4.4 BATCH MEANS TO ESTIMATE THE MEAN FROM STATIONARY DATA

We have seen that two problems present difficulties in estimating the stationary mean using data from a stable dynamic simulation: the initial transient period and autocorrelated observations. In this section, we will present a simple and useful method to compute a confidence interval for the mean using a single series of data from a dynamic simulation: the batch means method.

First, let's deal with the problem of the initial transient period. Because the initial observations are not drawn from the stationary distribution, we have little choice other than to discard them and only use observations that are generated after the system's behavior is close to the stationary behavior. We thus must determine how many observations to discard at the start of the series. A number of techniques have been proposed to formally choose a truncation point (Law, 1983). All of the techniques are fairly difficult to implement. Therefore, we will use a relatively simple graphical approach due to Welch (1981). As in the previous section, we will replicate the system many times, compute the mean of the series of observations from the replications, and smooth this series of means. Then by observing the pattern of this smoothed curve, we can visually select a point at which the stationary behavior seems to dominate. In the plot of Figure 4.2, one could truncate at approximately 100 observations. Even though the plot spends much of the time below the theoretical mean of 1.633 (information that is not generally available), the behavior has stabilized by this time and thus this looks like a good truncation point. We will not be too concerned about this point because, as we will later argue, the batch means method is fairly robust when it comes to errors in choosing the truncation point.

4.4.1 Batch Means Computation

Now that the initial transient observations have been removed, we assume that all remaining observations are sampled from the system operating in steady state. The next problem is to choose a method to compute a confidence interval for the mean. Several methods have been developed for this situation (Law, 1983; Alexopoulos and Seila, 1998). The simplest method, called the *batch means method* (Conway, 1963; Fishman, 1979; Law and Kelton, 1984), groups observations into several batches, computes the sample mean of the observations in each batch, and computes the confidence interval from these batch means using traditional statistical methods that apply to independent observations. As a simple example, suppose that our data, after removing the initial transient observations, consist of the following 12 observations:

1.2, 3.3, 2.6, 5.1, 4.4, .6, 0.0, 1.3, 1.5, 3.7, 3.5, 2.4

(Of course, in an actual application we would generate thousands of observations; this small dataset is just used to illustrate the computational procedure.) Suppose that we group the observations into three batches of four observations each. Then, the batches and their means would be

Batch 1	1.2, 3.3, 2.6, 5.1	Mean	$\overline{X}_1 = 3.05$
Batch 2	4.4, .6, 0.0, 1.3	Mean	$\overline{X}_2 = 1.58$
Batch 3	1.5, 3.7, 3.5, 2.4	Mean	$\overline{X}_3 = 2.78$

Now we will treat the three batch means as if they are a sequence of independent observations from the same population whose mean we want to estimate. The sample mean of the batch means is $\overline{\overline{X}} = 2.47$, and the standard deviation is

$$s_{\overline{X}} = .78$$

Thus, a 95 percent confidence interval for the mean is

$$\begin{aligned}\overline{\overline{X}} \pm t_{.025,2}\frac{s_{\overline{X}}}{\sqrt{3}} &= 2.47 \pm 4.303\,\frac{.78}{\sqrt{3}}\\ &= 2.47 \pm 1.95\\ &= (.52, 4.42)\end{aligned}$$

Recall that $t_{\alpha,k}$ is the $100(1-\alpha)$-percentage point on the Student's t-distribution with k degrees of freedom. A review of basic statistics is presented in Appendix A.

Estimating the mean of a stationary sequence of observations produced by a dynamic simulation is a much more difficult problem than estimating the mean using independent observations, which we discussed in Chapter 3. If the amount of data is too small or the method is not applied appropriately, the confidence coefficient of the resulting confidence interval will be smaller than the presumed value. For example, if a 95 percent confidence interval is computed, we presume that the coverage probability is .95, when in fact it could be considerably less if the method is applied with too few observations or not applied correctly. The actual 95 percent confidence interval would necessarily be larger. This tricks us into thinking that our estimate of the mean is more precise than it really is.

4.4.2 Some Guidelines to Applying the Batch Means Method

Studies (Law and Kelton, 1984) have shown that the batch means method is highly competitive with other methods in terms of the accuracy and reliability of confidence intervals if it is applied correctly. When applying the batch means method using a fixed total number of observations, we must decide how many batches to form and therefore how many observations each batch will have. Theoretical considerations show that, if the batch size is sufficiently large, the batch means from separate batches will be approximately uncorrelated, even though the individual observations are autocorrelated. A central limit theorem similar to the one for independent observations applies to many processes of observations from dynamic simulations. If this is also the case, then the batch means will be approximately normally distributed when the batch size is sufficiently large, and this implies that they will be approximately independent because normally distributed random variables that are uncorrelated are also independent. Thus, our objective in applying the batch means method is to select the batch size large enough that the batch means are approximately uncorrelated—but small enough that the maximum number of batches is formed. As a rule of thumb, one should normally trade batch size for number of batches. For example, if we have generated 2000 observations, then we would generally be better off forming 10 batches of 200 observations than 20 batches of 100 observations. We usually want the number of batches to be at least 10 so that the percentiles of the t-distribution will be relatively close to those of the normal distribution, which is the smallest they can be. We also want the size of each batch to be at least 100 under most circumstances. Experience has shown that for many moderately congested systems this is the minimum number of observations required to give any assurance that the batch means are approximately uncorrelated or approximately normally distributed. Using these values, one should still consider the confidence interval computed to be a "rough" estimate.

An obvious alternative to the batch means approach is to run several independent replications, compute a sample mean from each, and use these sample means to compute a confidence interval for the mean. This approach has the appealing property that, because each replication starts from the same initial condition and is run independently of the others, we are guaranteed that the sample means are mutually independent, not just approximately uncorrelated. Any consideration of whether the batches are large enough is moot. However, this approach has two critical shortcomings: First, we must allow the simulation to run through the initial transient period in *each replication*. The batch means method requires this to be done only once—at the beginning of the run. Obviously, using independent replications is more wasteful of computation time because all data in the initial transient period are discarded. Second, if the length of the initial transient period is misjudged too small, then some or perhaps all of the observations in each replication will be collected before the system has reached steady state. This will cause the sample mean in every replication to be biased. In the batch means method, however, the second batch comes after the first, and the third after the second, and so on, so any bias in the batch means will reduce and likely be eliminated in successive batches because they are farther from the start of the run. Indeed, since batch sizes must be large, it is likely that only the first batch mean has significant bias. Thus, the batch means method is robust with respect to errors in determining the length of the initial transient period.

A useful graphical method to judge the adequacy of the initial transient period is to plot the batch means as a time series. If the first batch mean is either smaller or larger than all of the others, then there is evidence that the transient period is too short. The transient period can then be lengthened by eliminating the first batch from the sample. This procedure should be applied until the sequence of batch means appears to be stationary.