Output Analysis (Chapter 9, Law)

- **Introduction**
  - The basic, most serious disadvantage of simulation is that we don’t get exact “answers”.
  - Two different runs of the same model (with different random numbers) lead to different answers.
  - Thus simulation output results are really observations from their probability distribution.

- **The Statistical Nature of Simulation Output**
  - Let $Y_1, Y_2, \ldots$ be an output process from a single simulation run. E.g.,
    - $Y_i = \text{delay in queue of } i^{th} \text{ arriving customer},$
    - $Y_i = \text{production in } i^{th} \text{ hour in a factory}.$
- $Y_i$'s are random variables that are generally neither independent nor identically distributed (nor normally distributed).
- So classical iid normal-theory statistical methods don’t apply directly to the $Y_i$'s (e.g. estimating confidence intervals or doing tests of hypothesis).
- Let $y_{11}, y_{12}, \ldots, y_{1m}$ be a realization of the random variables $Y_1, Y_2, \ldots, Y_m$ resulting from making a single simulation run of length $m$ observations, using a particular stream of underlying U(0, 1) random numbers (i.e. 1 replication).
- If we use another stream of random numbers in a second replication, we get a realization $y_{21}, y_{22}, \ldots, y_{2m}$.
- The realization $y_{21}, y_{22}, \ldots, y_{2m}$ is independent of, but identically distributed as $y_{11}, y_{12}, \ldots, y_{1m}$.
- Then, by making $n$ replications we obtain $n$ realizations.
  
  $y_{11}, y_{12}, \ldots, y_{1m}$
  $y_{21}, y_{22}, \ldots, y_{2m}$
  $\vdots \quad \vdots \quad \vdots$
  $y_{n1}, y_{n2}, \ldots, y_{nm}$
- Observations within rows are not iid, but observations across columns (e.g. $y_{12}, y_{22}, \ldots, y_{n2}$) are iid but may not be normally distributed.
This implies that summary measures across runs are also iid.

For example, the mean of the observations in different replication,

\[
\bar{y}_1 = \sum_{i=1}^{m} y_{1m}, \quad \bar{y}_2 = \sum_{i=1}^{m} y_{2m}, \ldots, \quad \bar{y}_n = \sum_{i=1}^{m} y_{nm}
\]

are iid.

- **Example 1.**

  Consider a bank with 5 tellers, one FIFO queue, open 9am-5pm, flush out before stopping. Interarrivals ~ expo (mean = 1 min.), service times ~ expo (mean = 4 min.)

  Simulating 10 days (replications) starting with no customers, the summary measures are as follows.

<table>
<thead>
<tr>
<th>Replication</th>
<th>Number served</th>
<th>Finish time (hours)</th>
<th>Average delay in queue (minutes)</th>
<th>Average queue length</th>
<th>Proportion of customers delayed &lt; 5 minutes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>484</td>
<td>8.12</td>
<td>1.53</td>
<td>1.52</td>
<td>0.917</td>
</tr>
<tr>
<td>2</td>
<td>475</td>
<td>8.14</td>
<td>1.66</td>
<td>1.62</td>
<td>0.916</td>
</tr>
<tr>
<td>3</td>
<td>484</td>
<td>8.19</td>
<td>1.24</td>
<td>1.23</td>
<td>0.952</td>
</tr>
<tr>
<td>4</td>
<td>483</td>
<td>8.03</td>
<td>2.34</td>
<td>2.34</td>
<td>0.822</td>
</tr>
<tr>
<td>5</td>
<td>455</td>
<td>8.03</td>
<td>2.00</td>
<td>1.89</td>
<td>0.840</td>
</tr>
<tr>
<td>6</td>
<td>461</td>
<td>8.32</td>
<td>1.69</td>
<td>1.56</td>
<td>0.856</td>
</tr>
<tr>
<td>7</td>
<td>451</td>
<td>8.09</td>
<td>2.69</td>
<td>2.50</td>
<td>0.783</td>
</tr>
<tr>
<td>8</td>
<td>486</td>
<td>8.19</td>
<td>2.86</td>
<td>2.83</td>
<td>0.782</td>
</tr>
<tr>
<td>9</td>
<td>502</td>
<td>8.15</td>
<td>1.70</td>
<td>1.74</td>
<td>0.873</td>
</tr>
<tr>
<td>10</td>
<td>475</td>
<td>8.24</td>
<td>2.60</td>
<td>2.50</td>
<td>0.779</td>
</tr>
</tbody>
</table>

Note that the “variability” of summary measures across replications.

One replication does not give the right answer.
• Types of Output Performance Measures
  ➢ Typical performance measures that are estimated include
    o Average time in system
    o Worst (longest) time in system
    o Average, worst time in queue(s)
    o Average, worst, best number of “good” pieces produced per day
    o Variability (standard deviation, range) of number of parts produced per day
    o Average, maximum number of parts in the system (WIP)
    o Average, maximum length of queue(s)
    o Proportion of time a machine is down, up and busy, up and idle
  ➢ Asking for different performance measures can change the way simulation is done. (E.g., estimating a certain measure may require more running time than another measure).

• Transient and Steady-State Behavior
  ➢ As aforementioned, the outputs, \( Y_1, Y_2, \ldots, Y_m \) of a simulation are not iid.
  ➢ However, in some situations, the distribution of \( Y_i \) could converge to a “steady-state” distribution for \( i \) large enough, independent of the initial conditions.
Specifically, let $F_i(y|I)$ be the distribution function of $Y_i$
given initial condition $I$ ($I$ could be, for example, the initial
number in the system).

If a steady state distribution exists, then

$$\lim_{i \to \infty} F_i(y|I) = F(y).$$

Roughly speaking, if there is a time index $k$ such that for
$i > k$, $F_i(y|I) \approx F(y)$, then the process is “in steady state”
after time $k$.

Even in steady state the $Y_i$’s are not independent.

The steady-state distribution (if it exists) does not depend on
initial conditions.

But the nature and rate of convergence of the transient
distributions can depend heavily on the initial conditions.
E.g., the following graph plots the convergence of the mean of expected delay of customer \( i \), \( E[D_i] \), of an \( M/M/1 \) queue with \( \lambda = 1 \) and \( \mu = 10/9 \), to the steady state mean \( W_q = \rho / [\mu(1-\rho)] = 8.1 \), for different values of the initial number in the system, \( s \).

- Types of Simulations with Regard to Output Analysis
  - Simulations can be classified as *terminating* and *nonterminating*.
  - In terminating simulation, there is a natural event that specifies then end of each replication.
E.g., simulating one day of operation of a retailer (from 8 to 5) starting empty and terminating when all customers arriving before 5 finish service, is a terminating simulation.

However, simulating a manufacturer where leftover work in one day is carried over to the next day is not a terminating simulation.

In a nonterminating simulation, there is no natural event that specifies the end of a replication. The interest is in “long-run” behavior characteristic of “normal” operation.

A measure of performance of a nonterminating simulation can be estimated from the steady-state distribution of the process at hand (if it exists).

- **Estimating means in terminating simulations**

  Estimating means in a terminating simulation is based on the approximate assumption that the observations from different replications are independent and normally distributed.

  Let $X_j$ be a random variable defined over the $j^{th}$ replication (e.g. $X_j =$ average delay in queue, machine utilization, etc.).

  Then, with $n$ independent replications, an *approximate* confidence interval for the mean of $X_j$ can be estimated based on the normal assumption.
A 100 \((1-\alpha)\)% “confidence interval” for the mean of \(X_j\) is

\[
\bar{X}(n) \pm t_{n-1,1-\alpha/2} \sqrt{\frac{S^2(n)}{n}}
\]

where \(\bar{X}(n) = \frac{\sum_{j=1}^{n} X_j}{n}\), \(S^2(n) = \frac{\sum_{j=1}^{n} (X_j - \bar{X}(n))^2}{n-1}\), are the sample mean and variance, and \(t_{n-1,1-\alpha}\) is such that

\[
P\{T_{n-1} < t_{n-1,1-\alpha/2}\} = 1-\alpha/2,
\]

where \(T_{n-1}\) is the Student’s \(t\) Distribution with \(n-1\) degrees of freedom.

The usage of the student distribution is based on the following fact.

**Fact.** If \(X_1, X_2, \ldots, X_n\) are iid normal random variables with unknown variance, and “true” mean \(\mu\), then

\[
(\bar{X}(n) - \mu) / (S(n) / \sqrt{n}) \text{ has a Student’s } t \text{ Distribution with } n-1 \text{ degrees of freedom.}
\]

When \(n\) is large the Student’s \(t\) distribution converges to the standard normal. That’s why the \(t\) distribution is often used for small samples (e.g. number of replications).

A 100\((1-\alpha)\)% confidence interval (CI) means that

100\((1-\alpha)\)% of the times the interval will contain \(\mu\).

That is, the probability that the confidence interval contains \(\mu\) is \(1-\alpha\).
See Law book for examples and for discussion of how accurate is the t-student distribution CI.

As a guidelines the normal approximation is accurate if $X_j$’s are some kind of an “average.”

**Example 2.**

For the bank in Example 1, let $X_j$ = mean delay in queue from the $j^{th}$ replications. Then, it can be seen that

$\bar{X}(10) = 2.03$ and $S^2(10) = 0.31$. Then, a 95% confidence interval for the mean delay in queue is

$$\bar{X}(10) \pm t_{9,0.95} \sqrt{\frac{S^2(10)}{10}} = 2.03 \pm 1.833 \sqrt{0.31/10} = 2.03 \pm 0.32,$$

where $t_{9,0.95} = 1.833$ is found from page 716 (Law).

Here, 2.03 is the “point estimate”, and 0.32 is the “half-length” of the confidence interval.

The half width is around 15% of the point estimate which may be reasonably accurate for some applications.

**How in Arena?**

Upon specifying a number of replications, Arena generates confidence intervals for output from terminating simulations as explained above. (See Section 6.3 in KSS book.)

If the runs are long enough, Arena may also give confidence interval for each replication separately based on the “batch means” as explained below.
• **Obtaining a specific precision**

  ➢ If the number $n$ of replication is chosen too small, the confidence intervals might too wide to be useful.
  ➢ How large should $n$ to get a good precision?
  ➢ Depends on how one defines precision.

  ➢ The idea is to make the half-width $\delta(\alpha, n) = t_{n-1,1-\alpha/2} \sqrt{\frac{S^2(n)}{n}}$ small enough.
  ➢ “Small enough” can be quantified in two ways.

  ➢ With an *absolute precision* approach, one specifies $\beta > 0$, and finds $n^*$ which is big enough to make, $|\bar{X}(n^*) - \mu| < \beta$, or equivalently $\delta(\alpha, n^*) \leq \beta$.

  ➢ The value of $n^*$ can be found sequentially by increasing $n$ (maybe from $n = 2$) and evaluating $S(n)$ and $\delta(\alpha, n)$.

  ➢ One quick way to estimate $n^*$ is used fixed estimates for $S(n)$ based on a small number of replications.

  ➢ For example, if we run the simulation for $n_1$ replications and we want to enhance the estimate in terms of absolute precision. Then, an approximate value for $n^*$ that will achieve an absolute precision $\beta$ is

  $$n^* = \min \left\{ i \geq n_1 : t_{i-1,1-\alpha/2} \sqrt{\frac{S^2(n_i)}{i}} \leq \beta \right\}.$$
With a *relative precision* approach, one specifies $0 < \gamma < 1$, and finds $n^*$ which is large enough to make

$$| \bar{X}(n^*) - \mu | / \mu < \gamma .$$

This translates into making $\delta(\alpha, n^*) / \bar{X}(n^*) < \gamma / (1 - \gamma)$.

Note that $\gamma$ is replaced by $\gamma (1 - \gamma)$ because we use $\bar{X}(n^*)$ as an estimate of $\mu$. (See text for details if you care.)

The value of $n^*$ can be determined sequentially or approximately similar to the absolute precision case.

- **Estimating other measures of performance**
  - Other measures of performance could be also estimated in a terminating simulation.
  - The most useful one are the variance, the proportions, and the quantiles.
  - Example 9.20 in Law book indicates that comparing alternatives on the basis of means only can be troublesome.
  - Confidence intervals for proportions and quantiles are more difficult to obtain than for the means.
  - See Law book for large sample estimation for proportions and quantiles.
• **Choosing Initial Conditions**
  - For terminating simulations, the initial conditions can affect the output performance measure, so the simulations should be initialized appropriately.
  - For example, suppose one wants to estimate expected delay in queue of bank customers who arrive and complete their service between noon and 1:00 pm (peak hour).
  - The bank is likely to be crowded already at noon, so starting empty and idle at noon will probably bias the results low.
  - There are two possible alternatives here.
    - If bank actually opens at 9:00 am, start the simulation empty and idle, let it run for 3 simulated hours, clear the statistical accumulators, and observe statistics for the next hour.
    - Fit a (discrete) distribution to number of customers at noon, and draw from this distribution to initialize the simulation.

• **Statistical Analysis for Steady-State Parameters**
  - In a nonterminating simulation, the interest is in estimating the long run “steady state” measures of performance.
  - However, initial conditions could bias the estimation.
  - To avoid this bias, the most common technique is to warm up the model, also called initial-data deletion.
Identify index $l$ (for discrete-time processes) or time $t_0$ (for continuous-time processes) beyond which the output appears not to be drifting any more statistically.

Then, consider data beyond the cut-off point to estimate the desired measure of performance.

For example, suppose the objective is to estimate the steady state mean of a process based on observed values, $Y_1, Y_2, \ldots, Y_m$, in a single replication.

Then, after removing $l$ initial transient data, an estimator for the mean is

$$\bar{Y}(m, l) = \frac{\sum_{i=l+1}^{m} Y_i}{m - l}. $$

There is no straightforward approach for choosing the warmup period $l$.

The simplest method for finding $l$ is a graphical procedure based on multiple replications.

(Many replications are used because it’s difficult to estimate $l$ from just one replication.)

The idea is to plot an estimate for $E[Y_i]$ versus $i$ and choose $l$ by eyeballing a point where the curves flattens out.

To estimate $E[Y_i]$, first an average across replications is taken leading to $\bar{Y}_i$ and then a moving average with window $w$ is applied to for further smoothing leading to $\bar{Y}_i(w)$.
A plot of $\bar{Y}_i(w)$ is then used to estimate $l$ as shown.

Arena’s companion package “Output Analyzer” can be used for applying the procedure above. (See Section 7.2.1 in KSS book for an example.)

Note finally that there is no easy way to specify the run length in a nonterminating simulation. You can (i) rely on a graphical procedure to judge if steady state is reached or (ii) keep increasing the length of the simulation until the variability in output measure is reasonable.

---

• **Replication/Deletion Approaches for Estimating Means in a Nonterminating Simulation**
  
  ➢ Assume that an appropriate warm-up period \( l \) has been identified.
  
  ➢ The replication/deletion technique starts with \( n \) replications each having \( m \) observations and delete the first \( l \) observations of each replication.
  
  ➢ Then, it proceeds just like the case of terminating simulation to estimate a CI for the mean based on the normal approximation and the t-distribution.

• **Other Approaches for Estimating Means in a Nonterminating Simulation**
  
  ➢ Other approaches for estimating means in a nonterminating simulation are based on having one long run (versus many short replications in Replication/Deletion).

<table>
<thead>
<tr>
<th></th>
<th>Many short runs</th>
<th>One long run</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Good</strong></td>
<td>Simple (same as terminating)</td>
<td>Less point-estimator bias</td>
</tr>
<tr>
<td></td>
<td>Get IID data</td>
<td>No restarts</td>
</tr>
<tr>
<td><strong>Bad</strong></td>
<td>Point-estimator bias (initial transient)</td>
<td>“Sample” of size 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hard to get variance estimate</td>
</tr>
</tbody>
</table>
The batch means approach considers a single run with \( m \) observations and divides the observations into \( n \) batches of length \( k \) each (\( m = nk \)). (Warm-up period data may be eliminated first.)

Let \( \overline{Y}_j(k) \) be the mean of the observations in batch \( j \).
The batch means approach is based on the approximate assumption that \( \bar{Y}_j(k) \) are iid normal and that they provide an unbiased estimator for the mean.

Then, a confidence interval for the mean is found as

\[
\bar{Y}(m) \pm t_{n-1,1-\alpha/2} \sqrt{\frac{S^2(n)}{n}}
\]

where \( \bar{Y}(m) = \frac{\sum_{j=1}^{n} \bar{Y}_j(k)}{n} \), \( S^2(n) = \frac{\sum_{j=1}^{n} (\bar{Y}_j(k) - \bar{Y}(m))^2}{n} \).

The main advantage of the batch means approach is its simplicity and the major drawback is the difficulty of choosing an appropriate batch size that reduces correlation.

**Batch Means in Arena**

- If the simulation run is long enough (> 320 observations), then Arena attempts to generate confidence intervals based on batch means (see Section 7.2.3 for details.)
- Arena utilizes a number of batched between 20 and 40.
- Arena implements a statistical test to check for correlation between batches.
- If the correlation test is not passed, Arena does not report an confidence interval. (Then, you can try to increase the batch size or the simulation run length.)
- Interestingly, you can specify a desired value of the half-width of the confidence interval as a stopping rule for the simulation (see KSS book, p. 320).