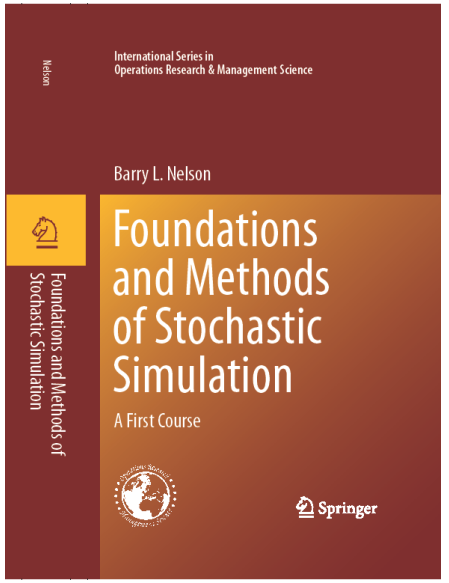


Chapter 9:
Simulation for Research

©Barry L. Nelson

Northwestern University

March 2013

Simulation as a research tool

- Simulation is often used as a research tool, even when the research has nothing to do with simulation.
  - Generating random test problems for optimization algorithms.
  - Testing the robustness of policies derived from simple stochastic models on complex systems.
  - Evaluating the accuracy of an approximation (e.g., queueing).
  - Establishing the properties of a new simulation-based estimator.
- Such studies are often done poorly, jeopardizing the publishability of otherwise good ideas.

Practitioner vs. Researcher

The *practitioner's experiment* solves a real problem, has a practical limit on time and effort, and leads to a decision.

Q: How much production capacity should we add?

The practitioner never knows for sure whether they got the "right answer."

The *researcher's experiment* is driven by a precisely formulated research question.

Q: Which optimization heuristic is better?

The researcher may repeatedly solve many problem instances, including ones for which they already know the answer.

Research experiments vs. illustrations

A *research experiment* should allow you to make statements about cases, scenarios or problems that you *did not* try based on ones you did try.

Ex: Optimization problems over a range of numbers of decision variables, numbers of constraints, tightness of the RHS, etc.

An *illustration* is a specific case, scenario or problem that helps us understand how a method is implemented and the results interpreted.

Ex: $M(t)/M/\infty$, $M/G/1$, *SAN and Asian option.*

Research Principle 1. *Completely random cases, scenarios or problems are not necessarily relevant ones. Instead, generate test cases, scenarios or problems that have features that are representative of the space of interest.*

Example: Hill and Reilly (2000) used randomly generated test problems to compare the speed and quality of solution of a heuristic and a convergent algorithm for solving two-dimensional knapsack problems:

$$\max \sum_{j=1}^{n} c_j x_j$$

$$\sum_{j=1}^{n} a_{ij} x_j \leq b_i, \; i = 1, 2$$

$$x_j \in \{0, 1\}, \; j = 1, 2, \ldots, n$$

where $c_i > 0$ and $a_{ij} \geq 0$.

Standard test problems

Fix the number of items at n and set b_i to be a specified fraction of $\sum_{j=1}^{n} a_{ij}$.

Need to generate test values of c_i and a_{ij}.

Select uniform distributions for c_j, a_{1j} and a_{2j}, then randomly and independently generate the necessary values.

These "completely random test problems" are certainly random, and seem fair, right?

Why not the "standard approach?"

These problems are neither realistic nor interesting.

- Positive correlation between a_{ij} and c_j implies that if the jth item is valuable then it tends to be costly, which makes the problem realistic and hard.
- Negative correlation between a_{1j} and a_{2j} means an item that is inexpensive relative to b_1 tends to be expensive relative to budget b_2; this also makes the problem realistic and hard.
- Turns out correlation structure affects the heuristic and convergent algorithms differently, which might be important in selecting which one to use.

We would see few of these realistic and interesting problems if generated "completely randomly."

Including correlation in the design

Hill and Reilly systematically controled the strengths of the correlations in a designed experiment.

This allowed them to make stronger conclusions than they could by simply averaging over the space of "completely random" test problems.

They also made sure the heuristic and convergent algorithm saw *exactly the same test problems.* This illustrates...

Research Principle 2. *If you are comparing things, use common random numbers.*

Random & relevant test problems

Test problems are often needed in our research.

Random $\neq$ Relevant

1. Think carefully about the space of real problems over which you want to draw inference.
2. Create test problems that randomly cover this relevant space.
3. Apply each method to the *same* test problems.
4. Look at a lot of test problems, especially if the space is high dimensional.

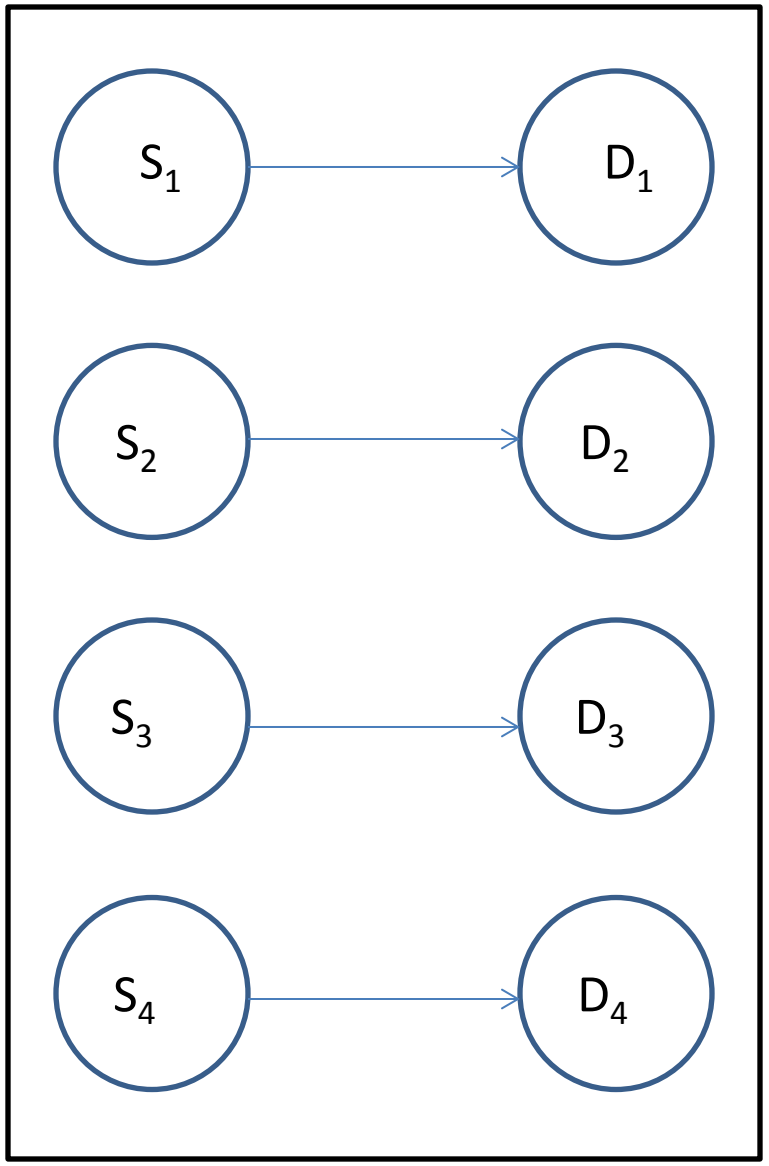
Research Principle 3. *Run a designed experiment that identifies and controls factors that might influence the outcome, both favorably and unfavorably, and that are actually encountered in practice.*

Iravani, Van Oyen and Sims (2005) wanted to create a simple measure of *structural flexibility (SF)*: the capability to satisfy multiple types of demand in the face of changing demand and resource capacity.

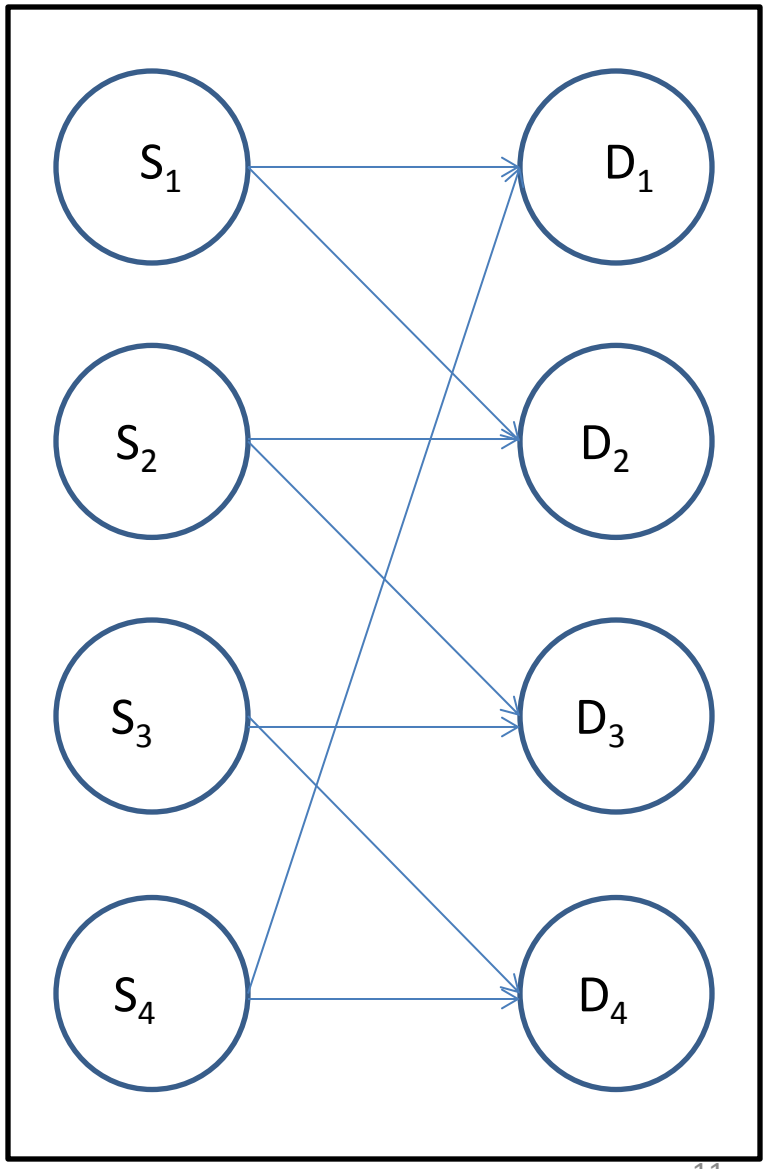
They represented a system as a graph from resources S_k to demand types D_j with an arc $S_k \rightarrow D_j$ if resource k can satisfy some demand of type j.

SF indices are functions of the number of paths through the network by which excess capacity for demand type i can be redirected to satisfy demand type j, and also the number of different resources that can satisfy demand of type j.

One factory to 1 car model



Two factories to each car model



Research question: Is a simple metric based only on structure really valid?

Approach: Choose pairs of system designs and compute their SF indices. Simulate the pairs to see if higher SF index predicts higher simulated productivity.

Key is designing an experiment that varies factors that are *not* inputs to the SF index, but do represent what occurs in the real world.

- Physical flow in the system (open or closed)
- Closeness of the system to capacity
- Uncertainty due to short-term variability and long-term shocks

Research Principle 4. *When there are no natural ranges of operability for your factors, link them so that they are large or small relative to each other.*

Example: Nelson, Swann, Goldsman and Song (2001) wanted to compare the computational efficiency of some ranking & selection (simulation optimization) procedures designed for up to $K = 500$ systems.

In addition to K, other factors that might affect efficiency include the true means $\mu_1, \ldots, \mu_K$, the true variances $\sigma_1^2, \ldots, \sigma_K^2$, the first-stage sample size n_0 and the indifference zone parameter δ.

How should these factors be set when there is no natural range of operability?

Linking factors

The key is to vary factors *relative to each other*.

- Anchor the means at $\mu_1 = \delta$ (bigger better).

  SK: $\mu_2 = \cdots = \mu_K = 0$

  MDM: $\mu_i = \mu_1 - (i-1)\delta/\tau$ with $\tau = 1, 2, 3$

- Anchor the variances to a common σ^2 or to the means.

  Common: $\sigma_2^2 = \cdots = \sigma_K^2 = \sigma^2$ and $\sigma_1^2 = \rho\sigma^2$ with $\rho = 1/2, 2$.

  Unequal: $\sigma_i = |\mu_i - \delta| + 1$ or $\sigma_i = 1/(|\mu_i - \delta| + 1)$.

- Set $\delta = d\sigma_1/\sqrt{n_0}$ with $d = 1/2, 1, 2$.

If we set $\sigma_1^2 = 1$, then there are only two remaining factors K and n_0 for which we can set reasonable ranges.

To be able to control all of these factors, Nelson et al. used normal and lognormal distributions to generate the simulation output data.

This might cause them to miss something, since real simulations have more complicated output processes.

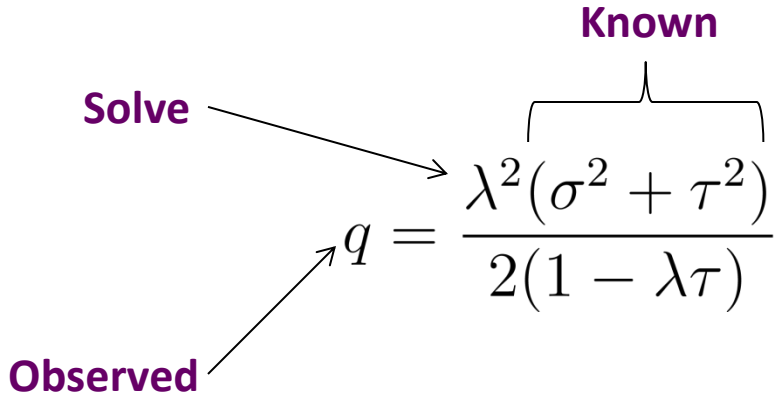
For this reason, we also want to include somewhat more realistic test cases, even though we have less control: Markovian queues, simple inventory models, AR(1), MA(1), SAN, Asian option, etc.

Research Principle 5. *Since it may not be possible to anticipate all important factors, include some realistic examples along with the controllable surrogate models.*

Research Principle 6. *If you reduce the standard error or confidence interval width enough, then the simulation estimate is effectively the same as the true value.*

Example: Whitt (1981) considered the situation where you can observe the congestion in a queue, and you know the service process, but you cannot observe the arrival process. Your goal is to predict what would happen to congestion if you replaced the current service process by a different one.

Simple example: $M/G/1$



Testing a useful approximation

Whitt wanted his method to work on queueing systems with very general, stationary arrival processes (meaning there could be dependence). Ex: $G/M/1$

He wanted to approximate the unknown arrival process G with a renewal process GI that yields a tractable queueing model. Ex: $GI/M/1$

A "good" GI means that the predicted results with a new service process should be right. Ex: $GI/M'/1 \approx G/M'/1$

Since he was using tractable queueing models as his approximations, he needed to use test cases that were *not* tractable (i.e., no known answer).

Using simulation to get "the truth"

Whitt considered his approximation good enough if

$$\frac{|q_{\text{approx}} - q_{\text{true}}|}{q_{\text{true}}} \leq 0.1$$

We do not know q_{true}, but simulation provides an estimate $\widehat{q}_{\text{true}} \pm H$.

The approximation is good enough if the relative error ratio holds *for every possible value* of q_{true} such that

$$\widehat{q}_{\text{true}} - H \leq q_{\text{true}} \leq \widehat{q}_{\text{true}} + H$$

We simulate until the $\pm H$ is small enough to know for sure.

Research Principle 7. *Measure and control the error in your research experiment using nested simulations.*

Example: A control-variate estimator $\widehat{\beta}_0$ is an alternative to the sample mean $\bar{Y}$ for estimating $\mu_Y = \mathrm{E}(Y)$ in a simulation.

As the number of replications goes to infinity, control-variate estimators are consistent and have smaller variance than $\bar{Y}$. In finite samples they may be biased and could have larger variance.

Nelson (1990) was interested in establishing properties of control-variate estimators, including how they perform in small samples.

Empirical evaluation

Do a designed experiment over relevant factors.

For each factor setting make m macroreplications—i.i.d. replications of the experiment—to estimate bias, variance and CI coverage. This is a nested experiment.

$$\begin{aligned}
\widehat{b} &= \frac{1}{m}\sum_{i=1}^{m}\widehat{\beta}_0^{(i)} - \mu_Y = \bar{\beta}_0 - \mu_Y \\
S^2_{\widehat{\beta}_0} &= \frac{1}{m-1}\sum_{i=1}^{m}\left(\widehat{\beta}_0^{(i)} - \bar{\beta}_0\right)^2 \\
\widehat{p} &= \frac{1}{m}\sum_{i=1}^{m} I\left\{\mu_Y \in \widehat{\beta}_0^{(i)} \pm t_{1-\alpha/2,n-2}\sqrt{\widehat{\Sigma}_{11}^{(i)}}\right\}.
\end{aligned}$$

m should not be chosen arbitrarily.

Number of macroreplications

Suppose that we set $1-\alpha = 0.95$, meaning 95% confidence intervals.

If the confidence interval has the desired coverage, then the standard error of $\widehat{p}$ is $\sqrt{(0.95)(0.05)/m}$.

To get approximately two decimal places of precision we need the number of macroreps m to satisfy

$$2\sqrt{\frac{(0.95)(0.05)}{m}} < 0.01$$

or $m \approx 1900$.

Macroreps can provide the "truth"

The control-variate variance estimator $\widehat{\Sigma}_{11}$ is expected to be biased, but we do not actually know what the true $\mathrm{Var}(\widehat{\beta}_0)$ is in small samples.

Note that $S^2_{\widehat{\beta}_0}$ is an unbiased estimator of $\mathrm{Var}(\widehat{\beta}_0)$, since it is based on i.i.d. observations of $\widehat{\beta}_0$.

An estimator of the bias of $\widehat{\Sigma}_{11}$ is therefore

$$\frac{1}{m}\sum_{i=1}^{m}\widehat{\Sigma}_{11}^{(i)} - S^2_{\widehat{\beta}_0}$$