Skart: A Skewness- and Autoregression-Adjusted Batch-Means Procedure for Simulation Analysis

Ali Tafazzoli, North Carolina State University
James R. Wilson, North Carolina State University
Emily K. Lada, SAS Institute Inc.
Natalie M. Steiger, University of Maine

www.isc.ecn.edu/jwilson

May 31, 2009
Outline

1. Motivation and Problem Statement
2. Skewness & Autocorrelation Adjustment to Student’s $t$-Statistic
3. Skart Procedure
4. Performance Evaluation of Skart
5. Conclusions and Recommendations
Why Not Use Classical Statistics for Steady-State Simulation Analysis?

Classical Confidence Interval Based on Student’s $t$-Statistic

If the simulation output process $\{X_i : i = 1, \ldots, n\}$ consists of independent and identically distributed (i.i.d.) normal random variables, then a valid $100(1 - \alpha)$% CI for $\mu_X$ is

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

where: $t_{1-\alpha/2,n-1}$ denotes the $1 - \alpha/2$ quantile of Student’s $t$-distribution with $n - 1$ degrees of freedom; and $S^2 = (n - 1)^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$ is the sample variance of the $\{X_i\}$.

Three fundamental problems arise in steady-state simulation output analysis:

- **Warm-up Problem**—Caused by transients in the initial sequence of responses owing to the system’s starting condition;

- **Correlation Problem**—Caused by stochastic dependencies among successive responses; and

- **Nonnormality Problem**—Caused by pronounced departures from normality in simulation-generated responses.
Method of Nonoverlapping Batch Means (NBM)

The conventional method of batch means requires a sufficiently large batch size $m$ so the adjacent, nonoverlapping batch means

$$Y_j(m) = \frac{1}{m} \sum_{i=m(j-1)+1}^{mj} X_i \quad \text{for } j = 1, \ldots, k$$

are i.i.d. normal. As $m \to \infty$, an asymptotically valid $100(1 - \alpha)\%$ CI for $\mu_X$ is

$$\bar{Y}(m, k) \pm t_{1-\alpha/2, k-1} \frac{S_{m,k}}{\sqrt{k}},$$

where $\bar{Y}(m, k)$ and $S_{m,k}$ are the sample mean and variance of the batch means, respectively.

The main difficulty with any batch-means procedure is reliable determination of the following:

(i) the length of the warm-up period; and

(ii) the batch size $m$. 


Skart: A New Batch-Means Procedure

- Skart is a new batch-means procedure that determines sufficiently large values for the batch size \( m \) and the length of the initial warm-up period so the truncated batch means are approximately a stationary first-order autoregressive process with mean \( \mu_X \).

- Skart determines \( k' \) truncated batch means \( \{Y_j(m)\} \) with batch size \( m \) beyond the warm-up period, with sample mean \( \bar{Y}(m, k') \) and sample variance \( S_{m,k'}^2 \).

- Skart determines \( k'' \) approximately i.i.d. spaced batch means with batch size \( m \) and spacer size nearly the same as the length of the warm-up period so that the spaced batch means have sample variance \( S_{m,k''}^2 \) and an approximately unbiased estimator of marginal skewness \( \hat{\beta}_{m,k''} \).

**Skart’s skewness- and autoregression-adjusted CI for \( \mu_X \)**

\[
\left[ \bar{Y}(m, k') - G(t_{1-\alpha/2,k''-1}) \sqrt{\frac{AS_{m,k''}^2}{k'}} , \bar{Y}(m, k') - G(t_{\alpha/2,k''-1}) \sqrt{\frac{AS_{m,k''}^2}{k'}} \right],
\]

where:

\[
G(\zeta) \equiv \frac{\sqrt[3]{1 + 6\beta(\zeta - \beta)} - 1}{2\beta}, \quad \text{with } \beta = \frac{\hat{\beta}_{m,k''}}{6\sqrt{k'}};
\]

\[
A = \left[ 1 + \hat{\phi}_{Y(m)} \right] \left[ 1 - \hat{\phi}_{Y(m)} \right],
\]

with \( \hat{\phi}_{Y(m)} \) denoting the sample lag-one correlation of the truncated batch means.
Skart’s Combined Skewness and Autocorrelation Adjustment to Student’s $t$-Statistic

If the truncated batch means $\{Y_j(m) : j = 1, \ldots, k'\}$ are correlated samples from a nonnormal distribution, then the usual batch-means $t$-statistic

$$ t = \frac{\bar{Y}(m,k') - \mu_X}{\sqrt{S^2_{m,k'} / k'}} $$

(1)

does not have Student’s $t$-distribution with $k' - 1$ degrees of freedom and cannot be used to compute exact CIs for $\mu_X$.

In steady-state simulation, the batch means are often substantially correlated and nonnormal → We need to adjust the batch-means $t$-statistic (1) by combining the following:

- an appropriate adjustment to the numerator of (1) in terms of

  $$ \hat{B}_{m,k''} = \text{estimated skewness of the } \{Y_j(m)\}, $$
  $$ S^2_{m,k''} = \text{estimated variance of the } \{Y_j(m)\}, $$

  where $\hat{B}_{m,k''}$ and $S^2_{m,k''}$ are computed from $k''$ spaced batch means separated by spacers having the same size as the warm-up period so the spaced batch means are approximately independent; and

- an appropriate correlation adjustment to the denominator of (1) in terms of

  $$ \hat{\phi}_{Y(m)} = \text{sample lag-one correlation of } \{Y_j(m) : j = 1, \ldots, k'\}. $$
To adjust the numerator of (1), we compute the skewness-adjustment coefficients

\[
C_0 = \frac{\hat{B}_{m,k''} S_{m,k''}}{6k'}, \quad C_2 = \frac{\hat{B}_{m,k''}}{3S_{m,k''}}, \quad C_3 = \frac{\hat{B}_{m,k''}^2}{27S_{m,k''}^2}.
\]

To adjust the denominator of (1), we compute the correlation-adjustment coefficient,

\[
A = \frac{1 + \hat{\phi}_{Y(m)}}{1 - \hat{\phi}_{Y(m)}}
\]

so that

\[
A S_{m,k''}^2 / k' \text{ is an estimator of } \text{Var}[\bar{Y}(m, k')] \approx \gamma_X / (mk') \text{ with } k'' - 1 \text{ degrees of freedom.} \tag{2}
\]

The modified Student's \( t \)-statistic used in Skart has the final form

\[
t_1 = \frac{[\bar{Y}(m, k') - \mu_X] + C_0 + C_2[\bar{Y}(m, k') - \mu_X]^2 + C_3[\bar{Y}(m, k') - \mu_X]^3}{\left[ A S_{m,k''}^2 / k' \right]^{1/2}},
\]

which has approximately Student's \( t \)-distribution with \( k'' - 1 \) degrees of freedom.
Overview of Skart

- Skart is an automated sequential procedure for \textit{on-the-fly} or \textit{offline} steady-state simulation output analysis.
- Skart requires the following user-supplied inputs:
  - An initial time series with length at least 1,280 observations from the simulation model whose steady-state mean $\mu_X$ is to be estimated;
  - A desired CI coverage probability $1 - \alpha$, where $0 < \alpha < 1$;
  - An upper bound $H^*$ on the final CI half-length, where $H^*$ is expressed in absolute or relative terms;
  - For time-persistent statistics, Skart asks the user for a sampling interval value $\Delta$ expressed in the basic time units of the simulation clock. Skart then computes time-weighted statistics at multiples of $\Delta$ during a pilot simulation run.

- Skart returns the following outputs:
  - A nominal $100(1 - \alpha)\%$ CI for $\mu_X$ that satisfies the specified precision requirement; or
  - A new, larger sample size to be used by Skart.

- The delivered CIs can be \textit{asymmetric} around the computed mean for the case of no precision requirement.
- When a precision level is specified, the CI’s half-length is considered to be equal to the length of the wider tail.
**Skart Procedure**

**Flow Chart of Skart**

1. **Start**
   - Collect observations, compute their sample skewness, and set batch size

2. **Compute nonspaced batch means and their sample skewness; set max batches per spacer**
   - Independence test passed?
     - Yes: Skip first spacer; reinflate batch count and reset batch size
     - No: Add another batch to each spacer; recompute spaced batch means

3. **Reached max batches per spacer?**
   - No: Compute nonspaced batch means and autoregression-adjusted variance estimator
   - Yes: Increase batch size; deflate batch count; set spacer size to zero and collect additional observations

4. **CI meets precision requirement?**
   - Yes: Deliver CI
   - No: Compute spaced batch means and associated skewness-adjusted t-ratio

5. **Compute skewness- and autoregression-adjusted CI**

---

Tafazzoli et al. – NC State
Skart: A Skewness- and Autoregression-Adjusted Batch-Means Procedure

9 / 16
Performance Evaluation

We examined the performance of Skart for both observation-based and time-persistent statistics with respect to:

- CI coverage probability;
- Mean and variance of the CI half-length; and
- Total sample size required to deliver the final CI.

We applied Skart to a series of test processes with known $\mu_X$ to “stress-test” Skart and to gauge its performance on processes with characteristics typical of large-scale applications:

- $M/M/1$ queue waiting times (empty-and-idle and extreme initial conditions);
- $M/M/1$ number-in-queue process;
- $M/H_2/1$ queue waiting times;
- $AR(1)$ process;
- Autoregressive-to-Pareto (ARTOP) Process;
- $M/M/1$ LIFO queue waiting times;
- $M/M/1$ SIRO queue waiting times;
- $M/M/1/M/1$ queue waiting times; and
- Two-state discrete-time Markov chains (symmetric, skewed, and highly skewed).

We compared the performance of Skart, ASAP3, WASSP and SBatch.
Performance of Skart in the $M/M/1$ queue-waiting-time process for different levels of server utilization (SU) and for 90% and 95% CIs

<table>
<thead>
<tr>
<th>Performance Measure</th>
<th>10% SU</th>
<th>20% SU</th>
<th>30% SU</th>
<th>40% SU</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>90%</td>
<td>95%</td>
<td>90%</td>
<td>95%</td>
</tr>
<tr>
<td>CI coverage</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>90% SU</td>
<td>89.8%</td>
<td>94.6%</td>
<td>89.7%</td>
<td>96.1%</td>
</tr>
<tr>
<td>20% SU</td>
<td>89.8%</td>
<td>94.6%</td>
<td>89.7%</td>
<td>96.1%</td>
</tr>
<tr>
<td>30% SU</td>
<td>89.8%</td>
<td>94.6%</td>
<td>89.7%</td>
<td>96.1%</td>
</tr>
<tr>
<td>40% SU</td>
<td>89.8%</td>
<td>94.6%</td>
<td>89.7%</td>
<td>96.1%</td>
</tr>
<tr>
<td>Avg. sample size</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>90% SU</td>
<td>20,650</td>
<td>20,650</td>
<td>14,597</td>
<td>14,597</td>
</tr>
<tr>
<td>20% SU</td>
<td>20,650</td>
<td>20,650</td>
<td>14,597</td>
<td>14,597</td>
</tr>
<tr>
<td>30% SU</td>
<td>20,650</td>
<td>20,650</td>
<td>14,597</td>
<td>14,597</td>
</tr>
<tr>
<td>40% SU</td>
<td>20,650</td>
<td>20,650</td>
<td>14,597</td>
<td>14,597</td>
</tr>
<tr>
<td>Avg. rel. precision</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>90% SU</td>
<td>6.74%</td>
<td>8.09%</td>
<td>12.21%</td>
<td>14.78%</td>
</tr>
<tr>
<td>20% SU</td>
<td>6.74%</td>
<td>8.09%</td>
<td>12.21%</td>
<td>14.78%</td>
</tr>
<tr>
<td>30% SU</td>
<td>6.74%</td>
<td>8.09%</td>
<td>12.21%</td>
<td>14.78%</td>
</tr>
<tr>
<td>40% SU</td>
<td>6.74%</td>
<td>8.09%</td>
<td>12.21%</td>
<td>14.78%</td>
</tr>
<tr>
<td>Avg. CI half-length</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>90% SU</td>
<td>0.0077</td>
<td>0.0093</td>
<td>0.0331</td>
<td>0.0413</td>
</tr>
<tr>
<td>20% SU</td>
<td>0.0077</td>
<td>0.0093</td>
<td>0.0331</td>
<td>0.0413</td>
</tr>
<tr>
<td>30% SU</td>
<td>0.0077</td>
<td>0.0093</td>
<td>0.0331</td>
<td>0.0413</td>
</tr>
<tr>
<td>40% SU</td>
<td>0.0077</td>
<td>0.0093</td>
<td>0.0331</td>
<td>0.0413</td>
</tr>
<tr>
<td>Var. CI half-length</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>90% SU</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>20% SU</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>30% SU</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>40% SU</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Notice the U-shaped trend in the sample size values as the server utilization increases.
Skart's Performance for $M/M/1$ Queue Waiting Times

- Waiting times in $M/M/1$ queue with 90% traffic intensity and empty-and-idle initial condition

<table>
<thead>
<tr>
<th>Precision Requirement</th>
<th>Performance Measure</th>
<th>Skart # replications</th>
<th>Skart CI coverage</th>
<th>Skart Avg. sample size</th>
<th>Skart Avg. CI half-length</th>
<th>Skart Var. CI half-length</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td># replications</td>
<td>1,000</td>
<td>87.9%</td>
<td>41,855</td>
<td>1.6741</td>
<td>0.2282</td>
</tr>
<tr>
<td></td>
<td>CI coverage</td>
<td>87.1%</td>
<td>18,090</td>
<td>0.2603</td>
<td>1.3864</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Avg. sample size</td>
<td>90%</td>
<td>18,090</td>
<td>3.0715</td>
<td>1.1905</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Avg. CI half-length</td>
<td>90.4%</td>
<td>18,090</td>
<td>0.0387</td>
<td>1.1556</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Var. CI half-length</td>
<td>90.4%</td>
<td>18,090</td>
<td>0.0387</td>
<td>1.1556</td>
<td></td>
</tr>
</tbody>
</table>

| ±15%                   | # replications       | 1,000                | 87.5%             | 70,473                 | 1.1905                   | 0.0217                   |
|                        | CI coverage          | 86.6%                | 92,049            | 0.0396                 | 1.1556                   |
|                        | Avg. sample size     | 87.2%                | 103,742           | 0.0439                 | 1.1556                   |
|                        | Avg. CI half-length  | 91%                  | 103,742           | 0.0387                 | 1.1556                   |
|                        | Var. CI half-length  | 91%                  | 103,742           | 0.0387                 | 1.1556                   |

| ±7.5%                  | # replications       | 1,000                | 90%               | 273,540                | 0.6396                   | 0.0013                   |
|                        | CI coverage          | 88.8%                | 388,000           | 0.0545                 | 0.6141                   |
|                        | Avg. sample size     | 90.4%                | 287,568           | 0.5866                 | 0.6141                   |
|                        | Avg. CI half-length  | 89.5%                | 287,568           | 0.0627                 | 0.6141                   |
|                        | Var. CI half-length  | 89.5%                | 287,568           | 0.0627                 | 0.6141                   |

Coverage of ABATCH’s nominal 90% CIs: no prec., 60%; ±15%, 72%; ±7.5%, 82%.

Coverage of Law-Carson’s nominal 90% CIs: no prec., 85%; ±15%, 85%; ±7.5%, 87%.

Coverage of Heidelberger-Welch’s 90% CIs: no prec., 68%; ±15%, 76%; ±7.5%, 77%.
Skart’s Performance for $M/M/1$/LIFO Queue Waiting Times

Waiting times in $M/M/1$/LIFO queue with 80% traffic intensity and empty-and-idle initial condition

<table>
<thead>
<tr>
<th>Precision Requirement</th>
<th>Performance Measure</th>
<th>Skart</th>
<th>Nominal 90% CIs</th>
<th>90% CIs</th>
<th>ASAP3</th>
<th>Nominal 95% CIs</th>
<th>95% CIs</th>
<th>ASAP3</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td># replications</td>
<td>1,000</td>
<td>1,000</td>
<td>400</td>
<td>400</td>
<td>1,000</td>
<td>1,000</td>
<td>400</td>
</tr>
<tr>
<td></td>
<td>CI coverage</td>
<td>87.1%</td>
<td>91.4%</td>
<td>93%</td>
<td>87%</td>
<td>92.5%</td>
<td>95.9%</td>
<td>96%</td>
</tr>
<tr>
<td></td>
<td>Avg. sample size</td>
<td>23,943</td>
<td>117,416</td>
<td>125,517</td>
<td>53,958</td>
<td>23,943</td>
<td>117,416</td>
<td>124,202</td>
</tr>
<tr>
<td></td>
<td>Avg. CI half-length</td>
<td>0.4318</td>
<td>0.1891</td>
<td>0.2650</td>
<td>0.1060</td>
<td>0.5522</td>
<td>0.2255</td>
<td>0.3350</td>
</tr>
<tr>
<td></td>
<td>Var. CI half-length</td>
<td>0.0326</td>
<td>0.0057</td>
<td>0.0230</td>
<td>0.1060</td>
<td>0.1015</td>
<td>0.0080</td>
<td>0.0310</td>
</tr>
<tr>
<td>±15%</td>
<td># replications</td>
<td>1,000</td>
<td>1,000</td>
<td>400</td>
<td>400</td>
<td>1,000</td>
<td>1,000</td>
<td>400</td>
</tr>
<tr>
<td></td>
<td>CI coverage</td>
<td>89.2%</td>
<td>91.3%</td>
<td>90.7%</td>
<td>86.8%</td>
<td>93.9%</td>
<td>94%</td>
<td>95.2%</td>
</tr>
<tr>
<td></td>
<td>Avg. sample size</td>
<td>27,671</td>
<td>118,209</td>
<td>124,512</td>
<td>54,017</td>
<td>32,985</td>
<td>119,903</td>
<td>126,682</td>
</tr>
<tr>
<td></td>
<td>Avg. CI half-length</td>
<td>0.3703</td>
<td>0.1812</td>
<td>0.2490</td>
<td>0.2600</td>
<td>0.4021</td>
<td>0.2117</td>
<td>0.2960</td>
</tr>
<tr>
<td></td>
<td>Var. CI half-length</td>
<td>0.0051</td>
<td>0.0021</td>
<td>0.0110</td>
<td>0.0040</td>
<td>0.0035</td>
<td>0.0022</td>
<td>0.0110</td>
</tr>
<tr>
<td>±7.5%</td>
<td># replications</td>
<td>1,000</td>
<td>1,000</td>
<td>400</td>
<td>400</td>
<td>1,000</td>
<td>1,000</td>
<td>400</td>
</tr>
<tr>
<td></td>
<td>CI coverage</td>
<td>91.9%</td>
<td>89.5%</td>
<td>90.2%</td>
<td>87.5%</td>
<td>95.6%</td>
<td>95.4%</td>
<td>96.2%</td>
</tr>
<tr>
<td></td>
<td>Avg. sample size</td>
<td>73,986</td>
<td>126,961</td>
<td>152,355</td>
<td>68,325</td>
<td>110,520</td>
<td>134,123</td>
<td>194,590</td>
</tr>
<tr>
<td></td>
<td>Avg. CI half-length</td>
<td>0.2263</td>
<td>0.1734</td>
<td>0.1860</td>
<td>0.2190</td>
<td>0.2253</td>
<td>0.1996</td>
<td>0.1990</td>
</tr>
<tr>
<td></td>
<td>Var. CI half-length</td>
<td>0.0002</td>
<td>0.0012</td>
<td>0.0020</td>
<td>0.0005</td>
<td>0.0002</td>
<td>0.0011</td>
<td>0.0010</td>
</tr>
</tbody>
</table>

- Coverage of ABATCH’s nominal 90% CIs: no prec., 75%; ±15%, 81%; ±7.5%, 89%.
- Coverage of Law-Carson procedure’s 90% CIs: no prec., 64%; ±15%, 76%; ±7.5%, 84%.
Conclusions

- We developed Skart, a robust and efficient implementation of the batch-means method that was designed to fill the gap between theory and practice.

  - Skart is a completely automated nonoverlapping batch-means procedure for constructing a skewness- and autoregression-adjusted CI for the steady-state mean of a simulation output process.
  
  - Skart incorporates some advantages of its predecessors ASAP3, WASSP, and SBatch while exploiting separate adjustments to the classical batch-means CI based on the corresponding effects of nonnormality and correlation of the delivered batch means.

- Our experimentation on a suite of test problems showed that:

  - Skart outperformed both SBatch and WASSP with respect to CI coverage probability and sampling efficiency.
  
  - Skart and ASAP3 produced comparable results in most of the studied test problems; however, in the problems involving the most extreme stress testing, Skart appeared to be a more robust procedure than ASAP3 overall.

  - In particular, Skart’s performance was better than ASAP3 for relatively high (coarse) precision levels when it was applied to processes with an exceptionally high correlation structure, like the AR(1) process, or to processes which are markedly nonnormal, like the $M/M/1/LIFO$ process.
The key advantages of Skart over other steady-state simulation output procedures are:

- Skart is specifically designed to handle time-persistent statistics as well as observation-based statistics.
- Skart has a nonsequential version (N-Skart) in which the user merely supplies a single simulation-generated time series of an arbitrary fixed length and requests a CI with a specific coverage probability based on the available data.
- Skart usually requires a smaller initial sample size compared with other well-known simulation analysis procedures.
- Skart removes the need for the normality test, which can sometimes result in excessive sample sizes, by applying Willink’s skewness adjustment to the classical NBM-based Student’s $t$-statistic.
- Skart efficiently incorporates different methods in the randomness-test and precision-requirement steps to control the sample size growth in highly correlated processes.
Thank You

Questions?