Proceedings of the 2015 Winter Simulation Conference
L. Yilmaz, W. K V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, eds.

INPUT UNCERTAINTY AND INDIFFERENCE-ZONE RANKING & SELECTION

Eunhye Song
Barry L. Nelson

Department of Industrial Engineering & Management Sciences
Northwestern University
Evanston, IL 60208, USA

L. Jeff Hong

Department of Economics and Finance
Department of Management Sciences
City University of Hong Kong
Kowloon, Hong Kong

ABSTRACT

The indifference-zone (IZ) formulation of ranking and selection (R&S) is the foundation of many procedures that have been useful for choosing the best among a finite number of simulated alternatives. Of course, simulation models are imperfect representations of reality, which means that a simulation-based decision, such as choosing the best alternative, is subject to model risk. In this paper we explore the impact of model risk due to input uncertainty on IZ R&S. "Input uncertainty" is the result of having estimated ("fit") the simulation input models to observed real-world data. We find that input uncertainty may force the user to revise, or even abandon, their objectives when employing a R&S procedure, or it may have very little effect on selecting the best system even when the marginal input uncertainty is substantial.

1 INTRODUCTION

One of the most elemental problems in stochastic simulation is searching for which of $k < \infty$ alternative system designs has the largest or smallest mean performance measure. Examples include selecting the inventory policy with the smallest long-run average cost; selecting the portfolio with the largest expected return; and selecting the manufacturing layout with the largest throughput. Ranking and selection (R&S) addresses this type of problem.

Assuming an environment in which we can make replications, the standard output model for R&S is

$$Y_{ij} = \mu_i + \varepsilon_{ij} \tag{1}$$

where Y_{ij} is the output from the jth replication of systems i, μ_i is its expected value, and the ε_{ij} are, for fixed i, mean zero, independent and identically distributed (i.i.d.) random variables whose distribution (and in particular variance) may depend on i. In this paper we will assume that the ε_{ij} are known to be normally distributed, so that only their variances and covariances matter. We also assume that a larger mean is better and (unknown to us) $\mu_k \geq \mu_{k-1} \geq \cdots \geq \mu_1$. The goal is to efficiently use the stochastic simulation outputs to correctly select system k, or a system whose mean is very close to μ_k, as best.

The joint distribution of the ε_{ij} and the true differences $\mu_k - \mu_i$, $i \neq k$ are critical to correctly selecting the best. The indifference-zone (IZ) formulation of R&S sidesteps the latter issue by delivering a compromise

guarantee:

$$\Pr\{\text{select } k \,|\, \mu_k - \mu_{k-1} \geq \delta\} \geq 1-\alpha$$

where $\delta > 0$ and $1/k < 1-\alpha < 1$ are user specified. The IZ parameter δ is interpreted as the smallest difference in mean performance that is practically significant, and $1-\alpha$ is the guaranteed probability of correct selection (PCS) when the best is at least δ better than the rest. The IZ condition frees the R&S procedure from depending on the true differences $\mu_k - \mu_i$. In fact if $\mu_k - \mu_i < \delta$ for some alternatives i, then selecting one of the close ones is also acceptable. Good general references for IZ R&S procedures are Bechhofer et al. (1995) and Kim and Nelson (2006).

Implicit in the use of a R&S procedure is the assumption that the simulation is a faithful representation of the real or conceptual system it was created to model. Of course, the simulation is never a perfect model of reality, so the decision of which alternative to select is subject to model risk. We will explore one aspect of that risk: input uncertainty.

Stochastic simulations require a collection of fully specified input-distribution models, which we denote generically by $\mathbf{F} = \{F_1, F_2, \ldots, F_L\}$. Interarrival-time and service-time distributions in queueing simulations, time-to-failure distributions in manufacturing simulations, and distributions of injury or illness type in emergency department simulations are typical examples of input models. The true but usually unknown real-world distributions are denoted by $\mathbf{F}^c$ (c is for "correct"). In this paper we consider the model risk due to having estimated or "fit" the input models; let $\widehat{\mathbf{F}}$ denote the fitted collection of input distributions. Unless we are very lucky, $\widehat{\mathbf{F}} \neq \mathbf{F}^c$. In this paper we compare k alternative system designs that may or may not share the same input distributions. Let $\mathbf{F}_i^c$ denote the true distributions for alternative i, and let $\widehat{\mathbf{F}}_i$ denote its fitted distribution.

To investigate this source of model risk, we extend Model (1) to include the dependence on the distributions used in the simulation:

$$Y_{ij}(\mathbf{F}_i) = \eta_i(\mathbf{F}_i) + \varepsilon_{ij}(\mathbf{F}_i) \tag{2}$$

where $\eta_i(\mathbf{F}_i)$ is the mean of the simulation output when distribution $\mathbf{F}_i$ is used to simulate alternative i. Thus, the mean responses depend on both the structural differences among the systems, represented by $\eta_i(\cdot)$, and the input distributions that are ultimately used in the simulations, $\mathbf{F}_i$. The true mean for alternative i is therefore $\mu_i = \eta_i(\mathbf{F}_i^c)$. Ideally we would like to correctly select $k^\star$ where

$$k^\star = \text{argmax}_{i=1,2,\ldots,k} \mathrm{E}[Y_i(\mathbf{F}_i^c)] = \text{argmax}_{i=1,2,\ldots,k} \eta_i(\mathbf{F}_i^c).$$

However, if we use $\widehat{\mathbf{F}}_i$ in place of $\mathbf{F}_i^c$, then without adjustment the R&S procedure only provides a statistical guarantee of selecting

$$\widehat{k} = \text{argmax}_{i=1,2,\ldots,k} \mathrm{E}[Y_i(\widehat{\mathbf{F}}_i)|\widehat{\mathbf{F}}_i] = \text{argmax}_{i=1,2,\ldots,k} \eta_i(\widehat{\mathbf{F}}_i).$$

Notice that the expectation in the objective function is now conditional on $\widehat{\mathbf{F}}_i$, so unconditionally $\widehat{k}$ is a random variable. Clearly, $\widehat{k}$ may not equal $k^\star$ with finite samples of real-world input data, in which case we may choose a wrong system as the best. This type of risk is known as *input-model risk* or *input uncertainty*. Being satisfied with $\widehat{k}$ as the best implies ignoring the input-model risk.

In this paper we investigate the impact of input uncertainty on the IZ R&S formulation. Our interest is not in proposing new procedures, but rather in revealing the inherent strengths and weaknesses of the IZ formulation in the face of input uncertainty. This will provide a foundation upon which procedures could perhaps be built. General overviews of the input uncertainty problem include Barton (2012) and Song et al. (2014). Corlu and Biller (2013) consider input uncertainty in subset selection, another approach to R&S that does not involve an IZ parameter. Fan et al. (2013) present a robust IZ R&S formulation that selects the best alternative with respect to the worst-case choices among a finite collection of possible input models.

The remainder of the paper is organized as follows. In the next section we introduce an output model that is useful for representing the effect of input uncertainty, and we exploit it to define some key concepts in Section 3. We specialize these concepts to two classes of IZ R&S procedures in Sections 4–5. Some conclusions are offered in Section 6.

2 MIXED-EFFECTS MODEL

As stated, Model (2) is not sufficient to examine the impact of input uncertainty on IZ R&S. To support our analysis, we expand the model further into a mixed-effects model of the simulation output (Kutner et al. 2004):

$$Y_{ij}(\widehat{\mathbf{F}}_i) = \mu_i + \gamma_i + \beta(\widehat{\mathbf{F}}_i) + (\mu\beta)_i(\widehat{\mathbf{F}}_i) + \varepsilon_{ij}(\widehat{\mathbf{F}}_i), \tag{3}$$

where

$$\begin{aligned}\mu_i &= \eta_i(\mathbf{F}_i^c),\\ \gamma_i &= \mathrm{E}[\eta_i(\widehat{\mathbf{F}}_i)] - \eta_i(\mathbf{F}_i^c).\end{aligned}$$

We also assume

$$\begin{aligned}&\beta(\widehat{\mathbf{F}}_i) \sim (0, \sigma_\beta^2),\\ &(\mu\beta)_i(\widehat{\mathbf{F}}_i) \sim (0, \sigma_{\mu\beta_i}^2),\\ &\varepsilon_{ij}(\widehat{\mathbf{F}}_i) \sim (0, \sigma_i^2(\widehat{\mathbf{F}}_i)).\end{aligned} \tag{4}$$

Under this model, $\mathrm{E}[\eta_i(\widehat{\mathbf{F}}_i)] = \mu_i + \gamma_i$, which is the fixed effect of the ith system on its output. The random effect $\beta(\widehat{\mathbf{F}}_i)$ captures the impact of $\widehat{\mathbf{F}}_i$ on the simulation output where the randomness is from the sampling distribution of the input model $\widehat{\mathbf{F}}_i$. Notice that its variance will be a decreasing function of the real-world sample size for input modeling. The interaction effect between the ith system and the estimated input model $\widehat{\mathbf{F}}_i$ is captured by $(\mu\beta)_i(\widehat{\mathbf{F}}_i)$. Simulation error $\varepsilon_{ij}(\widehat{\mathbf{F}}_i)$ depends on the system and the input model used to run the simulation.

Our goal in this paper is to gain insight. Therefore, to simplify the analysis, we make the following assumptions:

Assumption 1 All k systems share the same input models; i.e., $\mathbf{F}_1^c = \mathbf{F}_2^c = \cdots = \mathbf{F}_k^c$.

Thus, we are considering the case when the systems are different in structural ways (e.g., number of servers or inventory policy), but the underlying randomness is the same. Under Assumption 1 we can drop the subscript i from $\widehat{\mathbf{F}}_i$ and use $\widehat{\mathbf{F}}$ to represent the common input model. Assumption 1 implies that given $\widehat{\mathbf{F}}$, $\beta(\widehat{\mathbf{F}})$ is fixed for all k systems. Also, using common $\widehat{\mathbf{F}}$ implies the interaction effects $(\mu\beta)_i(\widehat{\mathbf{F}})$'s are correlated. The following assumption imposes a correlation structure among interaction effects.

Assumption 2 For any given $\widehat{\mathbf{F}}$, $\left\{(\mu\beta)_i(\widehat{\mathbf{F}})\right\}_{i=1}^{k-1}$ are jointly normally distributed with $\sum_{i=1}^{k}(\mu\beta)_i(\widehat{\mathbf{F}}) = 0$.

Assumption 2 implies that conditional on $\widehat{\mathbf{F}}$, $\{(\mu\beta)_i(\widehat{\mathbf{F}})\}_{i=1}^{k-1}$ act like fixed effects that account for any differences among the systems in how they are affected by $\widehat{\mathbf{F}}$. If we scale the variance of the interaction effects in (4) as $\sigma_{\mu\beta}^2(k-1)/k$ for algebraic simplicity, then under Assumption 2, $\mathrm{Cov}[(\mu\beta)_i(\widehat{\mathbf{F}}), (\mu\beta)_{i'}(\widehat{\mathbf{F}})] = -\sigma_{\mu\beta}^2/k$ for $i \neq i'$. The normality assumption on $\{(\mu\beta)_i(\widehat{\mathbf{F}})\}_{i=1}^{k-1}$ is helpful for inference on the distribution of $Y_{kj}(\widehat{\mathbf{F}}) - Y_{ij}(\widehat{\mathbf{F}})$.

We further assume the simulation error is normally distributed with variance depending on the system and the real-world sample size; for simplicity the sample size is not captured in our notation.

Assumption 3 $\varepsilon_{ij}(\widehat{\mathbf{F}}) \sim N(0, \sigma_i^2)$ for $i = 1, 2, \ldots, k,\ j = 1, 2, \ldots.$

Finally, we assume the finite sample bias to be the same for all i. Equivalently, this can be written as

Assumption 4 $\mathrm{E}[\eta_i(\widehat{\mathbf{F}})] - \mathrm{E}[\eta_\ell(\widehat{\mathbf{F}})] = \eta_i(F^c) - \eta_\ell(F^c)$ for $i \neq \ell$.

Assumption 4 is strong: it implies that the finite sample bias is the same for all systems. This will not be strictly true in general, but the assumption helps us analyze the problem structure.

Under these assumptions, we can rewrite Model (3) as

$$Y_{ij}(\widehat{\mathbf{F}}) = \mu_i + \gamma + \beta(\widehat{\mathbf{F}}) + (\mu\beta)_i(\widehat{\mathbf{F}}) + \varepsilon_{ij}(\widehat{\mathbf{F}}), \tag{5}$$

where

$$\begin{aligned}
&\mu_i = \eta_i(\mathbf{F}^c),\\
&\gamma_i = \gamma \text{ for all } i,\\
&\beta(\widehat{\mathbf{F}}) \sim (0, \sigma_\beta^2),\\
&(\mu\beta)_i(\widehat{\mathbf{F}}) \sim \mathrm{N}\left(0, \frac{k-1}{k}\sigma_{\mu\beta}^2\right),\\
&\mathrm{Cov}\left[(\mu\beta)_i(\widehat{\mathbf{F}}), (\mu\beta)_{i'}(\widehat{\mathbf{F}})\right] = -\frac{1}{k}\sigma_{\mu\beta}^2 \text{ for } i \neq i',\\
&\varepsilon_{ij}(\widehat{\mathbf{F}}) \sim \mathrm{N}(0, \sigma_i^2).
\end{aligned}$$

As a result of Assumption 1–4, $\{Y_{kj}(\widehat{\mathbf{F}}) - Y_{ij}(\widehat{\mathbf{F}})\}_{i=1}^{k-1}$ are jointly normally distributed with a well-defined covariance matrix under Model (5); we exploit this structure later.

Model (5) captures the key features of input uncertainty that are relevant to making comparisons among alternative systems. There is a bias, γ, due to using estimated input models; this bias arises even if the input models are unbiased for $\mathbf{F}^c$ because simulations are typically nonlinear transformations. The bias is averaged over the sampling distribution of $\widehat{\mathbf{F}}$, while the common random effect $\beta(\widehat{\mathbf{F}})$ is due to the specific input models that are actually realized. The extent to which the individual systems are affected differently by $\widehat{\mathbf{F}}$ is captured by $(\mu\beta)_i(\widehat{\mathbf{F}})$. Thus, the overall random effect of $\widehat{\mathbf{F}}$ on the output from system i is $\beta(\widehat{\mathbf{F}}) + (\mu\beta)_i(\widehat{\mathbf{F}})$, a common effect and a deviation specific to system i. The variance $\sigma_{\mu\beta}^2$ represents how heterogeneous this deviation might be; if $\sigma_{\mu\beta}^2 = 0$ then all systems are affected in the same way, and therefore not at all when looking at the differences between them. Finally, there is the usual stochastic output variance of each system represented by $\varepsilon_{ij}(\widehat{\mathbf{F}}) \sim \mathrm{N}(0, \sigma_i^2)$.

If we decide to use common random numbers (CRN) for replications across the systems for given $\widehat{\mathbf{F}}$, then the simulation errors $\varepsilon_{ij}(\widehat{\mathbf{F}})$ and $\varepsilon_{i'j}(\widehat{\mathbf{F}})$ are no longer independent for fixed j. However, the $\varepsilon_{ij}(\widehat{\mathbf{F}})$'s are the only part in Model (5) that is affected by CRN. Typically, we expect CRN to reduce $\mathrm{Var}[\varepsilon_{kj}(\widehat{\mathbf{F}}) - \varepsilon_{ij}(\widehat{\mathbf{F}})]$. In this paper, we assume CRN is not used unless otherwise noted.

3 KEY CONCEPTS

We focus on R&S procedures that guarantee a lower bound on the probability of selecting the best when the configuration of the means is in the so-called *preference zone* suggested by Bechhofer (1954): $\mathrm{PZ}_\delta = \{(\mu_1, \mu_2, \ldots, \mu_k) : \mu_k - \mu_i \geq \delta, \forall i \neq k\}$ for some $\delta > 0$. Again, δ is the *indifference-zone parameter*, which implies that the user is indifferent among systems whose means μ_i are within δ of μ_k.

Let $\mathrm{PCS}_{\delta,\mathbf{N}}$ denote the minimum PCS under PZ_δ when $\mathbf{N} = \{N_1, N_2, N_3, \ldots, N_k\}$ replications are obtained from the k systems. If there is no input-model risk, then for the normal-theory case there are many R&S procedures that will deliver $\mathbf{N}$ to guarantee $\mathrm{PCS}_{\delta,\mathbf{N}} \geq 1 - \alpha$. In the presence of input-model risk, however, the $\mathrm{PCS}_{\delta,\mathbf{N}}$ may not have a lower bound of $1 - \alpha$ no matter how large the N_i's are.

Suppose we obtain N_i observations from system i using input model $\widehat{\mathbf{F}}$ and compare the sample means. Under Model (5)

$$\bar{Y}_k(\widehat{\mathbf{F}}) - \bar{Y}_i(\widehat{\mathbf{F}}) = \mu_k - \mu_i + (\mu\beta)_k(\widehat{\mathbf{F}}) - (\mu\beta)_i(\widehat{\mathbf{F}}) + \bar{\varepsilon}_k(\widehat{\mathbf{F}}) - \bar{\varepsilon}_i(\widehat{\mathbf{F}}) \text{ for } i \neq k$$

where a "bar" indicates averaging. Notice that the common effect of the input model $\beta(\widehat{\mathbf{F}})$ disappears as all systems share the same $\widehat{\mathbf{F}}$. Compared to the no-input-model-risk case, we have an extra term $\tau_i(\widehat{\mathbf{F}}) = (\mu\beta)_k(\widehat{\mathbf{F}}) - (\mu\beta)_i(\widehat{\mathbf{F}})$ which represents the difference in how systems i and k are affected by the choice of input model. *If all systems are affected in precisely the same way then these terms are all zero and input uncertainty, no matter how large marginally, would have no impact when looking at differences.* More generally, $\tau_i(\widehat{\mathbf{F}})$ has expected value 0 with respect to the sampling distribution of $\widehat{\mathbf{F}}$, so it cannot be averaged out by obtaining more simulation replications. Given $\widehat{\mathbf{F}}$ we are stuck with the $\tau_i(\widehat{\mathbf{F}})$'s that it implies, and these terms cannot be estimated directly because they are confounded with the mean simulation response $\mu_k - \mu_i$. Hence, defining $\text{PCS}_{\delta,\mathbf{N}}$ as in a standard R&S procedure no longer makes sense.

Although we cannot directly estimate the $(\mu\beta)_i(\widehat{\mathbf{F}})$ terms, we might be able to characterize their distribution. For instance, under Model (5) they have mean zero and are jointly normally distributed with variance and covariance determined by $\sigma^2_{\mu\beta}$. If we can somehow approximate this distribution then we might, as a compromise, try to control the *average PCS*.

Definition 1 The average PCS is

$$\overline{\text{PCS}} \equiv \int_{\mathbb{R}^{k-1}} \Pr\{\bar{Y}_k(\widehat{\mathbf{F}}) > \bar{Y}_i(\widehat{\mathbf{F}}) \text{ for all } i \neq k | \tau(\widehat{\mathbf{F}}) = \mathbf{t}\} d\mathbf{G}(\mathbf{t}),$$

where $\tau(\widehat{\mathbf{F}}) = \{\tau_1(\widehat{\mathbf{F}}), \tau_2(\widehat{\mathbf{F}}), \ldots, \tau_{k-1}(\widehat{\mathbf{F}})\}^T$ whose joint cumulative distribution (cdf) is $\mathbf{G}$, and $\mathbf{t} = \{t_1, t_2, \ldots, t_{k-1}\}^T$.

The average PCS is in the same spirit as PCS itself: PCS is the fraction of correct selections averaged over many applications of the procedure. Average PCS extends this average to include the possible input models that could have been obtained.

Using $\overline{\text{PCS}}$ is different from the robust selection-of-the-best method suggested by Fan et al. (2013), where they have an "ambiguity set" $\mathscr{A}$ that contains q candidate distributions $\widehat{\mathbf{F}}$ and correct selection is defined as selecting system i that minimizes $\max_{\widehat{\mathbf{F}} \in \mathscr{A}} \eta_i(\widehat{\mathbf{F}})$. Unlike their paper, we do not concentrate on a particular candidate $\widehat{\mathbf{F}}$, but instead define $\overline{\text{PCS}}$ by averaging PCS over the sampling distribution of $\widehat{\mathbf{F}}$.

Based on Definition 1 we could lower-bound the average PCS over PZ_δ; call this $\overline{\text{PCS}}_{\delta,\mathbf{N}}$. For given $\mathbf{N}$, $\overline{\text{PCS}}_{\delta,\mathbf{N}} \leq \text{PCS}_{\delta,\mathbf{N}}$ since there is input-model risk represented by the $\tau_i(\widehat{\mathbf{F}})$'s. Although we cannot actually guarantee $\text{PCS}_{\delta,\mathbf{N}} \geq 1-\alpha$ when there is input-model risk, we might settle for a procedure that promises $\overline{\text{PCS}}_{\delta,\mathbf{N}} \geq 1-\alpha$. We suggest how this could be achieved in the next two sections.

Instead of compromising $\text{PCS}_{\delta,\mathbf{N}}$, we can think of compromising δ. Intuitively, as δ increases, $\overline{\text{PCS}}_{\delta,\mathbf{N}}$ also increases as μ_k is assumed to be farther away from μ_i for $i \neq k$ under PZ_δ and we are less affected by the input-model risk. Under input-model risk, we define the *effective indifference-zone parameter* δ_{eff} as follows:

Definition 2 Suppose $\mathbf{N}^\star = \{N_1^\star, N_2^\star, \ldots, N_k^\star\}$ are the numbers of replications specified by a procedure to achieve $\text{PCS}_{\delta,\mathbf{N}^\star} \geq 1-\alpha$ in the no-risk case. Then the effective indifference-zone parameter is

$$\delta_{\text{eff}} \equiv \inf\{\omega > 0 \colon \overline{\text{PCS}}_{\omega,\mathbf{N}^\star} \geq 1-\alpha\}.$$

In words, δ_{eff} is the minimum difference we can detect by using the same number of replications as in the no-risk case, but to obtain an average PCS greater than $1-\alpha$ when there is input-model risk. Thus, $\delta_{\text{eff}} \geq \delta$.

Notice that this definition implicitly assumes that the distribution of each $Y_i(\widehat{\mathbf{F}})$ does not change when we change the mean configuration (PZ_δ) by increasing δ. This assumption is not true in general as the distribution of $Y_i(\widehat{\mathbf{F}})$ is affected by the change in its mean unless it is a distribution that is invariant under mean translation. Under Model (5) this assumption holds because the normal distribution is a location family.

We have $\overline{\text{PCS}}_{\delta,\mathbf{N}^\star} \leq \text{PCS}_{\delta,\mathbf{N}^\star}$ for given $\mathbf{N}^\star$. Can we improve $\overline{\text{PCS}}_{\delta,\mathbf{N}^\star}$ by increasing $\mathbf{N}$? To answer this question, we define the *minimum indifference-zone parameter under input-model risk*, $\delta_{\min}$, as follows.

Definition 3 The minimum indifference-zone parameter under input model-risk is

$$\delta_{\min} \equiv \inf\{\omega > 0 \colon \overline{\text{PCS}}_{\omega,\infty} \geq 1-\alpha\},$$

where $\overline{\text{PCS}}_{\omega,\infty} = \lim_{N_i\to\infty,\forall i} \overline{\text{PCS}}_{\omega,\mathbf{N}}$.

Therefore, $\delta_{\min}$ is the smallest difference we can ever detect in the presence of input-model risk.

Under the mixed-effects model δ_{eff} and $\delta_{\min}$ depend on the interaction variance $\sigma^2_{\mu\beta}$; with larger $\sigma^2_{\mu\beta}$, both δ_{eff} and $\delta_{\min}$ increase because $\widehat{\mathbf{F}}$ has a more dramatically different effect on each alternative system. If $\delta > \delta_{min}$, then we can still achieve $\overline{\text{PCS}}_{\delta,\mathbf{N}} \geq 1-\alpha$. However, if $\delta \leq \delta_{min}$, then we cannot achieve $\overline{\text{PCS}}_{\delta,\mathbf{N}} \geq 1-\alpha$ because $\lim_{N_i\to\infty,\forall i} \overline{\text{PCS}}_{\delta,\mathbf{N}} < 1-\alpha$. In this case, we can either decrease $1-\alpha$ and compromise the probability of correct selection or increase δ. *To users, the former means that they are willing to accept less certain selection of the best while retaining the IZ parameter δ they chose, whereas the latter means that they are willing to select a near-best system with increased δ but with the same certainty.*

So far we have considered procedures that guarantee the desired PCS over the least-favorable configuration (LFC) in the preference zone. One criticism for this approach is that the resulting procedure tends to be conservative if in fact the mean configuration is a more favorable element of PZ_δ.

For instance, if $\mu_k - \mu_{k-1} \gg \delta$, then a procedure based on the LFC of PZ_δ may be too conservative. *In the no-risk case, conservatism makes us collect more replications than necessary to achieve* $\text{PCS} \geq 1-\alpha$*, but we do achieve it. With input-model risk, conservatism can lead to the conclusion that we cannot guarantee to select the best system when in fact we can, at least for the average case.* For example, under PZ_δ if $\delta \leq \delta_{min}$, then we would conclude that we cannot achieve $\overline{\text{PCS}} \geq 1-\alpha$ without increasing δ. However, if we have a favorable mean configuration, i.e., $\mu_k - \mu_{k-1} \gg \delta_{\min}$, then we might still achieve $\overline{\text{PCS}} \geq 1-\alpha$ without compromising δ. *Therefore, the true mean configuration has a more significant impact when we have input model risk than in the no-risk case.*

Suppose the true mean configuration $\mu = (\mu_1, \mu_2, \ldots, \mu_k)^\top$ and $\mathbf{F}^c$ are known in the no-risk case. Then the PCS under this configuration based on $\mathbf{N}$ replications is

$$\text{PCS}_{\mu,\mathbf{N}} = \Pr\{\bar{Y}_k(\mathbf{F}^c) - \bar{Y}_i(\mathbf{F}^c) - (\mu_k - \mu_i) > \mu_i - \mu_k, \text{ for all } i \neq k\}.$$

Clearly, if $\mu_k - \mu_{k-1} \gg \delta$, then $\text{PCS}_{\mu,\mathbf{N}} > \text{PCS}_{\delta,\mathbf{N}}$. The means of other systems also matter; the greater each $\mu_k - \mu_i$ is, the bigger $\text{PCS}_{\mu,\mathbf{N}}$ is. Notice that $\text{PCS}_{\mu,\mathbf{N}}$ does not involve δ as we do not assume PZ_δ anymore. If we want to take advantage of the user's indifference among systems up to mean difference of δ, then we can use the *probability of good selection* (PGS), which is the probability of selecting a system whose mean is within δ from the best (Nelson and Banerjee 2001). The PGS under the true mean configuration μ is

$$\text{PGS}_{\mu,\delta,\mathbf{N}} = \Pr\{\bar{Y}_k(\mathbf{F}^c) - \bar{Y}_i(\mathbf{F}^c) - (\mu_k - \mu_i) > -\max[\delta, \mu_k - \mu_i], \text{ for all } i \neq k\}. \tag{6}$$

Notice the dependence of PGS on δ. Nelson and Matejcik (1995) show that if the joint distribution of $\{Y_1, Y_2, \ldots, Y_k\}$ are invariant under mean translation, then $\text{PGS}_{\mu,\delta,\mathbf{N}} \geq \text{PCS}_{\delta,\mathbf{N}}$. The normality assumption of Model (5) satisfies the condition. Similar to Definition 1 we can define the *average PGS* under input model risk as follows.

Definition 4 The average PGS under the true mean configuration μ is

$$\overline{\mathrm{PGS}}_{\mu,\delta,\mathbf{N}} = \int_{\mathbb{R}^{k-1}} \Pr\{\bar{Y}_k(\widehat{\mathbf{F}}) - \bar{Y}_i(\widehat{\mathbf{F}}) - (\mu_k - \mu_i) > -\max[\delta, \mu_k - \mu_i] \text{ for all } i \neq k | \tau(\widehat{\mathbf{F}}) = \mathbf{t}\} d\mathbf{G}(\mathbf{t}),$$

where $\tau(\widehat{\mathbf{F}}) = \{\tau_1(\widehat{\mathbf{F}}), \tau_2(\widehat{\mathbf{F}}), \ldots, \tau_{k-1}(\widehat{\mathbf{F}})\}^T$ and $\mathbf{G}$ is its cdf, and $\mathbf{t} = \{t_1, t_2, \ldots, t_{k-1}\}^T$.
It is clear that $\overline{\mathrm{PGS}}_{\mu,\delta,\mathbf{N}}$ is an increasing function of $\mu_k - \mu_i$, if $\mu_k - \mu_i > \delta$. *Therefore, even when minimum indifference-zone parameter $\delta_{\min} > \delta$, there may still exist* $\mathbf{N}$ *such that* $\overline{\mathrm{PGS}}_{\mu,\delta,\mathbf{N}} \geq 1-\alpha$ *when the actual difference* $\mu_k - \mu_{k-1} > \delta_{\min}$.

In the next two sections we discuss how single-stage and sequential, eliminating R&S procedures are affected by, and might be modified to account for, input-model risk using the concepts defined in this section.

4 SINGLE-STAGE PROCEDURES

Many of the issues that arise when doing R&S in the presence of input uncertainty can be illustrated using the known and equal-variances case, as decribed in Bechhofer (1954), and these issues arise in very much the same way in two-stage procedures. All of the results assume Model (5) describes the simulation outputs.

Suppose we further assume that

$$\varepsilon_i(\widehat{\mathbf{F}}) \sim \mathrm{N}(0, \sigma^2) \text{ (equal variances)}$$
$$\sigma^2, \sigma^2_{\mu\beta} \text{ are known.}$$

Under these assumptions, we can devise a single-stage R&S procedure. Because of the equal-variance assumption, we have equal sample sizes across all k systems. Therefore, we replace $\mathbf{N}$ with N in this section. Given the IZ parameter δ and the assumptions above we can show that

$$\overline{\mathrm{PCS}}_{\delta,N} \geq \int_{-\infty}^{\infty} \Phi\left(\frac{\delta}{\sqrt{\sigma^2/N + \sigma^2_{\mu\beta}}} + z_0\right)^{k-1} \phi(z_0) dz_0, \tag{7}$$

where ϕ and Φ are the probability density function (pdf) and cdf of the standard normal distribution, respectively. If $\sigma_{\mu\beta} = 0$, then (7) gives the lower bound on the PCS for the no-risk case. Notice that for fixed N the right-hand side of (7) decreases as $\sigma^2_{\mu\beta}$ increases. Thus, we need larger N to guarantee $\overline{\mathrm{PCS}}$. *This follows the discussion in Section 3 that the greater the difference in how each system's mean is affected by the choice of input models, the less certain we are about the selection of the best system.*

Using (7), δ_{eff} can be obtained by solving

$$\int_{-\infty}^{\infty} \Phi\left(\frac{\delta_{\mathrm{eff}}}{\sqrt{\sigma^2/N^\star + \sigma^2_{\mu\beta}}} + z_0\right)^{k-1} \phi(z_0) dz_0 = 1-\alpha,$$

where $N^\star$ satisfies $\int_{-\infty}^{\infty} \Phi\left(\delta/\sqrt{\sigma^2/N^\star} + z_0\right)^{k-1} \phi(z_0) dz_0 = 1-\alpha$. Clearly, $\delta_{\mathrm{eff}} \geq \delta$ where the equality holds when $\sigma^2_{\mu\beta} = 0$. We can also obtain $\delta_{\min}$ by solving

$$\int_{-\infty}^{\infty} \Phi\left(\frac{\delta_{\min}}{\sigma_{\mu\beta}} + z_0\right)^{k-1} \phi(z_0) dz_0 = 1-\alpha.$$

This implies that once we are given $\sigma_{\mu\beta}$ and δ, we can tell if we would be able to achieve a given $\overline{\mathrm{PCS}}$ even before we simulate the systems. Also, as k increases with the same $\sigma^2_{\mu\beta}$, $\overline{\mathrm{PCS}}_{\delta,N}$ is lower when

we compare larger numbers of systems. Therefore, the $\delta_{\min}$ required to attain PCS $1-\alpha$ increases. This provides intuition about how input model risk affects problems with increasing numbers of alternatives.

To examine the effect of CRN, suppose that the stochastic noise across systems, $\{\varepsilon_{ij}(\widehat{\mathbf{F}}), i = 1,2,\ldots,k\}$, is jointly normally distributed with mean 0 and variance-covariance matrix

$$\sigma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}$$

where $\rho \geq 0$ represents the effect of CRN. Then the average PCS, given N replications, is

$$\overline{\mathrm{PCS}}_{\delta,N} \geq \int_{-\infty}^{\infty} \Phi\left(\frac{\delta}{\sqrt{\sigma^2(1-\rho)/N + \sigma^2_{\mu\beta}}} + z_0 \right)^{k-1} \phi(z_0) dz_0.$$

This demonstrates that CRN reduces the sample size required to achieve a given $\overline{\mathrm{PCS}}$*, but does not alter the* $\delta_{\min}$ *that can be detected. In other words, CRN does not enhance the best resolution that can be attained when there is input uncertainty.*

So far in this section we tried to lower-bound the PCS over the PZ_δ. If we assume the true μ is known, then we can derive $\overline{\mathrm{PCS}}_{\mu,N}$ as

$$\overline{\mathrm{PCS}}_{\mu,N} = \int_{-\infty}^{\infty} \Pi_{i=1}^{k-1} \Phi\left(\frac{\mu_k - \mu_i}{\sqrt{\sigma^2/N + \sigma^2_{\mu\beta}}} + z_0 \right) \phi(z_0) dz_0.$$

Similarly,

$$\overline{\mathrm{PGS}}_{\mu,N} = \int_{-\infty}^{\infty} \Pi_{i=1}^{k-1} \Phi\left(\frac{\max[\delta, \mu_k - \mu_i]}{\sqrt{\sigma^2/N + \sigma^2_{\mu\beta}}} + z_0 \right) \phi(z_0) dz_0.$$

We see that if the $\mu_k - \mu_i$ are large enough then any desired $\overline{\mathrm{PCS}}$ or $\overline{\mathrm{PGS}}$ can be obtained. Of course, μ is a property of the problem and is unknown. However, we can attempt to provide a lower bound for $\overline{\mathrm{PCS}}_{\mu,N}$ and $\overline{\mathrm{PGS}}_{\mu,N}$ by using estimates of μ. Nelson and Banerjee (2001) suggest providing $1-\alpha$ lower confidence bounds on $\mathrm{PCS}_{\delta,N}$ and $\mathrm{PGS}_{\delta,N}$ using joint lower confidence bounds for $\mu_k - \mu_i$ for $i = 1,2,\ldots,k-1$. Their approach relies on unconstrained multiple comparisons with the best intervals with fixed-width. We can take a similar approach to provide lower confidence bounds on $\overline{\mathrm{PCS}}_{\mu,N}$ and $\overline{\mathrm{PGS}}_{\mu,N}$ that better reflect the true mean differences.

5 SEQUENTIAL ELIMINATING PROCEDURES

The single-stage procedure in Section 4 does not select a winner until all observations from all k systems have been collected; the same is true of most two-stage procedures. Because the IZ formulation guards against a particularly difficult configuration of the means, such procedures may not be able to deliver a PCS guarantee in the face of input uncertainty. As we illustrated, exploiting a more favorable configuration can mitigate the effect input uncertainty. Fully sequential, eliminating R&S procedures are more sensitive to the actual differences in the means, and may therefore be a good choice for selecting the best when there is input-model risk. We investigate this conjecture here.

In a typical fully sequential, eliminating R&S procedure, we perform pairwise comparisons of sample means of all systems and sequentially rule out inferior systems as we obtain more replications from the

systems in contention. In the well-known KN procedure (Kim and Nelson 2001), the systems in contention at the rth replication are

$$I^r = \left\{ i\colon\ i \in I^{r-1} \text{ and } \bar{Y}_i(r) \geq \bar{Y}_\ell(r) - W_{i\ell}(r), \forall \ell \in I^{r-1} \backslash \{i\} \right\},$$

where $\bar{Y}_i(r)$ is the sample mean of system i up to the rth replication and $W_{i\ell}(r)$ indicates how far $\bar{Y}_i(r)$ can drop below other systems' sample means without being eliminated. The value of $W_{i\ell}(r)$ that guarantees PCS $\geq 1-\alpha$ in the preference zone is a function of r, δ, α and $S^2_{i\ell} = \sum_{j=1}^{n_0} (Y_{ij} - Y_{\ell j} - [\bar{Y}_i(n_0) - \bar{Y}_\ell(n_0)])^2/(n_0-1)$, which is an estimator of $\sigma^2_{i\ell} = \text{Var}(Y_{ij} - Y_{\ell j})$. This method relies on the fact that $(Y_{kj} - Y_{ij})/\sigma_{ik} \sim N(\Delta, 1)$ with $\Delta > 0$, so the sum of differences behaves like Brownian motion with drift Δ. Elimination occurs when this sum of differences drifts outside a continuation region, and which system is eliminated depends on the direction that it departs. When two systems $i \neq k$ are compared in isolation, the probability of correctly selecting system k is minimized over PZ_δ when $\mu_k = \mu_i + \delta$ in which case $\Delta = \delta/\sigma_{ik}$.

As seen in the previous section, when we have input uncertainty we have an extra term $\tau_i(\widehat{\mathbf{F}}) = (\mu\beta)_k(\widehat{\mathbf{F}}) - (\mu\beta)_i(\widehat{\mathbf{F}})$ in the pairwise difference. Conditional on the value of $\tau_i(\widehat{\mathbf{F}})$, the drift of $(Y_{kj} - Y_{ij})/\sigma_{ik}$ changes: $(Y_{kj} - Y_{ij})/\sigma_{ik} \sim N((\mu_k - \mu_i + \tau_i(\widehat{\mathbf{F}}))/\sigma_{ik}, 1)$. Assuming $\mu_k \geq \mu_i + \delta$, three different cases can arise:

1. $\tau_i(\widehat{\mathbf{F}}) \geq 0$
   Then the drift of $(Y_{kj} - Y_{ij})/\sigma_{ik}$ is larger than in the no-risk case, so if we applied the no-risk procedure we would need fewer replications to make a correct selection.
2. $-(\mu_k - \mu_i) < \tau_i(\widehat{\mathbf{F}}) < 0$
   Then the drift is smaller than in the no-risk case, but still positive. Whether or not the correct-selection guarantee holds depends on whether or not the drift is smaller than δ.
3. $\tau_i(\widehat{\mathbf{F}}) < -(\mu_k - \mu_i)$
   Then the drift is negative instead of positive, so we are actually more likely to choose system i rather than system k as the best. As $\sigma^2_{\mu\beta}$ increases the probability of Case 3 increases and it becomes more difficult to reach the correct selection by comparing the sample means of the systems.

Can we reflect the impact of the input model risk in the design of the continuation region? If we assume that $\sigma^2_{\mu\beta}$ and σ^2 are known, then $r(\bar{Y}_k(r) - \bar{Y}_i(r))/\sigma^2_{ik} \big| \tau_i(\widehat{\mathbf{F}})$ for $r = 1, 2, \ldots$ has the same joint distribution as observing the continuous time Brownian motion $B_{\mu_k - \mu_i + \tau_i(\widehat{\mathbf{F}})}(r/\sigma^2_{ik})$ at discrete times $r = 1, 2, \ldots$. Suppose we have a triangular continuation region as in Hong (2006). Then the probability of system $i \neq k$ eliminating system k in isolation (incorrect selection due to the ith system or ICS_i) is

$$\begin{aligned}
\Pr\{\text{ICS}_i | \tau_i(\widehat{\mathbf{F}})\} &\leq \Pr\left\{ \inf_{t \in (0,+\infty)} B_{\mu_k - \mu_i + \tau_i(\widehat{\mathbf{F}})}(t) < -a + bt \right\} \\
&\leq \Pr\left\{ \inf_{t \in (0,+\infty)} B_{\delta + \tau_i(\widehat{\mathbf{F}}) - b}(t) < -a \right\} \\
&= \min\left\{ \exp(-2a(\delta + \tau_i(\widehat{\mathbf{F}}) - b)), 1 \right\}.
\end{aligned}$$

Then,

$$\text{E}[\Pr\{\text{ICS}_i | \tau_i(\widehat{\mathbf{F}})\}] \leq \text{E}\left[\min\left\{ \exp(-2a(\delta + \tau_i(\widehat{\mathbf{F}}) - b)), 1 \right\} \right]. \tag{8}$$

By setting (8) equal to $\alpha/(k-1)$ for each i, we can guarantee $\overline{\text{PCS}} \geq 1-\alpha$. As we have two parameters, there are many combinations of (a, b) that satisfy such a constraint. One suggestion is to fix $a = -\log(\alpha/(k-1))/\delta$, which is the value we typically choose in the no-risk case, and solve (8) $= \alpha/(k-1)$ for b. Notice that a

is always positive, if $0 < \alpha < 1$ and $k \geq 2$. If $\sigma^2_{\mu\beta}$ is large, then we are more likely to have Case 3 above and b can be negative to meet $\overline{\text{PCS}} \geq 1 - \alpha$. Negative b means that the resulting continuation region is no longer triangular; rather, the continuation region becomes wider as the number of replications increases and the method essentially fails. *Therefore, even though a fully sequential, eliminating procedure better exploits the true mean differences, the use of an indifference-zone parameter in the design of the procedure can cause it to fail even when the true means are in a favorable configuration.*

6 CONCLUSIONS

In this paper we used a mixed-effects model of simulation output to study the impact of input uncertainty on IZ R&S procedures. We find that a straightforward application of IZ selection can be misleading, providing an invalid correct-selection guarantee. However, it may be possible to adjust IZ procedures to guarantee an average PCS, averaged over the sampling distribution of the input models. Unfortunately, there will be cases in which the IZ assumption leads to the conclusion that even an average PCS guarantee cannot be delivered, when in fact it could be delivered if we had known that we were in a more favorable configuration. Methods that try to exploit the true differences become even more desirable when there is input-model risk.

ACKNOWLEDGMENTS

This paper is based upon work supported by the National Science Foundation under Grant No. CMMI-1068473.

REFERENCES

Barton, R. R. 2012. "Tutorial: Input Uncertainty in Output Analysis". In *Proceedings of the 2012 Winter Simulation Conference*, edited by C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A. M. Uhrmacher, 1–12. Piscataway, NJ, USA: IEEE.

Bechhofer, R. E. 1954. "A Single-Sample Multiple Decision Procedure for Ranking Means of Normal Populations with Known Variances". *The Annals of Mathematical Statistics* 25 (1): 16–39.

Bechhofer, R. E., D. M. Goldsman, and T. J. Santner. 1995. *Design and Analysis of Experiments for Statistical Selection, Screening, and Multiple Comparisons*. New York: Wiley.

Corlu, C. G., and B. Biller. 2013. "A Subset Selection Procedure under Input Parameter Uncertainty". In *Proceedings of the 2013 Winter Simulation Conference*, edited by R. Pasupathy, S.-H. Kim, A. Tolk, R. Hill, and M. E. Kuhl, 463–473. Piscataway, NJ, USA: IEEE.

Fan, W., L. J. Hong, and X. Zhang. 2013. "Robust Selection of the Best". In *Proceedings of the 2013 Winter Simulation Conference*, edited by R. Pasupathy, S.-H. Kim, A. Tolk, R. Hill, and M. E. Kuhl, 868–876. Piscataway, NJ, USA: IEEE.

Hong, L. J. 2006. "Fully Sequential Indifference-Zone Selection Procedures with Variance-Dependent Sampling". *Naval Research Logistics* 53 (5): 464–476.

Kim, S.-H., and B. L. Nelson. 2001. "A Fully Sequential Procedure for Indifference-zone Selection in Simulation". *ACM Transactions on Modeling and Computer Simulation* 11 (3): 251–273.

Kim, S.-H., and B. L. Nelson. 2006. "Selecting the Best System". In *Handbook in Operations Research and Management Science: Simulation*, edited by S. G. Henderson and B. L. Nelson. Amsterdam: Elsevier.

Kutner, M., C. Nachtsheim, J. Neter, and W. Li. 2004. *Applied Linear Statistical Models*. 5 ed. McGraw-Hill/Irwin.

Nelson, B. L., and S. Banerjee. 2001. "Selecting a Good System: Procedures and Inference". *IIE Transactions* 33 (3): 149–166.

Nelson, B. L., and F. J. Matejcik. 1995. "Using Common Random Numbers for Indifference-Zone Selection and Multiple Comparisons in Simulation". *Management Science* 41 (12): 1935–1945.

Song, E., B. L. Nelson, and C. D. Pegden. 2014. "Advanced Tutorial: Input Uncertainty Quantification". In *Proceedings of the 2014 Winter Simulation Conference*, edited by A. Tolk, S. Y. Diallo, I. O. Ryzhov, L. Yilmaz, S. Buckley, and J. A. Miller, 162–176. Piscataway, NJ, USA: IEEE.

AUTHOR BIOGRAPHIES

EUNHYE SONG is a Ph.D. candidate in the Department of Industrial Engineering and Management Sciences at Northwestern University. Her research interests are input uncertainty quantification, design of simulation experiments and simulation optimization. Her e-mail address is eunhyesong2016@u.northwestern.edu.

BARRY L. NELSON is the Walter P. Murphy Professor in the Department of Industrial Engineering and Management Sciences at Northwestern University. He is a Fellow of INFORMS and IIE. His research centers on the design and analysis of computer simulation experiments on models of stochastic systems, and he is the author of *Foundations and Methods of Stochastic Simulation: A First Course*, from Springer. His e-mail address is nelsonb@northwestern.edu.

L. JEFF HONG is a chair professor in the Department of Economics and Finance, and the Department of Management Sciences, at the City University of Hong Kong. His research interests include Monte Carlo methods, financial engineering, and stochastic optimization. He is currently an associate editor for *Operations Research*, *Naval Research Logistics* and *ACM Transactions on Modeling and Computer Simulation*. His email address is jeffhong@cityu.edu.hk.