will not be detected within 5 samples is ($0.63545^2 = 0.40$), while for the chart based on Euclidean distances this probability is only 0.06413. This example illustrates the point that if it is important to detect changes of similar magnitude along various non-axial directions, one cannot rely on univariate schemes only to do the job - he should use an additional chart for monitoring appropriately defined "distances" between the observations and μ_0.

b) Σ is unknown

In this case, we are still able to control λ, though we do not have a fixed matrix Σ to be used in (B.7). Instead of this matrix, it would be natural to use the sample covariance matrix, S_i,

$$S = \frac{1}{n-1}\sum_{j=1}^{n}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)\,(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)^{\mathsf{T}}, \qquad (B.8)$$

where $\mathbf{x}_{ij}$ is the j-th vector of the i-th sample. The distribution of the Mahalanobis distance defined in the described way depends on the underlying Σ only via the parameter we are trying to control, λ; in particular, $n \times m^2$ corresponds to Hotelling's T^2 - distribution (which represents a slightly modified version of the non-central F - distribution, see Anderson (1984, p. 163)).[40] The price we pay for not using the "true" covariance matrix is directly related to the fact that an inverse of a random matrix, used in (B.7) instead of Σ inflates the variability of the sequence $\{m_i\}$. Indeed, in the situation discussed earlier, it is not difficult to see that one needs a sample of 12 instead of 5 to

---

[40] The use of T^2 - statistics in Shewhart - type control charts is well known and documented (ex. see Jackson (1959) or Woodall and Ncube (1985); the latter work also contains an extensive list of references on this subject). In the present work, we illustrate the use of this statistics in a Cusum-Shewhart setting. As one could expect, this procedure is much more powerful than its Shewhart counterpart.

achieve the same resolution power, if he uses Mahalanobis distances with respect to an estimated covariance matrix:

```
0 SETFIND 1000
'MAHAL' SETXPLR 0 0.5 1 1.5
'MAHAL DEGREES SAMPLE' ASSIGN 0.5 2 12
2 1.1 ONEXPLRD 'DFHOTEL'

The observ. represent Mahal. distance (Sigma unknown) with
MAHAL=0.5  DEGREES=2  SAMPLE=12
Step1: Search for H satisfying ARL=1000
Step1 complete; H=1.39, ARL=987.3
Analysis of upper Cusum scheme with parameters  H,K = 1.39 1.1
The level of discretization is 30
The changing parameter name is MAHAL

   MAHAL      ARL     SDRL        5       10
       0 26860.9 26860.3  .99982  .99963
     0.5   987.3   986.2  .99551  .99047
     1.0    12.6     9.7  .76886  .46905
     1.5     3.0     1.3  .04631  .00018
```

Note that the function DFHOTEL used in the above run has exactly the same parameters as DFMAHAL we used earlier. Use of the same reference value ($k=1.1$) was suggested by a simulation run of the type described earlier with the only difference that the function SIMHOTEL was used instead of SIMMAHAL.

Appendix C. Discretization of Cusum - Shewhart schemes

For purposes of analysis of the run length distribution, we discretize the one-sided Page's schemes as shown in Fig.C.1.

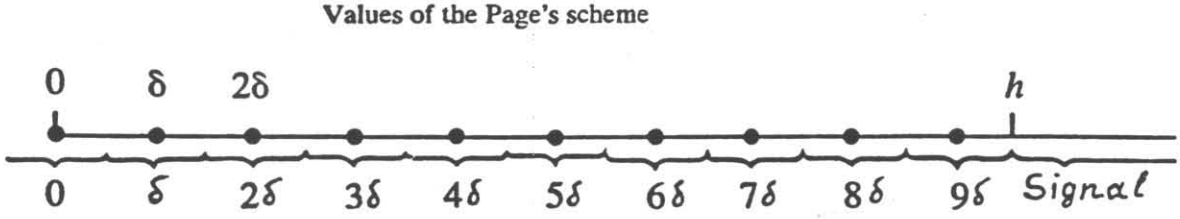

Fig. C.1. Discretization of the values of one-sided Cusum scheme.

In other words, the values of s_0, s_1, ... , will be rounded to the center of a corresponding group. The number of groups will be termed the *level of discretization* of the scheme and denoted by d; for example, in the case represented by Fig. C.1, the level of discretization is $d = 10$. The length of an interval corresponding to a single group, δ, will be called the *discretization interval;* is is always related to the level of discretization by means of the formula

$$\delta = h/(d - 0.5). \qquad (C.1)$$

Thus, the centers of the groups are at points $0, \delta, 2\delta, \ldots, (d-1)\delta$ and $h = (\delta/2)$ + (center of the last group). Such a method of discretization usually gives approximations of good quality and is recommended in many sources (ex. Brook and Evans (1972)). It is clear that by using high levels of discretization, we can approximate the characteristics of the Run Length distribution to any degree of accuracy. But how high is high? We performed extensive studies which indicate that levels of discretization of order $d \approx 30$ are satisfactory for most practical purposes. The reason for that is related to the fact that we discretize *the states* of the Page's schemes but *not* the observations themselves. Thus, relatively low sensitivity with respect to level of discretization is explained by compensation of roundoff errors when computing subsequent values of the scheme. As an example,

let us apply the scheme ($h = 3,\ k = 1,\ s_0 = 0$) to five sequences of normal observations corresponding to $\mu = 0,\ 0.5$ and 1 and $\sigma = 1$. Table C.1 contains the values of ARL as well as lower and upper 5% quantiles of the run length distribution (in parentheses) corresponding to levels of discretization ranging from 10 to 100. It indicates that levels of discretization as low as 10 enable one to roughly assess the properties of the run length distribution. Moreover, discretization does not represent something unnatural, since it is automatically assumed in any procedure involving roundoff to a certain number of digits after the decimal point.

d	$\mu =$ 0	0.5	1
10	1918(100, 5741)	117(8, 343)	17(3, 45)
30	1958(102, 5860)	117(9, 345)	17(3, 45)
100	1962(102, 5873)	118(9, 345)	17(3, 45)

Table C.1. Effect of the level of discretization on ARL and 5% quantiles (in parenthesis) corresponding to the scheme ($h = 3,\ k = 1,\ s_0 = 0$) The observations are iid normal with $\sigma = 1$. The entries are rounded to the nearest integer.

The case in which the observations $x_1,\ x_2,\ \ldots$ are integers is of special interest because of its relevance to the problems of controlling the process proportion of defectives (cumulative p-charts), the number of defects per produced unit (cumulative c-charts), etc. In this case a proper choice of the interval (or level) of discretization and scheme parameters can eliminate the roundoff error altogether. Indeed, let us pick the interval of discretization $\delta = 0.1$ and require that the reference value be some multiple of δ and the signal level be chosen in accordance with (C.1). Then the Page's scheme becomes "naturally" discretized and can be analyzed exactly. Of course, we are limited in our choice of the reference values; however in most practical situations it is not a serious limitation. Moreover, if needed, we can always take a shorter discretization interval and have additional possibilities for the choice of k . The price for doing that is related to an increase in the level of discretization, which in turn determines the size of the Markov transition matrix used in the analysis. In general, we would not recommend the user of CONTRD to use levels of discretization above 100 on a regular basis - it will just waste CPU time.

Appendix D. Steady state analysis of Cusum - Shewhart schemes

When analyzing the behaviour of a control scheme with respect to out-of-control state of the process generating observations, we usually assume that the deviation of the process from on-target conditions occur at time $i = 0$; this is also tacitly assumed when we talk about the Run Length. Under the above assumption, we are primarily concerned about the speed of detecting the presence of out-of-control conditions by using various control schemes.

Clearly, this approach may lead to a pessimistic estimate of detection speed in the "real life" situations. Indeed, we typically assume that the scheme is at 0 (worst case) at the onset of out-of-control conditions. However, in practice there is a possibility that at this moment the scheme has some "natural" headstart and, therefore, will signal earlier. Thus, one may be interested in analysis of the *Residual Run Length* corresponding to an assumption that deviations of the process from the target conditions occur after a substantial period of time, during which the process operated in on-target mode characterized by some distribution function of the observations, say F.

Steady state analysis is related to behaviour of Residual Run Length. It starts with an assumption that process is in control (i.e. the observations come from the distribution F), and then computes the probabilities of various values of the (discretized) control scheme after a "very long" period of time *given* that no out-of-control signals were *not* triggered during this period of time. The resulting steady state distribution (which is sometimes called "quasi - stationary" distribution in the literature) provides the weighting factors that are applied to a set of ARL's (or other run length characteristics) corresponding to appropriate headstarts.

It is clear that questions related to behaviour of the Residual Run Length remain of interest also in cases where headstarts are used; in fact, it is not difficult to see that results of the steady state analysis do not depend on the headstart(s). In CONTRD this type of analysis is currently available for one-sided schemes only; it is invoked by specifying a negative headstart.

More information about the steady state analysis can be found in Yashchin (1984 and 1985a).

Appendix E. List of functions for generating random variables

Every function for simulating a set of (independent and identically distributed) random variables has a left argument, L, representing the quantity of variables to be generated and a right argument, R, characterizing the parameters of the distribution. For example 20 SIMBINOM 50 0.3 will generate 20 binomial random variables with parameters n=50 p=0.3. Other possibilities are as follows:

Function Name	Parameters	Comment
SIMARMA	R[1] = μ R[2] = σ R[3] = p R[4] = q R[5, ... 4+p] = τ R[5+p, ... 4+p+q] = θ	Autoregressive Moving Average (ARMA) with parameters (p, q), level μ and st. dev. σ. τ and θ are the coefficients of the AR and MA parts, respectively. ARIMA process can also be generated (the second left component L[2], if present, corresponds to parameter d).
SIMBETA	R[1] = α R[2] = β	Beta with parameters α and β
SIMBINOM	R[1] = n R[2] = p	Binomial with parameters n and p
SIMGAMA	R[1] = α R[2] = β	Gamma with parameters α and β
SIMGEOM	R[1] = p	Geometric with parameter p. The mean of this variable is p^{-1} and its possible values are 1,2,...
SIMLOGN	R[1] = μ R [2] = σ	Lognormal with parameters μ and σ
SIMMULT	20 SIMMULT 'MNS COVAR' will generate 20 multivariate normal vectors (rows). Vector MNS and matrix COVAR must be set to contain the set of means and covariance matrix, respectively.	
SIMNCHI	R[1] = d R[2] = λ	Non-central chi-square with d degrees of freedom and non-centrality parameter $\lambda \geq 0$
SIMNEGB	R[1] = k R[2] = p	Negative binomial (sum of k geometric random variables with mean p^{-1}. The possible values of the variable are k, $k+1$, ...
SIMNORM	20 SIMNORM 0.1 2 will generate 20 normal variables with mean 0.1 and stand. deviation 2. 20 4 SIMNORM 0.1 0.5 2 will generate a matrix having 20 rows. Each row represents a sample of 4 normal variables with mean 0.1+0.5 Y (Y is an independent standard normal variable) and standard deviation 2. Thus, 0.1 is the grand mean, 0.5 is the between-rows standard deviation and 2 is the within-row standard deviation	

SIMHYPGEOM	R[1] = N R[2] = k R[3] = n	Hypergeometric with lot size N, number of defectives in the lot k and the sample size n
SIMPOIS	R[1] = λ	Poisson with mean λ
SIMPROP	R[1] = λ R[2] = p	Proportion of defectives corresponding to to random (Poisson with mean λ) sample size and probability of a defective unit p
SIMUNIF	R[1] = Lower bound R[2] = Upper bound	Uniform distribution
SIMWEIB	R[1] = Shape R[2] = Scale	Weibull random variable; if Shape=1, the the simulated variable is exponential with mean Scale.

The last two functions compute the d.f. of the Mahalanobis distance between the mulrivariate normal sample mean and some fixed point (ex. centroid of the target region), μ_0. The parameter λ represents the Mahalanobis distance (with respect to the covariance matrix) between the population mean and μ_0 (see (B.6)). p is the dimension of the multivariate observation and n is the sample size.

SIMMAHAL	R[1] = λ R[2] = p R[3] = n	Distribution of the Mahalanobis distance with respect to a "true" covariance matrix.
SIMHOTEL	R[1] = λ R[2] = p R[3] = n	Distribution of the Mahalanobis distance with respect to an esimated covariance matrix, S. Squared variables multiplied by n corresponds to a Hotelling's T^2 - distribution.

References:

1. Anderson, T. W. (1984), *An Introduction to Multivariate Statistical Analysis,* Wiley, New York.

2. Bagshaw, M. and Johnson, R.A. (1975, a), *The Effect of Serial Correlation on the Performance of CUSUM TESTS II,* Technometrics, 17, pp. 73-80.

3. Bissell, A. (1969), *Cusum Techniques for Quality Control,* Appl. Stat., 18, pp. 1-30.

4. British Standards Institution (1980-1983), *Guide to Data Analysis and Quality Control Using Cusum Techniques, Parts 1 - 4* .

5. Brook, D. and Evans, D.A. (1972), *An Approach to the Probability Distribution of Cusum Run Length,* Biometrika, 59, pp. 539-549.

6. van Dobben de Bruyn, D.S. (1968), *Cumulative Sum Tests Theory and Practice,* Hafner Publishing Co., N.Y.

7. Duncan, A. J. (1974), *Quality Control and Industrial Statistics,* Irwin, IL.

8. Efron, B. (1981), *The Jackknife, the Bootstrap and Other Resampling Plans,* CBMS-NSF Regional Conference Series in Applied Mathematics, Vol.39, Philadelphia, Pennsylvania.

9. Jackson, J.E. (1959), *Quality Control Methods for Several Related Variables,* Technometrics, 1, pp. 359-377.

10. Johnson, R.A. and Bagshaw, M. (1974), *The Effect of Serial Correlation on the Performance of CUSUM Tests,* Technometrics, 16, pp. 103-122.

11. Khan, R. (1984), *On Cumulative Sum Procedure and the SPRT with Applications,* J. Royal Stat. Soc. (B), 46, pp. 79-85.

12. Lorden, G. (1971), *Procedures for Reacting to a Change in Distribution,* Ann. Math. Stat., 42, pp. 1897-1908.

13. Lucas, J. M. (1982), *Combined Shewhart-Cusum Quality Control Schemes,* J. Qual. Technol., 14 (2), pp. 51-59.

14. Lucas, J. M. and Crosier, R. B. (1982), *Fast Initial Response for Cusum Quality Control Schemes: Give your Cusum a Head Start,* Technometrics, 24, pp. 199-205.

15. Lucas, J. M. (1985), *Counted Data Cusum's,* Technometrics, 27, pp. 129-144.

16. Page, E. (1954), *Continuous Inspection Schemes,* Biometrika, 41, pp. 100-115.

17. Walpole, R. E. and Myers, R. H. (1978) *Probability and Statistics for Engineers and Scientists,* Macmillan, New York.

18. Woodall, W. H. (1985a), *The Statistical Design of Quality Control Charts,* The Statistician, 34, pp. 155-160.

19. Woodall, W. H. (1985b), *The Design of Cusum Quality Control Charts,* Submitted for publication.

20. Woodall, W. H. and Ncube, M. N. (1985c), *Multivariate Cusum Quality Control Procedures,* Technometrics, 27, pp. 285-292.

21. Woodward, R. and Goldsmith, P.L. (1964) *Cumulative Sum Techniques,* ICI Monograph No. 3, Oliver and Boyd, London.

22. Yashchin, E. (1984), *Design, Analysis and Running of Cusum-Shewhart Control Schemes,* IBM Research Report RC 10869, 56 pages.

23. Yashchin, E. (1985a), *On the analysis and design of Cusum-Shewhart control schemes,* IBM Journ. Res. Devel., 29, pp. 377-391.

24. Yashchin, E. (1985b), *On a Unified Approach to the Analysis Of Two-sided Cumulative Sum Schemes With Headstarts,* Adv. Appl. Prob., 17, pp. 562-593.