Control charts for high-quality processes: *MAX* or *CUMAX*?

Willem Albers

Department of Applied Mathematics

University of Twente

P.O. Box 217, 7500 AE Enschede

The Netherlands

Abstract. For attribute data with (very) small failure rates control charts were introduced which are based on subsequent groups of r failure times, for some $r \geq 1$. Within this family, it was shown to be attractive to stop once the maximum of such a group is sufficiently small, because this choice allows a very satisfactory nonparametric adaptation. The question we address here is whether a cumulative approach offers even further improvement. Thus instead of fixed groups, we shall use the first sequence of r consecutive sufficiently small failure times to produce a signal. A further reason for considering this type of chart is the fact that it forms the nonparametric counterpart of the well-known sets method.

Keywords and phrases: Statistical Process Control, health care monitoring, geometric charts, average run length, estimated parameters, order statistics, sets method

2000 Mathematics Subject Classification: 62P10, 62C05, 62G15

1 Introduction and motivation

By virtue of ever increasing standards, high-quality processes are more and more common in industrial settings. Moreover, for health care monitoring they even form the standard, as in this area failures probabilities are tiny by nature: events such as malfunctioning equipment, surgical errors, recurrence of cancer and birth defects, should be avoided as much as possible. In both fields, application of control charts to improve and maintain quality is strongly advocated (see e.g. Sonesson and Bock (2003), Thor et al. (2007) and Shaha (1995) for some health care monitoring review papers). In view of the really small failure probabilities p involved it is quite common to use charts based on the geometrically distributed waiting times from one failure till the next. A group of size r ($r \geq 1$) of such waiting times is inspected and an alarm is raised if their sum (or alternatively, their maximum) falls below some boundary value. Of course, a two-sided version can also be used, but in practice the focus with respect to going out of control (*OoC*) typically is on increases of p , and thus on shorter waiting times than during in control (*IC*). The boundary value is determined such that the alarm rate during *IC* (the so-called false alarm rate (*FAR*)), remains sufficiently small (e.g. 0.001). Guidelines are available on how to optimally choose r in relation to the underlying parameters. See Albers (2011) for a description and further references.

Nevertheless, as discussed in Albers (2011) as well, a major remaining problem to be dealt with concerns the estimation issues involved. Usually such problems are conveniently ignored in practice. The typically unknown p is estimated using a so-called Phase I sample and the result $\hat{p}$ is simply plugged in. However, as the *FAR* is (very) small, the relative errors due to this estimation step are far from negligible for practical sample sizes and further analysis and

corrections are needed. Moreover, the problem may not stop at estimating a single parameter p . Health care applications often involve patients showing considerable heterogeneity, leading to overdispersion of the formerly geometric waiting times. A rigorous way to solve the resulting distributional estimation problems of course is to adopt a nonparametric approach. However, for sample sizes common in practice, the aforementioned relative estimation errors may now become huge: if e.g. the FAR should equal 0.001, a customary sample size like 100 is not very useful towards estimating the corresponding 0.001-quantile.

Consequently, further adaptation is required. The usual form of the chart is based on the sum of the r waiting times for each group, as in the homogeneous case this statistic is negative binomially distributed and moreover clearly optimal. However, as already mentioned above, an alternative is to adopt the maximum, rather than the sum, of the waiting times as our statistic. A simple example (roughly) illustrates the advantage: for $r = 3$, the probability that all waiting times in a group fall below their 0.1-quantile is $(0.1)^3 = 0.001$. But, unlike a 0.001-quantile, estimating a 0.1-quantile based on a sample of moderate size is quite feasible. Of course, solving the estimation problems in this way only makes sense if the step from sum to maximum merely causes a small loss of detection power when the assumption of a homogeneous case happens to hold after all. Fortunately, in Albers (2011) it is demonstrated that this is indeed true. Hence such a loss can be viewed as a small insurance premium to be paid for safeguarding against the risk of making substantial errors with the basic negative binomial chart once the ideal of homogeneity fails. Having settled this issue, the estimation aspects of the proposed nonparametric *MAX*-chart are subsequently dealt with in Albers (2011), resulting in a straightforward empirical version which is easy to understand and to apply.

The topic of the present paper now is the question whether there is still room for further improvement. The motivation for raising it is twofold. In the first place, essentially the same question received a positive answer in the corresponding continuous (and typically normal) case of controlling the mean of a process (see Albers and Kallenberg (2009)). Here the focus usually is on detecting increases in such a mean during *OoC*, implying that a signal typically should result if the minimum of a group of size r is too large. But obviously the analysis for a *MIN*-chart is essentially identical to that for its *MAX*-counterpart, so we can ignore this difference between the two situations. In Albers and Kallenberg (2009) it is suggested to replace the fixed-group approach by a sequential or cumulative version: produce a signal as soon as r consecutive observations all exceed a suitable boundary value. The resulting *CUMIN*-chart definitely looks better than the more rigid *MIN*-chart. It is an accelerated version, starting anew as soon as an observation falls below the boundary value and it thus makes no sense to complete the full group of size r .

However, note that matters are in fact a bit more subtle. For the *MIN*-chart, the average run length (*ARL*) during *IC* equals r/FAR . Making a fair comparison now requires that the *CUMIN*-chart matches this *ARL*-value during *IC*. But, as the runs of failed attempts in this latter chart are mostly shorter than r , a lower value of FAR is needed here to actually achieve this. In other words, the *CUMIN*-chart has to be more strict in the sense that it employs a higher boundary value to be exceeded. In view of this, it is not at all trivial anymore that it should be the winner. Nevertheless, in Albers and Kallenberg (2009) it is demonstrated that for the basic situation of normal underlying distributions, this typically indeed is the case, both empirically (Figure 1 and Table 1) and theoretically (asymptotic results in Lemma's 3.1 and 3.2). Consequently, it seems worthwhile to investigate whether this state of affairs also holds true for attribute rather than continuous data, i.e. when dealing with geometric rather than normal distributions. In other words, is it true that *CUMAX* beats *MAX*?

The second reason for posing the question is quite straightforward. As already remarked in Albers (2011), in the present attribute data setting the cumulative version of the *MAX*-chart is nothing but the well-known sets method, introduced by Chen (1978): a signal results once

all of r successive waiting times are too small. But, just as in Albers (2011), the focus here will be on showing how this type of approach can form the basis for a satisfactory nonparametric procedure, thus adequately solving the aforementioned serious underlying estimation issues. Even if these problems are not ignored in practice, usually at best the effect of estimating a single parameter is studied. Typically, the latter already turns out to be substantial; to give an example in the present context, Chen et al. (1997) mention a 30–90% increase in FAR for a 10 per cent bias in $\hat{p}$. Nevertheless, as the *CUMAX*-chart actually is the more prominent of the two proposals, it certainly makes sense to figure out whether it is an adequate competitor to (or even an improvement over) the less well-known *MAX*-chart from Albers (2011). In that case, its empirical nonparametric version, introduced here along the lines of Albers and Kallenberg (2009) and Albers (2011), offers an attractive robust alternative to many existing methods which rely with often unfounded optimism on negligibility of both stochastic and model errors.

In section 2 the charts are introduced and compared for the basic homogeneous case, i.e. where the underlying distributions are simply geometric. The comparisons of detection power are based on the commonly used form of *ARL*, which conveniently assumes that the process goes *OoC* prior to the onset of monitoring, or just as monitoring begins. Several authors (e.g. see Sego et al. (2008)) have pointed out that this is somewhat artificial. Indeed, in Albers and Kallenberg (2009) it was observed in this connection that the shift will rarely coincide precisely with the start of a new group and thus that the impact of going *OoC* will probably be delayed till the next group starts. For the *MIN*-chart with its groups of fixed size r , this effect will be more pronounced than for the more quickly reacting *CUMIN*-chart. In the present context this consequently translates into an (additional?) advantage of *CUMAX* over *MAX*. This effect, and more generally the impact of using a different form of *ARL*, will be studied in section 3. Having settled this issue, the empirical nonparametric version of the *CUMAX*-chart is the topic of section 4. For convenience, the conclusions reached, as well as a summary of the resulting procedure, are presented in section 5.

2 The homogeneous case

As explained in the introduction, the nonparametric chart, which is our ultimate goal, has to satisfy two requirements. First of all, the estimation effects should be manageable (i.e. not too large and, if desired, allowing suitable corrections). On the other hand, the price involved for this protection should be reasonable, in the sense that its corresponding counterpart in the ideal setting of a known, and even simply geometric, underlying distribution should have good detection power. Consequently, the homogeneous case will always be the yardstick from which we start.

Let $D_1, D_2, \dots$, be a sequence of independent identically distributed random variables (r.v.'s), with failure probability $P(D_1 = 1) = 1 - P(D_1 = 0) = p$ during *IC*. During *OoC*, this p becomes θp , for some $\theta > 1$, and the process should be stopped as quickly as possible. Based on this sequence of D 's, we consider geometric r.v.'s $X_i, i = 1, 2, \dots$, defined as the times from the $(i - 1)^{th}$ up to and including the i^{th} failure. Clearly,

$$P(X_i = k) = p(1 - p)^{k-1}, \quad (2.1)$$

where $k = 1, 2, \dots$. For some $r \geq 1$, the *MAX*-chart from Albers (2011) now gives a signal if $\max(X_1, \dots, X_r)$ is too small; otherwise, the next group of size r is considered, and so on. To calibrate matters, as a criterion for selecting the boundary value n it is proposed to use

$$FAR = P(\max(X_1, \dots, X_r) \leq n) = r\alpha. \quad (2.2)$$

for some small $\alpha > 0$. Then the ARL during IC has the same value $r/(r\alpha) = 1/\alpha$ for all r , allowing fair comparison. Since $P(X_1 \leq n) = 1 - (1 - p)^n$, it is immediate from (2.2) that

$$n = \frac{\log(1 - \{r\alpha\}^{1/r})}{\log(1 - p)}. \quad (2.3)$$

(Either use standard interpolation in (2.3) or let n be the largest integer such that $FAR \leq r\alpha$; in practice the differences involved will be negligible.)

During OoC the alarm rate changes into $\{1 - (1 - \theta p)^n\}^r$. In view of (2.3) this result then readily leads to the conclusion that in terms of detection power we have

$$ARL = ARL_{r,\theta} = \frac{r}{\{1 - (1 - \{r\alpha\}^{1/r})^{\frac{\log(1-\theta p)}{\log(1-p)}}\}^r}, \quad (2.4)$$

which for the small p we are interested in to high precision equals $r/\{1 - (1 - \{r\alpha\}^{1/r})^\theta\}^r$ (if desired, see the Appendix in Albers (2011) for details on this approximation step). Consequently, the dependence of this ARL on the actual p is negligible; hence the notation $ARL_{r,\theta}$ in (2.4). To analyze the behavior of $ARL_{r,\theta}$, note to begin with that it decreases from the prescribed value $1/\alpha$ at $\theta = 1$ to the lower limit r as θ becomes very large. In particular, $ARL_{1,\theta} \approx 1/\{1 - (1 - \alpha)^\theta\} \approx 1/(\theta\alpha)$, showing that the simple geometric chart performs rather poorly, unless θ is quite large. This suggests using larger r , the more so if the excess of θ over 1 is supposed to be smaller. In fact, a simple rule of thumb can be derived for r^{opt} , the value of r which approximately minimizes ARL for given α and θ (see Albers (2011) for details, among others Lemma 3.1):

$$r^{opt} = \frac{1}{\alpha(2.6\theta + 2) + 0.01(4\theta - 3)}. \quad (2.5)$$

Usually a truncated version like $\min(5, r^{opt})$ will suffice: most of the attainable improvement over $r = 1$ has already been achieved at $r = 5$ and using really large r may feel awkward in practice.

Next we use the set-up described above to introduce the *CUMAX*-chart (which, as mentioned in the Introduction, goes back to Chen (1978)). Now the idea is to let the process stop as soon as r consecutive X_i are all at most $\tilde{n}$, for some suitable boundary value $\tilde{n}$. As already argued in the Introduction, we expect $\tilde{n}$ to be smaller than n . Here we shall make matters precise. We have the following result:

Lemma 2.1 *Let $\tilde{n} = \log(1 - h^{-1}(\alpha))/\log(1 - p)$, in which h^{-1} is the inverse of $h(x) = (1 - x)x^r/(1 - x^r)$. Then the *CUMAX*-chart satisfies*

$$ARL_{r,\theta} = \frac{1}{h(1 - \{1 - h^{-1}(\alpha)\}^{\frac{\log(1-\theta p)}{\log(1-p)}})} \approx \frac{\frac{1}{[1 - (1 - h^{-1}(\alpha))^\theta]^r} - 1}{(1 - h^{-1}(\alpha))^\theta}, \quad (2.6)$$

with in particular $ARL_{r,1} = 1/\alpha$. Moreover, $h^{-1}(\alpha) \approx \{\alpha/(1 - \alpha^{1/r})\}^{1/r}$ and $\tilde{n} < n$ from (2.3).

Proof. It is well-known (see e.g. Kenett and Pollak (1983) or Albers and Kallenberg (2009)) that $ARL_{r,\theta} = 1/h(P_\theta(X_1 \leq \tilde{n}))$. As fairness dictates that the *CUMAX*-charts should satisfy $ARL = 1/\alpha$ as well during IC , it follows that $\tilde{n}$ should be selected such that $h^{-1}(\alpha) = P_1(X_1 \leq \tilde{n}) = 1 - (1 - p)^{\tilde{n}}$. Hence indeed $\tilde{n} = \log(1 - h^{-1}(\alpha))/\log(1 - p)$. From the definition of h and the fact that α is very small, it immediately follows that $h^{-1}(\alpha) \approx \{\alpha/(1 - \alpha^{1/r})\}^{1/r}$,

which is considerably smaller than $(r\alpha)^{1/r}$. Comparison of $\tilde{n}$ to (2.3) then shows that indeed $\tilde{n} < n$. It remains to note that $P_\theta(X_1 \leq \tilde{n}) = 1 - (1 - \theta p)^{\tilde{n}} = 1 - (1 - h^{-1}(\alpha))^{\log(1-\theta p)/\log(1-p)} \approx 1 - (1 - h^{-1}(\alpha))^\theta$ and to observe that $1/h(x) = \{1/x^r - 1\}/(1 - x)$ in order to arrive at (2.6). $\square$

The result from (2.6) now enables us to compare the performance of the *CUMAX*-chart to that of the *MAX*-chart as represented by (2.4). To avoid confusion, let us from this point on denote the *ARL* from (2.6) by *ARLCUM* and that from (2.4) by *ARLM*. The expressions involved - especially the one from (2.6), even in its approximate version - are too complicated to allow an exact analysis and hence we resort to a numerical study. The surprising conclusion is:

$$\text{'typically } \textit{ARLCUM} \text{ is somewhat larger than } \textit{ARLM}! \text{'}$$
 (2.7)

The surprise is twofold: as stated in the Introduction, crude intuition tells that *CUMAX* is definitely better. But even having noted that matters are more subtle, the fact that in the continuous case of controlling a normal mean *CUMIN* does beat *MIN* (see Albers and Kallenberg (2009)), suggests that the same will happen with *CUMAX* versus *MAX*. However, for the range of parameter values of interest, this is not the case.

As concerns this range, we typically are interested in values of r between 1 and 5 and values of α between 0.001 and 0.01, while p only needs to be small, e.g. at most 0.01, and plays no role otherwise (if desired, see Albers (2011) and its references for further discussion on these choices). If for such r and α we look at *ARLCUM*/*ARLM* as a function of θ , this ratio increases from its starting value 1 at $\theta = 1$ to a maximum slightly larger than 1 at a certain θ_{mr} , after which it again decreases towards 1. For the difference *ARLCUM* - *ARLM* the pattern is similar: rise from 0 at $\theta = 1$ to a maximum of a few observations at a certain θ_{md} , followed by decrease to 0. Obviously, as both *ARLCUM* and *ARLM* start at $1/\alpha$ and are decreasing, we have that $\theta_{mr} > \theta_{md}$. To add some explicitness to this qualitative characterization, we present a number of representative values in Table 2.1.

Table 2.1. Maximal values w.r.t. θ for ratios and differences of the *ARL*'s from (2.6) and (2.4) for various r and α . Given in each cell are for (i): $\arg\max \theta_{mr}$, together with corresponding values of the maximum of *ARLCUM*/*ARLM* and of *ARLM*; likewise for (ii): $\arg\max \theta_{md}$ and *ARLCUM* - *ARLM* and *ARLM* at that point.

(i) <i>ARLCUM</i> / <i>ARLM</i>									
$r \backslash \alpha$	0.001			0.005			0.01		
2	26	1.16	4.1	12	1.15	3.9	9	1.14	3.6
3	10	1.17	6.1	6	1.16	5.5	4.5	1.14	5.6
4	6	1.16	8.6	4	1.14	7.7	3.5	1.13	5.8
5	5	1.14	9.4	3.5	1.13	8.6	3	1.12	8.1
(ii) <i>ARLCUM</i> - <i>ARLM</i>									
$r \backslash \alpha$	0.001			0.005			0.01		
2	2.0	4.4	261	2.0	1.9	55.4	2.0	1.3	29.0
3	1.5	6.5	332	1.6	2.5	62.0	1.6	1.7	33.2
4	1.5	6.8	259	1.6	2.6	50.7	1.5	1.8	33.2
5	1.4	6.2	274	1.4	2.6	65.5	1.5	1.8	30.3

From Table 2.1 we infer for the ratio that both the maximum and its location θ_{mr} tend to decrease in r as well as in α . The maximal values obtained are rather stable and indicate that *ARLCUM* uses at most a fraction $1/6$ more observations than *ARLM*. As this maximum is reached at small values of *ARLM*, the absolute increase is only about $\frac{1}{2}$ to $1\frac{1}{2}$ observations. In fact, the second part of the table provides the maximal differences, which can be a bit larger, but nevertheless stay quite limited. Indeed θ_{md} is quite a bit smaller than θ_{mr} , implying that the

corresponding values of $ARLM$ are still (much) larger. An increase of less than 7 observations on a total of about 300 then is also not very shocking. A final remark is that the computations for Table 2.1 have been performed using the exact versions in (2.4) and (2.6). However, using the approximate versions in either case, leads to virtually the same outcomes.

Hence the conclusion so far is that $ARLM$ outperforms $ARLCUM$, albeit only marginally, and not the other way around. Perhaps it is good to point out that this is not uniformly true. If we let r continue to increase beyond values of practical interest, for small α a slight reversal may occur. In Table 2.2 we provide a counterexample. Indeed, both the maximal value of the ratio and its location θ_{mr} have further decreased in comparison to the values from Table 2.1, but now a small dip occurs as θ increases from 1 to θ_{mr} .

Table 2.2. An example of ' $ARLM > ARLCUM$ ' for $r = 16$, $\alpha = 0.001$ and small θ .

θ	1	1.2	1.4	2.6 ($= \theta_{mr}$)	5
$ARLCUM/ARLM$	1	0.99	1.00	1.06 (= max)	1.01
$ARLCUM$	1000	309.1	137.8	23.9	16.3
$ARLCUM$	1000	311.7	137.9	22.6	16.2

Since we are specifically dealing with small θ , some analytic explanation is feasible here. We have:

Lemma 2.2 *For $(\theta - 1)$ small, both (2.4) and (2.6) satisfy*

$$ARL_{r,\theta} = \frac{1 - (\theta - 1)c(r) + O((\theta - 1)^2)}{\alpha}, \quad (2.8)$$

with $c(r) = c_M(r) = r\{(1 - (r\alpha)^{1/r})/(r\alpha)^{1/r}\}\log\{1/(1 - (r\alpha)^{1/r})\}$ for $ARLM$ and $c(r) = c_{CUM}(r) = \{(r - (r + 1)h^{-1}(\alpha))/h^{-1}(\alpha)\}\log\{1/(1 - h^{-1}(\alpha))\}$ for $ARLCUM$.

Proof. For $\theta - 1$ small and $0 \leq x < 1$ we have that $1 - (1 - x)^\theta \doteq x - (\theta - 1)(1 - x)\log(1 - x)$, where ' $\doteq$ ' stands for equality up to the given order. Hence in view of (2.4) it follows that indeed $ARLM \doteq r/\{[(r\alpha)^{1/r}]^r(1 + r(\theta - 1)\{(1 - (r\alpha)^{1/r})/(r\alpha)^{1/r}\}\log(1 - (r\alpha)^{1/r})\} \doteq \{1 - (\theta - 1)c_M(r)\}/\alpha$. For $ARLCUM$ from (2.6) observe that $1/h(1 - (1 - x)^\theta) = \{(1 - (1 - x)^\theta)^{-r} - 1\}/(1 - x)^\theta \doteq \{x^{-r}[1 + (\theta - 1)r\{(1 - x)/x\}\log(1 - x)] - 1\}(1 - x) - 1\{1 - (\theta - 1)\log(1 - x)\} \doteq [1 + (\theta - 1)\{r - (r + 1)x/x\}\log(1 - x)]/h(x)$. For $x = h^{-1}(\alpha)$ this precisely produces $ARLCUM \doteq \{1 - (\theta - 1)c_{CUM}(r)\}/\alpha$, as desired. $\square$

The ratio $c_{CUM}(r)/c_M(r)$ obtained from (2.8) is always very close to 1, but it does increase from 0.99 at $r = 2$ to 1.00 at $r = 11$ and 1.01 at $r = 16$. Hence indeed a dip as mentioned above should occur near $\theta = 1$ for (from a practical point of view too) large r . (Note however that the result in (2.8) is very local indeed; it merely indicates the direction and by no means can produce reliable approximations for values in Table 2.2 at points such as $\theta = 1.2$.)

3 Choice of ARL

As remarked in the introduction, all derivations for ARL 's in section 2 are based on the convenient convention that the process is already OoC once our computation starts. This very much is the common practice; nevertheless this assumption is not always realistic. Often the change towards OoC will happen within a group and thus cause some delay. Albers and Kallenberg

(2009) mentioned this point for the continuous case and added that it will affect MIN more than $CUMIN$, as the latter has shorter groups. Since $CUMIN$ was already found to be better than MIN , this conclusion could thus be viewed as a small additional bonus in its favor. However, in the present case matters are more complicated: in section 2 we observed that in the region of interest MAX is in fact better than $CUMAX$, albeit only marginally. Hence the question now is to what extent this disadvantage of $CUMAX$ is balanced by its quicker reaction to onset of the OoC stage.

Note that selecting the change point within a group is actually a two-step procedure. The first consists of recognizing that going OoC can happen after any waiting time X_i (cf. (2.1)), and not just after the last one of a group. This step was made by Wolter (1987). However, as Sego et al. (2008) point out, it still assumes that changes only occur directly after a failure (i.e. a $D_i = 1$). Hence the second step is to allow it to happen at any D_i , and thus within an X_i . For clarity of presentation, we shall begin by analyzing the impact of the first step and subsequently update the results obtained to accommodate the effect of the second one. Consequently, all in all four cases will be considered.

3.1 MAX -chart; change directly after failure.

Hence we consider MAX from section 2, but now assume that the first group of r X_i that (partially) counts in the OoC stage still contains j IC (and thus $r - j$ OoC) items, for $j = 0, 1, \dots, r - 1$. We shall denote the corresponding ARL by $ARLM_j$, implying that the result from (2.4) thus corresponds to $ARLM_0$. Since for the MAX -chart we go through the sequence of D_i 's with steps of fixed size r , the only natural choice for an underlying distribution for the location of the change point is the uniform one, i.e. $P(\text{'change occurs at } j\text{'}) = 1/r$ for $j = 0, 1, \dots, r - 1$. Consequently, what Sego et al. (2008) call the steady-state- ARL , in the present case simply equals

$$SSARLM = r^{-1} \sum_{j=0}^{r-1} ARLM_j. \quad (3.1)$$

We then have the following result.

Lemma 3.1 *Let $\tau = p_1^*/p_\theta^*$, with $p_\theta^* = P_\theta(X_1 \leq n)$, then the difference*

$$\Delta_M = SSARLM - ARLM_0 = \frac{r+1}{2} - \frac{1-\tau^r}{1-\tau}. \quad (3.2)$$

Proof. Given that the first group contains j IC items, it contributes $r-j$ to $ARLM_j$. Moreover, the probability that it causes a signal equals $p_j = (p_1^*)^j (p_\theta^*)^{r-j}$. Consequently $ARLM_j = p_j(r-j) + (1-p_j)\{r-j + ARLM_0\} = r-j + (1-p_j)ARLM_0$. In view of (3.1) this implies that $SSARLM = (r+1)/2 + (1-r^{-1}\sum_{j=0}^{r-1}p_j)ARLM_0$. Since $\sum_{j=0}^{r-1}p_j = (p_\theta^*)^r \sum_{j=0}^{r-1}\tau^j$, while $ARLM_0 = r/(p_\theta^*)^r$, the result in (3.2) now readily follows. $\square$

The interpretation of (3.2) is rather straightforward: the difference between the steady state version and its commonly used counterpart $ARLM_0$ consists of two parts. The first is the penalty $(r+1)/2$ for wasting on the average half a group before 'full' OoC monitoring begins. On the other hand, the second term offers some compensation representing the possibility that the incomplete starting group nevertheless already manages to stop the process. Suppose that $\theta \rightarrow 1$, then $\tau \rightarrow 1$ and $\Delta_M \rightarrow -(r-1)/2$: if the process goes OoC only marginally, it indeed helps if we start on average in the middle of a group, and then the compensation outweighs the penalty. Of course, in this situation $ARLM_0$ is only slightly below $1/\alpha$, so relatively

speaking Δ_M is not very important. At the other opposite, as θ increases, $p_\theta^* \rightarrow 1$ and thus $\tau \rightarrow p_1^* = (r\alpha)^{1/r}$. Then $\Delta_M \rightarrow (r+1)/2 - (1-r\alpha)/(1-(r\alpha)^{1/r})$, which indeed is positive for the range of r and α considered (e.g. 1.48 for $\alpha = 0.001$ and $r = 5$). To see for which θ the difference Δ_M actually crosses 0, first note that equality $(r+1)/2 = (1-\tau^r)/(1-\tau)$ results for the values $\tau = \frac{1}{2}, 0.62, 0.69$ and 0.74 for $r = 2, 3, 4$ and 5 , respectively. Next, the relation $\tau = p_1^*/p_\theta^*$ gives that $\theta = \log(1-(r\alpha)^{1/r}/\tau)/\log(1-(r\alpha)^{1/r})$, from which the values of $\theta = \theta(r, \alpha)$ such that $\Delta_M = 0$ readily follow. The result decreases in r and increases in α , with as boundary values over the region of interest $\theta(2, 0.01) = 2.18$ and $\theta(5, 0.001) = 1.48$. Hence the difference Δ_M in (3.2) becomes positive for kind of 'average' θ : not really large ones, but also not very close to 1. All in all, the analysis above suggests that the effect will be quite moderate.

3.2 CUMAX-chart; change directly after failure.

At the onset of the *OoC* stage, the present number of $X_i \leq \tilde{n}$ will range from 0 to $r-1$. (The value r means a signal and the chart is reset to 0, so these two situations can be identified for the present purpose.) The corresponding *ARL* will now be denoted by $ARLCUM_j$ for $j = 0, 1, \dots, r-1$, and thus the result from (2.6) is $ARLCUM_0$. Note that assuming a uniform distribution over j is no longer justified. In fact, writing here $\tilde{p}_\theta = P_\theta(X_1 \leq \tilde{n})$ (cf. Lemma 3.1), the corresponding transition probabilities during *IC* simply equal $\tilde{p}_1$ for going from j to $j+1$ and $(1-\tilde{p}_1)$ for going from j to 0, for $j = 0, 1, \dots, r-2$. For $j = r-1$, the identification of states r and 0 simply leads to a transition probability 1 for going from $r-1$ to 0. The stationary distribution of this Markov chain can be obtained in a straightforward way, with as result

$$\pi_j = P(\text{'change occurs at } j') = \frac{(1-\tilde{p}_1)\tilde{p}_1^j}{1-\tilde{p}_1^r}, \quad (3.3)$$

for $j = 0, 1, \dots, r-1$. Hence for the present case the steady-state-*ARL* is given by $SSARLCUM = \sum_{j=0}^{r-1} \pi_j ARLCUM_j$ (cf. (3.2)). We have

Lemma 3.2 *Let $\tau = \tilde{p}_1/\tilde{p}_\theta$, then $\Delta_{CUM} = SSARLCUM - ARLCUM_0$ equals*

$$\Delta_{CUM} = \frac{1 - \frac{(1-\tilde{p}_1)(1-\tau^r)}{(1-\tilde{p}_1^r)(1-\tau)}}{1 - \tilde{p}_\theta}. \quad (3.4)$$

Proof. Given that we start from state j , we have a probability $\tilde{p}_\theta^{r-j}$ of finishing straight away in $r-j$ steps. If this fails, we will need on average $ARLCUM_0$, plus the steps wasted before the failed attempt stops. For this latter number, we are dealing with a geometric r.v. which is truncated at $r-j$ and has success probability $(1-\tilde{p}_\theta)$. It is a straightforward exercise to check that its expectation equals $E^* = 1/(1-\tilde{p}_\theta) - (r-j)\tilde{p}_\theta^{r-j}/(1-\tilde{p}_\theta^{r-j})$. Consequently,

$$ARLCUM_j = (r-j)\tilde{p}_\theta^{r-j} + (1-\tilde{p}_\theta^{r-j})\{E^* + ARLCUM_0\} = \frac{1-\tilde{p}_\theta^{r-j}}{(1-\tilde{p}_\theta)\tilde{p}_\theta^r}, \quad (3.5)$$

where the last step follows by first noting that for $j = 0$ we have $ARLCUM_0 = (1-\tilde{p}_\theta^r)\{1/(1-\tilde{p}_\theta) + ARLCUM_0\}$ and thus that $ARLCUM_0 = (1-\tilde{p}_\theta^r)/\{(1-\tilde{p}_\theta)\tilde{p}_\theta^r\}$. (Note that this latter result equals $1/h(\tilde{p}_\theta) = 1/h(P_\theta(X_1 \leq \tilde{n}))$, which indeed agrees with Lemma 2.1.) Next, combining (3.3) and (3.5), we obtain that $SSARLCUM = \sum_{j=0}^{r-1} \pi_j ARLCUM_j = \{1 - (1-\tilde{p}_1)\tilde{p}_\theta^r \sum_{j=0}^{r-1} \tau^j / (1-\tilde{p}_1^r)\} / \{(1-\tilde{p}_\theta)\tilde{p}_\theta^r\}$, from which (3.4) follows. $\square$

Once again, the difference consists of a penalty for having to wait till the present group has stopped before a full *OoC* group starts, but also of a compensation term, because this first group may happen to suffice. To understand the role of the factor $1/(1 - \tilde{p}_\theta)$ in (3.4), observe the following. Roughly speaking, while for $ARLM_0$ we are exactly dealing with r times a geometric r.v. with parameter $(p_\theta^*)^r$, for $ARLCUM_0$ we approximately have $1/(1 - \tilde{p}_\theta)$ times a geometric r.v. with parameter $\tilde{p}_\theta^r$ (see Albers and Kallenberg (2009) for details). Indeed $ARLCUM_0 = 1/h(\tilde{p}_\theta) = (1 - \tilde{p}_\theta^r)/\{(1 - \tilde{p}_\theta)\tilde{p}_\theta^r\} \approx 1/\{(1 - \tilde{p}_\theta)\tilde{p}_\theta^r\}$ (cf. Lemma 3.2). This $1/(1 - \tilde{p}_\theta)$ approximates the average length of a failed group, as opposed to length r in the fixed case. Consequently, this is the scale of the penalty, and hence the occurrence of this factor in (3.5). Subsequently, in the remaining factor its weight 1 is reduced, not only because of the probability that the very first group already suffices, but also due to the truncation effect (cf. $E^* = 1/(1 - (\tilde{p}_\theta) - (r - j)\tilde{p}_\theta^{(r-j)})/(1 - \tilde{p}_\theta^{(r-j)}) \leq 1/(1 - \tilde{p}_\theta)$ in Lemma 3.2).

To analyze the behavior of (3.4), we first note that at the one end, for $\theta \rightarrow 1$, $\tau \rightarrow 1$ and $\Delta_M \rightarrow 1/(1 - \tilde{p}_1) - r/(1 - \tilde{p}_1^r) \approx 1/(1 - \tilde{p}_1) - r$. Consequently, as $\tilde{p}_1 = h^{-1}(\alpha) \approx \{\alpha/(1 - \alpha^{1/r})\}^{1/r}$ (cf. Lemma 2.1), we start here as well with a negative value (e.g. -3.7 for $\alpha = 0.001$ and $r = 5$). As θ increases, so does Δ_{CUM} from (3.4), but note that in the limit, where $\tilde{p}_\theta \rightarrow 1$ and thus $\tau \rightarrow \tilde{p}_1$, we still have $(1 - \tilde{p}_1)(1 - \tau^r)/[(1 - \tilde{p}_1^r)(1 - \tau)] > 1$. In other words, Δ_{CUM} remains negative throughout. In fact, its upper limit to first order equals $r\tilde{p}_1^r/(1 - \tilde{p}_1^r) - \tilde{p}_1/(1 - \tilde{p}_1) \approx -\tilde{p}_1/(1 - \tilde{p}_1)$ (e.g. -0.4 for $\alpha = 0.001$ and $r = 5$). Hence in the present case the conclusion is quite clear: no head-start for the *CUMAX*-chart is realized by considering the customary $ARL = ARLCUM_0$ rather than *SSARLCUM*. In fact, the opposite is true: the advantage of starting with a group which only partially counts as being *OoC*, outweighs its lower probability of securing a signal.

3.3 MAX-chart; change at arbitrary point.

Obviously, matters may become different once we add the second step mentioned at the beginning of section 3, according to which going *OoC* can happen after any D_i , and not only just immediately after a failure (i.e., when $D_i = 1$). Note that now the scale on which we are working will in fact be that of the number of basic events D_i (e.g. patients examined), rather than that of the number of failure times X_i . However, according to Wald's lemma, the difference in terms of *ARL*'s between both scales is just a factor $1/p$ during *IC* and $1/(\theta p)$ during *OoC*. As we precisely start counting when the *OoC* stage kicks in, this difference will be constant. Hence for comparison between methods, it makes no difference which of the scales we use. Consequently, for simplicity we carry on with the *ARL*'s used so far, based on the numbers of X_i , and there is no need to add symbols for distinguishing between *D*- and *X*-based counts.

In the present situation three types of waiting times X_i occur: the *IC*-ones before the change, the *OoC*-ones after the change, and the special one inside which the change takes place. We have the following update of Lemma 3.1:

Lemma 3.3 *Let $\tau = p_1^*/p_\theta^*$, then the difference $\Delta_{M,c} = (r + 1)/2 - f\{(1 - \tau^r)/(1 - \tau)\}$, where*

$$f = \frac{\theta(1 - p)\tau - (1 - \theta p)}{\theta - 1} \approx \frac{\theta\tau - 1}{\theta - 1}. \quad (3.6)$$

Proof. The conditional probability that the first group causes a signal now becomes $p_j^* = (p_1^*)^j p_c (p_\theta^*)^{r-j-1}$, where $p_c = P(Z \leq n)$ and Z is the special waiting time containing the change. Hence here $SSARLM = (r + 1)/2 + (1 - r^{-1} \sum_{j=0}^{r-1} p_j^*) ARLM_0$ with $p_j^* = (p_c/p_\theta^*) p_j$, and thus Δ_M from (3.2) changes into $\Delta_{M,c} = (r + 1)/2 - (p_c/p_\theta^*) \{(1 - \tau^r)/(1 - \tau)\}$. To evaluate this correction factor, we first note that $Z + 1$ is distributed as the sum of two independent geometric r.v.'s with parameters p and θp , respectively (cf. e.g. Sego (2006)). Then it is

straightforward that $P(Z = j) = \theta p \{(1 - p)^j - (1 - \theta p)^j\} / (\theta - 1)$, $j = 1, 2, \dots$, and thus that $P(Z \leq j) = \{\theta(1 - p)[1 - (1 - p)^j] - (1 - \theta p)[1 - (1 - \theta p)^j]\} / (\theta - 1)$. In particular, $p_c = \{\theta(1 - p)p_1^* - (1 - \theta p)p_\theta^*\} / (\theta - 1)$, and hence $f = p_c/p_\theta^* = (\theta(1 - p)\tau - (1 - \theta p)) / (\theta - 1) \approx (\theta\tau - 1) / (\theta - 1)$. $\square$

Hence the penalty remains $(r + 1)/2$, but the compensation after the second step is reduced by the factor f from (3.6). For $\theta \rightarrow 1$, we have $f = \theta a / \{1 - (1 - a)^\theta\} - 1 \rightarrow 1 + \{(1 - a)\log(1 - a)\} / a \approx a/2 + a^2/6$, for $a = (r\alpha)^{1/r}$ small. Moreover, f increases in θ towards the limit a . For e.g. $r = 5$ and α ranging from 0.001 to 0.01, we have values of a between 0.35 and 0.55, implying that the reduction of the compensation term is substantial. Roughly speaking, for small r (like 2 and 3), the compensation becomes negligible and hence $\Delta_{M,c} \approx (r + 1)/2$, while for larger r (like 4 and 5), it lies between $\frac{1}{2}$ and 1 and thus $\Delta_{M,c}$ will be between $(r - 1)/2$ and $r/2$. Hence indeed some head-start effect results from using the basic $ARL = ARLM_0$, rather than the steady state $SSARLM$, but it remains limited.

3.4 CUMAX-chart; change at arbitrary point.

For this final case Lemma 3.2 is transformed into

Lemma 3.4 *In f from (3.6) now let $\tau = \tilde{p}_1/\tilde{p}_\theta$, then $\Delta_{CUM,c} = (1 - f) + f\Delta_{CUM}$.*

Proof. Let $\tilde{p}_c = P(Z \leq \tilde{n})$, then $SSARLCUM = 1 + (1 - \tilde{p}_c)ARLCUM_0 + \tilde{p}_c \sum_{j=0}^{r-1} \pi_j ARLCUM_{j+1}$. Since $\sum_{j=0}^{r-1} \pi_j ARLCUM_{j+1} = ARLCUM_0 + \{1 - (1 - \tilde{p}_1)(1 - \tau^r) / [(1 - \tilde{p}_1^r)(1 - \tau)] / \tilde{p}_\theta\} / (1 - \tilde{p}_\theta)$, it follows that $\Delta_{CUM,c} = 1 + \tilde{p}_c \{1 - (1 - \tilde{p}_1)(1 - \tau^r) / [(1 - \tilde{p}_1^r)(1 - \tau)] / \tilde{p}_\theta\} / (1 - \tilde{p}_\theta) = 1 - \tilde{p}_c/\tilde{p}_\theta + (\tilde{p}_c/\tilde{p}_\theta)\Delta_{CUM}$. In analogy to Lemma 3.3, we have that $\tilde{p}_c = \{\theta(1 - p)\tilde{p}_1 - (1 - \theta p)\tilde{p}_\theta\} / (\theta - 1)$. Hence the desired result follows from (3.6), with $\tau = \tilde{p}_1/\tilde{p}_\theta$. $\square$

The interpretation here is very simple: if the special waiting time during which the change occurs exceeds $\tilde{n}$, we can restart immediately, which means a single additional observation (on the X -scale). Otherwise, we are back at Δ_{CUM} from Lemma 3.2. Indeed, this latter outcome occurs at a weight $f = 1$, which corresponds to $\tilde{p}_c = \tilde{p}_\theta$ and thus to Z which behaves like X_i during OoC . But obviously, $\tilde{p}_c$ will typically be considerably smaller than $\tilde{p}_\theta$. In fact, just as in the previous subsection, f will increase from about $a/2 + a^2/6$ to a as θ increases, but now with $a = \tilde{p}_1 = h^{-1}(\alpha) \approx \{\alpha / (1 - \alpha^{1/r})\}^{1/r}$, as in subsection 3.2. For e.g. $r = 5$ and α ranging from 0.001 to 0.01, we now have values of a between 0.26 and 0.44. In analogy to the above, roughly speaking the compensation becomes negligible for small r (like 2 and 3) and lies around $\frac{1}{2}$ for larger r (like 4 and 5). As the penalty itself only equals 1 here, the head-start effect remains limited to about $\frac{1}{2}$ to 1 observation.

After having studied these four cases, it remains to draw conclusions. The overall one seems to be that head-start effects are either not (consistently) present or otherwise quite small. The first category occurs in subsections 3.1 and 3.2: here the effect is mixed (first negative, later positive) for MAX and even negative throughout for $CUMAX$. Shifting from ordinary ARL 's to steady state versions consequently does provide a small advantage (1 to 2 observations) for $CUMAX$ w.r.t. MAX . In the - probably more interesting - situation where both steps have been made and the change can occur anywhere, small head-start effects are indeed present. From subsections 3.3 and 3.4 we see that these range from $(r - 1)/2$ (for larger r) to $(r + 1)/2$ (for smaller r) in case of MAX and likewise from $\frac{1}{2}$ to 1 for $CUMAX$. Again a small advantage ($\frac{1}{2}$ to $1\frac{1}{2}$ observations) of $CUMAX$ over MAX results from changing to the steady state perspective. Hence we can now answer the question posed at the beginning of the

present section: the marginal advantage of *MAX* over *CUMAX* observed in section 2 is indeed (at least to considerable extent) balanced by the latter's quicker reaction to the start of the *OoC* stage. Consequently, from a steady state perspective, the performance of both procedures is very similar. As *CUMAX* dates back to Chen (1978) and as such is the more prominent of the two, it makes sense to introduce for this procedure as well its empirical nonparametric version. This will be the topic of the next section.

4 The nonparametric chart

After the excursion in section 3, we return to the standard setup from section 2. As argued at the beginning of that section, our starting point had to be the simple case of a known geometric distribution for the waiting times X_i . This was necessary to establish first that proposals such as *MAX* and *CUMAX* only lose little detection power compared to the optimal choice for this ideal setup. By now we can drop this artificial assumption and consider the case of main interest: an empirical nonparametric *CUMAX*-chart, providing an attractive robust alternative to the usual methods which fully ignore the often large (and sometimes huge) estimation effects. To avoid duplication, we shall be very brief here. The emphasis will be on the ideas and the actual implementation. For details and further properties, we refer to our earlier papers dealing with nonparametric proposals (Albers and Kallenberg (2009) and Albers (2011)).

Hence assumption (2.1) here no longer holds: the waiting times X_i during *IC* have an unknown distribution function F . Suppose a Phase I sample $X_1, \dots, X_m$ is available, then $F_m(x) = m^{-1} \# \{X_i \leq x\}$ is the corresponding empirical df and F_m^{-1} the quantile function, i.e. $F_m^{-1}(t) = \inf\{x | F_m(x) \geq t\}$. Note that $F_m^{-1}(t)$ equals $X_{(i)}$ for $(i-1)/m < t \leq i/m$, with $X_{(1)} < \dots < X_{(m)}$ the order statistics for the sample. Consequently, any q -quantile $k = F^{-1}(q)$ can in principle be estimated by

$$\hat{k} = F_m^{-1}(q) = X_{(s)}, \quad (4.1)$$

where $s = \{mq\}$, with $\{y\}$ denoting the smallest integer $\geq y$. For the *MAX*-chart, (2.2) now becomes $FAR = F^r(n) = r\alpha$ and thus the explicit boundary value from (2.3) in its turn is replaced by $n = F^{-1}((r\alpha)^{1/r})$. In view of (4.1), it immediately follows that the estimated version $\hat{n} = X_{(s)}$, with $s = \{m(r\alpha)^{1/r}\}$. Hence the empirical nonparametric *MAX*-chart simply checks after each incoming group of r waiting times $Y_1, \dots, Y_r$ whether $\max(Y_1, \dots, Y_r) \leq X_{(s)}$. If this is the case, a signal is given; otherwise the next group of size r is considered. Note that grouping indeed helps: for $r = 1$, typically $s = \{m\alpha\}$ will equal 1, while for $r > 1$ such use of extreme order statistics (and thus the occurrence of relatively huge stochastic errors) is effectively avoided.

For the *CUMAX*-chart, we can fortunately proceed in precisely the same manner. For unknown F we have (see also Lemma 2.1 and Lemma 3.2) that $F(\tilde{n}) = P_1(X_1 \leq \tilde{n}) = \tilde{p}_1$ has to equal $h^{-1}(\alpha) \approx \{\alpha/(1 - \alpha^{1/r})\}^{1/r}$. Hence here $\tilde{n} = F^{-1}(h^{-1}(\alpha))$ and

$$\hat{\tilde{n}} = X_{(s)} \text{ with } s = \{m h^{-1}(\alpha)\} \approx \left\{ m \left(\frac{\alpha}{1 - \alpha^{1/r}} \right)^{1/r} \right\}. \quad (4.2)$$

Again implementation is straightforward: for the empirical nonparametric version of the *CUMAX*-chart a signal follows as soon as r consecutive Y_i are all at most $\hat{\tilde{n}}$. Moreover, although $\hat{\tilde{n}}$ is somewhat smaller than $\hat{n}$, grouping is still effective. For example, $m = 100$, $r = 3$ and $\alpha = 0.001$ produces $s = \{14.4\} = 15$ for *MAX* and $s = \{10.4\} = 11$ for *CUMAX*, which both lead too much less extreme order statistics than $X_{(1)}$.

Clearly, as soon as estimation - even if this only means replacing p in (2.3) by some $\hat{p}$ - enters the picture, performance characteristics such as *FAR* and *ARL* become stochastic as well. In

the present situation during *IC* this implies that $ARL = 1/h(F(\tilde{n})) = 1/h(F(F^{-1}(h^{-1}(\alpha)))) = 1/\alpha$ becomes (cf. (4.2)) $\widehat{ARL} = 1/h(F(X_{(s)}))$, with $s = \{mh^{-1}(\alpha)\}$. Let $U_{(1)} < \dots < U_{(m)}$ be the order statistics for a sample of size m from the uniform distribution on $(0,1)$, then it follows that this $\widehat{ARL}$ is distributed as $1/h(U_{(s)})$ during *IC*, which shows that the resulting chart is indeed truly nonparametric. It is also clear that $U_{(s)} \xrightarrow{P} h^{-1}(\alpha)$ and thus that $\widehat{ARL} \xrightarrow{P} 1/\alpha$. Consequently, there is no model error and the stochastic error tends to 0 as $m \rightarrow \infty$. For $r = 1$ this convergence is too slow for practical purposes (cf. the remarks in the introduction), but for $r > 1$ the procedure is quite well-behaved, in the sense that the stochastic error becomes comparable in size to that of the parametric competitors (see the papers mentioned above for details). As the latter suffer from non-vanishing model errors, the nonparametric proposal definitely seems preferable, unless one is quite convinced about the correctness of the rigid parametric model.

The fact that the nonparametric stochastic error for larger r is no longer huge, but comparable to the parametric one, is of course nice. Nevertheless, even in the parametric case, such errors are still not negligible and the need is felt to offer possibilities for suitable corrections. Again we refer to the aforementioned papers for a detailed discussion of this issue. Here we just indicate some of the possibilities involved. For example, as mentioned above, the fixed $ARL = 1/\alpha$ has been replaced by $\widehat{ARL} = 1/h(U_{(s)})$. To control the extent to which this $\widehat{ARL}$ may fall short of the intended $1/\alpha$, we can bound the exceedance probability:

$$P\left(\widehat{ARL} < \frac{1}{\alpha(1+\varepsilon)}\right) \leq \beta, \quad (4.3)$$

for suitably chosen small positive ε and β . To be specific, suppose that $\alpha = 0.001$ and thus the intended $ARL = 1000$, then we can e.g. require that $P(\widehat{ARL} < 800)$ is at most 20%, which thus corresponds to $\varepsilon = 0.25$ and $\beta = 0.2$ in (4.3).

It is immediate that the probability in (4.3) equals $P(U_{(s)} > h^{-1}(\alpha(1+\varepsilon)))$, which in its turn translates into the binomial probability $P(\text{Bin}(m, h^{-1}(\alpha(1+\varepsilon))) < s)$. Next, using a normal approximation step (cf. Lemma 4.1 in Albers and Kallenberg (2009)), we finally arrive at $\Phi(-\varepsilon m^{1/2}v)$, in which Φ is the standard normal distribution function and $v = v(r, \alpha) = \{h^{-1}(\alpha)/(1 - h^{-1}(\alpha))\}^{1/2}/r$. For example, for $r = 1, 2, \dots, 5$, the factor $v(r, 0.001)$ equals 0.032, 0.091, 0.113, 0.120 and 0.121, respectively. This shows that taking $r > 1$ provides a very substantial improvement, but also that the probability in (4.3) remains considerable. For common values like $\varepsilon = 0.25$, $m = 100$ and $r = 3$ we e.g. obtain the value 0.41, which may very well be considered still unsatisfactory. The way to solve this problem is to apply a small correction: the value s should be replaced by a - typically slightly - smaller value s^* . More specifically, to achieve approximate equality in (4.3), it suffices to use $X_{(s^*)}$ with

$$s^* = s(1 + \frac{\varepsilon}{r}) - u_\beta \{s(1 - \frac{s}{m})\}^{\frac{1}{2}}, \quad (4.4)$$

in which $u_\beta = \Phi^{-1}(1 - \beta)$, the standard normal upper β -quantile. (For implementation, use interpolation between $X_{(\{s^*\}-1)}$ and $X_{(\{s^*\})}$.) Note from (4.4) that s^* increases in ε and β , as should be the case. Moreover, the positive correction will be smaller than the negative one, unless m is much larger than common in practice. For the values used by way of example so far ($m = 100$, $r = 3$, $\alpha = 0.001$ and $\varepsilon = 0.25$) e.g. the choice $\beta = 0.2$ leads to $s^* = 9.28$ in (4.4), as opposed to $s = 11$. Hence replacing $X_{(11)}$ by $0.72X_{(9)} + 0.28X_{(10)}$ reduces the exceedance probability from the value 0.41 obtained above to the intended value 0.20.

5 Conclusions and summary

Our starting point was the *MAX*-chart, which signals if all waiting times Y_i in a group of size r are at or below the boundary value $X_{(s)}$, a suitably chosen order statistic from a Phase I sample. This chart has good detection power and, being nonparametric, no model error. Moreover, for $r > 1$ its stochastic error is comparable to that of parametric competitors, which usually do have a non-vanishing model error. The question in the present paper was whether a cumulative approach offers still further improvement. This *CUMAX*-chart signals as soon as any sequence of r consecutive Y_i are all at most $X_{(s)}$. Such an accelerated version definitely looks better, but matters are not that simple. For example, fair comparison requires the ' s ' used for *CUMAX* to be smaller than the one for *MAX*. Indeed, in section 2 we demonstrate that over the parameter region of interest, *MAX* is in fact marginally better than *CUMAX* in terms of *ARL*. On the other hand, the customarily used *ARL* assumes the process to be *OoC* as monitoring starts. It can (and has been) argued that a steady state version is more realistic, as it removes possible head-start advantages provided by the ordinary *ARL*. It seems that for *CUMAX*, being the more flexible of the two, the head-start effect would be smaller than for *MAX*. In section 3 we demonstrate that this time the intuition is indeed correct. Once again the differences are rather marginal, and as such more or less balance the opposite effects noted in section 2. All in all, the conclusion is that both charts show very little difference in performance. Hence it remains largely a matter of taste which one is preferred. The advantage of *MAX* is its even greater simplicity, definitely as far as properties and corrections are concerned. For *CUMAX* speaks the fact that it is in fact a nonparametric counterpart of the well-known sets method. Hence the step towards this ultimate nonparametric version is supplied in section 4. To avoid duplication, only few details are given, as this step is quite similar to the one for *MAX*.

For convenience, we summarize the application of the *CUMAX*-chart:

1. Select a desired in-control $ARL = 1/\alpha$ and a degree of change θ in the average failure probability during *OoC* that should be optimally protected against.
2. Apply rule of thumb (2.5) to obtain the best r (typically truncate at 5 in practice).
3. Select size m (e.g. $m = 100$) and collect a Phase I sample of waiting times $X_1, \dots, X_m$.
4. Compute the smallest integer $s \geq mh^{-1}(\alpha)$ and find the order statistic $X_{(s)}$ (cf. (4.2)).
5. Start monitoring: signal as soon as r consecutive Y_i are all $\leq X_{(s)}$.
6. If desired, select small ε and β such that $P(\widehat{ARL} < 1/\{\alpha(1 + \varepsilon)\}) = \beta$ is acceptable.
7. Replace s by $s^* = s(1 + \varepsilon/r) - \Phi^{-1}(1 - \beta)\{s(1 - s/m)\}^{\frac{1}{2}}$ (cf. (4.4)) to achieve this bound.

Acknowledgement. The author is grateful to Erik van Doorn for his contributions to section 3.

References

- Albers, W. (2011). Empirical nonparametric control charts for high-quality processes. *J. Statist. Planning & Inference* **140**, 214-225.
- Albers, W. and Kallenberg, W.C.M. (2009). *CUMIN* charts. *Metrika* **70**, 111-130.
- Chen, R. (1978). A surveillance system for congenital malformations. *J. Amer. Statist. Ass.* **73**, 323-327.
- Chen, R., Connelly, R.R. and Mantel, N. (1997). The efficiency of the sets and the cuscore techniques under biased baseline rates. *Statist. in Med.* **16**, 1401-1411.
- Kenett, R. and Pollak, M. (1983). On sequential detection of a shift in the probability of a rare event. *J. Amer. Statist. Ass.* **78**, 389-395.

- Shaha, S. H. (1995). Acuity systems and control charting. *Qual. Manag. Health Care* **3**, 22-30.
- Sego, L.H. (2006). Applications of control charts in medicine and epidemiology. PhD-dissertation, Virginia Polytech. Inst. and State Univ.
- Sego, L.H, Woodall, W.H. and Reynolds, M.R. (2008) A comparison of surveillance methods for small incidence rates. *Statist. in Med.* **27**, 1225-1247.
- Sonesson, C. and Bock, D. (2003). A review and discussion of prospective statistical surveillance in public health. *J. R. Statist. Soc. A* **166**, 5-21.
- Thor, J., Lundberg, J., Ask, J., Olsson, J., Carli, C., Härenstam, K. P. and Brommels, M. (2007). Application of statistical process control in healthcare improvement: systematic review", *Qual. & Safety in Health Care* **16**, 387-399.
- Wolter C. (1987). Monitoring intervals between rare events: a cumulative score procedure compared with Rina Chen's sets technique. *Methods of Information in Medicine* **26**, 215-219.