

Control charts for health care monitoring under overdispersion

Willem Albers

Department of Applied Mathematics

University of Twente

P.O. Box 217, 7500 AE Enschede

The Netherlands

Abstract. An attractive way to control attribute data from high quality processes is to wait till $r \geq 1$ failures have occurred. The choice of r in such negative binomial charts is dictated by how much the failure rate is supposed to change during Out-of-Control. However, these results have been derived for the case of homogeneous data. Especially in health care monitoring, (groups of) patients will often show large heterogeneity. In the present paper we will show how such overdispersion can be taken into account. In practice, typically neither the average failure rate, nor the overdispersion parameter(s), will be known. Hence we shall also derive and analyze the estimated version of the new chart.

Keywords and phrases: Statistical Process Control, high-quality processes, geometric charts, average run length, estimated parameters, heterogeneity

2000 Mathematics Subject Classification: 62P10, 62C05, 62F12

1 Introduction and motivation

In this paper we consider high-quality processes, in which the proportion of nonconforming items is expected to be (very) small. First of all, due to constant efforts to improve quality in production, such a setup will be encountered more and more often in industrial settings. Moreover, in the quite different, but equally important, field of health care monitoring, this is in fact the standard situation: negative events (malfunctioning equipment, unsuccessful surgery, excessive delay before help arrives, detection of (the return of) a serious disease) should typically be (very) rare.

In review papers on health care monitoring (see e.g. Thor et al. (2007), Shaha (1995) and Sonesson and Bock (2003)), the use of SPC methods is strongly advocated, with special emphasis on control charts as the key tools. Now a standard choice for controlling attribute data is a p -chart, based on the number of failures in a series of given sampling intervals. However, for the really small proportions p we encounter in high-quality processes, substantial improvements can be achieved by applying a different type of chart, which goes by a variety of names, such as 'time-between-events' or 'geometric'. All such charts essentially employ the number of successes between failures, see e.g. Liu et al. (2004), Yang et al. (2002), Xie et al. (1998), Ohta et al. (2001) and Wu et al. (2001).

A known drawback of this geometric chart however is that it requires a rather long time to react to a moderate increase of the failure rate p . Only large deteriorations quickly produce an Out-of-Control (*OoC*) signal. Clearly, in particular for health care applications, this can

be quite unacceptable. Most of the authors quoted above therefore suggest as a remedy to essentially use a negative binomial chart: postpone the decision whether to stop until $r > 1$ failures have occurred. Some guidance on how to choose r in practice can be found in Ohta et al. (2001), but a systematic treatment of this issue was given in Albers (2008), resulting in a simple rule of thumb for choosing the optimal r as a function of the desired false alarm rate (FAR) and the supposed degree of increase of p compared to its value during In-Control (IC). As expected, the larger the increase one has in mind, the smaller r should be, with again the geometric chart ($r = 1$) as the ultimate result.

The second problem addressed in Albers (2008) concerns the estimation step involved. Note the general nature of this issue: typically, control charts have one or more unknown parameters which first have to be estimated on the basis of a so-called Phase I sample. Contrary to popular optimism, the effects of this estimation step are only negligible when (much) larger sample sizes are used than is customary in practice. Hence as a rule, such effects have to be taken into account and, if possible, corrections should be applied to the control limits to neutralize these. This program is indeed carried out in Albers (2008) for the negative binomial charts when p is unknown, and the result is a chart which is both simple to understand and to apply.

As such it thus offers a very satisfactory solution to the problem of monitoring high quality processes, characterized by an incoming sequence D_1, D_2, \dots , of independent identically distributed (i.i.d.) random variables (r.v.'s) with $P(D_1 = 1) = 1 - P(D_1 = 0) = p$, where p is (very) small. However, note the underlying homogeneity assumption, which is made explicit by this more formal description. For industrial processes this assumption usually is quite reasonable, although it will certainly not always be warranted. But in medical applications, patients will often show large heterogeneity, and we really have to take such variation between subjects into account on a rather regular basis.

Roughly speaking two types of situations should be distinguished. In the first, we essentially only know that such heterogeneity does occur. It is e.g. due to the existence of different subgroups, each with its own probability of failure, but we lack further information. The only way in which it becomes apparent, is through an increase of variance over what would be expected under the homogeneous model. This is the well-known phenomenon of overdispersion. See e.g. Poortema (1999) for a general review and Christensen et al. (2003) and Fang (2003) more specifically in connection with attribute control charts. The present paper will be devoted to demonstrating how negative binomial charts can be adapted to cover the overdispersion situation as well.

However, before addressing this issue, in passing we consider the second of the two situations mentioned above. Here we do have knowledge about the underlying structure. For example, incoming patients are classified into different risk categories, for each of which the corresponding p_i is known or can be estimated. This opens the possibility for so-called risk adjustment (see Grigg et al. (2004)): the base-line risk of each patient can be taken into account, thus allowing a more accurate appraisal of e.g. a surgeon's performance on a series of such patients. Clearly, this is an interesting option, giving rise to various questions. For what type of application is risk adjustment advisable, how should it be applied, what are the (typically larger!) estimation effects and how can these be controlled? As moreover the approach to be used will be quite different from what is needed in the overdispersion case, we prefer to treat risk-adjusted negative binomial charts in a separate, forthcoming paper.

In section 2 we introduce the negative binomial chart and describe its behavior in the homogeneous situation. Using this starting point, we demonstrate in section 3 how the extension to the overdispersion case can be made.

2 The homogeneous case

To introduce the ideas, as well as the notation involved, in this section we briefly (for full details and examples see Albers (2008)) consider the homogeneous case, where D_1, D_2, \dots , is a sequence of i.i.d. r.v.'s, with $P(D_1 = 1) = 1 - P(D_1 = 0) = p$ during *IC*. Once the process goes *OoC*, the failure probability p is replaced by θp for some $\theta > 1$ and a signal should follow as soon as possible. (Note that $\theta > 1$ is of primary interest, but a two-sided version can be derived in a completely similar way.). The 'time-between-events' approach means that we do not work with fixed-length blocks of D 's, but instead wait each time till the r^{th} failure occurs, for some $r \geq 1$. Let $X_i, i = 1, 2, \dots$ be the successive numbers of D 's involved, then these X_i clearly are i.i.d. copies of a negative binomial r.v. $X_{r,p}$ such that

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

where $k = r, r+1, \dots$. Unless confusion might occur, we suppress the indices whenever possible, here as well as in the sequel, and thus simply write X instead of $X_{r,p}$.

As $\theta > 1$, a signal should result when an r^{th} failure arrives too soon, i.e. at the first time an $X_i \leq n$, for some suitable lower limit $n = n_{r,p}$. In Albers (2008) it is argued that 'suitable' means that $F_{r,p}(n) = P(X_{r,p} \leq n) = r\alpha$, for some small $\alpha > 0$. In this way, the average run length (*ARL*) during *IC* will attain the same value $r/(r\alpha) = 1/\alpha$ for all r , thus allowing a fair comparison among the negative binomial charts for $r \geq 1$. It follows that $n = n_{r,p} = F_{r,p}^{-1}(r\alpha)$, the $r\alpha^{\text{th}}$ quantile of the negative binomial distribution function (df) $F_{r,p}$. (Either let n be the largest integer such that $F_{r,p}(n) \leq r\alpha$, or use standard interpolation.) Hence a numerical solution is easily obtained (e.g. using Maple), but to see how n behaves as a function of r, p and α , additional effort is needed.

To this end, in Albers (2008) use is made of the well-known relations

$$F_{r,p}(n) = P(X_{r,p} \leq n) = P(Y_{n,p} \geq r) \approx P(Z_{np} \geq r), \quad (2.2)$$

where $Y_{n,p}$ is a binomial r.v. with parameters n and p , while Z_{np} is a Poisson r.v. with parameter $\lambda = np$. The Poisson approximation in (2.2) requires n to be large, which will be the case for $r > 1$. (For $r = 1$, just use the explicit exact results $F_{1,p}(n) = 1 - (1-p)^n$ and $n = n_{1,p} = \log(1-\alpha)/\log(1-p)$.) Hence we observe that $n = n_{r,p} \approx \lambda/p$, with λ chosen such that $P(Z_\lambda \geq r) = r\alpha$. This leads to the next approximation step:

$$\tilde{n} = \frac{\tilde{\lambda}}{p}, \quad (2.3)$$

with $\tilde{\lambda}$ given by

Lemma 2.1. *Let $\alpha_r = (r!r\alpha)^{1/r}$, then λ such that $P(Z_\lambda \geq r) = r\alpha$ can be approximated for $p \leq 0.01, r \leq 5$ and $\alpha \leq 0.01$ by*

$$\tilde{\lambda} = \alpha_r(1 + \zeta_r), \quad \text{with } \zeta_r = \frac{\alpha_r}{r+1} + \frac{1}{2}\alpha_r^2 \frac{3r+5}{(r+1)^2(r+2)}. \quad (2.4)$$

Proof. See Albers (2008). □

As concerns the region of (p, r, α) covered by Lemma 2.1, note that p being small precisely is the point of departure for considering negative binomial charts and thus assuming $p \leq 0.01$ is

only natural. For r , interest in practice will focus on moderate values like $r \leq 5$, as continuing to wait for r failures tends to feel uncomfortable if r becomes too large. Finally, the upper bound 0.01 for α is amply sufficient: in this way values of $FAR = r\alpha$ as high as 0.05 can be reached. Since n is large, the error committed in the Poisson step is negligible and it does not matter at all whether $p = 0.01$ or e.g. $p = 0.0001$. In Albers (2008) it is demonstrated (see Table 2.1) that the approximation $\tilde{\lambda}$ from (2.4) is quite close to the 'exact' $\lambda^* = np = pF_{r,p}^{-1}(r\alpha)$ for the (α, r) from our region (and thus the phrase 'can be approximated' from Lemma 2.1 is validated). Hence application of the negative binomial chart for given (p, r, α) now has become very simple: just combine (2.3) and (2.4).

Next consider the *OoC* situation. Here we have $\theta > 1$ and thus

$$ARL = ARL_{r,\theta} = \frac{r}{F_{r,\theta p}(n_{r,p})} \approx \frac{r}{P(Z_{\theta\lambda} \geq r)}, \quad (2.5)$$

with λ such that $P(Z_\lambda \geq r) = r\alpha$. In analogy to Lemma 2.1 we obtain:

Lemma 2.2. *The exact ARL from (2.5) can be approximated for $p \leq 0.01$, $r \leq 5$, $\alpha \leq 0.01$ and $3/2 \leq \theta \leq 4$ by*

$$\widetilde{ARL} = \widetilde{ARL}_{r,\theta} = \frac{r}{1 - \exp(-\theta\alpha_r)[1 + \theta\alpha_r + \dots + \frac{(\theta\alpha_r)^{r-2}}{(r-2)!} + \frac{(\theta\alpha_r)^{r-1}(1-\theta\alpha_r\zeta_r)}{(r-1)!}]}, \quad (2.6)$$

with α_r and ζ_r as in (2.4).

Proof. See Lemma 3.1 from Albers (2008). □

The behavior of the $ARL_{r,\theta}$ from (2.5) is nicely illustrated by looking at

$$h_r = h_{r,\theta} = \frac{ARL_{1,\theta}}{ARL_{r,\theta}}, \quad (2.7)$$

These functions start at 1 for $\theta = 1$, then increase substantially before decreasing again towards the limiting value $1/r$. As expected, 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 (cf. Figure 3.2. in Albers (2008)). More specifically, for given α and θ , the value r^{opt} that minimizes ARL_r is adequately approximated by

$$\tilde{r}^{opt} = \frac{1}{\alpha(2.6\theta + 2) + 0.01(4\theta - 3)}, \quad (2.8)$$

(cf. Table 3.2 from Albers (2008)). The overall conclusion is that the major part of the improvement over the geometric chart usually is already achieved within the range $2 \leq r \leq 5$. Only when α and θ are both small, it might pay to go beyond $r = 5$. However, waiting for a too large number of failures before being allowed to stop, might be considered undesirable in practice anyhow.

As announced in the Introduction, it remains to deal with the estimation effects for the typical case of unknown p . Before the actual monitoring begins, a Phase I sample consisting of m geometric $X_{1,p}$'s (cf. (2.1)) is required now, leading for p to the standard estimator $\hat{p} = 1/\overline{X}$, with $\overline{X} = m^{-1}\sum_{i=1}^m X_i$. Hence the lower limit now becomes $\hat{n} = n_{r,\hat{p}} = F_{r,\hat{p}}^{-1}(r\alpha) \approx \lambda/\hat{p} = \lambda\overline{X}$, where λ is still such that $P(Z_\lambda \geq r) = r\alpha$. Combining this result with the approximation step from (2.3) produces $\tilde{n} = \tilde{\lambda}\overline{X}$, with $\tilde{\lambda}$ as in (2.4), and the chart can be easily applied again.

In Albers (2008) the impact of the estimation step is analyzed in some detail, so let us be

quite brief here. The performance characteristics now have become random: for each outcome \bar{x} of \bar{X} we have e.g. $\widehat{FAR} = FAR(\bar{x}) = P(X_{r,p} \leq \hat{n}|\bar{x})$, and hence in general a r.v.

$$\widehat{FAR} = FAR(\bar{X}) = P(X_{r,p} \leq \hat{n}|\bar{X}) \quad (2.9)$$

will result. Likewise, we have $ARL(\bar{X})$ instead of the fixed $ARL = 1/\alpha$. Writing $\bar{X} = (1/p)(1+U)$, with $U = p\bar{X} - 1$, and noting that $EU = 0$ and $EU^2 = (1-p)/m \approx 1/m$, Taylor expansion in powers of U readily reveals the extent to which e.g. $ARL(\bar{X})$ differs from $1/\alpha$. This can be done in terms of the bias $EARL(\bar{X}) - 1/\alpha$ (see Lemma 4.1 from Albers (2008)) or in terms of the exceedance probability $P(ARL(\bar{X}) < (1-\varepsilon)/\alpha)$, for some selected and small $\varepsilon > 0$ (see Lemma 4.3). Moreover, corrections can be derived to achieve either unbiasedness or a prescribed maximum exceedance probability. Just switch to a slightly more strict limit $n_{\varepsilon} = \hat{n}(1-c) = \lambda\bar{X}(1-c)$, for some small $c > 0$. Applying the expansions again, c can be easily selected to achieve these goals (see in Albers (2008) Lemma 4.2 for unbiasedness and Lemma 4.3 again for the exceedance case).

3 Overdispersion

From now on we drop the assumption of homogeneity, according to which the D_i were identically distributed. Instead, each D_i has its own p_i , but we have no further knowledge about the underlying mechanism. All that is clear is that overdispersion causes an inadequate fit for the single parameter homogeneous model. Hence we need to consider a larger parametric family, by at least adding one (overdispersion) parameter. Clearly, this wider family cannot be expected to be 'true' either: it also remains an approximation of the underlying unknown structure. But, being wider, it should provide a better approximation. Note that this modeling issue is not specific for the case at hand. For example, in the continuous case of controlling the mean of a process, lack of fit for the normality assumption inspired Albers, Kallenberg and Nurdyati (2004) to consider a wider parametric family than the normal one. In that case the additional parameter served to accommodate tail length, rather than overdispersion, but the idea is similar.

Bearing the above in mind, we proceed as follows. In the homogeneous case, stopping at the r^{th} failure led to the negative binomial $X_{r,p}$ from (2.1). To incorporate overdispersion, let P be a r.v. on $(0, 1]$ (or, more generally, on $(0, \infty)$, with $P(P > 1)$ negligible) such that

$$E\frac{p}{P} = 1, \quad \text{var}\frac{p}{P} = \tau, \quad (3.1)$$

where p is interpreted as the average failure rate and $\tau \geq 0$ is the overdispersion parameter. Typically, τ will be positive, but homogeneity (i.e. $\tau = 0$) is included as a boundary case. Next introduce $\tilde{X} = \tilde{X}_{r,p} = X_{r,P}$, i.e., given $P = p^*$, we have that $\tilde{X}_{r,p}$ is distributed as X_{r,p^*} . Clearly, $\tau = 0$ corresponds to the homogeneous case; in typical applications, τ will not be really large, but also not sufficiently small to be negligible. A straightforward calculation shows that

$$E\tilde{X}_{r,p} = EX_{r,p} = \frac{r}{p}, \quad \text{var}(\tilde{X}_{r,p}) = \text{var}(X_{r,p}) + \frac{r(r+1)\tau}{p^2} = \frac{r}{p^2}(1-p+\beta), \quad (3.2)$$

where $\beta = (r+1)\tau$. Hence the relative increase due to overdispersion is $\beta/(1-p) \approx \beta$, expressing the joint effect of the length of the waiting sequence and the variation in failure rates.

Next we use these new r.v.'s to extend the basic homogeneous model as follows: once again

we consider a sequence of i.i.d. r.v.'s, but now these will be copies of \tilde{X} rather than of X . In other words, for each 'time-between-events'-sequence of length r , a new realization of P is chosen independently. As already argued above in general terms, this is just a modeling step, without the intention of precisely grasping the true underlying structure. Acting as if the basic sequence of D_i 's conveniently selects a new value of P exactly if and only if an r^{th} failure occurs, clearly is a simplification of reality. The point is that it is a considerably less stringent simplification than assuming homogeneity.

The obvious advantage of the parameterization above is that it allows us to keep using the relations from (3.2). For, given $P = p^*$, we have $P(X_{r,p^*} \leq n) = P(Y_{n,p^*} \geq r) \approx P(Z_{np^*} \geq r)$, and thus taking expectations w.r.t. P leads to

$$P(\tilde{X}_{r,p} \leq n) \approx P(Z_T \geq r), \quad (3.3)$$

where Z is again Poisson and the r.v. $T = nP$. Note that with (3.3) we have arrived at a classical overdispersion setup: a Poisson r.v. Z with random parameter T .

For the next modeling step, by far the most prominent choice (see e.g. Poortema (1999)) is to let T be Gamma distributed, resulting in a (shifted) negative binomial r.v. Z_T . To be more precise, let $G(\zeta, \eta)$ denote the gamma distribution with density

$$f_G(x) = \frac{\eta^\zeta x^{\zeta-1} e^{-\zeta x}}{\Gamma(\zeta)}, \quad x > 0, \quad (3.4)$$

then we have for the present setup:

Lemma 3.1. *If T is $G(\zeta, \eta)$ then $Z_T + \zeta$ is $NB(\zeta, \eta/(\eta + 1))$. Moreover, $EZ_T = \zeta/\eta$ and $\text{var}(Z_T) = (\zeta/\eta)(1 + 1/\eta)$. Finally, for $\tau > 0$, let $v = 1 + \tau^{-1}$ and choose*

$$\zeta = v + 1, \eta = \frac{v}{np}, \quad (3.5)$$

then $P = T/n$ satisfies (3.1).

Proof. The first result is obtained directly through $P(Z_T = k) = \int_0^\infty P(Z_X = k) f_G(x) dx = \Gamma(\zeta + k)/\{k!\Gamma(\zeta)\}\{1/(\eta + 1)\}^k \{\eta/(\eta + 1)\}^\zeta$, $k = 0, 1, \dots$. Hence $Z_T + \zeta$ is distributed as the negative binomial $X_{\zeta, \eta/(\eta+1)}$ from (2.1). The moments of Z_T are also straightforward. Just observe the shift involved: $E(Z_T + \zeta) = \zeta(\eta + 1)/\eta$ and thus $EZ_T = \zeta/\eta$. For the final step, first note that $E(1/T) = \eta/(\zeta - 1)$, $E(1/T)^2 = \eta^2/\{(\zeta - 1)(\zeta - 2)\}$, and thus $\text{var}(1/T) = \{\eta/(\zeta - 1)\}^2/(\zeta - 2)$. As moreover $P = T/n$ is $G(\zeta, n\eta)$, it follows that

$$1 = E\frac{P}{P} = \frac{np\eta}{\zeta - 1}, \quad \tau = \text{var}\frac{P}{P} = \left(\frac{np\eta}{\zeta - 1}\right)^2$$

will hold if $1/(\zeta - 2) = 1/(v - 1) = \tau$ and $\eta = (\zeta - 1)/(np)$, i.e. if (3.5) is true. \square

If $\tau \rightarrow 0$ and thus $v \rightarrow \infty$, clearly $T \xrightarrow{P} np = \lambda$ from (2.2) and we are indeed back in the homogeneous case. Moreover, $E(P/p) = (1 + 2\tau)/(1 + \tau)$ and $\text{var}(P/p) = \tau(1 + 2\tau)/(1 + \tau)^2$, which illustrates that for τ small, P will generally be close to p . In addition, it follows that $P(P > 1) \leq \text{var}(P)/(1 - EP)^2 = O(\tau p^2)$. Since $p \leq 0.01$, this is indeed completely negligible for any given τ .

Using Lemma 3.1, we can now proceed from (3.3) by once again taking a step similar to the first one from (2.2), with Y indicating a binomial r.v.. If T is $G(\zeta, \eta)$, then

$$P(\tilde{X}_{r,p} \leq n) \approx P(Z_T \geq r) = P(X_{\zeta, \eta/(\eta+1)} \geq \zeta + r) = P(X_{\zeta, \eta/(\eta+1)} > \zeta + r - 1) = \quad (3.6)$$

$$P(Y_{\zeta+r-1, \eta(\eta+1)} < \zeta) = P(Y_{\zeta+r-1, 1(\eta+1)} > r-1).$$

If in addition (3.5) holds, it follows that in fact

$$P(\tilde{X}_{r,p} \leq n) \approx P(Y_{\zeta+r-1, 1/(\eta+1)} \geq r) = P(Y_{v+r, \lambda/(v+\lambda)} \geq r), \quad (3.7)$$

where $\lambda = np$. Clearly, if $\tau \rightarrow 0$, we have $v \rightarrow \infty$ and $Y_{v+r, \lambda/(v+\lambda)} \xrightarrow{P} Z_\lambda$ and we are back at the final step of (2.2). But of course, we shall not take such a step in (3.7): the difference between the binomial $Y_{v+r, \lambda/(v+\lambda)}$ and the Poisson Z_λ precisely reflects the overdispersion effect we want to quantify, and thus we shall want to hold on to it. In fact, we now look for $n_\tau = n_{\tau, r, p} \approx \lambda_\tau/p$, such that $P(Y_{v+r, \lambda/(v+\lambda)} \geq r) = r\alpha$ for $\lambda = \lambda_\tau$. Extending Lemma 2.1 we obtain as a final approximation step (cf. (2.3)):

$$\tilde{n}_\tau = \frac{\tilde{\lambda}_\tau}{p}, \quad (3.8)$$

with $\tilde{\lambda}_\tau$ given by

Lemma 3.2. *Let $\alpha_{r\tau} = v(r\alpha/\binom{v+r}{r})^{1/r}$, then λ such that $P(Y_{v+r, \lambda/(v+\lambda)} \geq r) = r\alpha$ can be approximated for $p \leq 0.01$, $r \leq 5$, $\alpha \leq 0.01$ and $\beta = (r+1)\tau \leq 1$ by $\tilde{\lambda}_\tau = \alpha_{r\tau}(1 + \zeta_{r\tau})$, with*

$$\zeta_{r\tau} = \alpha_{r\tau} \frac{v+r+1}{v(r+1)} + \frac{1}{2}(\alpha_{r\tau})^2 \left\{ \frac{(3r+5)(v+r+1)^2}{(r+1)^2(r+2)v^2} - \frac{(v+r+1)}{(r+2)v^2} \right\}. \quad (3.9)$$

Proof. This is a straightforward extension of the proof of Lemma 2.1, so we will be quite brief here. In Albers (2008), a result from Klar (2000) for Poisson probabilities is applied, which shows that the error committed by replacing $P(Z_\lambda \geq r)$ by $\sum_{j=r}^{r+2} P(Z_\lambda = j)$ is sufficiently small. But Klar (2000) contains a similar result for the binomial case, and this can be used here for $Y_{v+r, \lambda/(v+\lambda)}$ in precisely the same manner. The second step in the proof of Lemma 2.1 consists of expanding $\sum_{j=r}^{r+2} P(Z_\lambda = j)$ w.r.t. λ to third order. By equating the result obtained to $r\alpha$ and inverting w.r.t. λ , the expansion (2.4) follows. Again, the same procedure, be it a bit more laborious, can be applied here. To provide some details, note that after the first step we have the approximation

$$r\alpha = \binom{v+r}{r} \frac{\lambda^r v^v}{(v+\lambda)^{v+r}} \left\{ 1 + \frac{\lambda}{r+1} + \frac{(v-1)\lambda^2}{(r+1)(r+2)v} \right\}, \quad (3.10)$$

from which it is immediate that $(\alpha_{r\tau})^r = \lambda^r \{1 + O(\lambda)\}$, and thus $\lambda = \alpha_{r\tau}$ to first order. The refinement in (3.9) follows by solving (3.10) to third, rather than just first, order. \square

Note that in Lemma 3.2 the condition $\beta = (r+1)\tau \leq 1$ has been added to the broad area of interest for (p, r, α) from Lemma 2.1. This region for τ seems amply sufficient as well: after (3.2), it was remarked that the relative increase due to overdispersion is about β . Hence reaching this upper bound for τ means a doubling of the variance. Beyond this level, the overdispersion effect looks too strong to be accommodated by simply adapting the homogeneous approach, as we are proposing here. In such circumstances, it seems advisable to acquire more detailed information about the process at hand (cf. the remarks about risk adjusted charts from the Introduction).

Maybe it is useful to add a remark about the opposite end of the interval as well. For $\tau \rightarrow 0$,

and thus $v \rightarrow \infty$, comparison of Lemma's 2.1 and 3.2, in particular of (2.4) and (3.9), shows that $\alpha_{r\tau} \rightarrow \alpha_r$, $\zeta_{r\tau} \rightarrow \zeta_r$, $\tilde{\lambda}_\tau \rightarrow \tilde{\lambda}$, and hence $\tilde{n}_\tau \rightarrow \tilde{n}$ for the lower limits in (3.8) and (2.3), respectively. Hence for really small τ , the difference between the two models become negligible and the additional effort to accommodate overdispersion may no longer be worthwhile. Thus in addition to $\beta \leq 1$, a lower bound $\beta \geq \beta_0$ could be added, for some small value β_0 , like $\beta_0 = 0.05$. Since there is no technical necessity, we refrained from including it in Lemma 3.2, but the point will recur once estimation enters the picture in section 5.

For finite v , we observe that in fact $\alpha_{r\tau}/\alpha_r = v/\{(v+1)\dots(v+r)\}^{1/r} < 1$, and thus the leading term of $\tilde{\lambda}_\tau$ in (3.9) is smaller than the corresponding one from $\tilde{\lambda}$ in (2.4). Actually, a similar relation holds for the next coefficient in these expansions: $(\zeta_{r\tau}\alpha_{r\tau}^2)/(\zeta_r\alpha_r^2) = v(v+r+1)/\{(v+1)\dots(v+r)\}^{2/r} < 1$. To verify this last step, note that $v(v+r+1) < (v+r/2-s)(v+r/2+s+1)$ for $s < r/2$. Hence for r even, we immediately observe that $\{v(v+r+1)\}^{r/2} < \prod_{s=0}^{r/2-1} (v+r/2-s)(v+r/2+s+1) = (v+1)\dots(v+r)$. For r odd, a similar argument is used, together with the fact that $\{v(v+r+1)\}^{1/2} < v+r/2+1/2$. Hence $\tilde{\lambda}_\tau$ typically is smaller than $\tilde{\lambda}$, as should be the case, because overdispersion has a widening effect and thus forces us to lower the control limit \tilde{n}_τ in comparison to the result \tilde{n} from the homogeneous case.

Just as in Albers (2008), we now want to check the quality of the approximation $\tilde{\lambda}_\tau$ by comparing it to the solution $\lambda = \lambda_\tau$ of the equation

$$P(Y_{v+r,\lambda/(v+\lambda)} \geq r) = P(X_{r,\lambda/(v+\lambda)} \leq v+r) = r\alpha. \quad (3.11)$$

By way of illustration, first consider the geometric case $r = 1$, where a direct approach is feasible: here (3.11) boils down to $\alpha = P(X_{1,\lambda/(v+\lambda)} \leq v+1) = 1 - \{1 - \lambda/(v+\lambda)\}^{v+1}$. Hence $(1 + \lambda/v)^{-(v+1)} = 1 - \alpha$ and thus the solution $\lambda_\tau = v\{(1 - \alpha)^{-1/(v+1)}\} - 1$ is readily obtained. Indeed, expanding this expression leads to $\tilde{\lambda}_\tau = v\alpha/(v+1)\{1 + \frac{1}{2}(v+2)\alpha/(v+1) + (v+2)(2v+3)\alpha^2/[6(v+1)^2]\}$, which agrees with (3.9) for $r = 1$. In passing also observe the following. In the geometric case $r = 1$ we directly have that $P(\tilde{X}_{1,p} \leq n) = EP(X_{1,p} \leq n) = 1 - E(1 - P)^n = 1 - \sum_{k=0}^n \binom{n}{k} (-1)^k EP^k$, with $EP^k = p^k \{(v+1)\dots(v+k)\}/v^k$, as P is $G(v+1, v/p)$ -distributed (cf. (3.5)). Using the Poisson approximation subsequently gives $P(\tilde{X}_{1,p} \leq n) \approx EP(Z_{nP} \geq 1) = 1 - Ee^{-nP} = 1 - \{v/(v+np)\}^{v+1}$, which in its turn agrees with the result derived just above, using $P(X_{1,\lambda/(v+\lambda)} \leq v+1)$.

For $r > 1$, obtaining λ_τ is less straightforward. Let $\xi = \lambda/(v+\lambda)$, then for given ξ we have from (3.11) that $v+r = F_{r,\xi}^{-1}(r\alpha)$, the $r\alpha^{\text{th}}$ quantile of the negative binomial df $F_{r,\xi}$. (Here we shall use the interpolated version.) Consequently, we obtain

$$\lambda = \frac{\xi}{1-\xi} \{F_{r,\xi}^{-1}(r\alpha) - r\}, \quad \tau = \{F_{r,\xi}^{-1}(r\alpha) - r - 1\}^{-1}, \quad (3.12)$$

for given r , α and ξ . By adapting the value of ξ , selected values for $\beta = (r+1)\tau$ can be obtained iteratively in (3.12), and thus the corresponding $\lambda = \lambda_\tau$ as well. In Table 3.1 below some illustrative values are collected. Just as in Table 2.1 from Albers (2008), we use $\alpha = 0.001$, 0.005 and 0.001 . For the present purpose, the focus no longer is on comparing the various values of r , so we can restrict ourselves representative values like $r = 3$ and $r = 5$. The emphasis now is on the relative overdispersion increase β , for which we let the values range from 0 (homogeneous case) to 1 (upper limit in lemma 3.2).

Table 3.1. Comparison of the approximation $\tilde{\lambda}_\tau$ from (3.9) to λ_τ obtained through (3.11) for various α , r and β . The first value is λ_τ ; the second one is $\tilde{\lambda}_\tau$.

$r = 3$												
$\alpha \setminus \beta$	0		0.05		0.1		0.2		0.5		1	
0.001	.282	.281	.275	.275	.269	.269	.258	.258	.234	.234	.206	.206
0.005	.509	.506	.497	.496	.487	.486	.469	.467	.427	.425	.380	.378
0.01	.665	.660	.652	.647	.639	.634	.616	.611	.562	.557	.503	.497

$r = 5$												
$\alpha \setminus \beta$	0		0.05		0.1		0.2		0.5		1	
0.001	1.08	1.07	1.06	1.05	1.04	1.03	1.00	.99	.91	.90	.81	.80
0.005	1.62	1.58	1.59	1.55	1.57	1.52	1.52	1.47	1.40	.135	1.25	120
0.01	1.97	1.88	1.94	1.86	1.91	1.82	1.85	1.77	1.71	1.62	1.55	1.45

From Table 2.1 in Albers (2008) it was concluded that the approximation performs quite well over the region considered. Fortunately, Table 3.1 shows that this conclusion can be extended as well to the case of positive β , all the way to the upper limit 1. Note that another important result from Table 3.1 is the observation that the effect of β indeed can be considerable: as β grows, the resulting λ 's decrease quite a bit in comparison to the values for the homogeneous case $\beta = 0$. Remember once more that this decrease serves to accommodate the overdispersion effect and to maintain the value of FAR during IC at $r\alpha$. By way of illustration we show in Table 3.2 what happens to this FAR if the overdispersion is ignored and the λ for $\beta = 0$ is used while in fact β is positive.

Table 3.2. Realized FAR 's (in %) when using the homogeneous $\lambda = \lambda_0$ for various α , r and β .

r	3						5					
$\alpha \setminus \beta$	0	0.05	0.1	0.2	0.5	1	0	0.05	0.1	0.2	0.5	1
0.001	.300	.322	.341	.382	.501	.693	.500	.546	.590	.681	.973	1.49
0.005	1.50	1.59	1.68	1.85	2.34	3.07	1.50	2.68	2.85	3.20	4.21	5.83
0.01	3.00	3.16	3.32	3.62	4.50	5.75	5.00	5.30	5.58	6.14	7.76	10.1

Indeed, the realized FAR can be doubled, or even tripled, if overdispersion effects become substantial, thus producing on the average far too short runs during IC .

To illustrate that application of the resulting chart is still quite simple, we conclude this section with:

Example 3.1. Suppose an ARL of 200 is considered acceptable, i.e. $\alpha = 0.005$ is chosen. If we want to decide about stopping or continuing at each third failure, we should use $r = 3$. In the homogeneous case (cf. Example 2.1 from Albers (2008)), we used λ such that $P(Z_\lambda \geq 3) = 0.015$ here, leading to $\lambda = 0.509$ (or $\tilde{\lambda} = 0.506$). However, assume now that in fact $\tau = 1/8$, and thus $\beta = (r + 1)\tau = 1/2$. According to Table 3.2, using the homogeneous λ would produce $FAR = 0.0234$ rather than 0.0150. Hence we proceed by noting that $v = 1 + \tau^{-1} = 9$, and thus obtain λ_τ from solving $P(Y_{12,\lambda/(9+\lambda)} \geq 3) = 0.015$ (cf. (3.11)) or, more directly, $\tilde{\lambda}_\tau$ from (3.9), leading to $\lambda_\tau = 0.427$ and $\tilde{\lambda}_\tau = 0.425$ (cf. Table 3.1). To complete the example, fix a value of p as well, e.g. by letting $p = 0.001$. During IC , the third failure should then on average arrive after 3000 observations. In the homogeneous case, action is taken if this already happens after at most 509 (or 506) observations. Taking the overdispersion into account now actually lowers these limits to 427 (or 425) in the present case. \square

4 The OoC situation

In this section we let the process go *OoC*, in the sense that p is again replaced by θp , for some $\theta > 1$. In the homogeneous case we observed (cf. (2.5)) for this situation that $ARL = ARL_{r,\theta} = r/F_{r,\theta p}(n_{r,p}) \approx r/P(Z_{\theta\lambda} \geq r)$, where λ is such that $P(Z_\lambda \geq r) = r\alpha$. For the present case, this result can be adapted as follows. First we update (3.1) into $P(\tilde{X}_{r,\theta p} \leq n) \approx P(Z_{\theta T} \geq r)$. If T is $G(\xi, \eta)$, then θT is $G(\xi, \eta/\theta)$, from which we readily obtain through (3.6) and (3.7) that $P(\tilde{X}_{r,\theta p} \leq n) \approx P(Y_{v+r,\theta\lambda/(v+\theta\lambda)} \geq r)$. Consequently, under overdispersion we arrive at

$$ARL = ARL_{r,\theta} \approx \frac{r}{P(Y_{v+r,\theta\lambda/(v+\theta\lambda)} \geq r)}, \quad (4.1)$$

with λ such that $P(Y_{v+r,\lambda/(v+\lambda)} \geq r) = r\alpha$. Hence, just as in the homogeneous case, going out of control leads to replacement of the relevant λ by $\theta\lambda$. Not surprisingly, this means that Lemma 2.2 can be adapted in a straightforward manner to

Lemma 4.1. *The ARL from (4.1) can be approximated for $p \leq 0.01$, $r \leq 5$, $\alpha \leq 0.01$, $\beta \leq 1$ and $3/2 \leq \theta \leq 4$ by $\widetilde{ARL} = \widetilde{ARL}_{r,\theta,\tau} =$*

$$\frac{r}{1 - \frac{v}{(v+\theta\alpha_{r\tau})^{v+r}} \left[1 + \frac{\theta\alpha_{r\tau}(v+r)}{v} + \dots + \binom{v+r}{r-2} \left(\frac{\theta\alpha_{r\tau}}{v}\right)^{r-2} + \binom{v+r}{r-1} \left(\frac{\theta\alpha_{r\tau}}{v}\right)^{r-1} \left\{ 1 - \frac{\theta\alpha_{r\tau}\xi_{r\tau}(v+1)}{v+\theta\alpha_{r\tau}[1+\xi_{r\tau}]} \right\} \right]}, \quad (4.2)$$

with $\alpha_{r\tau}$ and $\xi_{r\tau}$ as in (3.9).

Proof. Apply the method of Lemma 3.1 from Albers (2008) to the relevant binomial rather than Poisson probabilities. \square

Clearly, as $\tau \rightarrow 0$, the result from (4.2) converges to the one in (2.6): Lemma 2.2 is in fact a boundary case of Lemma 4.1 (also cf. the relation between Lemma 2.1 and Lemma 3.2). The range of values of interest for θ obviously remains the same as in Albers (2008). Just as in that paper, we are interested in the quality of the approximation provided, but now the focus is on the behavior with respect to β . In Table 4.1 some illustrative values are collected, with α and r as in Tables 3.1 and 3.2 and θ as in Table 3.1 from Albers (2008). Since the behavior in β is again monotone (cf. Tables 3.1 and 3.2), we just present the boundary cases $\beta = 0$ and $\beta = 1$.

Table 4.1. Comparison of \widetilde{ARL} from (4.2) to ARL from (4.1) for various α , r , β and θ . In each 2×2 cell the upper values are ARL 's and the lower ones \widetilde{ARL} 's, while the left column is for $\beta = 0$ (homogeneity) and the right one for $\beta = 1$.

$\alpha \setminus \theta$		$r = 3$							
		3/2		2		3		4	
0.001		329	338	154	162	55.7	61.3	28.7	32.7
		332	344	155	164	56.2	62.1	28.9	33.0
0.005		71.2	74.5	36.0	39.1	15.1	17.5	9.04	10.7
		73.4	77.9	36.9	40.6	15.4	17.9	9.10	10.9
0.01		37.6	39.7	20.0	22.0	9.32	10.9	6.04	7.27
		39.3	42.2	20.7	23.3	9.47	11.2	6.06	7.37

		$r = 5$							
$\alpha \setminus \theta$	3/2		2		3		4		
0.001	203	224	73.7	88.0	22.2	29.1	11.6	15.7	
	233	160	82.1	69.6	23.5	25.4	11.8	14.4	
0.005	49.8	56.3	21.9	26.8	9.31	12.1	6.44	8.22	
	61.7	59.0	25.4	28.6	9.71	12.7	6.31	8.42	
0.01	28.2	32.1	13.9	17.0	7.12	8.96	5.60	6.74	
	36.3	39.4	16.2	20.2	7.21	9.87	5.30	7.05	

Several interesting observations can be made from Table 4.1. As expected, the required numbers of observations increase as β goes from 0 to 1. Do note that this fact should not be interpreted as a 'drawback' of the adjusted charts, in the sense that avoiding this adjustment would in fact have produced a lower *ARL* and thus a better *OoC* performance. From Table 3.2 it is evident that such an 'improvement' can only be obtained by cheating on the requirement that $ARL = 1/\alpha$ during *IC*. Nevertheless, it is gratifying to observe as well that the impact of changing β is much smaller under *OoC* than under *IC*. In the latter case, Table 3.2 shows that even tripling of the intended value can occur, while the relative increase in Table 4.1 is considerably smaller. Note that this phenomenon is of a general nature and by no means special for the present situation. In Albers and Kallenberg (2004) it was already pointed out that the fact that one is dealing during *IC* with very small probabilities, easily causes errors which may be small from an absolute point of view, but unpleasantly large when considered in a relatively sense. In addition, Table 4.1 shows that in general the approximation works well in the region considered, with again a decreasing quality as $r\alpha$ increases. Moreover, observe that for small α and θ at $r = 5$ the approximation no longer increases as β goes from 0 to 1, which also indicates that here the limits of its usefulness are reached.

Yet another conclusion is that the pattern with respect to the optimal choice of r for given θ obviously hardly changes in going from the homogeneous case $\beta = 0$ to the opposite end at $\beta = 1$. Consequently, there is no need to adapt the analysis from Albers (2008) at this point, and we can stick to the rule of thumb from that paper, quoted here in (2.8). In Table 3.2 of Albers (2008) it is demonstrated that this recommendation performs quite well. However, it is also remarked that comparison to Table 3.1 shows most of the gain has already been realized in the region $2 \leq r \leq 5$. Since working with too large values of r will often be considered undesirable anyhow in practice, the final conclusion is that generally speaking truncation of (2.8) to $r \leq 5$ will be just fine. Clearly, there is no reason to change this advice for the present overdispersion case. To illustrate matters, we conclude the present section with:

Example 4.1 Using Example 3.1 as a starting point, let once more $\alpha = 0.005$, $p = 0.001$ and $r = 3$. Homogeneity in this situation gave $\lambda = 0.509$ (or 0.506) and thus $n = 509$ (or 506). Suppose now that in fact $\tau = 1/4$, i.e. $\beta = 1$, then during *IC* this choice would actually produce $FAR = 3.07\%$, instead of 1.50%. Hence the corresponding *ARL* would be less than 100, instead of the intended 200. Consequently, we definitely prefer to repair this defect by lowering our limit to $n = 380$ (or 378). The price for this correction during *OoC* boils down at $\theta = 4$ to an increase in *ARL* from 9.04 to 10.7 (or from 9.10 to 10.9), which seems quite moderate. Even after correction, 3 to 4 blocks of 3 failures on the average will suffice for a signal to occur.

Next observe that (2.8) suggests $r = 5$ as optimal choice for $\alpha = 0.005$ and $\theta = 4$. Then the lower limit $n = 1620$ (or 1580) should be lowered to $n = 1280$ (or 1200), in order to avoid a rise of the *FAR* during *IC* from 2.50% to 5.83%. As a consequence, the *ARL* during *OoC* at $\theta = 4$ will rise from 6.44 to 8.22 (or from 6.31 to 8.42). Indeed some further improvement

over $r = 3$ is achieved: 1 to 2 blocks of 5 failures will now suffice on average.

Finally, to illustrate that most of the gain with respect to the geometric chart (i.e. $r = 1$) typically is achieved within the range $2 \leq r \leq 5$, note the following. The geometric chart has $ARL \approx 1/(\theta\alpha)$ (see (3.2) in Albers (2008)), which means an ARL of about 50 here. The step towards $r = 3$ gives the main reduction to 9.04, with a slight further improvement for $r = 5$ to 6.44. The latter two values are those for the homogeneous case. Accommodating overdispersion means a renewed increase to 10.7 and 8.22, respectively, which is very mild compared to the starting value of 50. Hence also in this respect, the price for correcting for overdispersion seems quite fair. \square

5 The estimated chart

Typically the underlying parameters of the chart will be unknown in practice. In the present setup not only the failure rate p involved, but also the overdispersion parameter τ from (3.1) (or equivalently, $\beta = (r + 1)\tau$ from (3.2) or $v = 1 + \tau^{-1}$ from (3.5)). Hence these will have to be estimated and a Phase I sample is needed before monitoring can start. Just as in Albers (2008), let m be the size of such a sample, in the sense that we observe the sequence D_1, D_2, \dots until m failures have been gathered. Note that m does not depend on the r we choose: in this way, also with respect to estimation, fairness in comparing charts for different r is preserved. Also observe that the r.v.'s involved are typically not simply distributed as $X_{r,p}$ from (2.1) for the homogeneous case, but also not necessarily as $\tilde{X}_{r,p}$ from (3.2), since this latter choice was proposed as a convenient modeling step (cf. the discussion in section 3). Hence we prefer to adopt the following general notation: for simplicity (and without essential loss of generality), let $k = m/r$ be an integer, then our Phase I sample consists of k r.v.'s $Y_{r,p}$. Here each $Y_{r,p}$ is an overdispersed waiting time till the r^{th} failure, so let us use here as well (cf. (3.2)) the notation

$$EY_{r,p} = \frac{r}{p}, \quad \text{var}(Y_{r,p}) = \frac{r}{p^2}(1 - p + \beta). \quad (5.1)$$

In this way, for both $Y_{r,p}$ and $\tilde{X}_{r,p}$, the relative increase due to overdispersion is denoted by $\beta/(1 - p) \approx \beta$.

For brevity's sake denote the k $Y_{r,p}$'s from Phase I by Y_1, \dots, Y_k and let

$$Y^* = m^{-1}\sum_{i=1}^k Y_i, \quad S_r^2 = (m - r)^{-1}\sum_{i=1}^k (Y_i - rY^*)^2. \quad (5.2)$$

Clearly, $Y^* = r^{-1}\bar{Y}$, with $\bar{Y} = k^{-1}\sum_{i=1}^k Y_i$, and thus Y^* is just the average waiting time till the first failure, with $EY^* = 1/p$. Moreover, $S_r^2 = r^{-1}\tilde{S}_r^2$, where $\tilde{S}_r^2 = (k - 1)^{-1}\sum_{i=1}^k (Y_i - \bar{Y})^2$, the sample variance of the Y_i 's. Consequently, $E\tilde{S}_r^2 = \text{var}(Y_1)\{1 - [k(k - 1)]^{-1}\sum_{i \neq j} \rho(Y_i, Y_j)\}$. Obviously, if the Y_i are distributed as in (2.1) (i.e. homogeneity holds after all), all correlations involved will be 0. More important, however, is the fact that this remains true if the Y_i are distributed according to (3.2), i.e. as $\tilde{X}_{r,p}$. Then not only all underlying D 's are independent, but also a new and independent P is drawn after each r^{th} failure.

Note that this observation indicates what will happen for general Y_i . Typically, the effect of the correlation terms in $E\tilde{S}_r^2$ will remain negligible, as the only contribution comes from carryover effects, due to carrying on for a while with the same p after an r^{th} failure. Only if the stretches involved are too large, problems will arise in this respect. However, as stated before, under such circumstances a closer scrutiny of the underlying process seems indicated (risk adjustment methods etc.). The present approach focuses on the simple setup where the information available essentially consists only of waiting times till r^{th} failures.

Hence we may assume that $ES_r^2 \approx r^{-1} \text{var}(Y_1) \approx (1 + \beta)/p^2$ (cf. (5.2)). Then it follows that $p = 1/EY^*$ and $\beta \approx ES_r^2/(EY^*)^2 - 1$, leading to the simple estimators

$$\hat{p} = \frac{1}{Y^*}, \hat{\beta} = \max(0, \frac{S_r^2}{(Y^*)^2} - 1), \quad (5.3)$$

and thus also to $\hat{\tau} = \hat{\beta}/(r + 1)$ and $\hat{v} = 1 + \hat{\tau}^{-1}$. The maximum in (5.3) has been included since nonpositive values of $S_r^2/(Y^*)^2 - 1$ can occur. However, this is a negligible complication, because it will typically only happen if the underlying β is really small. But, as remarked following Lemma 3.2, such β are not at all interesting and taking the trouble to accommodate the overdispersion effect can be reserved for e.g. $\beta \geq \beta_0 = 0.05$. Hence the proper reaction in practice to finding such a nonpositive value is to refrain from additional effort, i.e. to stick to the homogeneous approach described in section 2. That is precisely what (5.3) does: $\hat{\beta} = 0$ in that case.

Basically, the above is all that is needed to transform the chart into its estimated version: just replace p , β , τ and v in sections 3 and 4 by their estimated counterparts \hat{p} , $\hat{\beta}$, $\hat{\tau}$ and \hat{v} , respectively. For example, instead of the lower limit $n_\tau = \lambda_\tau/p$, with $\lambda = \lambda_\tau$ solving $P(Y_{v+r, \lambda/(v+\lambda)} \geq r) = r\alpha$, we now have

$$\hat{n}_\tau = \frac{\lambda_\tau^*}{\hat{p}}, \quad (5.4)$$

with $\lambda = \lambda_\tau^*$ such that $P(Y_{\hat{v}+r, \lambda/(\hat{v}+\lambda)} \geq r) = r\alpha$. Likewise, $\tilde{n}_\tau = \tilde{\lambda}_\tau/p$ from (3.8) becomes $\hat{\tilde{n}}_\tau = \tilde{\lambda}_\tau^*/\hat{p}$, where $\tilde{\lambda}_\tau^*$ is obtained from $\tilde{\lambda}_\tau$ in (3.9) by substituting \hat{v} for v everywhere in $\alpha_{r\tau}$ and $\zeta_{r\tau}$. Once such an estimated lower limit \hat{n}_τ (or $\hat{\tilde{n}}_\tau$) has been obtained from the Phase I sample, the actual monitoring can start again: each time we wait till the r^{th} failure, and if this occurs at or before this lower limit, a signal is given. Hence, straightforward application of the estimated chart remains easy.

However, it remains to note, just like in section 2 for the homogeneous case, that as a consequence of the estimation step the performance characteristics FAR and ARL will now be stochastic, rather than fixed at $r\alpha$ and $1/\alpha$, respectively. In analogy to (2.9), we e.g. have $\widehat{FAR} = FAR(Y^*, S_r^2) = P(\tilde{X}_{r,p} \leq \hat{n}_\tau | Y^*, S_r^2)$. In section 2 it was also remarked that for the homogeneous case the impact of the estimation step was analyzed in some detail in Albers (2008). (Even there not in full detail, i.e. with considerable mathematical rigor. That would have required a complete separate paper, such as in the continuous case of controlling the mean of a process: see Albers and Kallenberg (2004a, b)). To avoid repetition and undue lengthening of the present paper, we shall here be really brief.

Just as in Albers (2008), we include the possibility to apply a small correction c to the estimated limit \hat{n}_τ from (5.4). To this end, consider

$$\hat{n}_{\tau,c} = \hat{n}_\tau(1 - c). \quad (5.5)$$

Hence for $c = 0$ we again have the uncorrected case and $n_{\hat{\tau},0} = n_{\hat{\tau}}$. In addition, let $\widehat{FAR}_c = P(\tilde{X}_{r,p} \leq \hat{n}_{\tau,c} | Y^*, S_r^2)$ and

$$U = \frac{p}{\hat{p}} - 1, W = -\frac{(r - \lambda)(\hat{\beta} - \beta)}{(1 + \beta)(r + 1 + \beta)}, \quad (5.6)$$

then we have

Lemma 5.1. *To first order \widehat{FAR}_c equals*

$$r\alpha + r \frac{v}{v + \lambda} (U + W - c) P(Y_{v+r, \lambda/(v+\lambda)} = r), \quad (5.7)$$

in which $\lambda = \lambda_\tau$ solves $P(Y_{v+r, \lambda/(v+\lambda)} \geq r) = r\alpha$.

Proof. As $\lambda = \lambda_\tau^*$ is such that $P(Y_{\hat{v}+r, \lambda/(\hat{v}+\lambda)} \geq r)$ equals $r\alpha$ as well (cf. (5.4)), it follows that $(\hat{v} + r)\lambda_\tau^*/(\hat{v} + \lambda_\tau^*)$ to first order equals $(v + r)\lambda_\tau/(v + \lambda_\tau)$. Consequently, $\lambda_\tau^*/\lambda_\tau - 1 \approx (r - \lambda_\tau)(\hat{v} - v)/\{v(\hat{v} + r)\}$. Since $v = 1 + \tau^{-1}$, this expression transforms into $(r - \lambda_\tau)(\tau - \hat{\tau})/\{(1 + \tau)(1 + (r + 1)\hat{\tau})\}$. Using that $\beta = (r + 1)\tau$, it follows that $\lambda_\tau^*/\lambda_\tau - 1$ to first order equals W from (5.6). Next we obtain from (5.4) and (5.5) that $\hat{n}_{\tau, c}/n_\tau = \lambda_\tau^* p/(\lambda_\tau \hat{p})(1 - c) \approx 1 + U + W - c$. In view of (3.11), this implies that $\widehat{FAR} \approx P(Y_{v+r, \lambda/(v+\lambda)} \geq r)$, where now $\lambda = \lambda_\tau(1 + U + W - c)$. Since $\partial P(Y_{n, p} \geq r)/\partial p = (p/r)P(Y_{n, p} = r)$, a first order expansion around $\lambda_\tau/(v + \lambda_\tau)$ then produces the result in (5.7). \square

Indeed, for $\tau \rightarrow 0$, and thus $v \rightarrow \infty$, the expression (5.7) converges to $r\alpha + r\{U - c\}P(Z_\lambda = r)$, which agrees with the result for the homogeneous case from e.g. (4.6) in Albers (2008) (just note that $rP(Z_\lambda = r) = \lambda P(Z_\lambda = r - 1)$ and let $c = 0$).

The result in (5.7) can now be used to evaluate the exceedance probability $P(\widehat{FAR} > r\alpha(1 + \varepsilon))$ for the uncorrected case, and moreover to determine c such that, for some prescribed small δ

$$P(\widehat{FAR}_c > r\alpha(1 + \varepsilon)) \leq \delta. \quad (5.8)$$

In passing note that $P(\widehat{ARL} < (1 - \varepsilon)/\alpha) = P(r/\widehat{FAR} < (1 - \varepsilon)/\alpha) = P(\widehat{FAR} > r\alpha(1 + \tilde{\varepsilon}))$, where $\tilde{\varepsilon} = \varepsilon/(1 + \varepsilon)$. Hence control of \widehat{FAR}_c through (5.8) automatically provides that of \widehat{ARL}_c , and vice versa. Let u_δ be the upper δ -point of the standard normal d.f. Φ , i.e. $1 - \Phi(u_\delta) = \delta$, and denote the standard deviation of $(U + W)$ by $\sigma_{(U+W)}$. Finally, let

$$\gamma_\tau = \frac{v}{v + \lambda} \frac{P(Y_{v+r, \lambda/(v+\lambda)} = r)}{r\alpha}, \quad (5.9)$$

then we obtain along the lines of Lemma 4.3 from Albers (2008):

Lemma 5.2. *For γ_τ in (5.9) we have $1 - (v + r + 1)\lambda/\{(v + \lambda)(r + 1)\} < \gamma_\tau < v/(v + \lambda)$. Moreover*

$$P(\widehat{FAR} > r\alpha(1 + \varepsilon)) \approx 1 - \Phi\left(\frac{\varepsilon}{\gamma_\tau r \sigma_{(U+W)}}\right), \quad (5.10)$$

and equality in (5.8) is achieved by using $\hat{n}_{\tau, c}$ from (5.5) with

$$c = \sigma_{(U+W)} u_\delta - \frac{\varepsilon}{\gamma_\tau r}. \quad (5.11)$$

Proof. The result for γ_τ follows by once more using Klar (2000). Together, (5.7) and (5.9) imply that $\widehat{FAR}_c \approx r\alpha(1 + \gamma_\tau r\{U + W - c\})$. Hence the exceedance probability from (5.8) to first order equals $P(\gamma_\tau r\{U + W - c\} > \varepsilon)$. As $U + W$ is asymptotically normal with mean 0 and variance $\sigma_{(U+W)}^2$, this probability approximately equals $1 - \Phi(\{c + \varepsilon/(\gamma_\tau r)\}/\sigma_{(U+W)})$. For $c = 0$, this produces (5.10). If instead the prescribed δ should result, $c + \varepsilon/(\gamma_\tau r)$ has to equal $\sigma_{(U+W)} u_\delta$, and hence c should be chosen as in (5.11). \square

Once again, letting $\tau \rightarrow 0$ reproduces the results from the homogeneous case. In particular, $\gamma_\tau \rightarrow \gamma$ with $1 - \lambda/(r + 1) < \gamma < 1$ and $\sigma_{(U+W)} \rightarrow \sigma_U$, which for $\tau = 0$ simply equals $m^{-1/2}$ to first order.

In the present case, some effort is needed to obtain $\sigma_{(U+W)}$. The expressions involved are more complicated and an additional estimation step is required. In doing so, as before we assume that possible dependencies between the Y_i are negligible. For their marginal distribution, we might use that of $\tilde{X}_{r,p}$, and accordingly express the 3rd and 4th central moments involved in terms of r , p and τ . However, the resulting expressions are rather complicated. Moreover, simplification by using expansion w.r.t τ only works quite locally, as the coefficients of the higher order terms tend to grow considerably. But, apart from these technical aspects, it seems better anyhow not to rely on such an assumption and to just use moment estimators like $\hat{\mu}_j = k^{-1} \sum_{i=1}^k (Y_i - \bar{Y})^j$ for μ_j , $j = 3$ or 4 . Then we can proceed as follows: first note that W from (5.6) to first order can be written as $-a\{(1 + U^*)/(1 + U)^2 - 1\}$, where

$$a = \frac{r - \lambda}{r + 1 + \beta} \text{ and } U^* = \frac{p^2 \tilde{S}_r^2}{r(1 + \beta)} - 1. \quad (5.12)$$

Hence $U + W \approx (1 + 2a)U - aU^*$. From (5.6) and (5.3) it follows that $\sigma_U^2 = p^2 \text{var}(Y^*)$, which in view of (5.2) and (5.1) leads to $\sigma_U^2 = (p/r)^2 \text{var}(\bar{Y}) = (p/r)^2 (r/p^2) \{1 - p + \beta\}/k \approx m^{-1}(1 + \beta)$. Consequently, σ_U^2 can be estimated by $m^{-1}(1 + \hat{\beta})$. For $\text{Cov}(U, U^*)$ and $\sigma_{U^*}^2$ similar steps can be taken. We obtain that $\text{Cov}(U, U^*) \approx p^3/\{r^2(1 + \beta)\} \text{Cov}(\bar{Y}, \tilde{S}_r^2) = p^3/\{r^2(1 + \beta)\} \mu_3/k = m^{-1} p^3/\{r(1 + \beta)\} \mu_3$ and $\sigma_{U^*}^2 \approx p^4/\{r^2(1 + \beta)^2\} \text{Var}(\tilde{S}_r^2) = [p^4/\{r^2(1 + \beta)^2\} \mu_4 - 1]/k = m^{-1} [p^4/\{r(1 + \beta)^2\} \mu_4 - r]$. Hence $\sigma_{(U+W)}$ now readily follows, after which replacement of p , β and μ_j by their respective estimators gives the desired $\hat{\sigma}_{(U+W)}$. Note that $\hat{\sigma}_{(U+W)}$ still is of order $m^{-1/2}$, implying that the correction c from (5.11) will indeed be small if the Phase I sample size m is sufficiently large.

References

- Albers, W. (2008). Negative Binomial charts for monitoring high-quality processes. Techn. Report 1881, University of Twente.
- Albers, W. and Kallenberg, W. C. M. (2004a). Estimation in Shewhart control charts: effects and corrections. *Metrika* **59**, 207 -234.
- Albers, W. and Kallenberg, W. C. M. (2004b). Are estimated control charts in control? *Statistics* **38**, 67 - 79.
- Albers, W., Kallenberg, W.C.M. and Nurdiati, S. (2004). Parametric control charts. *J. Statist. Planning & Inference* **124**, 159 -184.
- Christensen, A., Melgaard, M., Iwersen, J., and Thyregod, P. (2003). Environmental Monitoring Based on a Hierarchical Poisson-Gamma Model. *J. Qual. Technol.* **35**, 275-285.
- Klar, B. (2000). Bounds on tail probabilities of discrete distributions. *Prob. Engin. & Inform. Science* **14**, 161-171.
- Fang, Y. (2003). c -Charts, X -Charts, and the Katz Family of Distributions. *J. Qual. Technol* **35**, 104-114.
- Grigg, O. and Farewell, V. (2004). An overview of risk-adjusted charts. *J. Royal Statist. Soc. A* **167**, 523-539.
- Liu, J. Y. Xie, M., Goh T.N. and Ranjan P. (2004). Time-Between-Events charts for on-line process monitoring. *Intern. Engin. Man. Conf.*, 1061-1065.
- Ohta, H., Kusukawa, E. and Rahim, A. (2001). A $CCC - r$ chart for high-yield processes. *Qual. & Reliab. Engin. Int.* **17**, 439-446.

- Poortema, K. (1999). On modelling overdispersion of counts. *Statist. Neerl.* **53**, 5-20.
- 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.
- 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.
- Wu, Z., Zhang, X. and Yeo, S.H. (2001). Design of the sum-of-conforming-run-length control charts. *Eur. J. of Oper. Res.* **132**, 187-196.
- Xie, M., Goh, T.N. and Lu, X.S. (1998). A comparative study of *CCC* and *CUSUM* charts. *Qual. Reliab. Eng. Int.* **14**, 339-345.
- Yang, Z., Xie, M., Kuralmani, V. and Tsui, K.-L. (2002). On the performance of geometric charts with estimated parameters. *J. Qual. Techn.* **34**, 448-458.