SIMULATION OUTPUT ANALYSIS

Dave Goldsman

School of ISyE
Georgia Tech
Atlanta, Georgia, USA

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

10/19/17

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.

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.

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 on Serial Correlation: Suppose that $Y_1, Y_2, \ldots, Y_n$ are identically distributed 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, recall the *covariance function*, $R_k \equiv \mathrm{Cov}(Y_1, Y_{1+k})$, $k = 0, 1, 2, \ldots$. Can show (next lesson) that

$$\mathrm{Var}(\bar{Y}_n) = \frac{1}{n}\left[R_0 + 2\sum_{k=1}^{n-1}\left(1 - \frac{k}{n}\right)R_k\right].$$

This is a problem, since the "classical" confidence interval for the mean μ requires i.i.d. observations and $\mathrm{Var}(\bar{Y}_n) = \mathrm{Var}(Y_i)/n$. Those extra covariances are trouble! ☹

The main point is that it's difficult to apply "classical" statistical techniques to the analysis of simulation output — in large part due to the presence of serial correlation.

This Module's Raison d'Être: 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:

- Mass transit system 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:

- 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.)

What's coming up? In §2, we'll go over some math results that we'll refer to occasionally.

§3 discusses techniques to analyze output from terminating simulations are based on the method of independent replications.

Additional problems arise for steady-state simulations...

§4 is concerned with the problem of starting the simulation — how should it be initialized at time zero, and how long must it be run before data representative of steady state can be collected?

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

Outline

We'll look at a few examples illustrating the fact that things turn out a little differently when you don't have i.i.d. observations.

Working Assumptions: For the remainder of this section, suppose that $Y_1, Y_2, \ldots, Y_n$ are *identically distributed* with mean μ, but *not independent*.

Such an assumption often applies in the context of steady-state simulation.

Let's get properties of the sample mean and variance — these guys perform very well in the i.i.d. case — but you have to be careful when using them in simulation output analysis.

Properties of the Sample Mean: First, $\mathrm{E}[\bar{Y}_n] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[Y_i] = \mu$, so the sample mean is still unbiased for μ.

Now recall the *covariance function*, $R_k \equiv \mathrm{Cov}(Y_1, Y_{1+k})$, $\forall k$. Then

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

How did we go from (1) to (2)?

Just add up the terms in the following matrix of covariances.

$$\begin{pmatrix} R_0 & R_1 & R_2 & \cdots & R_{n-3} & R_{n-2} & R_{n-1} \\ R_1 & R_0 & R_1 & \cdots & R_{n-4} & R_{n-3} & R_{n-2} \\ R_2 & R_1 & R_0 & \cdots & R_{n-5} & R_{n-4} & R_{n-3} \\ & & \vdots & \vdots & \vdots & & \\ R_{n-3} & R_{n-4} & R_{n-5} & \cdots & R_0 & R_1 & R_2 \\ R_{n-2} & R_{n-3} & R_{n-4} & \cdots & R_1 & R_0 & R_1 \\ R_{n-1} & R_{n-2} & R_{n-3} & \cdots & R_2 & R_1 & R_0 \end{pmatrix}$$

The result follows because there are n R_0 terms, $2(n-1)$ R_1 terms, $2(n-2)$ R_2 terms, $\ldots$, and $2(n-(n-1))$ R_{n-1} terms.

Equation (2) is important — it relates the variance of the sample mean to the covariances of the process. With this in mind, define

$$\sigma_n^2 \equiv n\mathrm{Var}(\bar{Y}_n) = R_0 + 2\sum_{k=1}^{n-1}\left(1 - \frac{k}{n}\right)R_k.$$

We also define the related *variance parameter*,

$$\sigma^2 \equiv \lim_{n\to\infty}\sigma_n^2 =^\star R_0 + 2\sum_{k=1}^{\infty} R_k = \sum_{k=-\infty}^{\infty} R_k,$$

where $=^\star$ holds if the R_k's decrease to 0 quickly as $k \to \infty$.

The variance parameter σ^2 is so pretty and turns up all over the place.

Notice that if the Y_i's are i.i.d., then for all $k \neq 0$, we have $R_k = 0$, in which case we have the familiar result $\sigma^2 = \sigma_n^2 = R_0 = \mathrm{Var}(Y_1)$.

But in the dependent case, $\sigma^2 = \sum_{k=-\infty}^{\infty} R_k$ adds in the effects of the covariances. In queueing applications, the covariances are positive and $\sigma^2 \doteq \sigma_n^2 \gg \mathrm{Var}(Y_1)$, which may be bigger than you'd think.

The ratio $\sigma_n^2/\mathrm{Var}(Y_1)$ is sort of the number of Y_i's needed to obtain the information equivalent to one "independent" observation.

Warning: $\sigma_n^2 \gg \mathrm{Var}(Y_1)$ causes the classical confidence interval (CI) for the mean μ to misbehave. Stay tuned.

Example: The first-order autoregressive process is defined by

$$Y_i \;=\; \phi Y_{i-1} + \varepsilon_i, \quad \text{for } i = 1, 2, \ldots,$$

where $-1 < \phi < 1$, $Y_0 \sim \text{Nor}(0, 1)$, and the ε_i's are i.i.d. $\text{Nor}(0, 1 - \phi^2)$ RV's that are independent of Y_0.

Recall that the Y_i's are all Nor(0,1) and $R_k = \phi^{|k|}$, $\forall k$.

After a little algebra, one can show that

$$\sigma^2 \;=\; \sum_{k=-\infty}^{\infty} \phi^{|k|} \;=\; \frac{1+\phi}{1-\phi}.$$

So, e.g., for $\phi = 0.9$, we have $\sigma^2 = 19$. $\square$

Properties of the Sample Variance:

$$S_Y^2 \equiv \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2.$$

If $Y_1, Y_2, \ldots, Y_n$ are i.i.d., then S_Y^2 is unbiased for $R_0 = \mathrm{Var}(Y_1)$. Moreover, S_Y^2 is also unbiased for $\sigma_n^2 = n\mathrm{Var}(\bar{Y}_n) = R_0$ and $\sigma^2 = \lim_{n\to\infty} \sigma_n^2 = R_0$. ☺

But if the Y_i's are dependent, then S_Y^2 may not be such a great estimator for $\mathrm{Var}(Y_1)$, σ_n^2, or σ^2. ☹

So let's again suppose that $Y_1, Y_2, \ldots, Y_n$ are identically distributed with mean μ and covariance function R_k, and see what happens....

Get the expected value of S_Y^2. The first steps should be familiar.

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

Now let's assume that the R_k's > 0. Equation (2) implies

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

Collecting these results shows that

$$\mathrm{E}[S_Y^2] \;<\; \mathrm{Var}(Y_1) \;\ll\; n\mathrm{Var}(\bar{Y}_n). \quad \square$$

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

So what happens if you dare to use it?

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

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

where $t_{\alpha/2,n-1}$ is a t-distribution quantile.

Since $\mathrm{E}[S_Y^2/n] \ll \mathrm{Var}(\bar{Y}_n)$, the CI will have true coverage $\ll 1-\alpha$!
Oops! This is why you have to be really careful with correlated data!

Outline

Goal: Simulate some system of interest over a *finite* time horizon and analyze the output.

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 random.

Example: Obtain the waiting times $Y_1, Y_2, \ldots, Y_{100}$ of the first 100 customers to show up at a store.

Or m could denote the random number of customers observed during a time period $[0, T]$, where T itself could be known or random.

For example, we might consider all customers from 10 a.m. to 2 p.m. (T is specified, but m is random). Or all customers from 10 a.m. until the cashier gets a call to pick up his kid at school (now T is random).

Alternatively, we might observe *continuous* simulation output $\{Y(t)|0 \leq t \leq T\}$ over an interval $[0, T]$, where T can again be known or random.

Example: $Y(t)$ could denote the number of customers in the queue over the time period 8 a.m. to 5 p.m. (T is constant). Or from 8 a.m. until the cashier has to pick his kid up (T is random).

Easiest Goal: Estimate the expected value of the sample mean of the observations, $\theta \equiv \mathrm{E}[\bar{Y}_m]$, where $\bar{Y}_m \equiv \frac{1}{m}\sum_{i=1}^{m} Y_i$.

Of course, the distribution of the Y_i's may change over the day. This is no problem — θ is simply the expected value of the average of all of these Y_i's.

In the continuous case, we'd be interested in estimating $\theta \equiv \mathrm{E}[\bar{Y}(T)]$, where $\bar{Y}(T) \equiv \frac{1}{T}\int_0^T Y(t)\,dt$.

Example: Estimate the average waiting time of the first 100 customers. Or estimate the time-averaged number of customers in line between 8 a.m. and 5 p.m.

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

But since the Y_i's are not necessarily i.i.d., then Equation (2) and the fact that the Y_i's may not be identically distributed show that we could have $\mathrm{Var}(\bar{Y}_m) \neq \mathrm{Var}(Y_1)/m$. ☹

And then the rest of the diatribe in the math interlude says that we'd better not use S_Y^2/m to estimate $\mathrm{Var}(\bar{Y}_m)$.

Whatever shall we do??

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

IR estimates $\mathrm{Var}(\bar{Y}_m)$ by conducting r 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 $j = 1, 2, \ldots, m$ from replication $i = 1, 2, \ldots, r$. E.g., $Y_{i,j}$ is customer j's waiting time from rep i.

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

Define the grand sample mean as $\bar{Z}_r \equiv \frac{1}{r}\sum_{i=1}^{r} Z_i$. The obvious point estimator for $\mathrm{Var}(\bar{Y}_m) = \mathrm{Var}(Z_i)$ is the sample variance of the Z_i's,

$$S_Z^2 \equiv \frac{1}{r-1}\sum_{i=1}^{r}(Z_i - \bar{Z}_r)^2.$$

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 $\mathrm{Var}(\bar{Y}_m) = \mathrm{Var}(Z_i)$ than is S_Y^2/m.

In light of the above, S_Z^2/r is a reasonable estimator for $\mathrm{Var}(\bar{Z}_r)$.

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

$$S_Z^2 \approx \frac{\text{Var}(Z_1)\chi^2(r-1)}{r-1}.$$

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

$$\theta \in \bar{Z}_r \pm t_{\alpha/2,r-1}\sqrt{S_Z^2/r}\,. \tag{3}$$

Example: Suppose we want to estimate the expected average waiting time for the first $m = 5000$ customers at the bank. We make $r = 5$ independent replications of the system, each 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 (3) gives the following 95% CI for the expected average waiting time for the first 5000 customers:

$$\theta \in 4.28 \pm (2.78)\sqrt{0.487/5} \;=\; [3.41, 5.15]. \quad \square$$

What if we'd like a smaller CI? We'll need to run more replications.

Let $H \equiv t_{\alpha/2,r-1}\sqrt{S_Z^2/r}$ denote the half-length of the current CI, and let $\epsilon \equiv t_{\alpha/2,r^\star-1}\sqrt{S_Z^2/r^\star}$ denote the *desired half-length* based on the same variance estimator S_Z^2 but more replications $r^\star > r$.

$$r^\star = \frac{t_{\alpha/2,r^\star-1}^2 S_Z^2}{\epsilon^2} < \frac{t_{\alpha/2,r-1}^2 S_Z^2}{\epsilon^2} = \left(H/\epsilon\right)^2 r.$$

So take $r^\star \leftarrow (H/\epsilon)^2 r$, run $r^\star - r$ more reps, and re-calculate the CI using all $r^\star$ reps. You'll *probably* get a CI with half-length close to ϵ.

If we want to reduce the length of the CI by a factor of 10, we'll need to increase the reps by a factor of 100.

What if we'd like a CI for other performance measures?

How about quantiles?

Definition: The *p-quantile* of a RV W with c.d.f. $F(w)$ is $\xi_p \equiv \min\{w|F(w) \geq p\}$. If W is continuous, then $F(\xi_p) = p$ and $\xi_p = F^{-1}(p)$.

Example: Suppose that $W \sim \text{Exp}(\lambda)$. Then (we've seen this a million times), $F(w) = 1 - e^{-\lambda w}$, so that $\xi_p = -\frac{1}{\lambda}\ell\text{n}(1-p)$. □

We can use the method of independent replications to obtain confidence intervals for quantiles.

Example: Let W_i denote the maximum waiting time that some customer experiences at the airport ticket counter between 8 a.m. and 5 p.m. during replication i of a simulation, $i = 1, 2, \ldots, r$.

Let's order the i.i.d. W_i's: $W_{(1)} \leq W_{(2)} \leq \cdots \leq W_{(r)}$.

Then the typical point estimator for the quantile ξ_p is

$$\widehat{\xi_p} \equiv W_{(\lfloor rp+0.5 \rfloor)},$$

where $\lfloor \cdot \rfloor$ is the floor (round-down) function.

Now that we have a point estimator for ξ_p, we'll get a CI for ξ_p....

A slightly conservative approximate nonparametric CI for ξ_p will turn out to be of the form $\xi_p \in [W_{(j)}, W_{(k)}]$ (Hahn and Meeker 1991).

To this end, first note that $P(W_i \leq \xi_p) = p$. So if we define A as the number of W_i's that are $\leq \xi_p$, then $A \sim \text{Bin}(r, p)$.

Second, the event that $\{j \leq A \leq k-1\}$ means that between j and $k-1$ of the W_i's are $\leq \xi_p$. This is equivalent to the event

$$\Big\{ \big(\xi_p \geq W_{(j)}\big) \text{ and } \big(\xi_p < W_{(k)}\big) \Big\}.$$

Putting this all together, we have. . . .

$$\begin{aligned} & P\big(W_{(j)} \le \xi_p < W_{(k)}\big) \\ &= P(j \le A \le k-1) \\ &= \sum_{\ell=j}^{k-1} \binom{r}{\ell} p^\ell (1-p)^{r-\ell} \\ &\doteq \Phi\left(\frac{k-0.5-rp}{\sqrt{rp(1-p)}}\right) - \Phi\left(\frac{j-0.5-rp}{\sqrt{rp(1-p)}}\right), \end{aligned}$$

where $\Phi(\cdot)$ is the Nor(0,1) c.d.f., the "0.5" terms are "continuity corrections", and the approximation requires rp and $r(1-p) \ge 5$.

In order to get an approximate $100(1-\alpha)\%$ CI of the form $\xi_p \in [W_{(j)}, W_{(k)}]$, we need to find j and k such that

$$\Phi\left(\frac{k-0.5-rp}{\sqrt{rp(1-p)}}\right) - \Phi\left(\frac{j-0.5-rp}{\sqrt{rp(1-p)}}\right) \;\geq\; 1-\alpha.$$

(The CI is a little conservative because of the binomial discretization.)

Suggested values for j and k arise by setting

$$\frac{j-0.5-rp}{\sqrt{rp(1-p)}} \;=\; -z_{\alpha/2} \quad \text{and} \quad \frac{k-0.5-rp}{\sqrt{rp(1-p)}} \;=\; z_{\alpha/2} \quad \Rightarrow$$

$$\begin{aligned} j &= \left\lfloor rp + 0.5 - z_{\alpha/2}\sqrt{rp(1-p)} \right\rfloor \\ k &= \left\lceil rp + 0.5 + z_{\alpha/2}\sqrt{rp(1-p)} \right\rceil. \end{aligned}$$

Sometimes — especially for "extreme" quantiles — you need a lot of replications to get reasonably small half-lengths.

Example: Let's suppose that we want a 95% CI for $\xi_{0.9}$ and we've made $r = 1000$ reps.

Point estimator for $\xi_{0.9}$ is $\widehat{\xi}_{0.9} = W_{(\lfloor 1000(0.9)+0.5 \rfloor)} = W_{(900)}$.

With the CI in mind,

$$j = \left\lfloor 900.5 - 1.96\sqrt{90} \right\rfloor = 881, \quad k = \left\lceil 900.5 + 1.96\sqrt{90} \right\rceil = 920,$$

so that the CI 95% is $[W_{(881)}, W_{(920)}]$. $\square$

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. You have to 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 independent replications, or construct CUSUM plots.

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

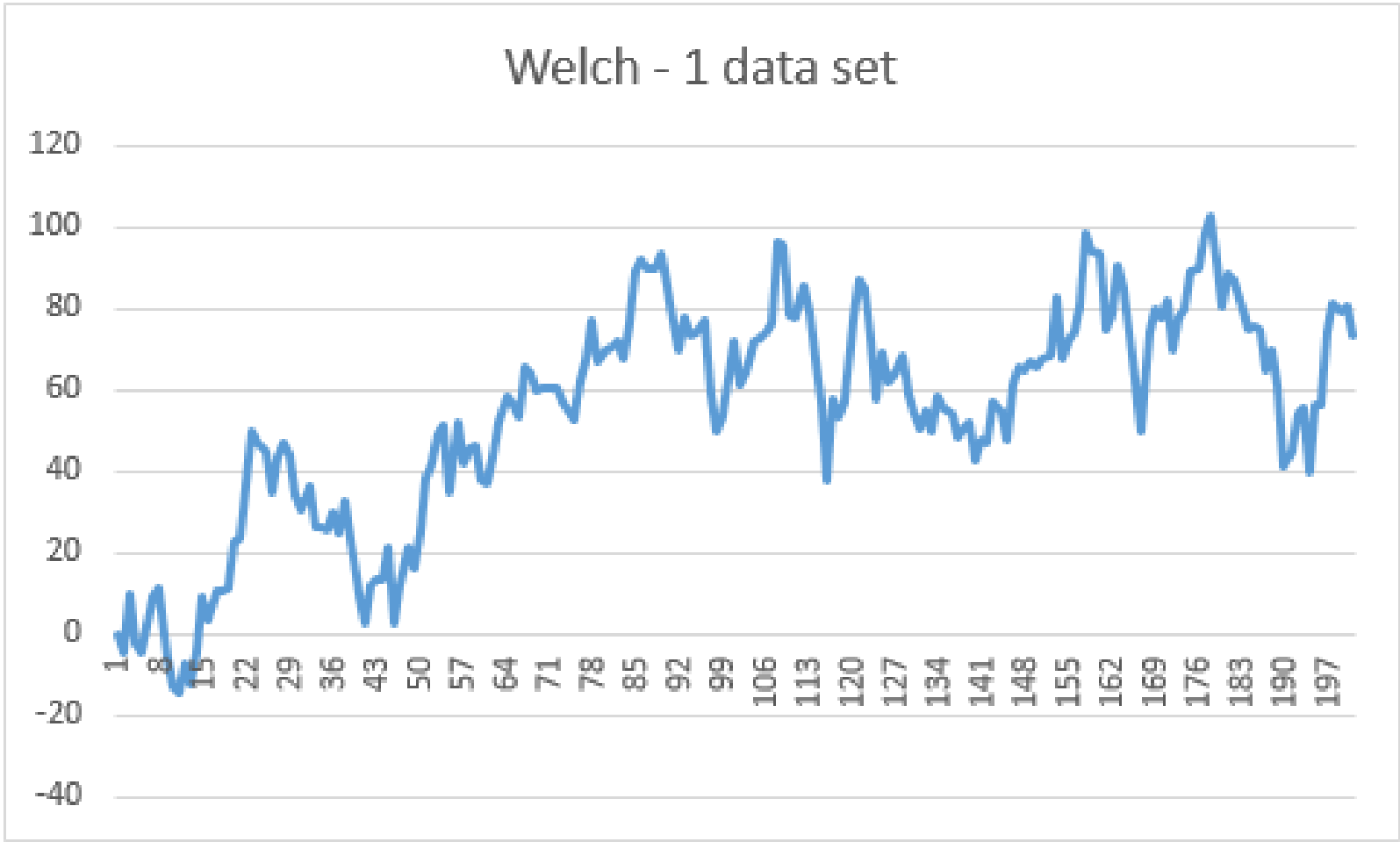
Output truncation is probably the most popular method for dealing with initialization bias; and all of the major simulation languages have built-in truncation functions.

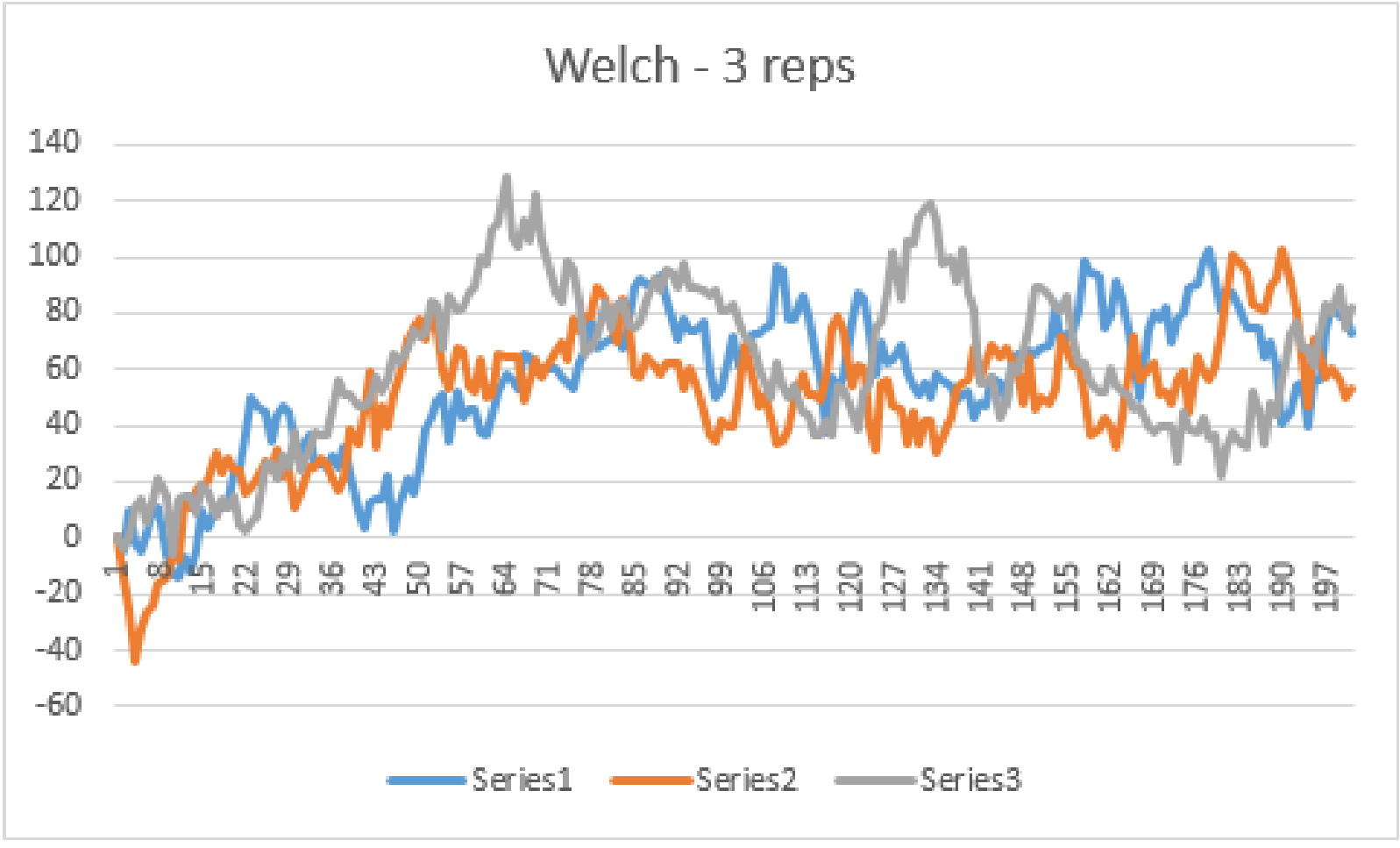
But how can one find a good truncation point? If the output is truncated "too early," significant bias might still exist in the remaining data. If truncated "too late," then good observations might be wasted.

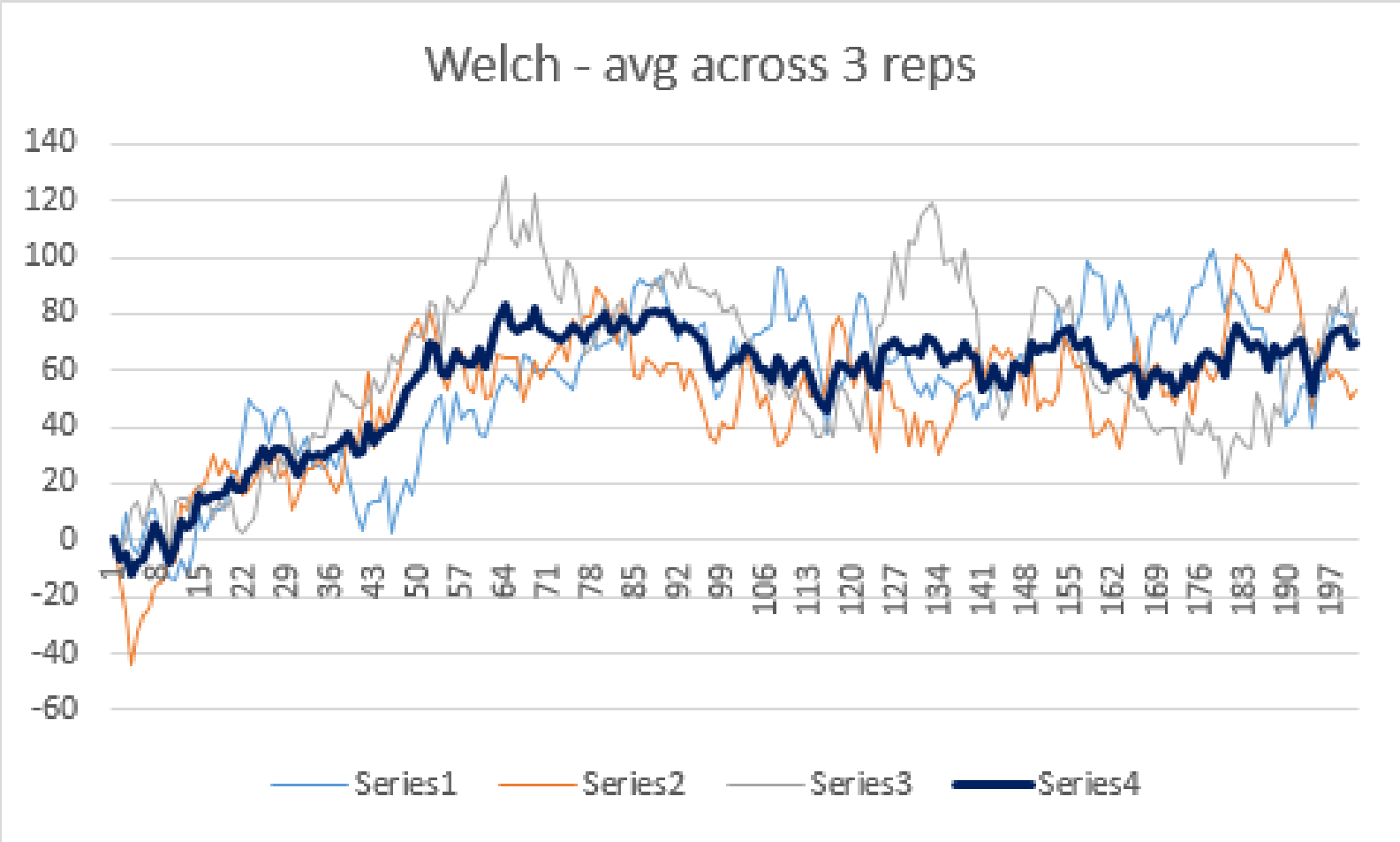
Unfortunately, all simple rules to determine truncation points do not perform well in general.

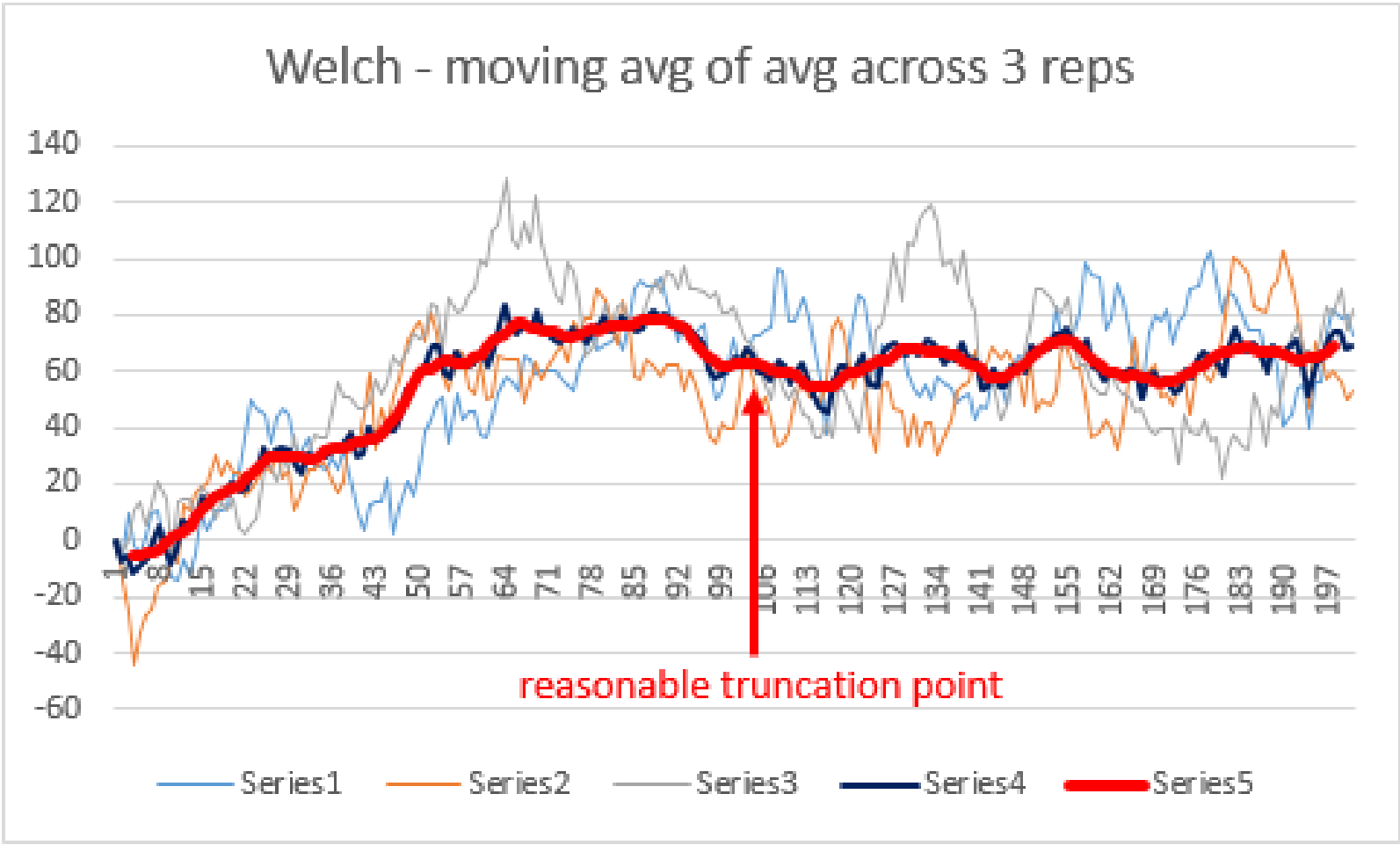
A reasonable practice is to average observations across several replications, and then visually choose a truncation point based on the averaged run; see Welch (1983) for a nice visual/graphical approach.

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


Welch - 1 data set
120
100
80
60
40
20
0
-20
-40
1
8
15
22
29
36
43
50
57
64
71
78
85
92
99
106
113
120
127
134
141
148
155
162
169
176
183
190
197


Welch - 3 reps
140
120
100
80
60
40
20
0
-20
-40
-60
1
8
15
22
29
36
43
50
57
64
71
78
85
92
99
106
113
120
127
134
141
148
155
162
169
176
183
190
197
Series1
Series2
Series3


Welch - avg across 3 reps
140
120
100
80
60
40
20
0
-20
-40
-60
1
8
15
22
29
36
43
50
57
64
71
78
85
92
99
106
113
120
127
134
141
148
155
162
169
176
183
190
197
Series1
Series2
Series3
Series4


Welch - moving avg of avg across 3 reps
140
120
100
80
60
40
20
0
-20
-40
-60
1
8
15
22
29
36
43
50
57
64
71
78
85
92
99
106
113
120
127
134
141
148
155
162
169
176
183
190
197
reasonable truncation point
Series1
Series2
Series3
Series4
Series5

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

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.

Bottom Line: Once initialization effects are dealt with, you can do *steady-state* analysis. . . .

Outline

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

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 / finite-horizon 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*,

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

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. Nor$(0, 1)$. Then $R_0 = 1 + \theta^2$, $R_{\pm 1} = \theta$, and $R_k = 0$, o'w.

By (2), $n\text{Var}(\bar{Y}_n) = R_0 + 2\sum_{k=1}^{n-1}\left(1 - \frac{k}{n}\right)R_k = (1+\theta)^2 - \frac{2\theta}{n}$, 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. Nor$(0, 1-\phi^2)$, $-1 < \phi < 1$, and $Y_0 \sim$ 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* (BM) 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,

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

Similar to IR (actually, more similar than I'm indicating, because I'm using different notation), we define the *batch means estimator* for $\sigma^2 = \lim_{n\to\infty} n\mathrm{Var}(\bar{Y}_n) = \lim_{m\to\infty} m\mathrm{Var}(\bar{Y}_{1,m})$ as

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

How good is $\widehat{V}_{\mathrm{B}}$ as an estimator of σ^2? Let's look at its mean and variance.

First of all, we have $\mathrm{E}[\widehat{V}_{\mathrm{B}}] \doteq \frac{\sigma^2}{b-1}\mathrm{E}[\chi^2(b-1)] = \sigma^2$, so $\widehat{V}_{\mathrm{B}}$ is asymptotically unbiased for σ^2 as the batch size $m \to \infty$.

More-Precise Result: It can be shown that

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

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

$$\mathrm{Var}(\widehat{V}_{\mathrm{B}}) \;\doteq\; \frac{\sigma^4}{(b-1)^2}\mathrm{Var}(\chi^2(b-1)) \;=\; \frac{2\sigma^4}{b-1}.$$

These facts immediately imply that for large m and b,

$$\mathrm{MSE}(\widehat{V}_{\mathrm{B}}) \;=\; \mathrm{Bias}^2 + \mathrm{Var} \;\doteq\; \frac{\gamma^2}{m^2} + \frac{2\sigma^4}{b}.$$

Now let's find m to minimize MSE. To do so, take $m = cn^{\delta}$, where $c > 0$ and $0 < \delta < 1$. Then

$$\mathrm{MSE}(\widehat{V}_{\mathrm{B}}) \;\doteq\; \frac{\gamma^2}{c^2 n^{2\delta}} + \frac{2c\sigma^4}{n^{1-\delta}}.$$

It's easy to see that the choice $\delta = 1/3$ yields the fastest convergence to 0 for $\mathrm{MSE}(\widehat{V}_{\mathrm{B}})$.

For $\delta = 1/3$, we have

$$\mathrm{MSE}(\widehat{V}_{\mathrm{B}}) \;\doteq\; \frac{1}{n^{2/3}}\left[\frac{\gamma^2}{c^2} + 2c\sigma^4\right].$$

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,

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

Of course, σ^2 and γ are unknown in practice and must be estimated (somehow) — this is not for today.

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. Nor$(\mu, \sigma^2/m)$ for large m, we get the following approximate $100(1-\alpha)\%$ CI for μ:

$$\mu \in \bar{Y}_n \pm t_{\alpha/2,b-1}\sqrt{\widehat{V}_{\mathrm{B}}/n}.$$

This equation is similar to (3) (though I'm using different notation). The difference is that BM divides one long run into a number of batches, whereas IR uses a number of independent shorter runs.

Consider the old IR numerical example but now pretend that the Z_i's are batch means (instead of replicate means); then the same numbers carry through the example if you note that $S_Z^2/r = \widehat{V}_{\mathrm{B}}/n$.

Some Properties of the Batch Means CI

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

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

Remarks: $\mathrm{E}[H]$ decreases in b, though it smooths out around $b = 30$. A common recommendation is to take $b \doteq 30$ and concentrate on increasing the batch size m as much as possible.

The technique of BM 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 §4.

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

Independent Replications

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

This problem is explicitly avoided by the method of IR, described in the context of terminating simulations in §3. In fact, the replicate means are independent by their construction.

Unfortunately, since *each* of the r reps has to be started properly, initialization bias presents more trouble when using IR for steady-state analysis than when using BM.

Recommendation: Because of initialization bias in each of the replications, *use batch means over independent reps.* (Alexopoulos and Goldsman 2004, "To Batch or not to Batch?")

Overlapping Batch Means

Suppose that we use the following *overlapping batches*

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

with overlapping batch means

$$
\bar{Y}^{\mathrm{o}}_{i,m} = \frac{1}{m} \sum_{j=i}^{i+m-1} Y_j, \quad i = 1, 2, \ldots, n-m+1.
$$

Turns out: Even though the $\bar{Y}^{\mathrm{o}}_{i,m}$'s are highly correlated, no problema!

The overlapping batch means (OBM) estimator for μ is $\bar{Y}_n$ (no surprise), and the OBM estimator for $\sigma^2 = \lim_{n\to\infty} n\mathrm{Var}(\bar{Y}_n)$ is

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

Facts: As n and m get large,

$$\frac{\mathrm{E}[\widehat{V}_{\mathrm{O}}]}{\mathrm{E}[\widehat{V}_{\mathrm{B}}]} \to 1 \quad \text{and} \quad \frac{\mathrm{Var}(\widehat{V}_{\mathrm{O}})}{\mathrm{Var}(\widehat{V}_{\mathrm{B}})} \to \frac{2}{3}.$$

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 $\widehat{V}_{\mathrm{O}}$ 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 $\widehat{V}_{\mathrm{O}} \approx \sigma^2 \chi^2(d)/d$, where $d = \frac{3}{2}(b-1)$. So you get 50% more d.f. than regular batch means.

Resulting CI: $\mu \in \bar{Y}_n \pm t_{\alpha/2,d}\sqrt{\widehat{V}_{\mathrm{O}}/n}$

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.