CHAPTER 8

Quality Improvement

Statistical methods for quality improvement have been used extensively in the United States, for example, especially during the past 25 years. Unlike a t-test to compare two means or ANOVA used to analyze data from an experimental design, some of these statistical methods are used repeatedly. When this is the case, the concept of power, which applies to a single analysis, is not appropriate.

For example, control chart usage may involve a small sample of items being taken every 30 minutes, with the objective of maintaining tight control over a process. The properties of a single sample are almost irrelevant since many samples will be obtained in a given day. Hence, sample size determination is viewed in a different way and different statistics are used. This is discussed in Section 8.1.3, in particular.

8.1 CONTROL CHARTS

There are two phases of control chart usage: analysis of past data and real-time process monitoring, and charts are constructed using either individual observations or subgroups, which might be groups of 4 or 5 observations obtained every 30 minutes, say, from some process. (The analysis of past data and real-time process monitoring are herein termed Stage 1 and Stage 2, respectively; they have also been designated as Phase 1 and Phase 2 by some writers.) For Stage 1, there is a question of how much past data to use in computing the necessary parameter estimates. If an insufficient amount of data is used, control chart performance could have both unacceptable variability and unacceptable properties, on average.

Sample Size Determination and Power, First Edition. Thomas P. Ryan.

Note that 3-sigma control chart limits are typically used. For a normal distribution, the tail areas outside $\mu \pm 3\sigma$ are each .000135. Since this is a very small value, we don't have a hypothesis testing situation with $\alpha = .05$ or .01. Rather, we can, loosely speaking, think of $\alpha = .000135$ for a one-sided test and $\alpha = 2(.00135) = .0027$ for a two-sided test. [There are dissimilarities between control chart usage and hypothesis testing, however; see Woodall (2000).]

8.1.1 Shewhart Measurement Control Charts

A measurement control chart is a chart for which a measured process characteristic is being monitored, such as a diameter. The standard 3-sigma control charts are often referred to as "Shewhart charts" since Dr. Walter A. Shewhart first sketched out the basic idea of a control chart in 1924, with the first chart proposed being a *p*-chart.

An attribute control chart, such as a *p*-chart, is one for which a unit is declared to be either conforming or nonconforming, or else the number of nonconformities of a certain type or all types are counted. (A *p*-chart is the former.) Sample size determination for attribute control charts is discussed in Section 8.1.2.

To illustrate the use of measurement charts, let's assume that the objective is to control the mean of some process characteristic, which we will assume has a normal distribution, and we will also assume that individual observations will be plotted on a (control) chart. The 3-sigma control limits will be at $\hat{\mu} \pm 3\hat{\sigma}$ and the chart is called an *X*-chart. The mean, μ, would logically be estimated by the average of the past data values, $\bar{X}$. The estimation of $\hat{\sigma}$ is not as simple, however. The usual approach in Stage 1 is to use $\hat{\sigma} = \overline{\text{MR}}\ /1.128$, with $\overline{\text{MR}}$ denoting the average of moving ranges of size two of the time-ordered past data and 1.128 being the constant that makes the estimator unbiased. For Stage 2, Cryer and Ryan (1990) showed that $\hat{\sigma} = S/c_4$ is the preferred estimator, as that estimator has a smaller variance than the estimator based on moving ranges, with s denoting the standard deviation of the Stage 1 (historical) data, and c_4 is the constant, which depends on the number of observations, that makes the estimator unbiased. Ideally, the moving range estimator should be used to help identify the data that have not come from the in-control distribution. These data would then be removed and the remaining data used to compute both $\bar{X}$ and S/c_4 for Stage 2. (The reason that the moving range estimator should be used in Stage 1 is that it is less sensitive to aberrant data than is the standard deviation, with the resultant control limits thus being better at helping the control chart user to identify data that is not from the in-control distribution.)

The control chart parameters should be estimated with sufficient data so that the point at which signals are received after a change in the process mean does not have high variability. That is, for example, we would not want to receive a signal on the 5th plotted point one time and on the 85th plotted point the next time. Control chart properties have high variability because of the small probability associated with 3σ limits, even when the parameters are assumed known; the problem should not be exacerbated by using a small amount of data to estimate the parameters.

The results given by Quesenberry (1993) are useful in determining the total number of observations to be used in Stage 1, which for subgroup data is the number of subgroups times the subgroup size and is the number of individual observations for a control chart of individual observations. As a rule-of-thumb, at least 100 observations should be used to compute trial control limits, with approximately 300 observations used for "permanent" limits. (Processes change over time, so parameters must eventually be reestimated.)

When data are in subgroups rather than as individual observations, subgroup averages can be plotted on a chart. When this is done, the chart is called an $\bar{X}$-chart. Then a decision must be made regarding the subgroup size as well as the number of subgroups of past data that will be used for parameter estimation. Historically, subgroups of 4 or 5 have been used, without any attention given to the power associated with each subgroup size for a given mean shift. Dockendorf (1992) made this point and provided two graphs that could be used as an aid in determining subgroup size. Unfortunately, however, an "allowable shift in standard deviation units" is on the horizontal scale of each of the graphs, with 1.5 in the middle of the scale. This greatly limits the value of the graphs as a shift of less than one standard deviation unit would typically be an acceptable shift, with CUSUM schemes, for example, typically designed with the focus of detecting (at least) a one standard deviation shift.

It should be noted that the sample size need not be constant, as it is reasonable to decrease the subgroup size or increase the sampling interval, or both, when a process seems to be running smoothly with a process disturbance creating an out-of-control situation being very unlikely. When that is the case, a variable sample size might be used. Such charts are discussed in Section 8.1.5.

Consider Table 8.1, which indicates how power changes with subgroup size changes for a given shift in the mean from μ to $\mu + \sigma$. Control chart users think in

Table 8.1 Power of Detecting a 1σ Upward Mean Shift with an $\bar{X}$-Chart with Subgroup Size n (μ and σ Assumed Known)

n	Power (ARL)
1	.02275 (43.96)
2	.05639 (17.73)
3	.10241 (9.76)
4	.15866 (6.30)
5	.22245 (4.50)
6	.29098 (3.44)
7	.36158 (2.77)

terms of average run lengths (ARLs), however, which is the expected number of points to be plotted after a change in the parameter(s) being monitored for a signal to be received from a point plotting outside the control limits. Although power is not typically mentioned when control charts are designed, the probability of a point plotting outside the control limits after the parameter that is being monitored (such as a mean) has changed can be termed the "power" of the control chart, loosely speaking. ARL is then equal to 1/power. Parameter changes for measurement charts are generally expressed in terms of the number of standard deviation units of what is being plotted, and ARLs are given for parameter changes expressed in this manner.

Thus, if subgroup averages are being plotted, which would be the case when an $\bar{X}$-chart is used, the change would be expressed in terms of a multiple of $\sigma_{\bar{X}}$. That won't work in trying to determine subgroup size, however, because the subgroup size is then "lost." Specifically, $Z = [\mu + 3\sigma_{\bar{X}} - (\mu + a\sigma)]/\sigma_{\bar{X}}$ is a function of n, whereas $Z = [\mu + 3\sigma_{\bar{X}} - (\mu + a\sigma_{\bar{X}})]/\sigma_{\bar{X}}$ is not a function of n, with $\mu + 3\sigma_{\bar{X}}$ denoting the upper control limit (UCL) of an $\bar{X}$-chart, and $(\mu + a\sigma)$ representing the new mean as a function of σ, and similarly for $\mu + a\sigma_{\bar{X}}$. Therefore, it is the first of these two forms for Z that was used in the power and ARL values that are shown in Table 8.1, which is based on the assumption of known μ and σ. (This assumption will later be relaxed.)

Note that none of the power values are even remotely close to .8 or .9, but that is of no consequence since, unlike a typical hypothesis testing problem, sampling is often done very frequently (such as every 15 minutes), and the ARL values show that the shift should be detected after just a few plotted points when the subgroup size is around five. Thus, power really isn't an issue when control charts are used, nor are very small differences in ARL values of any concern. Thus, the choice between, say, a subgroup size of 5 and a subgroup size of 6 should be based on factors other than those shown in Table 8.1.

Regarding the number of subgroups of past data to use, which is also a sample size determination problem of sorts, applying the Quesenberry (1993) results would lead to at least 20 subgroups of size 5 being used for Stage 1, or at least 25 subgroups of size 4 (i.e., at least 100 observations). This would be applied to measurement control charts such as $\bar{X}$-, R-, and X-charts, which are virtually the most frequently used control charts. Since individual observations are plotted on an X-chart, this means that at least 100 individual observations would be needed for Stage 1, and about 2 or 3 times that many when (semi-) permanent control limits are used in Stage 2 (the process monitoring stage). Sample size determination for other types of charts will be discussed later.

Regarding the choice of subgroup size, since we may view control chart usage as being similar to hypothesis testing since a decision of "in control" or "out of control" is made each time a point is plotted, a very large subgroup would be impractical and would likely result in changes being detected that are not of practical significance, for changes that are expressed as $\mu + a\sigma$ or $\mu - a\sigma$.

(Note that this is the same type of consideration that must be made when hypothesis testing is used, in general.) On the other hand, even small changes may be of significance in medical applications and the failure to detect such changes could have serious consequences.

We want the ARL to be large when the process is in control and to be small for changes that are deemed to be consequential, analogous to having both a small Type I error probability and a small Type II error probability for general hypothesis testing. The objective is to detect changes before they have serious consequences. There are many types of statistical process control charts and control procedures but the general objective should be the same with each one. That is, enough data should be used to estimate control chart parameters so that control chart performance does not have large variability, and the number of observations used at each time period when a point is plotted on the chart should be such that the control chart scheme has desirable ARL properties. That is, a control chart signal should be received very infrequently when the process is in statistical control (i.e., false signals should be rare), but a signal should be received very quickly when there is evidence of a process change that is of a magnitude that one wishes to detect quickly. Unfortunately, for decades ARLs were given based on the assumption of known parameter values, which is essentially analogous to acting as if an infinite sample size was used in estimating the parameters. Note, however, that the assumption of known parameters in determining ARLs is similar to assuming a known value of σ and determining a sample size, as was done in Chapters 2 and 3, for example. Recall from the discussion in Section 3.2, though, that it would be much better to utilize parameter estimation variability in such a way that a confidence interval on power is obtained. If very large numbers of observations are used in estimating parameters, confidence intervals on those parameters would be quite narrow and the parameter estimates would have very little sampling variability, which in turn would mean that power would have very little sampling variability and would render a confidence interval for power unnecessary. With control charts, there may be enough historical data available that a large amount of data could be used for parameter estimation, although it is important that such data be recent.

8.1.2 Using Software to Determine Subgroup Size

Release 11 of PASS can be used to determine subgroup size for various control charts, with PASS using simulation exclusively. Simulation wouldn't be necessary if the parameters were assumed to be known, as then the appropriate statistical theory would be used. Simulation *is* necessary when parameters are estimated, however. One very useful way in which PASS can be used is to vary the number of subgroups that are used for estimating the parameters and see how this directly affects the properties of the chart. This will essentially duplicate the work of Quesenberry (1993) but it can be a good learning tool.

8.1.2.1 $\bar{X}$-Chart

To illustrate, it is well known that the ARL for determining a $1\sigma_{\bar{X}}$ shift in the process mean when an $\bar{X}$-chart is used is 43.89 when the parameters are assumed to be known. Therefore, if an extremely large amount of historical data is used for parameter estimation, the ARL for such a shift obtained by simulation should be approximately 43.89 if a very large number of simulations are performed. Setting the number of observations for parameter estimation at 1000 in PASS and performing 50,000 simulations produces an ARL of 44.2 and a median run length of 30.0. (The difference between the last two numbers illustrates why ARL values are often supplemented with other information from the run length distribution, such as the median and various percentiles, because the run length distribution is very skewed and with highly skewed distributions the average is not a very good measure.) Here the ARL of 44.2 is slightly different from the theoretical value of 43.9 even though two large numbers were used: 1000 observations for parameter estimation and 50,000 simulations. Such differences can be explained as follows. Since 3-sigma limits on an $\bar{X}$-chart are being used and there are very small tail areas (.00135 on each end) when normality is assumed, even very small errors in parameter estimation can result in probabilities that differ more than slightly from .00135. When this occurs, run lengths can be more than slightly affected. Furthermore, the standard deviation of the run lengths will be of the same order of magnitude as the average run length, so run lengths exhibit considerable variability and some large run lengths invariably result when simulations are performed.

Thus, there are factors at work that will cause the simulated average run length to differ slightly from the theoretical value.

The in-control (i.e., no parameter change) ARL for the chart is known to be 370.37 in the parameters-known case. The run length standard deviation is, at 369.87, almost the same as the ARL, so some very large and some very small run lengths will be generated when simulations are performed, and for this reason the simulated ARL should not be expected to be very close to the theoretical ARL. The simulations will take some time since almost all of the in-control run lengths will be large. Using PASS, the simulated in-control ARL was 372.6 and the median run length was 258.0 when 20,000 simulations were performed and the parameters were estimated from 1000 observations. The simulations took 10.69 minutes.

The simulated in-control ARL, which was obtained in PASS by letting the out-of-control distribution be the same as the in-control distribution, is thus very close to the theoretical value.

All of this is fine for gaining an understanding of the importance of using a large amount of historical data in Stage 1 and obtaining a good idea of just how much data to use, but what if we are interested in determining the subgroup size? Let's again assume that there is a shift from $N(\mu, \sigma) = N(0, 2)$ to $N(1, 2)$, with N signifying a normal distribution with mean μ and standard deviation σ. It was stated earlier in this section that the simulated ARL was 44.2. Now let's assume that we want the ARL for detecting this change to have a target value of 20. How large should the subgroup size be? PASS gave $n = 8$ as the desired subgroup size, as this produced an ARL value of 17.8, not far from the target ARL of 20. This was obtained using 50,000 simulations (Monte Carlo samples in the PASS vernacular), with the chart parameters assumed known [i.e., $N(0, 2)$]. This result is almost identical when the parameters are estimated from the simulated data, but then the simulations take a very long time to run (1.17 hours when 20,000 simulations were run and for each simulated sample the parameters were estimated from 1000 observations).

The necessary sample size is easy to compute analytically for an $\bar{X}$-chart, under the assumption of normality and known parameter values. The mechanics involve computing the appropriate Z-statistic and computing the probability using the normal distribution. Specifically,

$$Z = \frac{\mu + 3\sigma_{\bar{x}} - \mu^*}{\sigma_{\bar{x}}} \tag{8.1}$$

with μ denoting the mean that was used in computing the control limits, and μ^* denoting the mean after the mean shift. Here those two means are 0 and 1, respectively. We want the value of the Z-statistic to be such that $P(Z > Z_0) = .05$ since $1/.05 = 20$, the desired ARL for this mean shift. We can usually generally ignore the other tail of the standard normal distribution (i.e., the tail that is on the opposite side of the shift), as is being done here, because that will usually have no effect on the solution. Substituting the two mean values and $\sigma = 2$ into Eq. (8.1), we obtain

$$\begin{aligned} Z &= \frac{0 + 3(2/\sqrt{n}) - 1}{2/\sqrt{n}} \\ &= 3 - \frac{\sqrt{n}}{2} \end{aligned}$$

Since we want $P(Z > Z_0) = .05$, it follows that $Z_0 = 1.645$. Thus, $1.645 = 3 - \sqrt{n/2}$, so that $\sqrt{n} = 2.710$ and $n = 7.34$. If we require that the in-control ARL be *at most* 20, we would round up to $n = 8$, which would give an ARL of 17.73, whereas $n = 7$ gives an ARL of 21.39. PASS follows this rule, which is quite reasonable and is in the same spirit as solving for a sample size such that the power is at least equal to the target value. That is, here the ARL is at least as desirable as the target value since the smaller a parameter-change ARL, the better.

8.1.2.2 S-Chart and S^2-Chart

Now let's assume that a process standard deviation is being monitored, still assuming a normal distribution, and the objective is to detect an increase from 0.1 to 0.2 within 20 samples, so the target ARL is 20. When a two-sided S-chart with .00135 probability limits is specified, using PASS with 20,000 simulations results in $n = 2$ being selected, under the assumption of normality for the individual observations. [Note that it is much better to use an S-chart with probability limits than an S-chart with 3-sigma limits because there will be no lower control limit (LCL) when the subgroup size is less than 6.]

We could try to obtain an analytical solution as was done with an $\bar{X}$-chart, but we encounter problems immediately. To illustrate, when the individual observations have a normal distribution, $W = (n{-}1)s^2/\sigma^2$ has a χ^2 distribution with $(n{-}1)$ degrees of freedom, so that $\sqrt{W}$ has a chi distribution. The upper control limit for the S-chart using .00135 probability limits is

$$\text{UCL} = \sigma\sqrt{\frac{\chi^2_{.00135,n-1}}{n-1}}$$

Since the numerator is a function of n, it can't be specified without knowing n, which is what we are trying to solve for! (Of course, we encounter the same problem when the t-distribution is involved in sample size determination.) The necessary solution can be obtained numerically, but that won't be pursued here. If a numerical solution must be obtained, which might be difficult for many control chart users without experience in obtaining numerical solutions, then one might as well use software such as PASS and use its capability to solve for n using simulation.

Since the control chart capabilities in PASS are only in the latest release (PASS 11), it seems reasonable to assume that not very many control chart users have used PASS for determining control chart subgroup sizes. Therefore, it is of interest to see how well the sample size approximation for an S^2-chart given by Bonett and Al-Sunduqchi (1994) works, and especially how it works relative to PASS, which can be regarded as the gold standard provided that a very large number of simulations is used, which of course would be the case for any software that uses simulations.

The approximation that Bonett and Al-Sunduqchi (1994) stated as being the best is the average of n_1 and n_2, with n_1 given by

$$n_1 = \left(\frac{k^{1/2}Z_{\alpha/2} + \sigma Z_\beta}{Z_{p1}}\right)^2$$

with $Z_{p1} = 2^{1/2}|k^{1/2}{-}\sigma|$ and $\sigma^2 = k$ is the hypothesized value of σ^2, which for an S^2-chart or S-chart would be the in-control value of σ^2. As indicated by Bonett and Al-Sunduqchi (1994), this sample size expression results from a simple approximation given by Duncan (1986, p. 601). They stated that Duncan's approximation is very accurate when $\alpha = \beta$, but only when $\alpha = \beta$. This condition almost certainly will not be met when an S^2-chart or S-chart is used, however, because we want α to be very small, such as .0027 as when probability limits are used, and if $\beta = .0027$, then the power would be .9973—a much larger value than a practitioner would want to use as, in particular, an extremely large sample size would often be required. Furthermore, as Table 8.1 indicates, we don't need a large power value in order to have a very small and desirable ARL.

The expression for n_2 is

$$n_2 = \left(\frac{Z_{\alpha/2} + Z_\beta}{Z_{p2}} \right)^2$$

with $\mathrm{Z}_{p2} = (|\ln(\sigma^2) - \ln(k)|)/2^{1/2}$. (It should be noted that although the title of their paper was on control charts, the methodology is presented as if the focus was on hypothesis testing. In particular, there is no mention of ARLs.)

■ EXAMPLE 8.1

Assume that the in-control value of σ^2 is 9, $\alpha = .0027$ (i.e., .00135 probability limits), we want to detect $\sigma^2 = 4$, as that would indicate significant improvement, and the desired power is .90. The value of n_1 is then

$$\begin{aligned} n_1 &= \left[\frac{9^{1/2}3 + (2)(1.28)}{2^{1/2}} \right]^2 \\ &= 66.8 \end{aligned}$$

It can be shown that the value of $n_2 = 86.0$. Of course, no one is going to use subgroup sizes this large in practice. Since we know that n_1 should not be a good approximation to the necessary sample size, what should the ARL be if we use a sample size of 86? Not surprisingly, PASS gives the ARL and median run length as 1.0, using 20,000 simulations. Thus, these sample size formulas are actually for a one-sample hypothesis test of σ^2, not for an S^2-chart, despite the way that the material is presented by Bonett and Al-Sunduqchi (1994). That is, the sample size is being determined for immediate rejection of $\sigma^2 = 9$ when $\sigma^2 = 4$. That is not of any particular interest in quality control work because immediate detection of a change in any parameter is not necessary. Apparently Bonett and Al-Sunduqchi (1994) were not aware of the fact that control chart procedures are evaluated using ARL properties.

Castagliola, Celano, and Chen (2009) considered the properties of an S^2-chart when the in-control process variance is estimated and concluded that at least 200

Stage 1 samples are needed so that the S^2-chart will have properties similar to the properties in the known variance case. They also derived new control limits when a small number of Stage 1 samples are used. ■

8.1.3 Attribute Control Charts

The situation with attribute charts is much different from the use of charts for measurement data (which are also called variables charts), because, for example, a large number of units of production might form a sample when nonconforming units are being charted, so the sample size may be in the hundreds, or even in the thousands.

The standard deviation of a sample proportion is a function of the sample size, so it would be preferable to think about detecting a specified change in the true (population) proportion, p, rather than thinking about detecting a change that is a function of $\sigma_{\hat{p}}$, the standard deviation of the sample proportion.

For example, since a p-chart should be used, in particular, to detect an improvement in product quality, we might want to detect a reduction in nonconforming (i.e., defective) items from, say, 2% to 1%. Of course, a control chart user might also be concerned about the possibility of p increasing to, say, .3, but the primary use of an attribute control chart should be to see if improvement is occurring, assuming that the quality has not already reached an acceptable level. Attribute charts simply provide a picture of quality over time; they can't be used to help identify problems that are causing quality to be less than desired.

A decision must be made regarding the form of the control limits. When p is approximately .01, the regression-based control limits of Ryan and Schwertman (1997) are superior to the customary use of 3-sigma limits since the use of the latter implies that the distribution of nonconforming units is approximately symmetric, which won't be the case when p is that small, regardless of the value of n. Those control limits for a p-chart are

$$\text{UCL} = 0.6195/n + 1.00523p + 2.983\sqrt{p/n}$$

$$\text{LCL} = 2.9529/n + 1.0195p - 3.2729\sqrt{p/n}$$

When p is assumed to be .02, for example, LCL $= 2.9529/n + 1.0195(.02) - 3.2729\sqrt{.02/n}$. Considering only the LCL for the moment, and assuming the desired power is .80, simply for illustration, we want

$$P(\hat{p}< 2.9529/n + 0.0204 - 0.4629/\sqrt{n} \mid p = .01) = .80 \qquad (8.2)$$

Note that Eq. (8.2) does not lend itself to an algebraic solution for n, which would have to be obtained iteratively.

Since the binomial distribution must be used directly, it would be preferable to convert Eq. (8.2) to a form corresponding to the number of nonconforming units being plotted rather than the proportion. That form, which is obtained by multiplying the inequality terms by n, is

$$P(X < 2.9529 + 0.0204n - 0.4629\sqrt{n}|p = .01) = .80 \qquad (8.3)$$

with $X \sim$ Binomial(n, .01) denoting the number of nonconforming units in a sample of size n. The iterative task can be simplified by starting with the solution obtained using available software, none of which uses the Ryan–Schwertman limits, but the solution should be a good starting point. The solution obtained using PASS is $n = 2460$ if we let $\alpha = .00135$. Substituting $n = 2460$ into Eq. (8.2) and solving for X produces $X = 30.18$. Since P ($X \leq 30$| $n = 2460$ and $p = .01) = .8819$, this is not the solution for n that satisfies Eq. (8.2), although it is at least in the ballpark, as expected. [It should also be noted that $P(X \leq 29$| $n = 2460$ and $p = .01) = .8401$,with 29 being the largest value of X that corresponds to a power of at least .80, so Eq. (8.2) misses the value of X by one unit for this example.]

Since there is apparently no software that uses the Ryan–Schwertman control limits, there is thus no software that can be used to solve for n given an initial starting value. It is not difficult to solve iteratively for n, however, such that Eq. (8.2) is satisfied. Using guided trial and error, if $n = 2165$, the LCL $= 25.58$ (giving a lower tail area of .00169), and $P(X \leq 25$| $n = 2165$ and $p = .01) = .800469$. Thus, with the Ryan–Schwertman approach, a sample size of $n = 2165$ could be used and it can be noted that this time the use of Eq. (8.2) to solve for X does give the correct value as the solution is $X = 25.58$, so that the LCL expressed in terms of X is $X = 26$. Note also that this gives the *actual* power here, conditioned on the assumed value of p, because the power is computed using the binomial distribution. (It should be noted that the Ryan–Schwertman approach was not intended to be used to determine sample size, however.)

This might seem like a very large sample size, but if p is small, a large sample would have to be used in order to observe any nonconforming units. This is not built into any sample size determination algorithm in software, but it is an important consideration when the control chart is used. For example, if $p = .001$, many samples of $n = 1000$ would not have any nonconforming units.

There is not much chance that a sample of size $n = 2165$ would be devoid of nonconforming units if $p = .01$, however, since the expected number of nonconforming units is 21.65. Therefore, it isn't necessary to use this large a sample in order to virtually ensure that some nonconforming units will be present. Furthermore, we don't need power as large as .80, since that would produce ARL $= 1/.80 = 1.25$. Although it would certainly be desirable to detect a 50% reduction in the true proportion this quickly, quality practitioners would almost certainly prefer slower detection and a smaller sample size.

It should be noted that samples are obtained repeatedly when control charts are used, so sample size considerations that are made when a single hypothesis test is to be performed do not apply to control charts.

Therefore, let's now assume that power of .20 would be acceptable, which would produce an ARL of 5. Unfortunately, using "Tests for One Proportion (Proportions)" in PASS won't work because when a power is less than .50, PASS uses 1 minus the specified power, and the same thing happens with the other PASS routines for proportions. MINITAB cannot be used to determine sample size for a p-chart in quality improvement work using the binomial distribution because it uses a normal approximation in its hypothesis testing for proportions, but that won't work when p is small. The use of nQuery, which uses the exact binomial test, shows a power of .2008 for $n = 905$, this apparently being as close as one can come to .20 and still be at least .20, this result being obtained through trial and error because in nQuery it is not possible to solve for the sample size for a stated value of the power in testing a proportion. Since the expected number of nonconforming units is 9.05 for a sample of this size, the sample size should be adequate.

Thus, if we lower the power substantially so as to produce a slightly larger ARL value, the sample size can be reduced considerably.

This approach can be compared with the methods discussed by Morris and Riddle (2008), although they gave only a lower bound on sample size for each of the methods that they presented and they did not consider hypothesis testing or changes in the value of p. Instead, they concentrated on the sample size necessary to make the probability of zero nonconforming units in a sample correspond to the probability that $Z \leq -3$, with $Z \sim N(0, 1)$. In other words, they appropriately focused on the use of a p-chart to detect quality improvement, with the latter detectable by having a LCL with the lower tail area being approximately what it would be for a chart with 3σ limits under the assumption of normality.

Morris and Riddle (2008) presented four methods, the first of which was based on the normal approximation to the binomial distribution, and they gave a lower bound for n, which was $n \geq 9(1-p)/p$. That isn't quite right, however, because it can easily be shown that the LCL on both a p-chart and an np-chart with 3σ limits will be zero when $n = 9(1-p)/p$. Zero cannot be a LCL because the definition of a LCL is that it is a value below which a plotted point can fall. So the lower bound on n in order for the LCL to exist is $n > 9(1-p)/p$.

The normal approximation won't work well at all when p is quite small, however, as was demonstrated by Ryan (2011, Chapter 8). (Of course, in quality improvement applications we hope that p is indeed very small!) Recognizing this shortcoming, Morris and Riddle (2008) presented the minimum sample size for the modified p-chart given by Winterbottom (1993), although they again erred slightly in giving the minimum sample size such that the LCL could be zero. Another approach that they gave was to determine the minimum sample size by working directly with the binomial distribution, which leads to the well-known result $n > \log(.00135)/\log(1-p)$, which results from setting $P\,(X = 0)$ to .00135, which is reasonable when p is very small. The other method that they gave was a rule-of-thumb based on the Poisson approximation to the binomial distribution, which was $n \geq 6.608/p$. They stated that it is better to use $n \geq 6.6/p$ when $p \geq 0025$. That rule-of-thumb won't work for a chart with 3σ limits, however, because the requirement that $n > 9(1-p)/p$ is not met when $p = .01$, for example, so there would not be a LCL.

Of the four methods that they presented, the authors indicated their preference for determining the minimum sample size directly from the binomial distribution.

Yang, Xie, Kuralmani, and Tsui (2002) considered the determination of sample size for a geometric chart, which is an alternative to a p-chart when p is extremely small (such as .001), as then no method of determining the LCL for a p-chart will result in the chart having satisfactory properties.

Nomograms are also available for determining sample sizes, but these of course are for a specific form of a hypothesis test statistic and, more specifically, standard forms. Many such nomograms are given by Brush (1988).

8.1.4 CUSUM and EWMA Charts

CUSUM and EWMA charts are alternatives to the use of Shewhart charts and differ from those charts in the sense that data are accumulated over time and used in computing the value of each statistic at each point in time, whereas with a Shewhart chart only the data at each point in time are used when a chart is constructed. The name "CUSUM" reflects the nature of the chart, as the acronym stands for "cumulative sum" chart, whereas "EWMA" stands for "exponentially weighted moving average" chart.

Both types of charts use recent data in addition to the current data, but they do so in different ways. A basic CUSUM scheme employs two cumulative sums

$$S_{H_i} = \max[0, (Z_i - k) + S_{H_{i-1}}] \qquad S_{Li} = \min\,[0, (Z_i + k) + S_{Li-1}]$$

with i denoting the current subgroup or individual observation and Z_i is the "Z-score" as used in introductory statistics courses (i.e., it is the number of standard deviation units that the subgroup mean or individual observation is from the assumed mean), and as used in Section 2.1. The first cumulative sum is for detecting mean increases and the second is for detecting mean decreases. The value of k is chosen to be one-half the mean shift, in standard deviation units, that one wishes to quickly detect. Thus, if the focus is on quickly detecting a one-sigma shift, then k would be set to 0.5. There are variations of the basic CUSUM, but those will not be discussed here. See Ryan (2011, Chapter 8).

An exponentially weighted moving average (EWMA) chart weights data based on their proximity to the current point in time. The weighting, when subgroup averages are used, is done as

$$w_t = \lambda \bar{x}_t + (1 - \lambda) w_{t-1} \tag{8.4}$$

with λ the weight applied to the current subgroup average (i.e., at time "t"), with w_{t-1} denoting the value of the EWMA statistic at the previous time period.

Both types of charts can be used either with individual observations or with subgroups, as indicated previously.

8.1.4.1 Subgroup Size Considerations for CUSUM Charts

When used with subgroups, the CUSUM user can determine a reasonable subgroup size by looking at the ARLs for various subgroup sizes. These are given in Table 8.2 for a basic CUSUM for the same range of possible subgroup sizes as was used in Table 8.1.

The ARL values in parentheses are the ones obtained using the `getarl.exe` program of Professsor Doug Hawkins, which can be accessed at `ftp://ftp.stat.umn.edu/pub/cusum`. Those ARL values are computed analytically and note the almost exact agreement between the simulated values using PASS and the analytical values, as would be expected when 50,000 simulations are used. (PASS can be used to solve for either the subgroup size or the run length distribution. The former is illustrated in Table 8.4 in Section 8.1.4.3.) It should also be noted that these are "zero start" ARLs; that is, the CUSUM values for S_H and S_L are assumed to equal zero at the time that the shift occurs. Of course, the values won't necessarily be zero and Hawkins's applet additionally shows the

Table 8.2 ARL for Detecting a 1σ Upward Mean Shift with a Basic CUSUM Chart with Subgroup Size *n* ($k = 0.5, h = 5$), Using PASS with 50,000 Simulations

n	ARL	Median RL	Value of a in $a\ \sigma_{\bar{x}}$ Mean Change
1	10.39 (10.38)	9.00	1.00
2	6.23 (6.22)	6.00	1.41
3	4.79 (4.78)	4.00	1.73
4	4.01 (4.01)	4.00	2.00
5	3.52 (3.52)	3.00	2.24
6	3.18 (3.18)	3.00	2.45
7	2.93 (2.93)	3.00	2.65
8	2.73 (2.73)	3.00	2.83
9	2.57 (2.57)	2.00	3.00
10	2.45 (2.44)	2.00	3.16

"steady state ARL," which is obtained after the CUSUM has been running and reached a "steady state" and is lower than the zero start ARL. Although PASS has some CUSUM options, it does not give steady state ARLs, so a comparison with the results of Hawkins's applet thus cannot be made for steady state ARLs.

The purpose of the last column is to show the magnitude of the mean shift in units of multiples of $\sigma_{\bar{x}}$ such that the upward mean shift equals $\mu + \sigma_x$, as stated. because this is how the shift is customarily given in research papers on CUSUM procedures and SPC books. In particular, this facilitates comparison with tabular values for the integer values of $a = 1, 2$, and 3 that are given in such sources and there is agreement for those values. The numbers in that column show that most of the mean shifts in the table are large ones in terms of $\sigma_{\bar{x}}$, so the corresponding ARL values are small.

Comparing Table 8.2 with Table 8.1, note that an $\bar{X}$-chart is more powerful than the Basic CUSUM once the sample size reaches 5. It is well-known that the $\bar{X}$-chart is more powerful than the Basic CUSUM at detecting large shifts, and by fixing the shift at σ, as is done here, rather than in terms of $\sigma_{\bar{X}}$ as is usually done, the shift relative to $\sigma_{\bar{X}}$ is thus increasing relative to the latter as n increases since $\sigma_{\bar{X}}$ decreases as n increases. So, viewed in this way, the shift is becoming "larger." For example, for a $3\sigma_{\bar{X}}$ upward shift in the mean, the ARL for the $\bar{X}$-chart is 2.0 since the probability of a point plotting above the UCL is .5, since the distribution with the new mean is centered at the UCL. This is less than the ARL for a Basic CUSUM for that shift, which is 2.57.

8.1.4.2 CUSUM and EWMA Variations

There are variations of Basic CUSUM and EWMA charts that, for example, try to capture the value of an $\bar{X}$-chart for detecting large mean shifts. This is accomplished by adding Shewhart-type limits, with the resultant charts being called Shewhart–CUSUM and Shewhart–EWMA charts, respectively. PASS will determine sample size for Shewhart–CUSUM and Shewhart–EWMA charts, as well as a fast initial response (FIR) CUSUM, which is protection against all of the assignable causes possibly not removed after a process is stopped and the search for assignable causes initiated. The sample size for each type of chart is determined using simulation.

8.1.4.3 Subgroup Size Determination for CUSUM and EWMA Charts and Their Variations

For an EWMA chart, the tables of Lucas and Saccucci (1990) can be used to select the subgroup size, the value of λ [in Eq. (8.4)], and the value of L for L-sigma control limits. Alternatively, PASS could be used to determine subgroup size and in general to design an EWMA chart. This can be illustrated and discussed relative to the Lucas and Saccucci (1990) tables, as follows. A small value of λ is often recommended and expanded tables beyond those given in Lucas and Saccucci (1990) suggest that a reasonable combination is $L = 3.00$ and $\lambda = .25$.

Table 8.3 ARL for Detecting a 1σ Upward Mean Shift with a Basic EWMA Chart with Subgroup size *n* ($L = 3.00$, $\lambda = .25$), Using PASS with 200,000 Simulations

n	ARL	Median RL	Value of a in $a\ \sigma_{\bar{x}}$ Mean Change
1	10.40	8.00	1.00
2	5.31	5.00	1.41
3	3.71	3.00	1.73
4	2.94	3.00	2.00
5	2.47	2.00	2.24
6	2.15	2.00	2.45
7	1.92	2.00	2.65
8	1.75	2.00	2.83
9	1.61	1.00	3.00
10	1.51	1.00	3.16

As in Table 8.2, we will assume a 1σ increase in the population mean and will look at ARLs for the same values of n (see Table 8.3).

It is of interest to run PASS and see what sample size it selects for each of these charts. Following Ryan (2011, p. 275), we will assume Shewhart limits of 3.7σ for the Shewhart–CUSUM chart (and also for the Shewhart–EWMA chart), with $h = 5.2$ and the headstart value for FIR CUSUM then $5.2/2 = 2.6$. The results are given in Table 8.4.

Jones, Champ, and Rigdon (2001) examined the properties of an EWMA chart with estimated parameters. One of their conclusions was that the effect of estimating σ was most pronounced when λ was small. In particular, they stated that when $\lambda = .1$, 400 samples of size 5 in Stage 1 are needed to achieve the same in-control performance as in the parameter known case. Jones et al. (2001) obtained their results analytically but we can, of course, also obtain the results by simulations in PASS. For example, when the mean and variance are estimated using 250 samples of size 5 and the EWMA chart parameters are $L = 3$ and

Table 8.4 Subgroup Sizes and ARLs for Basic CUSUM and EWMA Charts and Their Variations, Using PASS with 50,000 Simulations and a Target ARL of 3.0 for Detecting a 1σ Mean Shift

Chart Type	Sample Size	ARL
CUSUM	8	2.815
Shewhart–CUSUM	7	2.786
FIR CUSUM	3	2.873
EWMA	4	2.916
Shewhart–EWMA	4	2.939

$\lambda = .1$, the ARL for detecting a 1σ mean shift is 2.448 using 20,000 simulations. When the parameters are assumed to be known, the ARL is 2.446. Thus, the ARLs are essentially the same, as expected because of the number of samples that were used. (The results should be virtually the same with a larger number of simulations.) When only 40 preliminary samples are used, however, the ARL is 2.46 for the 20,000 simulations, thus showing a 0.57% ARL inflation. It is much worse when only 20 preliminary samples are used, as then the ARL is 2.513, which is a 2.7% inflation, and the ARL is 2.576 when only 10 preliminary samples of size 5 are used, again when 20,000 simulations are used (and 2.595 using 100,000 simulations). Practitioners may not be overly concerned about these differences, however.

8.1.4.4 EWMA Applied to Autocorrelated Data

Autocorrelated data frequently occur in practice and Lu and Reynolds (1999) briefly considered the number of observations that are needed for parameter estimation. They concluded that values of 100 or less that are typically used in applications are much too small and that the number "should be larger by an order of magnitude." Köksal, Kantar, Ula, and Testik (2008) also considered EWMA and other types of charts applied to autocorrelated data and addressed the question of how large the Stage 1 sample size should be for effective process monitoring in Stage 2.

8.1.5 Adaptive Control Charts

It was stated in Chapter 7 that adaptive clinical trials have become popular. The same general concept can be applied to control charts. That is, the sample size and sampling interval can be variable rather than fixed and be determined by recent control chart results. For example, if a process has stabilized and out-of-control conditions are very rare, there is no point in collecting data as frequently as data were collected when the process was not doing very well. There has been a moderate amount of research on designing control charts with variable sample size and variable sampling interval. Included in this category are papers by Wu and Luo (2004), Wu, Zhang, and Wang (2007), Tseng, Tang, and Lin (2007), Jensen, Bryce, and Reynolds (2008), and Zhou, Wang, and Li (2009). Jensen et al. (2008) recommended that adaptive control charts be used for "mature processes." Certainly, that is good advice as processes should be stable and have a recent history of stability so that their performance is fairly predictable.

8.1.6 Regression and Cause-Selecting Control Charts

A regression control chart was proposed by Mandel (1969) and is used for testing a process or product characteristic that naturally changes over time, but might change at an unexpectedly fast rate and thus be out of control. An example would be tool wear. Tools wear out but the wear at any point in time should be within expected limits.

The concept is similar to a prediction interval in simple linear regression, whose width depends partly on the sample size. Specifically, the centerline on the chart is at $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ and the control limits are given by $\hat{Y} \pm 2s$, rather than from

$$\hat{Y} + 2s\sqrt{1+\frac{1}{n}+\frac{\left(X_0-\bar{X}\right)^2}{\sum\left(X-\bar{X}\right)^2}}$$

which would seem to be the natural choice for obtaining the control limits. The argument given by Mandel (1969) is that the use of control limits obtained from $\hat{Y} \pm 2s$ will provide tighter control for extreme values of X_0, which is where tight control is generally desired. Thus, Y is controlled through $\hat{Y}$. Since the control limits are not a direct function of n, general guidelines, such as the 10:1 rough rule-of-thumb proposed by Draper and Smith (1998), could be employed to determine the sample size.

A cause-selecting control chart is similar to a regression control chart, with the former used to distinguish between quality problems that occur in one stage of a process from quality problems that occur at a previous stage. Letting Y denote the output from the "current" stage and X denote the output from the previous stage, Zhang (1984, 1985) termed a plot of the standardized residuals of Y on X a cause-selecting control chart. Since a residual that is outside the control limits could signify that either Y or X is out of control, or both, a control chart on X would be maintained to check on the control of X. As recommended by Ryan (2011, p. 420), the control limits for X should be constructed as $\bar{x} \pm 3s_x/c_4$ for Stage 2. Although c_4 is a function of n, the values of c_4 are almost constant in the neighborhood of sample sizes that are likely to be used. Therefore, sample sizes that are used should be determined by rules-of-thumb such as the one given at the beginning of this section.

Kim, Mahmoud, and Woodall (2003) proposed a method for monitoring a linear regression equation in Stage 2 (Phase 2, in their terminology) that utilized three separate EWMA charts: one for the Y-intercept, one for the slope, and one for the error variance. They did not address the sample size that should be used, however. Although an EWMA chart applied to data in subgroups would generally have the same subgroup size as an $\bar{X}$-chart, that won't work for monitoring a linear profile because it would be undesirable to try to estimate three parameters—β_0, β_1, and σ_ϵ^2 (the variance error) using only 4 or 5 observations. Therefore, even though subgroups would be obtained with the same frequency as when an $\bar{X}$-chart is used, it would be desirable to use a subgroup size of approximately 10 so that the estimates of the three parameters would not have a large sampling variability.

Although the control limits recommended by Kim et al. (2003) for monitoring β_0 and σ_ϵ^2 are a function of the subgroup size, this is not the case for the EWMA chart for monitoring β_1 since the variance of $\hat{\beta}_1$ is not a function of the sample size. Whether the control limits reflect the sample size or not, it is desirable to use enough observations to estimate the parameters so that the sampling variability will not be excessive and the performance of the chart will not have unacceptable variability.

This issue has apparently not been addressed in the quality improvement literature and the problem is bypassed by selecting a value for L with L-sigma limits such that the in-control ARL is acceptable.

8.1.7 Multivariate Control Charts

Multivariate analogues to the univariate control charts that have been presented in this chapter have been given in the literature, but unfortunately there is apparently no software that can be used to determine subgroup size for a multivariate chart for subgroup averages, or to solve for sample size for a multivariate attributes chart. Therefore, guidance must come from the literature.

Champ, Jones-Farmer, and Rigdon (2005) studied the properties of a T^2 control chart and gave their recommendations for the amount of historical data to use for parameter estimation in Stage 1, in addition to the subgroup size for Stage 2. One of their recommendations is that the subgroup size should move in the opposite direction as the number of variables used in a T^2 control chart, so that the subgroup size needed to achieve desired performance decreases as the number of variables increases.

They also pointed out that when parameters are estimated (which of course will almost always be the case), the T^2 control chart, which has only a UCL, will not have an in-control ARL of $1/\alpha$ when the mean vector and covariance matrix are estimated, with the UCL given as

$$\mathrm{UCL} = \frac{p\,(m+1)\,(n-1)}{m\,(n-1)+1-p} F_{1-\alpha,p,\,mn-m-p+1}$$

with p denoting the number of variables, m the number of subgroups, and n the subgroup size, and $F_{1-\alpha,p,\,mn-m-p+1}$ is the $(1-\alpha)$ percentile of the F-distribution with p and $(mn - m - p + 1)$ degrees of freedom.

In particular, their Table 1 gave in-control ARL values when the parameters were estimated using $m = 50$ subgroups of size $n = 5$. The ARLs ranged from 11% higher than the in-control ARL assuming known parameter values when $p = 10$ to 17% higher when $p = 2$. A somewhat nonintuitive result from their research was their statement that the out-of-control performance of the T^2-chart depends only on the "statistical distance between the in-control and out-of-control mean vectors" and thus does not depend on their parameter values or their estimates, and thus must also not depend on the sample size. This is true provided that the control limit is adjusted so that the in-control ARL with the adjusted limit matches the in-control ARL for the standard limit based on the assumption of known parameter values.

Therefore, they focused attention on the number of subgroups and subgroup size that should be used for a given number of variables so that the in-control ARL differs from the theoretical in-control ARL by a specified percentage and provided graphs that users could utilize in designing a T^2 control chart. For example, let $p =$ the number of variables, $m =$ number of subgroups, and $n =$ subgroup size. For $p = 10$, they recommended $mn > 550$, if the goal is to have the in-control ARL no more than 5% higher than the desired ARL of 200. As they stated, this might consist of 110 subgroups of size 5. Their recommendations for other numbers of variables monitored by the chart were given in their Table 3. Specifically, they recommended sample sizes greater than 900, 700, 650, and 600 for $p = 2$, 4, 6, and 8, respectively.

Champ et al. (2005) stated that their sample size recommendations call for the use of a larger amount of data than do the recommendations of Lowry and Montgomery (1995) and Nedumaran and Pignatiello (1999), and they explain that is due to goals being different.

8.2 MEDICAL APPLICATIONS

Certain applications, such as medical applications, present special challenges and problems that do not exist in manufacturing applications. For example, whereas a decision could be made to chart the average diameter of five ball bearings every 15 minutes when thousands are rolling off an assembly line every hour, it certainly would not be possible to chart the average of some characteristic relating to coronary bypass operations every 15 minutes, as the operations almost certainly would not be performed at virtually the same time. Therefore, as Shahian, Williamson, Svensson, Restuccia, and D'Agostino (1996) discussed, it is more convenient to analyze the patients on a time interval, such as a week, month, or quarter. This will invariably result in a varying sample size and the term "subgroup" is not particularly applicable since the measurements are not being made close together in time, although the authors did use that term, while recognizing that measurements should be made close enough together in time so that conditions are uniform. The authors stated: "In the instance of cardiac surgical morbidity and mortality, which ranges from 2% to 5% for most adverse outcomes, a subgroup size of 50 to 100 patients would be desirable." This led them to choose quarterly reporting of data.

The authors also stated that for subgroup sizes greater than nine, an *S*-chart (a chart of subgroup standard deviations) should be used instead of an *R*-chart (a chart of subgroup ranges). So here a decision is being made as to what chart to use or recommend based on the subgroup size. This illustrates some “suboptimal thinking” because an *S*-chart should be used for any subgroup size, with the two charts being equivalent when the subgroup size is two. When the subgroup size is greater than two, all of the data are not being used in calculating the range since the range is the largest value minus the smallest value. Efficiency is lost—with any statistical procedure—when data that have been recorded are not used. There is also a problem with the acceptable minimum number of subgroups that they mention, as they state the minimum as being 10. The minimum should depend on the subgroup size, as the total number of observations should be at least 100. In their case, they used 17 consecutive quarters, so more than enough data were used in determining control limits. Of course, many changes in conditions could occur over such a long period of time. They obviously recognized this possibility in stating that any changes in the patient population, staff, procedures, or equipment should be noted.

It is not clear from their article whether they are distinguishing between Stage 1 and Stage 2, however, and this is true of many articles on control charts in the applied literature.

Unfortunately, there have been many applications of control charts in which the number of observations used to determine the control limits is far less than desirable. For example, Mohammed, Worthington, and Woodall (2008) cited Carey (2003) as stating that at least 20–25 observations are needed to compute the control limits for an *X-MR* chart, which is a combined chart of individual observations and moving ranges. As indicated in Section 8.1.1, that is far less than the number of observations that should be used.

8.3 PROCESS CAPABILITY INDICES

Process capability indices are used to determine how well a process performs relative to engineering tolerances. Many such indices have been proposed during the past 25 years and no attempt will be made to review them here, other than to state that improved indices appeared starting in the mid-1980s and on through the 1990s. These indices were superior to the most frequently used index during the early 1980s, which was $C_p = (\text{USL} - \text{LSL})/6\sigma$, with USL and LSL denoting the upper specification limit and the lower specification limit, respectively. This index was of limited value because it was not a function of the process mean. A better capability index is C_{pk}, which is defined as $C_{pk} = \frac{1}{3}Z_{\min}$, with $Z_{\min}$ defined as the minimum of $Z_1 = (\text{USL}-\mu)/\sigma$ and $Z_2 = (\mu-\text{LSL})/\sigma$. C_{pk} is thus a function of a *z*-score, and if USL and LSL are a large distance from μ, relative to σ, so as to make Z_1 and Z_2 large, estimating C_{pk} is then almost like estimating an extreme percentile of a distribution, which generally requires thousands of observations. Here μ and σ are being estimated, however, not a percentile, but the effect is similar since, for example, if (USL$-\mu$) is large, misestimation of σ by a large percentage amount could have a sizable effect on the Z-score.

The reader interested in a detailed presentation of process capability indices is referred to Pearn and Kotz (2006), Kotz and Lovelace (1998), and Kotz and Johnson (1993).

Process capability indices must be estimated, so the necessary sample size for such estimation must be addressed. The following statements are made at `http://www.itl.nist.gov/div898/handbook/pmc/section1/pmc16.htm`: "Most capability indices estimates are valid only if the sample size used is 'large enough.' Large enough is generally thought to be about 50 independent data values." Most capability indices assume a normal distribution and the indices are sensitive to that assumption. It would be preferable to have somewhat more than 50 observations to test the normality assumption. Of course, "power" is not relevant here, because nothing is changing (presumably), so it is not a matter of having a large enough sample size to detect a change quickly. Instead, the objective should be to use a large enough sample so that a good picture of the present can be obtained. Sample size determination for one of the process capability indices, C_{pm}, has been considered by Zimmer, Hubele, and Zimmer (2001). See also Pearn and Shu (2003) and Shu and Wang (2006), as the latter reviewed existing formulas for the lower confidence bounds of certain process capability indices and proposed a new approach for sample size determination for C_p and C_{pm}.

8.4 TOLERANCE INTERVALS

The most commonly used type of tolerance interval is an interval on population values that a user can state, with a specified degree of confidence, will contain at least a certain proportion of population values. A lesser-known type of tolerance interval is one that contains at least a certain proportion of the center of the population with a specified degree of confidence, which is usually referred to as an equal-tailed tolerance interval (Krishnamoorthy and Mathew, 2009, p. 4).

Tolerance intervals are used in engineering statistics and in quality improvement work. Interest in determining sample size for tolerance intervals dates at least from the work of Wilks (1941), who presented both a nonparametric approach and a method based on the assumption of a normal distribution. With the former it was a matter of determining how large a sample should be so that the largest and smallest sample values can form the tolerance interval.

Of course, there has been much research work on tolerance intervals since 1941 and, for example, Chapter 9 of Hahn and Meeker (1991) is a 37-page chapter entitled "Sample Size Requirements for Tolerance Intervals, Tolerance Bounds, and Demonstration Tests," although the chapter consists almost entirely of nomograms.

The term “tolerance interval” has been used in a potentially confusing way in some places. For example, the Power and Precision software gives a “Tolerance Interval for the 95%, 1-tailed Confidence Interval” for the t-test option where the parameters are being estimated. First, there is no such thing as a one-tailed confidence interval, as one of the endpoints is either plus infinity or minus infinity, so we cannot have a finite confidence *interval*. The appropriate term is “confidence bound” or perhaps “confidence limit.” This is a somewhat common error, however.

Second, the terms “tolerance interval” and “confidence interval” should not be used in the same expression, as they are totally different. The Power and Precision software is simply indicating a range for the confidence intervals that would occur with repeated sampling. Although that is certainly of some interest as an aid in understanding confidence interval variability, only one sample is generally taken.

What the two types of intervals do have in common is that their width is a direct function of the variability of the estimator of σ, as s is used in constructing a tolerance interval just as it is used in constructing a confidence interval for the mean of a normal distribution. Specifically, the tolerance interval is $\bar{x} \pm ks$ if the interval is to be symmetric about $\bar{x}$, with k chosen from a table to give the desired coverage probability with the desired degree of confidence. Thus, since the expression for the endpoints of the interval is not a function of n, we can’t use the expression to solve for n.

The situation is different for a nonparametric tolerance interval, as there the objective is to take a large enough sample so that, for example, the largest and smallest observations in the sample can form the tolerance interval. Algorithms for determining sample size for this purpose date from at least Brooker and Shelby (1975), although the latter does not guarantee approximately equal tail probabilities, whereas the approach of Neave (1978) does so, as well as the tolerance interval approach given by Chou and Mee (1984). See also Chou and Johnson (1987), who gave tables of minimum sample sizes. See also the tables of sample sizes for normality-based and nonparametric tolerance intervals given by Kirkpatrick (1977). Another excellent source is Odeh and Owen (1980) but that book has been out of print for many years.

Although tolerance intervals are generally either based on the assumption of normality or else are nonparametric tolerance intervals, Chen and Yang (1999) gave Monte Carlo algorithms for estimating the sample size and coverage for tolerance intervals for nonnormal distributions.

A two-sided β-expectation tolerance interval is defined (see, e.g., Mee, 1984) in the following way for a normally distributed random variable, X: $E_{\hat{\mu}\hat{\sigma}}\{P_x[\hat{\mu} - k\hat{\sigma} < X < \hat{\mu} + k\hat{\sigma} | \hat{\mu}, \hat{\sigma}]\} = \beta$. In words, the expectation of the tolerance interval with endpoints $\hat{\mu} - k\hat{\sigma}$ and $\hat{\mu} + k\hat{\sigma}$ is β, so on average the coverage is β. There is no guarantee about the coverage in each tail of the distribution but Chou (1984) went a step further and determined sample sizes for β-expectation tolerance intervals that control the proportion in each tail of a normal distribution. Odeh, Chou, and Owen (1989) explained the difference between a β-expectation tolerance interval and a β-content tolerance interval, with the latter constructed to contain at least $100\beta\%$ of the population with a stated degree of confidence. This is the most frequently used of the two types.

Fountain and Chou (1991) determined minimum sample sizes for two-sided β-content tolerance intervals. This type of tolerance interval differs from a β-expectation tolerance interval in that it is defined as $P_{\hat{\mu}\hat{\sigma}}\{P_x[\hat{\mu} - k\hat{\sigma} < X < \hat{\mu} + k\hat{\sigma} \mid \hat{\mu}, \hat{\sigma}]\} \geq \beta$ for some confidence coefficient γ.

Early work on sample size determination for tolerance limits includes Faulkenberry and Weeks (1968) and Faulkenberry and Daly (1970). See also Sa and Razaila (2004), who proposed an algorithm for constructing one-sided tolerance limits continuously over time for any distribution. They showed that the sample size required by their method is reduced over time. Wang and Tsung (2009) considered tolerance intervals for binomial and Poisson random variables and proposed procedures for calculating the minimum and average coverage probabilities.

8.5 MEASUREMENT SYSTEM APPRAISAL

Measurement variability consists of variability due to materials and people, such as parts and operators. Variance components are estimated (such as σ^2_{parts}) and enough observations must be used in order to obtain good estimates of those components. When used in measurement system appraisal this is called a Gauge *R* and *R* study, with the two *R*s representing "reproducibility" and "repeatability," respectively.

Confidence intervals are often constructed for variance components and a desired maximum width of such intervals might be specified, which would lead to sample size determination. Readers interested in more information about confidence intervals for variance components are referred to Burdick and Larsen (1997). Their general recommendation is that it is important to obtain enough samples and operators, these being more important than replicates. Specifically, they stated that 10 samples, 3 operators, and 3 replicates is typical in published Gauge *R* and *R* studies.

8.6 ACCEPTANCE SAMPLING

Sample size determination is covered briefly here (and deliberately placed at the end of the chapter) because some practitioners still use it, not because it has any particular value relative to the current quality objectives of most organizations. As the prominent industrial statistician Harold F. Dodge (1893–1976) stated decades ago: "You can't inspect quality into a product."

Although tables have long been used to determine how many items to inspect and the criteria used to determine when to accept or reject a lot (such as the charts in Hahn, 1974), software can of course also be used for this purpose. STATGRAPHICS is one such software package, as it can be used for acceptance sampling with measurement data and for attribute data. These methods have historically been referred to as acceptance sampling for variables and acceptance sampling for attributes, respectively.

MINITAB can also be used to determine a sample size and sampling plan. To illustrate sample size determination for Acceptance Sampling by Variables, assume that the lot size is 25,000 units, the producer's risk (α) is to be .05; the consumer's risk (β) is set at .10; the acceptable quality level (AQL) in terms of nonconforming units per million is 10 and the rejectable quality level (RQL) is 50 nonconforming units per million. The USL is 100 and the LSL is 50. The necessary sample size given by MINITAB is 564. MINITAB could also be used to determine a sample size and sampling plan for Acceptance Sampling by Attributes.

Hauck and Shaikh (2001) stated that one approach employed in the pharmaceutical industry is to use a two-sided tolerance interval and accept a given batch if the tolerance interval falls entirely within an acceptance interval. They addressed the problem of determining power and sample size for such an approach.

8.7 RELIABILITY AND LIFE TESTING

Product reliability is also an important part of quality and quality improvement. Moura (1991) is a compact source of information on sample size determination for accelerated life testing, and Chapter 9 of Mathews (2010) contains some information on sample size determination for reliability. The Weibull probability distribution has historically played a key role in reliability analysis. Jeng (2003) gave a sample size determination approach that was based on a combination of simulation and asymptotic theory for a Weibull test plan when there is complete data or time censored data.

Since reliability is the industrial counterpart to survival analysis, reliability is also discussed briefly in Section 9.2.

8.8 SOFTWARE

Unfortunately, software developers in general have not focused attention on sample size determination for the tools that were presented in this chapter, with PASS being the most useful software and MINITAB having value in designing acceptance sampling schemes.

8.9 SUMMARY

Although control chart users have not concentrated on subgroup (sample) size determination, that doesn't mean that it shouldn't be done. Certainly, determining the amount of data to use for parameter estimation in Stage 1 is extremely important. The need for sample size determination in Stage 2 is lessened somewhat, however, by the fact that control chart usage is a continuous process, whereas hypothesis testing in general is not. Sample size determination has always been important in acceptance sampling, however, and in tolerance interval construction.

REFERENCES

Bonett, D. G. an M. S. Al-Sunduqchi (1994). Approximating sample size requirements for s^2 charts. *Journal of Applied Statistics*, **21**(5), 425–429.

Brooker, P. and M. J. P. Shelby (1975). Algorithm AS 92: The sample size for a distribution free tolerance interval. *Applied Statistics*, **24**, 388–390.

Brush, G. G. (1988). *How to Choose the Proper Sample Size*. Milwaukee, WI: American Society for Quality Control.

Burdick, R. K. and G. A. Larsen (1997). Confidence intervals on measures of variability in *R&R* studies. *Journal of Quality Technology*, **29**(3), 261–273.

Carey, R. G. (2003). *Improving Health Care with Control Charts: Basic and Advanced SPC Methods and Case Studies*. Milwaukee, WI: Quality Press.

Castagliola, P., G. Celano, and G. Chen (2009). The exact run length distribution and design of the S^2 chart when the in-control variance is estimated. *International Journal of Reliability, Quality, and Safety Engineering*, **16**(1), 23–28.

Champ, C. W., L. A. Jones-Farmer, and S. E. Rigdon (2005). Properties of the T^2 control chart when parameters are estimated. *Technometrics*, **47**(4), 437–445.

Chen, H. and T.-K. Yang (1999). Computation of the sample size and coverage for guaranteed-coverage nonnormal tolerance intervals. *Journal of Statistical Computation and Simulation*, **63**, 299–320.

Chou, Y. M. (1984). Sample sizes for β-expectation tolerance limits which control both tails of the normal distribution. *Naval Research Logistics Quarterly*, **31**(4), 601–607.

Chou, Y.-M. and G. M. Johnson (1987). Sample sizes for strong two-sided distribution-free tolerance intervals. *Statistical Papers*, **28**, 117–131.

Chou, Y.-M. and R. W. Mee (1984). Determination of sample sizes for setting β-content tolerance intervals controlling both tails of the normal distributions. *Statistics and Probability Letters*, **2**, 311–314.

Cryer, J. D. and T. P. Ryan (1990). Estimation of sigma for an X chart: $\overline{MR}/d_2$ or S/c_4? *Journal of Quality Technology*, **22**, 187–192.

Dockendorf, L. (1992). Choosing appropriate sample subgroup sizes for control charts. *Quality Progress*, October, 160.

Draper, N. R. and H. Smith (1998). *Applied Regression Analysis*, 3rd edition. New York: Wiley.

Duncan, A. J. (1986). *Quality Control and Industrial Statistics*, 5th edition. Homewood, IL: Irwin.

Faulkenberry, G. D. and J. C. Daly (1970). Sample size for tolerance limits on a normal distribution. *Technometrics*, **12**(4), 813–821.

Faulkenberry, G. D. and D. L. Weeks (1968). Sample size determination for tolerance limits. *Technometrics*, **10**(2), 343–348.

Fountain, R. L. and Y.-M. Chou (1991). Minimum sample sizes for two-sided tolerance intervals for finite populations. *Journal of Quality Technology*, **23**(2), 90–95.

Hahn, G. J. (1974). Minimum size sampling plans. *Journal of Quality Technology*, **6**(3), 121–127.

Hahn, G. J. and W. O. Meeker (1991). *Statistical Intervals: A Guide for Practitioners*. New York: Wiley.

Hauck, W. W. and R. Shaikh (2001). Sample sizes for batch acceptance from single- and multistage designs using two-sided normal tolerance intervals with specified content. *Journal of Biopharmaceutical Statistics*, **11**, 335–346.

Jeng, S. L. (2003). Exact sample size determination for a Weibull test plan when there is time censoring. *Journal of Statistical Computation and Simulation*, **73**, 389–408.

Jensen, W. A., G. R. Bryce, and M. R. Reynolds, Jr. (2008). Design issues for adaptive control charts. *Quality and Reliability Engineering International*, **24**, 429–445.

Jones, L. A., C. W. Champ, and S. E. Rigdon (2001). The performance of exponentially weighted moving average charts with estimated parameters. *Technometrics*, **43**(2), 156–167.

Kim, K., M. A. Mahmoud, and W. H. Woodall (2003). On the monitoring of linear profiles. *Journal of Quality Technology*, **35**, 317–328.

Kirkpatrick, R. L. (1977). Sample sizes to set tolerance limits. *Journal of Quality Technology*, **9**(1), 6–12.

Köksal, G., B. Kantar, T. Ali Ula, and C. Testik (2008). The effect of Phase I sample size on the run length performance of control charts for autocorrelated data. *Journal of Applied Statistics*, **35**, 67–87.

Kotz, S. and N. L. Johnson (1993). *Process Capability Indices*. New York: Chapman and Hall.

Kotz, S. and C. R. Lovelace (1998). *Process Capability Indices in Theory and Practice*. London: Hodder Arnold.

Krishnamoorthy, K. and T. Mathew (2009). *Statistical Tolerance Regions: Theory, Applications, and Computation*. Hoboken, NJ: Wiley.

Lowry, C. A. and D. C. Montgomery (1995). A review of multivariate control charts. *IIE Transactions*, **27**, 800–810.

Lu, C. W. and M. R. Reynolds (1999). EWMA control charts for monitoring the mean of autocorrelated processes. *Journal of Quality Technology*, **31**, 166–188.

Lucas, J. M. and M. S. Saccucci (1990). Exponentially weighted moving average control schemes: Properties and enhancements (with discussion). *Technometrics*, **32**, 1–29.

Mandel, B. J. (1969). The regression control chart. *Journal of Quality Technology*, **1**(1), 1–9.

Mathews, P. (2010). *Sample Size Calculations: Practical Methods for Scientists and Engineers*. Fairport Harbor, OH: Mathews, Malnar, and Bailey, Inc.

Mee, R. W. (1984). β-expectation and β-content tolerance limits for balanced one-way ANOVA random model. *Technometrics*, **26**(3), 251–254.

Mohammed, M. A., P. Worthington, and W. H. Woodall (2008). Plotting basic control charts: Tutorial notes for health care practitioners. *Quality and Safety in Health Care*, **17**, 137–145.

Morris, R. L. and E. J. Riddle (2008). Determination of sample size to detect quality improvement in p-charts. *Quality Engineering*, **20**(3), 281–286.

Moura, E. C. (1991). *How to Determine Sample Size and Estimate Failure Rate in Life Testing*. Milwaukee, WI: ASQC Quality Press.

Neave, H. (1978). Algorithm AS 124: Sample sizes for one-sided and strong two-sided distribution-free tolerance intervals. *Applied Statistics*, **27**, 188–190.

Nedumaran, G. and J. J. Pignatiello, Jr. (1999). On constructing T^2 control charts for on-line process monitoring. *IIE Transactions*, **31**, 529–536.

Odeh, R. E. and D. B. Owen (1980). *Tables for Normal Tolerance Limits, Sampling Plans, and Screening*. New York: Marcel Dekker.

Odeh, R. E., Y. M. Chou, and D. B. Owen (1989). Sample size-determination for two-sided β-expectation tolerance intervals for a normal distribution. *Technometrics*, **31**(4), 461–468.

Pearn, W. L. and S. Kotz (2006). *Encyclopedia and Handbook of Process Capability Indices: A Comprehensive Exposition of Quality Control Measures*. Series on Quality, Reliability, and Engineering Statistics, Vol. 12. Hackensack, NJ: World Scientific Publishing.

Pearn, W. L. and M.-H. Shu (2003). Lower confidence bounds with sample size information for C_{pm} applied to production yield assurance. *International Journal of Production Research*, **41**(15), 3581–3599.

Quesenberry, C. P. (1993). The effect of sample size on estimated limits for $\bar{X}$ and X control charts. *Journal of Quality Technology*, **25**, 237–247.

Ryan, T. P. (2011). *Statistical Methods for Quality Improvement*, 3rd edition. Hoboken, NJ: Wiley.

Ryan, T. P. and N. C. Sohwertman (1997). Optimal limits for attribute control charts. *Journal of Quality Technology*, **29**, 86–98.

Sa, P. and L. Razaila (2004). One-sided continuous tolerance limits and their accompanying sample size problem. *Journal of Applied Statistics*, **31**, 419–434.

Shahian, D. M., W. A. Williamson, L. G. Svensson, J. D. Restuccia, and R. S. D'Agostino (1996). Applications of statistical quality control to cardiac surgery. *Annals of Thoracic Surgery*, **62**, 1351–1358.

Shu, M. H. and C. H. Wang (2006). A review and extensions of sample size determination for estimating process precision and loss with a designated accuracy ratio. *The International Journal of Advanced Manufacturing Technology*, **27**(9/10), 1038–1046.

Tseng, S.-T., J. Tang, and C.-H. Lin (2007). Sample size determination for achieving stability of double multivariate exponentially weighted moving average controller. *Technometrics*, **49**(4), 409–419. Technical Note: *Technometrics*, **51**, 335–338.

Wang, H. and F. Tsung (2009). Tolerance intervals with improved coverage probabilities for binomial and Poisson variables. *Technometrics*, **51**, 25–33.

Wilks, S. S. (1941). Determination of sample sizes for setting tolerance limits. *The Annals of Mathematical Statistics*, **12**(1), 91–96.

Winterbottom, A. (1993). Simple adjustments to improve control limits on attribute charts. *Quality and Reliability Engineering International*, **9**, 105–109.

Woodall, W. H. (2000). Controversies and contradictions in statistical process control (with discussion). *Journal of Quality Technology*, **32**(4), 341–350.

Wu, Z. and H. Luo (2004). Optimal design of the adaptive sample size and sampling interval *np* control chart. *Quality and Reliability Engineering International*, **20**, 553–570.

Wu, Z., S. Zhang, and P. Wang (2007). A CUSUM scheme with variable sample sizes and sampling intervals for monitoring the process mean and variance. *Quality and Reliability Engineering International*, **23**, 157–170.

Yang, Z., M. Xie, V. Kuralmani, and K.-L. Tsui (2002). On the performance of geometric control charts with estimated control limits. *Journal of Quality Technology*, **34**(4), 448–458.

Zhang, G. X. (1984). A new type of control charts and a theory of diagnosis with control charts. In *ASQC Annual Quality Congress Transactions*, pp. 175–185. Milwaukee, WI. American Society for Quality Control.

Zhang, G. X. (1985). Cause-selecting control charts—A new type of quality control charts. *The QR Journal*, **12**, 221–225.

Zhou, M.-Y., X.-L. Wang, and X.-Y. Li (2009). The Effects of VSSI $\bar{X}$ chart when control limits are estimated. *Journal of Data Analysis*, **4**, 1–14 (in Chinese).

Zimmer, L. S., N. F. Hubele, and W. J. Zimmer (2001). Confidence intervals and sample determination of C_{pm}. *Quality and Reliability Engineering International*, **17**, 51–68.

EXERCISES

8.1. Derive the values for $n = 2$ that were given in Table 8.1 for power and ARL.

8.2. Control chart properties that are given in textbooks and in journal articles are based on the assumption that the observations have come from a common distribution, such as a normal distribution, or that the Central Limit Theorem can be invoked to assume approximate normality for subgroup means. The Central Limit Theorem does not apply, even approximately, for subgroup sizes that are typically used for variables control charts, such as an $\bar{X}$-chart. Assume that a Shewhart $\bar{X}$-chart with 3-sigma limits and a subgroup size of 5 is in operation. Use appropriate software to determine the ARL for detecting a mean increase of $1.5\sigma_{\bar{x}}$ when the individual observations are from a logistic distribution. Based on your answer, what would you suggest be done before control chart limits be constructed? Explain.