Comparisons of Estimators of Process Standard Deviation in Constructing Shewhart Control Charts with Unequal Subgroup Sizes

Nien Fan Zhang[1], Per Winkel[2]
Statistical Engineering Division, National Institute of Standards and Technology, Gaithersburg, MD 20899
[2]Copenhagen Trial Unit, Centre of Clinical Intervention Research, University Hospital of Copenhagen, Blegdamsvej, Rigshospitalet DF-2100, Denmark

Abstract

The $\bar{X}$ and S control charts with unequal subgroup sizes have been discussed in the literature and used in practice. Several estimators of the process standard deviation based on sample standard deviations have been proposed in constructing the charts. We discuss the properties of these estimators and make comparisons using the criterion of minimum mean squared error. Control charts for $\bar{X}$ and S based on the recommended estimator of the process standard deviation are also discussed

Key Words: Biased Estimator, Control limit, Linear combination, Minimum variance, Relative difference.

1. Introduction

Control charts for $\bar{X}$ and S with unequal subgroup sizes are often used in practice. When the subgroup size varies, the corresponding control chart for equal subgroup sizes needs to be modified. Burr (1969), (1976) discussed the use of the sample standard deviation as the estimator of the process standard deviation. SAS (1995)[1] and Hart and Hart (2002) also provided three approaches to estimate the process standard deviation when the subgroup size varies. Montgomery (2001) used $\bar{S}/c_4(n_i)$, which is related to one of these three estimators to estimate the process standard deviation. Based on these estimators, the corresponding $\bar{X}$ and S control charts can be constructed. However, the properties of these estimators of the process standard deviation have not been studied completely. In this article, we will compare these estimators and illustrate the construction of the corresponding $\bar{X}$ and S control charts using the best of these estimators.

2. Estimators of Process Standard Deviation

[1] A commercial product is identified in this paper to adequately describe the statistical methodology. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the product identified is necessarily the best available for the purpose.

We assume that the process variable X is normally distributed with a mean μ and a variance σ^2. We also assume that there are k subgroups. For the ith subgroup, the sample size is n_i (n_i > 1). Given subgroups with unequal sample sizes, SAS (1995), pp. 1214 - 1215 lists three estimators of σ, the process standard deviation. The first, called the default estimator, is also proposed in Levinson and Tumbelty (1997) and Hart and Hart (2002) and is given by

$$\hat{\sigma}_{\mathrm{DF}} = \frac{\sum_{i=1}^{k} \frac{S_i}{c_4(n_i)}}{k}, \tag{1}$$

where S_i is the sample standard deviation of the ith subgroup and $c_4(n_i)$ is defined as

$$c_4(n_i) = \sqrt{\frac{2}{n_i - 1}} \frac{\Gamma\left(\frac{n_i}{2}\right)}{\Gamma\left(\frac{n_i - 1}{2}\right)}. \tag{2}$$

The second estimator, called the minimum variance linear unbiased estimator (MVLUE), was proposed by Burr (1969) and is given by

$$\hat{\sigma}_{\mathrm{MVLUE}} = \frac{\sum_{i=1}^{k} \frac{h_i S_i}{c_4(n_i)}}{\sum_{i=1}^{k} h_i}, \tag{3}$$

where

$$h_i = \frac{[c_4(n_i)]^2}{1 - [c_4(n_i)]^2}. \tag{4}$$

Burr (1969) showed that $\hat{\sigma}_{\mathrm{MVLUE}}$ is unbiased and has minimum variance among all the estimators that are linear combinations of S_i ($i = 1, \ldots, k$). The third estimator is a weighted root-mean-square (RMSE) estimator of σ given by

$$\hat{\sigma}_{\mathrm{RMSE}} = \frac{\bar{S}}{c_4(H)}, \tag{5}$$

where

$$\bar{S} = \sqrt{\frac{\sum_{i=1}^{k} S_i^2 (n_i - 1)}{\sum_{i=1}^{k} n_i - k}} \tag{6}$$

and $H = \sum_{i=1}^{k} n_i - k + 1$. This estimator was also discussed in Winkel and Zhang (2007), pp. 70-76. Montgomery (2001) pp. 245 - 246 used a fourth estimator, $\bar{S}/c_4(n_i)$ to estimate the process standard deviation for the ith subgroup and to construct the corresponding control charts.

3. Comparisons of the Estimators Comparisons of the Estimators

Assuming X follows a normal distribution it follows from (1) and (3) that $E[S_i] = c_4(n_i)\sigma$ (Ryan (1989), p. 143). Thus $\hat{\sigma}_{\text{DF}}$ and $\hat{\sigma}_{\text{MVLUE}}$ are unbiased estimators of σ. We show that $\hat{\sigma}_{\text{RMSE}}$ is also an unbiased estimator of σ. Since

$$\frac{(n_i - 1)S_i^2}{\sigma^2} \sim \chi^2_{n_i - 1}, \tag{7}$$

it follows from (6) that

$$\bar{S}^2 = \frac{\sum_{i=1}^{k} S_i^2 (n_i - 1)}{\sum_{i=1}^{k} n_i - k} \sim \frac{\sum_{i=1}^{k} \sigma^2 \chi^2_{n_i - 1}}{\sum_{i=1}^{k} n_i - k} = \frac{\sigma^2 \chi^2_{\sum_{i=1}^{k} n_i - k}}{\sum_{i=1}^{k} n_i - k}, \tag{8}$$

where χ^2_m is a Chi-square statistic with m degrees of freedom. Thus,

$$\bar{S} \sim \frac{\sigma \chi_{\sum_{i=1}^{k} n_i - k}}{\sqrt{\sum_{i=1}^{k} n_i - k}}. \tag{9}$$

From Lancaster (1982), p. 439, we have

$$E[\chi^p_\nu] = 2^{\frac{p}{2}} \frac{\Gamma\left(\frac{\nu + p}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}, \tag{10}$$

where p and ν are positive integers. From (9), (10), and (2) it follows that

$$E[\bar{S}] = \frac{\sigma E\left[\chi_{\sum_{i=1}^{k} n_i - k}\right]}{\sqrt{\sum_{i=1}^{k} n_i - k}} = \sigma c_4(H). \tag{11}$$

Thus, from (5) $\hat{\sigma}_{\text{RMSE}}$ is an unbiased estimator of σ. From (11), we obtain

$$E\left[\frac{\bar{S}}{c_4(n_i)}\right] = \frac{c_4(H)}{c_4(n_i)}\sigma . \tag{12}$$

Since $H > n_i$ for any i, $c_4(H) \neq c_4(n_i)$. In addition, from Montgomery (2001), p.761, $c_4(H) > c_4(n_i)$. Thus, it is clear that $\bar{S}/c_4(n_i)$ proposed in Montgomery (2001) is a biased estimator of σ, which tends to overestimate σ.

Now we need to compare the variances of these estimators. From (1) we have

$$\text{Var}[\hat{\sigma}_{\text{DF}}] = \frac{\sum_{i=1}^{k} \frac{\text{Var}[S_i]}{c_4^2(n_i)}}{k^2} . \tag{13}$$

The variance of S_i is given by (see Ryan (1989), p. 143)

$$\text{Var}[S_i] = [1 - c_4^2(n_i)]\sigma^2 . \tag{14}$$

Inserting this result in equation (13) we obtain

$$\begin{aligned} \text{Var}[\hat{\sigma}_{\text{DF}}] &= \frac{\sigma^2}{k^2}\sum_{i=1}^{k}\frac{1-c_4^2(n_i)}{c_4^2(n_i)} \\ &= \frac{\sigma^2}{k^2}\sum_{i=1}^{k}\frac{1}{h_i} \end{aligned} . \tag{15}$$

Similarly, from (14) and (3) we obtain the variance of $\hat{\sigma}_{\text{MVLUE}}$

$$\text{Var}[\hat{\sigma}_{\text{MVLUE}}] = \frac{\sigma^2}{\sum_{i=1}^{k} h_i} . \tag{16}$$

To obtain the variance of $\hat{\sigma}_{\text{RMSE}}$, we first note from (10) that

$$\begin{aligned} \text{Var}[\chi_\nu] &= E[\chi_\nu^2] - E^2[\chi_\nu] \\ &= \nu - 2\frac{\Gamma^2\left(\frac{\nu+1}{2}\right)}{\Gamma^2\left(\frac{\nu}{2}\right)} \end{aligned} \tag{17}$$

From (9) and (17) it follows that

$$\mathrm{Var}[\bar{S}] = \frac{\sigma^2}{\sum_{i=1}^{k} n_i - k} \mathrm{Var}[\chi_{\sum_{i=1}^{k} n_i - k}]$$

$$= \frac{\sigma^2}{\sum_{i=1}^{k} n_i - k} \left[\sum_{i=1}^{k} n_i - k - 2 \frac{\Gamma^2\left(\frac{\sum_{i=1}^{k} n_i - k + 1}{2} \right)}{\Gamma^2\left(\frac{\sum_{i=1}^{k} n_i - k}{2} \right)} \right]. \qquad (18)$$

By (5), we obtain

$$\mathrm{Var}[\hat{\sigma}_{\mathrm{RMSE}}] = \frac{\sigma^2 \Gamma^2\left(\frac{\sum_{i=1}^{k} n_i - k}{2} \right)}{2\Gamma^2\left(\frac{\sum_{i=1}^{k} n_i - k + 1}{2} \right)} \left[\sum_{i=1}^{k} n_i - k - 2 \frac{\Gamma^2\left(\frac{\sum_{i=1}^{k} n_i - k + 1}{2} \right)}{\Gamma^2\left(\frac{\sum_{i=1}^{k} n_i - k}{2} \right)} \right]. \qquad (19)$$

Because $\hat{\sigma}_{\mathrm{MVLUE}}$ is unbiased and has minimum variance among all the estimators that are linear combinations of S_i $(i = 1,...,k)$,

$$\mathrm{Var}[\hat{\sigma}_{\mathrm{DF}}] \geq \mathrm{Var}[\hat{\sigma}_{\mathrm{MVLUE}}]. \qquad (20)$$

The equality in (20) holds, i.e., $\mathrm{Var}[\hat{\sigma}_{\mathrm{DF}}] = \mathrm{Var}[\hat{\sigma}_{\mathrm{MVLUE}}]$ when $n_i = n$ for $i = 1,...,k$. We now need to compare $\mathrm{Var}[\hat{\sigma}_{\mathrm{RMSE}}]$ with $\mathrm{Var}[\hat{\sigma}_{\mathrm{DF}}]$ and $\mathrm{Var}[\hat{\sigma}_{\mathrm{MVLUE}}]$. We are not able to do this analytically. However, for given n_i $(i = 1,...,k)$ k, and σ, the variances of the three estimators in (15), (16), and (19) may be calculated. They may then be compared numerically. Without loss of generality, we assume that $\sigma^2 = 1$. For equal subgroup size n and k subgroups we have

$$\mathrm{Var}[\hat{\sigma}_{\mathrm{DF}}] = \mathrm{Var}[\hat{\sigma}_{\mathrm{MVLUE}}] = \frac{1 - c_4^2(n)}{k c_4^2(n)} \sigma^2. \qquad (21)$$

and

$$\mathrm{Var}[\hat{\sigma}_{\mathrm{RMSE}}] = \frac{\Gamma^2\left(\frac{nk-k}{2}\right)\left[nk-k-2\frac{\Gamma^2\left(\frac{nk-k+1}{2}\right)}{\Gamma^2\left(\frac{nk-k}{2}\right)}\right]}{2\Gamma^2\left(\frac{nk-k+1}{2}\right)}\sigma^2 . \qquad (22)$$

For all combinations where n varied from 2 to 20 and k varied from 2 to 10, we found that $\mathrm{Var}[\hat{\sigma}_{\mathrm{DF}}]$ and $\mathrm{Var}[\hat{\sigma}_{\mathrm{MVLUE}}]$ are always larger than $\mathrm{Var}[\hat{\sigma}_{\mathrm{RMSE}}]$. Table 1 shows for selected values of n and k, the relative differences $\{\mathrm{Var}[\hat{\sigma}_{\mathrm{DF}}] - \mathrm{Var}[\hat{\sigma}_{\mathrm{RMSE}}]\}/\mathrm{Var}[\hat{\sigma}_{\mathrm{RMSE}}]$.

Table 1 $\{\mathrm{Var}[\hat{\sigma}_{\mathrm{DF}}] - \mathrm{Var}[\hat{\sigma}_{\mathrm{RMSE}}]\}/\mathrm{Var}[\hat{\sigma}_{\mathrm{RMSE}}]$ calculated for selected values of subgroup size n and number of subgroups k when the subgroup size is constant.

	k=2	3	4	5	6	7	8	9	10
n=2	0.044	0.068	0.083	0.093	0.100	0.105	0.109	0.113	0.115
4	0.030	0.041	0.048	0.052	0.054	0.056	0.058	0.059	0.060
6	0.021	0.028	0.032	0.035	0.036	0.037	0.038	0.039	0.040
8	0.016	0.021	0.024	0.026	0.027	0.028	0.029	0.029	0.029

Table 2 $\{\mathrm{Var}[\hat{\sigma}_{\mathrm{DF}}] - \mathrm{Var}[\hat{\sigma}_{\mathrm{RMSE}}]\}/\mathrm{Var}[\hat{\sigma}_{\mathrm{RMSE}}]$ (denoted d_1) and $\{\mathrm{Var}[\hat{\sigma}_{\mathrm{MVLUE}}] - \mathrm{Var}[\hat{\sigma}_{\mathrm{RMSE}}]\}/\mathrm{Var}[\hat{\sigma}_{\mathrm{RMSE}}]$ (denoted d_2) for various combinations of subgroup sizes as shown in row 1.

N	[2 3]	[2 3 4]	[2 3 4 5]	[2 3 4 5 6]	[2 3 4 5 6 7]	[2 3 4 5 6 7 8]	[2 3 4 5 6 7 8 9]
d_1	0.18	0.31	0.41	0.49	0.55	0.61	0.66
d_2	0.04	0.05	0.05	0.05	0.04	0.04	0.04

Table 2 shows the relative difference $\{\mathrm{Var}[\hat{\sigma}_{\mathrm{DF}}] - \mathrm{Var}[\hat{\sigma}_{\mathrm{RMSE}}]\}/\mathrm{Var}[\hat{\sigma}_{\mathrm{RMSE}}]$ (denoted d_1) and the relative difference $\{\mathrm{Var}[\hat{\sigma}_{\mathrm{MVLUE}}] - \mathrm{Var}[\hat{\sigma}_{\mathrm{RMSE}}]\}/\mathrm{Var}[\hat{\sigma}_{\mathrm{RMSE}}]$ (denoted d_2) for various combinations of subgroup sizes. The combinations examined are shown in row 1 of Table 2. For example, the vector [2 3 4 5] indicates that the n_i's are 2, 3, 4, and 5 and k = 4.

It is seen in Table 2 that the "default" estimator $\hat{\sigma}_{\mathrm{DF}}$ has the largest variance. Using $\hat{\sigma}_{\mathrm{DF}}$ in place of $\hat{\sigma}_{\mathrm{RMSE}}$ may increase the variance by as much as 66 %. Thus, for the combinations examined, $\hat{\sigma}_{\mathrm{RMSE}}$ has the minimum variance followed by $\hat{\sigma}_{\mathrm{MVLUE}}$. Since these three estimators are unbiased, $\hat{\sigma}_{\mathrm{RMSE}}$ also has the minimum mean square error (MSE) for the combinations examined.

As pointed earlier $c_4(H) > c_4(n_i)$. Therefore from (5) we obtain

$$\mathrm{Var}[\hat{\sigma}_{\mathrm{RMSE}}] < \mathrm{Var}\left[\frac{\bar{S}}{c_4(n_i)}\right] \tag{23}$$

for all i. From this result and the fact that $\bar{S}/c_4(n_i)$ is a biased estimator, it follows that the MSE of $\hat{\sigma}_{\mathrm{RMSE}}$ is smaller than that of $\bar{S}/c_4(n_i)$ for all i. Based on the discussion above we recommend to using $\hat{\sigma}_{\mathrm{RMSE}}$ as the estimator of the process standard deviation. In the following we derive the equations for control charts with unequal subgroup sizes using this estimator.

4. $\bar{X}$ and S Control Charts with Unequal Subgroup Sizes Using $\hat{\sigma}_{\mathrm{RMSE}}$

For the $\bar{X}$ chart, we plot the sample statistic $\bar{X}_i$ for $i = 1, 2, \ldots$. For subgroup i, with the subgroup size n_i, $\bar{X}_i$ has mean μ and variance σ^2/n_i. Assuming that the process variable is normally distributed and the parameters of this distribution are known, the 3σ control limits at the ith point would be

$$E[\bar{X}_i] \pm 3\sqrt{\mathrm{Var}[\bar{X}_i]} = \mu \pm 3\frac{\sigma}{\sqrt{n_i}}. \tag{24}$$

Given k subgroups, the process mean is estimated by

$$\bar{\bar{X}} = \frac{\sum_{i=1}^{k} n_i \bar{X}_i}{\sum_{i=1}^{k} n_i} \tag{25}$$

and σ by $\hat{\sigma}_{\mathrm{RMSE}}$ in (5). Thus the control limits of the $\bar{X}$ chart when the distribution parameters are known are defined by

$$\bar{\bar{X}} \pm 3\frac{\bar{S}}{c_4(H)\sqrt{n_i}}, \tag{26}$$

where $H = \sum_{i=1}^{k} n_i - k + 1$.

For the S chart, we plot the statistic S_i. As for the $\bar{X}$ chart, similar to (24) when the parameters of the distribution are known, the control limits are given by

$$E[S_i] \pm 3\sqrt{\mathrm{Var}[S_i]}. \tag{27}$$

It is known that $E[S_i] = c_4(n_i)\sigma$ and $\text{Var}[S_i] = (1 - c_4^2(n_i))\sigma^2$. Since $E[S_i]$ depends on n_i, the S chart does not have a common, central straight line. It follows from (5) that the central point for the ith subgroup ($E[S_i]$) is estimated by $\left[c_4(n_i)/c_4(H)\right]\overline{S}$.

From (27) and (14), using $\hat{\sigma}_{\text{RMSE}}$, the upper control limit for the ith subgroup is given by

$$\frac{c_4(n_i)\overline{S}}{c_4(H)}\left(1 + 3\frac{\sqrt{1 - c_4^2(n_i)}}{c_4(n_i)}\right). \tag{28}$$

Similarly, the lower control limit of the S chart for the ith subgroup is given by

$$max\left(0, \frac{c_4(n_i)\overline{S}}{c_4(H)}\left(1 - 3\frac{\sqrt{1 - c_4^2(n_i)}}{c_4(n_i)}\right)\right). \tag{29}$$

5. Conclusions

To build $\overline{X}$ and S control charts when subgroup sizes vary, an appropriate estimator of the process standard deviation has to be determined. Three unbiased and one biased estimators have been proposed in the literature. We studied their properties and showed that the weighted root-mean–square estimator has a minimum variance as well as the minimum MSE. The corresponding $\overline{X}$ and S control charts based on that estimator were also derived.

References

Burr, I. W. 1969. Control charts for measurements with varying sample sizes, Journal of Quality Technology, 1, 163-167.

Burr, I. W. 1976. Statistical Quality Control Methods, New York: Mercel Dekker.

Hart, M. and Hart, R. 2002. Statistical Process Control for Health Care, Pacific Grove, California: Duxbury.

Lancaster, H. O. 1982. Chi Distribution. Encyclopedia of Statistical Science, vol. 1, New York: John Wiley.

Levinson, W. A. and Tumbelty, F. 1997. SPC Essentials and Productivity Improvement – A manufacturing Approach, Milwaukee, Wisconsin: ASQC Quality Press.

Montgomery, D. C. 2001. Introduction to Statistical Quality Control, 4th edn., New York: John Wiley.

Ryan, T. P. 1989. Statistical Methods for Quality Improvement, New York: John Wiley.

SAS Institute Inc. 1995. SAS/QC Software: Usage and reference, Version 6, ed. 1, Vol. 2, Cary, NC.

Winkel, P. and Zhang, N. F. 2007. Statistical Development of Quality in Medicine, Southern Gate, Chichester, West Sussex : John Wiley.