Proceedings of the 2001 Winter Simulation Conference
B. A. Peters, J. S. Smith, D. J. Medeiros, and M. W. Rohrer, eds.

ANALYSIS OF SIMULATION EXPERIMENTS BY BOOTSTRAP RESAMPLING

Russell C.H. Cheng

Department of Mathematics
University of Southampton
Southampton, SO17 1BJ, U.K.

ABSTRACT

This tutorial considers some very general procedures for analysing the results of a simulation experiment using bootstrap resampling. Bootstrapping has come to be recognised in statistics as being far ranging and effective. However it is not so well known in simulation despite being ideally suited for use in such a context. We discuss aspects ranging from the elementary to the advanced. We describe the rationale and the simple steps needed to implement bootstrapping in (i) estimation of the distributional properties of the output and its dependence on factors of interest; (ii) model fitting; (iii) model selection; (iv) model validation; (v) sensitivity analysis.

1 INTRODUCTION

This tutorial considers some very general procedures for analysing the results of a simulation experiment The procedures to be described and discussed make use of resampling. Such methods have come to be recognised in statistics as far ranging and effective. However they are not so well known in simulation despite being ideally suited for use in such a context.

Though the underlying methodology is essentially well-known it is possible to take advantage of particular features of simulation experiments to simplify or to improve the accuracy or the efficiency of the analysis.

We discuss aspects ranging from the elementary to the advanced. We will discuss (i) estimation of the distributional properties of the output and its dependence on factors of interest; (ii) model fitting; (iii) model selection; (iv) model validation; (v) sensitivity analysis.

Although it is not strictly necessary, it is convenient to conduct this discussion in a unified framework by adopting a metamodelling approach which uses a regression model comprising a deterministic *regression function* and a stochastic *error process* to describe the behaviour of the output. All the aspects previously mentioned can be conveniently discussed within this framework.

In Section 2 we describe the classical bootstrap and also a parametric version based on a metamodelling approach. Section 3 describes the simulation experimental layout. Section 4 describes calculation of distributional properties of the simulation output. Section 5 describes goodness-of-fit problems and model selection. Section 6 describes model validation, Section 7 describes sensitivity analysis and Section 8 describes the assessment of sequential procedures ,such as subset selection.

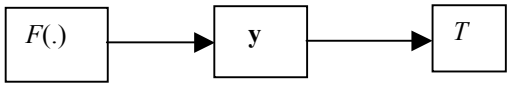
2 BOOTSTRAP METHOD

2.1 The Basic Bootstrap

Bootstrapping is an ideal method, well matched to simulation methodology, for estimating accuracy of simulation output and for carrying out sensitivity analysis. It provides a computer-based solution to the fundamental question of all statistical methodology:

> *A statistic, T, is calculated from a random sample,*
> $\mathbf{y} = (y_1, y_2, \ldots, y_m)$. *What is the distribution of T?*

The basic process is illustrated in Figure 1 where a random sample **y** is drawn from a (null) distribution $F(.)$, and a statistic of interest, T, is then calculated from this sample.



Null Distribution **Sample** **Test Stat.**

Figure 1: Basic Process

If we could repeat the basic process a large number of times, B say, (typically B = 100 at least, see Chernick, 1999) then we have a large sample $\{T^{(1)}, T^{(2)},\ldots, T^{(B)}\}$ of test statistics, and the (sampled) empirical distribution function (EDF) of these estimates the required distribution. This de-

sired process is illustrated in Figure 2 and the resulting EDF formed from the (ordered) sample $\{T^{(1)}, T^{(2)},..., T^{(B)}\}$ is depicted in Figure 3. This converges to the true distribution of T as B becomes indefinitely large.

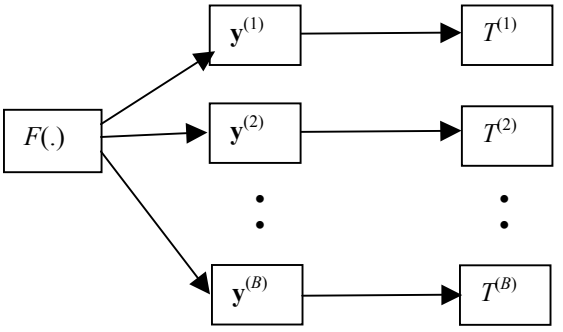


Figure 2: Desired Process

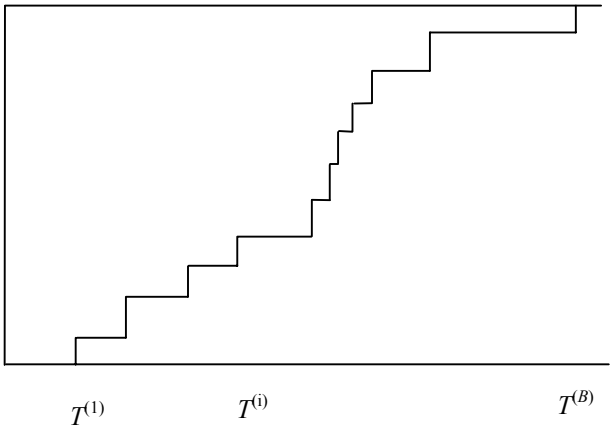


Figure 3: EDF of the Sample of Test Statistics

Obtaining the samples $\mathbf{y}^{(1)}, \mathbf{y}^{(2)},..., \mathbf{y}^{(B)}$ is usually too expensive to do. The bootstrap resampling method replaces $F(.)$ by the best estimate we have for it. This is simply the EDF formed from the original sample **y.** We denote this by $F_m(.)$.

The bootstrap version of the desired process is illustrated in Figure 4. Here and in what follows we use an asterisk to denote a bootstrap value. Thus $\mathbf{y}^{*(i)}$ denotes the ith bootstrap sample and $T^{*(i)}$ the test statistic calculated from this sample. The EDF formed from the (ordered) bootstrap sample $\{T^{*(1)}, T^{*(2)},..., T^{*(B)}\}$ estimates the (true) CDF of T. This is illustrated in Figure 5.

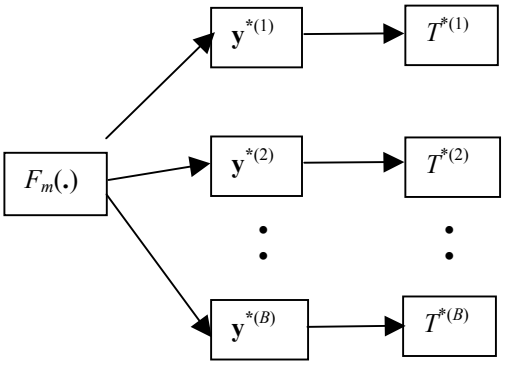


Figure 4: Bootstrap Process

Also depicted in Figure 5 is the test statistic, T, calculated from the original sample. Its p-value

$$p = \frac{\#\text{of } T^{*(i)} \geq T}{B}$$

can readily be read off from the bootstrap EDF, as indicated in the Figure.

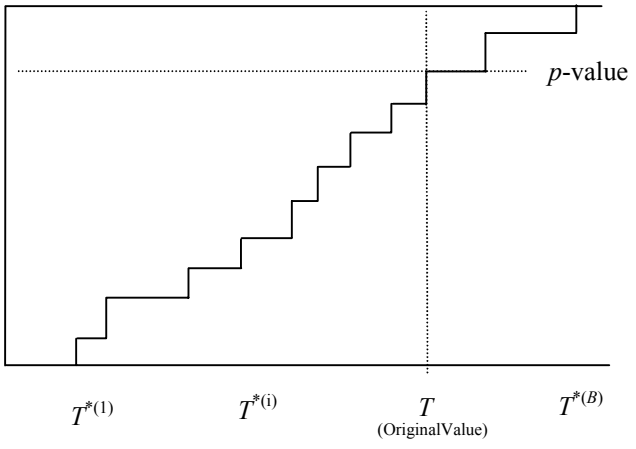


Figure 5: EDF of the Sample of Bootstrap Test Statistics

There is a huge literature both theoretical and practical showing how well the method works. Chernick (1999) gives some 1600 references. Hjorth (1994) gives a good introduction. See also Davison and Hinkley (1997). Expressed informally, for the method to work, we require two things. Firstly we require that $F_m(.)$ converges to $F(.)$. When **y** is a random sample this is guaranteed by the Glivenko-Cantelli theorem that ensures that $F_m(.) \rightarrow F(.)$ uniformly with probability one (See Chernick 1999 for details). Secondly if we regard the statistic T as estimating some characteristic property of the distribution (technically this property is a functional of the population distributions) then this characteristic must satisfy certain smoothness properties. Most characteristics such as moments

and percentiles are smooth and so their estimators have distributions that can be estimated by bootstrap resampling. Bootstrap resampling most often fails when T is a statistic estimating a functional that is not smooth. Thus for example maxima or minima of a sample have distributions that cannot usually be estimated by bootstrap resampling.

2.2 The Parametric Bootstrap

A variation of the Basic Process is depicted in Figure 6. The only difference is that the statistical quantity of interest depends on a vector of parameters $\boldsymbol{\theta}_0$ that has to be estimated from the sample **y.**

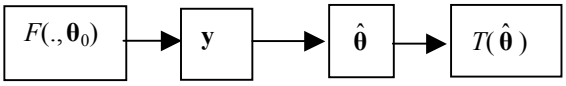


Figure 6: Basic Parametric Process

The sample is assumed to have a joint distribution with probability increment $DF(\mathbf{y}, \boldsymbol{\theta})$ depending on $\boldsymbol{\theta}$. If the sample is a random sample then the probability increment becomes a product of univariate increments: $DF(\mathbf{y}, \boldsymbol{\theta}) = \prod_{j=1}^{m} DF(y_j, \boldsymbol{\theta})$, but in general we do not necessarily have to assume this.

A good way of obtaining $\hat{\boldsymbol{\theta}}$ is to use maximum likelihood (ml) estimation (see for example Efron and Tibshirani, 1993). The likelihood is simply the joint distribution probability element evaluated at the observed values, and then treated as a function of $\boldsymbol{\theta}$. It is usually easier to work with its logarithm (loglikelihood):

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{m} \log DF(y_i, \boldsymbol{\theta}) . \qquad (1)$$

The ml estimate, $\hat{\boldsymbol{\theta}}$, is the value of $\boldsymbol{\theta}$ which maximizes this loglikelihood. When the loglikelihood is maximized at an interior point of the parameter space then $\hat{\boldsymbol{\theta}}$ satisfies the likelihood equation $L'(\boldsymbol{\theta}) = \mathbf{0}$. The observed and expected information are defined as $\mathbf{I}(\boldsymbol{\theta}) = -\partial^2 L(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}^2$ and $\mathbf{i}(\boldsymbol{\theta}) = E[\mathbf{I}(\boldsymbol{\theta})]$.

Under general regularity conditions the ml estimate has the important property that its distribution is asymptotically normal as the sample size tends to infinity:

$$\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}_0, \mathbf{i}(\boldsymbol{\theta}_0)^{-1}) . \qquad (2)$$

In practice, as $\boldsymbol{\theta}_0$ is unknown, use is made of one of the asymptotically equivalent versions:

$$\hat{\boldsymbol{\theta}} \sim N(\hat{\boldsymbol{\theta}}, \mathbf{i}(\hat{\boldsymbol{\theta}})^{-1}) \text{ or } \hat{\boldsymbol{\theta}} \sim N(\hat{\boldsymbol{\theta}}, \mathbf{I}(\hat{\boldsymbol{\theta}})^{-1}) . \qquad (3)$$

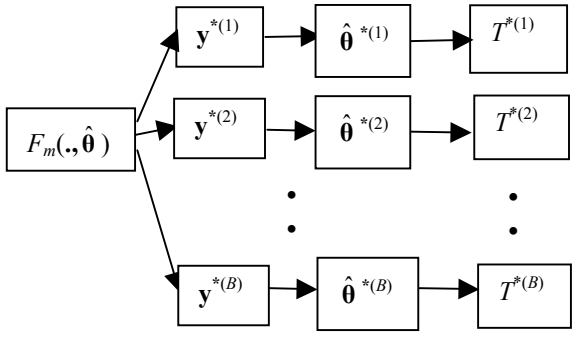


Figure 7: Parametric Bootstrap Process, First Version

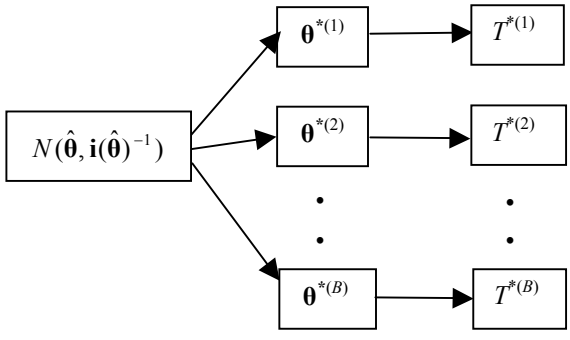


Figure 8: Parametric Bootstrap Process, Second Version

Bootstrapping can be done in two ways. The 'non-parametric' method is simply to consider the estimation of $\boldsymbol{\theta}$ as part of the process of calculating T. Bootstrapping can be done in two ways. The 'non-parametric'

Bootstrapping can be done in two ways. The 'non-parametric' method is simply to consider the estimation of $\boldsymbol{\theta}$ as part of the process of calculating T. Bootstrapping then proceeds precisely as before by resampling of **y**. This method is depicted in Figure 7.

However an alternative is possible if a method like ml is used to estimate $\boldsymbol{\theta}$. This is illustrated in Figure 8. In this alternative method we bypass the sampling of **y**, and instead resample values of $\hat{\boldsymbol{\theta}}$ using (3) directly.

3 THE SIMULATION MODEL

We assume that y, the output of interest from the simulation run, depends on a stream of uniform random variates $\mathbf{u} = (u_1, u_2, ...)$, on a vector $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_p)$ of p parameters and on a vector $\mathbf{x} = (x_1, x_2, ... x_q)$ of q design factors, i.e.

$$y = y(\mathbf{u}, \boldsymbol{\theta}, \mathbf{x}).$$

The design factor values are assumed known, but the parameters may not all be known. Typically the uniforms are transformed into random variates that involve some of the parameter values. Thus actual random variates used might take the form $w_i = \varphi_i(\mathbf{u}_i, \boldsymbol{\theta})$ where $\mathbf{u}_i$ denotes some subset of the uniforms used. In what follows we do not need to consider the w_i's explicitly. We regard the objective as being to estimate the expected value of y, and note that this is a function of $\boldsymbol{\theta}$ and $\mathbf{x}$ only:

$$\eta(\boldsymbol{\theta}, \mathbf{x}) = E(y, \boldsymbol{\theta}, \mathbf{x}) = \int y(\mathbf{u}, \boldsymbol{\theta}, \mathbf{x}) \mathrm{d}\mathbf{u} . \qquad (4)$$

The above formulation requires one further refinement. We need to distinguish between two types of parameter. In the above formulation $\boldsymbol{\theta}$ denotes those parameters that depend purely on processes that are exogenous to the simulation model. Examples of such parameters are arrival rates depending on arrival processes external to the simulation under consideration. Such parameters are part of the description of how the process behaves, and their values will have to be fixed before a simulation can be carried out. For this reason we call the components of $\boldsymbol{\theta}$ *input parameters* and $\boldsymbol{\theta}$ itself the *input parameter vector.*

However there is another way that parameters can enter the simulation. We let $\boldsymbol{\beta}$ denote those parameters which are added subsequent to the simulation experiment and which have been introduced to describe the distributional properties of the output y. Typically $\boldsymbol{\beta}$ as part of a regression metamodel. In fitting such a model, $\boldsymbol{\beta}$ will need to be estimated from the output y values obtained from the simulations. We shall call the vector $\boldsymbol{\beta}$ the *output parameter vector*, and its components, *output parameters*.

We consider two cases, depending on whether $\boldsymbol{\theta}$ is known or not. In either case $\boldsymbol{\beta}$ is unlikely to be known, so we shall only consider the case where it is unknown.

Case A: Here $\boldsymbol{\theta}$ is regarded as fixed and known.

We consider the overall simulation experiment as being made up of runs made at n design points with m_i *runs* made at $\mathbf{x}_i$ the ith design point. The responses or outputs from these runs will be written as:

$$y_{ij}=y(\mathbf{u}_{ij}, \boldsymbol{\theta}_0, \boldsymbol{\beta}, \mathbf{x}_i) = \eta(\boldsymbol{\theta}_0, \boldsymbol{\beta}, \mathbf{x}_i) + e(\mathbf{u}_{ij}, \boldsymbol{\theta}_0, \boldsymbol{\beta}, \mathbf{x}_i),$$

$$i = 1, 2, ..., n, j = 1, 2, ..., m_i, \qquad (5)$$

where $\boldsymbol{\theta}_0$ is the given value of $\boldsymbol{\theta}$. The 'error' variable e is the random difference between the simulation run output and $\eta(\boldsymbol{\theta}_0, \mathbf{x}, \boldsymbol{\beta})$. We shall assume E($e$) = 0. We have

$$\mathrm{E}[y_{ij}] = \eta(\boldsymbol{\theta}_0, \boldsymbol{\beta}, \mathbf{x}_i)$$

and the mean of the outputs $\bar{y}_i = \sum_{j=1}^{m_i} y_{ij} / m_i$ is an unbiased estimator of $\eta(\boldsymbol{\theta}_0, \boldsymbol{\beta}, \mathbf{x}_i)$.

To fully identify the distributional properties of the y_{ij} the unknown $\boldsymbol{\beta}$ will have to be estimated. We can regard the estimation of $\boldsymbol{\beta}$ as part of the simulation process with the estimate $\hat{\boldsymbol{\beta}}$ forming part of the output of the simulation. In status $\hat{\boldsymbol{\beta}}$ is like the output y itself.

Case B: Here $\boldsymbol{\theta}$ is not known but has to be estimated. As already remarked $\boldsymbol{\theta}$ denotes those parameters that depend purely on processes that are exogenous to the simulation model. We suppose therefore that there exist empirical data: $\mathbf{z} = (z_1, z_2, ...z_l)$ drawn from a distribution $G(., \boldsymbol{\theta}_0)$ where $\boldsymbol{\theta}_0$ is the unknown true parameter value. This value will thus have to be estimated from this data.

Again we assume that m_i runs are made at $\mathbf{x}_i$, the ith design point. The responses or outputs from these runs will be written as:

$$y_{ij}=y(\mathbf{u}_{ij}, \hat{\boldsymbol{\theta}}, \boldsymbol{\beta}, \mathbf{x}_i) = \eta(\hat{\boldsymbol{\theta}}, \boldsymbol{\beta}, \mathbf{x}_i) + e(\mathbf{u}_{ij}, \hat{\boldsymbol{\theta}}, \boldsymbol{\beta}, \mathbf{x}_i),$$

$$i=1,2,...,n, j = 1,2,...,m_i. \qquad (6)$$

Here the unknown $\boldsymbol{\theta}$ has been replaced by an estimate $\hat{\boldsymbol{\theta}}$.

We have included the vector $\boldsymbol{\beta}$ in (5) and (6), but it is best regarded as part of the output process. It can only be estimated after the simulations and this will need to be done in order to completely determine the distribution of y.

The mean of the outputs $\bar{y}_i = \sum_{j=1}^{m_i} y_{ij} / m_i$ is still an unbiased estimator of $\eta(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{x}_i)$.

For notational simplicity, when the y are assumed to have the form (5) or (6), we write the distribution of y as $F(., \boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{x})$.

We now consider each of the problems listed in the introduction.

4 DISTRIBUTIONS AT A FIXED DESIGN POINT

The situation here is as given in (5) or (6) depending on whether $\boldsymbol{\theta}$ is known or not. If all simulation runs are conducted at the same design point then n = 1, and the sample is $\mathbf{y}_1 = (y_{11}, y_{12}, ..., y_{1m})$. As there is just one sample we suppress the first subscript and write it simply as $\mathbf{y} = \mathbf{y}_1 = (y_1, y_2, ..., y_m)$. We also write $\mathbf{x}_0 = \mathbf{x}_1$. Of likely interest is the estimation of typical properties of the distribution of y such as its mean, variance and selected percentiles. We illustrate with the case of the mean. Here $T = \bar{y}$.

Consider first Case A, where $\boldsymbol{\theta}$ is known. Let $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. In this case there are two possible ways of estimating the distribution of T.

The first is the 'non-metamodelling' approach where we ignore the structure assumed in the right-hand side of equation (5) or (6) and deal with the y_j directly. We therefore estimate the distribution of T by the bootstrap process of Figure 4.

The other is where we adopt the structure assumed in the right-hand side of (5) or (6). We write the distribution of y as $F(., \boldsymbol{\theta}_0, \hat{\boldsymbol{\beta}}, \mathbf{x}_0)$, where $\hat{\boldsymbol{\beta}}$ has been estimated from the sample $\mathbf{y}$, by ml, say. The distribution of T can thus still be estimated by the bootstrap process of Figure 4 except that the bootstrap samples $\mathbf{y}^{*(i)}$ are each drawn from $F(., \boldsymbol{\theta}_0, \hat{\boldsymbol{\beta}}, \mathbf{x}_0)$. The only point of note is that the calculation of each bootstrap $T^{*(i)}$ requires calculation of a corresponding estimate, $\hat{\boldsymbol{\beta}}^{*(i)}$, of $\boldsymbol{\beta}$.

Consider now Case B, where $\boldsymbol{\theta}$ is not known. In this case the simulation runs must use an estimate, $\hat{\boldsymbol{\theta}}$, of $\boldsymbol{\theta}$. The non-metamodelling approach cannot be used as the original sample $\mathbf{y}$ does not include the variation arising from this estimation process.

We can however use the metamodelling approach. To incorporate the added uncertainty arising from the estimation of $\boldsymbol{\theta}$, each bootstrap sample $\mathbf{y}^{*(i)}$ is now drawn from a separate distribution, namely $F(., \hat{\boldsymbol{\theta}}^{*(i)}, \hat{\boldsymbol{\beta}}, \mathbf{x}_0)$, in which $\hat{\boldsymbol{\theta}}^{*(i)}$ has been sampled by the bootstrap process of Figure 7 or 8 (except that in Figure 7 $\mathbf{y}$ needs to be replaced by $\mathbf{z}$, and F_m by G_l) so that it varies from sample to sample.

5 GOODNESS-OF-FIT AND MODEL SELECTION

5.1 Goodness-of-Fit

The discussion of Section 4 extends naturally to the study of how well the simulation output is represented by some given parametric model. We consider again the situation where $n = 1$ with the sample written as $\mathbf{y} = \mathbf{y}_1 = (y_1, y_2, ..., y_m)$. and where we write $\mathbf{x}_0 = \mathbf{x}_1$. The situation is precisely that of equation (5) or (6), with $n = 1$, but where we are uncertain whether we have the correct form for $\eta(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{x}_0)$ or for $e(\mathbf{u}, \boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{x}_0)$. This is a question of goodness-of-fit of the model and is a problem that can be difficult to handle using traditional techniques. Resampling methods offer a particularly good solution to goodness-of-fit problems (See Cheng *et al.* 1996).

The situation is basically that of Figure 6 but where we are uncertain if the null distribution of y is $F(., \boldsymbol{\theta}_0, \hat{\boldsymbol{\beta}}, \mathbf{x}_0)$ when we have Case A or is $F(., \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}, \mathbf{x}_0)$ when we have Case B. In either case we simply take T to be a goodness of fit statistic. A good choice is the Anderson - Darling statistic (Anderson and Darling, 1952)

$$A^2 = -m - m^{-1}\sum_{j=1}^{m}(2j-1)\{\ln(z_j) + \ln(1 - z_{m-j+1})\}$$

where $z_j = F(y_{(j)}, \boldsymbol{\theta}, \hat{\boldsymbol{\beta}}, \mathbf{x}_0)$, and $y_{(j)}$ denotes the jth observation of the *ordered* $\mathbf{y}$ sample. We shall use $T = A^2$ in this case.

The method now precisely follows that of Section 4, but using the metamodelling approach. The bootstrap process of Figure 7 (except that $F_m(., \hat{\boldsymbol{\theta}})$ in the Figure is replaced by $F(., \boldsymbol{\theta}_0, \hat{\boldsymbol{\beta}}, \mathbf{x}_0)$ in Case A and by $F(., \hat{\boldsymbol{\theta}}^{*(i)}, \hat{\boldsymbol{\beta}}, \mathbf{x}_0)$ in Case B) is applied to obtain an estimate of the distribution of the test statistic T under the assumption that the fitted model is the correct one. Thus, in calculating $A^{*(i)2}$, the bootstrap value of A^2 obtained from the ith bootstrap sample, we use $z_j^{*(i)} = F(y_{(j)}^{*(i)}, \boldsymbol{\theta}_0, \hat{\boldsymbol{\beta}}, \mathbf{x}_0)$ in Case A and $z_j^{*(i)} = F(y_{(j)}^{*(i)}, \hat{\boldsymbol{\theta}}^{*(i)}, \hat{\boldsymbol{\beta}}, \mathbf{x}_0)$ in Case B.

We then examine the p-value - as measured by this estimated distribution - of the T statistic calculated from the *original* sample $\mathbf{y}$, i.e.

$$p = \frac{\#\text{of } T^{*(i)} \geq T}{B}.$$

If this p-value is too small the model is rejected as being a poor fit.

The method above is described for just one design point but extends in an obvious way if we wish to consider the goodness-of-fit of the model as a whole, including all design points. All that is needed is some appropriate portmanteau goodness-of-fit test statistic covering all design points simultaneously.

5.2 Model Selection

Model selection is easily handled if we regard it simply as an extension of goodness-of-fit to the situation where we may be unsure as to which of a number of alternative models may be the best representation of the output.

Assume that we have a number, k say, of contending models, each like (5) or (6). We write these models as:

$$F_j(., \boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{x}),\ j = 1, 2, ..., k$$

We calculate the p-value of the goodness-of-fit statistic for each of these contending models in exactly the way described in the above goodness-of-fit procedure. Models with p-values that are too small can be ruled out as being

not good fits. If just one model has to be chosen, then that corresponding to the largest *p*-value is the natural choice.

6 MODEL VALIDATION

We consider only the basic situation where the simulation output is being compared with existing output from some real or previously validated model. The situation is similar to that of model selection when just two models are being compared so that essentially the same techniques can be applied here.

We consider the case where $\mathbf{y}_0 = (y_{01}, y_{02}, \ldots y_{0m})$ is the observed output from a real system and we wish to compare this with $\mathbf{y}_1 = (y_{11}, y_{12}, \ldots y_{1m})$ the observed output from a set of simulation runs. We use a method proposed by Kleijnen *et al.* (2001).

We consider a statistic $T=T(\mathbf{y}_0, \mathbf{y}_1)$ that measures the difference between the two samples in some way. This might simply be $T = \bar{y}_1 - \bar{y}_0$ or a quantity such as $T = \int (F_{0m} - F_{1m})^2 dF_{0m}$ that measures the average squared difference between the EDF's of the two samples. Here $F_{0m}(.)$ and $F_{1m}(.)$ denote the EDF's of the two samples $\mathbf{y}_0$ and $\mathbf{y}_1$ respectively.

We now make a second set of simulation runs, $\mathbf{y}_2$, under the same conditions used for obtaining $\mathbf{y}_1$. Thus if we used $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ in Case A or $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ in Case B when obtaining $\mathbf{y}_1$ then we use the same value for $\boldsymbol{\theta}$ in the simulations that yield $\mathbf{y}_2$. The key point is that $T(\mathbf{y}_0, \mathbf{y}_1)$ and $T(\mathbf{y}_2, \mathbf{y}_1)$ will be identically distributed if all three samples are drawn from the same distribution. The distribution of $T(\mathbf{y}_0, \mathbf{y}_1)$ or of $T(\mathbf{y}_2, \mathbf{y}_1)$ can be obtained by bootstrapping.

The distribution of $T(\mathbf{y}_2, \mathbf{y}_1)$ is possibly the easier to use as it can be estimated by bootstrapping using either the non-metamodelling or the metamodelling approach.

In the non-metamodelling approach we can obtain bootstrap versions of $\mathbf{y}_1$, $\mathbf{y}_2$ by resampling with replacement to yield: $\mathbf{y}_1^{*(i)}$, $\mathbf{y}_2^{*(i)}$, for $i = 1, 2, \ldots, m$; and from these pairs we can calculate bootstrap $T^{*(i)} = T(\mathbf{y}_1^{*(i)}, \mathbf{y}_2^{*(i)})$, for $i = 1, 2, \ldots, m$. The *p*-value of $T(\mathbf{y}_0, \mathbf{y}_1)$, as given by the EDF formed from the $T^{*(i)}$, is a measure of the similarity between the output of the real system and that of the simulation, with a small value of *p* indicating dissimilarity.

In the metamodelling approach we can obtain bootstrap versions of $\mathbf{y}_1$, $\mathbf{y}_2$ by sampling from the fitted metamodel, this being $F(., \boldsymbol{\theta}_0, \hat{\boldsymbol{\beta}}, \mathbf{x}_0)$ in Case A and $F(., \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}, \mathbf{x}_0)$ in Case B. The procedure is then exactly as for the non-metamodelling case.

7 SENSITIVITY ANALYSIS

We consider Case B only. The study of the distribution of parameter estimates, and the way they affect estimates of the simulation output, can be regarded as a problem of sensitivity analysis. The effect of randomness (i.e. that due to $\mathbf{u}$) built into the model may also be regarded as a sensitivity issue. In fact using the regression metamodel approach, with the output regarded as being of the form assumed in equation (6), these two problems can be considered as two parts of the same problem. More explicitly, we are interested in how overall variability depends respectively on $\hat{\boldsymbol{\theta}}$ and on the $\mathbf{u}_{ij}$. A number of authors have discussed this issue, see for example Helton (1993, 1994).

Cheng and Holland (1995, 1997) show that, to first order, Var(y) is made up of two separate contributions, one depending purely on the variation arising from $\hat{\boldsymbol{\theta}}$ being random, the other due to variation in $\mathbf{u}_{ij}$. We call these two components the *parameter* and *simulation uncertainties* respectively.

Assessing the parameter uncertainty using classical statistical techniques requires estimation of the sensitivity coefficients:

$$\beta_i = \frac{\partial \eta(\boldsymbol{\theta}, \mathbf{x})}{\partial \theta_i}, \; i = 1,2,\ldots, p$$

This can be expensive when *p* is large.

We can overcome the problem using bootstrapping. We consider the case $n = 1$ and write $\mathbf{y} = \mathbf{y}_1$ as before.

Consider first the non-metamodelling approach. We begin by assessing the simulation uncertainty on its own (i.e. that due just to the $\mathbf{u}_j$). This is done by making the basic set of simulation runs using $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ in all runs. The sample variance

$$s^2 = (m-1)^{-1} \sum_{j=1}^{m} (y_j - \bar{y})^2 \qquad (7)$$

calculated from the *m* runs estimates the simulation uncertainty (at the given design point) only.

We then make a second set of simulation runs but now with $\boldsymbol{\theta}$ resampled separately in each run using the sampling scheme of either Figure 7 or 8 (Except that $\mathbf{y}$ needs to be replaced by $\mathbf{z}$, and F_m by G_l in Figure 7). The variance estimated from these runs, using exactly the same formula of equation (7), and which we denote by v_i^2, measures the parameter and simulation uncertainties combined. The difference

$$t_i^2 = v_i^2 - s_i^2$$

now estimates the parameter uncertainty at the *i*th design point.

A metamodelling approach is also possible and follows along the same lines as discussed for other problems. Details are left as an exercise.

If the variability of some quantity other than the mean is of interest then the entire set of subruns at the design point may be needed to calculate it. For example if the y's are not normal and an estimate of a percentage point is required, then this can be calculated from the EDF of the y values observed at the design point. In this case no separate estimate of the variability of this estimate is available as there was for the mean $\bar{y}$. The variance of such an estimate will have to be carried out by bootstrapping of the entire set of runs B times and calculating the estimate of interest from each bootstrap set of runs.

8 SEQUENTIAL PROCEDURES

8.1 Selection of the Best System Out of n

Suppose that the output (6) arises from a comparison of n different systems where the objective is to choose the best system according to a given performance index, assumed to be y. Thus the n 'design points' are actually distinct systems being compared.

Even if the runs have arisen from a sequential procedure, then, providing there are a reasonable number of runs m_i for each system, we can assess the robustness of the process that has produced a given best selection. Suppose for example that the sequential procedure has resulted in selection of the first system as being the best. We can test the reliability of this selection simply by repeating the selection process, only using bootstrap resampling each time, rather than actual simulations to generate the sequence of y's used. Thus, whenever a new y value from the ith design point is required by the selection procedure, we simply sample this from the set of $\mathbf{y}_i$ values initially obtained. Sampling is done with replacement of course, with values sampled each time from the full sample $\mathbf{y}_i$.

The entire sequential process is repeated B times. The proportion of times that the originally selected best system comes out on top provides an estimate of the confidence that this system is actually the best.

The above might be regarded as being the 'non-metamodelling' approach because resampling is from the original observations.

An alternative resampling strategy can be employed that makes explicit use of the fitted metamodel. This would be especially advantageous in situations where some of the m_i are too small for resampling to be done with confidence. In this case each bootstrap experiment is still a repeat of the entire sequential procedure, the only difference being that, when we are called to sample from the distribution of the ith model, we simply sample from the fitted distribution: $F(., \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}, \mathbf{x}_i)$, $i = 1, 2, ..., n$.

8.2 Important Factors

The methodology of Section 8.1 applies to any other sequential procedures. For example suppose the regression function $\eta(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{x})$ incorporates the effects of a number different factors and we use a sequential procedure to select those that are considered to be important.

We can use bootstrapping to assess how stable is the choice of those factors deemed to be important. We simply bootstrap the entire sequential procedure B times. In each bootstrap, the y's that are used can again be generated in either of the two ways indicated in Section 8.1.

In the first way resampling is simply carried out using the y samples obtained from the original simulations. The resampling is done with replacement.

In the second way we resample from the fitted distributions $F(., \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}, \mathbf{x}_i)$, $i = 1, 2, ..., n$.

9 SUMMARY

The above discussion indicates the flexibility of the bootstrap procedure and how it can be readily applied to many problems of practical interest in simulation modelling and experimentation. For reasons of space the discussion has focused on describing the procedures in sufficient detail to enable each to be easily and unambiguously implemented by the interested reader. It is hoped to illustrate these procedures with numerical examples in the Conference presentation.

REFERENCES

Anderson, T.W. and D.A. Darling. 1952. Asymptotic theory of certain goodness of fit criteria based on stochastic processes, *Annals of Mathematical Statistics*, 23: 193-212.

Cheng, R.C.H. and W. Holland. 1995. The effect of input parameters on the variability of simulation output. In *UKSS95, Proceedings of the 2nd Conference of the UK Simulation Society*, eds. R.C.H. Cheng and R.B. Pooley, 29-36. Edinburgh: Edinburgh University Reprographics Press.

Cheng, R.C.H. 1995. Bootstrap Methods in Computer Simulation Experiments . In *Proceedings of the 1995 Winter Simulation Conference* ed. W.R. Lilegdon, D. Goldsman, C. Alexopoulos and K. Kang, 171-177. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers..

Cheng, R.C.H., W. Holland, and N.A. Hughes. 1996. Selection of input models using bootstrap goodness-of-fit. In *Proceedings of the 1996 Winter Simulation Conference*, ed. J.M. Charnes, D.J. Morrice, D.T. Brunner and J.J. Swain, 317-322. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Cheng, R.C.H. and W. Holland. 1997. Sensitivity of computer simulation experiments to errors in input data. *Journal of Statistical Computation and Simulation,* 57: 219-241.

Chernick, M.R. 1999. *Bootstrap Methods A Practitioner's Guide*. New York, Wiley.

Davison, A.C. and D.V. Hinkley. 1997. *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press.

Efron, B. and R.J. Tibshirani. 1993. *An Introduction to the Bootstrap*. New York and London: Chapman & Hall.

Helton, J.C. 1993. Uncertainty and Sensitivity Analysis Techniques for Use in Performance Assessment for Radioactive Waste Disposal. *Reliability Engineering and System Safety*, 42: 327-367.

Helton, J.C. 1994. Treatment of Uncertainty in Performance Assessments for Complex Systems. *Risk Analysis*, 14: 483-511.

Hjorth, J.S.U. 1994. *Computer Intensive Statistical Methods*. London: Chapman & Hall.

Kleijnen, J.P.C., R.C.H. Cheng and B. Bettonvil. 2001. Validation of Trace-Driven Simulation Models. *Management Science*. To appear.

AUTHOR BIOGRAPHY

RUSSELL C. H. CHENG is Professor and Head of Operational Research at the University of Southampton. He has an M.A. and the Diploma in Mathematical Statistics from Cambridge University, England. He obtained his Ph.D. from Bath University. He is a former Chairman of the U.K. Simulation Society, a Fellow of the Royal Statistical Society, Member of the Operational Research Society. His research interests include: variance reduction methods and parametric estimation methods. He is Joint Editor of the *IMA Journal of Management Mathematics*. His email and web addresses are <R.C.H.Cheng@maths.soton.ac.uk> and <www.maths.soton.ac.uk/staff/Cheng>.