The Rate of False Signals in $\overline{X}$ Control Charts with Estimated Limits

DIANE P. BISCHAK

University of Calgary, Calgary, AB T2N 1N4, Canada

DAN TRIETSCH

American University of Armenia, Yerevan, Armenia

The in-control statistical properties of $\overline{X}$ charts have usually been studied from the perspective of the average run length (ARL) until the first (false) signal, known as the in-control ARL. We argue that the ARL is a confusing concept when used with charts with estimated limits and that the rate of false signals (RFS), which focuses on the behavior of charts during extended use, is more intuitive. We use the RFS to illustrate graphically the dangers of using too few subgroups to estimate control limits. We also discuss diffidence charts, which make the inherent uncertainty concerning RFS observable to the practitioner and thus help the practitioner determine what is an acceptable number of subgroups for a given charting application.

Key Words: Average Run Length; Control Limits; Diffidence; Estimated Parameters.

BECAUSE process parameters are seldom, if ever, known with certainty, the statistical properties of $\overline{X}$ control charts that are based on estimates of process parameters should be of particular interest to practitioners. Often in practice, the true process mean μ and standard deviation σ are estimated from some number m of initial subgroups of size n each. The upper and lower control limits, UCL and LCL, are then a function of these estimates, for example, $\overline{\overline{X}} \pm k\,\overline{S}\,/(c_4\sqrt{n})$, where $\overline{\overline{X}}$ is the average of the m subgroup averages $\overline{X}_i$, $i = 1, \ldots, m$, $\overline{S}$ is the average of the m subgroup standard deviations S_i, c_4 is the value such that $E[S_i/c_4] = \sigma$ for the given subgroup size n, and k is the desired number of standard deviations to use for process control. Thus the control limits are developed from the m subgroups gathered in Phase I, while the subsequent use of the chart to control the process takes place in Phase II.

From the practitioner's viewpoint, it would be desirable to use only a small number of subgroups to estimate μ and σ in Phase I and get on to Phase II, charting the process, as soon as possible. However, as a number of authors have discussed, basing the chart on estimates of μ and σ from a small number of subgroups may give rise to some unexpected and undesirable effects. An early paper addressing the problems associated with estimated control limits was Hillier (1964) (we will refer to this and related articles by Hillier (1969), Yang and Hillier (1970), and their predecessor Proschan and Savage (1960) as H&Y), which computes the probability that a Phase II subgroup mean will fall outside the control limits when the process is in control and shows that, for small m, this probability can be much larger than in the known-limits case. Thus, a chart with limits estimated from only a few subgroup means will tend, on average, to produce a greater number of false signals, chasing after which will add to the cost of production. Several authors have attempted to circumvent this problem; see the discussion of various procedures and their properties in Del Castillo et al. (1996) and, in particular, H&Y's computation of an adjusted constant for the number of process standard deviations enclosed by the control limits, the self-starting CUSUM charts of Hawkins (1987), the Q-charts of Quesenberry (1991) that allow charting to be done

Dr. Bischak is an Associate Professor in the Haskayne School of Business. Her email address is diane.bischak@haskayne.ucalgary.ca.

Dr. Trietsch is a Professor in the Department of Industrial Engineering. His email address is trietsch@aua.am.

from the beginning of data collection, the method of detecting a process shift proposed by Hawkins et al. (2003), the development of wider control limits for the exponentially weighted moving average (EWMA) chart by Jones (2002), and the recommended setting of control limits for multiple-valued EWMA charts given by Sullivan and Jones (2002).

Most papers following H&Y have studied the statistical properties of charts with limits estimated from small samples through examination of the *average run length*, or ARL, rather than the probability that H&Y considered. (In this paper, we address in-control, or false, signals only, so we reserve the term ARL for the in-control ARL.) We believe that this change in focus was detrimental and should be reversed. We feel that, in the estimated-limits case, a concept related to the probability studied by H&Y that we call the *rate of false signals*, or RFS, is more intuitive than the ARL. It is also a useful way to help determine a reasonable number of subgroups to sample in order to construct control limits. In the remainder of this paper, we show that ARL is a somewhat peculiar measure and is potentially confusing. We discuss, and reject, some historical arguments that were used to justify the focus on ARL and then discuss the properties of the distribution of the RFS. We also present a probability-based version of *diffidence charts*, which were introduced by Trietsch (1999) (henceforth, T99). Diffidence charts graphically demonstrate the amount of uncertainty hidden in estimated limits. Such charts can be useful to the practitioner in deciding whether to collect more data before using a particular chart. (Relevant downloadable files including excerpts from T99, a spreadsheet application discussed later, and previous results by these authors are given in https://webdisk.ucalgary.ca/~dbischak/JQT2007/BischakTrietsch.htm.)

The Average Run Length for Charts with Estimated Limits: A Source of Confusion

A summary of key notation and definitions for the remainder of the paper is given in Table 1. We will assume the following scenario for establishing an $\bar{X}$ control chart. (Although we focus on the $\bar{X}$ control

TABLE 1. A Summary of Key Notation

Notation	Definition
n	The sample size
m	The number of subgroups sampled in Phase I
k	The number of standard deviations used to set the control chart limits
$\bar{X}_i$	The ith subgroup mean
$\bar{\bar{X}}$	The average of the m subgroup means in Phase I
$\widehat{\text{LCL}}, \widehat{\text{UCL}}$	For a particular chart, control limits that are a function of estimated parameters and are therefore random variables
$\hat{l}, \hat{u}$	For a particular chart, numbers that are realizations of the control limits
L	The run length: when the process is in control, the number of subgroups after any given signal, up to and including the next signal that occurs
L_1	The first run length: the number of subgroups recorded from the beginning of chart use until a signal is first seen (including the signal itself) when the process is in control
$E_{\text{II}}[\cdot]$	An expectation taken over the distribution of the specified random variable in Phase II (before process parameters are estimated in Phase I)
$E_{\text{II}\|\text{I}}[\cdot]$	An expectation taken over the distribution of the specified random variable in Phase II conditional upon the values of the process parameter estimates obtained in Phase I
ARL	The average run length, defined as $E_{\text{II}}[L_1]$
RFS	The rate of false signals, a random variable defined as the rate at which false signals will occur in Phase II (with the distribution established before process parameters are estimated in Phase I)
RFS_p	The pth quantile of the RFS distribution
H	$\widehat{\text{UCL}} - \bar{\bar{X}}$

chart, the S-chart and other charts for monitoring dispersion could be handled in a similar fashion.) In Phase I, we sample m independent subgroups of n items each and create a trial control chart with limits $\hat{l}$ and $\hat{u}$ based on k standard deviations. (Without loss of generality, these values of m, n, and k remain fixed for the remainder of the paper.) Assuming that no assignable causes were found for any points that were outside the control limits, we declare the control chart to be ready for use in Phase II (otherwise, we take action to obtain a "good" chart). In Phase II, we start collecting data to be charted. Subgroups are sampled from the process and their means are plotted on the chart. When the process has not yet shifted from the in-control condition, occasionally one of these means falls outside the interval $[\hat{l}, \hat{u}]$; this constitutes a (false) "signal."

Our discussion centers on the run length for a Phase II process that has not yet gone out of control. We will define the random variable *run length*, L, as the number of subgroups after any given signal, up to and including the next signal that occurs. Of particular interest is the first run length seen in a charted process, denoted by the random variable L_1, which is the number of subgroups recorded from the beginning of chart use until a signal is first seen (including the signal itself) when the process is in control.

Control charts were originally developed under the assumption that the parameters of the underlying in-control process were known. There was no Phase I to estimate process parameters, as none was required; Phase II charting began immediately. In this case, there is no reason to discuss L in connection with a particular realization of a chart or to limit discussion to the first run on a chart except for expository purposes; in the known-parameters case, a run length is a run length is a run length. When the process parameters are known, L is a geometrically distributed random variable. The parameter of this geometric random variable is the probability of seeing a subgroup mean fall outside the control limits. With known limits and normally distributed subgroup means, this probability is simply the area of the appropriate normal distribution falling outside the control limits, e.g., if $k = 3$, the probability is 0.0027 and the resulting expected value of L is $1/0.0027 = 370.4$. This expected value was given the name average run length, or ARL. The term ARL is usually associated in particular with the expected value of the first run length, L_1, and we formally define the ARL as $E_{\mathrm{II}}[L_1]$, the subscript indicating that we are referring to the expectation over the distribution of L_1 in Phase II. Throughout the remainder, we will use the terms ARL and $E_{\mathrm{II}}[L_1]$ interchangeably.

Here we consider the case that the parameters of the underlying in-control process are not known, so that a Phase I is needed to estimate them. In this case, it is necessary to carefully distinguish among the following four expected values, which are all equal in the known-parameters case: $E_{\mathrm{II}}[L_1]$ (the ARL as defined above); $E_{\mathrm{II}|\mathrm{I}}[L_1] = E[L_1$ in Phase II $|$ the values of the process parameter estimates in Phase I], i.e., the expected value of the first run length L_1, given particular values of the Phase I data; $E_{\mathrm{II}|\mathrm{I}}[L]$, the expected value of any in-control run length L, given particular values of the Phase I data; and $E_{\mathrm{II}}[L]$, the expected value of L before any data are obtained in Phase I. The distinction indicated by the subscripts II and II $|$ I is that, in the latter, one of the sources contributing to the variation in the Phase II random variable has been controlled—namely, the Phase I estimates of the process parameters—and hence the conditional and unconditional expected values of the random variable may be different. A good deal of the confusion in the literature concerning the properties of estimated charts stems from the failure to distinguish among these quantities.

Although we will go into it in more detail below, we now specify the measure that we propose as an alternative to the ARL: the rate of false signals, RFS. The random variable RFS is defined as the rate at which false signals will occur in Phase II. The probability that H&Y studied was the prospective, pre-Phase I expected value of RFS, $E_{\mathrm{II}}[\mathrm{RFS}]$. After Phase I takes place and the chart is created, the random variable RFS takes on the realized value $r = P($a subgroup mean will be a false signal $|$ the process parameter estimates from Phase I). If we happen to have perfect estimates of the process parameters for an $\bar{X}$ chart and we have set $k = 3$, then $r = 0.0027$, the probability of a false signal when the process parameters are known.

We summarize three relationships among these expectations that will be discussed below:

$$E_{\mathrm{II}|\mathrm{I}}[L_1] = E_{\mathrm{II}|\mathrm{I}}[L]; \tag{1}$$

for finite m,

$$E_{\mathrm{II}}[L_1] \neq E_{\mathrm{II}}[L]\colon E_{\mathrm{II}}[L_1] > E_{\mathrm{II}}[L]; \tag{2}$$

and

$$E_{\mathrm{II}}[L] = 1/E_{\mathrm{II}}[\mathrm{RFS}]. \tag{3}$$

The first relationship states that, for any given chart based on particular estimates of the process parameters, the first run length and any other run length will have the same expectation. This is clearly true, as these run lengths have the same geometric distribution with parameter r. Given this relationship, one might think the same would be true for the unconditional expected run lengths as well, i.e., that $E_{\mathrm{II}}[L_1]$ (the ARL) is equal to $E_{\mathrm{II}}[L]$. But this is not the case: for any finite number of Phase I subgroups m, $E_{\mathrm{II}}[L_1] > E_{\mathrm{II}}[L]$. To understand this, consider a large number of Phase II charts all resulting from different Phase I estimates. Each of these charts can be thought of as contributing a single run length toward $E_{\mathrm{II}}[L_1]$. However, *all* the run lengths on these charts, not just one per chart, contribute to $E_{\mathrm{II}}[L]$, and a chart with a small $E_{\mathrm{II|I}}[L]$ will contribute more run lengths than a chart with a large $E_{\mathrm{II|I}}[L]$ will contribute. Hence, $E_{\mathrm{II}}[L]$ will be smaller than $E_{\mathrm{II}}[L_1]$, and the second relationship holds.

The third relationship can be explained as follows. Suppose we create C independent charts, each with different estimates of the process parameters, and run each of them for B subgroups in Phase II. Suppose further that chart c contributes J_c false signals (where J_c is a random variable). Then as B and C approach infinity, $\sum_c J_c/BC$ converges in probability to $E_{\mathrm{II}}[\mathrm{RFS}]$, and $BC/\sum_c J_c$ will be approximately equal to $E_{\mathrm{II}}[L]$. Hence, $E_{\mathrm{II}}[L]$ and $E_{\mathrm{II}}[\mathrm{RFS}]$ are reciprocals.

We can also use Jensen's inequality with the third relationship to demonstrate the second relationship. Observe that (with obvious notation) $E_{\mathrm{II|I}}[L_1 \mid \mathrm{RFS} = r] = 1/r$. Letting $f_{\mathrm{RFS}}(r)$ be the density function for the random variable RFS, we have

$$\begin{aligned} E_{\mathrm{II}}[1/\mathrm{RFS}] &= \int_0^1 \frac{1}{r} f_{\mathrm{RFS}}(r)dr \\ &= \int_0^1 E_{\mathrm{II|I}}[L_1 \mid \mathrm{RFS} = r] f_{\mathrm{RFS}}(r)dr \\ &= E_{\mathrm{II}}[L_1]. \end{aligned} \tag{4}$$

By Jensen's inequality, $E_{\mathrm{II}}[1/\mathrm{RFS}] > 1/E_{\mathrm{II}}[\mathrm{RFS}]$, which, with Equation (3), implies that $E_{\mathrm{II}}[L_1] > E_{\mathrm{II}}[L]$.

The relationship among $E_{\mathrm{II}}[L_1]$, $E_{\mathrm{II}}[L]$, and $E_{\mathrm{II}}[\mathrm{RFS}]$ can be seen in the simulation results of Table 2. For each of 2000 replications, a 3-standard-deviation $\bar{X}$ chart was created from process parameter estimates based on m Phase I subgroups of size $n = 5$, and the resulting chart was run for 25,000 Phase II subgroups. The experiment was repeated with $n = 10$. The average of the first run length on each chart was recorded as an estimate of $E_{\mathrm{II}}[L_1]$, the average of all the run lengths across the replications was recorded as an estimate of $E_{\mathrm{II}}[L]$, and the number of false signals divided by the total number of subgroups (50,000,000) was recorded as an estimate of $E_{\mathrm{II}}[\mathrm{RFS}]$. It can be seen that, for small m, the estimates of $E_{\mathrm{II}}[L_1]$ are much larger than the estimates of $E_{\mathrm{II}}[L]$, and as m increases, the two estimates converge from above and below, respectively, toward the known-parameters average run-length value of 370.4. The estimate of $E_{\mathrm{II}}[\mathrm{RFS}]$ can also be seen to be close to the reciprocal of the estimate of $E_{\mathrm{II}}[L]$.

TABLE 2. Estimates of $E_{\mathrm{II}}[L_1]$, $E_{\mathrm{II}}[L]$, and $E_{\mathrm{II}}[RFS]$ for Various m and n

m	$E_{\mathrm{II}}[L_1]$	$E_{\mathrm{II}}[L]$	$E_{\mathrm{II}}[\mathrm{RFS}]$
		$n = 5$	
5	690	77	0.0126
10	527	143	0.0068
25	411	245	0.0040
50	403	291	0.0034
100	374	324	0.0030
200	378	345	0.0029
400	365	354	0.0028
1600	369	364	0.0027
		$n = 10$	
5	401	109	0.0090
10	342	179	0.0055
25	347	271	0.0036
50	373	314	0.0031
100	367	340	0.0029
200	388	350	0.0028
400	368	359	0.0027
1600	376	363	0.0027

The preceding illustrates our assertion that run lengths are potentially quite ambiguous. To appreciate how damaging this can be, we now cite a case where researchers confused the two and thus reached erroneous conclusions, including the rejection of H&Y's conclusions concerning a correction factor for estimated limits. Ghosh et al. (1981) (henceforth G81) developed an expression for the expected long-run average economic cost per unit time

(p. 1802), and stated that it is a function of the ARL ($E_{\text{II}}[L_1]$) as well as of the expected time until a subgroup mean falls outside the control limits, given that the process is out of control (known as the out-of-control ARL). They thus conclude that the ARL is "the primary measure of the effectiveness of control procedures" (p. 1803). However, their derivation relies on the well-known result for renewal reward processes (see, for example, Ross (2003), p. 417) that the long-run average cost per unit time is equal to the ratio of the expected cost per cycle to the expected length of a cycle. They define a cycle to be the elapsed time from when the control procedure is first applied (assuming that the process starts in control) until the time that the process is returned to the in-control state after a shift to an out-of-control state has occurred and has been detected, and their concern is with the average cost per unit time incurred during "an infinite sequence of these in-control–out-of-control cycles" (p. 1801). That is, they are deriving the long-run average cost per unit time for the use of a particular control chart that is used long enough to generate many false signals. Hence, their long-run average cost should be a function of $E_{\text{II}}[L]$, not the ARL. The G81 authors proceed to study the properties of ARL instead of studying $E_{\text{II}}[L]$, as would be necessary to investigate the long-run average cost performance of a given control procedure.

A further error stemming from the confusion of the ARL with $E_{\text{II}}[L]$ is seen in one of G81's main conclusions, as follows:

> [T]he effect of using limits based on the t-distribution in place of the three sigma limits is to increase the ARL values even further. This suggests that *if the desired ARL values are those calculated for three-sigma limits when σ is known*, then the limits for the case where σ is estimated should be reduced rather than increased. (Note that this is contrary to the recommendation of Yang and Hillier (1970).)
>
> (G81, pp. 1811–1812. Emphasis added.)

Our results in Table 2, as well as the computational evidence of a number of other authors (Quesenberry (1993), Chen (1997), Albers and Kallenberg (2004)), suggest that, if μ and σ must be estimated, the resulting ARL is larger than the ARL in the known-parameters case. Nonetheless, the practical conclusion the G81 authors draw is flawed. What H&Y calculated in their papers was $E_{\text{II}}[\text{RFS}]$, the average risk of a false signal the user faces, and they showed that estimating μ and σ from a small number of samples results in a higher $E_{\text{II}}[\text{RFS}]$ than would be the case if these parameters were known—hence, their recommendation to reduce $E_{\text{II}}[\text{RFS}]$, the RFS for an "average" chart, by increasing the width of the control limits by a factor that is a function of m. Practitioners who would follow G81's advice and reduce this width instead of widening it would face an increased $E_{\text{II}}[\text{RFS}]$ over and above that resulting from estimation. We conclude that trying to match ARL to a desired benchmark is, in at least some applications, a dubious objective.

Albers and Kallenberg (2004, Remarks 2.1 and 2.4) discuss the seemingly contradictory result that, for small m, both the ARL and $E_{\text{II}}[\text{RFS}]$ are larger than in the known-parameters case. This paradox is resolved by the fact that the ARL and $E_{\text{II}}[\text{RFS}]$ are not reciprocals but rather ARL $> 1/E_{\text{II}}[\text{RFS}]$ (derived from the relationships given in Equations (2) and (3) above). For example, Quesenberry (1993) has, for $m = 10$, an estimated ARL of 611 (greater than the known-parameters ARL of 370.4), while Hillier's $E_{\text{II}}[\text{RFS}]$ is 0.0067 (greater than the known-parameters $E_{\text{II}}[\text{RFS}]$ of 0.0027). With the correct measure replacing $E_{\text{II}}[L_1]$ (the ARL), we get an expected run length of $E_{\text{II}}[L] = 1/0.0067 = 149 <$ 370.4.

With such confusion between $E_{\text{II}}[L]$ and $E_{\text{II}}[L_1]$, erroneous conclusions are likely, and indeed have occurred. To prevent such errors, we have a choice between two approaches. One is to use $E_{\text{II}}[L_1]$ and $E_{\text{II}}[L]$ together in a way that will require users to distinguish carefully between them. For example, $E_{\text{II}}[L_1]/E_{\text{II}}[L]$, which is the product of the ARL and $E_{\text{II}}[\text{RFS}]$, is such a measure, which decreases asymptotically to 1 as m approaches infinity. Another way is to revert back to consideration of the random variable RFS, thus avoiding discussion of expected or average run lengths entirely.

Consider charts such as CUSUM and EWMA charts. Here RFS is *not* the probability of a false signal in any given subgroup while the process is in control. This probability varies over the charting period because it depends on the accumulated results that precede it. However, when considered for long periods of chart usage, it still has an intuitive interpretation as the long-term rate of false signals (Hwang and Trietsch (1996)). Indeed, even for these charts, the use of ARL may lead to counterintuitive results, and RFS or a similar measure is likely to do better. For example, in a study of multivariate EWMA charts, Sullivan and Jones (2002) show that attempting to match a desired in-control ARL value will give spurious results. As an alternative, they propose matching the cumulative probability that a chart will signal

within, say, 100 subgroups to a desired benchmark such as $1 - (1 - 0.0027)^{100} = 0.2369$, where 0.0027 is again the rate of false signals for an $\bar{X}$ chart with known parameters and $k = 3$. This criterion, which requires wider control limits, is more robust and effective for setting limits and leads to better chart performance. So despite the very different research framework, here too the ARL must be allowed to be larger in order to achieve desired chart performance in terms of risk. In another paper supporting this approach, although assuming known parameters, Margavio et al. (1995) demonstrate that it is possible and useful to design EWMA and CUSUM charts in such a way that, at any given subgroup, the probability of a false signal at the next subgroup will match any desired profile; for example, this "alarm rate" could be set to a constant. More research is required to adapt their method for the estimated parameters case, but conceptually it should be possible either to match $E_{\text{II}}[\text{RFS}]$ to a desired profile or to match some quantile of RFS to a desired profile.

The excerpt from G81 above is, to our knowledge, the first criticism of H&Y in the literature, and this criticism was rooted in error. Some of G81's observations have also been made by Quesenberry (1993)—henceforth Q93—and although he did not explicitly confuse $E_{\text{II}}[L_1]$ with $E_{\text{II}}[L]$, he too found fault with H&Y's results, as follows. Let Y_i be the mean of subgroup i of a chart, and let $\widehat{\text{UCL}}$ denote a UCL that is a function of estimated parameters and is therefore a random variable (due to symmetry, we may ignore the LCL). Both G81 and Q93 state that the random events $\{\bar{X}_i - \widehat{\text{UCL}} > 0\}$ and $\{\bar{X}_j - \widehat{\text{UCL}} > 0\}$, i.e., false signals on subgroups i and j, are dependent—which is true because the random variable $\widehat{\text{UCL}}$ influences both. Q93 criticized H&Y because they did not take the dependence discussed above into account and then concluded that "to assess the performance of these charts the properties of the distribution of run lengths *must* be studied" (p. 239; emphasis added).

Our approach, however, looks at $\{\bar{X}_i - \hat{u} > 0\}$ and $\{\bar{X}_j - \hat{u} > 0\}$, where $\hat{u}$ is a realization from the distribution of $\widehat{\text{UCL}}$, and concludes, equally correctly, and perhaps more intuitively, that the signals on any *particular* chart are independent of each other. From false signals on particular charts, we can move to studying the distribution of RFS or its expectation, $E_{\text{II}}[\text{RFS}]$, across charts. Depending on whether by "run lengths" Q93 is referring to L or to L_1, the results of our study and his may or may not lead to the same conclusions. Nonetheless, as did G81, in the remainder of his paper, Q93 did not consider the distribution of L but of L_1. Later, because T99 (Ch. 8) also took the "independent" position (and developed it toward a study of the distribution of RFS), he drew criticism from Quesenberry (2001), whose implicit claim is that the *only* way to study the performance of sample-based charts is through this dependence, and furthermore, that this implies that only a focus on the ARL is legitimate. (Trietsch thanks Professor Quesenberry for a useful correspondence about this issue, prior to the publication of T99. Trietsch also regrets a valid but unnecessary remark in T99 about a potential misuse of Q-charts.) We believe, however, that this dependence is merely a symptom, not a cause. Below, we demonstrate that it is not necessary to consider this dependence in order to study the distributions of RFS and the ARL. Furthermore, G81, Chen (1997), and Chakraborti (2000) all develop expressions for the ARL without actually using this dependence, and even Q93's own simulation study of the ARL does not make explicit use of this dependence.

The Distribution of the Rate of False Signals

H&Y examined $E_{\text{II}}[\text{RFS}]$, the pre-Phase I expected value of the probability, or rate, of false signals. As well as studying its expectation, it would be natural to study the variation of this rate, i.e., if we generate many charts using the same Phase I choices of m, n, and k, we may ask how the random variable RFS will be distributed across charts. For example, we would be able to say: "If you use 5 subgroups of size 5 in Phase I, estimate σ by $\bar{S}$, and set k to 3, then the expected rate of false signals is about 1.2%. With probability 0.95 the rate of false signals will be below about 4.75% and with probability 0.05 it will be above about 0.026%." If the mean looks too high to a practitioner, increasing k can reduce it to any value desired, but if the interval between the two quantiles looks too wide, the only guaranteed remedy is to increase m.

Our proposed use of $E_{\text{II}}[L_1]/E_{\text{II}}[L]$ (i.e., ARL multiplied by $E_{\text{II}}[\text{RFS}]$) as a measure notwithstanding, we believe that it is best to have knowledge of the complete distribution of the RFS, L_1, or L. On this point, we are in agreement with Quesenberry (1993), who reported values of the L_1 distribution as a function of m (given n and k) and, through this, avoided drawing the incorrect practical conclusions at which

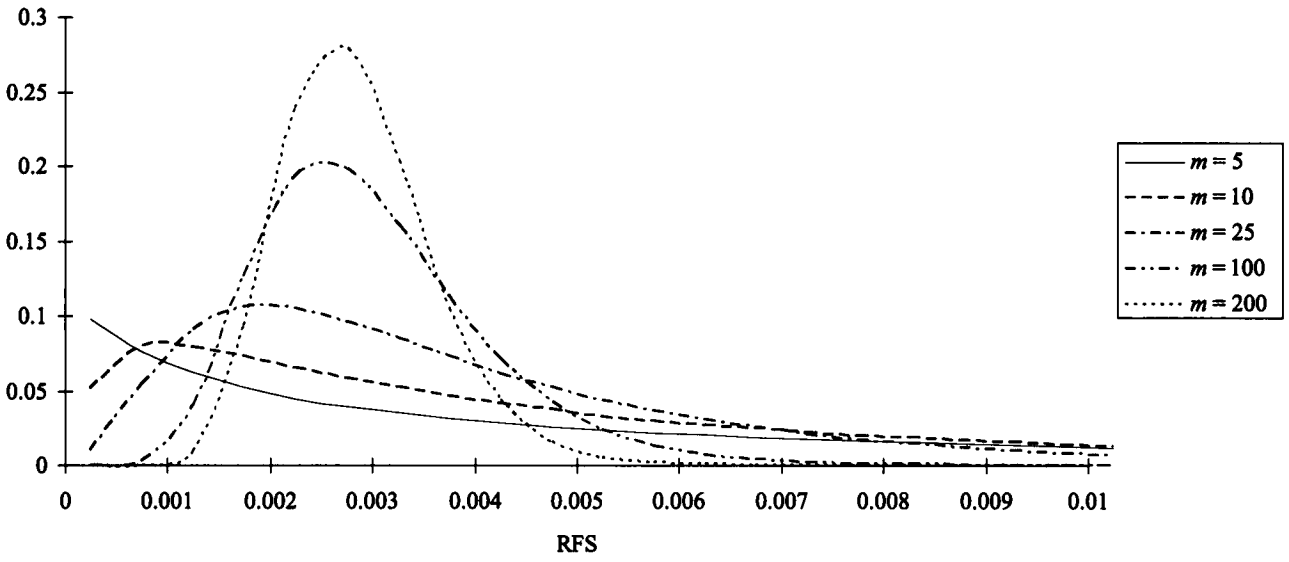

FIGURE 1. Probability Density Function of RFS for Various Values of m, for $n = 5$ and $k = 3$.

G81 arrived. (In a similar vein, Jones et al. (2004) provide an enlightening discussion of the in-control distribution of L_1 and its impact on the ARL for CUSUM charts with estimated parameters.) Recall from the integral in Equation (4) that, if we have the distribution of RFS, we can obtain the ARL by computing $E_{\text{II}}[1/\text{RFS}]$. Thus, even if our main concern is the ARL, our study of the RFS would provide all the information we need. Our preference for the RFS stems from its clearer definition and from the fact that it is less likely to be used incorrectly because it cannot easily be confused with other measures.

Figure 1 shows the probability density function of RFS for a chart with $n = 5$ and $k = 3$, using various values of m. These graphs were created from smoothed histograms of simulated data; hence, the graphs do not begin at the origin. The number of replications was 100,000 for each value of m except $m = 5$, where 1,000,000 replications were run. At small values of m, the density functions display an extremely large spread, indicating the wide range of possible false-signal rates achievable across charts. As m increases, the density gradually tightens up; if m were to go to infinity, the density function would become a single spike at 0.0027.

Table 3 provides simulated numerical values of quantiles of the RFS for $m = 5$ and 10; for $m \geq 25$, the values given are derived from Equation 7 given subsequently in this paper. By the table, 1% of charts created from 25 subgroups, a number recommended by many authors, have an RFS that is less than 0.0006 and 10% have an RFS less than 0.0013. Even at $m = 100$, 10% of charts will have an RFS of 0.0018 or less, that is, at most two-thirds of the "planned" 0.0027 value. These results have implications for the power of the chart to detect shifts in the process. At the other end of the table, the tendency of a chart based on small m to produce false alarms can be seen: a chart based on 25 subgroups has approximately a 20% chance of having an RFS that is more than twice as high as 0.0027. Similarly, there is a 5% chance that the RFS will be 0.96% or more.

We now provide a new approximate expression for the distribution of RFS (which is exact in the special case where the centerline is not estimated). Quesenberry (1993) presented a clever approximation for $E_{\text{II}}[\text{RFS}]$, based on approximating the width of the control limits by a normal random variable. For this reason, the approximation is valid for large m. Similar results, without the limitation for large m but based on a chi-square approximation instead, which is only exact for the case in which the process variance is estimated by $\overline{S^2}$, were presented by H&Y. Although Quesenberry dismissed these results and his own derivation, stating that "these probabilities have only limited usefulness when assessing the properties of charts using $\widehat{LCL}$ and $\widehat{UCL}$ as control limits," Quesenberry's equation is actually the key to the approximate computation of the RFS distribution. We present it as an explicit function of m, n,

TABLE 3. Quantiles of RFS for Various m and p, for $n = 5$ and $k = 3$

	p										
m	0.01	0.05	0.1	0.2	0.25	0.5	0.75	0.8	0.9	0.95	0.99
5	0.00005	0.00026	0.00053	0.00127	0.00167	0.00494	0.01391	0.01786	0.03011	0.04741	0.12448
10	0.00012	0.00042	0.00073	0.00134	0.00174	0.00399	0.00845	0.01005	0.01570	0.02210	0.03712
25	0.00058	0.00099	0.00130	0.00180	0.00203	0.00326	0.00515	0.00574	0.00763	0.00959	0.01449
50	0.00088	0.00128	0.00155	0.00194	0.00212	0.00297	0.00414	0.00448	0.00552	0.00654	0.00890
100	0.00121	0.00156	0.00178	0.00210	0.00223	0.00283	0.00359	0.00380	0.00442	0.00500	0.00628
200	0.00152	0.00182	0.00200	0.00223	0.00233	0.00277	0.00327	0.00341	0.00380	0.00415	0.00490
400	0.00179	0.00203	0.00217	0.00235	0.00242	0.00273	0.00308	0.00317	0.00343	0.00365	0.00411
800	0.00202	0.00220	0.00231	0.00244	0.00249	0.00272	0.00296	0.00302	0.00319	0.00334	0.00363
1600	0.00220	0.00234	0.00241	0.00251	0.00255	0.00271	0.00288	0.00292	0.00303	0.00313	0.00333
(∞)	0.00270	0.00270	0.00270	0.00270	0.00270	0.00270	0.00270	0.00270	0.00270	0.00270	0.00270

and k and with a cosmetic change:

$$E_{\mathrm{II}}[\mathrm{RFS}] \approx 2\Phi\left[\frac{-k}{\sqrt{1+\frac{1}{m}\left(1+\frac{k^2(1-c_4(n)^2)}{c_4(n)^2}\right)}}\right], \tag{5}$$

where $\Phi(x)$ is the area under the standard normal curve from $-\infty$ to x. The approximation is based on the observation that a subgroup mean falls above $\widehat{UCL}$ if one normal random variable, $Y = \bar{X}_i - \mu$, exceeds the sum of two independent random variables, $D = \bar{\bar{X}} - \mu$ and $H = \widehat{UCL} - \bar{\bar{X}}$, i.e., D is the deviation of the centerline from the true mean and H is the half-width of the control limits. D is normal by assumption, and for large m, say $m \geq 25$, H is approximately normal. Whether a signal occurs depends on $W = Y - D - H$, which is approximately normal. Then due to symmetry with $\widehat{LCL}$, $E_{\mathrm{II}}[\mathrm{RFS}] = 2 \cdot P(\text{signal above } \widehat{UCL}) = 2 \cdot P(W > 0) = 2(1 - \Phi(-\mu_W/\sigma_W))$, where μ_W and σ_W are the expected value and standard deviation of W, respectively, and the result follows (see Q93 for details). Note that this is not an exact result *only* because H is not a normal random variable.

To obtain $F_{\mathrm{RFS}}(p)$, the distribution function of the RFS, approximately for $m \geq 25$, let h_p denote the value of H such that $F_H(h_p) = 1 - p$, i.e., h_p is the $(1-p)$th quantile of the distribution of chart half-widths. Note that when the $\overline{S^2}$ statistic is used to estimate the process variance, H is distributed proportionally to the square root of a chi-square random variable with $m(n-1)$ degrees of freedom. H&Y and Chen (1997) provide approximations of H for small m where $\bar{S}$ or R (the range) is used, also by a square root of a chi-square random variable; for large m, H is distributed approximately normally. With any one of these, it is straightforward to identify h_p for any $p \in (0,1)$. Because a proportion $1-p$ of chart half-widths will be smaller than h_p, it follows that a proportion $1-p$ of charts will have expected rates of false signals *greater* than the expected rate obtained with H set to h_p, i.e., a proportion p will have smaller rates. (We use "expected" rate here because of the unknown centerline and the process variance.) Given h_p, we can calculate

$$E[\mathrm{RFS} \mid H = h_p] = 2 \cdot P(W > 0 \mid H = h_p) = 2\Phi\left(\frac{-h_p}{\frac{\sigma}{\sqrt{n}}\sqrt{1+\frac{1}{m}}}\right),$$

a calculation that resembles the one performed for Equation (5), and setting $\sigma/\sqrt{n} = E[H]/k$ (since $E[H] = k\sigma/\sqrt{n}$), we have, for $m \geq 25$, the approximate pth quantile of the RFS distribution,

$$\mathrm{RFS}_p \approx 2\Phi\left[\frac{-k\frac{h_p}{E[H]}}{\sqrt{1+\frac{1}{m}}}\right]. \tag{6}$$

Comparing this expression with that given in Equation (5), the numerator is multiplied by $h_p/E[H]$; the denominator drops the contribution of

$\mathrm{Var}(H)$ because we assume that $H = h_p$. The result is only an approximation because of the effects of estimating the centerline. For the case where the centerline is given (i.e., when the process is adjustable and we want the chart to signal if the adjustment is off-center, as discussed in T99), we drop the square root of $1 + 1/m$ and Equation (6) becomes an exact expression: $\mathrm{RFS}_p = 2\Phi(-kh_p/E[H])$. T99 suggests using the result in this special case as a bound on the true RFS variation. For the case where the centerline is estimated, however, and $m \geq 25$, we can use Equation (6) with the normal approximation for H. To this end, define z_p to be the pth quantile of the standard normal distribution. Then $h_p/E[H] \approx 1 - z_p\sqrt{1-c_4^2}/(c_4\sqrt{m})$, so that

$$\mathrm{RFS}_p \approx 2\Phi\left[\frac{-k\left(1 - \frac{z_p}{\sqrt{m}}\frac{\sqrt{1-c_4^2}}{c_4}\right)}{\sqrt{1+\frac{1}{m}}}\right]. \qquad (7)$$

As $m \to \infty$, this converges to the theoretical RFS (and the density function of RFS becomes a spike). Because $F_{\mathrm{RFS}}(p)$ is a monotonically increasing function, its inverse—characterized in Equation (6) and approximated in Equation (7)—is sufficient for obtaining $F_{\mathrm{RFS}}(p)$ for any p.

Because, as was shown in Equation (4), ARL $= E_{\mathrm{II}}[1/\mathrm{RFS}]$, this result should be of equal interest to those who prefer the ARL as a measure and to those who, like us, prefer the RFS. Readers may find a spreadsheet application using this relationship on the authors' web site (see the URL given above). The spreadsheet can be used to provide approximate confidence intervals for either $E_{\mathrm{II}}[\mathrm{RFS}]$ or $E_{\mathrm{II}}[L_1]$ as a function of m, n, and k. As mentioned above, Table 3 summarizes some values based on Equation (7) for $m \geq 25$ and also simulation results for $m < 25$. Although we do not report the details, we tested the validity of Equation (7) by simulation as well, and the results were in good agreement.

Probability-Based Diffidence Charts

Diffidence charts (T99) are a special type of self-starting chart, the purpose of which is to avoid the commingling of estimation errors with the process variation, so that the considerable variation in the performance of estimated charts constructed with a given Phase I choice of m, n, and k can be revealed. Self-starting charts are an intuitive way to utilize small amounts of data even while collecting more data. Typically, instead of demarcating a clear borderline between Phase I and Phase II, self-starting charts dynamically consider the previous subgroups as Phase I for the purpose of adjudicating the next subgroup. This allows for charting a process even while we are in fact still collecting baseline data. However, self-starting charts may be misinterpreted as ostensibly reliable early on, while this is not formally the case at all. For example, take Q-charts for variables (Quesenberry (1991)), which look at the distribution of a special transformation of the former baseline data as combined with the current subgroup data such that the in-control result is standard normal. The implicit claim is that, if a pattern of such transformed points behaves like a pattern of in-control standard deviates, then the data that we transformed are also in control. In reality, Q-charts no longer depict the original process at all, and instead they depict a related random variable that combines the effects of process variation with the effects of estimation. A point may show out of control either because it is indeed atypical for the process or because it is judged by parameter estimates that are far from their correct values due to sampling error (e.g., when m is too small) or both; the Q-chart provides no clue which. Furthermore, it is more likely that a point may be falsely adjudicated in or out of control due to sampling error with small samples, but self-starting charts such as Q-charts mask this, and some users may think they are equally reliable regardless of the current sample size. In principle, similar observations apply for most self-starting charts.

The purpose of diffidence charts is to highlight this unavoidable fact and show explicitly how much of the observed variation may plausibly be due to the estimation process. This is done by creating confidence intervals for the control limits that the chart would have if the true parameter values were known. We thus obtain three regions into which each subgroup mean may fall: if a point is closer to the center than to the inner control limit (or the inner boundary of the control limit confidence interval), we say that it is "strongly in control"; if it falls outside the outer limit, it is "strongly out of control"; and if it falls in between, it is "diffident." Diffident points are exactly the ones for which the sample is too small to tell clearly whether or not they are in control. If most points are not diffident, then increasing the sample size is not important because all the necessary decisions can be made without doing so.

Figure 2 depicts a diffidence chart of a process

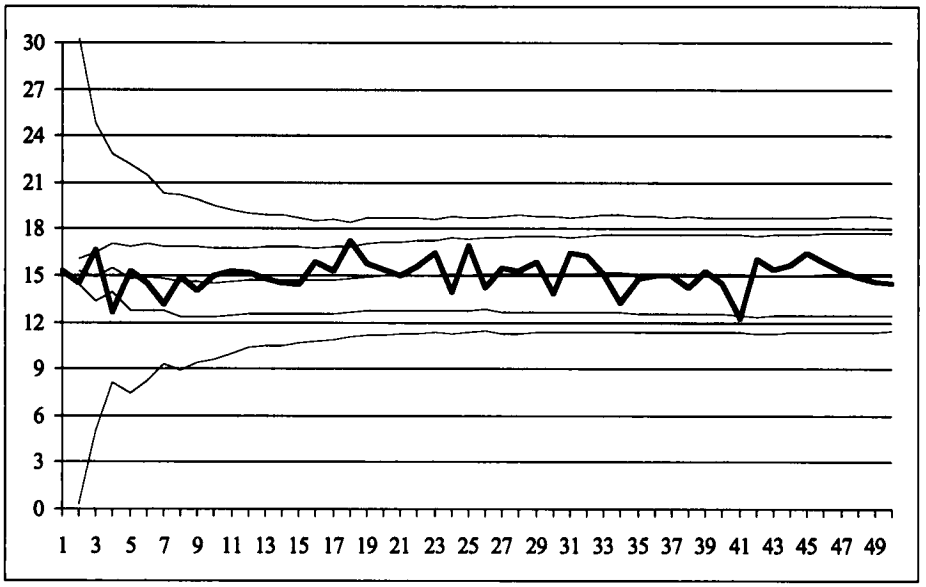

FIGURE 2. Diffidence Chart of a Process.

with $p = 0.0027$, i.e., using a $100(1-p)\% = 99.73\%$ confidence interval on the control limits. The thick line plots the subgroup means. From top to bottom, the other five lines shown are the upper and lower diffidence limits for UCL, the estimated centerline, and the upper and lower diffidence limits for LCL. Points 3, 4, 18, and 41 are diffident. Of course, one should not be too concerned with points 3 and 4 because the sample size is too small to conclude anything without further data being obtained. Indeed, with the exception of point 41, all the diffident points become "strongly in control" as the limits are dynamically updated, but there is simply not enough data yet to say the same about point 41—although its true deviation is only 2.72 standard deviations, so it would be deemed in control with known-parameters limits. (The spreadsheet used to create this figure is available on the authors' web site. Readers can experiment with other p values for the same Phase I sample, which was generated randomly without any attempt to select "good" results.)

When a sample is strongly in control, it is very likely that both the process and the measurement are in control. But a point is diffident exactly when it is impossible to determine confidently whether the process is in control or out of control, and likewise it is impossible to tell whether the data are representative, as the sample is too small. As the number of samples increases, the diffident zones narrow. When judged against the updated diffident zones, points that have been diffident may be resolved retroactively as being in control or out of control. Until a diffident point is resolved as out of control, however, it is probably safer to include it in the dynamic Phase I sample. When points rarely fall in the diffident zone, the chart is satisfactory. This happens when the diffidence band is narrow enough. Until this is achieved, we may treat diffident points as warranting caution. Also, if we choose to check for an assignable cause of a diffident point, it is useful to know how likely it is to be indeed out of control, so we can conduct the check more rationally—a point near the lower diffidence limit will justify a briefer check than one farther out, near the outer limit. Another idea is to treat all diffident points as warranting a careful check but not an interruption of the process. In this way, they would be similar to the 2σ warning limits often used in practical applications (Montgomery (2001), pp. 165–166). Nonetheless, the best resolution of diffident points is an open research question at the time of this writing. Finally, although we limit our discussion here to diffidence charts for the sample mean, diffidence charts for s, R, p, or u can be set up in a similar manner (T99).

Determining Probability-Based Diffidence Limits

As we discussed, diffidence charts include bounds for the control chart limits UCL and LCL. These bounds may be based on the $\pm k\sigma$ method (see T99

for examples based on this method), or on probabilistic analysis. Trietsch and Hwang (1997) use the latter approach for RFS limits of pattern tests, and we will discuss this approach for $\bar{X}$ charts here. Create a chart based on given values m, n, and k, and let h denote the realization of the random variable half-width H in this chart. Because we are using unbiased estimators, the correct value of h to use would be $E[H]$, with the half-width anchored at the correct location of the centerline. However, we know neither H nor the centerline with certainty (cases where the centerline is given are easier—see T99). Define $w_p = h_p/E[H]$. Then a $100(1-p)\%$ confidence interval for $E[H]$ is given by

$$h/w_{p/2} \leq E[H] \leq h/w_{1-p/2}.$$

In order to be pronounced "strongly in control," a point must be between the centerline and the UCL's lower diffidence limit $ucl_{p/2}$ drawn at a distance of $h/w_{p/2}$ from the centerline (and similarly, between the centerline and the LCL's upper limit $lcl_{p/2}$). In contrast, if we want to characterize points that are "strongly out of control" correctly, placing $ucl_{1-p/2}$ equal to $h/w_{1-p/2}$ is *not* sufficient because we want to be sure that the effect of error in the centerline placement is not the true culprit. So we must use $ucl_{1-p/2} = h\sqrt{1+1/m}/w_{1-p/2}$ (and similarly, the LCL's lower limit $lcl_{1-p/2}$). If we draw control limits at the two extremes implied by these limits, we effectively obtain an approximate confidence interval for the correct location of the control limits. Points between the two extremes are "diffident with the prescribed probability." A useful way to think about this is that the diffidence probability is the confidence level associated with the confidence interval for the control limit.

What characterizes diffident points is that we need a larger Phase I sample to resolve them confidently. The diffidence bands highlight the range where the fact that process variation and measurement variation are confounded has operational consequences, while points outside the diffidence bands are those where the signal is clear in spite of the confounding. Because the width of the confidence interval decreases with m (approximately proportionally to $\sqrt{m}$), a chart with dynamic control limits including such confidence intervals can serve to show graphically how reliable the location of the limits is.

Conclusion

In this paper, we have suggested that the in-control ARL can be a confusing measure of doubtful value in some applications. We introduced the RFS (rate of false signals) as a measure of the performance of an $\bar{X}$ chart and argued that it is a useful measure because of its clarity to practitioners. We also discussed diffidence charts, which graphically present the inherent uncertainty that is concealed when control limits are computed from subgroup means.

We believe that what is of interest to the practitioner is the probability that she has obtained a "bad" chart, that is, one that, due to the luck of the draw, has estimated control limits that either result in an excessive rate of false signals or lack power. Knowledge of the distribution of the RFS and the use of diffidence charts can help determine this probability for given data and they can also help predict this probability for a chart in advance of sampling.

Acknowledgments

We gratefully acknowledge the time and effort of the former editor, Joe Sullivan, whose comments greatly improved this paper. We also thank the anonymous referees for their comments and another former editor, William Woodall, for his receptiveness to our paper. The first author's research was supported by the Natural Sciences and Engineering Research Council of Canada under Grant No. 239153.

References

ALBERS, W. and KALLENBERG, W. C. M. (2004). "Estimation in Shewhart Control Charts: Effects and Corrections." *Metrika* 59, pp. 207–234.

CHAKRABORTI, S. (2000). "Run Length, Average Run Length, and False Alarm Rate of Shewhart $\bar{X}$ Chart: Exact Derivations by Conditioning." *Communications in Statistics: Simulation and Computation* 29, pp. 61–81.

CHEN, G. (1997). "The Mean and Standard Deviation of the Run Length Distribution of $\bar{X}$ Charts when Control Limits Are Estimated." *Statistica Sinica* 7, pp. 789–798.

DEL CASTILLO, E.; GRAYSON, J. M.; MONTGOMERY, D. C.; and RUNGER, G. C. (1996). "A Review of Statistical Process Control Techniques for Short Run Manufacturing Systems." *Communications in Statistics: Theory and Methods* 25, pp. 2723–2737.

GHOSH, B. K.; REYNOLDS, M. R.; and VAN HUI, Y. (1981). "Shewhart $\bar{X}$-Charts with Estimated Process Variance." *Communications in Statistics: Theory and Methods* A10, pp. 1797–1822.

HAWKINS, D. M. (1987). "Self-Starting CUSUM Charts for Location and Scale." *The Statistician* 36, pp. 299–315.

HAWKINS, D. M.; QIU, P.; and KANG, C. W. (2003). "The Changepoint Model for Statistical Process Control." *Journal of Quality Technology* 35, pp. 355–366.

HILLIER, F. S. (1964). "$\bar{X}$ Chart Control Limits Based on a Small Number of Subgroups." *Industrial Quality Control* 20, pp. 24–29.

HILLIER, F. S. (1969). "$\bar{X}$ and R-Chart Control Limits Based on a Small Number of Subgroups." *Journal of Quality Technology* 1, pp. 17–26.

HWANG, F. K. and TRIETSCH, D. (1996). "A Simple Relation Between the Pattern Probability and the Rate of False Signals in Control Charts." *Probability in the Engineering and Informational Sciences* 10, pp. 315–323.

JONES, L. A. (2002). "The Statistical Design of EWMA Control Charts with Estimated Parameters." *Journal of Quality Technology* 34, pp. 277–288.

JONES, L. A.; CHAMP, C. W.; and RIGDON, S. E. (2004). "The Run Length Distribution of the CUSUM with Estimated Parameters." *Journal of Quality Technology* 36, pp. 95–108.

MARGAVIO, T. M.; CONERLY, M. D.; WOODALL, W. H.; and DRAKE, L. G. (1995), "Alarm Rates for Quality Control Charts." *Statistics and Probability Letters* 24, pp. 219–224.

MONTGOMERY, D. C. (2001). *Introduction to Statistical Quality Control*, 4th edition. Wiley, New York, NY.

PROSCHAN, F. and SAVAGE, I. R. (1960). "Starting a Control Chart." *Industrial Quality Control* 17, pp. 12–13.

QUESENBERRY, C. P. (1991). "SPC Q Charts for Start-Up Processes and Short or Long Runs." *Journal of Quality Technology* 23, pp. 231–224.

QUESENBERRY, C. P. (1993). "The Effect of Sample Size on Estimated Limits for $\bar{X}$ and X Control Charts." *Journal of Quality Technology* 25, pp. 237–247.

QUESENBERRY, C. P. (2001). "Review of *Statistical Quality Control: A Loss Minimization Approach*, by D. Trietsch." *Journal of Quality Technology* 33, pp. 119–122.

ROSS, S. M. (2003). *Introduction to Probability Models*, 8th edition. Academic Press, San Diego, CA.

SULLIVAN, J. H. and JONES, L. A. (2002). "A Self-Starting Control Chart for Multivariate Individual Observations." *Technometrics* 44, pp. 24–33.

TRIETSCH, D. (1999). *Statistical Quality Control: A Loss-Minimization Approach*. World Scientific Publishing Company, Singapore.

TRIETSCH, D. and HWANG, F. K. (1997). "Notes on Pattern Tests for Special Causes." *Quality Engineering* 9(3), pp. 467–477.

YANG, C.-H. and HILLIER, F. S. (1970). "Mean and Variance Control Chart Limits Based on a Small Number of Subgroups." *Journal of Quality Technology* 2(1), pp. 9–16.