Control charts for health care monitoring under intermittent Out-of-Control behavior

Willem Albers,Department of Applied Mathematics, University of Twente, P.O.Box 217, 7500 AE Enschede,The Netherlands

Abstract. Health care monitoring typically concerns attribute data with very low failure rates. Efficient control charts then signal if the waiting time till r (e.g. $r\leqslant 5$) failures is too small. An interesting alternative is the *MAX*-chart, which signals if all the associated r waiting times for a single failure are sufficiently small. In comparing these choices, the usual change point set-up has been used, in which going Out-of-Control (*OoC*) means that the failure rate suddenly jumps up and then stays at this higher level. However, another situation of interest is intermittent *OoC* behavior. In industrial settings, an *OoC* process can be adjusted to return to In-Control (*IC*), but with health care monitoring this usually is no option and stretches of *OoC* and *IC* behavior may alternate. Comparison of such intermittent alternatives to the change point situation shows that the former can be characterized as tail alternatives, in the sense that the difference w.r.t. the *IC*-distribution becomes more concentrated in the lower tail. This suggests to generalize the *MAX*-chart as follows: now signal if all but 1 (or 2) out of r individual waiting times are too small. A numerical study shows that this approach indeed works well.

Keywords: Statistical Process Control, high-quality processes, tail alternatives, average run length
2000 Mathematics Subject Classification: 62P10, 62C05, 62G15

1. Introduction and motivation

In health care monitoring we are typically dealing with attribute data for which the rate of failure (malfunctioning equipment, surgical error, recurrence of cancer) p should be (very) small. For some review material, see e.g. Sonesson and Bock(2003) and Shaha(1995). Suitable control charts in this situation can be based on waiting times till r failures have occurred, with usually $1\leq r\leq 5$. If such a negative binomially distributed waiting time is too small, a signal is given. In Albers(2010) the optimal choice of r is derived and estimation of the typically unknown p on the basis of a Phase I sample is discussed. Because of the small false alarm rates (*FAR*'s) involved, the estimation effects involved are not at all negligible and suitable corrections are derived.

Quite often p will vary from patient to patient and such heterogeneity will lead to overdispersion. Assuming a negative binomial distribution then no longer is correct and a nonparametric approach becomes attractive. However, the small probabilities involved often ruin this possibility: Phase I samples will usually be much too small to avoid huge estimation effects in a fully nonparametric approach. In Albers(2011) it is demonstrated how this obstacle can be overcome in a very satisfactory manner by applying a *MAX*-chart instead. Here no longer the total waiting time till the r^{th} failure is used. Instead a signal results if each of the r underlying waiting times for a single failure is sufficiently small. If homogeneity happens to be true after all, this *MAX*-chart is only slightly less efficient than the optimal negative binomial one. Once these ideal conditions no longer hold, the estimated version of the *MAX*-chart, being nonparametric, remains valid, which obviously does not hold true for its negative binomial competitor.

In a sense, the results above already provide a quite satisfactory solution for the monitoring problem described. However, we should realize that so far all optimality considerations and comparisons have been exclusively based on the change point model. Here the process is In-Control (*IC*) till a given point, at which it suddenly goes Out-of-Control (*OoC*), as the defective rate jumps from p to θp, for some $\theta>1$. This change point model is the common or standard choice, so it makes

sense to begin by following it. Nevertheless, we should realize that other possibilities do occur as well. One of these, which will be the topic of the present paper, concerns intermittent *OoC* behavior. Here the process is not permanently showing a higher defective rate under *OoC*, but stretches governed by p and by θp keep alternating. Note that especially with health care applications this possibility seems likely to occur. In industrial settings, signals typically lead to direct interventions, forcing the process back into the *IC* state; in the present type of application this is often (less) feasible.

We shall begin by analyzing how intermittent behavior of the process affects the resulting waiting times on which the charts are based. It turns out that in the intermittent case the difference in the waiting time distribution between *IC* and *OoC* becomes more concentrated in the lower tail in comparison to what happens in the change point situation. The phenomenon of so-called tail alternatives is known from other contexts as well, see e.g. Albers et al.(2001). Some optimality considerations readily suggest how to adapt the *MAX*-chart to this situation. It will be argued that the proper way to react to tail behavior is to signal already if all but j, with typically j=0, 1 or 2, of the r underlying waiting times for a single failure are sufficiently small. Note that in a sense the term 'already' is misleading: to achieve the same *FAR* in each case, an increase of j will imply a decrease of the value which is considered 'sufficiently small'. This decrease is in line with the desire to improve the detection power against tail alternatives. A numerical study is presented to illustrate that the suggested approach indeed turn out to work well.

The paper is organized as follows: in section 2 the notation and methodology is introduced by means of the *MAX*-chart. The intermittent case and the corresponding proposal for a chart are the subject of section 3.

2. The *MAX*-chart

Our starting point is the change point situation: a sequence D_1, D_2, … 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 quickly be stopped. For this purpose consider the subsequent waiting times till a failure occurs, denoted by X_i, i=1, 2,... Clearly, these r.v.'s will be geometric: $P(X_i=k)=p(1-p)^{k-1}$, k=1,2,.... In Albers(2011) a *MAX*-chart is proposed, which for some r (typically $r\leq 5$) signals if $\max(X_1, ..., X_r)$ is too small (and otherwise continues towards the next group of r X_i's). As mentioned there, the proposal in itself is far from new, as it essentially goes back to the so-called sets method from Chen (1978). However, the focus in Albers(2011) is on using it to obtain a satisfactory nonparametric procedure. In this way the serious underlying estimation issues are solved, which in practice are typically conveniently ignored almost completely.

The lower limit n should be chosen such that during *IC* for some small $\alpha>0$

$$P(\max(X_1, ..., X_r) \leq n) = r\alpha. \tag{2.1}$$

In this way a fair comparison for varying r can be made, as the average run length (*ARL*) during *IC* will have the same value $r/(r\alpha)=1/\alpha$ for all r. From (2.1) it follows that $1-(1-p)^n = \{r\alpha\}^{1/r}$ and thus $n = \log(1- \{r\alpha\}^{1/r})/\log(1-p)$. During OoC the alarm rate becomes$\{1-(1-\theta p)^n\}^r$, and consequently we obtain

$$ARL_{r,\theta} = r/\{1 - (1-\{r\alpha\}^{1/r})^{\log(1-\theta p)/\log(1-p)}\}^r \approx r/\{1 - (1-\{r\alpha\}^{1/r})^{\theta}\}^r. \tag{2.2}$$

As p is (very) small, the last step in (2.2) provides a really close approximation (see Albers(2011) for details).

Using (2.2), the performance of the *MAX*-chart for various r is studied in Albers(2011). For r=1, the simple geometric chart has $ARL_{1,\theta} \approx 1/(\theta\alpha)$, which decreases rather slowly as θ increases. Unless θ is very large, increasing r indeed greatly improves matters, as can e.g. be demonstrated by studying $ARL_{1,\theta}/ARL_{r,\theta}$. As functions of θ these ratios start at 1, increase to a maximum and then slowly decrease again to $1/r$. For larger r the peak is higher and it occurs for lower θ. On the other hand, the decline is also faster as r increases. A simple rule of thumb is to use $\min(5,r^{opt})$, where $r^{opt}=1/\{\alpha(2.6\theta+2) + 0.01(4\theta-3)\}$. The truncation is added here because using too high values of r may feel awkward in practice and moreover almost all the improvement over r=1 has already been realized at r=5.

Finally, let $ARL^*_{r,\theta}$ correspond to the negative binomial chart, then studying $ARL_{r,\theta}/ARL^*_{r,\theta}$ shows how much the *MAX*-chart looses w.r.t. to this optimal choice if homogeneity happens to hold after all. As is demonstrated in Albers(2011), the excess of these ratios over 1 stays quite limited. Hence it does makes sense to use the *MAX*-chart to provide a robust alternative. This is achieved in the estimation step: let $X_{(1)} < \ldots < X_{(m)}$ be the order statistics of a Phase I sample of size m, then the lower bound n can be replaced by its estimated counterpart $X_{(s)}$, where s is the smallest integer $\geq m(r\alpha)^{1/r}$. Note that as r increases, this s indeed moves away from the extreme value 1, thus allowing estimation with 'normal' instead of huge errors. See Albers(2011) for further details.

3. The intermittent case

In the change point situation the process went *OoC* once the failure probability jumped from p to θp, for some θ>1. Here we will instead consider the case where going *OoC* means that for stretches of the D_i still p holds, while for the remainder the failure rate has become $\kappa\theta p$, for some suitable κ>1. To get some feeling for how this may work out in comparison to the change point case, we model the impact on the waiting times as follows: for some γ with $0 \leq \gamma \leq 1$, we have

$$P_{\theta,\gamma}(X_i=k) = \gamma p(1-p)^{k-1} + (1-\gamma)\kappa\theta p(1-\kappa\theta p)^{k-1}. \qquad (3.1)$$

Hence for simplicity we assume in (3.1) that the D_i leading to a certain waiting time are either all based on p, or all on $\kappa\theta p$. Obviously, this is not meant as an exact description of the underlying reality, but merely as a better approximation than, and hence improvement over, the simple geometric model (cf. Box's remark 'all models are wrong, but some are useful').

To allow a meaningful comparison, we first align the intermittent situation from (3.1) with the constant *OoC* rate θp from the change point case. This implies that we require in addition $\gamma/p + (1-\gamma)/(\kappa\theta p) = 1/(\theta p)$, and thus that $\gamma=(\kappa-1)/(\kappa\theta-1)$ (or equivalently $\kappa=(1-\gamma)/(1-\gamma\theta)$). Note in particular that now $\gamma<1/\theta$: during OoC, the expected waiting time should be reduced by a factor $1/\theta$, so the fraction γ with the original waiting time indeed needs to be smaller than this $1/\theta$. On the other hand, observe that γ=0 (i.e. κ=1) reproduces the change point case $\theta p(1-\theta p)^{k-1}$.

The comparison can subsequently be made by considering the likelihood ratio

$$P_{\theta,\gamma}(X_i=k)/P_{1,0}(X_i=k) = \gamma + (1-\gamma)\kappa\theta\{(1-\kappa\theta p)/(1-p)\}^{k-1}. \qquad (3.2)$$

This ratio decreases more sharply in k if γ (and thus κ) is larger. More precisely, it crosses the value 1 for k such that $\kappa\theta\{(1-\kappa\theta p)/(1-p)\}^{k-1}=1$. Hence k-1 then equals $-\log(\kappa\theta)/\log\{(1-\kappa\theta p)/(1-p)\} \approx \log(\kappa\theta)/\{(\kappa\theta-1)p\}$, and it follows that this k decreases in κ. Consequently, in comparison to the change point case, the intermittent situation can be characterized as a so-called tail alternative. The notion of tail alternatives has been used before, see e.g. Albers et al.(2001) and the references contained in that paper. A typical application occurs when comparing two medical treatments where a possible difference in effect (e.g. expected additional survival time) is not constant over the whole range of outcomes, but instead (more or less) concentrated at one end. This can be due to a multitude of causes, such as postoperative mortality.

The characterization above will help us to find a suitable chart for intermittent behavior. To show this, we proceed as follows. If we have $(X_1, ..., X_r)$ at our disposal in the change point situation, a simple Neyman-Pearson argument shows that using the sum of these X_i, and thus the negative binomial chart, is optimal (cf. (3.2)). Indeed, as mentioned in the Introduction, the *MAX*-chart then is (fortunately only slightly) less efficient and 'only' chosen for its superior robustness properties. But let us for a moment reverse matters and ask for what type of alternative the *MAX*-chart would in fact be optimal. The answer is a bit exotic, but also straightforward: suppose

$$P(X_i=k) = bp(1-p)^{k-1} \text{ for } k\leq n \text{ and } P(X_i=k) = ap(1-p)^{k-1} \text{ for } k>n, \qquad (3.3)$$

where $b>1>a$ (and obviously a and b for given n chosen such that these probabilities do sum to 1). For this type of alternative another look at the likelihood ratio readily shows that here it is optimal to reject for large values of Y, with $Y = \#\{X_i \leq n,\ i=1, .., r\}$. The extreme case $Y=r$ in fact precisely results in using the *MAX*-chart with $FAR=\{1-(1-p)^n\}^r$, which boils down to $r\alpha$ if in (3.3) we let $n = \log(1- \{r\alpha\}^{1/r})/\log(1-p)$.

However, the result obtained covers more than just *MAX*. In fact, for each j=0, 1,2,...,r-1, giving a signal if $Y\geq r-j$ produces a chart. Observe that Y is bin(r,c) with $c=P(X_i \leq n)$ and denote $P(Y\geq i)$ by $B^*(r,c,i)$, then the corresponding $FAR=B^*(r,c,r-j)$. As before, fair comparison requires the same $ARL=1/\alpha$ to be used in all cases. Obviously, $B^*(r,c,r-j)$ increases in j and hence $c=c_j$ has to simultaneously decrease, in order to keep the level at $r\alpha$. Since $c_j=P(X_i \leq n_j)$, it follows that then $n_j=\log(1-c_j)/\log(1-p) \approx -\log(1-c_j)/p$ decreases in j as well. To give an example, if ARL=1000 and r=5, we obtain that $B^*(5,0.347,5) = B^*(5,0.185,4) = B^*(5,0.083,3) =0.005$ and consequently that $n_jp\approx-\log(1-c_j)= 0.426$, 0.205 and 0.087 for j=0,1 and 2, respectively. Hence each of these charts with FAR=$r\alpha$ is optimal against an alternative of the form (3.3). As j increases, the alternative in question more and more takes the form of a tail alternative.

The excursion to the alternatives (3.3) thus has shown how the *MAX*-chart should be adapted to become suitable for alternatives with tail character, such as the ones from (3.1): not only signal if all r individual waiting times are sufficiently small, but already do so if this holds for r-1, or possibly even for r-2, such times. To compensate for this more liberal attitude, the criterion for being 'sufficiently small' of course has to strengthened accordingly. Precisely this latter effect produces the suitability for detecting tail behavior. In principle we could use $r-j$ for j all the way up to r-1 (which upper boundary case leads to a *MIN*-chart, which is not really a good idea), but numerical evidence, like the example above, suggests that $j<r/2$ suffices. Since typically $r\leq 5$, this leads to concentrating on j=1 or j=2 as alternatives to j=0 (the *MAX*-chart).

To implement the charts, we observe that the requirement $FAR=B^*(r,c_j,r-j)=r\alpha$ through expansion w.r.t. c_j produces the following 3rd order approximation

$$c_j = c_{0j}\Big\{1+\frac{j}{r+1-j}c_{0\,j} + j\frac{(r+1+2\,j)(r+1-j)+2\,j}{2(r+1-j)^2(r+2-j)}c_{0\,j}^2\Big\}, \qquad (3.4)$$

where $c_{0j} = \{r\alpha/\binom{r}{j}\}^{1/(r-j)}$. For j=0, the solution $c_0=c_{00}=(r\alpha)^{1/r}$ is exact, while for j>0 the value of c_j is sufficiently small to make (3.4) quite accurate. For $r\leq 5$ and $\alpha\leq 0.01$ the error resulting from using $c_1=c_{01}\{1 + c_{01}/r + (r+2)c_{01}^2/(2r^2)\}$ with $c_{01}=\alpha^{1/(r-1)}$ is at most 2% and from using $c_2=c_{02}\{1 + 2c_{02}/(r-1) + (r^2+4r+1)c_{02}^2/(r(r-1)^2)\}$ with $c_{02}=(2\alpha/(r-1))^{1/(r-2)}$ at most 1% (for the 2nd order approximation we find 7% and 3%, respectively). As $n_j=\log(1-c_j)/\log(1-p) \approx -\log(1-c_j)/p$ continues to hold for j>0, the approximation for the lower bound n_j is immediate from (3.4).

It remains to analyze the resulting *OoC*-behavior. In analogy to (2.2) it now follows from (3.1) that

$$c_{\theta,\gamma,j} = P_{\theta,\gamma}(X_i \leq \log(1-c_j)/\log(1-p)) \approx \gamma c_j + (1-\gamma)\{1-(1-c_j)^{\kappa\theta}\}, \qquad (3.5)$$

with again $\kappa=(1-\gamma)/(1-\gamma\theta)$. Hence e.g. $c_{1,\gamma}=c_j$ (*IC*-case) and $c_{\theta,0}=1-(1-c_j)^{\theta}$ (change point *OoC*-case). The final step is to observe that then

$$ARL_{\theta,\gamma,j} = r/B^*(r,c_{\theta,\gamma,j},r-j), \qquad (3.6)$$

with $c_{\theta,\gamma,j}$ as in (3.5), c_j as in (3.4) and B^* again the binomial exceedance probability. In particular, $ARL_{1,\gamma,j}=1/\alpha$, $ARL_{\theta,\gamma,0}=r/\{\gamma(r\alpha)^{1/r}+(1-\gamma)[1-(1-(r\alpha)^{1/r})^{\theta}]\}^r$ (cf. (2.2)).

By way of illustration, we collect some representative outcomes in Table 1 below. For r we take the illustrative value 5 and for α the small 0.01 and the very small 0.001. In Albers(2010) values of θ up to 4 are used; here we have a mixture of 1 and $\kappa\theta$, so we stop at the smaller value 2 for θ and let κ run as far as 7.

Table 1. Values of $ARL_{\theta,\gamma,j}$ from (3.6) for $r=5$ and various α, θ *and* $\kappa=(1-\gamma)/(1-\gamma\theta)$. The upper, middle and lower values in each cell are for j =0, 1 and 2, respectively.

$\theta = 3/2$

α \ κ	1	2	3	4	5	6	7
	214	115	88.0	78.3	74.8	73.9	74.2
0.001	260	126	81.8	60.8	50.3	44.2	40.4
	337	170	106	74.5	57.0	46.2	39.0
	30.3	24.3	24.0	24.8	25.5	26.2	26.7
0.01	34.2	22.4	18.7	17.3	16.7	16.6	16.6
	40.2	25.0	18.9	15.9	14.2	13.1	12.5

$\theta = 2$

α \ κ	1	2	3	4	5	6	7
	80.9	39.7	31.5	29.5	29.3	29.8	30.4
0.001	108	43.1	27.0	20.8	17.8	16.3	15.4
	161	65.2	37.7	26.1	20.1	16.7	14.5
	15.6	13.3	13.9	14.7	15.3	15.7	16.0
0.01	17.8	11.5	9.97	9.58	9.53	9.59	9.69
	22.5	12.9	9.86	8.57	7.93	7.59	7.39

Note that Table 1 nicely shows that the actual behavior is as predicted on the basis of our theoretical considerations. For κ=1 we are in the change point situation and the *MAX*-chart (j=0) clearly beats using j=1 and (certainly) j=2. However, to keep matters in perspective, note that all three charts easily defeat the geometric chart (r=1), for which $ARL \approx 1/(\theta\alpha)$. As soon as κ start to grow, the choice j=1 tends to become better than the *MAX*-chart. E.g. at α=0.01 intermittent behavior between θ either equal to 1 or to 4 has ARL=11.5 for j=1 and ARL=13.3 for j=0. The larger values of κ have been added mainly to show that indeed j=2 becomes better than the *MAX*-chart as well, and that this choice eventually also beats j=1. A qualitative conclusion that suggests itself is that in case of intermittent behavior j=1 provides a good alternative to using a *MAX*-chart. If the tail character is quite extreme, j=2 might even be the sensible choice.

A final point to consider is the (nonparametric) estimation aspect. As argued in the Introduction, this is quite important, as reliance on e.g. homogeneity quite often is rather dubious. Nevertheless, we can be extremely brief about this topic here: everything just goes on as in the *MAX*-case. No longer conclude from $c_j = P(X_i \le n_j)$ that the lower bound $n_j = \log(1-c_j)/\log(1-p)$, but instead use its estimated counterpart $X_{(s)}$, where s is the smallest integer $\ge mc_j$ and $X_{(1)} <...< X_{(m)}$ are the order statistics from a Phase I sample of size m.

For convenience we summarize the application of the new chart as follows:

1. Select a desired in-control $ARL = 1/\alpha$ and an average degree of change θ during *OoC* that should be optimally protected against.
2. As a suitable choice for r, use $\min(5, r^{opt})$, with $r^{opt} = 1/\{\alpha(2.6\theta+2) + 0.01(4\theta-3)\}$.
3. For low, average or strong intermittent behavior use j=0 (=*MAX*), 1 or 2, respectively.
4. Compute c_j from (3.4) and collect a Phase I sample X_1, ..., X_m (take e.g. m=100).
5. Find the smallest integer $s \ge mc_j$ and the corresponding order statistic $X_{(s)}$.
6. Now start monitoring: collect the next r waiting times X_{m+1}, ..., X_{m+r}.
7. Signal if at least $r-j$ of these X_i are $\le X_{(s)}$; otherwise collect the next group of size r.

References

Albers, W. (2010). The optimal choice of negative binomial charts for monitoring high-quality processes. *J. Statist. Planning & Inference* **140**, 214-225.

Albers, W. (2011). Empirical nonparametric control charts for high-quality processes. To appear in *J.Statist.Planning & Inference.*

Albers, W., Kallenberg, W.C.M. and Martini, F. (2001). Data driven rank tests for classes of tail alternatives. *J.Amer.Statist.Ass.* **96**, 685−696.

Chen, R. (1978). A surveillance system for congenital malformations. *J. Amer. Statist. Ass.***73**, 323–327.

Shaha, S H. (1995). Acuity systems and control charting. *Qual. Manag. Health Care* **3**, 22–30.

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.