Proceedings of the 1997 Winter Simulation Conference
ed. S. Andradóttir, K. J. Healy, D. H. Withers, and B. L. Nelson

SELECTING THE BEST SYSTEM: A DECISION-THEORETIC APPROACH

Stephen E. Chick

Department of Industrial and Operations Engineering
The University of Michigan
1205 Beal Avenue
Ann Arbor, Michigan 48109-2117, U.S.A.

ABSTRACT

The problem of selecting the best system from a finite set of alternatives is considered from a Bayesian decision-theoretic perspective. The framework presented is quite general, and permits selection from two or more systems, with replications that use either independent or common random numbers, with unknown mean and covariance for the output, and permits Gaussian or non-Gaussian simulation output. For the case of unknown means and variance with common random numbers, the framework provides a probability of correct selection that does not suffer from problems associated with the Bonferroni inequality. We indicate some criteria for which the Bayesian approach and other approaches are in general agreement, or disagreement. The probability of correct selection can be calculated either by quadrature or by Monte Carlo simulation from the posterior distribution of the parameters of the statistical distribution of the simulation output. We also comment on expected-value decision-making versus optimization criteria based on other functionals of the distribution of the output.

1 MOTIVATION

One important application for stochastic simulation is the selection of the best system from a set of alternatives (Law and Kelton, 1991). Goldsman and Nelson (1994) provide a comprehensive literature review of ranking, selection, and comparison techniques for selecting the best system and related problems (screening a large number of systems, comparing all systems to a standard system and comparing all systems to a default).

Two features of many of the techniques are (1) estimations of means (or means of differences) of performance measures for the systems and (2) a measure of the evidence that the correct decision was made, expressed in terms of P-values of hypothesis tests. (Multiple comparison procedures do not fall into this category and are discussed in Section 6.)

When the mean and variance are unknown and common random numbers are employed, the measure of evidence is usually based on P-values of multiple comparisons and the Bonferroni inequality. This inequality can significantly misstate the overall confidence of a selection. More generally, the use of P-values as a measure of evidence had been called into question by a number of researchers (see Appendix A). The present work provides an alternate probabilistic measure of evidence that extends a recently proposed Bayesian framework for analyzing the simulation output of a single system (Chick, 1997). The relevant assumptions required from that paper are given in Section 2.

The selection problem for independent simulation replications is covered in Section 3. An extension to handle dependencies such as common random numbers is given in Section 4. Arbitrary correlation between the output of each system is permitted, and no assumptions are made regarding the values of the mean and variance, other than that they are unknown quantities. Approximations that indicate a relation with frequentist techniques is given. An example of selecting the best system with CRN is given in Section 5. The relevance of BEM for estimating the probability of correct selection is discussed in Section 6. Selection criteria using functionals of the output distribution other than the mean are discussed in Section 7. Section 8 concludes with some comments on the approach, particularly with respect to decision theory.

Rather than speaking of mean values of output, as is typical in a simulation context, this paper speaks of expected utility as the output of interest in order to remain consistent with the decision theory terminology.

2- ASSUMPTIONS

Suppose there are K different simulated systems, and the objective is to select the system with the maximum expected utility. We make the following additional assumptions for the case of independent replications. Modified assumptions for dependent replications are given in Section 4.

1. The simulation output $O_{k,r}$ for system k, replication r ($k = 1, \ldots, K; r = 1, \ldots, R_k$), is independent from replication to replication.

2. The distribution of $O_{k,r}$ has density

$$p_{O_k|\theta_k}(o_{k,r}) = f_{O_k|\theta_k}(o_{k,r})do_k$$

with continuous (possibly multidimensional) parameters θ_k, $k = 1, \ldots K$.

3. The θ_k are initially unknown, with prior probability $p_{\Theta_1,\ldots,\Theta_K}(\theta_1, \ldots, \theta_K) = \prod_{k=1}^{K} p_{\Theta_k}(\theta_k)$.

4. A real-valued utility function u measures the utility $u(k, o_k)$ of an outcome o_k for system k.

5. The parameters for the input distributions to the simulation are fixed.

Assumption 5 simplifies the presentation by neglecting uncertainty about the distributions that describe randomness in the system behavior.

Define $\vec{v}(\theta) = (v_1(\theta_1), \ldots, v_K(\theta_K))$ to be the expected conditional utilities of each system, given θ, where $v_k(\theta_k) = E[u(k, O_k) \mid \theta_k]$. Because θ_k is unknown, the expected conditional utility is a random variable whose distribution depends on the distribution of θ_k. Write upper case $\Upsilon_k = E[u(k, O_k) \mid \Theta_k]$ for the random variable, and lower case v_k for its outcome. As more replications are run, more is learned about the distribution of Θ_k, and therefore about the unconditional expected utility.

The distribution of $\vec{\Upsilon} = (\Upsilon_1, \ldots, \Upsilon_K)$ is used to select the best system and provide a probabilistic measure of evidence of correct selection.

3- INDEPENDENT REPLICATIONS

An important case of selecting the best of K systems arises when all replications for all systems are run with independent random variables.- Suppose R_k independent replications are run for system k, $k = 1, \ldots K$. The assumptions in Section 2 lead to the following.

The expected utility v_k of a system is unknown, whose distribution can be determined from the likelihood $f_{O_k|\theta_k}(o_k)$, the prior distribution $p_{\Theta_k}(\theta_k)$ and Bayes' rule.

$$p_{\Theta_k|\mathcal{D}}(\theta_k) = A_k(\mathcal{D}) p_{\Theta_k}(\theta_k) \prod_{r=1}^{R_k} f_{O_k|\theta_k}(o_{k,r}), \quad (1)$$

where $\mathcal{D} = \{o_{k,r} \mid \text{all } k, r\}$ is the output from the simulation replications, and $A_k(\mathcal{D})$ does not depend on θ_k. The posterior predictive distribution for a future value $o_{k,R_{k+1}}$ of output is

$$p_{O_k|\mathcal{D}}\left(o_{k,R_{k+1}}\right) = \int_{\theta_k} f(o_{k,R_{k+1}} \mid \theta_k) p_{\Theta_k|\mathcal{D}}(\theta_k)$$

Finally, the posterior marginal distribution for the expected utility of system k is found by:

$$p_{\Upsilon_k|\mathcal{D}}(v_k) = \int_{o|U(k,o)=v_k} p_{O|\mathcal{D}}(o) \quad (2)$$

The integral takes a particularly simple form if the utility is the output (i.e., $p_{\Upsilon|\mathcal{D}}(v) = p_{O|\mathcal{D}}(o)$).

One is now in a position to describe which system is best with which probability from the posterior distributions for each of the systems. By the independence assumption and Equation (2), system i is better than system k (the event $\{v_i \geq v_k\}$) with probability

$$p_{\Upsilon_i,\Upsilon_k|\mathcal{D}}(v_i \geq v_k) = \iint_{v_i \geq v_k} p_{\Upsilon_i|\mathcal{D}}(v_i)\, p_{\Upsilon_k|\mathcal{D}}(v_k).$$

The probability that system i is best is then

$$\begin{aligned} p(i \text{ best} \mid \mathcal{D}) &= p_{\vec{\Upsilon}|\mathcal{D}}(v_i \geq v_k, \text{ for all } k \neq i) \\ &= \int_{\mathcal{A}_i} p_{\vec{\Upsilon}|\mathcal{D}}(\vec{v}), \end{aligned} \quad (3)$$

where the domain of integration $\mathcal{A}_i = \{\vec{u} \mid \vec{u}_i \geq \vec{u}_k \text{ for all } k \neq i\}$ is the set of points where the utility of system i is at least as great as the utility of the other systems. Note that in general $p(i \text{ best} \mid \mathcal{D}) \neq \prod_{k|k\neq i} p_{\Upsilon_i,\Upsilon_k|\mathcal{D}}(v_i \geq v_k)$, in spite of the independence of the Υ_k.

3.1- Approximations

Suppose the output is the univariate utility, $u(k, o_{k,r}) = o_{k,r}$. Approximate the distribution of the output of each system with a Gaussian distribution with unknown mean v_k and variance $\sigma_k^2 = \tau_k^{-1}$. Further suppose that τ_k has a gamma prior distribution $\mathbf{Ga}(\alpha_k, \beta_k)$, and that v_k, given $\sigma_k^2 = \tau_k^{-1}$, has Gaussian conditional distribution $\mathbf{N}_1\left(\mu_{0k}, \sigma_k^2/n_{0k}\right)$. This

generalizes the work of Andrews and Schriber (1983) and Andradóttir and Bier (1997), where the variance is assumed known.

Then the posterior distribution of the expected (conditional) utility has Student distribution

$$p_{\Upsilon_k|\mathcal{D}}(\upsilon_k) \sim \mathbf{St}_1\left(\mu_{R_k}, \frac{\zeta_{R_k}}{\beta_{R_k}}, 2\alpha + R_k\right) \qquad (4)$$

where

$$\begin{aligned}
\mu_{R_k} &= \frac{n_{0k}\mu_{0k} + R\bar{o}_k}{n_{0k} + R_k} \\
\zeta_{R_k} &= (n_{0k} + R_k)(\alpha_k + \frac{R_k}{2}) \\
\beta_{R_k} &=' \beta + \frac{s_k}{2} + \frac{n_{0k}R_k(\mu_{0k} - \bar{o}_k)^2}{2(n_{0k} + R_k)} \\
\bar{o}_k &= \sum_{r=1}^{R_k} o_{k,r}/R_k \\
s_k &= \sum_{r=1}^{R_k} (o_{k,r} - \bar{o}_k)^2.
\end{aligned}$$

The Student distribution with similar parameters plays a role in a standard frequentist confidence interval for the mean (Chick, 1997). Thus, inferences based on approximating the output with a normal distribution, together with the conjugate normal-gamma prior, can be related to classical inferences.

4- DEPENDENT REPLICATIONS

Techniques which induce dependent output from system to system between replications, such as common random numbers (CRN), can lead to significant savings of computational effort for system analysis. It is therefore desirable to extend the techniques of Section 3 to account for such dependencies.

If two systems see similar stochastic inputs (e.g. services times, inter-arrival times) it is plausible that a large realization of utility for system i, may increase the probability that the realized utility for system k will be large. For example, particular sequences of customer arrival times and service times may lead to a large utility for both an M/M/1 system and an M/M/2 system (where utility might be the negative of the waiting time).

4.1- General Theory

Assume that replications for each system can be paired, set R to be the number of replications of each system, set $\vec{o}_r = (o_{1,r}, \ldots, o_{K,r})$ to be the vector of outputs from the r-th replication of each system. Further assume that the vectors $\vec{o}_r, \vec{o}_{r'}$ are independent when $r \neq r'$, but that the components $o_{1,r}, \ldots, o_{K,r}$ are not necessarily independent for a given r. This assumption permits CRN to be analyzed.

Set $\vec{\theta} = (\theta_1, \ldots, \theta_K, \theta_{K+1})$ to be the vector of parameters for the output distributions, where θ_{K+1} determines dependencies between output from different systems. When $K > 2$, θ_{K+1} will generally be multivariate. When output is correlated, the expected conditional utilities $\vec{v} = (\upsilon_1, \ldots, \upsilon_K)$ are correlated. Let $p_{\vec{\Theta}}\left(\vec{\theta}\right)$ be the prior distribution for $\vec{\theta}$. Then the posterior distribution for $\vec{\Theta}$ is

$$p_{\vec{\Theta}|\mathcal{D}'}\left(\vec{\theta}\right) \propto p_{\vec{\Theta}}\left(\vec{\theta}\right) \prod_{r=1}^{R} f_{\vec{O}_r|\vec{\theta}}(\vec{o}_r), \qquad (5)$$

where $\mathcal{D}'$ is the correlated output from the simulation replications, and $f_{\vec{O}_r|\vec{\theta}}(\vec{o}_r)$ is the conditional probability of seeing the given outputs for a given set of parameters for the output distribution.

By analogy with the independent replication case, output from future replications can be predicted with

$$p_{\vec{O}|\mathcal{D}'}(\vec{o}_{R+1}) = \int_{\vec{\theta}} f_{\vec{O}|\vec{\theta}}(\vec{o}_{R+1}) p_{\vec{\Theta}|\mathcal{D}'}\left(\vec{\theta}\right)$$

The induced posterior distribution for the expected utility $\vec{v}$ of all systems is found by:

$$p_{\vec{\Upsilon}|\mathcal{D}'}(\vec{v}) = \int_{\vec{\theta}|\vec{\Upsilon}(\vec{\theta})=\vec{v}} p_{\vec{\Theta}|\mathcal{D}'}\left(\vec{\theta}\right)$$

The probability that system i is the best system is then calculated from the joint distribution of $\vec{v}$ as

$$p(\text{system } i \text{ is best}) = \int_{\mathcal{A}_i} p_{\vec{\Upsilon}|\mathcal{D}'}(\vec{v}) \qquad (6)$$

where the domain of integration $\mathcal{A}_i = \{\vec{u} \mid \vec{u}_i \geq \vec{u}_k \text{ for all } k \neq i\}$ is the set of points where the utility of system i is at least as great as the utility of the other systems.

The definition of $\mathcal{A}_i$ may be modified to allow for more general statements such as 'system i has an expected utility which is not less than 1 below the maximum utility', $\{\vec{u} \mid \vec{u}_i \geq \vec{u}_k - 1 \text{ for all } k \neq i\}$. Thus the present theory can be modified to handle a more general definition of 'best'.

4.2- Approximations and Asymptotic Results

The selection result of Section 4.1 requires a large number of parameters to be estimated simultaneously, and imposes difficult numerical integration problems before the best system can be selected. This section examines simplifications and approximations.

Appendix B contains reference information regarding the probability distributions used in this section.

First assume that the simulation output is the utility, $\vec{U} = \vec{O}$. Second, assume that the distribution of $\vec{U}$ is joint Gaussian with unknown parameter $(\vec{v}, \Sigma)$. This is often a reasonable assumption, as in cases where the utility satisfies a functional central limit theorem. Third, assume the appropriate multivariate generalization of the approximation in Section 3.1, a conjugate normal-Wishart distribution for $(\vec{v}, \Sigma)$,

$$\begin{aligned} p_{\vec{U}_r|\vec{v},\Sigma}(\vec{u}_r) &\sim \mathbf{N}_K(\vec{v}, \Sigma) \\ \pi(\tau) &\sim \mathbf{W}_K(\alpha_0, \beta_0) \\ \pi(\vec{v} \mid \Sigma) &= \mathbf{N}_K(\mu_0, \Sigma/n_0) \end{aligned}$$

where $\vec{u}_r$ is the correlated output from simulation replication r from each system, $(\vec{v}, \Sigma)$ are the unknown parameters of the Gaussian output distribution, and whose values are to be inferred through simulation analysis, and $\tau = \Sigma^{-1}$ is the inverse of the unknown covariance matrix. Prior information on τ is represented with a Wishart distribution $\mathbf{W}_K(\alpha_0, \beta_0)$, for some α_0, β_0 specified by the analyst, and prior information on $\vec{v}$ given $\tau = \Sigma^{-1}$ is represented by a Gaussian distribution with parameters $\mu_0, \Sigma/n_0$, where μ_0, n_0 are specified by the analyst.

Construct sample statistics from R simulation replications, the sample K-variant mean $\bar{o}$ and a sample $K \times K$ covariance S, where

$$\begin{aligned} \bar{o} &= \sum_{r=1}^{R} \frac{o_r}{R} \\ S &= \sum_{r=1}^{R} (o_r - \bar{o})(o_r - \bar{o})^t. \end{aligned} \tag{7}$$

The posterior distribution for $\tau, \vec{v}$ is (Bernardo and Smith, 1994)

$$\begin{aligned} p(\tau \mid \bar{o}, S) &\sim \mathbf{W}_K(\alpha_0 + R/2, W_1(\bar{o}, S)) \\ p(\vec{v} \mid \Sigma, \bar{o}, S) &\sim \mathbf{N}_K\left(\frac{n_0\mu_0 + R\bar{o}}{n_0 + R}, \frac{\Sigma}{n_0 + R}\right) \end{aligned}$$

where

$$W_1(\bar{o}, S)^{-1} = \beta_0 + \frac{1}{2}S + \frac{n_0 R(\bar{o} - \mu_0)(\bar{o} - \mu_0)^t}{2(n_0 + R)}.$$

The posterior marginal distribution for the expected utilities can be shown to be a multivariate Student distribution (Bernardo and Smith, 1994),

$$p(\vec{v} \mid \mathcal{D}') \sim \mathbf{St}_K\left(\frac{n_0\mu_0 + R\bar{o}}{n_0 + R}, \lambda_p, 2\alpha_0 + R - K + 1\right), \tag{8}$$

where

$$\lambda_p = (n_0 + R)\left(\alpha_0 + \frac{R - K + 1}{2}\right) W_1(\bar{o}, S).$$

Several comments can be made about Equation (8). First, the correlation of the utilities for different systems due to a given common-random number scheme is inferred from the simulation output, and no assumptions regarding the precise values of the correlation were required. Second, the expected value $\frac{n_0\mu_0 + R\bar{o}}{n_0 + R}$ of the unknown expected utility approaches the sample mean $\bar{o}$ asymptotically. Third, the multivariate Student distribution is shown to play a role for the Bayesian framework of selecting the best system when a Gaussian approximation for the output and a normal-Wishart distribution for the prior are taken. Nelson and Matejcik (1995) anticipated this for the special case of unknown means and a known covariance.

5 EXAMPLE

A well-known example (Law and Kelton, 1991) of using common random numbers to compare two systems is the Zippytel (one expensive automated teller) versus Klunkytel (two automated tellers that are half as expensive and half as fast) problem. Zippytel (system 1) is an M/M/1 queue, and Klunkytel (system 2) is an M/M/2 queue. Mean inter-arrival time is assumed to be 1, and both have the same utilization, $\rho = 0.9$. Utility $\vec{u}$ in this case is taken to be the negative of average delay in queue of the first 100 customers. Analytical results (Kelton and Law, 1985) indicate that the theoretical expected utilities are $\upsilon_1 = -4.13$ and $\upsilon_2 = -3.70$.

We simulate to evaluate the performance of the results in Section 4.2 in determining that the Klunkytel has better performance, $p(\upsilon_2 \geq \upsilon_1 \mid \mathcal{D})$, based on simulation output $\mathcal{D}$. $R = 100$ replications of each system with independent and synchronized CRN were performed. Sample means and standard errors (Table 1) are consistent with those published in (Law and Kelton, 1991).

Posterior distributions were taken from Equation (4) for independent replications, and from Equation (8) with CRN. Prior distributions assumed $\mu_0 = (4, 4)$, $n_0 = 1$, $\alpha = 2.5, \beta = \text{diag}(2, 2)$. This results in a prior mean for $\vec{v}$ near half the steady-state mean for Speedy, an expected mean for the covariance matrix of $\text{diag}(2, 2)$, and a variance in the mean which is $\text{diag}(50, 50)$. The assumptions are rather conservative, and the results of the analysis are somewhat insensitive to the prior distribution selected. The

Table 1: Analysis of Average Waiting Time, $R = 100$

System	Theory	$\bar{o}$	SE
Speedytel	4.13	3.80	.304
Klunkytel(Ind)	3.7	3.49	.318
Klunkytel(CRN)	3.7	3.40	.299

Table 2: Posterior Probability Klunkytel is Better and P-value for H_0 : Same Performance

Variates	$p(\upsilon_2 \geq \upsilon_1 \mid \mathcal{D})$	P-value for H_0
Indep.	.71	.59
CRN	$1 - 1 \times 10^{-30}$	1.5×10^{-34}

Bayesian results are compared with frequentist P-values for the hypothesis H_0 : Klunkytel and Speedytel have the same performance.

The correlation of the output for the two systems is extremely high, with S in Equation (7) calculated as

$$S = \begin{bmatrix} 917.0 & 898.3 \\ 898.3 & 884.2 \end{bmatrix}$$

The posterior distribution $p(\upsilon \mid \mathcal{D})$ has Student distribution with mean and variance given by:

$$\begin{aligned} E[\upsilon \mid \mathcal{D}] &= (-3.800, -3.404)^t \\ \mathrm{Var}[\upsilon \mid \mathcal{D}] &= \begin{bmatrix} 0.0891 & 0.0872 \\ 0.0872 & 0.0860 \end{bmatrix} \end{aligned}$$

This is relatively similar to frequentist intuition: a standard error of 0.3 (see Table 1) corresponds to a variance of 0.09, and both systems have marginal variances of about 0.09.

A summary of some of the results is presented in Table 2. Not surprisingly, neither the Bayesian nor frequentist approach presented strong evidence that Klunkytel was better than Speedytel, based on the replications with independent variates. When common random numbers were used, the Bayes approach made it quite clear that Klunkytel outperforms Speedytel. The frequentist P-value, together with the sample means, indicate that Klunkytel is the clear winner as well.

Note that the P-value does *not at all* correspond to the probability $p(\upsilon_2 = \upsilon_1 \mid \mathcal{D})$. The latter probability is 0 in the Bayesian framework proposed here. On the other hand, because of the properties of the Student distribution, one will observe that the $1 -$ P-value/2 and $p(\upsilon_2 \geq \upsilon_1 \mid \mathcal{D})$ will be close when: (1) the prior distribution is uninformative in some sense, or the number of replications is large, (2) the Gaussian approximations for the output are assumed, and (3) the estimated mean for system 2 is better than for system 1 (the factor of 2 is because the t-test is two-sided. The result holds because the difference of two t-distributions with the same degrees of freedom (the posteriors from the paired independent replications) is a t-distribution (analogous to frequentist estimation of the mean of differences). Under these conditions, the P-value provides a measure of evidence which is roughly parallel to a Bayesian posterior distribution.

If these three conditions are not satisfied, or if the point hypothesis testing framework of Berger and Sellke (1987) were used, the P-value would *not* be an appropriate measure of evidence for the reasons detailed in Berger and Sellke (1987) and briefly summarized in Appendix A.

6. COMMENTS ON THE BEM

Alternate selection processes exist in the literature. One is the BEM multinomial selection procedure described by Bechhofer, Elmaghraby, and Morse (1959). Miller, Nelson and Reilly (1996) describe an extension of BEM, the All Vector Comparison (AVC), along with analysis, simulations, and conjectures.

The BEM can be used to estimate the probability that a given system is best in cases when the posterior for the expected utility Equation (6) or Equation (8) is too difficult to handle analytically or by quadrature.

Sample random variables from the posterior distribution for the expected utility, from Equation (8) for example (and not from the original simulations of the K systems). Sampling from the original simulated systems would give incorrect results, as observed utilities, not expected utilities would be sampled. (For instance, if system 1 always outputs 10, and system B outputs 9 with probability .9 and 20 with probability .1, then system A would be declared better with probability .9 if output was sampled, but B has a better expected value.)

The BEM approach has a natural Bayesian representation. Let $\vec{p} = (p_1, \ldots p_K)$ be the multinomial parameters for the BEM selection problem. Any proper prior is possible for $\vec{p}$, but if the conjugate Dirichlet prior

$$\pi(\vec{p}) \sim \mathbf{Di}_{K-1}(\alpha_1, \ldots, \alpha_K)$$

is chosen, then the posterior distribution is tractable analytically. Specifically, if $(n_1, \ldots, n_K)$ is a vector representing that system k had the highest sampled

expected value n_k times, the posterior distribution is

$$p(\vec{p} \mid \mathcal{D}) \sim \mathbf{Di}_{K-1}(\alpha_1 + n_1, \ldots, \alpha_K + n_K)$$

The maximum a posteriori probability $\tilde{p}$ (Bayesian analog of MLE) is then given by

$$\tilde{p}_k = \frac{\alpha_k + n_k - 1}{\sum_{i=1}^{K} \alpha_i + n_i - 1}.$$

for $k = 1, \ldots, K-1$, and $\tilde{p}_K = 1 - \sum_{i=1}^{K-1} \tilde{p}_i$. These estimates can be used as a measure of belief that a given system is best.

7- FUNCTIONALS OF DISTRIBUTIONS

It is sometimes of interest to evaluate functionals of the output distributions of the various systems other than expected utility. For instance, a comparison of quantiles of the systems can provide insight for a decision-maker. This section discusses the selection of the 'best' system when best is defined by optimizing a functional of the output distributions other than expected utility.

A Bayesian analysis for output distributions is compatible with either expected-utility or functional optimization decision making. The analysis is quite similar to the analysis for expected utility decision making. The primary difference is that the functional $\mathcal{L}_k$ of interest is no longer $E[U_k]$, but takes a more general form that maps a probability distribution $p(\cdot)$ into a figure of merit ℓ_k.

$$\ell_k = \mathcal{L}_k(p_{O_k|\theta_k}(\cdot)) \text{-} \tag{9}$$

Here, the notation $p_{O_k|\theta_k}(\cdot)$ is used to emphasize that $\mathcal{L}_k$ is a function of the entire probability distribution of the output, rather than the distributions density at a specific value of the output. As with the expected utility case, $p_{O_k|\theta_k}(o_k)$ depends on an unknown parameter θ_k, and therefore ℓ_k is a random variable whose distribution is determined by the posterior distribution of θ_k from Equation (1) or Equation (5).

The selection of the 'best' system then follows exactly as above for selecting the system with the maximal utility, except that ℓ_k takes the place of υ_k.

Selecting the best system by optimizing a functional other than expected utility is non-optimal from a decision-theoretic viewpoint, unless one or more of the axioms of decision theory has been rejected. See for example deGroot (1970), or Bernardo and Smith (1994) for axioms for decision making.

8 CONCLUSIONS

A unified framework for the problem of selecting the best simulated system was presented from a Bayesian perspective. The framework assumes that simulation output is described by a parametric statistical distribution. The parameters are inferred from the output of the simulation. Uncertainty in the parameters induces a distribution on the conditional mean value of the output, conditional on the parameter of the output distribution. The classical approach of assuming Gaussian output and Student distributions for the mean value of the output was shown to have similar asymptotic properties to a special case of the Bayesian framework: namely, Gaussian output with a normal-Wishart prior.

Benefits of this Bayesian framework are selection from two or more systems, with either independent or common random numbers, with unknown (or known) means and/or covariances for the output, and Gaussian or non-Gaussian simulation output. For the case of unknown mean and variance with common random numbers, the framework provides a probability of correct selection that does not suffer from problems associated with the Bonferroni inequality.

We indicated some criteria for which the Bayesian approach and other approaches are in general agreement, or disagreement, in Section 5.

The framework can be adapted for selecting the best system, comparing systems to a standard, subset selection, and indifference zone techniques merely by changing a domain of integration of the posterior probability given in Equation (3), Equation (6), or Equation (8). Should these integrals resist evaluation by quadrature, the BEM multivariate selection criteria and Monte Carlo techniques can be applied to approximate the probability that a given system is the best. We also comment on expected-value decision-making versus optimization criteria based on other functionals of the distribution of the output.

The current paper did not discuss the determination of the number of replications required to achieve a specific probability P^* of correct selection. This constitutes an important area for further research. Another interesting application for further research is simulation output analysis when some replications are correlated, but not all replications can be paired (e.g., the missing data problem).

Although the focus has been on decision-theoretic perspective of Bayesian statistics and expected utility, other uses are possible. For instance, it is possible to use Bayesian statistics with expected values of other simulation outputs, or with other functionals of the output distribution. Relatively little research has

been focused in this domain of research, in contrast with the relatively large amount of frequentist-based work on estimation of means and other functionals, notably quantiles.

ACKNOWLEDGMENT

Thanks are due to Shane Henderson, who made valuable comments during the writing of this paper.

APPENDIX A: EVIDENCE

The P-value has been a standard measure of evidence in hypothesis testing for determining the best system. On the other hand, there has been considerable discussion in the statistical literature regarding the potential pragmatic pitfalls of using a P-value as a measure of belief that a decision based on confidence intervals is correct.

Berger and Sellke (1987) describe a series of experiments each of which could be analyzed using a hypothesis test of $H_0 : \theta_i = 0$ versus $H_1 : \theta_i \neq 0$, where each hypothesis seemed to be equally likely based on past experience. They pose two questions: Among experiments for which the P-value is around 0.05, what portion correspond to a true H_0? In other words, what is the posterior probability $p_{H|x}(H_0)$ that H_0 is true, given that data x has been observed? What is the same result for a P-value of 0.01? Such low P-values are typically used to indicate that H_0 is almost certain to be wrong. Suppose a Bayesian analysis were applied, and that each hypothesis has a prior probability of 1/2. They provide surprising lower bounds to those questions: 0.24 for the first question, and 0.07 for the second, regardless of the prior distribution $p_{\Theta_i|H_1 \text{ true}}(\theta_i)$ chosen. The relation between posterior probability and P-value of a hypothesis can take on different forms for different tests (Casella and Berger, 1987).

In the example of Section 5, $1 - \text{P-value}/2$ and $p(\upsilon_2 \geq \upsilon_1 \mid \mathcal{D})$ were close. More generally, for $K = 2$ systems, the P-value and the probability of correct selection (using the framework above) are off by a linear transformation. This is because a two-sided test is being used for a one-sided question.

The use of P-values as a measure of evidence, therefore, may be misleading. In general, the P-value for a classical test that examines test statistics $T(X)$ for observed data $X = x$ is $p(T(X) \geq T(x) \mid H_0)$, where H_0 specifies the true parameter. Simulations to determine coverage probabilities for approximations to confidence intervals similarly assume the null hypothesis. On the other hand, the posterior probability of a hypothesis given data x is $p(H_0 \mid x)$, a very different quantity - the probability that the hypothesis is true, given the data. The P-value refers to the probability of getting extreme test statistics for a given, known value of a parameter. In other words, the P-value conditions on events that never happen. $p(H_0 \mid x)$ refers to the probability of a proposition given certain data. Although this is generally clear to statisticians, it is a subtle point that is often overlooked in practice, where P-values are often misinterpreted as probabilities that a hypothesis is true. Further, introductory simulation and statistical text books tend not to cover this issue.

APPENDIX B: DISTRIBUTIONS

Densities for the multivariate distributions used in this paper are presented here for convenience. Additional information is available in Bernardo and Smith (1994) and Robert (1994).

Dirichlet distribution: A $(K-1)\times 1$ vector X is said to have a $\mathbf{Di}_{K-1}(\alpha_1, \ldots, \alpha_K)$ distribution when $X_i \geq 0$, $\sum_{i=1}^{K-1} X_i \leq 1$, and the density function $f(x \mid \alpha_1, \ldots, \alpha_K)$ is

$$f(x \mid \alpha_1, \ldots, \alpha_K) = c\left(1 - \sum_{i=1}^{K-1} x_i\right)^{\alpha_K - 1} \prod_{i=1}^{K-1} x_i^{\alpha_i - 1}$$

where $c = \Gamma(\sum_{i=1}^{K} \alpha_i) / \prod_{i=1}^{K} \Gamma(\alpha_i)$, and $\alpha_i > 0$. $E[X] = (\alpha_1, \ldots, \alpha_K) / \sum_{i=1}^{K} \alpha_i$.

Gamma distribution: A real-valued random variable X is said to have a $\mathbf{Ga}(\alpha, \beta)$ distribution when the density function $f(x \mid \alpha, \beta)$ is

$$f(x \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}.$$

$E[X] = \alpha\beta^{-1}$ and $\text{Var}[X] = \alpha\beta^{-2}$.

Multivariate Gaussian distribution: A K-dimensional vector random variable X is said to have a $\mathbf{N}_K(\mu, \Sigma)$ distribution when the density function $f(x \mid \mu, \Sigma)$ is

$$f(x \mid \mu, \Sigma) = \frac{e^{-\frac{(x-\mu)^t \Sigma^{-1} (x-\mu)}{2}}}{(2\pi)^{K/2} |\Sigma|^{1/2}}$$

where μ is the K-dimensional mean, $|\Sigma|$ is the determinant of Σ, and and Σ is the covariance matrix. $E[X] = \mu$ and $\text{Var}[X] = \Sigma$.

Multivariate Student distribution: A K-dimensional vector random variable X is said to have a $\mathbf{St}_K(\mu, \lambda, \alpha)$ distribution when the density function

$f(x \mid \mu, \lambda, \alpha)$ is

$$f(x \mid \mu, \lambda, \alpha) = c \left[1 + \frac{(x-\mu)^t \lambda (x-\mu)}{\alpha}\right]^{-\frac{\alpha+K}{2}}$$

where μ is the K-dimensional mean, λ is a symmetric, positive-definite $K \times K$ matrix, $\alpha > 0$ is the degrees of freedom, and $c = |\lambda|^{1/2}\Gamma((\alpha + K)/2)/\left(\Gamma(\alpha/2)(\alpha\pi)^{K/2}\right)$ is a normalizing constant. $E[X] = \mu$ and $\mathrm{Var}[X] = \lambda^{-1}\alpha/(\alpha - 2)$.

Wishart distribution: The Wishart distribution is an appropriate generalization for the χ^2 distribution to multiple dimensions. A symmetric, positive-definite $K \times K$ matrix X is said to have a $\mathbf{W}_K(\alpha, \beta)$ distribution when the density function $f(x \mid \alpha, \beta)$ is

$$f(x \mid \alpha, \beta) = c|x|^{\frac{\alpha-(K+1)}{2}} e^{-\mathrm{tr}(\beta x)}$$

where $\alpha > (K-1)/2$ is real-valued, β is a symmetric, non-singular $K \times K$ matrix, $|A|$ is the determinant of A, and $\mathrm{tr}(A)$ is the trace of A, and $c = |\beta|^{\alpha}/\left(\pi^{K(K-1)/4}\prod_{k=1}^{K}\Gamma((2\alpha+1-k)/2)\right)$ is a normalizing constant. When $K = 1$ the Gamma distribution is recovered.

REFERENCES

Andradóttir, S., and V. M. Bier. 1997. Applying Bayesian ideas in simulation. Department of Industrial Engineering, University of Wisconsin-Madison, Technical Report 97-1.

Andrews, R. W., and T. J. Schriber. 1983. A Bayesian batch means methodology for analysis of simulation output. In *Proceedings of the Winter Simulation Conference*, ed. S. Roberts, J. Banks, and B. Schmeiser, 37–38. Institute of Electrical and Electronics Engineers, Inc.

Bechhofer, R. E., S. Elmaghraby, and N. Morse. 1959. A single-sample multiple-decision procedure for selecting the multinomial event which has the highest probability. *Annals of Mathematical Statistics* 30:102–119.

Berger, J. O., and T. Sellke. 1987. Testing a point null hypothesis: The irreconcilability of P-values and evidence. *Journal of the American Statistical Association* 82(397):112–122.

Bernardo, J. M., and A. F. M. Smith. 1994. *Bayesian theory.* Chichester, UK: Wiley.

Casella, G., and R. L. Berger. 1987. Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *Journal of the American Statistical Association* 82(397):106–111.

Chick, S. E. 1997. Bayesian analysis for simulation input and output. In *Proceedings of the Winter Simulation Conference*, ed. S. Andradóttir, K. J. Healy, D. H. Withers, and B. L. Nelson. Institute of Electrical and Electronics Engineers, Inc. Piscataway, New Jersey.

de Groot, M. H. 1970. *Optimal statistical decisions.* New York: McGraw-Hill, Inc.

Goldsman, D., and B. L. Nelson. 1994. Ranking, selection, and multiple comparisons in computer simulation. In *Proceedings of the Winter Simulation Conference*, ed. J. D.Tew, S. Manivannan, D. A. Sadowski, and A. F. Seila, 192–199. Institute of Electrical and Electronics Engineers, Inc.

Kelton, D., and A. Law. 1985. The transient behavior of the M/M/s queue, with implications for steady-state simulation. *Operations Research* 33:378–396.

Law, A. M., and W. D. Kelton. 1991. *Simulation modeling & analysis.* 2nd ed. New York: McGraw-Hill, Inc.

Miller, J. O., B. L. Nelson, and C. H. Reilly. 1996. Getting more from the data in a multinomial selection problem. In *Proceedings of the Winter Simulation Conference*, ed. J. M. Charnes, D. J. Morrice, D. T. Brunner, D. T., and J. J. Swain, 287–294. Institute of Electrical and Electronics Engineers, Inc.

Nelson, B. L., and F. J. Matejcik. 1995. Using common random numbers for indifference-zone selection and multiple comparisons in simulation. *Management Science* 41:1935–1945.

Robert, C. P. 1994. *The Bayesian choice.* New York: Springer-Verlag.

AUTHOR BIOGRAPHY

STEPHEN E. CHICK is an assistant professor of Industrial and Operations Engineering at the University of Michigan, Ann Arbor. In addition to simulation, his research interests include engineering probability, Bayesian statistics in system design, reliability, decision analysis, and computational methods in statistics. His work experience includes several years of using simulation analysis for material handling system design in the automotive industry.