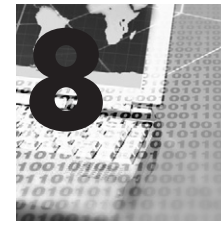
Statistical Analysis of Simulated Data



Introduction

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 amassed 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. The average of these k values, $\overline{X} = \sum_{i=1}^{k} X_i/k$, is then used as an estimator, or approximator, of θ.

In this chapter we 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.

The final section of this chapter shows how we can estimate the quality of more complicated estimators than the sample mean—by using an important statistical technique known as "bootstrap estimators."

8.1 The Sample Mean and Sample Variance

Suppose that $X_1, \ldots, X_n$ are independent random variables having the same distribution function. Let θ and σ^2 denote, respectively, their mean and

Simulation. DOI: http://dx.doi.org/10.1016/B978-0-12-415825-2.00008-5

variance—that is, $\theta = E[X_i]$ and $\sigma^2 = \text{Var}(X_i)$. The quantity

$$\overline{X} \equiv \sum_{i=1}^{n} \frac{X_i}{n}$$

which is the arithmetic average of the n data values, is called the *sample mean*. When the population mean θ is unknown, the sample mean is often used to estimate it.

Because

$$\begin{aligned} E[\overline{X}] &= E\left[\sum_{i=1}^{n} \frac{X_i}{n}\right] \\ &= \sum_{i=1}^{n} \frac{E[X_i]}{n} \\ &= \frac{n\theta}{n} = \theta \end{aligned} \tag{8.1}$$

it follows that $\overline{X}$ is an unbiased estimator of θ, where we say that an estimator of a parameter is an unbiased estimator of that parameter if its expected value is equal to the parameter.

To determine the "worth" of $\overline{X}$ as an estimator of the population mean θ, we consider its mean square error—that is, the expected value of the squared difference between $\overline{X}$ and θ. Now

$$\begin{aligned} E[(\overline{X} - \theta)^2] &= \text{Var}(\overline{X}) \quad (\text{since } E[\overline{X}] = \theta) \\ &= \text{Var}\left(\frac{1}{n}\sum_{1}^{n} X_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} \text{Var}(X_i) \quad (\text{by independence}) \\ &= \frac{\sigma^2}{n} \quad (\text{since } \text{Var}(X_i) = \sigma^2) \end{aligned} \tag{8.2}$$

Thus, $\overline{X}$, the sample mean of the n data values $X_1, \ldots, X_n$, is a random variable with mean θ and variance σ^2/n. Because a random variable is unlikely to be too many standard deviations—equal to the square root of its variance—from its mean, it follows that $\overline{X}$ is a good estimator of θ when $\sigma/\sqrt{n}$ is small.

Remark The justification for the above statement that a random variable is unlikely to be too many standard deviations away from its mean follows from both the Chebyshev inequality and, more importantly for simulation studies, from the

central limit theorem. Indeed, for any $c > 0$, Chebyshev's inequality (see Section 2.7 of Chapter 2) yields the rather conservative bound

$$P\left\{|\overline{X} - \theta| > \frac{c\sigma}{\sqrt{n}}\right\} \leqslant \frac{1}{c^2}$$

However, when n is large, as will usually be the case in simulations, we can apply the central limit theorem to assert that $(\overline{X} - \theta)/(\sigma/\sqrt{n})$ is approximately distributed as a standard normal random variable; and thus

$$\begin{aligned} P\{|\overline{X} - \theta| > c\sigma/\sqrt{n}\} &\approx P\{|Z| > c\}, \quad \text{where Z is a standard normal} \\ &= 2[1 - \Phi(c)] \end{aligned} \tag{8.3}$$

where Φ is the standard normal distribution function. For example, since $\Phi(1.96) = 0.975$, Equation (8.3) states that the probability that the sample mean differs from θ by more than $1.96\sigma/\sqrt{n}$ is approximately 0.05, whereas the weaker Chebyshev inequality only yields that this probability is less than $1/(1.96)^2 = 0.2603$. □

The difficulty with directly using the value of σ^2/n as an indication of how well the sample mean of n data values estimates the population mean is that the population variance σ^2 is not usually known. Thus, we also need to estimate it. Since

$$\sigma^2 = E[(X - \theta)^2]$$

is the average of the square of the difference between a datum value and its (unknown) mean, it might seem upon using $\overline{X}$ as the estimator of the mean that a natural estimator of σ^2 would be $\sum_{i=1}^{n} (X_i - \overline{X})^2/n$, the average of the squared distances between the data values and the estimated mean. However, to make the estimator unbiased (and for other technical reasons) we prefer to divide the sum of squares by $n - 1$ rather than n.

Definition *The quantity S^2, defined by*

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n - 1}$$

is called the sample variance.

Using the algebraic identity

$$\sum_{i=1}^{n} (X_i - \overline{X})^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2 \tag{8.4}$$

whose proof is left as an exercise, we now show that the sample variance is an unbiased estimator of σ^2.

Proposition

$$E[S^2] = \sigma^2$$

Proof Using the identity (8.4) we see that

$$\begin{aligned}(n-1)E[S^2] &= E\left[\sum_{i=1}^{n} X_i^2\right] - nE[\overline{X}^2] \\ &= nE\left[X_1^2\right] - nE[\overline{X}^2] \qquad (8.5)\end{aligned}$$

where the last equality follows since the X_i all have the same distribution. Recalling that for any random variable Y, $\mathrm{Var}(Y) = E[Y^2] - (E[Y])^2$ or, equivalently,

$$E[Y^2] = \mathrm{Var}(Y) + (E[Y])^2$$

we obtain that

$$\begin{aligned}E\left[X_1^2\right] &= \mathrm{Var}(X_1) + (E[X_1])^2 \\ &= \sigma^2 + \theta^2\end{aligned}$$

and

$$\begin{aligned}E[\overline{X}^2] &= \mathrm{Var}(\overline{X}) + (E[\overline{X}])^2 \\ &= \frac{\sigma^2}{n} + \theta^2 \quad [\text{from}(8.2)\text{and}(8.1)]\end{aligned}$$

Thus, from Equation (8.5), we obtain that

$$(n-1)E[S^2] = n(\sigma^2 + \theta^2) - n\left(\frac{\sigma^2}{n} + \theta^2\right) = (n-1)\sigma^2$$

which proves the result. □

We use the sample variance S^2 as our estimator of the population variance σ^2, and we use $S = \sqrt{S^2}$, the so-called sample standard deviation, as our estimator of σ.

Suppose now that, as in a simulation, we have the option of continually generating additional data values X_i. If our objective is to estimate the value of $\theta = E[X_i]$, when should we stop generating new data values? The answer to this question is that we should first choose an acceptable value d for the standard deviation of our estimator—for if d is the standard deviation of the estimator $\overline{X}$, then we can, for example, be 95% certain that $\overline{X}$ will not differ from θ by more than $1.96d$. We should then continue to generate new data until we have generated n data values for which our estimate of $\sigma/\sqrt{n}$—namely, $S/\sqrt{n}$—is less than the acceptable value d. Since the sample standard deviation S may not be a particularly

good estimate of σ (nor may the normal approximation be valid) when the sample size is small, we thus recommend the following procedure to determine when to stop generating new data values.

A Method for Determining When to Stop Generating New Data

1. Choose an acceptable value d for the standard deviation of the estimator.
2. Generate at least 100 data values.
3. Continue to generate additional data values, stopping when you have generated k values and $S/\sqrt{k} < d$, where S is the sample standard deviation based on those k values.
4. The estimate of θ is given by $\overline{X} = \sum_{i=1}^{k} X_i/k$.

Example 8a Consider a service system in which no new customers are allowed to enter after 5 P.M. 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.

To satisfy the above requirement it is necessary that we continually generate data values relating to the time at which the last customer departs (each time by doing a simulation run) until we have generated a total of k values, where k is at least 100 and is such that $1.96S/\sqrt{k} < 15$—where S is the sample standard deviation (measured in seconds) of these k data values. Our estimate of the expected time at which the last customer departs will be the average of the k data values. □

In order to use the above technique for determining when to stop generating new values, it would be valuable if we had a method for recursively computing the successive sample means and sample variances, rather than having to recompute from scratch each timea new datum value is generated. We now show how this can be done. Consider the sequence of data values $X_1, X_2, \ldots$, and let

$$\overline{X}_j = \sum_{i=1}^{j} \frac{X_i}{j}$$

and

$$S_j^2 = \sum_{i=1}^{j} \frac{(X_i - \overline{X}_j)^2}{j-1}, \quad j \geqslant 2$$

denote, respectively, the sample mean and sample variance of the first j data values. The following recursion should be used to successively compute the current value of the sample mean and sample variance.

With $S_1^2 = 0$, $\overline{X}_0 = 0$,

$$\overline{X}_{j+1} = \overline{X}_j + \frac{X_{j+1} - \overline{X}_j}{j+1} \tag{8.6}$$

$$S_{j+1}^2 = \left(1 - \frac{1}{j}\right) S_j^2 + (j+1)(\overline{X}_{j+1} - \overline{X}_j)^2 \tag{8.7}$$

Example 8b If the first three data values are $X_1 = 5$, $X_2 = 14$, $X_3 = 9$, then Equations (8.6) and (8.7) yield that

$$\begin{aligned}
\overline{X}_1 &= 5 \\
\overline{X}_2 &= 5 + \frac{9}{2} = \frac{19}{2} \\
S_2^2 &= 2\left(\frac{19}{2} - 5\right)^2 = \frac{81}{2} \\
\overline{X}_3 &= \frac{19}{2} + \frac{1}{3}\left(9 - \frac{19}{2}\right) = \frac{28}{3} \\
S_3^2 &= \frac{81}{4} + 3\left(\frac{28}{3} - \frac{19}{2}\right)^2 = \frac{61}{3}
\end{aligned}$$

□

The analysis is somewhat modified when the data values are Bernoulli (or 0, 1) random variables, as is the case when we are estimating a probability. That is, suppose we can generate random variables 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,

$$\text{Var}(X_i) = p(1 - p)$$

there is no need to utilize the sample variance to estimate $\text{Var}(X_i)$. Indeed, if we have generated n values $X_1, \ldots, X_n$, then as the estimate of p will be

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

a natural estimate of $\text{Var}(X_i)$ is $\overline{X}_n(1 - \overline{X}_n)$. Hence, in this case, we have the following method for deciding when to stop.

1. Choose an acceptable value d for the standard deviation of the estimator.
2. Generate at least 100 data values.

3. Continue to generate additional data values, stopping when you have generated k values and $[\overline{X}_k(1-\overline{X}_k)/k]^{1/2} < d$.
4. The estimate of p is $\overline{X}_k$, the average of the k data values.

Example 8c Suppose, in Example 8a, we were interested in estimating the probability that there was still a customer in the store at 5:30. To do so, we would simulate successive days and let

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

We would simulate at least 100 days and continue to simulate until the kth day, where k is such that $[p_k(1-p_k)/k]^{1/2} < d$, where $p_k = \overline{X}_k$ is the proportion of these k days in which there is a customer present at 5:30 and where d is an acceptable value for the standard deviation of the estimator p_k. □

8.2 Interval Estimates of a Population Mean

Suppose again that $X_1, X_2, \ldots, X_n$ are independent random variables from a common distribution having mean θ and variance σ^2. Although the sample mean $\overline{X} = \sum_{i=1}^n X_i/n$ is an effective estimator of θ, we do not really expect that $\overline{X}$ will be equal to θ but rather that it will be "close." As a result, it is sometimes more valuable to be able to specify an interval for which we have a certain degree of confidence that θ lies within.

To obtain such an interval we need the (approximate) distribution of the estimator $\overline{X}$. To determine this, first recall, from Equations (8.1) and (8.2), that

$$E[\overline{X}] = \theta, \quad \text{Var}(\overline{X}) = \frac{\sigma^2}{n}$$

and thus, from the central limit theorem, it follows that for large n

$$\sqrt{n}\frac{(\overline{X}-\theta)}{\sigma} \dot{\sim} N(0, 1)$$

where $\dot{\sim} N(0, 1)$ means "is approximately distributed as a standard normal." In addition, if we replace the unknown standard deviation σ by its estimator S, the sample standard deviation, then it still remains the case (by a result known as Slutsky's theorem) that the resulting quantity is approximately a standard normal. That is, when n is large

$$\sqrt{n}(\overline{X}-\theta)/S \dot{\sim} N(0, 1) \tag{8.8}$$

Now for any α, $0 < \alpha < 1$, let z_α be such that

$$P\{Z > z_\alpha\} = \alpha$$

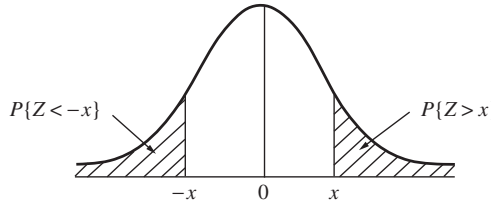


Figure 8.1. Standard normal density.

where Z is a standard normal random variable. (For example, $z_{.025} = 1.96$.) It follows from the symmetry of the standard normal density function about the origin that $z_{1-\alpha}$, the point at which the area under the density to its right is equal to $1 - \alpha$, is such that (see Figure 8.1)

$$z_{1-\alpha} = -z_\alpha$$

Therefore (see Figure 8.1)

$$P\{-z_{\alpha/2} < Z < z_{\alpha/2}\} = 1 - \alpha$$

It thus follows from (8.8) that

$$P\left\{-z_{\alpha/2} < \sqrt{n}\frac{(\overline{X} - \theta)}{S} < z_{\alpha/2}\right\} \approx 1 - \alpha$$

or, equivalently, upon multiplying by -1,

$$P\left\{-z_{\alpha/2} < \sqrt{n}\frac{(\theta - \overline{X})}{S} < z_{\alpha/2}\right\} \approx 1 - \alpha$$

which is equivalent to

$$P\left\{\overline{X} - z_{\alpha/2}\frac{S}{\sqrt{n}} < \theta < \overline{X} + z_{\alpha/2}\frac{S}{\sqrt{n}}\right\} \approx 1 - \alpha \tag{8.9}$$

In other words, with probability $1 - \alpha$ the population mean θ will lie within the region $\overline{X} \pm z_{\alpha/2}S/\sqrt{n}$.

Definition *If the observed values of the sample mean and the sample standard deviation are $\overline{X} = \overline{x}$ and $S = s$, call the interval $\overline{x} \pm z_{\alpha/2}s/\sqrt{n}$ an (approximate) $100(1 - \alpha)$ percent confidence interval estimate of θ.*

Remarks

1. To clarify the meaning of a "$100(1-\alpha)$ percent confidence interval," consider, for example, the case where $\alpha = 0.05$, and so $z_{\alpha/2} = 1.96$. Now before the data are observed, it will be true, with probability (approximately) equal to 0.95, that the sample mean $\overline{X}$ and the sample standard deviation S will be such that θ will lie between $\overline{X} \pm 1.96S/\sqrt{n}$. After $\overline{X}$ and S are observed to equal, respectively, $\overline{x}$ and s, there is no longer any probability concerning whether θ lies in the interval $\overline{x} \pm 1.96s/\sqrt{n}$, for either it does or it does not. However, we are "95% confident" that in this situation it does lie in this interval (because we know that over the long run such intervals will indeed contain the mean 95 percent of the time).
2. (A technical remark.) The above analysis is based on Equation (8.8), which states that $\sqrt{n}(\overline{X}-\theta)/S$ is approximately a standard normal random variable when n is large. Now if the original data values X_i were themselves normally distributed, then it is known that this quantity has (exactly) a t-distribution with $n-1$ degrees of freedom. For this reason, many authors have proposed using this approximate distribution in the general case where the original distribution need not be normal. However, since it is not clear that the t-distribution with $n-1$ degrees of freedom results in a better approximation than the normal in the general case, and because these two distributions are approximately equal for large n, we have used the normal approximation rather than introducing the t-random variable.

Consider now the case, as in a simulation study, where additional data values can be generated and the question is to determine when to stop generating new data values. One solution to this is to initially choose values α and l and to continue generating data until the approximate $100(1-\alpha)$ percent confidence interval estimate of θ is less than l. Since the length of this interval will be $2z_{\alpha/2}S/\sqrt{n}$ we can accomplish this by the following technique.

1. Generate at least 100 data values.
2. Continue to generate additional data values, stopping when the number of values you have generated—call it k—is such that $2z_{\alpha/2}S/\sqrt{k} < l$, where S is the sample standard deviation based on those k values. [The value of S should be constantly updated, using the recursion given by (8.6) and (8.7), as new data are generated.]
3. If $\overline{x}$ and s are the observed values of $\overline{X}$ and S, then the $100(1-\alpha)$ percent confidence interval estimate of θ, whose length is less than l, is $\overline{x} \pm z_{\alpha/2}s/\sqrt{k}$.

A Technical Remark The more statistically sophisticated reader might wonder about our use of an approximate confidence interval whose theory was based on the assumption that the sample size was fixed when in the above situation

the sample size is clearly a random variable depending on the data values generated. This, however, can be justified when the sample size is large, and so from the viewpoint of simulation we can safely ignore this subtlety. □

As noted in the previous section, the analysis is modified when $X_1, \ldots, X_n$ are Bernoulli random variables such that

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

Since in this case $\text{Var}(X_i)$ can be estimated by $\overline{X}(1 - \overline{X})$, it follows that the equivalent statement to Equation (8.8) is that when n is large

$$\sqrt{n}\frac{(\overline{X} - p)}{\sqrt{\overline{X}(1 - \overline{X})}} \dot{\sim} N(0, 1) \tag{8.10}$$

Hence, for any α,

$$P\left\{-z_{\alpha/2} < \sqrt{n}\frac{(\overline{X} - p)}{\sqrt{\overline{X}(1 - \overline{X})}} < z_{\alpha/2}\right\} = 1 - \alpha$$

or, equivalently,

$$P\left\{\overline{X} - z_{\alpha/2}\sqrt{\overline{X}(1 - \overline{X})/n} < p < \overline{X} + z_{\alpha/2}\sqrt{\overline{X}(1 - \overline{X})/n}\right\} = 1 - \alpha$$

Hence, if the observed value of $\overline{X}$ is p_n, we say that the "$100(1 - \alpha)$ percent confidence interval estimate" of p is

$$p_n \pm z_{\alpha/2}\sqrt{p_n(1 - p_n)/n}$$

8.3 The Bootstrapping Technique for Estimating Mean Square Errors

Suppose now that $X_1, \ldots, X_n$ are independent random variables having a common distribution function F, and suppose we are interested in using them to estimate some parameter $\theta(F)$ of the distribution F. For example, $\theta(F)$ could be (as in the previous sections of this chapter) the mean of F, or it could be the median or the variance of F, or any other parameter of F. Suppose further that an estimator of $\theta(F)$—call it $g(X_1, \ldots, X_n)$—has been proposed, and in order to judge its worth as an estimator of $\theta(F)$ we are interested in estimating its mean square error. That is, we are interested in estimating the value of

$$\text{MSE}(F) \equiv E_F[(g(X_1, \ldots, X_n) - \theta(F))^2]$$

[where our choice of notation MSE(F) suppresses the dependence on the estimator g, and where we have used the notation E_F to indicate that the expectation is to be taken under the assumption that the random variables all have distribution F]. Now whereas there is an immediate estimator of the above MSE—namely, S^2/n—when $\theta(F) = E[X_i]$ and $g(X_1, \ldots, X_n) = \overline{X}$, it is not at all that apparent how it can be estimated otherwise. We now present a useful technique, known as the bootstrap technique, for estimating this mean square error.

To begin, note that if the distribution function F were known then we could theoretically compute the expected square of the difference between θ and its estimator; that is, we could compute the mean square error. However, after we observe the values of the n data points, we have a pretty good idea what the underlying distribution looks like. Indeed, suppose that the observed values of the data are $X_i = x_i, i = 1, \ldots, n$. We can now estimate the underlying distribution function F by the so-called empirical distribution function F_e, where $F_e(x)$, the estimate of $F(x)$, the probability that a datum value is less than or equal to x, is just the proportion of the n data values that are less than or equal to x. That is,

$$F_e(x) = \frac{\text{number of } i\text{: } X_i \leqslant x}{n}$$

Another way of thinking about F_e is that it is the distribution function of a random variable X_e which is equally likely to take on any of the n values $x_i, i = 1, \ldots, n$. (If the values x_i are not all distinct, then the above is to be interpreted to mean that X_e will equal the value x_i with a probability equal to the number of j such that $x_j = x_i$ divided by n; that is, if $n = 3$ and $x_1 = x_2 = 1, x_3 = 2$, then X_e is a random variable that takes on the value 1 with probability $\frac{2}{3}$ and 2 with probability $\frac{1}{3}$.)

Now if F_e is "close" to F, as it should be when n is large [indeed, the strong law of large numbers implies that with probability 1, $F_e(x)$ converges to $F(x)$ as $n \to \infty$, and another result, known as the Glivenko–Cantelli theorem, states that this convergence will, with probability 1, be uniform in x], then $\theta(F_e)$ will probably be close to $\theta(F)$—assuming that θ is, in some sense, a continuous function of the distribution—and MSE(F) should approximately be equal to

$$\text{MSE}(F_e) = E_{F_e}[(g(X_1, \ldots, X_n) - \theta(F_e))^2]$$

In the above expression the X_i are to be regarded as being independent random variables having distribution function F_e. The quantity MSE(F_e) is called the *bootstrap approximation to the mean square error* MSE(F).

To obtain a feel for the effectiveness of the bootstrap approximation to the mean square error, let us consider the one case where its use is not necessary—namely, when estimating the mean of a distribution by the sample mean $\overline{X}$. (Its use is not necessary in this case because there already is an effective way of estimating the mean square error $E[(\overline{X} - \theta)^2] = \sigma^2/n$—namely, by using the observed value of S^2/n.)

Example 8d Suppose we are interested in estimating $\theta(F) = E[X]$ by using the sample mean $\overline{X} = \sum_{i=1}^{n} X_i/n$. If the observed data are $x_i, i = 1, \ldots, n$, then the empirical distribution F_e puts weight $1/n$ on each of the points $x_1, \ldots, x_n$ (combining weights if the x_i are not all distinct). Hence the mean of F_e is $\theta(F_e) = \overline{x} = \sum_{i=1}^{n} x_i/n$, and thus the bootstrap estimate of the mean square error—call it MSE (F_e)—is given by

$$\text{MSE}(F_e) = E_{F_e}\left[\left(\sum_{i=1}^{n} \frac{X_i}{n} - \overline{x}\right)^2\right]$$

where $X_1, \ldots, X_n$ are independent random variables each distributed according to F_e. Since

$$E_{F_e}\left[\sum_{i=1}^{n} \frac{X_i}{n}\right] = E_{F_e}[X] = \overline{x}$$

it follows that

$$\begin{aligned}\text{MSE}(F_e) &= \text{Var}_{F_e}\left(\sum_{i=1}^{n} \frac{X_i}{n}\right) \\ &= \frac{\text{Var}_{F_e}(X)}{n}\end{aligned}$$

Now

$$\begin{aligned}\text{Var}_{F_e}(X) &= E_{F_e}[(X - E_{F_e}[X])^2] \\ &= E_{F_e}[(X - \overline{x})^2] \\ &= \frac{1}{n}\left[\sum_{i=1}^{n}(x_i - \overline{x})^2\right]\end{aligned}$$

and so

$$\text{MSE}(F_e) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n^2}$$

which compares quite nicely with S^2/n, the usual estimate of the mean square error. Indeed, because the observed value of S^2/n is $\sum_{i=1}^{n}(x_i - \overline{x})^2/[n(n-1)]$, the bootstrap approximation is almost identical. □

If the data values are $X_i = x_i, i = 1, \ldots, n$, then, as the empirical distribution function F_e puts weight $1/n$ on each of the points x_i, it is usually easy to compute the value of $\theta(F_e)$: for example, if the parameter of interest $\theta(F)$ was the variance of the distribution F, then $\theta(F_e) = \text{Var}_{F_e}(X) = \sum_{i=1}^{n}(x_i - \overline{x})^2/n$. To determine the bootstrap approximation to the mean square error we then have to compute

$$\text{MSE}(F_e) = E_{F_e}[(g(X_1, \ldots, X_n) - \theta(F_e))^2]$$

However, since the above expectation is to be computed under the assumption that $X_1, \ldots, X_n$ are independent random variables distributed according to F_e, it follows that the vector $(X_1, \ldots, X_n)$ is equally likely to take on any of the n^n possible values $(x_{i_1}, x_{i_2}, \ldots, x_{i_n})$, $i_j \in \{1, 2, \ldots, n\}$, $j = 1, \ldots, n$. Therefore,

$$\text{MSE}(F_e) = \sum_{i_n} \cdots \sum_{i_1} \frac{[g(x_{i_1}, \ldots, x_{i_n}) - \theta(F_e)]^2}{n^n}$$

where each i_j goes from 1 to n, and so the computation of MSE (F_e) requires, in general, summing n^n terms—an impossible task when n is large.

However, as we know, there is an effective way to approximate the average of a large number of terms, namely, by using simulation. Indeed, we could generate a set of n independent random variables $X_1^1, \ldots, X_n^1$ each having distribution function F_e and then set

$$Y_1 = \left[g\left(X_1^1, \ldots, X_n^1\right) - \theta(F_e)\right]^2$$

Next, we generate a second set $X_1^2, \ldots, X_n^2$ and compute

$$Y_2 = \left[g\left(X_1^2, \ldots, X_n^2\right) - \theta(F_e)\right]^2$$

and so on, until we have collected the variables $Y_1, Y_2, \ldots, Y_r$. Because these Y_i are independent random variables having mean $\text{MSE}(F_e)$, it follows that we can use their average $\sum_{i=1}^{r} Y_i/r$ as an estimate of $\text{MSE}(F_e)$.

Remarks

1. It is quite easy to generate a random variable X having distribution F_e. Because such a random variable should be equally likely to be $x_1, \ldots, x_n$, just generate a random number U and set $X = x_I$, where $I = \text{Int}(nU) + 1$. (It is easy to check that this will still work even when the x_i are not all distinct.)
2. The above simulation allows us to approximate $\text{MSE}(F_e)$, which is itself an approximation to the desired $\text{MSE}(F)$. As such, it has been reported that roughly 100 simulation runs—that is, choosing $r = 100$—is usually sufficient. □

The following example illustrates the use of the bootstrap in analyzing the output of a queueing simulation.

Example 8e Suppose in Example 8a that we are interested in estimating the long-run average amount of time a customer spends in the system. That is, letting W_i be the amount of time the ith entering customer spends in the system, $i \geqslant 1$, we are interested in

$$\theta \equiv \lim_{n \to \infty} \frac{W_1 + W_2 + \cdots + W_n}{n}$$

To show that the above limit does indeed exist (note that the random variables W_i are neither independent nor identically distributed), let N_i denote the number of customers that arrive on day i, and let

$$D_1 = W_1 + \cdots + W_{N_1}$$
$$D_2 = W_{N_1+1} + \cdots + W_{N_1+N_2}$$

and, in general, for $i > 2$,

$$D_i = W_{N_1+\cdots+N_{i-1}+1} + \cdots + W_{N_1+\cdots+N_i}$$

In words, D_i is the sum of the times in the system of all arrivals on day i. We can now express θ as

$$\theta = \lim_{m\to\infty} \frac{D_1 + D_2 + \cdots + D_m}{N_1 + N_2 + \cdots + N_m}$$

where the above follows because the ratio is just the average time in the system of all customers arriving in the first m days. Upon dividing numerator and denominator by m, we obtain

$$\theta = \lim_{m\to\infty} \frac{(D_1 + \cdots + D_m)/m}{(N_1 + \cdots + N_m)/m}$$

Now as each day follows the same probability law, it follows that the random variables $D_1, \ldots, D_m$ are all independent and identically distributed, as are the random variables $N_1, \ldots, N_m$. Hence, by the strong law of large numbers, it follows that the average of the first m of the D_i will, with probability 1, converge to their common expectation, with a similar statement being true for the N_i. Therefore, we see that

$$\theta = \frac{E[D]}{E[N]}$$

where $E[N]$ is the expected number of customers to arrive in a day, and $E[D]$ is the expected sum of the times those customers spend in the system.

To estimate θ we can thus simulate the system over k days, collecting on the ith run the data N_i, D_i, where N_i is the number of customers arriving on day i and D_i is the sum of the times they spend in the system, $i = 1, \ldots, k$. Because the quantity $E[D]$ can then be estimated by

$$\overline{D} = \frac{D_1 + D_2 + \cdots + D_k}{k}$$

and $E[N]$ by

$$\overline{N} = \frac{N_1 + N_2 + \cdots + N_k}{k}$$

it follows that $\theta = E[D]/E[N]$ can be estimated by

$$\text{Estimate of } \theta = \frac{\overline{D}}{\overline{N}} = \frac{D_1 + \cdots + D_k}{N_1 + \cdots + N_k}$$

which, it should be noted, is just the average time in the system of all arrivals during the first k days.

To estimate

$$\text{MSE} = E\left[\left(\frac{\sum_{i=1}^{k} D_i}{\sum_{i=1}^{k} N_i} - \theta\right)^2\right]$$

we employ the bootstrap approach. Suppose the observed value of D_i, N_i is $d_i, n_i, i = 1, \ldots, k$. That is, suppose that the simulation resulted in n_i arrivals on day i spending a total time d_i in the system. Thus, the empirical joint distribution function of the random vector D, N puts equal weight on the k pairs $d_i, n_i, i = 1, \ldots, k$. That is, under the empirical distribution function we have

$$P_{F_e}\{D = d_i, N = n_i\} = \frac{1}{k}, \quad i = 1, \ldots, k$$

Hence,

$$E_{F_e}[D] = \overline{d} = \sum_{i=1}^{k} d_i/k, \qquad E_{F_e}[N] = \overline{n} = \sum_{i=1}^{k} n_i/k$$

and thus,

$$\theta(F_e) = \frac{\overline{d}}{\overline{n}}$$

Hence,

$$\text{MSE}(F_e) = E_{F_e}\left[\left(\frac{\sum_{i=1}^{k} D_i}{\sum_{i=1}^{k} N_i} - \frac{\overline{d}}{\overline{n}}\right)^2\right]$$

where the above is to be computed under the assumption that the k pairs of random vectors D_i, N_i are independently distributed according to F_e.

Since an exact computation of $\text{MSE}(F_e)$ would require computing the sum of k^k terms, we now perform a simulation experiment to approximate it. We generate k independent pairs of random vectors $D_i^1, N_i^1, i = 1, \ldots, k$, according to the empirical distribution function F_e, and then compute

$$Y_1 = \left(\frac{\sum_{i=1}^{k} D_i^1}{\sum_{i=1}^{k} N_i^1} - \frac{\overline{d}}{\overline{n}}\right)^2$$

We then generate a second set D_i^2, N_i^2 and compute the corresponding Y_2. This continues until we have generated the r values $Y_1, \ldots, Y_r$ (where $r = 100$ should suffice). The average of these r values, $\sum_{i=1}^{r} Y_i/r$, is then used to estimate $\text{MSE}(F_e)$, which is itself our estimate of MSE, the mean square error of our estimate of the average amount of time a customer spends in the system. □

Remark The Regenerative Approach The foregoing analysis assumed that each day independently followed the same probability law. In certain applications, the same probability law describes the system not over days of fixed lengths but rather over cycles whose lengths are random. For example, consider a queueing system in which customers arrive in accordance with a Poisson process, and suppose that the first customer arrives at time 0. If the random time T represents the next time that an arrival finds the system empty, then we say that the time from 0 to T constitutes the first cycle. The second cycle would be the time from T until the first time point after T that an arrival finds the system empty, and so on. It is easy to see, in most models, that the movements of the process over each cycle are independent and identically distributed. Hence, if we regard a cycle as being a "day," then all of the preceding analysis remains valid. For example, θ, the amount of time that a customer spends in the system, is given by $\theta = E[D]/E[N]$, where D is the sum of the times in the system of all arrivals in a cycle and N is the number of such arrivals. If we now generate k cycles, our estimate of θ is still $\sum_{i=1}^{k} D_i / \sum_{i=1}^{k} N_i$. In addition, the mean square error of this estimate can be approximated by using the bootstrap approach exactly as above.

The technique of analyzing a system by simulating "cycles," that is, random intervals during which the process follows the same probability law, is called the regenerative approach.

Exercises

1. For any set of numbers $x_1, \ldots, x_n$, prove algebraically that

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

where $\overline{x} = \sum_{i=1}^{n} x_i / n$.

2. Give a probabilistic proof of the result of Exercise 1, by letting X denote a random variable that is equally likely to take on any of the values $x_1, \ldots, x_n$, and then by applying the identity $\text{Var}(X) = E[X^2] - (E[X])^2$.

3. Write a program that uses the recursions given by Equations (8.6) and (8.7) to calculate the sample mean and sample variance of a data set.

4. Continue to generate standard normal random variables until you have generated n of them, where $n \geqslant 100$ is such that $S/\sqrt{n} < 0.1$, where S is the sample standard deviation of the n data values.

(a) How many normals do you think will be generated?
(b) How many normals did you generate?
(c) What is the sample mean of all the normals generated?
(d) What is the sample variance?
(e) Comment on the results of (c) and (d). Were they surprising?

5. Repeat Exercise 4 with the exception that you now continue generating standard normals until $S/\sqrt{n} < 0.01$.

6. Estimate $\int_0^1 \exp(x^2)dx$ by generating random numbers. Generate at least 100 values and stop when the standard deviation of your estimator is less than 0.01.

7. To estimate $E[X]$, $X_1, \ldots, X_{16}$ have been simulated with the following values resulting: 10, 11, 10.5, 11.5, 14, 8, 13, 6, 15, 10, 11.5, 10.5, 12, 8, 16, 5. Based on these data, if we want the standard deviation of the estimator of $E[X]$ to be less than 0.1, roughly how many additional simulation runs will be needed?

Exercises 8 and 9 are concerned with estimating e.

8. It can be shown that if we add random numbers until their sum exceeds 1, then the expected number added is equal to e. That is, if

$$N = \min\left\{n: \sum_{i=1}^{n} U_i > 1\right\}$$

then $E[N] = e$.

(a) Use this preceding to estimate e, using 1000 simulation runs.
(b) Estimate the variance of the estimator in (a) and give a 95 percent confidence interval estimate of e.

9. Consider a sequence of random numbers and let M denote the first one that is less than its predecessor. That is,

$$M = \min\{n: U_1 \leqslant U_2 \leqslant \cdots \leqslant U_{n-1} > U_n\}$$

(a) Argue that $P\{M > n\} = \frac{1}{n!}, n \geqslant 0$.
(b) Use the identity $E[M] = \sum_{n=0}^{\infty} P\{M > n\}$ to show that $E[M] = e$.
(c) Use part (b) to estimate e, using 1000 simulation runs.
(d) Estimate the variance of the estimator in (c) and give a 95 percent confidence interval estimate of e.

10. Use the approach that is presented in Example 3a of Chapter 3 to obtain an interval of size less than 0.1, which we can assert, with 95 percent confidence, contains π. How many runs were necessary?

11. Repeat Exercise 10 when we want the interval to be no greater than 0.01.

12. To estimate θ, we generated 20 independent values having mean θ. If the successive values obtained were

102, 112, 131, 107, 114, 95, 133, 145, 139, 117
93, 111, 124, 122, 136, 141, 119, 122, 151, 143

how many additional random variables do you think we will have to generate if we want to be 99 percent certain that our final estimate of θ is correct to within ± 0.5?

13. Let $X_1, \ldots, X_n$ be independent and identically distributed random variables having unknown mean μ. For given constants $a < b$, we are interested in estimating $p = P\{a < \sum_{i=1}^{n} X_i/n - \mu < b\}$.

(a) Explain how we can use the bootstrap approach to estimate p.
(b) Estimate p if $n = 10$ and the values of the X_i are 56, 101, 78, 67, 93, 87, 64, 72, 80, and 69. Take $a = -5$, $b = 5$.

In the following three exercises $X_1, \ldots, X_n$ is a sample from a distribution whose variance is (the unknown) σ^2. We are planning to estimate σ^2 by the sample variance $S^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2/(n-1)$, and we want to use the bootstrap technique to estimate $\mathrm{Var}(S^2)$.

14. If $n = 2$ and $X_1 = 1$ and $X_2 = 3$, what is the bootstrap estimate of $\mathrm{Var}(S^2)$?

15. If $n = 15$ and the data are

$$5, 4, 9, 6, 21, 17, 11, 20, 7, 10, 21, 15, 13, 16, 8$$

approximate (by a simulation) the bootstrap estimate of $\mathrm{Var}(S^2)$.

16. Consider a single-server system in which potential customers arrive in accordance with a Poisson process having rate 4.0. A potential customer will only enter if there are three or fewer other customers in the system when he or she arrives. The service time of a customer is exponential with rate 4.2. No additional customers are allowed in after time $T = 8$. (All time units are per hour.) Develop a simulation study to estimate the average amount of time that an entering customer spends in the system. Using the bootstrap approach, estimate the mean square error of your estimator.

Bibliography

Bratley, P., B. L. Fox, and L. E. Schrage, *A Guide to Simulation*, 2nd ed. Springer-Verlag, New York, 1988.

Crane, M. A., and A. J. Lemoine, *An Introducion to the Regenerative Method for Simulation Analysis*. Springer-Verlag, New York, 1977.

Efron, B., and R. Tibshirani, *Introduction to the Bootstrap*. Chapman-Hall, New York, 1993.

Kleijnen, J. P. C., *Statistical Techniques in Simulation*, Parts 1 and 2. Marcel Dekker, New York, 1974/1975.

Law, A. M., and W. D. Kelton, *Simulation Modelling and Analysis*, 3rd ed. McGraw-Hill, New York, 1997.