American Research Journal of Mathematics **Original Article**
ISSN 2378-704X **Volume 1,Issue1, Feb-2015**

Control Charts Based on Robust Scale Estimators

Nuri CELIK[1]

Bartin University, Department of Statistics

Abstract: Control charts are one of the most common techniques that have been used to observe and control the process deviations in the industry. The easiest and the most prevalent method of control charts is the Shewhart S-control chart. This method is based on the normal distribution assumption. However, there are a lot of inferences in literature that non-normal distributions are much more common than the normal distribution. When the normality assumption is not satisfied robust methods are preferred. In this paper, we determine some approaches by using robust scale estimators instead of simple standard deviation in order to apply S-control charts. Furthermore the performances of these different scale estimators are compared by Monte Carlo simulation study. Numerical examples are given at the end of the paper

Keywords: S-control charts, Scale estimators, Robust estimation, Non-normal distributions

I. INTRODUCTION

Control charts can be used to determine if a process has been in a state of statistical control by examining last data (Ryan, 1989). There are several advantages of using control charts; they improve the communication by applying a common methodology and enhance self-control. These charts identify the source of the problem and control charts can be determined lasting documents which provide information about the process performance.

Control charts generally contain central line (CL) and two control limits that are drawn horizontal and symmetrically below and above the central line. The central line shows the target value and the control limits determine the control area. If the process is between the upper control limit (UCL) and the lower control limit (LCL), it is said to be in-control. The final purpose of a control chart is to give a data-driven tool to help businesses to bring an out-of-control process back into an in control state.

Many different types of control charts have been studied in statistics literature. The Shewhart control charts have been the most frequently used control charts. The control chart based on the general theory proposed by Walter A. Shewhart is called as Shewhart control charts. These charts are the control graphs that plot sample standard deviations (*S*) in order to control the variability of a variable. The central line and the control limits of S-charts are calculated as,

$$LCL = \bar{S} - 3\frac{\bar{S}}{c_4}\sqrt{1-c_4^2} \quad (1)$$

$$CL = \bar{S} \quad (2)$$

$$UCL = \bar{S} + 3\frac{\bar{S}}{c_4}\sqrt{1-c_4^2} \quad (3)$$

Where $\bar{S} = \frac{\sum_{i=1}^{m} S_i}{m}$ and m is the number of subgroups. S is not an unbiased estimator of population standard deviation σ. Therefore, c_4 is a constant that makes S unbiased and depends on the sample size n.

When the underlying distribution is normal then S is an efficient estimator of σ. However, in literature, there are several studies saying that non-normal distributions are more prevalent than the normal distribution in practice, see for example, Pearson (1932), Geary (1947), Huber (1981) and Tan and Tiku (1999). In addition, observations in a sample which are too small or too large as compared to the bulk of observations are called outliers. Since their presence adversely affects the efficiency of most statistical procedures (Tiku and Akkaya, 2004). S is very sensitive to the outliers. In order to handle these difficulties robust estimation methods have been studied for last decades.

Shewhart control charts are based on the sample mean ($\bar{x}$) and S. Therefore, they are not resistant for non normality and/or presence of outliers. Ferrell (1953) proposed median and median midranges as control limits. Langenberg and

[1] *Corresponding Author: ncelik@bartin.edu.tr*

Iglewicz (1986) used trimmed means as determining control limits. Iglewicz and Hoaglin (1987) proposed plotting subgroup box plots involving inter quartile range (IQR). Rocke (1989) suggested another way of calculating resistant control limits. Abu-Shawiesh (2008) determined a simple approach to robust estimation of the process standard deviation by using Median Absolute Deviation (MAD). They showed that MAD has a better performance than *S* in heavy tailed distributions and moderate sample sizes. In this paper, we expand this approach by using other robust scale estimator in the technique of control charts.

The rest of the paper organizes as follows: in chapter 2, we determine the robust methods of control charts by using several robust scale estimators. In chapter 3, we apply these on different real data examples. Lastly, we compare the performance of the estimators by using the Monte Carlo simulation study. A conclusion is given at the end of the paper.

1. Control Charts with Robust Scale Estimators

Most classical statistical analysis are based on some assumptions, especially that the underlying distribution is normal. According to Geary (1947), normality is a myth, there never was, and never will be a normal distribution. In addition, observations in a sample which are too small or too large as compared to the bulk of observations are called outliers. Their presence seriously affects the performance of the normal theory procedures. Therefore, robust estimation and testing methods are needed. An estimator is said to be robust if it is fully efficient (or nearly so) for an assumed distribution but maintains high efficiency for plausible alternatives (Tiku and Akkaya, 2004). In this section, we give small information about some robust scale estimators and we construct robust quality control charts based on them.

1.1. Mean Absolute Deviation

The MAD was first introduced by Hampel (1974) as a robust alternative to the sample standard deviation. This estimator is very simple and easy to compute. For more robust and efficient estimators, MAD is used for initial estimator for iterative procedures. The MAD is calculated as

$$MAD = 1.4826\, med\, \{|x_i - med(x_i)|\} \tag{4}$$

The MAD has important robustness properties

- It has a maximum breakdown point which is % 50.
- The gross error sensitivity is 1.167 which is the smallest value that can be obtained with any dispersion estimator in normal distribution.
- The influence function of MAD is bounded.
- The efficiency of MAD is % 37 at normal distribution.

The control limits and central line for the Shewhart-S control chart based on the MAD are calculated as follows:

$$LCL = c_4\hat{\sigma} - 3\hat{\sigma}\sqrt{1 - c_4^2} = B_5^*\overline{MAD}$$

$$CL = c_4\hat{\sigma} = c_4^*\overline{MAD}$$

$$UCL = c_4\hat{\sigma} + 3\hat{\sigma}\sqrt{1 - c_4^2} = B_6^*\overline{MAD} \tag{5}$$

The values of the control limit factors c_4^*, B_5^* and B_6^* are given in the paper Omar and Abu-shawiesh (2008). They showed that in normal distribution, MAD has sameperformance with sample standard deviation. However, in non normal distribution, especially heavy tailed distribution, and for moderate sample sizes the robust method leads better performance than corresponding normal theory.

1.2. S_n and Q_n Estimators

Rousseeuw and Croux(1993) introduced new robust scale estimators alternative to the MAD. As determined in the paper the MAD has low efficiency at Gaussian distributions and uses symmetric way of variation. The estimator S_n does not need any locationestimation and is defined by

$$S_n = 1.1926\, med_i\{med_j\, |x_i - x_j\,|\} \tag{6}$$

S_n Estimator is a very powerful alternative to the MAD and has better robustness properties like;

- It has also a maximum breakdown point which is % 50.

- The influence function is also bounded.
- The efficiency of S_n estimator is % 58.23 at normal distribution which is better than the MAD.

Q_n Estimator, on the other hand, shares the same propertieswith S_n estimator also it has a smooth and bounded influence function. It is calculated by,

$$Q_n = 2.219\,\{|x_i - x_j|; i < j\}_k \quad (7)$$

Where $k = \binom{h}{2} \approx \binom{n}{2}/4$. Q_n estimators'robustness properties

- It has also a maximum breakdown point which is % 50.
- The influence function is also bounded and has no discrete part.
- The efficiency of Q_n estimator is % 82 at normal distribution which is better than the MAD and S_n estimator.

In small sample sizes S_n estimator performs better than Q_n estimator (Rousseeuw and Croux, 1993).

The limits and the center line of the Shewhart-S control chart based on S_n ($\overline{S_n} = \frac{\sum_{i=1}^{m} S_{n(i)}}{m}$)and Q_n ($\overline{Q_n} = \frac{\sum_{i=1}^{m} Q_{n(i)}}{m}$)estimator are found as follows

$$CL_S = c_4\hat{\sigma} = c_4 d_n \overline{S_n} = c_4^S \overline{S_n} \quad and \quad CL_Q = c_4\hat{\sigma} = c_4 e_n \overline{Q_n} = c_4^Q \overline{Q_n}$$

$$LCL_S = c_4\hat{\sigma} - 3\hat{\sigma}\sqrt{1-c_4^2} = c_4 d_n \overline{S_n} - 3 d_n \overline{S_n}\sqrt{1-c_4^2} \quad = B_{SL}^* \overline{S_n}$$

$$LCL_Q = c_4\hat{\sigma} - 3\hat{\sigma}\sqrt{1-c_4^2} = c_4 e_n \overline{Q_n} - 3 e_n \overline{Q_n}\sqrt{1-c_4^2} \quad = B_{QL}^* \overline{Q_n}$$

$$UCL_S = c_4\hat{\sigma} + 3\hat{\sigma}\sqrt{1-c_4^2} = c_4 d_n \overline{S_n} + 3 d_n \overline{S_n}\sqrt{1-c_4^2} \quad = B_{SU}^* \overline{S_n}$$

$$UCL_Q = c_4\hat{\sigma} + 3\hat{\sigma}\sqrt{1-c_4^2} = c_4 e_n \overline{Q_n} + 3 e_n \overline{Q_n}\sqrt{1-c_4^2} \quad = B_{QU}^* \overline{Q_n} \quad (8)$$

where$d_n$and $e_n$are the correction factors that make these estimators unbiased for normal distribution for $S_n$and$Q_n$respectively.

The values of the control limit factors d_n,e_n, c_4^S, c_4^Q, B_{SL}^*, B_{QL}^*, B_{SU}^* and B_{QU}^* are given in Table 1.

1.3. τEstimator

τestimator is introduced by Yohai and Zamar (1988) and isalso proposed by Maronna and Zamar (2002) as a starting robust estimator for estimating iterative variance covariance matrix. Asymptotically, τestimate is equivalent to an M estimatewith a function given by a weighted average of two psi-functions, one corresponding to a very robust estimate and the other to a highly efficient estimate (Yohai and Zamar, 1988). The estimate is calculated as follows: Define the functions;

$$W_c(x) = \left(1 - \left(\frac{x}{c}\right)^2\right)^2 \; I(|x| \le c) \quad and \quad \rho_c(x) = min(x^2, c^2)$$

Table1. *The control limit factors for S_n and Q_n estimators*

n	d_n	c_4^S	B_{SL}^*	B_{SU}^*	e_n	c_4^Q	B_{QL}^*	B_{QU}^*
2	1.175	0.937	0.000	3.062	1.192	0.951	0.000	3.108
3	1.482	1.313	0.000	3.372	1.491	1.322	0.000	3.393
4	1.349	1.243	0.000	2.815	1.355	1.248	0.000	2.827
5	1.244	1.169	0.000	2.439	1.275	1.199	0.000	2.501
6	1.197	1.139	0.036	2.241	1.259	1.198	0.038	2.358
7	1.156	1.109	0.134	2.084	1.211	1.162	0.141	2.184
8	1.130	1.090	0.196	1.984	1.173	1.132	0.203	2.061

9	1.123	1.089	0.260	1.917	1.134	1.099	0.263	1.936
10	1.097	1.067	0.299	1.834	1.101	1.071	0.300	1.841
11	1.096	1.068	0.337	1.799	1.099	1.071	0.338	1.804
12	1.076	1.052	0.372	1.731	1.078	1.054	0.373	1.735
13	1.075	1.052	0.401	1.702	1.073	1.051	0.400	1.701
14	1.066	1.046	0.428	1.664	1.065	1.045	0.428	1.663
15	1.059	1.039	0.441	1.639	1.059	1.040	0.448	1.640
16	1.056	1.039	0.472	1.606	1.056	1.039	0.472	1.606
17	1.047	1.031	0.485	1.578	1.047	1.031	0.485	1.578
18	1.045	1.030	0.501	1.559	1.046	1.031	0.502	1.560
19	1.040	1.026	0.517	1.535	1.042	1.028	0.518	1.538
20	1.039	1.025	0.517	1.535	1.039	1.025	0.516	1.534
21	1.039	1.026	0.535	1.517	1.036	1.023	0.533	1.514
22	1.038	1.026	0.554	1.497	1.035	1.023	0.552	1.494
23	1.036	1.024	0.552	1.496	1.034	1.022	0.551	1.493
24	1.033	1.022	0.571	1.473	1.032	1.021	0.571	1.472
25	1.033	1.022	0.571	1.473	1.030	1.019	0.569	1.469

Let $X = [x_1, x_2, \dots, x_n]$ be a univariate sample and put

$$\sigma_0 = MAD(x) \quad and \;\; w_i = W_{ci} = \left(\frac{x_i - med(x)}{\sigma_0}\right)$$

Then the location and scale statistics are defined

$$\mu(x) = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i}$$

$$\sigma(x)^2 = \frac{\sigma_0^2}{n} \sum_{i=1}^{n} \rho_{c2}\left(\frac{x_i - \mu(x)}{\sigma_0}\right) \tag{9}$$

To combine robustness and efficiency, $c_1 = 4.5$and $c_2 = 3$ are taken (Maronna and Zamar, 2002). τ estimators have the following robustness properties

- It has also a maximum breakdown point which is % 50,
- They are qualitatively robust,
- They are highly efficient for normal distribution.

The limits and the center line of the Shewhart-S control chart based on$\tau(\bar{\tau} = \frac{\sum_{i=1}^{m} \tau_i}{m})$estimator are found as follows

$$CL_\tau = c_4 \hat{\sigma} = c_4 k_n \bar{\tau} = c_4^\tau \bar{\tau}$$
$$LCL_\tau = c_4 \hat{\sigma} - 3\hat{\sigma}\sqrt{1 - c_4^2} = c_4 k_n \bar{\tau} - 3k_n \bar{\tau}\sqrt{1 - c_4^2} \;\; = B_{\tau L}^* \bar{\tau}$$
$$UCL_\tau = c_4 \hat{\sigma} + 3\hat{\sigma}\sqrt{1 - c_4^2} = c_4 k_n \bar{\tau} + 3k_n \bar{\tau}\sqrt{1 - c_4^2} \;\; = B_{\tau U}^* \bar{\tau} \tag{10}$$

The values of the control limit factors k_n, $c_4^\tau, B_{\tau L}^*$ and $B_{\tau U}^*$ are given in Table 2. ($k_n$is the correction factor)

2. Numerical Example

The numerical example is about a company uses a process to paint refrigerators with a coat of enamel. During each shift, a sample of 5 refrigerators is selected and the thickness of the paint (in mm) is determined. If the enamel is too thin, it will not provide enough protection. If it is too thick it will result in an uneven appearance with running and wasted paint. Table 3 shows the measurements from 20 consecutive shifts and the calculated statistics.

Table2. *The control limit factors for τ estimators*

n	k_n	c_4^τ	$B_{\tau L}^*$	$B_{\tau U}^*$
2	1.192	0.951	0.000	3.102
3	1.373	1.217	0.000	3.125

4	1.241	1.144	0.000	2.590
5	1.178	1.108	0.000	2.310
6	1.143	1.088	0.035	2.141
7	1.121	1.076	0.130	2.022
8	1.111	1.071	0.192	1.950
9	1.103	1.069	0.255	1.883
10	1.090	1.060	0.297	1.822
11	1.081	1.054	0.333	1.775
12	1.074	1.050	0.372	1.728
13	1.072	1.050	0.400	1.700
14	1.065	1.045	0.428	1.663
15	1.064	1.045	0.442	1.647
16	1.064	1.045	0.475	1.617
17	1.060	1.044	0.490	1.597
18	1.057	1.042	0.507	1.577
19	1.053	1.048	0.523	1.554
20	1.051	1.037	0.523	1.552
21	1.048	1.035	0.539	1.530
22	1.047	1.035	0.559	1.512
23	1.045	1.033	0.557	1.509
24	1.044	1.032	0.577	1.489
25	1.043	1.032	0.577	1.488

The performances of all types of estimators are shown in Figure 1. Figure 1 indicates that, all types of control charts are out-of control. However, the first and second control charts based on *S* and MAD respectively show that two points are out of control. On the other hand, the control charts based on other robust estimators indicates that one point is out of control.

Table3. *The Thickness of Refrigerators and Calculated Statistics*

Shift No						$\bar{X}$	S	MAD	S_n	Q_n	τ
1	2.7	2.3	2.6	2.4	2.7	2.54	0.182	0.148	0.119	0.222	0.164
2	2.6	2.4	2.6	2.3	2.8	2.54	0.195	0.297	0.239	0.222	0.174
3	2.3	2.3	2.4	2.5	2.4	2.38	0.084	0.148	0.119	0.222	0.075
4	2.8	2.3	2.4	2.6	2.7	2.56	0.207	0.297	0.239	0.222	0.186
5	2.6	2.5	2.6	2.1	2.8	2.52	0.259	0.148	0.119	0.222	0.223
6	2.2	2.3	2.7	2.2	2.6	2.40	0.235	0.148	0.119	0.222	0.218
7	2.2	2.6	2.4	2.0	2.3	2.30	0.224	0.148	0.239	0.222	0.200
8	2.8	2.6	2.6	2.7	2.5	2.64	0.114	0.148	0.119	0.222	0.102
9	2.4	2.8	2.4	2.2	2.3	2.42	0.228	0.148	0.119	0.222	0.208
10	2.6	2.3	2.0	2.5	2.4	2.36	0.230	0.148	0.239	0.222	0.210
11	3.1	3.0	3.5	2.8	3.0	3.08	0.259	0.148	0.119	0.222	0.223
12	2.4	2.8	2.2	2.9	2.5	2.56	0.288	0.445	0.358	0.222	0.258
13	2.1	3.2	2.5	2.6	2.8	2.64	0.404	0.297	0.358	0.444	0.361
14	2.2	2.8	2.1	2.2	2.4	2.34	0.279	0.148	0.119	0.222	0.222
15	2.4	3.0	2.5	2.5	2.0	2.48	0.356	0.148	0.119	0.222	0.284
16	3.1	2.6	2.6	1.8	2.1	2.64	0.365	0.297	0.239	0.444	0.326
17	2.9	2.4	2.9	1.3	1.8	2.26	0.702	0.741	0.596	0.666	0.629
18	1.9	1.6	2.6	3.3	3.3	2.54	0.783	1.038	0.835	1.095	0.700
19	2.3	2.6	2.7	2.8	3.2	2.72	0.327	0.148	0.239	0.222	0.270
20	1.8	2.8	2.3	2.0	2.9	2.36	0.483	0.741	0.596	0.444	0.432
Mean							0.311	0.296	0.263	0.321	0.273
CL							0.311	0.336	0.307	0.384	0.303
LCL							0.000	0.000	0.000	0.000	0.000
UCL							0.647	0.702	0.639	0.802	0.631

3. Simulation Study

In this section, we compare the performances of different control charts for different distributions. All the simulation results are based on *[100,000/n]* Monte Carlo runs and 100 subgroups and 10 sample sizes. We use some models, which distributed near normal or contains outliers namely Dixon's outlier model -$(n-1)$ observation come from normal distribution and one outlier (not known which one) comes from normal distribution with higher standard deviation value than the distribution that the bulks of data come from-, Contamination model *-(1-p)%* of the observation come from normal distribution with standard deviation *S* and *p%* of the observation come from normal distribution with standard deviation not equal to *S*, where *p* is a proportion differs from *[0,1]*- and Mixture model -*(1-p)%* of the observation come from normal distribution with standard deviation *S* and *p%* of the observation come from other distributions near normal-. We use the following sample models to represent a large number of plausible alternatives.

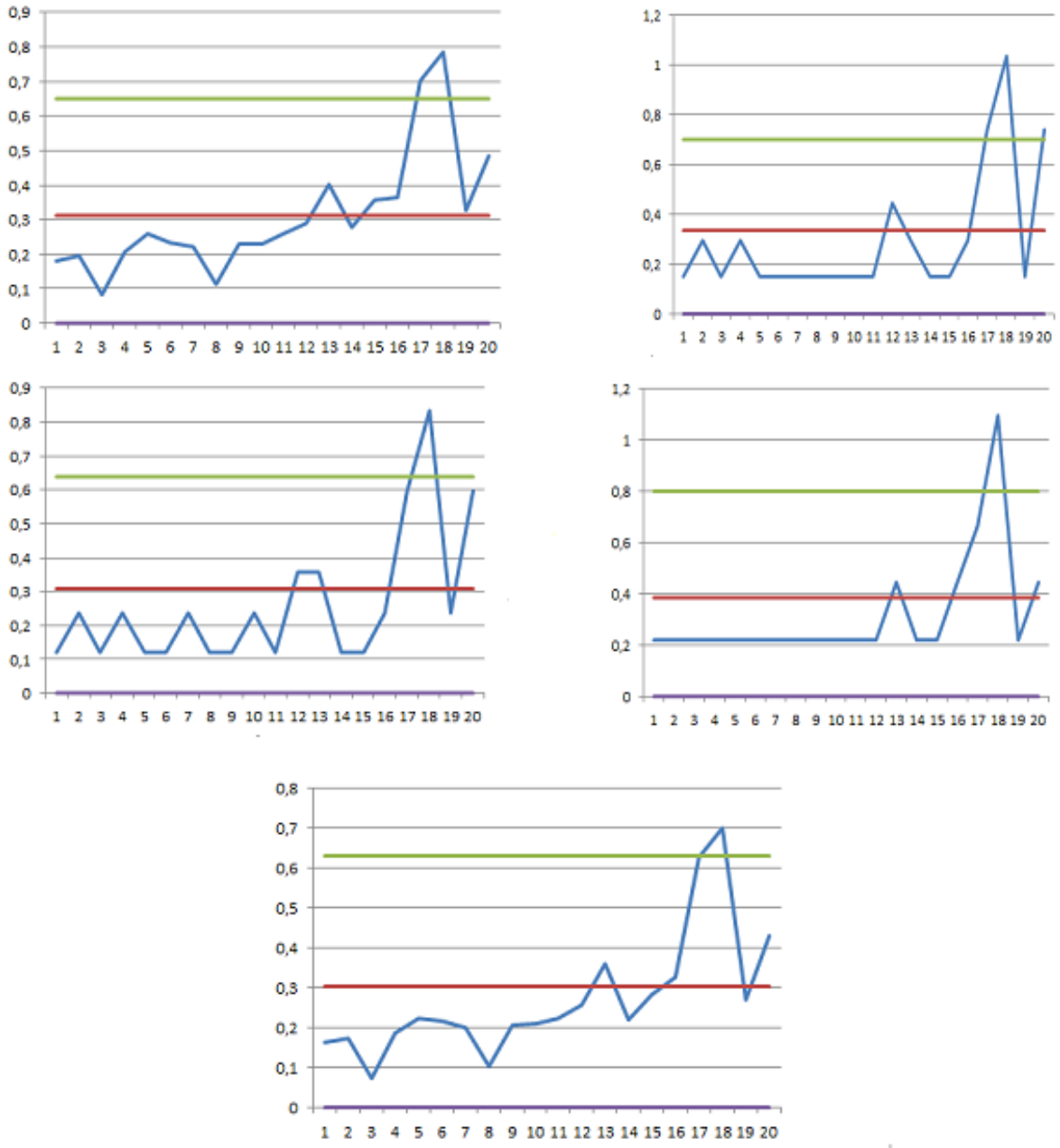

Fig1. *Control charts based on scale estimators*

Table4. *Control limits of control charts based on scale estimators*

Model (1)					
	S	MAD	S_n	Q_n	τ
LCL	0.264	0.266	0.272	0.274	0.269
CL	0.963	0.897	0.971	0.980	0.963
UCL	1.662	1.668	1.670	1.684	1.656
Out-of-Control	1	1	1	1	1
Model (2)					
LCL	0.740	0.494	0.319	0.338	0.534
CL	2.694	1.742	1.136	1.208	1.909
UCL	4.647	2.991	1.953	2.076	3.281
Out-of-Control	15	7	2	4	6
Model (3)					
LCL	0.421	0.318	0.319	0.325	0.377
CL	1.532	1.127	1.137	1.160	1.347
UCL	2.643	1.934	1.954	1.995	2.315
Out-of-Control	8	2	1	1	1
Model (4)					
LCL	0.290	0.238	0.279	0.291	0.289
CL	1.056	0.838	0.998	1.039	1.033
UCL	1.821	1.439	1.716	1.786	1.776
Out-of-Control	3	2	1	1	1
Model (5)					
LCL	0.738	0.519	0.344	0.353	0.578
CL	2.686	1.828	1.228	1.262	2.063
UCL	4.634	3.130	2.112	2.169	3.547
Out-of-Control	16	11	2	2	5

Sample Models

- Model (1): Normal Distribution N(0,1)
- Model (2): Dixon's outlier model: (n-1) observations come from N(0,1) but one observation (we do not know which one) comes from N(0,10)
- Model (3): Contamination model: 0.90N(0,1) +0.10N(0,4)
- Model (4): Mixture model: 0.90N(0,0.01) +0.10 t(2)
- Model (5): Cauchy Model

Simulation results are given in Table 4.

As indicated in Table 4, in normal distribution, all estimators leads approximately to the same control limits and also all have the same number of points falling outside the control limits. However, in model (2) and model (3), control chart based on sample standard deviation have the biggest control interval and also have the most number of out of control points because of the outliers. Control charts based onS_n, $Q_n$and τestimators have better performances with respect to *S* and MAD estimators. There are approximately same results in model (4) and (5).

II. CONCLUSION

Traditionally, control charts for monitoring the process variability are based on some assumptions. There are several forms of control charts. The most common control chart is Shewhart *S* control charts. These control charts are drawn with calculating standard deviations of subgroups. However, *S* is not a robust estimator. For this reason, we presented different control charts based on robust scale estimators. The result of numerical example and Monte Carlo simulations show that the proposed robust methods leads approximately to the same performance in the presence of normality and for alternative models, such as outlier, contamination and mixture model robust methods have better performance than traditionally used method. For heavy tailed distribution such as Cauchy, the robust

methods also have better performances as expected. When we compare robust methods, Q_n and S_n estimators have better properties than the other robust methods. Therefore, in the case of non normality or outliers, it is recommended to use proposed robust control charts as an alternative to *S* control charts.

REFERENCES

[1] Abu-Shawiesh, M.O.A., (2008). A simple Robust Control Chart Based on MAD, Journal of Mathematics and Statistics, 4(2), 102-107.

[2] Ferrell, E.B.,(1953). Control Charts Using Midranges and Median, Industrial Quality Control, 9, 30-34.

[3] Geary, R.C., (1947). Testing for normality, Biometrika,34, 209-242.

[4] Hampel, F.R., (1974). The influence curve and its role in robust estimation, Journal of the American Stat. Assoc., 69, 383 393.

[5] Huber, P.J., Robust Statistics, Jonh Wiley, New York, 1981.

[6] Iglewicz, B. and Hoaglin, D., (1987). Use of Boxplots for Process Evaluation, Journal of Quality Technology, 19, 180-190.

[7] Langenberg, P. and Iglewicz, B., (1986). Trimmed mean $\overline{X}$ and R charts, J. Qual. Technol., 18, 152-161.

[8] Maronna, R.A. and Zamar, R.H., (2002). Robust estimates of location and dispersion for high-dimensional datasets, Technometrics, 44-4, 307-317.

[9] Pearson, E.S., (1932). The analysis of variance in cases of nonnormal variation, Biometrika, 23, 114.133.

[10] Rocke, D.M., (1989). Robust Control Charts, Technometrics, 31, 173-193.

[11] Rousseeuw, P.J. and Croux, C., (1993). Alternatives to the madian absolute deviation, Journal of the American Stat. Assoc., 80, 1273-1283.

[12] Ryan, T.P. Statistical Methods for Quality Improvement, New York, 1989.

[13] Shewhart, W.A. Economic Control of Quality of Manufactured Product, Milwaukee, 1931.

[14] Tan, W.Y. and Tiku, M.L. Sampling distributions in terms of Laguerre Polynomials with applications, New Age International (formerly, Wiley Eastern), New Delhi, 1999.

[15] Tiku, M.L. and Akkaya, A.D. Robust Estimation and Hypothesis Testing. New Age International (P) Limited, Publishers (2004), New Delhi, 337pp.

[16] Yohai, V.J. and Zamar, R.H. (1988). High Breakdown Point Estimates Regression by Means of the Minimization of an Efficient Scale, Journal of the American Stat. Assoc.,83, 406-413.