A CONTROL CHART FOR HEAVY TAILED DISTRIBUTIONS

K. Thaga

Department of Statistics
University of Botswana, Botswana
thagak@mopipi.ub.bw

ABSTRACT

Standard control charts with control limits determined by the mean and standard error of the mean are constructed based on the assumption that the distribution of the quality characteristic being monitored follows a normal distribution. However, this assumption is not always valid. It is proposed to use a chart based on computing the control limits using the process mean and the standard error of the least absolute deviation for the case where the process quality characteristics follow a heavy tailed *t* distribution. Such a control chart is more effective than the normal distribution based chart since it has a low out-of-control average run length for both small and large values of process shift.

OPSOMMING

'n Kontrolekaart wat gebruik maak van kontrolelimiete gebaseer op die standaardafwyking van die geringste absolute limiet word ontwerp vir 'n t-verdeling met 'n betekenisvolle stert. Simulasietoetse vir vergelyking van die voorgestelde kontrolekaart met normaalverdeelde kontrolekaarte toon dat korter gemiddelde looplengtes voor diagnose van beheerverlies uitgewys word, bereik word.

1. INTRODUCTION

Statistical process control charts such as the Shewhart control chart (Shewhart [8]), the cumulative sum control chart (Page [5]), and the exponentially weighted moving average control chart (Roberts [6]), are used to monitor product quality and detect special events that may be indicators of out-of-control situations. Such charts are designed on the assumption that a process being monitored will produce measurements that can be modeled with an independent and identically distributed normal distribution, when only the inherent sources of variability are present in the system. However, in certain applications the process may produce measurements that can be represented with heavy tailed distributions. In this case, the standard control charts based on normality assumptions will not rapidly detect out-of-control situations, since the control limits will be stretched – particularly when the nature of the products is such that one cannot take large samples to be able to use the central limit theorem.

Several charts based on outlier resistant statistics have been proposed for use when there are outliers in the process measurements. These include, among others, the charts whose control limits are calculated using the median midrange and median range by Ferrell [1]. Langenberg and Iglewicz [4] proposed charts whose control limits are determined by the trimmed mean of the subgroup means and the trimmed mean of the ranges. White and Schroeder [11] proposed a chart constructed by plotting subgroup box plots. Such a chart uses the subgroup median and subgroup interquartile range. Rocke [7] proposed a series of robust control charts that use combinations of subgroup trimmed and untrimmed mean, median range and interquartile range.

Most of the charts discussed above use resistant statistics to determine the control limits, and then monitor the subgroup means for the occurrence of out-of-control signals. The median charts are less sensitive to process shifts since the median is not affected by outliers or extreme value. The chart that plots the mean and range with control limits determined from the subgroup means and the interquartile ranges is more effective in detecting mean shifts.

For a heavy tailed distribution, the extreme observations are not necessarily outliers or signs of the presence of assignable causes of variation. Thaga [9] proposed a chart based on the least absolute deviation that is effective in monitoring the process whose quality measurement follows a heavy tailed distribution. The chart is effective for monitoring processes with quality characteristics that are autocorrelated.

It is proposed and shown in this article that, for independent processes, when process variables follow a heavy tailed distribution, a chart where control limits are determined using the standard errors of the least absolute deviation estimators performs better than the chart where control limits are calculated using the standard errors of the ordinary least squares estimators. It is also proposed that the ratio of mean deviation to standard deviation should be used to determine the appropriateness of this chart in relation to the standard normal distribution based control chart.

2. LEAST ABSOLUTE DEVIATIONS ESTIMATORS

Consider a regression model of the form

$$Y_i = h(X_i, \theta) + \varepsilon_i \tag{1}$$

The $\varepsilon_i's$ are assumed to be independent for *i = 1, 2,..., n* and have a symmetric distribution. A least absolute deviation (LAD) estimator of θ is a solution to the problem

$$\min \sum_{i=1}^{n} |Y_i - h(X_i, \theta)|. \tag{2}$$

Here the discrepancy between the response variable Y_i and its approximation $h(X_i, \theta)$ provided by the model is measured by the L_1 distance instead of the usual L_2 distance, when studying the least squares estimate. A difficulty with the least absolute deviation method arises from the nondifferentiability of the objective function in equation (2). This function, however, is continuous and continuously differentiable at every point except at zero, where left and right derivatives exist. Because of such properties, an approach similar to that of Thavaneswaran and Heyde [10] can be followed, provided that the derivative at every point is replaced by the right derivative. This derivative is given by

$$\frac{\partial^{+}|x|}{\partial x} = I_{x \geq 0} - I_{x<0} \tag{3}$$

i.e. the least absolute deviation estimating function is given by

$$gLAD(Y, \tilde{\theta}) = \sum_{i=1}^{n} \frac{\partial h(X_i, \hat{\theta}_n)}{\partial \theta} (I_{(Y_i - h(X_i, \hat{\theta}_n) \geq 0)} - I_{(Y_i - h(X_i, \hat{\theta}_n) < 0)} \tag{4}$$

Let f be the conditional density function of Y/X such that $f(0)>0$. Under suitable regularity conditions (Gourieroux and Monfort [2]), it can be shown that the information associated with the estimating function is

$$\frac{I^{-1}}{4f^2(0)}, \tag{5}$$

Where

$$I = \sum_{i=1}^{n} E\left(\frac{\partial h(X_i, \hat{\theta}_n)}{\partial \theta} \frac{\partial h(X_i, \hat{\theta}_n}{\partial \theta} \right)$$

In a simple linear regression model, $h(x_i, \theta) = \theta x_i$ and

$$y_i = \theta x_i + \varepsilon_i \tag{6}$$

Then $Var(\hat{\theta}_{LSE}) = \dfrac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$ and $I = \sum_{i=1}^{n} x_i^2$.

Therefore $Var(\tilde{\theta}) = \dfrac{1}{4f^2(0)\sum_{i=1}^{n} x_i^2}$, where $\hat{\theta}$ is the ordinary least square (OLS) estimator and $\tilde{\theta}$ is the least absolute deviation estimator. The least absolute deviation estimating function is more efficient than the ordinary least squares estimating function if

the distribution of the error term is such that $4f^2(0) \geq 1/\sigma^2$, where σ^2 is the variance of the error term. For the case with normally distributed errors, the ordinary least square estimating function is more efficient than the least absolute deviation estimating function. When the errors have a Cauchy distribution, it can be shown that $4f^2(0) \geq 0$ (Gourieroux and Monfort [2]), and the least absolute deviation estimating function is more efficient than the least squares estimating function.

3. THE NEW CONTROL CHART

When the observations are independent identically distributed normal random variables, the observation at time *t* can be represented as $y_t = \theta + \varepsilon_t$. The Shewhart control chart for the process mean is developed with the following control limits:

$$\begin{aligned} \mathrm{LCL} &= \hat{\theta} - 3\frac{\sigma}{\sqrt{n}} \\ \mathrm{CL} &= \hat{\theta} \\ \mathrm{UCL} &= \hat{\theta} + 3\frac{\sigma}{\sqrt{n}}, \end{aligned} \tag{7}$$

where $\hat{\theta}$ is the process mean and $\frac{\sigma}{\sqrt{n}}$ is the standard error of the ordinary least square estimator $\hat{\theta}$.

As mentioned earlier, when the process measurements follow a heavy tailed distribution, the ordinary least square estimator has a greater standard error than the least absolute deviation estimator. Therefore a control chart based on the standard error of the ordinary least square estimator has wide control limits. A chart based on the least absolute deviation estimator is proposed as follows: The control limits for the chart are given as:

$$\begin{aligned} \mathrm{LCL} &= \hat{\theta} - 3\frac{\sigma}{2f(0)\sqrt{n}} \\ \mathrm{CL} &= \hat{\theta} \\ \mathrm{UCL} &= \hat{\theta} + 3\frac{\sigma}{2f(0)\sqrt{n}}, \end{aligned} \tag{8}$$

where $\frac{1}{2f(0)\sqrt{n}}$ is the standard error of the least absolute deviation estimator. The chart may be constructed by plotting the subgroup means against time or sample number with control limits given in equation (8). If the process measurements have a *t* distribution with v degrees of freedom, then

$$4f^2(0) = 4\frac{(\Gamma((v+1)/2))^2}{v\pi(\Gamma(v/2))^2} \tag{9}$$

The variance of the error term is finite for $v \geq 3$ and is given by $1/\sigma^2 = (v-2)/v$. It can be shown that for $v = 1$, the variance is infinite and $4f^2(0) = 4/\pi^2$. Similarly, for $v = 2$ the variance is infinite and $4f^2(0) = 0.5$ (Thavaneswaran and Heyde [10]). This shows that for heavy tailed *t* distributions with infinite variance, the least absolute deviation approach is superior. It can also be shown that for $v = 3$, $4f^2(0) \approx 0.54$ and for $v = 4$, 6. In these cases, the least absolute deviation estimating function provides more information than the ordinary least squares estimating function. For a *t* distribution with $v \geq 5$ with thin tails the least squares estimating function provides more information.

Table 1 shows the control limits for data simulated for a process that follows a *t* distribution with various degrees of freedom. 15,000 subgroups of three observations each have been simulated. The statistic $\bar{x}$ is the sample ordinary least square estimator, and $\tilde{x}$ is the sample least absolute deviation estimator. It can be seen that the least absolute deviation estimator has a small standard error when the process produces measurements that follow a heavy tailed *t* distribution, while the ordinary least square estimator has a small standard error for a *t* distribution with five or more degrees of freedom. Therefore, for heavy tailed data, the control limits for charts based on the normality assumption are wider than those computed using the least absolute deviation estimator. Wider control limits result in the control chart not being able to detect process shifts rapidly – particularly small shifts. Therefore the chart based on ordinary least square estimators is not recommended for heavy tailed distributions.

					OLS		LAD	
df.	$\bar{x}$	$\tilde{x}$	s.e($\bar{x}$)	s.e($\tilde{x}$)	LCL	UCL	LCL	UCL
1	2.19	1.13	1.24	0.71	0	5.91	0.06	4.32
2	2.50	1.17	0.86	0.63	0	5.08	0.61	4.39
3	1.26	1.02	0.73	0.61	0	3.45	0	3.09
4	1.26	0.91	0.61	0.60	0	3.09	0	3.06
5	1.06	0.84	0.45	0.59	0	2.41	0	2.83
6	1.04	0.82	0.40	0.58	0	2.24	0	2.78
7	0.82	0.72	0.30	0.58	0	1.72	0	2.56

Table 1: Control limits for the charts based on OLS and LAD estimators for a process following *t* distributions.

To decide which procedure to use, it is recommended that one should first calculate the ratio of the mean absolute deviation to the standard deviation. When the process quality characteristic follows a normal distribution, one expects this ratio to be 0.707. However, since the least absolute deviation estimator performs better than the ordinary least square estimator for a *t* distribution with four or less degrees of freedom, it is recommended that the chart based on the least absolute deviation estimator be used when the ratio is 0.707 or less. When the quality characteristic follows a *t* distribution with four degrees of freedom, the ratio of the mean deviation to the standard deviation is 0.707. For a *t* distribution with three degrees of freedom, the ratio is 0.637 (Johnson and Kotz [3]).

The average run lengths, which are the average number of points that must be plotted by the chart before a point falls beyond the control limits, are commonly used statistical process control measures for comparing the performance of control charts. When a point falls outside the control limit(s), the chart issues an out-of-control signal, indicating the possible presence of an assignable cause of variation in the process. It is therefore desirable for a chart to have a small average run length when the process has shifted, and a large average run length when the process is in control.

Figure 1 shows the average run length curves for the control charts for the least absolute deviation and ordinary least square estimators. These charts are compared by adjusting their control limits so that the two control charts have the same in-control average run length of 370. Viewing the out-of-control average run length, it can be concluded that the chart based on the least absolute deviation estimator is more sensitive than the chart based on the ordinary least square estimator in detecting both small and large shifts in the process for heavy tailed processes. For example, to detect a 0.5σ shift in the process mean, the chart based on the least absolute deviation estimator detects this shift on the 270th sample, while the chart based on the ordinary least square estimator signals this shift on the 295th sample. For a 2σ shift in the mean, a chart based on the ordinary least square estimator detects this shift on the 91st sample, while the chart based on the least absolute deviation estimator detects this shift on the 39th sample. And for a 3σ shift in the mean, a chart based on the ordinary least square estimator detects this shift on the 37th sample, while the chart based on the least absolute deviation estimator detects this shift on the 14th sample.

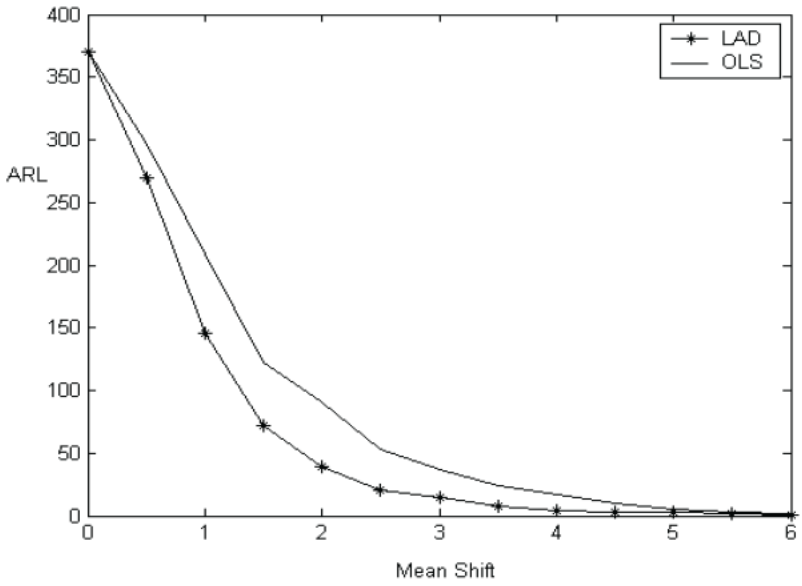

Figure1: The ARL curves for the OSL estimator and the LAD estimator based control charts.

4. AN EXAMPLE

To provide a picture of how the standard Shewhart chart and the proposed chart respond to various types of process change, a set of simulated data is used. Specific process changes are introduced into the data, and the two charts are plotted to monitor these changes.

Sixty samples of size four have been simulated for a process whose quality measurements follow a *t* distribution with two degrees of freedom. These data are used to construct the charts shown in Figures 2 and 3. Figure 2 shows the standard Shewhart chart, and Figure 3 shows the chart constructed using the least absolute deviation estimator. The figures show that the process is in control.

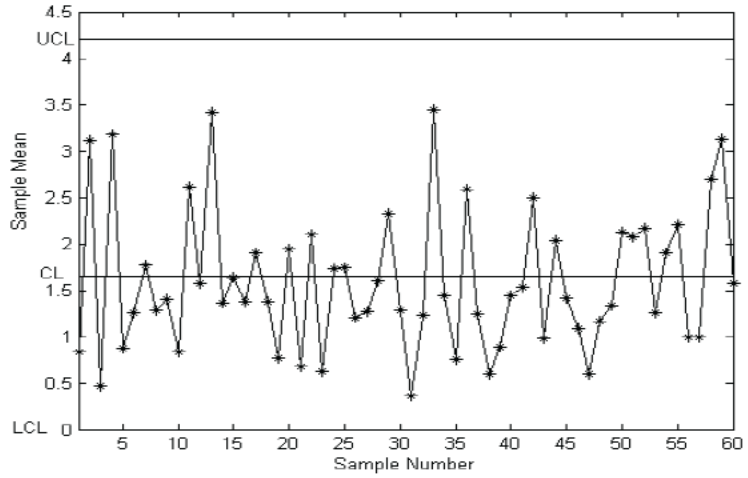

Figure 2: Shewhart chart for an in-control heavy tailed distribution

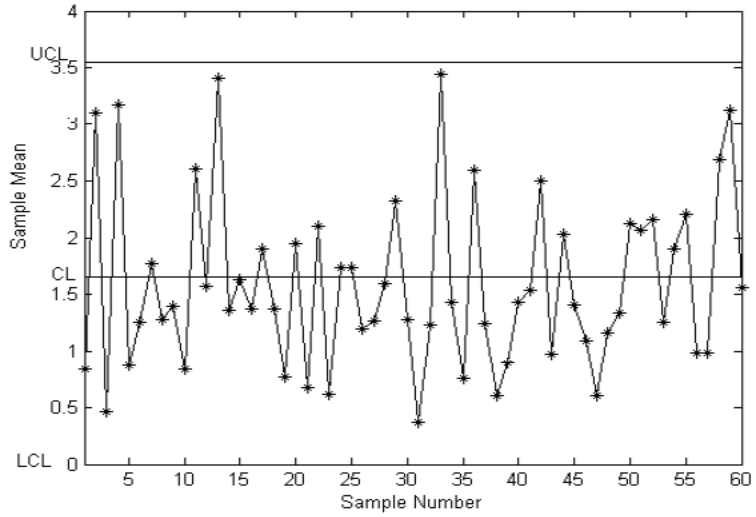

Figure 3: LAD-based chart for an in-control heavy tailed distribution

To introduce a process change, the next 40 observations were simulated using a *t* distribution with three degrees of freedom. This data was added to the data simulated above; the results are plotted in Figures 4 and 5.

The standard Shewhart chart shown in Figure 4 does not signal a shift in the process distribution. The chart that uses the least absolute deviation estimator signals a shift in the distribution for the first time on the 86th observation, as shown in Figure 5. The Shewhart does not detect this shift because it has wider control limits for heavy tailed data.

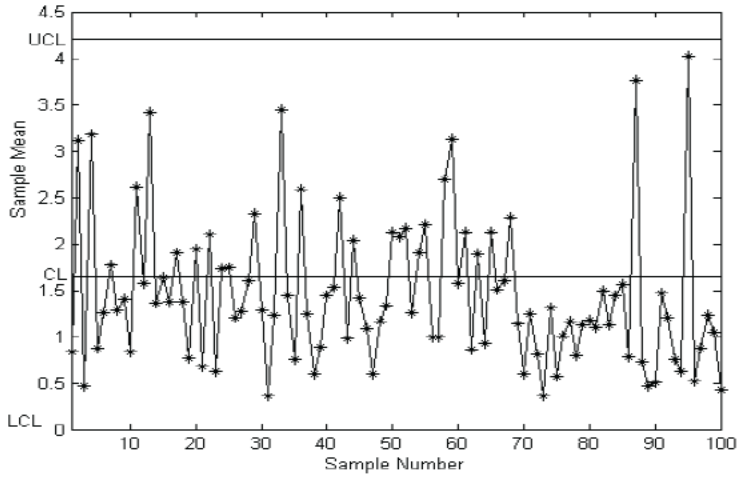

Figure 4: Shewhart chart for an out-of-control heavy tailed distribution

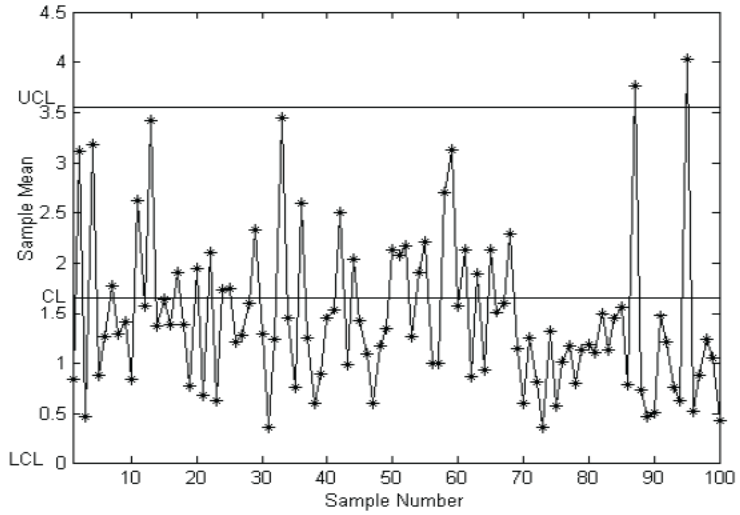

Figure 5: LAD-based chart for an out-of-control heavy tailed distribution

5. CONCLUSION

A control chart that is more effective than the standard Shewhart chart in detecting shifts in the process for heavy tailed distributions is proposed. This chart uses the standard error of the least absolute deviation to estimate the process variability. We use the least absolute deviation estimator because it provides more information about the process than the least squares estimator when the process follows heavy tailed distributions.

6. REFERENCES

[1] **Ferrell, E.B.** 1953. Control charts using midranges and medians. *Industrial Quality Control*, **9**, 30-34.

[2] **Gourieroux, C. and Monfort, A.** 1995. *Statistics and econometric models*, **1**. Cambridge University Press, Cambridge.

[3] **Johnson, N. L. and Kotz, S.** 1970. *Continuous univariate distributions*, Vols. **1 & 2.** Houghton Mifflin Company, Boston.

[4] **Langenberg, P. and Iglewicz, B.** 1986. Trimmed mean $\overline{X}$ and R charts, *Journal of Quality Technology*, **18**, 152-161.

[5] **Page, E.S.** 1961. Cumulative sum charts. *Technometrics*, **3**, 1-9.

[6] **Roberts, S.W.** 1959. Control charts test based on geometric moving averages. *Technometrics*, **1**, 239-250.

[7] **Rocke, D.M.** 1989. Robust control charts. *Technometrics*, **31**, 173-184.

[8] **Shewhart, W.A.** 1931. *Economic control of quality of manufactured product*. Van Nortrand Inc., New York.

[9] **Thaga, K.** 2008. Control chart for autocorrelated processes with heavy tailed distributions. *Economic Quality Control*. **23**, 1-10.

[10] **Thavaneswaran, A. and Heyde, C.C.** 1999. Prediction via estimating functions. *Journal of Statistical Planning and Inference*, **77**, 89-101.

[11] **White, E.M. and Schroeder, R.** 1987. A simultaneous control chart. *Journal of Quality Technology*, **19**, 1-10.