STATISTICAL INSIGHTS FOR SIMULATIONISTS

Dave Goldsman

School of ISyE
Georgia Tech
Atlanta, Georgia, USA

sman@gatech.edu
www.isye.gatech.edu/∼sman

ASIM Conference, Dortmund, Germany

September 22, 2015

Outline

Introduction

Steps in a Simulation Study:

- Preliminary Analysis of the System
- Model Building
- Verification & Validation
- Experimental Design & Simulation Runs
- Statistical Analysis of Output Data
- Implementation

Input processes driving a simulation are random variables (e.g., interarrival times, service times, and breakdown times).

Must regard the output from the simulation as random.

Runs of the simulation only yield *estimates* of measures of system performance (e.g., the mean customer waiting time).

These estimators are themselves random variables, and are therefore subject to sampling error.

Must take sampling error must be taken into account to make valid inferences concerning system performance.

Lots of measures you could be interested in:

- Means — what is the mean customer waiting time?

  Means aren't enough. If I have one foot in boiling water and one foot in freezing water, on average, I'm fine. So....
- Variances — how much is the waiting time liable to vary?
- Quantiles — what's the 99% quantile of the line length in a certain queue?
- Success probabilities — will my job be completed on time?

Would like point estimators and confidence intervals for the above measures.

Problem: simulations almost never produce raw output that is independent and identically distributed (i.i.d.) normal data.

Example: Customer waiting times from a queueing system...

(1) Are not independent — typically, they are serially correlated. If one customer at the post office waits in line a long time, then the next customer is also likely to wait a long time.

(2) Are not identically distributed. Customers showing up early in the morning might have a much shorter wait than those who show up just before closing time.

(3) Are not normally distributed — they are usually skewed to the right (and are certainly never less than zero).

Archetypal Example: Suppose that $Y_1, Y_2, \ldots, Y_n$ are stationary (i.e., the joint distribution doesn't change over time) but *not independent*. Is the sample mean $\bar{Y}_n \equiv \frac{1}{n}\sum_{i=1}^{n} Y_i$ a good estimator for $\mu = \mathrm{E}[Y_i]$?

$\mathrm{E}[\bar{Y}_n] = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}[Y_i] = \mu$, so it's still unbiased.

On the other hand, let's define the *covariance function*, $R_k \equiv \mathrm{Cov}(Y_1, Y_{1+k})$, $k = 0, 1, 2, \ldots$. Then

$$\begin{aligned} \mathrm{Var}(\bar{Y}_n) &= \mathrm{Cov}(\bar{Y}_n, \bar{Y}_n) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{Cov}(Y_i, Y_j) \qquad (1) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} R_{|i-j|} = \frac{1}{n}\left[R_0 + 2\sum_{k=1}^{n-1}\left(1 - \frac{k}{n}\right) R_k\right], \end{aligned}$$

where the last step follows after some algebra.

This is a very important equation, which relates the variance of the sample mean to the covariances of the process. If the Y_i's are *i.i.d.*, then $\mathrm{Var}(\bar{Y}_n) = \frac{\mathrm{Var}(Y_1)}{n} = \frac{R_0}{n}$, which is not a surprise.

In the non-i.i.d. case, the other terms in the summation reflect the effects of the covariances. In most queueing applications, the covariances are positive and die off slowly as k becomes large. In these cases, $\mathrm{Var}(\bar{Y}_n) \gg \frac{\mathrm{Var}(Y_1)}{n}$, and the ratio $\mathrm{Var}(\bar{Y}_n)/(\mathrm{Var}(Y_1)/n)$ is regarded as the number of Y_i's that you need in order to get one "independent" observation.

The main point is that you have to have to be really careful in the presence of correlation.

Thus, for this and other reasons, it's difficult to apply "classical" statistical techniques to the analysis of simulation output.

Our purpose: Give methods to perform statistical analysis of output from discrete-event computer simulations.

Why all the fuss?

- Beware — improper statistical analysis can invalidate all results
- Tremendous applications if you can get it right
- Lots of cool research problems out there

Types of Simulations

To facilitate the presentation, we identify two types of simulations with respect to output analysis: Finite-Horizon (Terminating) and Steady-State simulations.

Finite-Horizon Simulations: The termination of a finite-horizon simulation takes place at a specific time or is caused by the occurrence of a specific event. Examples are:

- Mass transit system between during rush hour.
- Distribution system over one month.
- Production system until a set of machines breaks down.
- Start-up phase of any system — stationary or nonstationary

Steady-state simulations: The purpose of a steady-state simulation is the study of the long-run behavior of a system. A performance measure is called a *steady-state parameter* if it is a characteristic of the equilibrium distribution of an output stochastic process. Examples are:

- Continuously operating communication system where the objective is the computation of the mean delay of a packet in the long run.
- Distribution system over a long period of time.
- Many Markov chains.

(Some people don't regard s-s simulation as interesting as finite-horizon — because in steady-state, you're always dead.)

Techniques to analyze output from terminating simulations are based on the method of indep. replications (discussed in §2).

Additional problems arise for steady-state simulations. . .

Must now worry about the problem of starting the simulation — how should it be initialized at time zero (§3), and

How long must it be run before data representative of steady state can be collected?

§4 deals with point and confidence interval estimation for steady-state simulation performance parameters.

§5 concerns the problem of comparing a number of competing systems.

Outline

Simulate some system of interest over a *finite* time horizon.

For now, assume we obtain *discrete* simulation output $Y_1, Y_2, \ldots, Y_m$, where the number of observations m can be a constant or a random variable.

Example: The experimenter can specify the number m of customer waiting times $Y_1, Y_2, \ldots, Y_m$ to be taken from a queueing simulation. Or m could denote the random number of customers observed during a specified time period $[0, T]$.

Alternatively, we might observe *continuous* simulation output $\{Y(t)|0 \leq t \leq T\}$ over a specified interval $[0, T]$.

Example: Estimate the time-averaged number of customers waiting in a queue during $[0, T]$. Then the quantity $Y(t)$ would be the number of customers in the queue at time t.

Easiest Goal: Estimate the expected value of the sample mean of the observations,

$$\theta \equiv \mathrm{E}[\bar{Y}_m],$$

where the sample mean in the discrete case is

$$\bar{Y}_m \equiv \frac{1}{m}\sum_{i=1}^{m} Y_i$$

(with a similar expression for the continuous case).

Example: We might be interested in estimating the expected average waiting time of all customers at a shopping center during the period 10 a.m. to 2 p.m.

Although $\bar{Y}_m$ is an unbiased estimator for θ, a proper statistical analysis requires that we also provide an estimate of $\mathrm{Var}(\bar{Y}_m)$.

Since the Y_i's are not necessarily i.i.d., then Equation (1) in §1 showed that we could have $\mathrm{Var}(\bar{Y}_m) \neq \mathrm{Var}(Y_i)/m$.

Similarly, the familiar sample variance,

$$S_Y^2 \equiv \frac{1}{m-1} \sum_{i=1}^{m} (Y_i - \bar{Y}_m)^2,$$

may be a poor estimator for $\mathrm{Var}(Y_i)$, let alone for $m\mathrm{Var}(\bar{Y}_m)$.

In fact, if the Y_i's are positively correlated, then it may be the case that $\mathrm{E}[S_Y^2/m] \ll \mathrm{Var}(\bar{Y}_m)$.

Here's why.... To keep things easy, let's suppose that all of the Y_i's have the same distribution with mean $\theta = \mathrm{E}[Y_i]$ (though they may be *positively correlated*). Then

$$\begin{aligned}
\mathrm{E}[S_Y^2] &= \frac{1}{m-1}\,\mathrm{E}\left[\sum_{i-1}^{m}(Y_i - \bar{Y}_m)^2\right] \\
&= \frac{1}{m-1}\,\mathrm{E}\left[\sum_{i=1}^{m} Y_i^2 - m\bar{Y}_m^2\right] \\
&= \frac{m}{m-1}\left(\mathrm{E}[Y_1^2] - \mathrm{E}[\bar{Y}_m^2]\right) \\
&= \frac{m}{m-1}\left(\{\mathrm{Var}(Y_1) + (\mathrm{E}[Y_1])^2\} - \{\mathrm{Var}(\bar{Y}_m) + (\mathrm{E}[\bar{Y}_m])^2\}\right) \\
&= \frac{m}{m-1}[\mathrm{Var}(Y_1) - \mathrm{Var}(\bar{Y}_m)] \quad \text{(since } \theta = \mathrm{E}[Y_1] = \mathrm{E}[\bar{Y}_m]\text{)}.
\end{aligned}$$

If we assume that the R_i's > 0, then Equation (1) implies

$$\begin{aligned}
\mathrm{E}[S_Y^2] &= \frac{m}{m-1}\left\{R_0 - \frac{1}{m}\left[R_0 + 2\sum_{i=1}^{m-1}\left(1-\frac{i}{m}\right)R_i\right]\right\} \\
&= R_0 - \frac{2}{m-1}\sum_{i=1}^{m-1}\left(1-\frac{i}{m}\right)R_i \\
&< R_0 \\
&\ll R_0 + 2\sum_{i=1}^{m-1}\left(1-\frac{i}{m}\right)R_i.
\end{aligned}$$

Collecting these results shows that

$$\mathrm{E}[S_Y^2] \;<\; \mathrm{Var}(Y_i) \;\ll\; m\mathrm{Var}(\bar{Y}_m). \quad \square$$

Thus, one should *not* use S_Y^2/m to estimate $\mathrm{Var}(\bar{Y}_m)$.

So what happens if you dare to use it?

Here's a typical $100(1-\alpha)$% confidence interval for the mean μ of i.i.d. normal observations with unknown variance:

$$\mu \in \bar{Y}_m \pm t_{\alpha/2,m-1}\sqrt{S_Y^2/m},$$

where $t_{\alpha/2,m-1}$ is a t-distribution quantile, and $1-\alpha$ is the desired coverage level.

Since $\mathrm{E}[S_Y^2/m] \ll \mathrm{Var}(\bar{Y}_m)$, the confidence interval will have true coverage $\ll 1-\alpha$! Oops!

The way around the problem is via the method of *independent replications* (IR).

IR estimates $\mathrm{Var}(\bar{Y}_m)$ by conducting b independent simulation runs (replications) of the system under study, where each replication consists of m observations.

It is easy to make the replications independent — just re-initialize each replication with a different pseudo-random number seed.

Notation and Stuff.

Denote the sample mean from replication i by

$$Z_i \equiv \frac{1}{m}\sum_{j=1}^{m} Y_{i,j},$$

where $Y_{i,j}$ is observation j from replication i, for $i = 1, 2, \ldots, b$ and $j = 1, 2, \ldots, m$.

If each run is started under the same operating conditions (e.g., all queues empty and idle), then the replication sample means $Z_1, Z_2, \ldots, Z_b$ are *i.i.d.* random variables.

Then the obvious point estimator for $\mathrm{Var}(\bar{Y}_m) = \mathrm{Var}(Z_i)$ is

$$S_Z^2 \equiv \frac{1}{b-1}\sum_{i=1}^{b}(Z_i - \bar{Z}_b)^2,$$

where the grand mean is defined as $\bar{Z}_b \equiv \frac{1}{b}\sum_{i=1}^{b} Z_i$.

Note that the forms of S_Z^2 and S_Y^2/m resemble each other. But since the replicate sample means are i.i.d., S_Z^2 is usually much less biased for $\mathrm{Var}(\bar{Y}_m)$ than is S_Y^2/m.

In light of the above, we see that S_Z^2/b is a reasonable estimator for $\mathrm{Var}(\bar{Z}_b)$.

If the number of observations per replication, m, is large enough, a central limit theorem tells us that the replicate sample means $Z_1, Z_2, \ldots, Z_b$ are approximately i.i.d. $\mathrm{Nor}(\theta, \mathrm{Var}(Z_i))$, and

$$S_Z^2 \approx \frac{\mathrm{Var}(Z_i)\chi^2(b-1)}{b-1}.$$

Then after the usual baby stats manipulations, we have the approximate IR $100(1-\alpha)\%$ two-sided confidence interval (CI) for θ,

$$\theta \in \bar{Z}_b \pm t_{\alpha/2,b-1}\sqrt{S_Z^2/b}\,. \tag{2}$$

Example: Suppose we want to estimate the expected average waiting time for the first 5000 customers in a certain queueing system. We will make five independent replications of the system, with each run initialized empty and idle and consisting of 5000 waiting times. The resulting replicate means are:

i	1	2	3	4	5
Z_i	3.2	4.3	5.1	4.2	4.6

Then $\bar{Z}_5 = 4.28$ and $S_Z^2 = 0.487$. For level $\alpha = 0.05$, we have $t_{0.025,4} = 2.78$, and (2) gives $[3.41, 5.15]$ as a 95% CI for the expected average waiting time for the first 5000 customers.

Independent replications can be used to calculate variance estimates for statistics other than sample means.

Then the method can be used to get CI's for quantities other than $\mathrm{E}[\bar{Y}_m]$, e.g., quantiles.

See any of the standard simulation texts for additional uses of independent replications.

Research Issue: Sequential procedures that deliver a CI of fixed size.

Outline

Before a simulation can be run, one must provide initial values for all of the simulation's state variables.

Since the experimenter may not know what initial values are appropriate for the state variables, these values might be chosen somewhat arbitrarily.

For instance, we might decide that it is "most convenient" to initialize a queue as empty and idle.

Such a choice of initial conditions can have a significant but unrecognized impact on the simulation run's outcome.

Thus, the *initialization bias* problem can lead to errors, particularly in steady-state output analysis.

Examples of problems concerning simulation initialization.

- Visual detection of initialization effects is sometimes difficult — especially in the case of stochastic processes having high intrinsic variance such as queueing systems.
- How should the simulation be initialized? Suppose that a machine shop closes at a certain time each day, even if there are jobs waiting to be served. One must therefore be careful to start each day with a demand that depends on the number of jobs remaining from the previous day.
- Initialization bias can lead to point estimators for steady-state parameters having high mean squared error, as well as CI's having poor coverage.

Since initialization bias raises important concerns, how do we detect and deal with it? We first list methods to detect it.

Attempt to detect the bias visually by scanning a realization of the simulated process. This might not be easy, since visual analysis can miss bias. Further, a visual scan can be tedious. To make the visual analysis more efficient, one might transform the data (e.g., take logs or square roots), smooth it, average it across several indep. replications, or construct CUSUM plots.

Conduct statistical tests for initialization bias. Various procedures check to see if mean or variance of process changes over time: ASAP3 (Wilson et al.), change point detection from statistical literature, etc.

If initialization bias is detected, one may want to do something about it. Two simple methods for dealing with bias. . .

(a) *Truncate the output* by allowing the simulation to "warm up" before data are retained for analysis.

Experimenter hopes that the remaining data are representative of the steady-state system.

Output truncation is probably the most popular method for dealing with initialization bias; and all of the major simulation languages have built-in truncation functions.

But how can one find a good truncation point? If the output is truncated “too early,” significant bias might still exist in the remaining data. If it is truncated “too late,” then good observations might be wasted.

Unfortunately, all simple rules to determine truncation points do not perform well in general.

A reasonable practice is to average observations across several replications, and then visually choose a truncation point based on the averaged run; see Welch (1983) for a nice visual/graphical approach.

This is where the new, sophisticated sequential change-point detection algorithms come into play.

(b) *Make a very long run* to overwhelm the effects of initialization bias.

This method of bias control is conceptually simple to carry out and may yield point estimators having lower mean squared errors than the analogous estimators from truncated data (see, e.g., Fishman 1978).

However, a problem with this approach is that it can be wasteful with observations; for some systems, an excessive run length might be required before the initialization effects are rendered negligible.

Outline

Now assume that we have on hand stationary (steady-state) simulation output, $Y_1, Y_2, \ldots, Y_n$.

Our goal: Estimate some parameter of interest, e.g., the mean customer waiting time or the expected profit produced by a certain factory configuration.

In particular, suppose the mean of this output is the unknown quantity μ. We'll use the sample mean $\bar{Y}_n$ to estimate μ.

As in the case of terminating simulations (where we used the method of IR), we must accompany the value of any point estimator with a measure of its variance.

Instead of $\mathrm{Var}(\bar{Y}_n)$, we can estimate the *variance parameter*,

$$\begin{aligned}
\sigma^2 &\equiv \lim_{n\to\infty} n\mathrm{Var}(\bar{Y}_n) \\
&= \lim_{n\to\infty} \left[R_0 + 2\sum_{k=1}^{n-1} \left(1 - \frac{k}{n}\right) R_k \right] \quad \text{by (1)} \\
&= \sum_{k=-\infty}^{\infty} R_k \quad \text{(if the } R_k\text{'s decrease quickly as } k \to \infty\text{)}.
\end{aligned}$$

Thus, σ^2 is simply the sum of all covariances!

σ^2 pops up all over the place: simulation output analysis, Brownian motions, financial engineering applications, etc.

Example: MA(1) process, $Y_{i+1} = \theta\epsilon_i + \epsilon_{i+1}$, where the ϵ_i's are i.i.d. $\text{Nor}(0, 1)$. Then $R_0 = 1 + \theta^2$, $R_{\pm 1} = \theta$, and $R_k = 0$, o'w.

By (1), $\text{Var}(\bar{Y}_n) = \frac{1}{n}\left[R_0 + 2\sum_{k=1}^{n-1}\left(1 - \frac{k}{n}\right)R_k\right] = \frac{(1+\theta)^2}{n} - \frac{2\theta}{n^2}$, and so $\sigma^2 = \lim_{n\to\infty} n\text{Var}(\bar{Y}_n) = (1+\theta)^2$. □

Example: AR(1) process, $Y_{i+1} = \phi Y_i + \epsilon_{i+1}$, where the ϵ_i's are i.i.d. $\text{Nor}(0, 1-\phi^2)$, $-1 < \phi < 1$, and $Y_0 \sim \text{Nor}(0, 1)$.

For the AR(1), $R_k = \phi^{|k|}\ \forall k$. Then after lots of algebra, it turns out that $\sigma^2 = (1+\phi)/(1-\phi)$; so σ^2 explodes as $\phi \to 1$. □

Many methods for estimating σ^2 and for conducting steady-state output analysis in general: batch means, IR, standardized time series, spectral analysis, regeneration, ARMA time series modeling, etc.

The method of **batch means** is often used to estimate σ^2 and to calculate confidence intervals for μ.

Idea: Divide one long simulation run into a number of contiguous *batches*, and then appeal to a central limit theorem to assume that the resulting batch sample means are approximately i.i.d. normal.

In particular, suppose that we partition $Y_1, Y_2, \ldots, Y_n$ into b nonoverlapping, contiguous batches, each consisting of m observations (assume that $n = bm$).

$$\underbrace{Y_1, \ldots, Y_m}_{\text{batch 1}}, \underbrace{Y_{m+1}, \ldots, Y_{2m}}_{\text{batch 2}}, \ldots, \underbrace{Y_{(b-1)m+1}, \ldots, Y_{bm}}_{\text{batch } b}$$

The ith batch mean is the sample mean of the m observations from batch $i = 1, 2, \ldots, b$,

$$\bar{Y}_{i,m} \equiv \frac{1}{m} \sum_{j=1}^{m} Y_{(i-1)m+j}.$$

The batch means are correlated for small m, but for large m,

$$\bar{Y}_{1,m}, \ldots, \bar{Y}_{b,m} \approx \text{ i.i.d. } \text{Nor}(\mu, \text{Var}(\bar{Y}_{i,m})) \approx \text{Nor}(\mu, \sigma^2/m).$$

Similar to independent replications (actually, more similar than I'm indicating, because I'm using different notation), we define the batch means estimator for $\bar{Y}_{i,m}$ as

$$\hat{V}_B \equiv \frac{m}{b-1} \sum_{i=1}^{b} (\bar{Y}_{i,m} - \bar{Y}_n)^2 \approx \frac{\sigma^2 \chi^2(b-1)}{b-1}.$$

How good is $\hat{V}_B \approx \frac{\sigma^2 \chi^2(b-1)}{b-1}$ as an estimator of σ^2?

First of all, we have $\mathrm{E}[\hat{V}_B] \approx \frac{\sigma^2}{b-1}\mathrm{E}[\chi^2(b-1)] = \sigma^2$, so $\hat{V}_B$ is asymptotically unbiased for σ^2.

More-precise result: It can be shown that

$$\mathrm{E}[\hat{V}_B] \;=\; \sigma^2 + \frac{\gamma(b+1)}{mb} + o(1/m),$$

where $\gamma \equiv -2\sum_{k=1}^{\infty} kR_k$ and $o(1/m)$ is a function that goes to 0 faster than rate $1/m$ as m gets big.

We also have $\mathrm{Var}(\hat{V}_B) \approx \frac{\sigma^4}{(b-1)^2}\mathrm{Var}(\chi^2(b-1)) = \frac{2\sigma^4}{b-1}$.

These facts immediately imply that

$$\mathrm{MSE}(\hat{V}_B) \;=\; \mathrm{Bias}^2 + \mathrm{Var} \;\approx\; \frac{\gamma^2}{m^2} + \frac{2\sigma^4}{b}.$$

Now take $m = cn^\delta$, where $c > 0$ and $0 < \delta < 1$. It's easy to see that the choice $\delta = 1/3$ yields the fastest convergence to 0 for $\mathrm{MSE}(\hat{V}_B)$.

Minimizing the above expression for MSE with respect to c, we get the "optimal" batch size $m^\star \equiv (\gamma^2 n/\sigma^4)^{1/3}$, and then the resulting "optimal" MSE,

$$\mathrm{MSE}^\star(\hat{V}_B) \;\equiv\; 3(\gamma\sigma^4/n)^{2/3}.$$

Now the batch means confidence interval for μ.

Since the batch means $\bar{Y}_{1,m}, \ldots, \bar{Y}_{b,m} \approx$ i.i.d. $\mathrm{Nor}(\mu, \sigma^2/m)$ for large m, we get the following approximate $100(1-\alpha)\%$ CI for μ:

$$\mu \in \bar{Y}_n \pm t_{\alpha/2,b-1}\sqrt{\hat{V}_B/n}.$$

This equation is similar to (2) (even though I'm using different notation). The difference is that batch means divides one long run into a number of batches, whereas IR uses a number of independent shorter runs.

Consider the old IR example from §2 with the understanding that the Z_i's must now be regarded as batch means (instead of replicate means); then the same numbers carry through the example if you note that $S_Z^2/b = \hat{V}_B/n$.

Some Properties of the Batch Means CI.

Define the *half-length* as $H \equiv t_{\alpha/2,b-1}\sqrt{\hat{V}_B/n}$. Then as $m \to \infty$, it can be shown that

$$\begin{aligned}
\sqrt{mb}\,H &\approx \sigma\, t_{\alpha/2,b-1}\frac{\chi(b-1)}{\sqrt{b-1}} \\
\sqrt{mb}\,\mathrm{E}[H] &\to \sigma\, t_{\alpha/2,b-1}\sqrt{\frac{2}{b-1}}\frac{\Gamma(\frac{b}{2})}{\Gamma(\frac{b-1}{2})} \\
mb\,\mathrm{Var}(H) &\to \sigma^2\, t^2_{\alpha/2,b-1}\left\{1-\frac{2}{b-1}\left[\frac{\Gamma(\frac{b}{2})}{\Gamma(\frac{b-1}{2})}\right]^2\right\},
\end{aligned}$$

where $\chi(b-1)$ is the chi-distribution.

Fact: $\mathrm{E}[H]$ decreases in b, though it smooths out around $b = 30$. Schmeiser recommends taking $b \approx 30$ and concentrating on increasing the batch size m as much as possible.

The technique of batch means is intuitively appealing and easy to understand.

But problems can come up if the Y_j's are not stationary (e.g., if significant initialization bias is present), if the batch means are not normal, or if the batch means are not independent.

If any of these assumption violations exist, poor confidence interval coverage may result — unbeknownst to the analyst.

To ameliorate the initialization bias problem, the user can truncate some of the data or make a long run as discussed in §3.

In addition, the lack of independence or normality of the batch means can be countered by increasing the batch size m.

Of the difficulties encountered when using batch means, the possibility of correlation among the batch means might be the most troublesome.

This problem is explicitly avoided by the method of **independent replications**, described in the context of terminating simulations in §2. In fact, the replicate means are independent by their construction.

Unfortunately, since *each* of the b reps has to be started properly, initialization bias presents more trouble when using IR than when using batch means.

Recommendation: Because of initialization bias in each of the replications, *use batch means over independent reps*. (Alexopoulos and Goldsman, “To Batch or not to Batch?”)

Suppose that we use the following **overlapping batches**

$$
\begin{array}{ccccccccc}
Y_1 & Y_2 & Y_3 & Y_4 & \cdots & Y_m & & & \\
 & Y_2 & Y_3 & Y_4 & Y_5 & \cdots & Y_m & Y_{m+1} & \\
 & & Y_3 & Y_4 & Y_5 & Y_6 & \cdots & Y_m & Y_{m+1} & Y_{m+2} \\
 & & & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots
\end{array}
$$

with overlapping batch means

$$
\bar{Y}^{\mathrm{o}}_{i,m} = \frac{1}{m} \sum_{j=i}^{i+m-1} Y_j \quad i = 1, \ldots, n-m+1.
$$

Turns out: Even though the $\bar{Y}^{\mathrm{o}}_{i,m}$'s are highly correlated, we can deal with it!

The OBM estimator for μ is $\bar{Y}_n$ (no surprise), and the OBM estimator for $\mathrm{Var}(\bar{Y}_n)$ is

$$\hat{V}_O = \frac{m}{n-m+1} \sum_{i=1}^{n-m+1} (\bar{Y}_{i,m}^{\mathrm{o}} - \bar{Y}_n)^2.$$

Facts: As n and m get large,

$$\frac{\mathrm{E}(\hat{V}_O)}{\mathrm{E}(\hat{V}_B)} \to 1 \quad \text{and} \quad \frac{\mathrm{Var}(\hat{V}_O)}{\mathrm{Var}(\hat{V}_B)} \to \frac{2}{3}$$

So OBM has the same bias as, but lower variance than regular BM — great! (Meketon and Schmeiser 1984, "Overlapping Batch Means: Something for Nothing?")

Note that no attempt was made to make the overlapping batch means independent.

This is related to the fact that $\hat{V}_O$ is almost identical to what is known as Bartlett's spectral estimator for σ^2.

Fact: For large m and $b = n/m$, it can be shown that $\hat{V}_O \approx \sigma^2 \chi^2(d)/d$, where $d = \frac{3}{2}(b-1)$. So you get 50% more d.f. than regular batch means.

Resulting CI: $\mu \in \bar{Y}_n \pm t_{\alpha/2,d}\sqrt{\hat{V}_O/n}$

Recommendation: For large m and n/m use OBM instead of BM!

Several other methods for obtaining variance estimators for the sample mean and CI's for the steady-state process mean μ.

Spectral Estimation. This method estimates $\mathrm{Var}(\bar{Y}_n)$ (as well as the analogous CI's for μ) in a manner completely different from that of batch means.

This approach operates in the so-called *frequency domain*, whereas batch means uses the *time domain*.

Spectral estimation sometimes takes a little effort, but it works well enough to suggest that the reader consult the relevant references, e.g., Lada and Wilson's work on WASSP.

Regeneration. Many simulations can be broken into i.i.d. blocks that probabilistically "start over" at certain *regeneration* points.

Example: An M/M/1 queue's waiting time process, where the i.i.d. blocks are defined by groups of customers whose endpoints have zero waiting times.

Regeneration uses the i.i.d. structure and, under certain conditions, gives great estimators for $\mathrm{Var}(\bar{Y}_n)$ and CI's for μ.

The method effectively eliminates any initialization problems.

On the other hand, it may be difficult to define natural regeneration points, and *extremely* long simulation runs are often needed to obtain a reasonable number of i.i.d. blocks.

Standardized Time Series. One often uses the central limit theorem to standardize i.i.d. random variables into an (asymptotically) normal random variable.

Schruben and various colleagues generalize this idea in many ways by using a *process* central limit theorem to standardize a stationary simulation process into a *Brownian bridge* process.

Properties of Brownian bridges are then used to calculate a number of good estimators for $\mathrm{Var}(\bar{Y}_n)$ and CI's for μ.

This method is easy to apply and has some asymptotic advantages over batch means.

Research Issue: Combine various strategies together to obtain even-better variance estimators.

Outline

One of the most important uses of simulation output analysis regards the comparison of competing systems or alternative system configurations.

Example: Evaluate two different “re-start” strategies that an airline can evoke following a major traffic disruption such as a snowstorm in the Northeast — which policy minimizes a certain cost function associated with the re-start?

Simulation is uniquely equipped to help the experimenter conduct this type of comparison analysis.

Many techniques: (i) classical statistical CI’s, (ii) common random numbers, (iii) antithetic variates, (iv) and ranking and selection procedures.

With our airline example in mind, let $Z_{i,j}$ be the cost from the jth simulation replication of strategy i, $i = 1, 2$, $j = 1, 2, \ldots, b_i$.

Assume that $Z_{i,1}, Z_{i,2}, \ldots, Z_{i,b_i}$ are i.i.d. normal with unknown mean μ_i and unknown variance, $i = 1, 2$. Justification?...

(a) Get independent data by controlling the random numbers between replications.

(b) Get identically distributed costs between reps by performing the reps under identical conditions.

(c) Get approximately normal data by adding up (or averaging) many sub-costs to get overall costs for both strategies.

Goal: Obtain a $100(1-\alpha)\%$ CI for the difference $\mu_1 - \mu_2$.

Suppose that the $Z_{1,j}$'s are independent of the $Z_{2,j}$'s and define

$$\bar{Z}_{i,b_i} \equiv \frac{1}{b_i}\sum_{j=1}^{b_i} Z_{i,j}, \quad i = 1, 2,$$

and

$$S_i^2 \equiv \frac{1}{b_i - 1}\sum_{j=1}^{b_i}(Z_{i,j} - \bar{Z}_{i,b_i})^2, \quad i = 1, 2.$$

An approximate $100(1-\alpha)\%$ CI is

$$\mu_1 - \mu_2 \in \bar{Z}_{1,b_1} - \bar{Z}_{2,b_2} \pm t_{\alpha/2,\nu}\sqrt{\frac{S_1^2}{b_1} + \frac{S_2^2}{b_2}}$$

where the (approx.) d.f. ν is given in any statistics text.

Suppose (as in airline example) that small cost is good. Then if the interval lies entirely to the left [right] of zero, then system 1 [2] is better; if the interval contains zero, then the two systems must be regarded, in a statistical sense, as about the same.

An alternative classical strategy: Use a CI that is analogous to a paired-t test.

Here take b replications from *both* strategies and set the difference $D_j \equiv Z_{1,j} - Z_{2,j}$ for $j = 1, 2, \ldots, b$.

Calculate the sample mean and variance of the differences:

$$\bar{D}_b \equiv \frac{1}{b}\sum_{j=1}^{b} D_j \text{ and } S_D^2 \equiv \frac{1}{b-1}\sum_{j=1}^{b}(D_j - \bar{D}_b)^2.$$

The resulting $100(1-\alpha)$% CI is

$$\mu_1 - \mu_2 \in \bar{D}_b \pm t_{\alpha/2,b-1}\sqrt{S_D^2/b}.$$

These paired-t intervals are very efficient if $\mathrm{Corr}(Z_{1,j}, Z_{2,j}) > 0$, $j = 1, 2, \ldots, b$.

Idea behind the above trick: Use *common random numbers*, i.e., use the same pseudo-random numbers in exactly the same ways for corresponding runs of each of the competing systems.

Example: Use the same customer arrival times when simulating different proposed configurations of a job shop.

By subjecting the alternative systems to identical experimental conditions, we hope to make it easy to distinguish which systems are best even though the respective estimators are subject to sampling error.

Consider the case in which we compare two queueing systems, A and B, on the basis of their expected customer transit times, θ_A and θ_B — the smaller θ-value corresponds to the better system.

Suppose we have estimators $\hat{\theta}_A$ and $\hat{\theta}_B$ for θ_A and θ_B, resp.

We'll declare A as the better system if $\hat{\theta}_A < \hat{\theta}_B$. If $\hat{\theta}_A$ and $\hat{\theta}_B$ are simulated independently, then the variance of their difference,

$$\mathrm{Var}(\hat{\theta}_A - \hat{\theta}_B) \;=\; \mathrm{Var}(\hat{\theta}_A) + \mathrm{Var}(\hat{\theta}_B),$$

could be very large; then our declaration might lack conviction.

If we could reduce $\mathrm{Var}(\hat{\theta}_A - \hat{\theta}_B)$, then we could be much more confident about our declaration.

CRN sometimes induces a high positive correlation between the point estimators $\hat{\theta}_A$ and $\hat{\theta}_B$.

Then we have

$$\begin{aligned} \mathrm{Var}(\hat{\theta}_A - \hat{\theta}_B) &= \mathrm{Var}(\hat{\theta}_A) + \mathrm{Var}(\hat{\theta}_B) - 2\mathrm{Cov}(\hat{\theta}_A, \hat{\theta}_B) \\ &< \mathrm{Var}(\hat{\theta}_A) + \mathrm{Var}(\hat{\theta}_B), \end{aligned}$$

and we obtain a savings in variance.

Antithetic random numbers. Alternatively, if we can induce *negative* correlation between two unbiased estimators, $\hat{\theta}_1$ and $\hat{\theta}_2$, for some parameter θ, then the unbiased estimator $(\hat{\theta}_1 + \hat{\theta}_2)/2$ might have low variance.

Most simulation texts give advice on how to run the simulations of the competing systems so as to induce positive or negative correlation between them.

Consensus: if conducted properly, CRN and ARN can lead to tremendous variance reductions.

Ranking, selection, and multiple comparisons methods form another class of statistical techniques used to compare alternative systems.

Here, the experimenter is interested in selecting the best of a number (≥ 2) of competing processes.

Specify the desired probability of correctly selecting the best process, especially if the best process is significantly better than its competitors.

These methods are simple to use, fairly general, and intuitively appealing. (see Bechhofer, Santner, and Goldsman 1995).

Outline

- Use of more-sophisticated variance estimators
- Automated sequential run-control procedures that control for initialization bias and deliver valid confidence intervals of specified length
- Change-point detection algorithms for initialization bias tests
- Incorporating combinations of variance reduction tools
- Multivariate confidence intervals
- Better ranking and selection techniques

If you like this stuff, here are some General References. . .

Alexopoulos, C. and A. F. Seila. 1998. Output data analysis, Chapter 7 in *Handbook of Simulation: Principles, Methodology, Advances, Applications and Practice*, J. Banks, Ed., John Wiley, New York.

Goldsman, D. and B. L. Nelson. 1998. Comparing systems via simulation, Chapter 7 in *Handbook of Simulation: Principles, Methodology, Advances, Applications and Practice*, J. Banks, Ed., John Wiley, New York.

Law, A. M. 2014. *Simulation Modeling and Analysis*, 5th Ed., McGraw-Hill, New York.