

Statistical analysis of simulation output

Harry Perros

Computer Science Department

NC State University

hp@csc.ncsu.edu

<http://www.csc.ncsu.edu/faculty/perros//index.html>

Statistical estimation

- Let us assume now that the simulation model has been built and debugged.
- At this stage, the simulation model can be exercised to generating a sample of observations for a particular parameter of interest.
- From this sample we can then compute a point estimate and its confidence interval.

Statistical estimation

- For instance, the repairman queue simulation can be run so that to obtain a sequence of successive waiting times:

$$W_1, W_2, \dots, W_n$$

- From this sequence, using simple statistics, we can compute the mean waiting time and its confidence intervals.

Transient vs steady-state simulation

- A simulation model can be used to estimate an operational parameter of interest during the *transient state* or the *steady state*.
- Typically, we are interested in steady-state simulations, though transient state simulations are also of interest.

Steady-state simulation

- A simulation always starts assuming an *initial condition*, i.e.
 - A seed for the generator
 - Initial state of the simulation, i.e. system is empty at the beginning.
- The behaviour of the system is affected by the initial conditions for an initial period T .
- After that period, the simulation reaches a *steady state*. That is, its statistical behaviour becomes independent of the initial conditions.

Transient state analysis

- In this case, we only simulate the system for a short period (before it reaches its steady state), in order to see the effects of an initial condition.
- Also, transient analysis maybe the only choice if the system does not have a steady state.

Estimation techniques for steady-state simulation

- Most of the performance measures are related to the probability distribution of an endogenously created random variable, assuming steady state.
- The most common measures are:
 - Mean
 - Variance
 - Percentiles

Estimation of the mean of a random variable

- Let x_1, x_2, \dots, x_n be n consecutive endogenously obtained observations of a random variable. Then

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

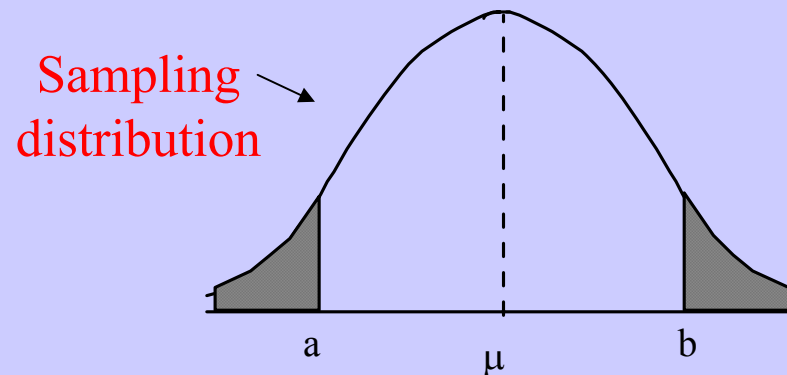
is an unbiased estimate of the true population mean, i.e., the expectation of the random variable.

- **However, it is not sufficient to just calculate the mean, without its confidence interval**

What's a confidence interval?

- Assume a set of observations x_1, x_2, \dots, x_n .
- These observations come from a population, known as the *parent population*. Let μ and σ^2 be the mean and the variance of the parent population.
- We want to estimate μ from the sample mean \bar{x} i.e.
$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$
- How good is our estimate??

- From central limit theorem we have: $\bar{x} \sim N(\mu, \sqrt{\sigma})$



- We fix points a and b , so that 95% of the observations (i.e. sample means) fall in between these points. Using the standard Normal distribution (We assume that $n > 30$), we have:

$$a = \mu - 1.96 \frac{\sigma}{\sqrt{n}}, b = \mu + 1.96 \frac{\sigma}{\sqrt{n}}$$

- Therefore, μ will lie within the interval 95%

$$\left(\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}}\right)$$

- If σ is not known, it can be replaced by s .
- The confidence interval for estimating μ (also known as interval estimate) is:

$$\left(\bar{x} - 1.96 \frac{s}{\sqrt{n}}, \bar{x} + 1.96 \frac{s}{\sqrt{n}}\right)$$

- The above confidence interval applies when the sample size n is at least 30.
- If the sample size is less than 30 (typical when we do replications), then we can construct a similar confidence interval using the *t-student* statistic with $n-1$ degrees of freedom.
- We have:

$$\left(\bar{x} - t_{.95} \frac{s}{\sqrt{n}}, \bar{x} + t_{.95} \frac{s}{\sqrt{n}} \right)$$

- The value 95% is known as the confidence of the estimation.
- Most typical values are: 99%, 95%, 90%.

**NO SIMULATION ESTIMATE SHOULD
BE GIVEN WITHOUT CONFIDENCE
INTERVALS**

Calculating the standard deviation

- The problem with constructing confidence intervals lies in the estimation of the standard deviation s .
- We have two cases:
 - The observations is the sample x_1, x_2, \dots, x_n are independent of each other.
 - The observations are correlated
- **BOTH CASES ARE POSSIBLE !!**

Case 1: Independent observations

If the observations x_1, x_2, \dots, x_n are independent of each other, then the standard deviation is given by:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

and, therefore, we obtain the confidence interval

$$\left(\bar{x} - 1.96 \frac{s}{\sqrt{n}}, \bar{x} + 1.96 \frac{s}{\sqrt{n}} \right)$$

Case 2: Correlated observations

- In general, the observation endogenously created by a simulation tend to be correlated. (Successive waiting times in a queue are correlated.)
- There are various methods for calculating the standard deviation
 - Estimation of the autocorrelation function
 - Batch means
 - Replications
 - Regenerative method

Replications

- Run the simulation several times, each time use a different seed for the generator.

replication 1: $x_{11}, x_{12}, \dots, x_{1m}$ $\longrightarrow \bar{x}_1$

replication 2: $x_{21}, x_{22}, \dots, x_{2m}$ $\longrightarrow \bar{x}_2$

....

replication n: $x_{n1}, x_{n2}, \dots, x_{nm}$ $\longrightarrow \bar{x}_n$

- This gives a set of independent observations (sample means)

- These sample means are treated as a sample of independent observations, from where a super mean can be calculated and its confidence interval.
- We have:

$$\bar{\bar{x}} = \frac{1}{n} \sum_{i=1}^n \bar{x}_i \quad s^2 = \frac{1}{n-1} \sum_{i=1}^n (\bar{x}_i - \bar{\bar{x}})^2$$

- Typically n is very small (less than 30), and we use the t-student statistic to construct the confidence interval:

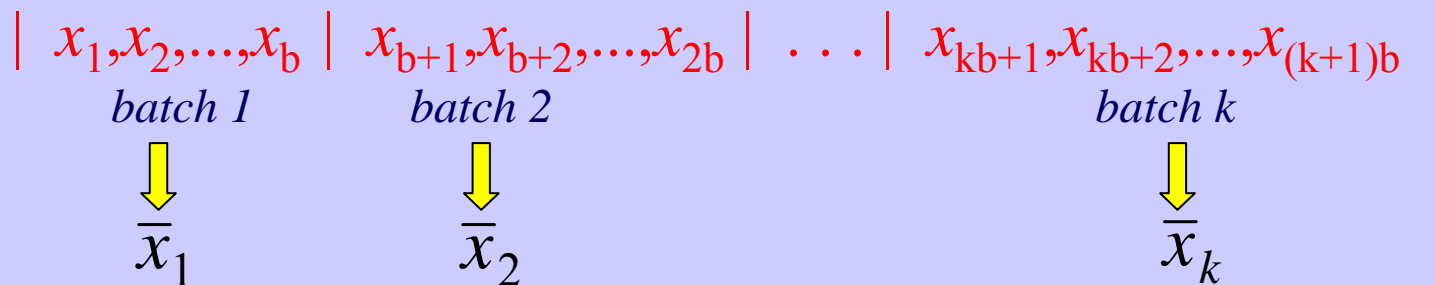
$$\left(\bar{\bar{x}} - t_{0.95} \frac{s}{\sqrt{n}}, \bar{\bar{x}} + t_{0.95} \frac{s}{n} \right)$$

Procedure

1. Run the simulation for an initial warm-up period.
2. Ignore collected statistics.
3. Then run for a long period to collect the first set of observations.
4. Change seed and repeat above steps, for a small number of times (definitely, less than 30)

The Batch means method

- This is a fairly popular technique and very easy to implement.
- Simulation is run for a long period and the successive observations are grouped into k ($k \geq 30$) *batches* of b observations, after allowing for an initial warm-up period:



- These sample means are treated as a sample of independent observations, from where a super mean can be calculated and its confidence interval.
- We have:

$$\bar{\bar{x}} = \frac{1}{k} \sum_{i=1}^k \bar{x}_i$$

$$s^2 = \frac{1}{k-1} \sum_{i=1}^k (\bar{x}_i - \bar{\bar{x}})^2$$

- The confidence interval is:

$$\left(\bar{\bar{x}} - 1.96 \frac{s}{\sqrt{k}}, \bar{\bar{x}} + 1.96 \frac{s}{\sqrt{k}} \right)$$

How long should a batch be?

- The batch size b has to be long enough so that the observations in one batch do not affect those in the adjacent batch (though some observations near the end of one batch will be correlated with some observations in the next batch).
- The batch size can be estimated by doing a *correlogram*.
- The batch size is a lot smaller than the simulation run in the case of replications.

Correlation of two variables

- Let X and Y be two variables with mean and variance: $(\mu_X, \sigma_X^2), (\mu_Y, \sigma_Y^2)$

- The *covariance* $\text{Cov}(X, Y)$ is defined as:

$$\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

- It reflects the dependency between X and Y .
 - If uncorrelated, then $\text{Cov}(X, Y) = 0$
 - If negatively correlated, then $\text{Cov}(X, Y) < 0$
 - If positively correlated, then $\text{Cov}(X, Y) > 0$

- The $\text{Cov}(X, Y)$ takes values in $(-\infty, +\infty)$. Also, it is not dimensionless.
- The *correlation* ρ_{XY} is given by the expression:

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

- It is dimensionless, ranges in $(-1, +1)$, and it is used as a measure of dependency.
 - If ρ_{XY} is close to $+1$ (-1), then X and Y are highly positively (negatively) correlated
 - If ρ_{XY} is close to -1 , then X and Y are highly negatively correlated

Autocorrelation

- Let us assume that we have n observations:

$$x_1, x_2, \dots, x_n$$

- We can form the following $n-1$ pairs:

$$(x_1, x_2), (x_2, x_3), \dots, (x_{n-1}, x_n)$$

- We regard the first observation in each pair as coming from a variable X , and the second observation as coming from variable Y .
- Then, we can calculate the correlation ρ_{XY} , known as the *autocorrelation lag 1*.

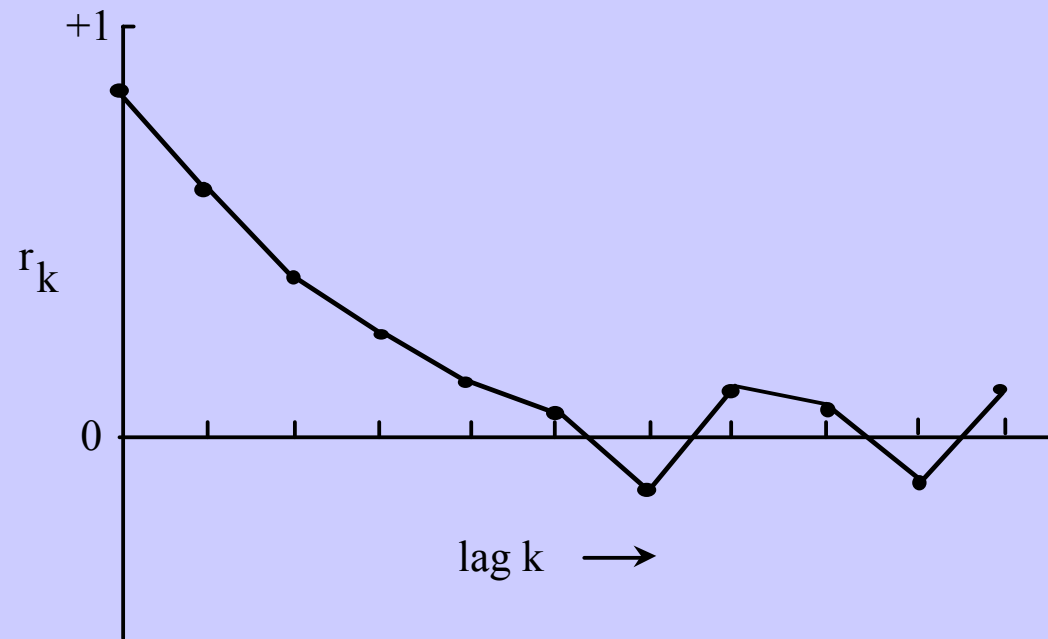
- Autocorrelation lag 1:

$$r_1 = \frac{\sum_{i=1}^{n-1} (x_i - \bar{X})(x_{i+1} - \bar{Y})}{\sum_{i=1}^{n-1} (x_i - \bar{X})^2} \quad \bar{X} = \frac{1}{n} \sum_{i=1}^n x_i$$

- Autocorrelation of lag k

$$r_k = \frac{\sum_{i=1}^{n-k} (x_i - \bar{X})(x_{i+k} - \bar{Y})}{\sum_{i=1}^{n-1} (x_i - \bar{X})^2}$$

Correlogram



It is a plot of autocorrelations against lag k

Fixing the batch size..

- Typically, the batch size b has to be large enough so that the successive batch means are not correlated.
- An estimate of b can be obtained by plotting out the correlogram from a preliminary simulation run.
- b is fixed to 5 times b' for which $r_{b'}$ is approximately zero