Proceedings of the 2002 Winter Simulation Conference
E. Yücesan, C.-H. Chen, J. L. Snowdon, and J. M. Charnes, eds.

COLLECTING DATA AND ESTIMATING PARAMETERS FOR INPUT DISTRIBUTIONS

Mike Freimer

School of Operations Research and
Industrial Engineering
Cornell University
276 Rhodes Hall
Ithaca, NY 14853, U.S.A.

Lee Schruben

Department of Industrial Engineering and
Operations Research
University of California at Berkeley
4135 Etcheverry Hall
Berkeley, CA 94720, U.S.A.

ABSTRACT

An early stage of a simulation study often consists of collecting data in order to parameterize the model. This paper addresses the question of how much data to collect, and from what sources. We use designed experiments to identify important unknown parameters, taking into account the current level of information about them. We develop approaches based on two common forms of analysis of variance: a fixed effects model, and a random effects model.

1 INTRODUCTION

From a high-level perspective, a simulation study consists of data collection, model analysis, and decision making. Each of these is generally viewed as a separate activity, the interactions of which are usually not considered. In this paper we provide a feedback mechanism from model analysis to data collection. The goal is to determine how much real-world data should be collected, and from what sources.

To a certain extent, such a feedback mechanism is already standard practice in the simulation community. Law and Kelton (1991) suggest, "Sensitivity analyses can be used to determine which parameters, distributions, or subsystems will have the greatest impact on the desired measures of performance. Given a limited amount of time for model development, one should obviously concentrate on the most important factors."

At its simplest, sensitivity analysis may consist of varying each parameter (one-at-a-time) from a low value to a high one and examining the effect on the simulation output. A more structured approach might involve running a factorial experiment and analyzing the results using analysis of variance (ANOVA). For example, if the number of unknown parameters k is not too large, a common technique is the 2^k factorial design. The analyst fixes two values of each parameter ("low" and "high") and performs simulation runs at each of the 2^k combinations of values. One then estimates the main effects and interactions of the parameters, and tests their significance. If the number of simulation parameters is very large, the 2^k factorial design involves a prohibitively large number of replications. In this case there are several screening designs available for determining a subset of the k parameters that are significant. Kleijnen (1987) discusses these designs in detail.

Our approach will take into account the current level of knowledge about each parameter. In particular, we may be less concerned with a sensitive parameter for which we have a precise estimate than with a less-sensitive parameter about which we have little information. There is an important tradeoff between our uncertainty about a parameter and its sensitivity with respect to the simulation output. We account for this tradeoff in the experimental design by selecting appropriate parameter levels to test. These levels reflect our uncertainty concerning the true value of the parameter.

We develop two approaches, based on the two most common variants of ANOVA. The first uses a *fixed effects* model and a 2^k factorial design. The treatment levels correspond to the endpoints of the confidence intervals (CI) for the unknown parameters, estimated from actual data. If the ANOVA detects a significant difference between the endpoints of the CI for a particular parameter, the conclusion is that more data must be collected for this parameter. Collecting more data reduces the width of the CI, reducing the effect of the parameter. In the limit, the CI narrows to a single point, our uncertainty about the parameter vanishes, and the parameter is of no further significance with respect to data collection.

A second approach is based on the *random effects* version of ANOVA. This approach has the advantage that it does not require a CI for each unknown parameter, however it can be much more computationally intensive.

2 FIXED EFFECTS MODEL

In this section we develop the approach based on the fixed effects version of ANOVA.

2.1 Review of the Fixed Effects ANOVA

We begin with a brief review of the fixed effects version of ANOVA for a single factor experiment. The notation is based on Hines and Montgomery (1990). Suppose the factor has a levels (treatments), and there are n observations per level: y_{ij}, $i = 1, \ldots, a$; $j = 1, \ldots, n$. We assume the observations, y_{ij}, satisfy the following statistical model:

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}. \tag{1}$$

Here μ is an overall mean, τ_i is the effect due to treatment level i, and ε_{ij} is a normally distributed error term with mean 0 and variance σ^2. The τ_i's are assumed to be deviations from the overall mean μ, so $\sum_{i=1}^{a} \tau_i = 0$. The treatment levels are specifically chosen by the analyst, so this is called a *fixed effects* model.

We want to test whether the factor treatments have a significant effect. The null and alternative hypotheses are:

$$H_0: \ \tau_1 = \tau_2 = \ldots = \tau_a = 0$$
$$H_1: \ \tau_i \neq 0 \ \text{ for at least one } i.$$

The test procedure is based on the following well-known equation, which partitions the total sum of squares of the data:

$$\begin{aligned} &\sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij} - \frac{1}{an}\sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}\right)^2 \\ &= n\sum_{i=1}^{a}\left(\frac{1}{n}\sum_{j=1}^{n} y_{ij} - \frac{1}{an}\sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}\right)^2 \\ &\quad + \sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij} - \frac{1}{n}\sum_{j=1}^{n} y_{ij}\right)^2. \end{aligned} \tag{2}$$

(For a derivation of (2), see Hines and Montgomery 1990.) We label the components of this equation as: $SS_T = SS_{treatments} + SS_E$. Under the null hypothesis, $SS_{treatments}/\sigma^2$ and SS_E/σ^2 are independently distributed chi-squared random variables with a-1 and $a(n$-1) degrees of freedom respectively. Therefore the test statistic:

$$F_0 = \frac{SS_{treatment}/(a-1)}{SS_E/a(n-1)}$$

has an $F_{a\text{-}1,a(n\text{-}1)}$ distribution. We reject H_0 if the test statistic is too large, i.e. if $F_0 > F_{\alpha,a\text{-}1,a(n\text{-}1)}$.

2.2 Data Collection for a Single Unknown Parameter

We begin with the simplest case, a single unknown parameter θ. Presumably we already have some information about θ, expressed as a confidence interval $[B_1, B_2]$. Suppose we have a vector of observations $\mathbf{X} = [X_1, \ldots, X_{n0}]$ that are random variables with mean θ and variance σ_X^2. A CI based on the central limit theorem is:

$$[B_1, B_2] \equiv \bar{X}(n_0) \pm z_{1-\alpha/2}\sqrt{S^2(n_0)/n_0}, \tag{3}$$

where $\bar{X}(n_0)$ is the sample mean of the n_0 observations, and $S^2(n_0)$ is the sample variance.

Define $L(\theta,\omega)$ to be a realization of the simulation model at fixed parameter value θ. Usually $L(\theta,\omega)$ is a performance measure (e.g. average waiting time). The expressions $E_L[L(\theta)]$ and $Var_L[L(\theta)]$ will refer the expectation and variance of the simulation output, where the subscript L refers to the randomness due to the simulation model. (This is as opposed to the randomness due to the observations $\mathbf{X}$.)

We would like to test whether θ has a significant effect on the expected simulation output, $E_L[L(\theta)]$, as θ varies over the range $[B_1, B_2]$. The approach will be to repeat n simulation replications at both θ=B_1 and θ=B_2, and perform an ANOVA test as described in the last section. Suppose $E_L[L(\theta)]$ is monotone in θ over $[B_1, B_2]$, so if the difference between treatments θ=B_1 and θ=B_2 is not significant, we can assume it is also not significant for any other pair of θ values in $[B_1, B_2]$. In this case our current estimator for θ, $\bar{X}(n_0)$, is precise enough; we cannot distinguish the effect of changing the parameter level within the range of our confidence interval. If the difference between treatments θ=B_1 and θ=B_2 is significant, the implication is that we should collect more observations X_i, so as to narrow the CI until the difference is no longer significant. Later in this section we will discuss the number of additional observations required. First, however, we discuss some technical considerations related to the use of ANOVA.

One issue with the procedure outlined in the last paragraph is that the error term ε_{ij} from model (1) is assumed to have a variance σ^2 that is independent of the factor level i. (This is known as *homoscedasticity*.) In the present framework this is equivalent to assuming $Var_L[L(\theta)] = \sigma^2$ for θ in a neighborhood of its true value. It is well known that $Var_L[L(\theta)]$ may change significantly with θ. We should therefore check the sample variances at θ=B_1 and θ=B_2. If they are different, a number of non-linear variance-stabilizing transformations are available (Casella and Berger 1990). Another approach, described by Kleijnen (1987), is to perform more replications at values of θ that show higher variance and average these to obtain "observations" with lower variance.

A second issue with the procedure is our assumption that $E[L(\theta)]$ is monotone in θ over $[B_1, B_2]$. In practice this is probably a mild assumption, especially if the width of the CI is small. However the benefit we gain is significant; the

assumption allows us to make an inference about an infinite number of values for θ, the entire range $[B_1, B_2]$. Furthermore, we might use a technique such as infinitesimal perturbation analysis to check the derivatives of $E[L(\theta)]$ at B_1 and B_2 and see if the assumption seems reasonable.

A third issue related to the procedure described above is the choice of n, the number of simulation replications made at $\theta=B_1$ and $\theta=B_2$. Recall that τ_1 is the effect of the treatment $\theta=B_1$, and τ_2 is the effect of the treatment $\theta=B_2$. Increasing the value of n raises the power of the hypothesis test, so if $\tau_1 \neq 0$ and/or $\tau_2 \neq 0$, then a large enough value for n will allow us to detect the difference, but how large a difference is it important to detect? This issue is addressed with a set of operating characteristic curves, each of which graphs the probability of type II error (β) for a particular sample size n against the following measure (Hines and Montgomery 1990):

$$\Phi^2 \equiv \frac{n(\tau_1^2 + \tau_2^2)}{2\sigma^2}.$$

These probabilities are derived from the fact that under the alternative hypothesis, F_0 has a noncentral F distribution. The parameter Φ^2 is a measure of the difference in means (relative to σ^2, the variability caused by the simulation) that it is important to detect. The analyst first determines an acceptable power for the test (1-β). Then, given a lower limit for Φ^2 that is important to detect, the analyst works backward through the operating characteristic curves to find a sample size n that will achieve (1-β). A set of curves is given by Pearson and Hartley (1972).

It turns out that we can interpret Φ^2 as a ratio of the variance due to parameter uncertainty and the variance due to simulation variability. In our construction τ_1 and τ_2 are random variables depending on B_1 and B_2. If we assume $\tau_1 = -\tau_2$, we have:

$$\begin{aligned}
\tau_1 &= E_L[L(B_1)] - \frac{1}{2}\{E_L[L(B_1)] + E_L[L(B_2)]\} \\
&= \frac{1}{2}E_L[L(B_1)] - \frac{1}{2}E_L[L(B_2)] \\
\tau_2 &= E_L[L(B_2)] - \frac{1}{2}\{E_L[L(B_1)] + E_L[L(B_2)]\} \\
&= \frac{1}{2}E_L[L(B_2)] - \frac{1}{2}E_L[L(B_1)].
\end{aligned}$$

Roughly speaking, the expectation of $(\tau_1^2 + \tau_2^2)/2$ is an upper bound on the variability of the simulation output due to parameter uncertainty. We derive this bound in the case where $E_L[L(\cdot)]$ is linear and $\alpha \leq 0.3174$. Suppose $E_L[L(x)] \equiv ax + b$. Using the expressions for τ_1 and τ_2 from the last paragraph and (3), we have:

$$\begin{aligned}
E_{\mathbf{X}}\left[\frac{\tau_1^2 + \tau_2^2}{2}\right] &= E_{\mathbf{X}}\left[\left(\frac{E_L[L(B_1)] - E_L[L(B_2)]}{2}\right)^2\right] \\
&= E_{\mathbf{X}}\left[\left(\frac{aB_1 - aB_2}{2}\right)^2\right] \\
&= \frac{a^2}{4}E_{\mathbf{X}}\left[\left(2z_{1-\alpha/2}\sqrt{\frac{S^2(n_0)}{n_0}}\right)^2\right] \\
&= \frac{a^2 z_{1-\alpha/2}^2 \sigma_X^2}{n_0}.
\end{aligned}$$

The last equality uses the fact that $E_{\mathbf{X}}(S^2(n_0)) = \sigma_X^2$. Now consider the variability of the simulation output due to parameter uncertainty, $Var_{\mathbf{X}}[E_L[L(\bar{X}(n_0))]]$. Again using $E_L[L(x)] \equiv ax + b$, we have:

$$\begin{aligned}
&Var_{\mathbf{X}}[E_L[L(\bar{X}(n_0))]] \\
&= E_{\mathbf{X}}\left\{\left(E_L[L(\bar{X}(n_0))] - E_{\mathbf{X}}[E_L[L(\bar{X}(n_0))]]\right)^2\right\} \\
&= E_{\mathbf{X}}\left\{\left(a\bar{X}(n_0) + b - E_{\mathbf{X}}[a\bar{X}(n_0) + b]\right)^2\right\} \\
&= a^2 E_{\mathbf{X}}\left\{\left(\bar{X}(n_0) - E_{\mathbf{X}}[\bar{X}(n_0)]\right)^2\right\} \\
&= a^2 Var_{\mathbf{X}}\{\bar{X}(n_0)\} \\
&= \frac{a^2\sigma_X^2}{n_0}.
\end{aligned}$$

If $\alpha \leq 0.3174$, then $z_{1-\alpha/2} \geq 1$, and:

$$E_{\mathbf{X}}\left[\frac{(\tau_1^2 + \tau_2^2)}{2}\right] \geq Var_{\mathbf{X}}[E_L(L(\bar{X}(n_0)))].$$

If the expectation of $(\tau_1^2 + \tau_2^2)/2$ is an upper bound on the variability due to parameter uncertainty, then the expectation of $(\tau_1^2 + \tau_2^2)/2\sigma^2$ is an upper bound for the ratio of variability caused by parameter uncertainty to variability caused by the simulation. The analyst determines a limit for this ratio (say 0.05), and an acceptable power for the test (1-β), and works backwards through the operating characteristic curves to find an appropriate sample size n.

Returning to the issue of data collection, if the factor effects at $\theta=B_1$ and $\theta=B_2$ are significant, we must collect additional observations X_i. How many more should we collect? The idea is to narrow the CI sufficiently so that we can no longer distinguish between the effects of the two parameter levels. Once we have determined how narrow the CI must be, we can estimate the number of additional observations required using (3). Recalling that the CI given by (3) is symmetric and centered at $\bar{X}(n_0)$, we define

the following procedure. First, let $\delta_0 \equiv z_{1-\alpha/2}\sqrt{S^2(n_0)/n_0}$, and choose a sequence $\{\delta_i\}$ such that for every i, $0 < \delta_{i+1} < \delta_i$. Set index i equal to zero. Now:

1. Perform $2n$ simulation replications, n each at $\bar{X}(n_0) \pm \delta_{i+1}$. Compute the ANOVA test statistic $F_0(\delta_{i+1})$, and compare this with $F_{\alpha,1,2(n-1)}$.
2. If $F_0(\delta_{i+1}) > F_{\alpha,1,2(n-1)}$, increment i and return to the first step; otherwise stop.

If the expected simulation output, $E[L(\theta)]$, is strictly monotone in θ over $[B_1,B_2]=[\bar{X}(n_0)-\delta_0, \bar{X}(n_0)+\delta_0]$, at each stage of the procedure the null hypothesis $H_0 : \tau_1(\delta_i) = \tau_2(\delta_i) = 0$ is, in fact, false. However since we have fixed n, eventually δ_i becomes so small that the test is not powerful enough to detect the difference between (τ_1, τ_2) and zero. If we choose a sequence $\{\delta_i\}$ that converges to zero, the procedure terminates with probability one. The output of the procedure is a half-width δ small enough that the analysis of variance is unable to distinguish between the treatments.

A useful choice for the sequence $\{\delta_i\}$ is one such that the number of additional real-world observations required to narrow the CI half-width from δ_i to δ_{i+1} is approximately constant. For a fixed positive integer η, let $\delta_i = z_{1-\alpha/2}\sqrt{S^2(n_0)/(n_0+i\eta)}$. If $S^2(n_0) \approx S^2(n_0+i\eta)$, then δ_i is approximately the half-width of the interval we would obtain by collecting $n_0+i\eta$ observations. Furthermore, from this definition it is apparent that $\{\delta_i\}$ converges to zero. If the procedure terminates at stage i_t, the number of additional observations required beyond n_0 is $i_t\eta$.

To illustrate this technique, we adapt a problem given by Law and Kelton (1991). A bank plans to install a new automated teller having a mean service time of 0.9 minutes. The bank has a midday busy period, during which the customer arrival rate is one per minute. The bank manager is interested in the performance of the new system, which he models as an m/m/1 queue. Suppose the service rate of the new machine has been supplied by the manufacturer, but the bank manager does not know the exact arrival rate, and he has collected n_0=1000 observations of the inter-arrival time. (These observations were drawn independently from an exponential distribution with mean 1.) He wants to know whether these data are sufficient for his simulation model, and if not, roughly how many more observations are required.

We begin by computing a 95% confidence interval for the mean of the arrival data, using (3). The information is listed in the table below.

Table 1: Summary Statistics for Inter-Arrival Data

# of Observations	Mean	Standard Deviation	CI Lower Limit	CI Upper Limit
1000	1.008	0.955	0.949	1.067

The next step is to perform a number of simulation replications at each extreme of the confidence interval. Suppose the α-value for the ANOVA test is 0.05, and we would like the power of the test to be at least 0.95 when the ratio $(\tau_1^2+\tau_2^2)/2\sigma^2 = 0.05$. Using Table 30 from Pearson and Hartley (1972), we see that when n = 126, then $\Phi = \sqrt{126 \cdot 0.05} = 2.51$, and the power of the test is 0.95. Therefore we perform 126 simulation replications at each extreme of the confidence interval for the mean inter-arrival time. This constitutes the first stage of the procedure (Step 0 in Table 2). We compute the F_0 statistic from the simulation data, which is 7.75. Since $F_{0.05,1,\infty}$ = 3.84, we reject the null hypothesis (H_0: $\tau_1 = \tau_2 = 0$) and conclude that we do not yet have sufficient real-world data.

Table 2: An Application of the Procedure of Section 2

Step i	δ_i	SS_T	SS_{treat}	SS_E	F_0
0	0.059	1579.12	47.49	1531.63	7.75
1	0.048	2285.23	205.86	2079.38	24.75
2	0.042	2056.68	99.68	1957.00	12.73
3	0.037	1991.19	35.61	1955.57	4.55
4	0.034	1983.84	30.05	1953.79	3.84

Next we narrow the confidence interval to half-width δ_1 and perform additional simulation replications. Suppose we choose an increment η = 500 observations, and let $\delta_i = z_{0.975}\sqrt{S^2(n_0)/(n_0+i\eta)}$. (The values of δ_i are listed in Table 2.) Again we compute the statistic F_0; the value is 24.75, and we conclude that 1000+η = 1500 observations are not yet enough. We continue in this way until i = 4, at which point F_0 = 3.84, and we no longer have sufficient evidence to reject the null hypothesis. The conclusion is that roughly 1000+4η = 3000 inter-arrival time observations will be sufficient data.

We close this section with a comment on the choice of $\{\delta_i\}$, $\delta_i = z_{0.975}\sqrt{S^2(n_0)/(n_0+i\eta)}$. Given this choice, the difference between δ_i and δ_{i+1} decreases as i increases. In fact, the ratio of δ_{i+1} to δ_i is:

$$\frac{\delta_{i+1}}{\delta_i} = \frac{z_{0.975}\sqrt{S^2(n_0)/(n_0+(i+1)\eta)}}{z_{0.975}\sqrt{S^2(n_0)/(n_0+i\eta)}} = \sqrt{\frac{n_0+i\eta}{(n_0+i\eta)+\eta}},$$

which approaches 1 as i approaches infinity. Therefore while the procedure is guaranteed to terminate, with this choice of $\{\delta_i\}$ it may take a long time to do so.

Another choice for $\{\delta_i\}$ is $\delta_{i+1} = \gamma\delta_i = \gamma^{i+1}\delta_0$ for a fixed constant γ. In this case the number of additional real-world observations required to narrow the CI half-width from δ_i to δ_{i+1} is no longer constant. To see this, we first find the number of observations n_i required to achieve a CI of half-width δ_i by solving $\delta_i = z_{0.975}\sqrt{S^2(n_i)/n_i}$ for n_i. If the sample variance $S^2(n)$ is roughly constant in the number of observations n, then we have:

$$\begin{aligned} n_i &= S^2(n_i)\cdot\left(\frac{z_{0.975}}{\delta_i}\right)^2 \\ &\approx S^2(n_0)\cdot\left(\frac{z_{0.975}}{\delta_i}\right)^2 \\ &= S^2(n_0)\cdot\left(\frac{z_{0.975}}{\gamma^i\cdot z_{0.975}\sqrt{S^2(n_0)/n_0}}\right)^2 \\ &= \frac{n_0}{\gamma^{2i}} \end{aligned}$$

Likewise $n_{i+1} \approx \frac{n_0}{\gamma^{2(i+1)}} \approx \frac{n_i}{\gamma^2}$, so to narrow the CI half-width from δ_i to δ_{i+1} we must multiply the overall number of observations by γ^{-2}.

2.3 Data Collection for Multiple Unknown Parameters

We next turn to the case where there are two unknown parameters, $\boldsymbol{\theta} = [\theta_1, \theta_2]$. The approach will be similar to that taken in the last section. Suppose we have observations $\{X_{i,1} : i = 1, \ldots, n_{0,1}\}$ that are random variables with mean θ_1 and variance $\sigma^2_{X,1}$, and observations $\{X_{i,2} : i = 1, \ldots, n_{0,2}\}$ that are random variables with mean θ_2 and variance $\sigma^2_{X,2}$. Let $[B_{11}, B_{21}]$ and $[B_{12}, B_{22}]$ be confidence intervals for θ_1 and θ_2 respectively, of the form given by:

$$\bar{X}(n_0) \pm z_{1-\alpha/4}\sqrt{S^2(n_0)/n_0}\,.$$

Let $\Xi = [B_{11}, B_{21}] \times [B_{12}, B_{22}]$, so by the Bonferroni inequality, Ξ is a (1-α)100% confidence interval for $\boldsymbol{\theta}$. We would like to test whether $\boldsymbol{\theta}$ has a significant effect on the expected simulation output, $E[L(\boldsymbol{\theta})]$, as it varies over Ξ. We will repeat n simulation experiments at each of the four combinations of levels determined by the endpoints of the confidence intervals, and perform a two-way ANOVA. (For a description of the two-way ANOVA, refer to Hines and Montgomery 1990.) This time we will test for the main effects of θ_1 and θ_2, as well as an interaction effect. If none of these are significant, we can assume our current estimators for θ_1 and θ_2 are sufficiently precise; we cannot distinguish the effect of parameter levels within the range of our confidence intervals. If the main or interaction effects are significant, the implication is that we should collect more real-world observations.

Let $\gamma_{0,1} \equiv z_{1-\alpha/4}\sqrt{S_1^2(n_{0,1})/n_{0,1}}$ and $\gamma_{0,2} \equiv z_{1-\alpha/4}\sqrt{S_2^2(n_{0,2})/n_{0,2}}$, which are the half-widths of the original CIs. Choose sequences $\{\gamma_{i,1}\}$ and $\{\gamma_{i,2}\}$ such that for every i, $0 < \gamma_{i+1,1} < \gamma_{i,1}$ and $0 < \gamma_{i+1,2} < \gamma_{i,2}$. We also define variables $\varphi_1(i)$ and $\varphi_2(i)$ that will be indices into the sequences $\{\gamma_{i,1}\}$ and $\{\gamma_{i,2}\}$, respectively. Let $\varphi_1(0) = \varphi_2(0) = 0$. We define the following procedure. Let $i = 0$.

1. Let $\delta_{i,1} = \gamma_{\varphi_1(i),1}$ and $\delta_{i,2} = \gamma_{\varphi_2(i),2}$.
2. Perform 4n simulation replications, n each at the four combinations of $\theta_1 = \bar{X}_1(n_{0,1}) \pm \delta_{i,1}$ and $\theta_2 = \bar{X}_2(n_{0,2}) \pm \delta_{i,2}$. Compute the ANOVA test statistics for the main and interaction effects of θ_1 and θ_2.
3. If none of the effects are significant then stop. Otherwise:
   - If the main effect of θ_1 is significant, set $\varphi_1(i+1) = \varphi_1(i) + 1$. Otherwise set $\varphi_1(i+1) = \varphi_1(i)$.
   - If the main effect of θ_2 is significant, set $\varphi_2(i+1) = \varphi_2(i) + 1$. Otherwise set $\varphi_2(i+1) = \varphi_2(i)$.
   - If neither of the main effects are significant but the interaction effect is, compare $(\gamma_{\varphi_1(i),1} - \gamma_{\varphi_1(i)+1,1})$ and $(\gamma_{\varphi_2(i),2} - \gamma_{\varphi_2(i)+1,2})$. If the former is larger, set $\varphi_1(i+1) = \varphi_1(i) + 1$ and $\varphi_2(i+1) = \varphi_2(i)$. Otherwise set $\varphi_1(i+1) = \varphi_1(i)$ and $\varphi_2(i+1) = \varphi_2(i) + 1$.
4. Increment the counter i and return to step 1.

If the sequences $\{\gamma_{i,1}\}$ and $\{\gamma_{i,2}\}$ both converge to zero, the procedure terminates with probability one. Again it may be convenient to choose $\{\gamma_{i,1}\}$ and $\{\gamma_{i,2}\}$ so that the number of observations required to narrow the CI from one step to the next is roughly constant. Thus for fixed values η_1 and η_2, let $\gamma_{i,1} = z_{1-\alpha/4}\sqrt{S_1^2(n_0)/(n_0 + i\eta_1)}$ and $\gamma_{i,2} = z_{1-\alpha/4}\sqrt{S_2^2(n_0)/(n_0 + i\eta_2)}$. (Given per-observation costs for each parameter, the values η_1 and η_2 might be chosen so that the cost of a step for either parameter is the same.)

This procedure can be easily extended to $k>2$ parameters. However one difficulty is that a prohibitively large number of simulation replications may be required: $2^k \cdot n$ replications at each stage of the procedure. The assumption must be that simulation data is much less expensive than real world observations. When this assumption is false, we can think about using other experimental designs. For example, if we assume that high-order interaction effects are negligible, we can use 2^{k-p} fractional designs that require many fewer replications. (See Hines and Montgomery 1990 and Kleijnen 1987 for details.) Other techniques for screening extremely large numbers of factors are described by Kleijnen (1987). These include random designs, in which factor levels are selected for testing with equal probabilities, and group-screening designs, in which the effects of several factors are grouped so that they all may be eliminated (deemed insignificant) at once.

3 RANDOM EFFECTS MODEL

An assumption made in Section 2 was that the expected simulation output, $E[L(\theta)]$, was monotone in θ over the parameter's confidence interval. This assumption was made so that by testing the extremes of the CI, we could draw conclusions about values of θ within the CI. As an alternative to this assumption, we can apply the random effect model for ANOVA. This model is used when the number of levels for a particular factor is infinite (as it is in our case), and the analyst wants to make inferences about the entire population of factor levels.

3.1 Review of the Random Effects ANOVA

We summarize the description of the random effects model for a single factor, given by Hines and Montgomery (1990). The analyst performs n simulation replications at each of a randomly selected factor levels. (The levels are selected with equal probability from the population of factor levels.) We assume the observations $\{y_{ij} : i = 1,\ldots,a;\ j = 1,\ldots,n\}$ satisfy the following statistical model:

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}. \tag{4}$$

Here μ is the overall mean, and τ_i and ε_{ij} are independent normally distributed random variables with mean 0 and variances σ_τ^2 and σ^2 respectively. We would like to test whether the treatments have any effect. Since the treatments are identical if $\sigma_\tau^2 = 0$, the null hypothesis is:

$$H_0 : \sigma_\tau^2 = 0.$$

Equation (2) still holds: $SS_T = SS_{treatments} + SS_E$, where $SS_{treatments}$ and SS_E are defined as before. One can show that SS_E/σ^2 has a $\chi^2(na-a)$ distribution, and under H_0 the distribution of $SS_{treatments}/\sigma^2$ is $\chi^2(a-1)$, independent of SS_E/σ^2. Therefore the test statistic is:

$$F_0 = \frac{SS_{treatments}/(a-1)}{SS_E/(na-a)},$$

which under the null hypothesis is distributed $F_{a-1,na-a}$. We reject the null hypothesis if $F_0 > F_{\alpha,a-1,na-a}$.

3.2 Data Collection for a Single Unknown Parameter

Again we will start with the simplest case, a single unknown simulation parameter. Suppose we have already collected n_0 real-world observations (each with mean θ and variance σ_X^2). Again let $\mathbf{X} = [X_1,\ldots,X_{n_0}]'$ be the vector of observations, and let $\bar{X}(n_0)$ be the sample mean of the observations. We would like to use $\bar{X}(n_0)$ as a surrogate for θ in our study since presumably the value of $\bar{X}(n_0)$ will be close θ. If we think of performing the simulation study over and over with different realizations of $\bar{X}(n_0)$, we would like to know whether there would be any significant effect on the simulation output. Clearly as n_0 increases any such effect will diminish since $\bar{X}(n_0)$ converges to θ. Therefore if there is a significant effect at n_0, we also want to know how many additional observations are required for this effect to become undetectable.

The idea is to use the random effects model to test whether the expected simulation output is constant over the population of realizations of $\bar{X}(n_0)$. We therefore must generate realizations of $\bar{X}(n_0)$, which requires us to make draws from its distribution. Unfortunately this distribution depends on the unknown variable θ. However this is precisely the situation where bootstrapping is appropriate. Therefore we approximate the distribution by drawing bootstrap samples from the vector of observations $\mathbf{X}$.

A bootstrap sample of size n is a set of n observations drawn with replacement from the elements of $\mathbf{X}$. In other words, it is a set of n observations drawn from the empirical distribution function corresponding to the observations in $\mathbf{X}$. (See Efron and Tibshirani 1993 for details.) A bootstrap replication of $\bar{X}(n_0)$, represented as $\bar{X}^*(n_0)$, is the sample average of a bootstrap sample of size n_0.

We now define the following procedure for determining the approximate number of observations required in addition to n_0. First, choose a positive integer η, which

will be the incremental number of observations at each stage of the procedure. Let $N = n_0$.

1. Generate a bootstrap samples of size N from the data in $\mathbf{X}$, and compute the bootstrap replication $\bar{X}^*(N)$ for each. Perform n simulation replications at each of the a values of $\bar{X}^*(N)$.
2. Compute the test statistic F_0 from the data generated in the first step. If $F_0 < F_{\alpha,a-1,na-a}$, stop. Otherwise let $N = N + \eta$ and return to the first step.

This procedure terminates with probability one since eventually the $\bar{X}^*(N)$'s will converge to the expectation of the empirical distribution, at which point there is no difference in treatments. The output of the procedure is the number of additional observations, (N-n_0), required for the analysis of variance to be unable to detect the effect of the uncertainty of $\bar{X}(N)$.

As with the procedure of Section 2, there is an issue concerning how to choose n, the number of simulation replications. Once again this issue is addressed with a set of operating characteristic curves, which for particular values of n plot the probability of type II error against the following measure:

$$\lambda = \sqrt{1 + \frac{n\sigma_\tau^2}{\sigma^2}}.$$

The analyst chooses a power for the test and a value of λ that is important to detect, and then works backward through the operating characteristic curves to find an appropriate n. (See Hines and Montgomery 1990 for details.) In this case the measure λ is easy to interpret, for σ_τ^2/σ^2 is the ratio of variance due to parameter uncertainty to variance due to the simulation. A set of operating characteristic curves for the random effects model is given by Hines and Montgomery (1990).

We mention in passing an advantage of this procedure over the one given in Section 2: we are not required to produce joint confidence intervals for each unknown parameter. In some cases it may be difficult or impossible to compute a CI, or to produce good joint CIs for multiple parameters from a single data set. The procedure in this section is applicable to any set of parameters for which we can generate bootstrap samples. On the other hand, the computational effort can be much greater. For the fixed effects model with power 0.95 and $(\tau_1^2 + \tau_2^2)/2\sigma^2 = 0.05$, the number of simulation replications n required at each factor level is 126; we make 2×126 = 252 replications at each step of the procedure. For the random effects model with a = 2, power 0.95, and $\sigma_\tau^2/\sigma^2 = 0.05$, the value of n is 16,800; we make 2×16,800 = 33,600 replications at each step of the procedure. The reason for this disparity is that the random effects model draws inferences about an infinite number of factor levels (parameter values). The fixed effects model draws inferences about only two factor levels; in order to make inferences about factor levels within the range of the CI, we add the assumption that the expected simulation response is monotone.

To demonstrate this procedure we return to the bank teller example, where we will use the same 1000 "observed" inter-arrival times. Again suppose the α-value for the ANOVA test is 0.05, and we would like the power of the test to be at least 0.95 when the ratio $\sigma_\tau^2/\sigma^2 = 0.05$. Using Appendix VIII from Hines and Montgomery (1990), we see that when n = 240 and a = 5, then λ = 3.6, and the power of the test is 0.95. Therefore we will perform 240 simulation replications at each of five bootstrap replications of the mean inter-arrival time.

Table 3: An Application of the Procedure of Section 3

Step i	Bootstrap #1	Bootstrap #2	Bootstrap #3	Bootstrap #4	Boot strap #5	F_0
0	1.020	0.997	1.056	1.041	0.991	3.94
1	0.996	1.016	1.003	1.020	0.980	2.86
2	1.015	0.990	1.047	1.035	1.013	6.26
3	1.009	1.031	1.026	1.003	1.006	1.59

The five bootstrap replications of size 1000 are listed in the row of Table 3 labeled "Step 0." After performing the simulation runs at these values, we compute the F_0 statistic, which is 3.94. Since $F_{0.05,4,\infty}$ = 2.37, we reject the null hypothesis $(H_0 : \sigma_\tau^2 = 0)$ and conclude that we do not yet have sufficient real-world data.

Next we increase the size of the bootstrap samples used to compute the bootstrap replications of the mean inter-arrival time. Suppose we again choose an increment η = 500 observations, so the size of the new bootstrap samples will be 1000+η = 1500 observations. The new bootstrap replications of the mean inter-arrival time are listed in the row of Table 5.5 labeled "Step 1." Again the value of F_0 is greater than $F_{0.05,4,\infty}$ = 2.37, and we conclude that 1500 observations are not enough. We continue in this manner until Step 3, when the value of the statistic F_0 is 1.59, which is less than $F_{0.05,4,\infty}$. We no longer have sufficient evidence to reject the null hypothesis, and the conclusion is that 1000+3η = 2500 observations of the inter-arrival time are sufficient.

3.3 Data Collection for Multiple Unknown Parameters

The procedure for two unknown parameters, $\boldsymbol{\theta} = [\theta_1, \theta_2]$, is similar to the one presented in the previous section. (See

Hines and Montgomery 1990 for a description of the random effects two-way ANOVA.) Suppose we have observations $\{X_{i,1} : i = 1,\ldots,n_{0,1}\}$ that are random variables with mean θ_1 and variance $\sigma^2_{X,1}$, and observations $\{X_{i,2} : i = 1,\ldots,n_{0,2}\}$ that are random variables with mean θ_2 and variance $\sigma^2_{X,2}$. Choose positive integers η_1 and η_2, which will be the incremental numbers of observations of the two parameters at each stage of the procedure. Let $N_1 = n_{0,1}$ and $N_2 = n_{0,2}$.

1. Generate a bootstrap samples of size N_1 from $\{X_{i,1} : i = 1,\ldots,n_{0,1}\}$, and compute bootstrap replications of $\overline{X}_1$ for each: $\overline{X}^*_{1,1}(N_1),\ldots,\overline{X}^*_{a,1}(N_1)$. Generate b bootstrap samples of size N_2 from $\{X_{i,2} : i = 1,\ldots,n_{0,2}\}$, and compute bootstrap replications of $\overline{X}_2$ for each: $\overline{X}^*_{1,2}(N_2),\ldots,\overline{X}^*_{b,2}(N_2)$.
2. Perform n simulation replications at each of the ab combinations of

$$\{\overline{X}^*_{1,1}(N_1),\ldots,\overline{X}^*_{a,1}(N_1)\}\times\{\overline{X}^*_{1,2}(N_2),\ldots,\overline{X}^*_{b,2}(N_2)\}.$$

3. Compute the test statistics from the data generated in the second step. If none of the effects are significant then stop. Otherwise
   - If the main effect of θ_1 is significant, set $N_1 = N_1 + \eta_1$.
   - If the main effect of θ_2 is significant, set $N_2 = N_2 + \eta_2$.
   - If neither of the main effects are significant but the interaction effect is, set $N_1 = N_1 + \eta_1$ and set $N_2 = N_2 + \eta_2$.
   - Return to step one.

This procedure again terminates with probability one. The output is the number of additional observations, (N_1-$n_{0,1}$) and (N_2-$n_{0,2}$), required for parameters $\boldsymbol{\theta} = [\theta_1,\theta_2]$.

4 TOPICS FOR FUTURE RESEARCH

We are working to compare the effectiveness of the two procedures developed in Sections 2 and 3. We are also investigating the connection between these procedures and the techniques for screening large number of factors mentioned in Section 1. Finally, these techniques have been developed under the implicit assumption that the cost of collecting information about each parameter is the same. We are also constructing a model that will relate the costs of data collection to the cost of the uncertainty in the simulation output caused by parameter estimation.

ACKNOWLEDGMENTS

The research reported here was partially supported by a joint src (fj-490) and nsf (dmi-9713549) research project in semiconductor operations modeling.

REFERENCES

Casella G., and R. L. Berger. 1990. *Statistical Inference.* Belmont, CA: Duxbury Press.

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

Hines, W., and D. Montgomery. 1990. *Probability and Statistics in Engineering and Management Science.* 3d ed. New York: John Wiley and Sons.

Kleijnen, J.P.C. 1987. *Statistical Tools for Simulation Practitioners.* New York: Marcel Dekker.

Law, A., and W.D. Kelton. 1991. *Simulation Modeling & Analysis.* New York: McGraw-Hill.

Pearson, E. S., and H. O. Hartley. 1972. *Biometrika Tables for Statisticians: Volume II.* Cambridge, Great Britain: Cambridge University Press.

AUTHOR BIOGRAPHIES

MIKE FREIMER is a Visiting Assistant Professor at Cornell's School of Hotel Administration and a lecturer at the School of Operations Research and Industrial Engineering. He received a Ph.D. from the School of ORIE in 2001. His undergraduate degree is in mathematics from Harvard. Prior to attending graduate school he worked for an operations research consulting firm, Applied Decision Analysis, Inc. in Menlo Park, CA. His research interests are in simulation modeling and operations management. His email address is <mfreimer@orie.cornell.edu>.

LEE SCHRUBEN is a Professor in the Department of Industrial Engineering and Operations at the University of California at Berkeley. His research interests are in statistical design and analysis of simulation experiments and in graphical simulation modeling methods. His simulation application experiences and interests include semiconductor manufacturing, dairy and food science, health care, banking, and the hospitality industry. His email address is <schruben@ieor.berkeley.edu>.