Resampling methodologies in Phase I control charts

Fernanda Otília Figueiredo[1] and M. Ivette Gomes[2]

[1] Faculdade de Economia da Universidade do Porto, and CEAUL, Rua Dr. Roberto Frias, 4200-464 Porto, Portugal
(E-mail: otilia@fep.up.pt)
[2] FCUL, DEIO and CEAUL, Universidade de Lisboa, Bloco C6, Piso 4 - Campo Grande, 1749-016 Lisboa, Portugal
(E-mail: ivette.gomes@fc.ul.pt)

Abstract. The bootstrap methodology is used in Phase I of control charting to estimate the nominal process parameters, together with the use of robust estimates. We evaluate the performance of the Mean-chart with estimated parameters for monitoring the process location, where the estimates are obtained on the basis of a simple reference sample or via bootstrapping from such sample. The run-length distribution of the corresponding charts is obtained by Monte Carlo simulations.

Keywords: Bootstrap, Control charts, Robust statistics, Statistical Process Control.

1 Introduction

The control charts, introduced by Shewhart in 1924, are one of the main tools in Statistical Process Control (SPC), but their domain has been successively enlarged, with applications to areas as diverse as Health, Medicine, Genetics, Biology, Environmental Sciences, Finance, Metrology, Sports and Justice, among others. For an overview of standard and non-standard applications of control charts see, for instance, Montgomery[18], Woodall and Montgomery[27],[28], MacCarthy and Wasusri[17], Dull and Tegarden[11], Vardeman et al.[26], and references therein.

To develop any control chart the nominal process parameters must be either assumed known or estimated. In practice the distribution of the process data as well as the process parameters are usually unknown, being the process parameters usually estimated from an in-control Phase I reference sample, made up of m subgroups of size n, before we proceed to the building of a (non-)parametric control chart.

A strong emphasis has been given to the analysis of the real performance of control charts implemented on the basis of estimated parameters, and to the effect of the non-normality in the performance of the usual control charts. Apart

from the pioneer works of Schilling and Nelson[25], Balakrishnan and Kocherlakota[1], Chan *et al.*[3], Rocke[21],[22], Quesenberry[20], Chen[8], Nedumaran and Pignatiello[19], Champ and Jones[4], Chakraborti[5],[6],[7], and Jensen *et al.*[16], we mention, among others, the recent works of Zhang and Castagliola[29], Schoonhoven *et al.*[23],[24], and Castagliola and Figueiredo[2]. From these studies we easily conclude that to obtain control charts implemented with estimated control limits with the same run-length properties as the corresponding charts with true limits, the choice of the number of subgroups, m, and the sample size, n, cannot be heuristic. Besides the need of a very large number m of subgroups, which is a limitation from a practical point of view, and in some cases even impossible, we must determine the control limits in a robust way. Other approach consists of modifying the chart parameters' in order to take into consideration the variability introduced by the estimation of the nominal process values in Phase I, allowing that way to maintain the expected false alarm rate.

Our aim in this paper is only to investigate the benefits of using the bootstrap methodology in Phase I of control charting to obtain a larger reference sample to estimate the nominal process parameters, together with the use of robust estimates. More precisely, from an in-control reference sample of m subgroups (20 or 30) of size $n = 5$, we set out to construct a larger reference sample of M_b subgroups (100, 500 or 1000) of size n by bootstrapping from the pooled sample of size $m \times n$. The nominal process parameters are then estimated through the use of a few location and scale statistics. As an illustration we only consider the traditional Mean-chart with estimated control limits implemented to monitor the mean value of a normal process. The paper is organized as follows. Section 2 provides some information about the implementation of the Mean-chart with estimated control limits, the bootstrap methodology and the statistics considered in the estimation of the nominal process parameters. Section 3 presents some relevant parameters of the run-length distribution of the traditional Mean-chart implemented on the basis of previous estimates, obtained by Monte Carlo simulations, and Section 4 concludes with some comments about the performance of the implemented control charts.

2 Mean-chart with estimated control limits based on bootstrap estimates

Let Y be a random variable associated with a normal process, being the in-control mean value, μ_0, and the in-control standard deviation, σ_0, both unknown. The most popular control chart for the process location monitoring is the Mean-chart with estimated control limits, $\overline{Y}$, obtained by plotting the sample means of the Phase II samples $(Y_{i,1}, \ldots, Y_{i,n})$, $i = 1, 2, \ldots$, of n independent normal random variables, $N(\mu_0 + \delta\sigma_0, \sigma_0)$, where i is the subgroup number and δ is the magnitude of the standardized mean shift. If $\delta = 0$ the process is in-control, otherwise the process is out-of-control due to a shift in the mean process.

The (estimated) control limits (CL's) of the $\overline{Y}$-chart are random variables, which can be written in the form

$$\widehat{CL}'s = \hat{\mu}_0 \pm K\hat{\sigma}_0. \tag{1}$$

where the chart parameter K depends on the sample size n, and is determined in order to obtain a given in-control performance, say, a fixed in-control Average Run-Length (ARL). For instance, the $\overline{Y}$-chart with exact 3-sigma control limits, $CL's = \mu_0 \pm \frac{3}{\sqrt{n}}\sigma_0$, leads to an in-control ARL=370.4. If we consider $K = 3/\sqrt{n}$ in (1), the corresponding Mean-chart does not have the same performance of the chart with exact 3-sigma control limits, unless the process nominal values μ_0 and σ_0 are adequately estimated.

The standard procedure is to estimate μ_0 and σ_0 from $m = 20, 30$ subgroups $(X_{i,1}, \ldots, X_{i,n})$, $i = 1, \ldots, m$ of size n, usually 4 or 5, assuming independence between and within subgroups, and that $X_{i,j} \sim N(\mu_0, \sigma_0)$. However, the literature refer that for an adequate estimation of μ_0 and σ_0, the number m of initial subgroups must be very large, at least $400/n$ (see, for instance, Quesenberry[20] and Castagliola and Figueiredo[2]).

In this study we apply the bootstrap methodology to the pooled sample of size $m \times n$ in order to obtain a larger number $M_b = 100, 500, 1000$ of subgroups of size n that will be used for the estimation of μ_0 and σ_0. How does the bootstrap methodology work?

Let $(W_1, \ldots, W_n)$ be a random sample of size n from a d.f. $F(.)$. The bootstrap sample, $(W_1^*, \ldots, W_n^*)$, is obtained by randomly sampling n times, with replacement, from the observed sample $(w_1, \ldots, w_n)$. These variables W_i^* are independent and identically distributed (i.i.d.) replicates from a random variable W^*, with d.f. equal to the empirical d.f. of our observed sample, given by

$$F_n^*(w) := \frac{1}{n}\sum_{i=1}^n I_{\{w_i \leq w\}}. \tag{2}$$

where I_A denotes the indicator function of the set A. For other details about the bootstrap methodology see, for instance, Davison and Hinkley[10], Efron[12] and Efron and Tibshirani[13].

In our case, by bootstrapping from the empirical d.f. associated to the pooled reference sample of size $m \times n$, $(x_{1,1}, \ldots, x_{1,n}, \ldots, x_{m,1}, \ldots, x_{m,n})$, we generate M_b random samples of size n, say $(X_{r,1}^*, \ldots, X_{r,n}^*)$, $r = 1, \ldots, M_b$. In the sequel $(X_{i,1}, \ldots, X_{i,n})$ denotes the i-th subgroup of size n used in the estimation of the nominal values and let $X_{i,(j)}$ be the j-th ascending order statistics (o.s.) associated to the subgroup $(X_{i,1}, \ldots, X_{i,n})$.

To estimate the nominal process parameters under consideration, i.e., the in-control mean value μ_0 and the in-control standard deviation σ_0, we have carried out the following procedure:

1. From k subgroups of size $n = 5$, with k denoting either m (20, 30) or M_b (100, 500, 1000), we compute k partial estimates, $\hat{\mu}_{0i}$ and $\hat{\sigma}_{0i}$, $i = 1, \cdots, k$;

2. Then, we consider the overall estimates $\widehat{\mu}_0 = \sum_{i=1}^{k} \widehat{\mu}_{0i}/k$ and $\widehat{\sigma}_0 = \sum_{i=1}^{k} \widehat{\sigma}_{0i}/k$, to be used in the 3-sigma control limits of the $\overline{Y}$-chart.

To obtain the partial estimates $\widehat{\mu}_{0i}$, we consider, for $n = 5$, the *sample mean*,

$$\overline{X}_i = \frac{1}{5} \sum_{j=1}^{5} X_{i,j},$$

and the *total median*, defined by

$$TMd_i = 0.058 \left(X_{i,(1)} + X_{i,(5)} \right) + 0.366 X_{i,(3)} + 0.259 \left(X_{i,(2)} + X_{i,(4)} \right).$$

To obtain the partial estimates $\widehat{\sigma}_{0i}$, unbiased whenever the underlying model is normal, we consider, for $n = 5$, the following statistics divided by the scale constant c (into brackets): the *sample standard deviation*,

$$S_i = \sqrt{\frac{1}{4} \sum_{j=1}^{5} \left(X_{i,j} - \overline{X}_i \right)^2} \qquad (c = c_4 = 0.940),$$

the *sample range*,

$$R_i = X_{i,(5)} - X_{i,(1)} \qquad (c = d_2 = 2.326).$$

and the *total range*, defined by

$$TR_i = 0.737 \left(X_{i,(5)} - X_{i,(1)} \right) + 0.263 \left(X_{i,(4)} - X_{i,(2)} \right) \qquad (c = 1.801).$$

The statistics TMd and TR are resistant to changes in the underlying model, and are similar to a special trimmed-mean, in which the ideal percentage of trimming does not depend on the data distribution. The distributional behaviour of the TMd and the TR estimators has already been investigated, and these statistics have revealed to be efficient and robust estimators of the mean value and the standard deviation, respectively. Details about these estimators can be found in Cox and Iguzquiza[9], Figueiredo[14] and Figueiredo and Gomes[15]. In the sequel the two overall estimates of μ_0 will be denoted by $\overline{\overline{X}}$ and $\overline{TMd}$, and the three overall estimates of σ_0 will be denoted by $\overline{S}/c_4$, $\overline{R}/d_2$ and $\overline{TR}/c$.

3 Run-length distribution of the proposed $\overline{Y}$-charts

The ability of a control chart to detect process changes is usually measured by the expected number of samples taken before the chart signals, i.e., by its ARL (*Average Run Length*), together with the standard deviation of the Run Length distribution, SDRL. When we have to estimate some process parameters to determine the control limits of the chart, the RL variable (i.e., the number of samples taken before the chart signals) has not a geometric distribution as it happens in the known parameters case, but a more right-skewed distribution.

Some authors, Chakraborti[5],[6],[7] and Jensen *et al.*[16], for instance, refer that for a more complete understanding of the chart performance, they suggest the analysis of the conditional RL distribution, i.e., the RL distribution conditional on the observed estimates, together with the analysis of the marginal RL distribution. Such a marginal distribution is computed by integrating the conditional RL distribution over the range of the parameter estimators and takes thus into account the random variability introduced into the charting procedure through parameter estimation without requiring the knowledge of the observed estimates.

In order to get information about the in-control and the out-of-control performance of the previous $\overline{Y}$ charts with estimated 3-sigma control limits to monitor normal data, we compute the (conditional) RL distribution of the $\overline{Y}$-charts by Monte Carlo simulation, using 250000 runs in the simulation experiment.

Table 1 presents the estimates of the most commonly used measures of performance of a control chart: the mean (ARL) and the standard deviation (SDRL) of the in-control RL distribution, for the case of known nominal process values (exact limits obtained by replacing $\mu_0 = 0$ and $\sigma_0 = 1$), and when the estimated control limits are based on the overall estimates $(\overline{\overline{X}}, \overline{S}/c_4)$, $(\overline{\overline{X}}, \overline{R}/d_2)$ and $(\overline{TMd}, \overline{TR}/c)$, obtained from a reference sample of m (20 and 30) subgroups of size $n = 5$ and from M_b (100, 500 and 1000) subgroups of size $n = 5$ obtained by bootstrapping from the pooled reference sample of size $m \times n$.

Table 2 presents the ARL and the SDRL of the $\overline{Y}$-charts with estimated and exact control limits for samples of size $n = 5$, when the process is out-of-control due to a shift in the mean value from $\mu = \mu_0 = 0$ to $\mu = \delta = 0.3, 0.5, 0.7, 1.0, 1.5$, and for $m = 30$ and $M_b = 500$ subgroups of size $n = 5$.

Mean-chart with		ARL			SDRL		
Estimated CL's # subgroups		Estimates			Estimates		
m	M_b	$(\overline{\overline{X}}, \overline{S}/c_4)$	$(\overline{\overline{X}}, \overline{R}/d_2)$	$(\overline{TMd}, \overline{TR}/c)$	$(\overline{\overline{X}}, \overline{S}/c_4)$	$(\overline{\overline{X}}, \overline{R}/d_2)$	$(\overline{TMd}, \overline{TR}/c)$
20		458	444	445	891	817	836
	100	404	386	375	728	713	676
	500	388	366	370	671	576	627
	1000	378	365	368	608	602	615
30		415	418	409	640	695	628
	100	394	374	375	633	620	588
	500	373	363	364	543	494	508
	1000	378	364	367	533	499	518
Exact CL's		371			371		

Table 1. In-control ARL and SDRL of the 3-sigma $\overline{Y}$-chart for samples of size $n = 5$. For the estimation we consider $m = 20, 30$ or $M_b = 100, 500, 1000$ subgroups of size $n = 5$.

Mean-chart with Estimated CL's # subgroups			ARL			SDRL		
			Estimates			Estimates		
m	M_b	δ	$(\overline{X}, \overline{S}/c_4)$	$(\overline{X}, \overline{R}/d_2)$	$(\overline{TMd}, \overline{TR}/c)$	$(\overline{X}, \overline{S}/c_4)$	$(\overline{X}, \overline{R}/d_2)$	$(\overline{TMd}, \overline{TR}/c)$
30		0.3	132.4	134.4	135.2	209.5	216.4	224.0
		0.5	42.3	42.5	42.9	62.3	64.1	65.4
		0.7	15.7	15.9	15.8	20.3	20.5	20.9
		1.0	5.0	5.0	4.9	5.4	5.5	5.2
		1.5	1.6	1.6	1.6	1.1	1.1	1.1
	500	0.3	124.8	121.5	123.3	196.0	181.5	188.2
		0.5	39.7	39.4	39.3	57.4	56.0	56.6
		0.7	14.9	14.7	14.8	18.8	18.9	18.6
		1.0	4.8	4.7	4.8	5.2	5.0	5.0
		1.5	1.6	1.6	1.6	1.1	1.0	1.0
Exact CL's		0.3	99.2			98.4		
		0.5	33.1			32.5		
		0.7	13.2			12.8		
		1.0	4.5			4.0		
		1.5	1.6			0.9		

Table 2. Out-of-control ARL and SDRL of the 3-sigma $\overline{Y}$-chart for samples of size $n = 5$. For the estimation we consider $m = 30$ or $M_b = 500$ subgroups of size $n = 5$. The process mean changed from $\mu = \mu_0 = 0$ to $\mu = \delta$.

4   Analysis of the performance of the proposed $\overline{Y}$-charts

As expected, for all the different combinations of the number of subgroups used in the estimation m or M_b and the estimates $(\hat{\mu}_0, \hat{\sigma}_0)$, the estimation of the nominal values have effect on the ARL and on the SDRL of the $\overline{Y}$-charts. However, the effect on the in-control and out-of-control RL behaviour becomes small when m increases, and specially if we consider a large number M_b of subgroups obtained by bootstrapping from the initial m subgroups of the reference sample. As m and M_b increases the ARL value of the chart with estimated control limits tends faster than the SDRL to the corresponding values obtained when the $\overline{Y}$-chart is implemented with exact CL's. For instance, if we consider $M_b = 500$ or 1000 subgroups obtained by bootstrapping from the initial $m = 20, 30$ subgroups of the reference sample, we obtain an in-control ARL approximately equal to 370.4, although the SDRL value maintains yet larger than 370.4. For detecting small shifts in the process mean value, we also get some improvements in terms of performance if we consider, for instance, $M_b = 500$ subgroups for the estimation of (μ_0, σ_0), by bootstrapping from the initial $m = 30$ subgroups of the reference sample. Concerning the different estimates of the nominal values here considered, the results are qualitatively the same, at least when monitoring normal data. As m and M_b increases, the upper percentiles become closer to the corresponding percentiles of the RL distribution of the $\overline{Y}$-chart with exact CL's. Finally, when it is not possible to consider a large reference sample or there is not available a modified chart parameter, K, that take into consideration the variability introduced by the estimation of

the nominal process values, the use of the bootstrap methodology should be explored because it can lead to some improvements in the performance of the chart.

Aknowledgments

References

1. Balakrishnan, N. and Kocherlakota, S., "Effects of nonnormality on $\overline{X}$ charts: single assignable cause model", *Dankhya: The Indian Journal of Statistics*, 48(B), 439–444 (1986).
2. Castagliola, P. and Figueiredo, F., "The median chart with estimated parameters", *European J. Industrial Engineering*, 7(5), 594–614 (2013).
3. Chan, L. K., Hapuarachchi, K. P., and Macpherson, B. D., "Robustness of $\overline{X}$ and R charts", *IEEE Transactions on Reliability*, 37(1), 117–123. (1988).
4. Champ, W. C. and Jones, A. L., "Design Phase I $\overline{X}$ charts with Small Sample Sizes", *Quality and Reliability Mathematical Engineering International*, 20, 497–510 (2004).
5. Chakraborti, S., "Run length, average run length and false alarm rate of Shewhart $\overline{X}$ chart: exact derivations by conditioning", *Communications in Statistics – Simulation and Computation*, 29(1), 61–81 (2000).
6. Chakraborti, S., "Parameter Estimation and Design Considerations in Prospective Applications of the $\overline{X}$ Chart", *Journal of Applied Statistics*, 33(4), 439–159 (2006).
7. Chakraborti, S., "Run Length Distribution and Percentiles: The Shewhart Chart With Unknown Parameters", *Quality Engineering*, 19, 119–127 (2007).
8. Chen, G., "The mean and standard deviation of the run length distribution of $\overline{X}$ charts when control limits are estimated", *Statistica Sinica*, 7(3), 789–798 (1997).
9. Cox, M. G. and Iguzquiza, E. P., "The total median and its uncertainity", in *Advanced Mathematical and Computational Tools in Metrology*, Ciarlini et al., Eds, 106–117 (2001).
10. Davison, A. and Hinkley, D.V., *Bootstrap Methods and their Application*, Cambridge University Press (2006).
11. Dull, R.B. and Tegarden, D.P., "Using control charts to monitor financial reporting of public companies", *Int. J. of Accounting Information Systems*, 5, 109–127 (2007).
12. Efron, B., "Bootstrap methods: another look at the jackknife", *Ann. Statist.*, 7(1), 1–26 (1979).
13. Efron, B. and Tibshirani, R.J., *An Introduction to the Bootstrap*, Chapman and Hall (1993).
14. Figueiredo, F., "Robust estimators for the standard deviation based on a bootstrap sample", *Proc. 13th European Young Statisticians Meeting*, 53–62 (2003).
15. Figueiredo, F. and Gomes, M.I., "The total median is Statistical Quality Control", *Applied Stochastic Models in Business and Industry*, 20, 339–353 (2004).
16. Jensen, W. A., Jones-Farmer, L. A., Champ, C. H., and Woodall, W.H., "Effects of Parameter Estimation on Control Chart Properties: A Literature Review", *J. Quality Technology*, 38, 349–364 (2006).

17. MacCarthy, B.L. and Wasusri, T., "A review of non-standard applications of statistical process control (SPC) charts", *Int. J. of Quality and Reliability Management*, 19, 295–320 (2002).

18. Montgomery, D.C., *Introduction to Statistical Quality Control: a Modern Introduction*, 6th edition, John Wiley & Sons (2009).

19. Nedumaran, G. and Pignatiello, J.J., "On Estimating $\overline{X}$ Control Limits", *J. Quality Technology* 33(2), 206–212 (2001).

20. Quesenberry, D.C., "The Effect of Sample Size on Estimated Limits for $\overline{X}$ and X Control Charts", *J. Quality Technology*, 25(4), 237–247 (1993).

21. Rocke, D.M., "Robust control charts", *Technometrics*, 31(2), 173–184 (1989).

22. Rocke, D.M., "$\overline{X}_Q$ and R_Q charts: robust control charts", *The Statistician*, 41, 97–104 (1992).

23. Schoonhoven, M., Riaz, M., and Does, R.J.M.M., "Design and Analysis of Control Charts for Standard Deviation with Estimated Parameters", *J. Quality Technology*, 43(4), 307–333 (2011).

24. Schoonhoven, M., Nazir, H.Z., Riaz, M., and Does, R.J.M.M., "Robust Location Estimators for the $\overline{X}$ Control Chart", *J. Quality Technology*, 43(4), 363–379 (2011).

25. Schilling, E.G. and Nelson, P.R., "The Effect of Non-Normality on the Control Limits of the $\overline{X}$ Charts", *J. Quality Technology*, 8(4), 183–188 (1976).

26. Vardeman, S., Hamada, M.S., Burr, T., Morris, M., Wenderberger, J., Jobe, J.M., Moore, L., and Wu, H., "an Introduction to Statistical Issues and Methods in Metrology for Physical Science and Engineering", *J. Quality Technology*, 46(1), 33–62 (2014).

27. Woodall, W.H. and Montgomery, D.C., "Research Issues and Ideas in Statistical Process Control", *J. Quality Technology*, 31(4), 376–386 (1999).

28. Woodall, W.H. and Montgomery, D.C., "Some Current Directions in the Theory and Application of Statistical Process Control", *J. Quality Technology*, 46(1), 78–94 (2014).

29. Zhang, Y. and Castagliola, P., "Run Rules $\overline{X}$ Charts When Process Parameters Are Unknown", *Int. J. of Reliability, Quality and Safety Engineering*, 17(4), 381–399 (2010).