APPENDIX C - Analytical Derivation of Decision Interval

The Cumulative Sum (CUSUM) statistical process control chart is used to detect small changes in the mean of a random process. For quality control applications, it is desirable to detect any changes in the process mean as quickly as possible. For example, a manufacturing process may experience a change in mean as a result of tool wear, breakage, or adjustment, or any number of other unknown causes. The process is more likely to produce a product that does not meet quality specifications while the process mean is operating at its changed value. The product that does not meet quality specifications represents a financial loss to the company in terms of scrap product or re-working costs.

Methods that attempt to detect a change in a random process can sometimes signal that a change may have occurred, when in fact the process is still in-control. This is referred to as a false alarm. The probability of a false alarm occurring is sometimes referred to as Type I error. A false alarm in a manufacturing process can also represent financial loss for a company. If the company halts the process to search for a potential change that does not exist, the company is still paying for labor and overhead, while no product is being produced. Therefore, quality engineers must strike a balance between the probability of false alarm and the rapid detection of changes.

The determination of an appropriate balance between false alarm and rapid detection requires an expression that relates the probability of false alarm with control chart parameters. This was easily done with the Shewhart (1927) X-bar control chart, where an observation was compared against decision intervals set at $\mu \pm L\sigma$, where μ is the mean of the process, σ is the standard deviation, and L is the width parameter. The probability of false alarm, α, can be calculated from the expression, $\alpha = 2 * \int_L^\infty f(x)dx$, where $f(x)$ is the assumed symmetric probability density function of the process. The CUSUM control chart (Page 1954), on the other hand, was derived from the sequential probability ratio test (Wald, 1947), therefore the control chart statistic at each time point is conditioned on the previous time points. The CUSUM control chart statistic is given by, $C_t = \max\{0, Z_t - k + C_{t-1}\}$, where Z_t is the standardized observation at time t and k is an optimality constant. When the value of $C_t > h$, where h is the control chart's decision interval, the chart signals that a change in the process mean may have occurred. As a result of the max operator and the C_{t-1} expression, an analytical expression would somehow need to account for the nested conditional probability and the results are likely to be non-intuitive.

Several attempts have been made to provide quality engineers with insight into understanding the false alarm probability of the CUSUM. In situations where it is not necessary to know the precise probability of false alarm, but acceptable and rejectable quality levels have been established, an expression for the value k can be determined (Kemp, 1967). The optimal value of k is $k = (m_a + m_r)/2$, where $m \le m_a$ is an acceptable value from the random process, m, and $m \ge m_r$ is a rejectable value of m. It

has also been shown that the CUSUM is the most powerful test for detecting a change in the process mean of $2*k*\sigma$ (Moustakides, 2002). These results for the parameter k still do not provide a relationship between the parameter k, the decision interval, h, and the probability of false alarm. Expressions have been proposed that relate k, h, and a Brownian approximation to the expected number of observations until an in-control process signals a false alarm (Nadler and Robbins, 1971; Reynolds, 1975). This approximation was shown to overestimate the probability of false alarm (Reynolds, 1975). Thus, an accurate relationship between k, h, and the probability of false alarm has not yet been proposed.

In this chapter, Monte Carlo simulation is used to simulate the performance of the CUSUM on a random process consisting of independent and identically distributed observations for a range of false alarm probabilities between 0.001 and 0.05. A hybrid function is fit to the simulated values. The function provides a good fit to the simulated data with an R^2 value of 99.07%. Methods for using the newly proposed function for establishing CUSUM control chart parameters are discussed.

C.1 Method

Monte Carlo simulation was used to estimate the expected number of observations until an in-control CUSUM control chart signaled a false alarm. Results from the simulation were averaged over 100,000 independently seeded runs. Values of k ranged from 0.05 to 1.25 in increments of 0.05. Values for h ranged from 3.0 to 5.0. The specific values of h were adjusted for each setting of k to produce an expected number of observations until false alarm that fell within a range of 20 to 1000. This range was chosen for pragmatic reasons. The standard deviation of the number of observations until false alarm can be almost as large as the expected number of observations for an in-control process (Ewan and Kemp, 1960; Brook and Evans, 1972). Expected number of observations exceeding 1000 would, therefore, have such a deviation in the probability of false alarm as to be impractical. I submit that most practical applications using a probability of false alarm between 0.005 and 0.05.

The simulated data was plotted on a contour plot to observe trends in the data. The dependent variable was the decision interval. The optimality parameter, k, was an independent variable, and the expected number of observations until false alarm were contours. Initial observation suggested that the data exhibited the characteristics of exponential growth, where increasing h and k would lead to significant increases in the number of observations until false alarm. Three candidate functions were investigated: the exponential, power, and logarithmic functions, given by,

$$h(k) = \beta_1 e^{\beta_2 k}, \tag{1}$$

$$h(k) = \beta_1 k^{\beta_2}, \tag{2}$$

$$h(k) = \beta_1 Ln(k) + \beta_2. \tag{3}$$

Each function was fit to the data by the method of least square error. After analyzing the sum of square error and the R^2 values for the three functions, it was found

that the exponential and logarithmic functions estimated the decision interval better for the lower numbers of observation until false alarm, and the power function estimated the larger values better. In order to create a consistent function that provided accurate estimations for all values, a combination of the functions, or hybrid function was constructed to fit the full range of simulated data.

Several combinations of functions were investigated. These include linear combinations of two functions. The first hybrid function was the combination of the exponential and power function, given by,

$$h(k) = \left(\frac{1}{\beta_1 k}\right) e^{\beta_2 k} + \beta_3 k^{\beta_4}. \tag{4}$$

The alteration of the coefficient of the exponential function to $(1/\beta_1 k)$ allows the exponential portion of the equation to become insignificant as the optimality constant increases. This was applied because small changes in the *k* value cause larger changes in the decision interval as the expected number of observations until false alarm gets larger. This function was also fit to the data by the method of least squares.

A second hybrid function was created to combine the logarithmic and power functions. This function is also a linear combination of two functions, where the coefficient of the logarithmic function is replaced with the value $(1/\beta_1 k)$ and is given by,

$$h(k) = \left(\frac{1}{\beta_1 k}\right) \ln(k) + \beta_2 k^{\beta_3} \tag{5}$$

Both of these hybrid functions were designed to incorporate the strengths of both functions involved in the linear combination to create the best estimate of *h* based on the simulated data.

C.2 Results

The logarithmic-power hybrid combination provided the best fit to the simulated data. The function for the decision interval is therefore given by,

$$h(\lambda, k) = \left(\frac{\lambda^{0.1}}{5k}\right) \ln(k) + \left(0.53 \ln(\lambda) + \left(\frac{\pi}{10}\right)\right) k^{-0.89}, \tag{6}$$

where λ represents the expected number of observations until a false alarm occurs. Since all of the simulated data was used to fit the function, the performance of the function was evaluated using a 10-fold cross validation. The coefficient of determination, R^2, ranged from 98.79% to 99.81% with an average value of 99.07% across the 10 folds. The newly proposed function therefore provides a good approximation of the required decision

interval based on the optimality parameter and expected number of observations until a false alarm over a wide range of potential values.

The function can also be expressed in terms of the probability of false alarm. The probability of false alarm, α, is equivalent to the reciprocal of the expected number of observations until a false alarm, $\alpha = 1/\lambda$. Substituting α, in Equation 6 and simplifying provides an alternate expression for the decision interval given by,

$$h(\alpha,k) = \left(\frac{1}{5k\alpha^{0.1}}\right)\ln(k) - \left(0.53\ln(\alpha) + \left(\frac{\pi}{10}\right)\right)k^{-0.89}. \qquad (7)$$

This alternate expression provides an estimate for the decision interval based on a desired optimality parameter and the probability of false alarm. Both expressions are equivalent.

C.3 Discussion

A quality engineer can determine the parameters of the CUSUM control chart for a specific application, using the newly proposed expression in Equation 6 or Equation 7. First the engineer should investigate the costs associated with a change in the process. Does the product need to be scrapped, reworked, or sold at a lower price for certain quality characteristics? Based on these costs the engineer can determine the maximum acceptable and minimum rejectable process means for the process. The optimality parameter, k, should be set half way between these values and expressed in standardized units according to Ewan and Kemp (1960). The engineer must then decide on an acceptable risk level for false alarms. The engineer should choose a probability of false alarm, α, between 0.001 and 0.05 for most applications. The engineer could alternatively choose an expected number of observations until a false alarm, λ, if this is more intuitive for deciding upon a value. The choice of λ should be between 20 and 1000. Finally, the quality engineer can use the expressions proposed in this paper to determine an appropriate decision interval, h, for the CUSUM control chart.

Without these expressions, the quality engineer would either look up candidate values for h, k, and λ, as published in Van Dobben de Bruyn (1968), Nadler and Robbins (1971), Bagshaw and Johnson (1975), Vance (1986), Fellner (1990), Luceno (1999), or McCulloh (2004). Unfortunately, the published values of h, k, and λ, do not conform to the more ideal values associated with a specific process as determined in the method described above or according to the procedure laid out by Kemp (1962). Alternatively, a quality engineer could use an expression that approximates λ, under certain conditions as published in Reynolds (1975), or Luceno and Puig-Pey (2000). The expressions presented here more accurately estimate h for a range of k spanning 0.05 to 1.25, and a range of λ spanning 20 to 1000.

Perhaps the most useful applications of the newly proposed expressions are in software used to automate statistical process control. Many manufacturing processes include automated sampling, measurement, and control charting. The specific parameters for the process are still usually set by the quality engineer, however. With the analytical

expression for the decision interval, the parameterization can be automated as well. In addition, statistical process control is finding applications outside of manufacturing. The CUSUM has been used to identify changes in the organizational behavior of Al-Qaeda[8] (McCulloh et al, 2007), shifts in the membership commitment within on-line communities of practice (Galbreath, 2008), and changes within the semantic content of e-mail messages in the Enron corpus (McCulloh et al, 2008). For these new applications of statistical process control, determining appropriate control chart parameters is less clear. An analytic expression for the decision interval is necessary to broaden the potential application areas for statistical process control in general and for the CUSUM in particular.

[8] The Shewhart X-bar and Cumulative Sum statistical process control charts have been implemented in the software package Organizational Risk Analyzer (ORA) available from the Center for Computational Analysis of Social and Organizational Systems (CASOS) at Carnegie Mellon University, www.casos.cs.cmu.edu. ORA is used for social and dynamic network analysis.