# Graduate Lectures and Problems in Quality Control and Engineering Statistics: Theory and Methods

To Accompany

*Statistical Quality Assurance Methods for Engineers*

by

Vardeman and Jobe

Stephen B. Vardeman

V2.0: January 2001

# Chapter 2

# Process Monitoring

Chapters 3 and 4 of V&J discuss methods for process monitoring. The key concept there regarding the probabilistic description of monitoring schemes is the run length idea introduced on page 91 and specifically in display (3.44). Theory for describing run lengths is given in V&J only for the very simplest case of geometrically distributed $T$. This chapter presents some more general tools for the analysis/comparison of run length distributions of monitoring schemes, namely discrete time finite state Markov chains and recursions expressed in terms of integral (and difference) equations.

## 2.1 Some Theory for Stationary Discrete Time Finite State Markov Chains With a Single Absorbing State

These are probability models for random systems that at times $t = 1, 2, 3\ldots$ can be in one of a finite number of states

$$\mathrm{S}_1, \mathrm{S}_2, \ldots, \mathrm{S}_m, \mathrm{S}_{m+1} \ .$$

The "Markov" assumption is that the conditional distribution of where the system is at time $t+1$ given the entire history of where it has been up through time $t$ only depends upon where it is at time $t$. (In colloquial terms: The conditional distribution of where I'll be tomorrow given where I am and how I got here depends only on where I am, not on how I got here.) So called "stationary" Markov Chain (MC) models employ the assumption that movement between states from any time $t$ to time $t+1$ is governed by a (single) matrix of (one-step) "transition probabilities" (that is independent of $t$)

$$\underset{(m+1)\times(m+1)}{\boldsymbol{P}} = (p_{ij})$$

where

$$p_{ij} = P[\text{system is in } \mathrm{S}_j \text{ at time } t+1 \,|\, \text{system is in } \mathrm{S}_i \text{ at time } t] \ .$$

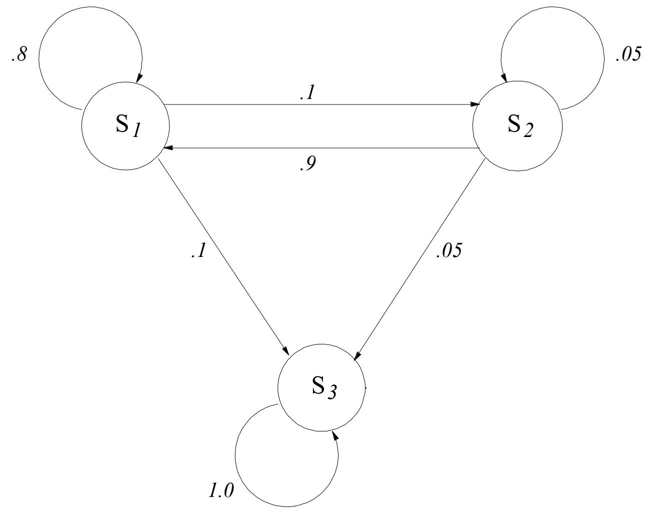


Figure 2.1: Schematic for a MC with Transition Matrix (2.1)

As a simple example of this, consider the transition matrix

$$\underset{3\times 3}{\boldsymbol{P}} \doteq \begin{pmatrix} .8 & .1 & .1 \\ .9 & .05 & .05 \\ 0 & 0 & 1 \end{pmatrix} . \tag{2.1}$$

Figure 2.1 is a useful schematic representation of this model.

The Markov Chain represented by Figure 2.1 has an interesting property. That is, while it is possible to move back and forth between states 1 and 2, once the system enters state 3, it is "stuck" there. The standard jargon for this property is to say that $S_3$ is an *absorbing state*. (In general, if $p_{ii} = 1$, $S_i$ is called an absorbing state.)

Of particular interest in applications of MCs to the description of process monitoring schemes are chains with a single absorbing state, say $S_{m+1}$, where it is possible to move (at least eventually) from any other state to the absorbing state. One thing that makes these chains so useful is that it is very easy to write down a matrix formula for a vector giving the mean number of transitions required to reach $S_{m+1}$ from any of the other states. That is, with

$$L_i = \text{the mean number of transitions required to move from } S_i \text{ to } S_{m+1} ,$$

$$\underset{m\times 1}{\boldsymbol{L}} = \begin{pmatrix} L_1 \\ L_2 \\ \vdots \\ L_m \end{pmatrix} , \quad \underset{(m+1)\times(m+1)}{\boldsymbol{P}} = \begin{pmatrix} \underset{m\times m}{\boldsymbol{R}} & \underset{m\times 1}{\boldsymbol{r}} \\ \underset{1\times m}{\boldsymbol{0}} & \underset{1\times 1}{1} \end{pmatrix} , \quad \text{and} \quad \underset{m\times 1}{\boldsymbol{1}} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

it is the case that

$$\boldsymbol{L} = (\boldsymbol{I} - \boldsymbol{R})^{-1}\boldsymbol{1} . \tag{2.2}$$

To argue that display (2.2) is correct, note that the following system of $m$ equations "clearly" holds:

$$\begin{aligned} L_1 &= (1+L_1)p_{11} + (1+L_2)p_{12} + \cdots + (1+L_m)p_{1m} + 1 \cdot p_{1,m+1} \\ L_2 &= (1+L_1)p_{21} + (1+L_2)p_{22} + \cdots + (1+L_m)p_{2m} + 1 \cdot p_{2,m+1} \\ &\vdots \\ L_m &= (1+L_1)p_{m1} + (1+L_2)p_{m2} + \cdots + (1+L_m)p_{mm} + 1 \cdot p_{m,m+1} \ . \end{aligned}$$

But this set is equivalent to the set

$$\begin{aligned} L_1 &= 1 + p_{11}L_1 + p_{12}L_2 + \cdots + p_{1m}L_m \\ L_2 &= 1 + p_{21}L_1 + p_{22}L_2 + \cdots + p_{2m}L_m \\ &\vdots \\ L_m &= 1 + p_{m1}L_1 + p_{m2}L_2 + \cdots + p_{mm}L_m \end{aligned}$$

and in matrix notation, this second set of equations is

$$\boldsymbol{L} = \boldsymbol{1} + \boldsymbol{R}\boldsymbol{L} \ . \tag{2.3}$$

So

$$\boldsymbol{L} - \boldsymbol{R}\boldsymbol{L} = \boldsymbol{1} \ ,$$

i.e.

$$(\boldsymbol{I} - \boldsymbol{R})\boldsymbol{L} = \boldsymbol{1} \ .$$

Under the conditions of the present discussion it is the case that $(\boldsymbol{I} - \boldsymbol{R})$ is guaranteed to be nonsingular, so that multiplying both sides of this matrix equation by the inverse of $(\boldsymbol{I} - \boldsymbol{R})$ one finally has equation (2.2).

For the simple 3-state example with transition matrix (2.1) it is easy enough to verify that with

$$\boldsymbol{R} = \begin{pmatrix} .8 & .1 \\ .9 & .05 \end{pmatrix}$$

one has

$$(\boldsymbol{I} - \boldsymbol{R})^{-1}\boldsymbol{1} = \begin{pmatrix} 10.5 \\ 11 \end{pmatrix} \ .$$

That is, the mean number of transitions required for absorption (into $S_3$) from $S_1$ is 10.5 while the mean number required from $S_2$ is 11.0.

When one is working with numerical values in $\boldsymbol{P}$ and thus wants numerical values in $\boldsymbol{L}$, the matrix formula (2.2) is most convenient for use with numerical analysis software. When, on the other hand, one has some algebraic expressions for the $p_{ij}$ and wants algebraic expressions for the $L_i$, it is usually most effective to write out the system of equations represented by display (2.3) and to try and see some slick way of solving for an $L_i$ of interest.

It is also worth noting that while the discussion in this section has centered on the computation of mean times to absorption, other properties of "time to absorption" variables can be derived and expressed in matrix notation. For example, Problem 2.22 shows that it is fairly easy to find the variance (or standard deviation) of time to absorption variables.

## 2.2 Some Applications of Markov Chains to the Analysis of Process Monitoring Schemes

When the "current condition" of a process monitoring scheme can be thought of as discrete random variable (with a finite number of possible values), because

1. the variables $Q_1, Q_2, ...$ fed into it are intrinsically discrete (for example representing counts) and are therefore naturally modeled using a discrete probability distribution (and the calculations prescribed by the scheme produce only a fixed number of possible outcomes),

2. "discretization" of the $Q$'s has taken place as a part of the development of the monitoring scheme (as, for example, in the "zone test" schemes outlined in Tables 3.5 through 3.7 of V&J), or

3. one approximates continuous distributions for $Q$'s and/or states of the scheme with a "finely-discretized" version in order to approximate exact (continuous) run length properties,

one can often apply the material of the previous section to the prediction of scheme behavior. (This is possible when the evolution of the monitoring scheme can be thought of in terms of movement between "states" where the conditional distribution of the next "state" depends only on a distribution for the next $Q$ which itself depends only on the current "state" of the scheme.) This section contains four examples of what can be done in this direction.

As an initial simple example, consider the simple monitoring scheme (suggested in the book *Sampling Inspection and Quality Control* by Wetherill) that signals an alarm the first time

1. a single point $Q$ plots "outside 3 sigma limits," or

2. two consecutive $Q$'s plot "between 2 and 3 sigma limits."

(This is a simple competitor to the sets of alarm rules specified in Tables 3.5 through 3.7 of V&J.) Suppose that one assumes that $Q_1, Q_2, \ldots$ are iid and

$$q_1 = P[Q_1 \text{ plots outside 3 sigma limits}]$$

and

$$q_2 = P[Q_1 \text{ plots between 2 and 3 sigma limits}] \ .$$

Then one might think of describing the evolution of the monitoring scheme with a 3-state MC with states

$$\begin{aligned} \mathrm{S}_1 &= \text{"all is OK,"} \\ \mathrm{S}_2 &= \text{"no alarm yet and the current } Q \text{ is between 2 and 3 sigma limits," and} \\ \mathrm{S}_3 &= \text{"alarm."} \end{aligned}$$

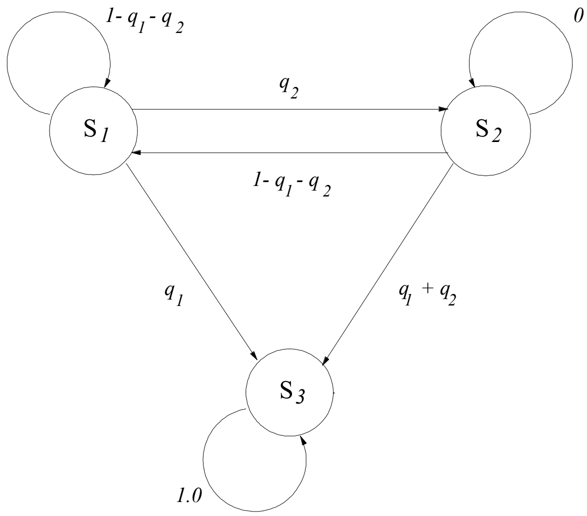


Figure 2.2: Schematic for a MC with Transition Matrix (2.4)

For this representation, an appropriate transition matrix is

$$\boldsymbol{P} = \begin{pmatrix} 1-q_1-q_2 & q_2 & q_1 \\ 1-q_1-q_2 & 0 & q_1+q_2 \\ 0 & 0 & 1 \end{pmatrix} \tag{2.4}$$

and the ARL of the scheme (under the iid model for the $Q$ sequence) is $L_1$, the mean time to absorption into the alarm state from the "all-OK" state. Figure 2.2 is a schematic representation of this scenario.

It is worth noting that a system of equations for $L_1$ and $L_2$ is

$$\begin{aligned} L_1 &= 1 \cdot q_1 + (1+L_2)q_2 + (1+L_1)(1-q_1-q_2) \\ L_2 &= 1 \cdot (q_1+q_2) + (1+L_1)(1-q_1-q_2) \ , \end{aligned}$$

which is equivalent to

$$\begin{aligned} L_1 &= 1 + L_1 \cdot (1-q_1-q_2) + L_2 q_2 \\ L_2 &= 1 + L_1(1-q_1-q_2) \ , \end{aligned}$$

which is the "non-matrix version" of the system (2.3) for this example. It is easy enough to verify that this system of two linear equations in the unknowns $L_1$ and $L_2$ has a (simultaneous) solution with

$$L_1 = \frac{1+q_2}{1-(1-q_1-q_2)-q_2(1-q_1-q_2)} \ .$$

As a second application of MC technology to the analysis of a process monitoring scheme, we will consider a so-called "Run-Sum" scheme. To define such a scheme, one begins with "zones" for the variable $Q$ as indicated in Figure 3.9 of V&J. Then "scores" are defined for various possible values of $Q$. For $j = 0, 1, 2$ a score of $+j$ is assigned to the eventuality that $Q$ is in the "positive $j$-sigma to $(j+1)$-sigma zone," while a score of $-j$ is assigned to the eventuality that $Q$ is in the "negative $j$-sigma to $(j+1)$-sigma zone." A score of $+3$ is assigned to any $Q$ above the "upper 3-sigma limit" while a score of $-3$ is assigned to any $Q$ below the "lower 3-sigma limit." Then, for the variables $Q_1, Q_2, \ldots$ one defines corresponding scores $Q_1^*, Q_2^*, \ldots$ and "run sums" $R_1, R_2, \ldots$ where

$$R_i = \text{"the 'sum' of scores } Q^* \text{ through time } i \text{ under the provision that a new sum is begun whenever a score is observed with a sign different from the existing Run-Sum."}$$

(Note, for example, that a new score of $Q^* = +0$ will reset a current Run-Sum of $R = -2$ to $+0$.) The Run-Sum scheme then signals at the first $i$ for which $|Q_i^*| = 3$ or $|R_i| \geq 4$.

Then define states for a Run-Sum process monitoring scheme

$$\begin{aligned}
\text{S}_1 &= \text{"no alarm yet and } R = -0,\text{"} \\
\text{S}_2 &= \text{"no alarm yet and } R = -1,\text{"} \\
\text{S}_3 &= \text{"no alarm yet and } R = -2,\text{"} \\
\text{S}_4 &= \text{"no alarm yet and } R = -3,\text{"} \\
\text{S}_5 &= \text{"no alarm yet and } R = +0,\text{"} \\
\text{S}_6 &= \text{"no alarm yet and } R = +1,\text{"} \\
\text{S}_7 &= \text{"no alarm yet and } R = +2,\text{"} \\
\text{S}_8 &= \text{"no alarm yet and } R = +3,\text{" and} \\
\text{S}_9 &= \text{"alarm."}
\end{aligned}$$

If one assumes that the observations $Q_1, Q_2, \ldots$ are iid and for $j = -3, -2, -1, -0, +0, +1, +2, +3$ lets

$$q_j = P[Q_1^* = j] \ ,$$

an appropriate transition matrix for describing the evolution of the scheme is

$$\boldsymbol{P} = \begin{pmatrix}
q_{-0} & q_{-1} & q_{-2} & 0 & q_{+0} & q_{+1} & q_{+2} & 0 & q_{-3}+q_{+3} \\
0 & q_{-0} & q_{-1} & q_{-2} & q_{+0} & q_{+1} & q_{+2} & 0 & q_{-3}+q_{+3} \\
0 & 0 & q_{-0} & q_{-1} & q_{+0} & q_{+1} & q_{+2} & 0 & q_{-3}+q_{-2}+q_{+3} \\
0 & 0 & 0 & q_{-0} & q_{+0} & q_{+1} & q_{+2} & 0 & q_{-3}+q_{-2}+q_{-1}+q_{+1} \\
q_{-0} & q_{-1} & q_{-2} & 0 & q_{+0} & q_{+1} & q_{+2} & 0 & q_{-3}+q_{+3} \\
q_{-0} & q_{-1} & q_{-2} & 0 & 0 & q_{+0} & q_{+1} & q_{+2} & q_{-3}+q_{+3} \\
q_{-0} & q_{-1} & q_{-2} & 0 & 0 & 0 & q_{+0} & q_{+1} & q_{-3}+q_{+2}+q_{+3} \\
q_{-0} & q_{-1} & q_{-2} & 0 & 0 & 0 & 0 & q_{+0} & q_{-3}+q_{+1}+q_{+2}+q_{+3} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{pmatrix}$$

and the ARL for the scheme is $L_1 = L_5$. (The fact that the 1st and 5th rows of $\boldsymbol{P}$ are identical makes it clear that the mean times to absorption from $\text{S}_1$ and $\text{S}_5$

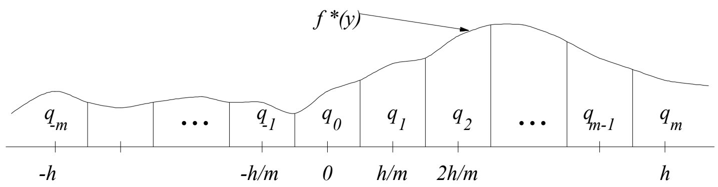


Figure 2.3: Notational Conventions for Probabilities from Rounding $Q - k_1$ Values

must be the same.) It turns out that clever manipulation with the "non-matrix" version of display (2.3) in this example even produces a fairly simple expression for the scheme's ARL. (See Problem 2.24 and Reynolds (1971 *JQT*) and the references therein in this final regard.)

To turn to a different type of application of the MC technology, consider the analysis of a high side decision interval CUSUM scheme as described in §4.2 of V&J. Suppose that the variables $Q_1, Q_2, \ldots$ are iid with a continuous distribution specified by the probability density $f(y)$. Then the variables $Q_1 - k_1, Q_2 - k_1, Q_3 - k_1, \ldots$ are iid with probability density $f^*(y) = f(y + k_1)$. For a positive integer $m$, we will think of replacing the variables $Q_i - k_1$ with versions of them rounded to the nearest multiple of $h/m$ before CUSUMing. Then the CUSUM scheme can be thought of in terms of a MC with states

$$\mathrm{S}_i = \text{“no alarm yet and the current CUSUM is } (i-1)\left(\frac{h}{m}\right)\text{”}$$

for $i = 1, 2, \ldots, m$ and

$$\mathrm{S}_{m+1} = \text{“alarm.”}$$

Then let

$$q_{-m} = \int_{-\infty}^{-h+\frac{1}{2}\left(\frac{h}{m}\right)} f^*(y)dy = P[Q_1 - k_1 \leq -h + \frac{1}{2}\left(\frac{h}{m}\right)] \ ,$$

$$q_m = \int_{h-\frac{1}{2}\left(\frac{h}{m}\right)}^{\infty} f^*(y)dy = P[h - \frac{1}{2}\left(\frac{h}{m}\right) < Q_1 - k_1] \ ,$$

and for $-m < j < m$ take

$$q_j = \int_{j\left(\frac{h}{m}\right)-\frac{1}{2}\left(\frac{h}{m}\right)}^{j\left(\frac{h}{m}\right)+\frac{1}{2}\left(\frac{h}{m}\right)} f^*(y)dy \ . \tag{2.5}$$

These notational conventions for probabilities $q_{-m}, \ldots, q_m$ are illustrated in Figure 2.3.

In this notation, the evolution of the high side decision interval CUSUM scheme can then be described in approximate terms by a MC with transition matrix

$$\underset{(m+1)\times(m+1)}{\boldsymbol{P}} = \begin{pmatrix} \sum_{j=-m}^{0} q_j & q_1 & q_2 & \cdots & q_{m-1} & q_m \\ \sum_{j=-m}^{-1} q_j & q_0 & q_1 & \cdots & q_{m-2} & q_{m-1}+q_m \\ \sum_{j=-m}^{-2} q_j & q_{-1} & q_0 & \cdots & q_{m-3} & q_{m-2}+q_{m-1}+q_m \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ q_{-m}+q_{-m+1} & q_{-m+2} & q_{-m+3} & \cdots & q_0 & \sum_{j=1}^{m} q_j \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix} .$$

For $i = 1, \ldots, m$ the mean time to absorption from state $\mathrm{S}_i$ ($L_i$) is approximately the ARL of the scheme *with head start* $(i-1)\left(\frac{h}{m}\right)$. (That is, the entries of the vector $\boldsymbol{L}$ specified in display (2.2) are approximate ARL values for the CUSUM scheme using various possible head starts.) In practice, in order to find ARLs for the original scheme with non-rounded iid observations $Q$, one would find approximate ARL values for an increasing sequence of $m$'s until those appear to converge for the head start of interest.

As a final example of the use of MC techniques in the probability modeling of process monitoring scheme behavior, consider discrete approximation of the EWMA schemes of §4.1 of V&J where the variables $Q_1, Q_2, \ldots$ are again iid with continuous distribution specified by a pdf $f(y)$. In this case, in order to provide a tractable discrete approximation, it will not typically suffice to simply discretize the variables $Q$ (as the EWMA calculations will then typically produce a number of possible/exact EWMA values that grows as time goes on). Instead, it is necessary to think directly in terms of rounded/discretized EWMAs. So for an odd positive integer $m$, let $\Delta = (UCL_{EWMA} - LCL_{EWMA})/m$ and think of replacing an (exact) EWMA sequence with a rounded EWMA sequence taking on values $a_i$ defined by

$$a_i \doteq LCL_{EWMA} + \frac{\Delta}{2} + (i-1)\Delta$$

for $i = 1, 2, \ldots, m$. For $i = 1, 2, ..., m$ let

$$\mathrm{S}_i = \text{“no alarm yet and the rounded EWMA is } a_i\text{”}$$

and

$$\mathrm{S}_{m+1} = \text{“alarm.”}$$

And for $1 \leq i, j \leq m$, let

$$\begin{aligned} q_{ij} &= P[\text{moving from S}_i \text{ to S}_j] \ , \\ &= P[a_j - \frac{\Delta}{2} \leq (1-\lambda)a_i + \lambda Q \leq a_j + \frac{\Delta}{2}] \ , \\ &= P[\frac{a_j - (1-\lambda)a_i}{\lambda} - \frac{\Delta}{2\lambda} \leq Q \leq \frac{a_j - (1-\lambda)a_i}{\lambda} + \frac{\Delta}{2\lambda}] \ , \\ &= P[a_i + \frac{(j-i)\Delta}{\lambda} - \frac{\Delta}{2\lambda} \leq Q \leq a_i + \frac{(j-i)\Delta}{\lambda} + \frac{\Delta}{2\lambda}] \ , \\ &= \int_{a_i + \frac{(j-i)\Delta}{\lambda} - \frac{\Delta}{2\lambda}}^{a_i + \frac{(j-i)\Delta}{\lambda} + \frac{\Delta}{2\lambda}} f(y)dy \ . \end{aligned} \tag{2.6}$$

Then with

$$\boldsymbol{P} = \begin{pmatrix} q_{11} & q_{12} & \cdots & q_{1m} & 1 - \sum_{j=1}^{m} q_{1j} \\ q_{21} & q_{22} & \cdots & q_{2m} & 1 - \sum_{j=1}^{m} q_{2j} \\ \vdots & \vdots & & \vdots & \vdots \\ q_{m1} & q_{m2} & \cdots & q_{mm} & 1 - \sum_{j=1}^{m} q_{mj} \\ 0 & 0 & \cdots & 0 & 1 \end{pmatrix}$$

the mean time to absorption from the state $\text{S}_{(m+1)/2}$ (the value $L_{(m+1)/2}$) of a MC with this transition matrix is an approximation for the EWMA scheme ARL with $EWMA_0 = (UCL_{EWMA} + LCL_{EWMA})/2$. In practice, in order to find the ARL for the original scheme, one would find approximate ARL values for an increasing sequence of $m$'s until those appear to converge.

The four examples in this section have illustrated the use of MC calculations in the second and third of the two circumstances listed at the beginning of this section. The first circumstance is conceptually the simplest of the three, and is for example illustrated by Problems 2.25, 2.28 and 2.37. The examples have also all dealt with iid models for the $Q_1, Q_2, \ldots$ sequence. Problem 2.26 shows that the methodology can also easily accommodate some kinds of dependencies in the $Q$ sequence. (The discrete model in Problem 2.26 is itself perhaps less than completely appealing, but the reader should consider the possibility of discrete approximation of the kind of dependency structure employed in Problem 2.27 before dismissing the basic concept illustrated in Problem 2.26 as useless.)

## 2.3 Integral Equations and Run Length Properties of Process Monitoring Schemes

There is a second (and at first appearance quite different) standard method of approaching the analysis of the run length behavior of some process monitoring schemes where continuous variables $Q$ are involved. That is through the use of integral equations, and this section introduces the use of these. (As it turns out, by the time one is forced to find numerical solutions of the integral equations, there is not a whole lot of difference between the methods of this section and those of the previous one. But it is important to introduce this second point of view and note the correspondence between approaches.)

Before going to the details of specific schemes and integral equations, a small piece of calculus/numerical analysis needs to be reviewed and notation set for use in these notes. That concerns the approximation of definite integrals on the interval $[a, a+h]$. Specification of a set of points

$$a \leq a_1 \leq a_2 \leq \cdots \leq a_m \leq a+h$$

and weights

$$w_i \geq 0 \quad \text{with} \quad \sum_{i=1}^{m} w_i = h$$

so that

$$\int_a^{a+h} f(y)dy \quad \text{may be approximated as} \quad \sum_{i=1}^{m} w_i f(a_i)$$

for "reasonable" functions $f(y)$, is the specification of a so-called "quadrature rule" for approximating integrals on the interval $[a, a+h]$. The simplest of such rules is probably the choice

$$a_i \doteq a + \left(\frac{i-\frac{1}{2}}{m}\right) h \quad \text{with} \quad w_i \doteq \frac{h}{m} \ . \tag{2.7}$$

(This choice amounts to approximating an integral of $f$ by a sum of signed areas of rectangles with bases $h/m$ and (signed) heights chosen as the values of $f$ at midpoints of intervals of length $h/m$ beginning at $a$.)

Now consider a high side CUSUM scheme as in §4.2 of V&J, where $Q_1, Q_2, \ldots$ are iid with continuous marginal distribution specified by the probability density $f(y)$. Define the function

$$L_1(u) \doteq \text{the ARL of the high side CUSUM scheme using a head start of } u \ .$$

If one begins CUSUMing at $u$, there are three possibilities of where he/she will be after a single observation, $Q_1$. If $Q_1$ is large ($Q_1 - k_1 \geq h - u$) then there will be an immediate signal and the run length will be 1. If $Q_1$ is small ($Q_1 - k_1 \leq -u$) the CUSUM will "zero out," one observation will have been "spent," and on average $L_1(0)$ more observations are to be faced in order to produce a signal. Finally, if $Q_1$ is moderate ($-u < Q_1 - k_1 < h - u$) then one observation will have been spent and the CUSUM will continue from $u + (Q_1 - k_1)$, requiring on average an additional $L_1(u + (Q_1 - k_1))$ observations to produce a signal. This reasoning leads to the equation for $L_1$,

$$\begin{aligned} L_1(u) &= 1 \cdot P[Q_1 - k_1 \geq h - u] + (1 + L_1(0))P[Q_1 - k_1 \leq -u] \\ &\quad + \int_{k_1-u}^{k_1+h-u} (1 + L_1(u + y - k_1))f(y)dy \ . \end{aligned}$$

Writing $F(y)$ for the cdf of $Q_1$ and simplifying slightly, this is

$$L_1(u) = 1 + L_1(0)F(k_1 - u) + \int_0^h L_1(y)f(y + k_1 - u)dy \ . \qquad (2.8)$$

The argument leading to equation (2.8) has a twin that produces an integral equation for

$$L_2(v) \doteq \text{the ARL of a low side CUSUM scheme using a head start of } v \ .$$

That equation is

$$L_2(v) = 1 + L_2(0)\left(1 - F(k_2 - u)\right) + \int_{-h}^0 L_2(y)f(y + k_2 - v)dy \ . \qquad (2.9)$$

And as indicated in display (4.20) of V&J, could one solve equations (2.8) and (2.9) (and thus obtain $L_1(0)$ and $L_2(0)$) one would have not only separate high and low side CUSUM ARLs, but ARLs for some combined schemes as well. (Actually, more than what is stated in V&J can be proved. Yashchin in a *Journal of Applied Probability* paper in about 1985 showed that with iid $Q$'s, high side decision interval $h_1$ and low side decision interval $-h_2$ for nonnegative $h_2$, if $k_1 \geq k_2$ and

$$(k_1 - k_2) - |h_1 - h_1| \geq \max\left(0, u - v - \max(h_1, h_2)\right) \ ,$$

for the simultaneous use of high and low side schemes

$$ARL_{\text{combined}} = \frac{L_1(0)L_2(v) + L_1(u)L_2(0) - L_1(0)L_2(0)}{L_1(0) + L_2(0)} \ .$$

It is easily verified that what is stated on page 151 of V&J is a special case of this result.) So in theory, to find ARLs for CUSUM schemes one need "only" solve the integral equations (2.8) and (2.9). This is easier said than done. The one case where fairly explicit solutions are known is that where observations are exponentially distributed (see Problem 2.30). In other cases one must resort to numerical solution of the integral equations.

So consider the problem of approximate solution of equation (2.8). For a particular quadrature rule for integrals on $[0, h]$, for each $a_i$ one has from equation (2.8) the approximation

$$L_1(a_i) \approx 1 + L_1(a_1)F(k_1 - a_i) + \sum_{j=1}^m w_j L_1(a_j)f(a_j + k_1 - a_i) \ .$$

That is, at least approximately one has the system of $m$ linear equations

$$\begin{aligned}
L_1(a_1) &= 1 + L_1(a_1)[F(k_1 - a_1) + w_1 f(k_1)] + \sum_{j=2}^{m} L_1(a_j) w_j f(a_j + k_1 - a_1) \ , \\
L_1(a_2) &= 1 + L_1(a_1)[F(k_1 - a_2) + w_1 f(a_1 + k_1 - a_2)] + \sum_{j=2}^{m} L_1(a_j) w_j f(a_j + k_1 - a_2) \ , \\
&\vdots \\
L_1(a_m) &= 1 + L_1(a_1)[F(k_1 - a_m) + w_1 f(a_1 + k_1 - a_m)] + \sum_{j=2}^{m} L_1(a_j) w_j f(a_j + k_1 - a_m)
\end{aligned}$$

in the $m$ unknowns $L_1(a_1), \ldots, L_1(a_m)$. Again in light of equation (2.8) and the notion of numerical approximation of definite integrals, upon solving this set of equations (for approximate values of $(L_1(a_1), \ldots, L_1(a_m))$) one may approximate the function $L_1(u)$ as

$$L_1(u) \approx 1 + L_1(a_1) F(k_1 - u) + \sum_{j=1}^{m} w_j L_1(a_j) f(a_j + k_1 - u) \ .$$

It is a revealing point that the system of equations above is of the form (2.3) that was so useful in the MC approach to the determination of ARLs. That is, let

$$\boldsymbol{L} = \begin{pmatrix} L_1(a_1) \\ L_1(a_2) \\ \vdots \\ L_1(a_m) \end{pmatrix}$$

and

$$\boldsymbol{R} = \begin{pmatrix} F(k_1 - a_1) + w_1 f(k_1) & w_2 f(a_2 + k_1 - a_1) & \cdots & w_m f(a_m + k_1 - a_1) \\ F(k_1 - a_2) + w_1 f(a_1 + k_1 - a_2) & w_2 f(k_1) & \cdots & w_m f(a_m + k_1 - a_2) \\ \vdots & \vdots & & \vdots \\ F(k_1 - a_m) + w_1 f(a_1 + k_1 - a_m) & w_2 f(a_2 + k_1 - a_m) & \cdots & w_m f(k_1) \end{pmatrix}$$

and note that the set of equations for the "$a_i$ head start approximate ARLs" is exactly of the form (2.3). With the simple quadrature rule in display (2.7) note that a generic entry of $\boldsymbol{R}$, $r_{ij}$, for $j \geq 2$ is

$$r_{ij} = w_j f(a_j + k_1 - a_i) = \left(\frac{h}{m}\right) f\left((j-i)\left(\frac{h}{m}\right) + k_1\right) \ .$$

But using again the notation $f^*(y) = f(y + k_1)$ employed in the CUSUM example of §2.2, this means

$$r_{ij} = \left(\frac{h}{m}\right) f^*\left((j-i)\left(\frac{h}{m}\right)\right) \approx \int_{(j-i)\left(\frac{h}{m}\right) - \frac{1}{2}\left(\frac{h}{m}\right)}^{(j-i)\left(\frac{h}{m}\right) + \frac{1}{2}\left(\frac{h}{m}\right)} f^*(y) dy = q_{j-i}$$

(in terms of the notation (2.5) from the CUSUM example). The point is that whether one begins from a "discretize the $Q - k_1$ distribution and employ the MC material" point of view or from a "do numerical solution of an integral equation" point of view is largely immaterial. Very similar large systems of linear equations must be solved in order to find approximate ARLs.

As a second application of integral equation ideas to the analysis of process monitoring schemes, consider the EWMA schemes of §4.1 of V&J where $Q_1, Q_2, \ldots$ are iid with a continuous distribution specified by the probability density $f(y)$. Let

$$L(u) = \text{the ARL of a EWMA scheme with } EWMA_0 = u \ .$$

When one begins a EWMA sequence at $u$, there are 2 possibilities of where he/she will be after a single observation, $Q_1$. If $Q_1$ is extreme ($\lambda Q_1 + (1-\lambda)u > UCL_{EWMA}$ or $\lambda Q_1 + (1-\lambda)u < LCL_{EWMA}$) then there will be an immediate signal and the run length will be 1. If $Q_1$ is moderate ($LCL_{EWMA} \leq \lambda Q_1 + (1-\lambda)u \leq UCL_{EWMA}$) one observation will have been "spent" and on average $L(\lambda Q_1 + (1-\lambda)u)$ more observations are to be faced in order to produce a signal. Now the event

$$LCL_{EWMA} \leq \lambda Q_1 + (1-\lambda)u \leq UCL_{EWMA}$$

is the event

$$\frac{LCL_{EWMA} - (1-\lambda)u}{\lambda} \leq Q_1 \leq \frac{UCL_{EWMA} - (1-\lambda)u}{\lambda} \ ,$$

so this reasoning produces the equation

$$\begin{aligned} L(u) &= 1 \cdot \left(1 - P[\frac{LCL_{EWMA} - (1-\lambda)u}{\lambda} \leq Q_1 \leq \frac{UCL_{EWMA} - (1-\lambda)u}{\lambda}]\right) \\ &\quad + \int_{\frac{LCL_{EWMA}-(1-\lambda)u}{\lambda}}^{\frac{UCL_{EWMA}-(1-\lambda)u}{\lambda}} \left(1 + L(\lambda y + (1-\lambda)u)\right) f(y)dy \ , \end{aligned}$$

or

$$L(u) = 1 + \int_{\frac{LCL_{EWMA}-(1-\lambda)u}{\lambda}}^{\frac{UCL_{EWMA}-(1-\lambda)u}{\lambda}} L(\lambda y + (1-\lambda)u) f(y)dy \ ,$$

or finally

$$L(u) = 1 + \frac{1}{\lambda} \int_{LCL_{EWMA}}^{UCL_{EWMA}} L(y) f\left(\frac{y - (1-\lambda)u}{\lambda}\right) dy \ . \qquad (2.10)$$

As in the previous (CUSUM) case, one must usually resort to numerical methods in order to approximate the solution to equation (2.10). For a particular quadrature rule for integrals on $[LCL_{EWMA}, UCL_{EWMA}]$, for each $a_i$ one has from equation (2.10) the approximation

$$L(a_i) \approx 1 + \frac{1}{\lambda} \sum_{j=1}^{m} w_j L(a_j) f\left(\frac{a_j - (1-\lambda)a_i}{\lambda}\right) \ . \qquad (2.11)$$

Now expression (2.11) is standing for a set of $m$ equations in the $m$ unknowns $L(a_1), \ldots, L(a_m)$ that (as in the CUSUM case) can be thought of in terms of the matrix expression (2.3) if one takes

$$\boldsymbol{L} = \begin{pmatrix} L(a_1) \\ \vdots \\ L(a_m) \end{pmatrix} \quad \text{and} \quad \underset{m \times m}{\boldsymbol{R}} = \left( \frac{w_j f\left(\frac{a_j-(1-\lambda)a_i}{\lambda}\right)}{\lambda} \right) . \qquad (2.12)$$

Solution of the system represented by equation (2.11) or the matrix expression (2.3) with definitions (2.12) produces approximate values for $L(a_1), \ldots, L(a_m)$ and therefore an approximation for the function $L(u)$ as

$$L(u) \approx 1 + \frac{1}{\lambda}\sum_{j=1}^{m} w_j L(a_j) f\left(\frac{a_j - (1-\lambda)u}{\lambda}\right) .$$

Again as in the CUSUM case, it is worth noting the similarity between the set of equations used to find "MC" ARL approximations and the set of equations used to find "integral equation" ARL approximations. With the quadrature rule (2.7) and an odd integer $m$, using the notation $\Delta = (UCL_{EWMA} - LCL_{EWMA})/m$ employed in §2.2 in the EWMA example, note that a generic entry of $\boldsymbol{R}$ defined in (2.12) is

$$r_{ij} = \frac{w_j f\left(\frac{a_j-(1-\lambda)a_i}{\lambda}\right)}{\lambda} = \frac{\Delta f\left(a_i + \frac{(j-i)\Delta}{\lambda}\right)}{\lambda} \approx \int_{a_i+\frac{(j-i)\Delta}{\lambda}-\frac{\Delta}{2\lambda}}^{a_i+\frac{(j-i)\Delta}{\lambda}+\frac{\Delta}{2\lambda}} f(y)dy = q_{ij} \ ,$$

(in terms of the notation (2.6) from the EWMA example of §2.2). That is, as in the CUSUM case, the sets of equations used in the "MC" and "integral equation" approximations for the "$EWMA_0 = a_i$ ARLs" of the scheme are very similar.

As a final example of the use of integral equations in the analysis of process monitoring schemes, consider the $X/MR$ schemes of §4.4 of V&J. Suppose that observations $x_1, x_2, \ldots$ are iid with continuous marginal distribution specified by the probability density $f(y)$. Define the function

$L(y)$ = "the mean number of additional observations to alarm, given that there has been no alarm to date and the current observation is $y$."

Then note that as one begins $X/MR$ monitoring, there are two possibilities of where he/she will be after observing the first individual, $x_1$. If $x_1$ is extreme ($x_1 < LCL_x$ or $x_1 > UCL_x$) there will be an immediate signal and the run length will be 1. If $x_1$is not extreme ($LCL_x \le x_1 \le UCL_x$) one observation will have been spent and on average another $L(x_1)$ observations will be required in order to produce a signal. So it is reasonable that the ARL for the $X/MR$ scheme is

$$ARL = 1 \cdot (1 - P[LCL_x \le x_1 \le UCL_x]) + \int_{LCL_x}^{UCL_x} (1 + L(y))f(y)dy \ ,$$

that is

$$ARL = 1 + \int_{LCL_x}^{UCL_x} L(y)f(y)dy \ , \tag{2.13}$$

where it remains to find a way of computing the function $L(y)$ in order to feed it into expression (2.13).

In order to derive an integral equation for $L(y)$ consider the situation if there has been no alarm and the current individual observation is $y$. There are two possibilities of where one will be after observing one more individual, $x$. If $x$ is extreme or too far from $y$ ($x < LCL_x$ or $x > UCL_x$ or $|x - y| > UCL_R$) only one additional observation is required to produce a signal. On the other hand, if $x$ is not extreme and not too far from $y$ ($LCL_x \leq x \leq UCL_x$ and $|x - y| \leq UCL_R$) one more observation will have been spent and on average another $L(x)$ will be required to produce a signal. That is,

$$\begin{aligned} L(y) &= 1 \cdot (P[x < LCL_x \quad \text{or} \quad x > UCL_x \quad \text{or} \quad |x-y| > UCL_R]) \\ &\quad + \int_{\max(LCL_x, y-UCL_R)}^{\min(UCL_x, y+UCL_R)} (1 + L(x))f(x)dx \ , \end{aligned}$$

that is,

$$\begin{aligned} L(y) &= 1 + \int_{\max(LCL_x, y-UCL_R)}^{\min(UCL_x, y+UCL_R)} L(x)f(x)dx \\ &= 1 + \int_{LCL_x}^{UCL_x} I[|x-y| \leq UCL_R]L(x)f(x)dx \ . \end{aligned} \tag{2.14}$$

(The notation $I[A]$ is "indicator function" notation, meaning that when $A$ holds $I[A] = 1$, and otherwise $I[A] = 0$.) As in the earlier CUSUM and EWMA examples, once one specifies a quadrature rule for definite integrals on the interval $[LCL_x, UCU_x]$, this expression (2.14) provides a set of $m$ linear equations for approximate values of $L(a_i)$'s. When this system is solved, the resulting values can be fed into a discretized version of equation (2.13) and an approximate ARL produced. It is worth noting that the potential discontinuities of the integrand in equation (2.14) (produced by the indicator function) have the effect of making numerical solutions of this equation much less well-behaved than those for the other integral equations developed in this section.

The examples of this section have dealt only with ARLs for schemes based on (continuous) iid observations. It therefore should be said that:

1. The iid assumption can in some cases be relaxed to give tractable integral equations for situations where correlated sequences $Q_1, Q_2, \ldots$ are involved (see for example Problem 2.27),

2. Other descriptors of the run length distribution (beyond the ARL) can often be shown to solve simple integral equations (see for example the integral equations for CUSUM run length second moment and run length probability function in Problem 2.31), and

3. In some cases, with discrete variables $Q$ there are difference equation analogues of the integral equations presented here (that ultimately correspond to the kind of MC calculations illustrated in the previous section).