Output Analysis

R.B. Lenin
(rblenin@daiict.ac.in)

Autumn 2007

Introduction

- We will consider the statistical aspects of a simulation study.
- We have to realize that doing a simulation is nothing else than doing an experiment.
- The outcome of a simulation will be, like the outcome of an experiment, an **estimator** of the performance measure of interest.
- Hence we also say something about the quality of the estimator, for example, by constructing a **confidence interval** of the performance measure of interest.
- First we describe how to construct an estimator and a confidence interval for the unknown mean of a random variable, given that we have a number of independent realizations of the random variable.

Introduction · · ·

- After that, we will study how we can use these results in the output analysis of a simulation study.
- A simulation study is usually undertaken to determine the value of some quantity θ connected with a particular stochastic model.
- A simulation of the relevant system results in the output data X, a random variable whose expected value is the quantity of interest θ.
- A second independent simulation – that is, a second simulation run – provides a new and independent random variable having mean θ.
- This continues until we have a total of k runs – and the k independent random variables $X_1, \ldots, X_k$ – all of which are identically distributed with mean θ.

Introduction · · ·

- The average of these k values

$$\bar{X} = \frac{1}{k} \sum_{i=1}^{k} X_i$$

  is then used as an **estimator**, or **approximator** of θ.
- We will consider the problem of deciding when to stop the simulation study. That is, deciding on the appropriate value of k.
- To help us decide when to stop, we will find it useful to consider the quality of our estimator of θ.
- In addition, we will also show how to obtain an interval in which we can assert that θ lies, with a certain degree of confidence.

Unbiased sample mean and sample variance

- Let $X_1, \ldots, X_n$ be independent random variables having the same distribution function.
- Let θ and σ^2 denote their mean and variance, respectively. That is,

$$\theta = E[X_i] \qquad \text{and} \qquad \sigma^2 = \text{Var}[X_i], \quad i = 1, 2 \ldots, n.$$

- Suppose that $X_1, \ldots, X_n$ are output data of a simulation of a system. For example, X_i may denote W_s in simulating a queueing model in the i^{th} run.
- Then $\bar{X} := \frac{1}{n} \sum_{i=1}^{n} X_i$ is known as the **unbiased sample mean**.

Unbiased sample mean and sample variance $\cdots$

- *Since a random variable is unlikely to be too many standard deviations (square root of variance) from its mean, it follows that* $\bar{X}$ *is a good estimator of* θ *when the standard deviation* $\frac{\sigma}{\sqrt{n}}$ *is small.*
- **But** σ **is unknown**.
- Define $S^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ - called the **unbiased sample variance**.

Unbiased sample mean and sample variance $\cdots$

- Note that $\bar{X}$ is a random variable with mean θ and variance $\frac{\sigma}{\sqrt{n}}$.
- For sufficiently large n, as will usually the case in simulations (number of runs), we can apply the central limit theorem to assert that

$$\frac{\bar{X} - \theta}{\sigma/\sqrt{n}}$$

  is approximately distributed as a standard normal random variable.
- We use S as an estimate of σ.

$$\therefore \frac{\bar{X} - \theta}{S/\sqrt{n}}$$

  is approximately $N(0, 1)$ for large n.

Unbiased sample mean and sample variance $\cdots$

- If $Z \sim N(0,1)$, then

$$\begin{aligned} Pr\{|Z| < c\} &= Pr\{-c < Z < c\} = Pr\{Z > -c\} - Pr\{Z > c\} \\ &= 1 - Pr\{Z < -c\} - Pr\{Z > c\} \\ &= 1 - 2Pr\{Z > c\} \qquad (\because Pr\{Z < -c\} = Pr\{Z > c\}) \\ &= 1 - 2(1 - Pr\{Z < c\}) = 2Pr\{Z < c\} - 1 \\ &= 2F_Z(c) - 1, \qquad (F_Z(x) - \text{ standard normal dist. fun..}) \end{aligned}$$

- Therefore,

$$\begin{aligned} Pr\left\{\left|\frac{\bar{X} - \theta}{S/\sqrt{n}}\right| < c\right\} &\approx 2F_z(c) - 1 \\ \Rightarrow Pr\left\{|\bar{X} - \theta| < c\frac{S}{\sqrt{n}}\right\} &\approx 2F_z(c) - 1. \qquad (1) \end{aligned}$$

Unbiased sample mean and sample variance $\cdots$

- Equation (1) states that the probability that the sample mean differs from θ by less than $c\frac{S}{\sqrt{n}}$ is approximately $2F_z(c) - 1$.
- For example, from the normal table we know $F_z(1.96) = 0.975$.

$$\therefore Pr\left\{|\bar{X} - \theta| < 1.96\frac{S}{\sqrt{n}}\right\} \approx 2 \times 0.975 - 1 = 0.95.$$

- That is, 95% certain that our estimated answer will not differ from the theoretical (true) value θ by more than $1.96\frac{S}{\sqrt{n}}$.

Unbiased sample mean and sample variance $\cdots$

- From the normal table we have $F_z(2.58) \approx 0.995$.

$$\therefore Pr\left\{|\bar{X} - \theta| < 2.58\frac{S}{\sqrt{n}}\right\} \approx 2 \times 0.995 - 1 = 0.99.$$

- That is, 99% certain that our estimated answer will not differ from the theoretical value by more than $2.58\frac{S}{\sqrt{n}}$.
- In simulation we continuously generate additional data values X_i.
- If our objective is to estimate $\theta = E[X_i]$, then when to stop generating new data values? that is, how many times do we need to repeat the simulation?

Unbiased sample mean and sample variance $\cdots$

- We should first choose an acceptable value d for the absolute difference between the estimator $\bar{X}$ and the true value θ.
- Suppose we want to be, for example, at least 95% certain that $\bar{X}$ will not differ from θ by not more than d.
- Then we should continue to generate new data until we have generated n data values for which our estimate $\frac{S}{\sqrt{n}}$ of $\frac{\sigma}{\sqrt{n}}$ is such that

$$1.96\frac{S}{\sqrt{n}} \leq d.$$

A method for when to stop generating new data

1. Choose an acceptable value d for the absolute difference between the estimator $\bar{X}$ and the true value θ.
2. Generate at least 100 data values.
3. Continue to generate additional data values, stopping when you have generated k values ($k > 100$) and $c\frac{S}{\sqrt{k}} \leq d$, where S is the sample standard deviation based on the k values.
4. The estimate of θ is then given by

$$\bar{X} = \frac{1}{k}\sum_{i=1}^{k} X_i.$$

A method for when to stop generating new data · · ·

Example

- Consider a service system where no new customers are allowed to enter after 5PM.
- Suppose that each day follows the same probability law and that we are interested in estimating the expected time at which the last customer departs the system.
- Furthermore, suppose we want to be at least 95% certain that our estimated answer will not differ from the true value by more than 15 seconds.

A method for when to stop generating new data $\cdots$

Example (Continues $\cdots$)

- To satisfy the above requirement it is necessary that we continuously generate data values relating to the time at which the last customer departs, by repeating the simulation run, for different seeds, until we have generated a total of k ($k \geq 100$) values and is such that

$$1.96\frac{S}{\sqrt{k}} \leq 15,$$

  where S is the sample standard deviation (measured in seconds) of these k data values.
- For 99% certain, we have to repeat the k ($k \geq 100$) times such that

$$2.58\frac{S}{\sqrt{k}} \leq 15.$$

- Our estimate of the expected time at which the last customer departs will be the average of the k data values $\frac{1}{k}\sum_{i=1}^{k} X_i$.

Computing $\bar{X}$ and S^2 Recursively

- Define

$$\bar{X}_j := \frac{1}{j}\sum_{i=1}^{j} X_i, \qquad S_j^2 := \frac{1}{j-1}\sum_{i=1}^{j}(X_i - \bar{X}_j)^2, \qquad j \geq 2.$$

  with $S_1^2 = 0$ and $\bar{X}_0 = 0$.
- Now,

$$\begin{aligned}
\bar{X}_{j+1} &= \frac{1}{j+1}\sum_{i=1}^{j+1} X_i = \frac{1}{j+1}\sum_{i=1}^{j} X_i + \frac{X_{j+1}}{j+1} \\
&= \frac{j}{j+1}\cdot\frac{1}{j}\sum_{i=1}^{j} X_i + \frac{X_{j+1}}{j+1} = \frac{j}{j+1}\bar{X}_j + \frac{X_{j+1}}{j+1} \\
&= \frac{j\bar{X}_j + X_{j+1}}{j+1} = \bar{X}_j + \frac{X_{j+1} - \bar{X}_j}{j+1}, \qquad j \geq 2.
\end{aligned}$$

- Similarly, using the above result, we get

$$S_{j+1}^2 = \left(1 - \frac{1}{j}\right) S_j^2 + (j+1)\left(\bar{X}_{j+1} - \bar{X}_j\right)^2.$$

A method for when to stop generating new data $\cdots$

- Suppose we are interested in simulating a Bernoulli random variable X, such that

$$X_i = \begin{cases} 1, & \text{with probability } p \\ 0, & \text{with probability } 1-p \end{cases}$$

  and suppose we are interested in estimating $E[X_i] = p$.
- Since in this situation, the formula for variance is known in terms of $\bar{X}$, there is no need to utilize the sample variance S^2 to estimate $\text{Var}[X_i]$.
- If we have generated n values $X_1, \ldots, X_n$, then as the estimate of p will be

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- A natural estimate of $\text{Var}[X_i]$ is

$$\bar{X}_n(1-\bar{X}_n).$$

A method for when to stop generating new data ···

- Therefore in this case we have the following method for deciding when to stop:
  1. Choose an acceptable value d for the absolute difference between $\bar{X}$ and p.
  2. Generate k ($k \geq 100$) data values $X_1, \ldots, X_k$ until

$$c\sqrt{\frac{\bar{X}_k(1-\bar{X}_k)}{k}} < d.$$

  3. The estimate of p is then given by

$$\bar{X}_k = \frac{1}{k}\sum_{i=1}^{k} X_i.$$

A method for when to stop generating new data $\cdots$

Example

- Suppose in example 1 we were interested in estimating the probability that there was still a customer in the store at 5:30PM.
- To do so, we would simulate successive days and let

$$X_i = \begin{cases} 1, & \text{if there is a customer present at 5:30PM on day } i \\ 0, & \text{otherwise} \end{cases}$$

- We will have to repeat the simulation until the k^{th} day ($k \geq 100$), where k is such that

$$c\sqrt{\frac{\bar{X}_k(1-\bar{X}_k)}{k}} < d.$$

where d is an acceptable value for the absolute difference between $\bar{X}$ and θ.

A method for when to stop generating new data $\cdots$

- Suppose we want to be at least 95% certain that our estimated answer $\bar{X}_k$ will not differ from the true value by more than 0.01.
  - Then we keep repeating the simulation run until k times, where k is such that

$$1.96\sqrt{\frac{\bar{X}_k(1-\bar{X}_k)}{k}} < 0.01.$$

- Suppose we want to be at least 99% certain that our estimated answer $\bar{X}_k$ will not differ from the true value by more than 0.01.
  - Then we repeat the simulation until k times, where k is such that

$$2.58\sqrt{\frac{\bar{X}_k(1-\bar{X}_k)}{k}} < 0.01.$$

- **Note that the quantity $\frac{S}{\sqrt{n}}$ is a decreasing sequence as n increases.**