Negative Binomial charts for monitoring high-quality processes

Willem Albers

Department of Applied Mathematics

University of Twente

P.O. Box 217, 7500 AE Enschede

The Netherlands

Abstract. Good control charts for high quality processes are often based on the number of successes between failures. Geometric charts are simplest in this respect, but slow in recognizing moderately increased failure rates p. Improvement can be achieved by waiting until $r > 1$ failures have occurred, i.e. by using negative binomial charts. In this paper we analyze such charts in some detail. On the basis of a fair comparison, we demonstrate how the optimal r is related to the degree of increase of p. As in practice p will usually be unknown, we also analyze the estimated version of the charts. In particular, simple corrections are derived to control the non-negligible effects of this estimation step.

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

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

1 Introduction and motivation

For decades a lot of effort has been devoted to improving quality in production. As a result, nowadays we are often dealing with high-quality industrial processes in which the proportion of nonconforming products is really very small. Hence it makes sense to pay special attention to control methods which seriously exploit this aspect. This holds even more strongly as a similar situation is also very common in health care monitoring. For most applications in this area, the occurrence of some type of failure (a malfunctioning instrument, help which arrives too late) or discovery of some kind of defect (a potentially fatal disease, a congenital defect) should be a rare phenomenon indeed. For some recent review papers on health care monitoring, see e.g. Sonesson and Bock (2003) and Thor et al. (2007). As mentioned in the latter paper, control charts are a core tool in the application of statistical process control (SPC) to healthcare quality improvement. This sentiment is also expressed in Shaha (1995), where *SPC* is mentioned as one of the most powerful quality management tools, with control charts as most notable among these tools.

The traditional approach to monitoring the nonconforming proportion in attribute data is to use a p-chart, which is based on the number of events in a series of sampling intervals. However,

e-mail: w.albers@utwente.nl;
running title: Negative Binomial charts for monitoring high-quality processes.

by now it is well-known that for really small proportions p, it is decidedly better to use so-called time-between-events (TBE) charts (Liu et al. (2004)), which e.g. look at the number of successes between failures. Consequently, the term geometric (or exponential) chart is also used (Yang et al. (2002)). In fact, quite a variety of names occurs: Liu et al. (2004) use the term cumulative quantity control (CQC) chart, Xie et al. (1998) and Ohta et al. (2001) speak about cumulative count of conforming (CCC) charts, while Wu et al. (2001) prefer conforming run length (CRL) charts. Nevertheless, the key quantity in all these papers essentially is the number of successes between failures.

Several interesting questions arise while studying charts of this type. To begin with, it was soon recognized (see e.g. Yang et al. (2002)) that a geometric chart is unfortunately quite slow to pick up relatively mild deteriorations of the process. Only if p increases considerably, a signal is quickly given. Especially in health care monitoring, this can be quite unacceptable: a certain tiny rate of failure is considered unavoidable, but nonnegligible increases above this level should really not remain unnoticed. Several authors suggested a way towards solving this problem: rather than deciding after a single failure whether or not to stop, it is better to postpone this decision until r failures have occurred. Hence a negative binomial chart is used, with the geometric chart as a special case for $r = 1$. This type of extension is discussed by Liu et al. (2004) as a CQR-r chart, by Ohta et al. (2001) as a CCC-r chart, while Wu et al. (2001) use the term sum of conforming run lengths (SCRL) chart.

The question remains of course how to choose r. A partial answer is given by Ohta et al. (2001), by using a simplified optimal design method within a given profit function framework. However, a broader analysis of this topic would definitely be worthwhile and it is one of the aims of the present paper to provide such information. In this connection, it is quite useful to note that a similar issue arises in the context of continuous data where normal charts, such as Shewhart's $\overline{X}$chart, are used to monitor the mean of the underlying process. For this case, ample information is already given by Albers and Kallenberg (AK for short)(2006) and the set-up contained in that paper can be used here as well.

By way of qualitative introduction, let us sketch the continuous situation as follows. Typically, if such a normal process strongly goes out-of-control (OoC) (e.g. a shift occurs of d standard deviations and this $d \geq 3$), the Shewhart X-chart, which uses individual observations, is just fine and an OoC signal will occur after 1 or at most 2 step(s). But if the shift size d gets smaller, it becomes better to wait until r (with e.g. $r = 4$ or 5) observations have arrived and to subsequently apply the $\overline{X}$chart based on their mean. In AK (2006) it is demonstrated how such charts for different r can be compared in a fair manner, and subsequently which r is optimal for given d. As can be intuitively expected, this optimal r turns out to decrease in d. In other words, the lower the extent to which the process goes OoC, the larger r should be.

Not surprisingly, the conclusion from the previous paragraph will turn out to hold for the negative binomial charts considered here as well: the more p increases above the in-control (IC) level, the smaller r should be. Hence eventually, i.e. for a major distortion of the process, the geometric chart is optimal again. Of course, this is merely a qualitative description. In the sections after this Introduction we shall provide a detailed account. First, in section 2 we introduce the negative binomial chart and study its IC behavior. Section 3 will be devoted to what happens during OoC. In particular, guidelines for choosing r will be given.

After thus having dealt with the first interesting question concerning negative binomial charts, we subsequently address the second one. To begin with, suppose that the nonconforming propor-

tion p during IC is known. Then for each given r, a lower limit, say $n = n_r$, can be evaluated such that the event that r failures are reached within n observations, has some prescribed probability. When this event indeed occurs, a signal is given and thus the corresponding probability is the false alarm rate (FAR). Hence the value chosen will typically be quite small, e.g. between 0.001 and 0.01. Next, once p increases (i.e. the process goes OoC), it will become easier to fall below n and the alarm rate rises, as should be the case. However, the assumption that p is known is often unrealistic. Common practice then is to use a so-called Phase I sample in order to estimate p, and thus n as well. But, as both p and the intended FAR are quite small, the resulting relative errors are uncomfortably large, unless an unrealistically large Phase I sample is available, as was demonstrated for the geometric case by Yang (2002).

Hence here as well a question remains: how to deal with the estimation effects for the negative binomial chart, i.e. for $r > 1$? This is even more pertinent as we already saw above that the geometric chart is unsatisfactory unless p rises sharply during OoC. Note that the answer should be twofold: step one consists of assessing the severity of the estimation errors, while step two offers suitable corrections for the estimated n, in order to compensate these errors in a meaningful way. Once more, it is quite helpful to observe that for continuous data and normal charts completely similar issues arise. These have been addressed in considerable detail by AK (2004a, 2004b) for the case $r = 1$ and in AK (2008) for $r > 1$. A similar approach will be very useful here.

Again, by way of illustration, let us briefly describe the ideas already developed for these normal charts. In a Shewhart X-chart, popular control limits are $\mu \pm 3\sigma$, with μ and σ the underlying normal mean and standard deviation, respectively. In case of unknown parameters, a Phase I sample provides estimated limits $\hat{\mu} \pm 3\hat{\sigma}$. As a result, for the resulting estimated chart performance characteristics such as the FAR or the average run length (ARL) should be considered conditional on $(\hat{\mu}, \hat{\sigma})$, and thus these quantities are in fact random. Hence comparison of this FAR to the intended value, say FAR_0, or similarly of this ARL to some ARL_0, requires choosing a criterion. One possibility is to select expectation, as in AK (2004a), and to investigate e.g. the bias $E(FAR(\hat{\mu}, \hat{\sigma})) - FAR_0$. Another is to consider exceedance probabilities like $P(ARL(\hat{\mu}, \hat{\sigma}) < ARL_0/(1 + \varepsilon))$ for some small positive ε, as in AK (2004b).

The first step mentioned above then entails to investigate the sizes of the Phase I sample needed to obtain sufficiently small biases and exceedance probabilities. It turns out that quite large sizes may be necessary before acceptably small values result. Hence for samples sizes of a more moderate magnitude such as encountered in practice, a logical second step is to derive corrections to the estimated limits that enforce the charts to behave in an acceptable manner after all. To give an explicit illustration, let $ARL_0 = 1000$. Typically, the fluctuations of $ARL(\hat{\mu}, \hat{\sigma})$ around 1000 can be quite large, meaning that unpleasantly small values like 500 or less are by no means unlikely. Suppose we choose $\varepsilon = 0.25$, i.e. we want to control the exceedance probability $P(ARL(\hat{\mu}, \hat{\sigma}) < 800)$. In AK (2004b) it is then demonstrated how to find a small correction δ such that broadening $\hat{\mu} \pm 3\hat{\sigma}$ somewhat to $\hat{\mu} \pm (3 + \delta)\hat{\sigma}$ reduces this probability to an acceptable prescribed value like 0.20.

In section 4 we shall carry out this program for the negative binomial charts. Hence the impact will be studied of replacing n from the case of known p by an estimate $\hat{n}$. Moreover, corrections will be derived to control the resulting charts, both w.r.t. bias and exceedance probability.

2 The negative binomial chart

Suppose we want to monitor a process in which the incoming observations each have a small probability p (e.g. $p \leq 0.01$) of being nonconforming. More formally, consider a sequence $D_1, D_2, \ldots$, of independent identically distributed (i.i.d.) random variables (r.v.'s) with $P(D_1 = 1) = 1 - P(D_1 = 0) = p$. However, we need to face the possibility that this IC situation only holds till a certain, unknown, point in this sequence. There the process goes OoC, in the sense that p is replaced by θp for some $\theta > 1$. The purpose of our monitoring clearly is to pick up this change as quickly as possible. (Note that we concentrate on the one-sided case $\theta > 1$, which seems to be of primary interest. But the two-sided case $\theta \neq 1$ can be dealt with in a completely similar manner.)

The traditional approach would be to consider blocks of incoming observations and give a signal if the fraction nonconforming in such a block is deemed too large. As argued in the Introduction, for small p it is better to use the TBE approach. Hence we use $D_1, D_2, \ldots$ to define a new sequence $X_1, X_2, \ldots$. Here X_1 is the number of D_i observed when the $r-th$ nonconforming item occurs, for some given $r \geq 1$. After this point we wait anew till r such failures have occurred and denote the corresponding number of D_i involved by X_2, etc. Clearly, the X_i are i.i.d. copies of a negative binomial r.v. $X_{r,p}$ such that

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

where $k = r, r+1, \ldots$. If no confusion is likely, we will simply write X instead of $X_{r,p}$. In fact, we will use this convention as well for other quantities to be introduced below: indices will only be dragged along when necessary or illuminating.

Since we are concerned about the case $\theta > 1$, a signal should be produced when too few successes proceed the occurrence of an $r-th$ failure. In other words, we should stop as soon as an $X_i \leq n$ for some suitably chosen lower limit $n = n_{r,p}$. In view of (2.1), to obtain this n it now only remains to specify an intended value FAR_0 and to solve $F_{r,p}(n) = P(X_{r,p} \leq n) = FAR_0$. But note that one aspect still needs attention: the choice of this FAR_0 should be made in such a way that the charts for various r can be compared in a meaningful and fair way. In line with AK (2006), we apply the following simple approach. The larger r, the longer it takes before the possibility of producing a signal arises. This problem is mentioned e.g. by Ohta (2001): on the one hand, increasing r leads to higher sensitivity for detecting a moderate rise in p, but on the other hand, the cost is higher, as more observations are needed.

Hence, to compensate for this effect, FAR_0 should be made to increase in r as well: if a longer wait is required to reach a possible stopping point, this can be balanced by allowing a larger probability that stopping then indeed happens. Take the case $r = 1$, i.e. the geometric chart, as a starting point and denote its intended FAR by α. Consequently, its corresponding ARL will be $1/\alpha$ and a fair way of matching matters is to impose this value for $r > 1$ as well. Clearly, as these charts take r steps at a time, we should simply set $FAR_0 = r\alpha$ in order to obtain $ARL_0 = r/(r\alpha) = 1/\alpha$ again. In this way, the charts for various r are comparable in IC behavior and thus it makes sense, as we will do in Section 3, to compare their performance for $\theta > 1$.

The matching obtained in this way is simple and intuitively attractive, and consequently quite optimal. But of course, differences do remain. Mainly this concerns the 'blocking effect': for a given r, one has to wait till the full block of r failures has finished. Especially for larger r, this can be considered to some extent as a drawback. Hence in practice there may be a reluctance to let r

grow too much, even if this looks promising from the point of view of discriminatory power. Quite often, values like $r = 3$, 4 or 5 may be felt to offer a sound compromise. Once again, essentially nothing new happens here: in the continuous situation with Shewhart charts completely similar considerations play a role. Groups of size 3-5 are popular in that context as well and in AK (2006) procedures using such values of r are providing the motivating examples. Having thus explained the background of our choice, we shall now investigate it in more detail. As a starting point we have for our lower limit the result

$$n = n_{r,p} = F_{r,p}^{-1}(r\alpha), \tag{2.2}$$

i.e. n equals the $r\alpha - th$ quantile of the negative binomial distribution function (df) $F_{r,p}$. (Of course, as $F_{r,p}$ is discrete, there is the usual element of choice involved. Either we let n be the largest integer such that $F_{r,p}(n) \leq r\alpha$, or we use standard interpolation to solve (2.2) exactly. We shall pay attention to such standard details only when it is unavoidable). In principle, (2.2) is all we need, as it provides (through e.g. Maple) the exact solution for the lower limit n for each given r, p and α. However, this result is not very helpful in understanding how n behaves as a function of the underlying parameters. For that purpose, further analysis is needed, involving suitable approximations. Here the term 'suitable' contains two elements. First of all, the approximations should be transparent, i.e. shed light on the relation between n and the triple (r, p, α). Moreover, they should also be accurate, a property which obviously cannot hold over a completely arbitrary region of parameter values.

Hence we shall first specify what parameter values will be considered. For r we have $r = 1$ as the boundary value from the geometric chart on the one hand, and $r > 1$ as the competing negative binomial ones on the other. As explained above, in principle r can attain arbitrary integer values, but some restraint will follow from practical considerations. Consequently, some emphasis on values from 2-5 will seem natural. About p the main observation is that it should be small, as this is the motivating reason for considering TBE-charts rather than traditional p-charts. We already mentioned $p \leq 0.01$, but this is mainly to fix ideas. Anyway, as long p is small, its actual value will only have a marginal effect on the accuracy of the approximations. Hence $p = 0.01$ is fine, but the same holds for e.g. $p = 0.0001$. For the basic geometric FAR value α, the situation is different: typically, the smaller α, the better the approximation. In the context of the normal control chart, customary values for a one-sided FAR are of the order of magnitude 0.001 (e.g. $0.00135 = 1/740$ as the exceedance probability in the traditional '3σ-bound'). Such values can be used here as well, but is also conceivable that, in view of the long waiting times encountered for very small p, somewhat larger α also are of interest. As an upper bound we therefore propose to use 0.01. Note that this is really quite large if we combine it with e.g. $r = 5$. According to (2.2), we then already have a 5% probability of stopping during IC. It does not seem very useful to go beyond this level.

Summarizing, we will let $p \leq 0.01$, (typically) $r \leq 5$ and $\alpha \leq 0.01$. In this region, adequate approximations are feasible. First take a separate look at the easy case $r = 1$. As $F_{1,p}(n) = 1 - (1-p)^n$, it is immediate that (2.2) boils down to

$$n = n_{1,p} = \frac{\log(1-\alpha)}{\log(1-p)}. \tag{2.3}$$

As -log$(1-x) \doteq x + x^2/2 + x^3/3$ for $x \to 0$ (where $\doteq$ stands for equality up to the given order),

approximate results such as $n \doteq (\alpha + \alpha^2/2 + \alpha^3/3)/p \doteq (\alpha + \alpha^2/2)/p \doteq \alpha/p$ are immediate from (2.3).

For $r > 1$, we use the well-known relations

$$F_{r,p}(n) = P(X_{r,p} \leq n) = P(Y_{n,p} \geq r) \approx P(Z_{np} \geq r), \tag{2.4}$$

where $Y_{n,p}$ is a binomial r.v. with parameters n and p, while Z_{np} is a Poisson r.v. with parameter $\lambda = np$. Hence, in addition to the above mentioned $F_{1,p}(n) = 1 - (1-p)^n$, we now also have results like $F_{2,p}(n) = 1 - (1-p)^n - np(1-p)^{n-1} \approx 1 - \exp(-\lambda)[1+\lambda]$. Moreover, note that n will typically be large, as required for the Poisson step in (2.4). An exception may occur when $r = 1$: from (2.3) it is evident that in the given parameter range small $n_{1,p}$ can arise (e.g. $\alpha = p = 0.001$ gives $n_{1,p} = 1$). But clearly this is no problem, as (2.3) already provides the explicit exact answer for this geometric case. Hence in general we will be able to use

$$n = n_{r,p} \approx \frac{\lambda}{p}, \tag{2.5}$$

where λ is such that $P(Z_\lambda \geq r) = r\alpha$.

Of course, the exact λ involved can easily be obtained numerically, but that in itself is not very interesting, as we could have used (2.2) for this purpose straightaway. The use of (2.5) is that it opens the way for further analysis of the behavior of n. Typically, $r\alpha$ should be small and thus λ should be only a small fraction of r, i.e. $\lambda = \eta r$ with η small and thus $n \approx \eta r/p$. If r is not too large, λ itself will be small as well and hence the Poisson probability involved admits further approximation steps. These we collect in the following result:

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*

$$\tilde{\lambda} = \alpha_r(1 + \zeta_r), \text{ with } \zeta_r = \frac{\alpha_r}{r+1} + \frac{1}{2}\alpha_r^2 \frac{3r+5}{(r+1)^2(r+2)}. \tag{2.6}$$

Proof. From Klar (2000) we have that for $k \geq 1$ and $r > \lambda - 1$

$$1 - \frac{\lambda^k}{\Pi_{j=1}^k (r+j)} < \frac{\Sigma_{j=r}^{r+k-1} P(Z_\lambda = j)}{P(Z_\lambda \geq r)} < 1. \tag{2.7}$$

Hence for $k = 3$, this ratio lies between $\{1 - \lambda^3/\Pi_{j=1}^3 (r+j)\}$ and 1. Since we aim at situations $\lambda = \eta r$ with η small, this typically means that the ratio from (2.7) is sufficiently close to 1 to allow us to solve

$$\frac{e^{-\lambda}}{r!}\lambda^r \left(1 + \frac{\lambda}{r+1} + \frac{\lambda^2}{(r+1)(r+2)}\right) = r\alpha \tag{2.8}$$

rather than $P(Z_\lambda \geq r) = r\alpha$. In addition note that, as $P(Z_\lambda \geq r)$ is increasing in λ, the solution from (2.8) will provide an upper bound for the true λ. The second step involves expanding $\exp(-\lambda)$: as $|\exp(-\lambda) - (1 - \lambda + \frac{1}{2}\lambda^2)| \leq \lambda^3/6$ for $\lambda > 0$, the error involved here will also be acceptable for small λ. Hence (2.8) leads to e.g. the first order result $\lambda^r/r! \doteq r\alpha$ and thus to

$\lambda \doteq \alpha_r$; using expansion to third order w.r.t. λ and inverting the result w.r.t. α_r produces (2.6) in a straightforward manner. □

Hence in addition to the exact result for n from (2.2) we now have, in view of (2.5) and Lemma 2.1, the approximation

$$\tilde{n} = \frac{\tilde{\lambda}}{p}, \tag{2.9}$$

with $\tilde{\lambda}$ as given in (2.6). For the boundary case $r = 1$, we simply find $\alpha_r = \alpha$ and (2.9) reduces to $\tilde{n} = (\alpha + \alpha^2/2 + \alpha^3/3)/p$ (cf. the result after (2.3)). However, note that α_r sharply increases in r for given α : e.g. let $\alpha = 0.01$, then $\alpha_2 = 0.20$ and $\alpha_4 = 0.99$. Consequently, n will indeed be large as soon as $r > 1$, which means that the error due to the Poisson step (cf. (2.4) and (2.5)) will be negligible for all p involved. Hence the actual value of p (as long as it is at most 0.01), will play almost no role as far as the approximation quality is concerned and in studying the behavior of the negative binomial chart we can focus on comparing $\tilde{\lambda}$ from (2.6) for various (α, r) to the 'exact' $\lambda^* = np = pF_{r,p}^{-1}(r\alpha)$ obtained from (2.2). In Table 2.1 below some illustrative values are collected. By way of boundary values, $\alpha = 0.001$ and $\alpha = 0.01$ were mentioned before; here we add $\alpha = 0.005$ as an intermediate value. In principle no upper bound exists for r, but on practical grounds, as discussed before as well, we stop after $r = 5$.

Table 2.1. Comparison of the approximation $\tilde{\lambda}$ from (2.6) to $\lambda^* = pF_{r,p}^{-1}(r\alpha)$ (cf. (2.2)) for various α and r. The first value is λ^*; the second one is $\tilde{\lambda}$.

α \ r	1		2		3		4		5	
0.001	0.001	0.001	0.065	0.065	0.281	0.281	0.631	0.628	1.08	1.07
0.005	0.005	0.005	0.149	0.148	0.508	0.506	1.02	1.00	1.62	1.58
0.01	0.01	0.01	0.215	0.213	0.665	0.660	1.27	1.24	1.97	1.89

From Table 2.1 we see that the approximation works quite well in the region considered and thus the statement from Lemma 2.1 that λ 'can be approximated' is substantiated. Moreover, its quality decreases as $r\alpha$ increases, which is of course evident from the two approximation steps applied in Lemma 2.1. Consequently, for small values of α, like 0.001, values of r beyond 5 could be used as well. Application of the negative binomial chart now has become very simple, as the following example illustrates:

Example 2.1. 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 $r = 3$. Thus we need λ such that $P(Z_\lambda \geq 3) = 0.015$, leading to $\lambda = 0.508$ and $\tilde{\lambda} = 0.506$ (hence $\eta = \lambda/r = 0.169$ is indeed small here). If p is the supposed IC-value of the process, the lower limit to be used then is $n = 0.508/p$, with $\tilde{n} = 0.506/p$ as the approximation. Although the role of p is rather trivial, let us for completeness' sake select a value for it as well, e.g. $p = 0.001$. Consequently, the third failure is expected during IC after about 3000 observations and a signal should be produced if it already arrives after at most 508 (or, in approximation, 506) observations. □

3 The *OoC* situation

As discussed in the previous section, at some unknown point in the sequence $D_1, D_2, \ldots$, the process may go OoC, in the sense that $p = P(D_i = 1)$ is replaced by θp for some $\theta > 1$. Just as we did for r, p and α, we shall first figure out what range of values of θ is of primary interest. As mentioned in the Introduction, especially in health care monitoring it is important to be able to pick up non-negligible increases above the level p from IC. Hence a lower value for this range such as $\theta = 3/2$ or $\theta = 2$ seems reasonable. Setting an upper value is somewhat arbitrary: on the one hand, a really large θ may be felt to represent an unrealistically large disturbance of the process. On the other hand, eventually the geometric chart becomes optimal again and some curiosity exists about the actual size of θ required for this event to happen. So let us choose e.g. $\theta = 4$ as a typical upper value of practical interest, but allow occasional excursions beyond this value. (Compare the similar 'soft' restriction on r: typically we focus on $r \leq 5$.)

During OoC the probability of a signal is given by $F_{r,\theta p}(n_{r,p})$, and therefore

$$ARL = ARL_{r,\theta} = \frac{r}{F_{r,\theta p}(n_{r,p})}. \tag{3.1}$$

In view of (2.2), all charts start at $ARL_{r,1} = 1/\alpha$. At the opposite end, we have as a lower limit $ARL_{r,1/p} = r$, which illustrates that for very large θ it is definitely better to take small r. Also observe that for $r = 1$ the result from (3.1) in view of (2.3) boils down to

$$ARL_{1,\theta} = \frac{1}{1-(1-\alpha)^{\log(1-\theta p)/\log(1-p)}} \doteq \frac{1}{1-(1-\alpha)^{\theta}} \doteq \frac{1}{\theta\alpha}. \tag{3.2}$$

Hence for the geometric chart the rate at which the $ARL = 1/\alpha$ decreases simply equals (to first order) the rate at which p increases. To appreciate that this is really quite slow, once more look at the continuous normal case. The '3σ-chart' mentioned before has a one-sided $ARL = 740$, which is already lowered to $ARL = 44$ for a moderate shift $d = 1$ (and even to $ARL = 2$ for a large shift $d = 3$). This decrease corresponds to a factor 17 (or even to 370), which in terms of the present θ would mean a major change indeed.

Consequently, there is ample reason to also consider ARL's for $r > 1$. However, for this case, it is no longer as straightforward as in (3.2) to interpret (3.1). Of course, (3.1) readily admits numerical computation, but the individual outcomes are even less illuminating than those of (2.2). Hence application of suitable approximations is once more in order. Let Z_μ denote a Poisson r.v. with parameter μ, then in analogy to (2.5) we have that

$$ARL \approx \frac{r}{P(Z_{\theta\lambda} \geq r)}, \tag{3.3}$$

with λ such that $P(Z_\lambda \geq r) = r\alpha$. An immediate consequence from (3.3) is the following. Since $\partial\{P(Z_{\theta\lambda} \geq r)/r\}/\partial\theta = P(Z_{\theta\lambda} = r)/\theta$, it follows that the derivative w.r.t. θ of the right-hand side of (3.3) at $\theta = 1$ equals -$r^2P(Z_\lambda = r)/P^2(Z_\lambda \geq r)$. In view of (2.7), this quantity lies between $-r/\alpha$ and $-r\{1-\lambda/(r+1)\}/\alpha$, which indicates that the rate of decrease of $ARL_{r,\theta}$ for small θ can indeed be greatly improved by using larger r.

To obtain more detailed information, we adopt the approach from Lemma 2.1.

Lemma 3.1. *The exact ARL from* (3.1) *can be approximated for* $p \leq 0.01$, $r \leq 5$, $\alpha \leq 0.01$ *and* $3/2 \leq \theta \leq 4$ *by*

$$A\tilde{R}L = A\tilde{R}L_{r,\theta} = \frac{r}{1 - \exp(-\theta\alpha_r)\{1 + \theta\alpha_r + \ldots + (\theta\alpha_r)^{r-1}\frac{1-\theta\alpha_r\zeta_r}{(r-1)!}\}} \tag{3.4}$$

with α_r *and* ζ_r *as in* (2.6).

Proof. From Lemma 2.1 it follows that in (3.3) we can replace λ by $\tilde{\lambda} = \alpha_r(1 + \zeta_r)$ from (2.6). Since $dP(Z_\mu \geq r)/d\mu = P(Z_\mu = r - 1)$, we have that $P(Z_{\mu(1+\zeta_r)} \geq r) \doteq P(Z_\mu \geq r) + \zeta_r\mu P(Z_\mu = r - 1) = 1 - \exp(-\mu)[1 + \mu + \ldots + \mu^{r-1}(1 - \zeta_r\mu)/(r - 1)!]$. Application of this result with $\mu = \theta\alpha_r$ immediately produces (3.4). □

In Table 3.1 illustrative values for the region covered by Lemma 3.1 are presented.

Table 3.1. Comparison of the approximation $A\tilde{R}L$ from (3.4) to the exact ARL from (3.1) for various α, r and θ. The upper value is ARL; the lower one is $A\tilde{R}L$.

		$\theta = 3/2$				$\theta = 2$			
α \ r	2	3	4	5	2	3	4	5
0.001	459	330	253	203	264	154	102	73.7
	454	332	266	233	261	155	106	82.2
0.005	93.6	71.3	58.2	49.8	55.3	36.1	26.8	21.9
	93.5	73.4	64.5	61.7	55.2	36.9	29.0	25.4
0.01	47.8	37.6	31.8	28.2	28.8	20.0	16.0	13.9
	47.9	39.2	36.2	36.3	28.8	20.7	17.5	16.2

		$\theta = 3$				$\theta = 4$			
α \ r	2	3	4	5	2	3	4	5
0.001	122	55.9	32.3	22.2	71.6	28.8	16.2	11.6
	121	56.2	33.3	23.5	71.0	28.9	16.5	11.8
0.005	27.0	15.2	11.0	9.31	16.7	9.04	6.90	6.44
	27.0	15.4	11.4	9.71	16.7	9.10	6.93	6.31
0.01	14.7	9.32	7.58	7.11	9.43	6.04	5.37	5.60
	14.7	9.47	7.77	7.21	9.43	6.06	5.29	5.30

Just as in Table 2.1, we may conclude that the approximation works quite well in the region considered, with a decreasing quality as $r\alpha$ increases. Hence the remark 'can be approximated' in Lemma 3.1 is justified. Moreover, it is evident that increasing r indeed leads to large improvements compared to the geometric case, where ARL virtually equals $1/(\theta\alpha)$ (cf. (3.2). By way of illustration we consider an explicit example.

Example 3.1. Let $\alpha = 0.005$, i.e. under IC all charts involved have $ARL = 200$. Suppose we are interested in recognizing a doubling of the value of p from IC, i.e. the case $\theta = 2$. If this occurs, the ARL of the geometric chart merely goes down to 100. However, using $r = 3$ gives a value 36.1, while $r = 5$ even leads to 21.9. Hence the individual chart on the average carries on for another 100 steps, while the grouped ones mentioned here stop after on average about 12 blocks of size 3,

or after 4 to 5 blocks of size 5. This same conclusion is reached if the approximations $A\tilde{R}L$ are used instead. □

Having demonstrated that increasing r is very worthwhile, it remains to provide further guidance on how to actually choose r. To this end, first consider for $r \geq 2$ the functions

$$h_r = h_{r,\theta} = \frac{ARL_{1,\theta}}{ARL_{r,\theta}}, \tag{3.5}$$

which nicely illustrate the relative gain that can be achieved by taking $r > 1$. It may be useful to recall here our convention of making explicit only those subscripts which are relevant at the given point. In view of (3.1) and (2.2), for $h_r = h_{r,\theta}$ from (3.5) also α and p play a role, so actually we have $h_{r,\theta,\alpha,p}$. As argued before, the impact of p is very marginal, an illustration of which is offered in Figure 3.1: only for very large θ, minor differences become visible. Hence in what follows, p will remain invisible again, i.e. the Poisson step is taken for granted.

A typical picture of h_r for various r is presented in Figure 3.2. After starting at 1 for $\theta = 1$, there is a substantial increase of h_r before the decline sets in 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. Nevertheless, it still takes quite long before h_r hits 1 again, i.e. the geometric chart start to dominate. E.g. with $\alpha = 0.01$, this occurs for h_5 at about $\theta = 22$, while h_3 requires $\theta = 40$ and h_2 even needs $\theta = 68$. Moreover, smaller α will produce still larger θ.

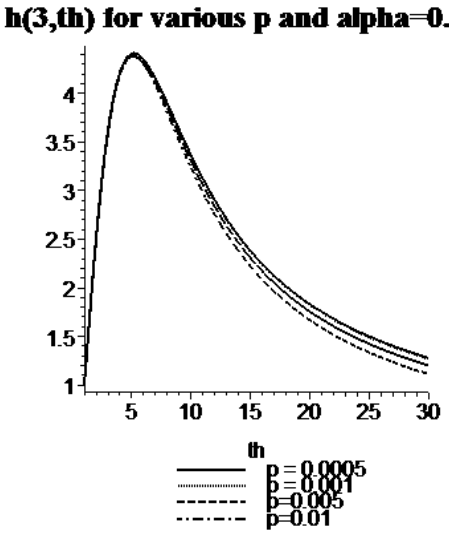

Figure 3.1

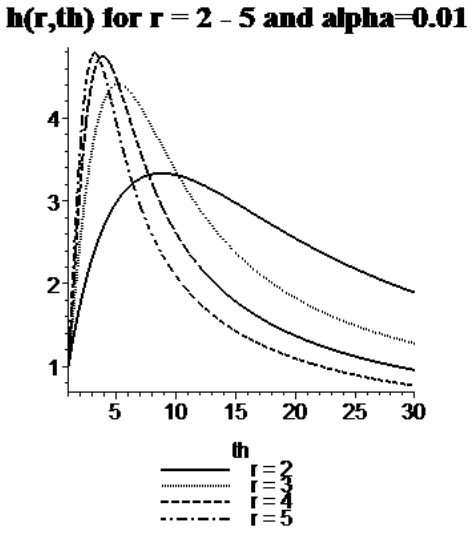

Figure 3.2

In view of Figure 3.2, the following result is of interest:

Lemma 3.2. *Under the conditions of* Lemma 3.1 *the value* $\theta^{max} = \theta_r^{max}$ *at which* h_r *from* (3.5) *reaches its maximum is adequately approximated by*

$$\tilde{\theta}^{max} = \frac{\tilde{\mu}}{\tilde{\lambda}}, \text{ with } \tilde{\mu} = \tilde{\mu}_r \text{ such that } rP(Z_{\tilde{\mu}} = r) = P(Z_{\tilde{\mu}} \geq r), \tag{3.6}$$

and $\tilde{\lambda}$ *as given by* (2.6).

Proof. From (3.5), together with (3.1)-(3.3), we obtain that $h_{r,\theta} \approx P(Z_{\theta\lambda} \geq r/\{r(1-(1-\alpha)^{\theta}\}$, with λ such that $P(Z_\lambda \geq r) = r\alpha$. Since $\partial\{P(Z_{\theta\lambda} \geq r)/r\}/\partial\theta = P(Z_{\theta\lambda} = r)/\theta$, it follows that $\partial h_{r,\theta}/\partial\theta = 0$ requires

$$\frac{rP(Z_{\theta\lambda} = r)}{P(Z_{\theta\lambda} \geq r)} = \frac{-\theta(1-\alpha)^{\theta}\log(1-\alpha)}{1-(1-\alpha)^{\theta}} \doteq 1 - \frac{\theta\alpha}{2} \doteq 1.$$

Hence $\theta\lambda$ approximately equals $\tilde{\mu}$ from (3.6). As moreover λ can be approximated according to Lemma 2.1 by $\tilde{\lambda}$ from (2.6), the desired result for $\tilde{\theta}^{max}$ follows. □

Example 3.2. It is easily verified that $\tilde{\mu}_2 = 1.79$, $\tilde{\mu}_3 = 3.38$, $\tilde{\mu}_4 = 4.88$ and $\tilde{\mu}_5 = 6.32$, which together with (2.6) and/or Table 2.1 immediately gives the desired $\tilde{\theta}^{max}$ from (3.6). For $\alpha = 0.01$ and $r = 3$ we e.g. have $\tilde{\theta}^{max} = 3.38/0.660 = 5.12$, while for $r = 5$ we find $\tilde{\theta}^{max} = 6.32/1.89 = 3.34$. These are indeed quite close to the corresponding exact values θ^{max}, which are 5.19 (with corresponding maximum 4.41 for h_3) and 3.23 (with maximum 4.78 for h_5), respectively. □

Note that by now we have accumulated quite a bit of information on the OoC behavior of the negative binomial chart. Its exact ARL_r is given in (3.1), with an adequate approximation $A\tilde{R}L_r$ in (3.4). Moreover, its relative behavior w.r.t. the geometric chart, as captured by h_r from (3.5), is depicted in Figures 3.1 and 3.2 and further characterized by (3.6). For briefness' sake we shall not go into more details. Instead we conclude this section by presenting a simple rule of thumb for finding r^{opt}, the value of r for which ARL_r is minimal in the region of interest. For given α and θ, let

$$\tilde{r}^{opt} = \frac{1}{\alpha(2.6\theta + 2) + 0.01(4\theta - 3)}. \tag{3.7}$$

Illustrative values obtained from (3.7) are assembled in Table 3.2.

Table 3.2. Comparison of the approximation $\tilde{r}^{opt}$ from (3.7) to the exact r^{opt} for various α and θ. The first value is r^{opt}, the second one is $\tilde{r}^{opt}$, while the third value is the realized ARL_r for $r = r^{opt}$.

α \ θ	3/2			2			3			4		
0.001	33	28	50.8	16	17	24.4	10	10	12.6	7	7	9.1
0.005	17	17	29.2	10	12	15.5	7	7	8.7	5	5	6.4
0.01	12	11	21.5	8	8	12.2	5	5	7.1	4	4	5.4

From Table 3.2 we may conclude that the rule of thumb (3.7) works quite well. In that respect, note that discrepancies between r^{opt} and $\tilde{r}^{opt}$ mainly occur in case of large values. But by that time the gain in decrease of ARL_r achieved by further increase of r has become marginal. E.g. for $\alpha = 0.001$ and $\theta = 3/2$ we have $ARL_{28} = 51.8$, which hardly differs from the value 50.8 given in the table for ARL_{33}. Moreover, note that Table 3.2 offers the opportunity to compare the optimal ARL's presented there to the ones listed in Table 3.1 for $r = 2 - 5$. The overall conclusion of such a comparison seems to be that the major part of the improvement w.r.t. the geometric chart usually is already achieved within the range $2 \leq r \leq 5$. Only for small α together with small θ it can be worthwhile to go beyond $r = 5$. As mentioned after Table 2.1, our approximations do

allow this. The question remains, however, which block length r is considered still acceptable in practice. To illustrate matters, we conclude with:

Example 3.3. Let us focus on the choice $r = 5$. Suppose first that we use $\alpha = 0.01$ (and thus $ARL_1 = 100/\theta$). Then $r = 5$ is just fine: for $\theta = 4$ it is even slightly over the top, as $r = 4$ is already marginally better, with an ARL-value of 5.4 as opposed to 5.6. For $\theta = 3$ it is optimal, while for $\theta = 2$ it gives 13.9 instead of the optimal 12.2 for $r = 8$. Even for $\theta = 3/2$ its 28.2 is not bad compared to 21.5 for $r = 12$, especially if we take into account that the starting point at $r = 1$ here is 66.7. In other words, most of the gain has indeed already been realized at $r = 5$. For $\alpha = 0.005$, the picture is only slightly less optimistic: for $\theta = 4$ it is optimal, for $\theta = 3$ it loses negligibly with 9.3 versus 8.7 for $r = 7$, while for $\theta = 2$ its 21.9 against 15.5 for $r = 10$ is also quite fair. Even for $\theta = 3/2$ the difference between its 49.8 and the optimal 29.2 seems bearable, in particular if this is once more compared to the starting point for $r = 1$, which here equals 133. Only for $\alpha = 0.001$ the remaining gap becomes a bit wider for the smaller θ: at $\theta = 2$ we can go from 500 at $r = 1$ through 73.7 for $r = 5$ quite a bit further down to 24.4 for $r = 16$, while for $\theta = 3/2$ we have 667 at $r = 1$, 203 at $r = 5$ and 50.8 at $r = 33$. Nevertheless, note that e.g. taking $r = 33$ is a dangerous option: if θ happens to be not that small after all, we are stuck with an unnecessarily large ARL. □

4 The estimated chart

Situations do occur where the value of p during IC is known. This happens for example if p is simply prescribed on the basis of an external minimal quality requirement. But generally p will be unknown in practice and a Phase I sample will have to precede the actual monitoring. Let m be the size of such a sample, in the sense that we observe the sequence $D_1, D_2, \ldots$ until m geometric r.v.'s $X_{1,p}$ (cf. (2.1)) - or equivalently a single negative binomial r.v. $X_{m,p}$ - have been gathered. Note that at this point r plays no role yet: regarding the estimation aspect as well, we want the comparison between the charts to be fair. Hence we do not sample m times an $X_{r,p}$ during Phase I, as this would unduly favor larger r. Let $\overline{X} = m^{-1}\Sigma_{i=1}^{m} X_i$ be the corresponding sample mean (i.e. $\overline{X} = m^{-1}X_{m,p}$), then it is straightforward that $E\overline{X} = 1/p$ and $\mathrm{var}(\overline{X}) = (1-p)/(mp^2)$. It is also standard that the unknown p will be estimated by $\hat{p} = 1/\overline{X}$. By plugging this result into (2.2), we immediately have an estimate $\hat{n} = n_{r,\hat{p}} = F_{r,\hat{p}}^{-1}(r\alpha)$ for n as well. But to be able to see what the impact of this step is, we have to resort again to our approximations. In view of (2.5), we obtain that

$$\hat{n} \approx \frac{\lambda}{\hat{p}} = \lambda\overline{X}, \tag{4.1}$$

where still λ is such that $P(Z_\lambda \geq r) = r\alpha$. Note that (4.1) is already easy to interpret: $\hat{n} = \eta r\overline{X}$, where $\eta = \lambda/r$ again is the small fraction discussed in section 2 (e.g. having value 0.169 in Example 2.1) and $r\overline{X}$ estimates the block length r/p (e.g. having value 3000 in Example 2.1). Obviously, combining (4.1) with the approximation step from (2.9) readily produces $\hat{\tilde{n}} = \tilde{\lambda}\overline{X}$, with $\tilde{\lambda}$ as in (2.6). Now the chart can be applied as before: after Phase I, rv's $X_{r,p}$ are formed again and we wait until such a r.v. is at most $\hat{\tilde{n}}$ (or $\hat{n}$).

However, note that it remains to analyze the effect of this estimation step. In particular, we should figure out how large m as to be to make this influence sufficiently small. A complication already pointed out in the Introduction, is the fact that the performance characteristics are now random. For each realization $\overline{x}$ of $\overline{X}$ we need to consider $\widehat{FAR} = FAR(\overline{x}) = P(X_{r,p} \leq \hat{n}|\overline{x})$, and hence in general we deal with the r.v.

$$\widehat{FAR} = FAR(\overline{X}) = P(X_{r,p} \leq \hat{n}|\overline{X}), \tag{4.2}$$

and likewise with $\widehat{ARL} = ARL(\overline{X})$. Consequently, no unique criterion exists to appraise relative errors such as

$$W_1 = \frac{\widehat{FAR} - r\alpha}{r\alpha} \text{ or } W_2 = \frac{\widehat{ARL} - \frac{1}{\alpha}}{\frac{1}{\alpha}}. \tag{4.3}$$

A first possibility is to use the bias involved and to require that e.g. EW_1 is sufficiently small. This was done in detail in AK (2004a) for the normal case. According to this mild criterion, the behavior of the chart is assessed in the long run, i.e. over a large number of subsequent applications of the procedure. Note that once m is sufficiently large for satisfactory overall behavior, it may still happen that individual applications of the chart lead to unpleasantly large errors. In other words, a well-behaved average still allows ample variation around this value. To control that aspect as well, a second, stronger, criterion has to be invoked: exceedance probabilities such as $P(W_1 > \varepsilon)$ or $P(W_2 < -\varepsilon)$), with ε some small positive number, should be made sufficiently small. This approach was extensively studied for the normal counterpart in AK (2004b).

Note that from (4.1) it follows that $\hat{n} \approx \hat{\lambda}/p$, where

$$\hat{\lambda} = \lambda(1 + U), \text{ with } U = p\overline{X} - 1. \tag{4.4}$$

Hence we can now write $\widehat{FAR} \approx P(Z_{\hat{\lambda}} \geq r|\overline{X})$, with λ such that $P(Z_\lambda \geq r) = r\alpha$. This enables us to obtain the following result.

Lemma 4.1. *To first approximation the relative bias of* $\widehat{FAR}$ *equals*

$$EW_1 = \frac{\gamma r(r - 1 - \lambda)}{2m}, \tag{4.5}$$

where $\gamma = P(Z_\lambda = r)/(r\alpha)$ *satisfies* $1 - \lambda/(r + 1) < \gamma < 1$.

Proof. Since $dP(Z_\mu \geq r)/d\mu = P(Z_\mu = r - 1)$ and $dP(Z_\mu = r)/d\mu = P(Z_\mu = r - 1) - P(Z_\mu = r)$, it follows from (4.2) and (4.3) that

$$\widehat{FAR} \doteq P(Z_\lambda \geq r) + \lambda U P(Z_\lambda = r - 1) + \frac{1}{2}\lambda^2 U^2\{P(Z_\lambda = r - 2) - P(Z_\lambda = r - 1)\}, \tag{4.6}$$

and thus that $\widehat{FAR} \doteq r\alpha + P(Z_\lambda = r)\{rU + \frac{1}{2}r(r - 1 - \lambda)U^2\}$. (Note that with the obvious interpretation $P(Z_\lambda = k) = 0$ as soon as k is negative, this result remains correct for $r = 1$

as well.) As $EU = 0$ and $EU^2 = (1-p)/m \approx 1/m$, we obtain in view of (4.3) that $EW_1 \approx \frac{1}{2}r(r-1-\lambda)P(Z_\lambda = r)/(mr\alpha)$ and hence the equality in (4.5) follows. As $P(Z_\lambda \geq r) = r\alpha$, the bounds on γ are an immediate consequence of applying (2.7) for $k = 1$. □

Remark 4.1. Obviously, more attention could be devoted to precise derivations for remainder terms etc. Such a detailed analysis can be found in AK (2004a) for the normal case. To avoid repetition we refrain from doing this here as well. Moreover, note that we already used approximations in the case of known p. This makes striving for very precise results at the subsequent estimation step rather pointless. Similar comments are in order below as well, e.g. when deriving correction terms. □

As remarked before, we have $\lambda = \eta r$, in which η is small (cf. Table 2.1) and thus the bias in (4.5) will be positive, unless of course $r = 1$. An illustration is provided by:

Example 4.1. Suppose $r = 3$ and $\alpha = 0.01$, then according to Table 2.1 $\lambda = 0.665$. Hence according to (4.5) $EW_1 = 2.00\gamma/m$, with $0.84 < \gamma < 1$. Hence to have a relative bias of e.g. at most 10% we need m to be at least 17, and possibly 20. For $r = 5$ and $\alpha = 0.01$, we have $\lambda = 1.08$ and thus $EW_1 = 7.30\gamma/m$, with $0.82 < \gamma < 1$. Now a 10% relative bias requires an m between 59 and 73.

In view of this example it might be worthwhile to correct $\hat{n}$ from (4.1) in order to remove the bias. So let us replace $\hat{n}$ by the slightly more strict limit

$$\hat{n}_c = \hat{n}(1-c) = \lambda\overline{X}(1-c), \tag{4.7}$$

for some small $c > 0$. We have:

Lemma 4.2. *The $\widehat{FAR}$ is unbiased to first order if we choose*

$$c = \frac{r-1-\lambda}{2m}. \tag{4.8}$$

Proof. From (4.7) it follows in view of (4.4) that $\hat{\lambda}_c = \hat{n}_c p = \hat{\lambda}(1-c) = \lambda(1+U)(1-c)$. Now replacement of $\hat{\lambda}$ by $\hat{\lambda}_c$ in $\widehat{FAR}$ produces in the expansion (4.6), after taking expectations and ignoring the contribution of c in its last term, an additional term $-c\lambda P(Z_\lambda = r-1) = -crP(Z_\lambda = r)$. Consequently, $E(\widehat{FAR}) \doteq r\alpha + P(Z_\lambda = r)\{-cr + \frac{1}{2}r(r-1-\lambda)/m$, which shows that the bias vanishes to first order for c as in (4.8). □

Example 4.1 (cont.) For $r = 3$ and $\alpha = 0.01$ the value $c = 0.67/m$ suffices, i.e $\hat{n}_c = 0.665\overline{X}(1 - 0.67/m)$, while $r = 5$ and $\alpha = 0.001$ gives $c = 1.46/m$, and thus $\hat{n}_c = 1.08\overline{X}(1 - 1.46/m)$. □

In principle, the same approach could be used to evaluate the relative bias of $\widehat{ARL}$, and also to subsequently correct it. However, we shall not pursue this. One reason of course is to avoid repetition, but, more importantly, it is also felt to be of little interest from a practical point of view. Quite often, practitioners are more interested in the likelihood of relatively short runs during

IC. This automatically leads us to our second criterion, based on exceedance probabilities. Here we can fortunately treat $\widehat{FAR}$ and $\widehat{ARL}$ simultaneously. To see this, note that according to (4.3), $P(W_2 < -\varepsilon) = P((r/\widehat{FAR} - 1/\alpha)/(1/\alpha) < -\varepsilon) = P(r\alpha/\widehat{FAR} < 1-\varepsilon) = P(\widehat{FAR}/(r\alpha) - 1 > \varepsilon/(1-\varepsilon)) = P(W_1 > \tilde{\varepsilon})$ with $\tilde{\varepsilon} = \varepsilon/(1-\varepsilon) \approx \varepsilon$. In other words, if too small values of $\widehat{ARL}$ are sufficiently rare, the same holds for too large values of $\widehat{FAR}$, and vice versa. Hence without loss of generality, we focus on $\widehat{FAR}$.

Specifically, we figure out how large the exceedance probability $P(W_1 > \varepsilon) = P(\widehat{FAR} > r\alpha(1+\varepsilon))$ can get. In addition, we employ a more strict limit $\hat{n}_c$ such as in (4.7) to ensure that, for some prescribed small β,

$$P(\widehat{FAR} > r\alpha(1+\varepsilon)) \leq \beta. \tag{4.9}$$

Let Φ be the standard normal df and let u_β denote its upper β-point, i.e. $1 - \Phi(u_\beta) = \beta$, then we have the following result:

Lemma 4.3. *Using* $\hat{n}$ *from* (4.1) *and* $\gamma = P(Z_\lambda = r)/(r\alpha)$ *from* (4.5) *leads to*

$$P(\widehat{FAR} > r\alpha(1+\varepsilon)) \approx 1 - \Phi\left(\frac{m^{\frac{1}{2}}\varepsilon}{\gamma r}\right). \tag{4.10}$$

Equality in (4.9) is achieved by using $\hat{n}_c$ from (4.7) with

$$c = m^{-\frac{1}{2}}u_\beta - \frac{\varepsilon}{\gamma r}. \tag{4.11}$$

Proof. From (4.6) and the proof of Lemma 4.2 it follows that the use of $\hat{n}_c$ implies that $\widehat{FAR} \doteq r\alpha + (U-c)rP(Z_\lambda = r)$. Hence the left-hand side of (4.9) to first order equals

$$P((U-c)P(Z_\lambda = r) > \varepsilon\alpha) = P(U > c + \frac{\varepsilon\alpha}{P(Z_\lambda = r)}). \tag{4.12}$$

As U is asymptotically normal with mean 0 and variance $1/m$, the latter probability in (4.12) approximately equals $1 - \Phi(m^{1/2}\{c + \varepsilon\alpha/P(Z_\lambda = r)\})$. The uncorrected version of $\hat{n}$ corresponds to $c = 0$ and hence (4.10) follows. If instead the prescribed β should result, $m^{1/2}\{c + \varepsilon\alpha/P(Z_\lambda = r)\})$ should equal u_β, and hence c should be chosen as in (4.11). □

Note that the results (4.10) and (4.11) are very transparent indeed. The exceedance probability can obviously be lowered by taking either a larger sample size m, or by becoming more liberal by allowing a larger ε. On the other hand, it is increased again when a larger r is used, which is yet another argument against going beyond e.g. $r = 5$. Note that (4.11) also implies that $c = 0$ will result for $m = (\gamma r u_\beta/\varepsilon)^2$. Hence for this sample size equality is reached in (4.9) without correction. As usual, some numerical illustration is given:

Example 4.2. Coming back to the motivating example mentioned in the Introduction, suppose we are interested in controlling the probability that $\widehat{ARL}$ falls more than 20% short of its intended value. Hence we use $\varepsilon = 0.20$ in $P(W_2 < -\varepsilon)$, which according to the above equals $P(W_1 > \tilde{\varepsilon})$, with $\tilde{\varepsilon} = \varepsilon/(1-\varepsilon) = 0.25$. In addition, let $\beta = 0.20$ and hence $u_\beta = 0.842$. Then no correction is needed anymore for $m = (\gamma r u_\beta/\varepsilon)^2 = 11.3(\gamma r)^2$. As $1-\lambda/(r+1) < \gamma < 1$, a safe upper bound thus is $11.3r^2$, i.e. $m = 284$ for $r = 5$. Suppose that the actual $m = 100$, then $c = 0.084 - 0.25/(\gamma r)$, which implies that for $r \geq 4$ some correction is indeed necessary. For e.g. $r = 5$ again we have that c is at most $0.084 - 0.050 = 0.034$, which is still quite small. If in addition $\alpha = 0.001$, we use $\hat{n}_c = 1.08\overline{X}(1 - 0.034)$. □

Summarizing, we have now successfully analyzed the impact of estimation on the negative binomial charts and also derived simple corrections in (4.8) and (4.11) to control for these effects. A final question that remains concerns the impact of such corrections during OoC. Obviously, lowering the limit $\hat{n}$ somewhat to improve the behavior during IC will also affect the OoC behavior. However, as is amply demonstrated in AK (2004a, 2004b), these effects are fortunately quite small. Thus they form no reason to avoid the use of such corrections. Hence, to avoid repetition, we shall be quite brief about this issue here and restrict ourselves to pointing out the basic explanation of this phenomenon.

Actually, it is quite simple: remember that the reason for the problems concerning the estimation step lies in the fact that during IC such small probabilities, like $r\alpha$, need to be estimated. The errors involved may be small in an absolute sense, but not in comparison to this $r\alpha$. During OoC, the alarm rate should rise sharply, hence we are no longer dealing with small probabilities and the estimation effect reduces to 'normal' proportions again. To be a bit more specific, a step like $\widehat{FAR} \doteq r\alpha + (U-c)rP(Z_\lambda = r)$ from Lemma 4.3 is now replaced by

$$P(Z_{\theta\hat{\lambda}_c} \geq r) \doteq P(Z_{\theta\lambda} \geq r) + (U-c)rP(Z_{\theta\lambda} = r). \tag{4.13}$$

Hence the expected relative impact of using c equals $RC = EP(Z_{\theta\hat{\lambda}_c} \geq r)/P(Z_{\theta\lambda} \geq r) - 1 = -\xi c$, with $\xi = rP(Z_{\theta\lambda} = r)/P(Z_{\theta\lambda} \geq r)$. Indeed, this ξ decreases in θ: it starts for $\theta = 1$ at $r\gamma$, with $1 - \lambda/(r+1) < \gamma < 1$, and e.g. passes 1 at θ^{max} from (3.6). Consequently, the effect of using a $c > 0$ diminishes considerably as θ increases.

References

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. and Kallenberg, W. C. M. (2006). Alternative Shewhart-type charts for grouped observations. *Metron* **LXIV(3)**, 357 - 375.

Albers,W. and Kallenberg, W. C. M. (2008). Minimum control charts. *J.Statist. Planning & Inference* **138**, 539-551.

Klar, B. (2000). Bounds on tail probabilities of discrete distributions. *Prob. Engin. & Inform. Science* **14**, 161-171.

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.

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., Hrenstam, K.P. and Brommels, M. (2007). Application of statistical process control in healthcare improvement: systematic review", *Qual. & Safety in Health Care* **16**, 387-399.

Zhang Wu , Xiaolan Zhang and Song Huat Yeo (2001). Design of the sum-of- conforming-run-length control charts. *Eur. J. of Oper. Res.* **132**, 187-196.

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.

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.