

DO MEAN-BASED RANKING AND SELECTION PROCEDURES CONSIDER SYSTEMS' RISK?

Demet Batur
F. Fred Choobineh

Department of Industrial and Management Systems Engineering
University of Nebraska-Lincoln
Lincoln, NE 68588, U.S.A.

ABSTRACT

The legacy simulation approach in ranking and selection procedures compares systems based on a mean performance metric. The best system is most often deemed as the one with the largest (or smallest) mean performance metric. In this paper, we discuss the limitations of the mean-based selection approach. We explore other selection criterion and discuss new approaches based on stochastic dominance using an appropriate section of the distribution function of the performance metric. In this approach, the decision maker has the flexibility to determine a section of the distribution function based on the specific features of the selection problem representing either the downside risk, upside risk, or central tendency of the performance metric. We discuss two different ranking and selection procedures based on this new approach followed by a small experiment and present some open research problems.

1 INTRODUCTION

Discrete-event simulation is used in the analysis and comparison of complex stochastic systems, e.g., manufacturing systems and communications networks. Stochastic systems are compared based on one or more performance metrics of interest that are outputs of simulations models. Since simulation outputs are random, statistically-sound ranking and selection (R&S) techniques are needed to choose the best system. R&S procedures are a collection of statistical procedures for comparing a finite set of stochastic systems with the goal of finding the best among them.

The legacy approach in R&S procedures compares systems based on the mean dominance of a performance metric of interest. Recent developments in the mean-based R&S procedures when the samples are independent and identically (IID), normally distributed are the works of [Kim and Nelson \(2001\)](#), [Hong \(2006\)](#), [Pichitlamken et al. \(2006\)](#), [Chick and Inoue \(2001\)](#), and [Chen et al. \(2000\)](#). [Kim and Nelson \(2001\)](#) propose a fully-sequential procedure *KN* that obtains observations from each system one at a time until there is enough evidence that a system's mean is dominated by one of the others. The ultimate objective is to select a system with a guarantee on the probability of correct selection. [Hong \(2006\)](#) presents a computationally more efficient version of the *KN* procedure where new samples are allocated to systems according to their variances. [Pichitlamken et al. \(2006\)](#) propose a fully-sequential mean-based procedure for evaluating neighborhood solutions in a simulation optimization algorithm when partial or complete information on solutions previously visited is maintained. [Chen et al. \(2000\)](#) and [Chick and Inoue \(2001\)](#) use the Bayesian approach to the mean-based selection problem. Their procedures choose the best system (largest or smallest mean) such that the posterior probability of correct selection is maximized while satisfying a simulation budget constraint. An extensive comparison of the mean-based R&S procedures is given in [Branke et al. \(2007\)](#). For a comprehensive review of the R&S procedures in simulation, refer to [Kim and Nelson \(2006\)](#).

There are two types of discrete-event simulations: terminating and steady-state. In terminating simulations, the simulation starts at time zero under some well-specified initial conditions, and there is a natural ending event that often specifies the finite horizon that the system operates. Since the interest is on the behavior of the system over a finite-time horizon, the initial conditions would have a large impact on the performance of the system. However, for the steady-state simulations, often the interest is in the most representative behavior of the system over a long period. Since the interest is on the performance of the system over a long-time horizon, the impact of the initial conditions on the performance of the system is negligible.

Let X be a random variable representing the performance metric of interest over a finite or an infinite horizon. For example, assume that the daily nominal operation time of a job shop is from 8 am to 4 pm with no job in the shop at the beginning of the day and all jobs to be completed by the end of the day with overtime if needed. Further assume the daily job arrivals are random and identically distributed. The job shop may be represented by a terminating simulation model. In this situation, for instance, X may be the number of hours per day that the shop is operating. However, if the job shop closes at 4 pm, some jobs may not get finished at the end of the day, and the ending condition of one day becomes the starting condition of the next day. If the arrival of the jobs to the job shop is a stationary process, then the shop may be represented by a steady-state simulation model, and the long-term behavior of the shop is of interest. For instance, X may be the steady-state flow time of a job through the shop.

The legacy mean dominance is the prevailing selection criterion for both terminating and steady-state simulation models. The mean represents how systems perform on the average, and it is a good measure only if the consideration of risk is not important. For the job shop example, the expected number of hours per day that the shop is operating does not provide very useful information for planning for overtime operation nor for estimating the risk of cost overrun. A common misconception is to confuse *estimation error* with *risk* (Henderson and Nelson 2006). The confidence interval of a mean performance metric is only a measure of the estimation error. It is not related to the risk associated with the performance metric. For example, as the simulation run length or the number of simulation replications increase, the estimation error decreases; however, the risk associated with the performance metric of interest remains the same.

A more prudent approach for choosing the best system is to consider systems' risk. Since the variance is a surrogate for risk, an improvement over the legacy mean selection criterion is to consider both the mean and variance as the selection criteria. Suppose $E_A(X)$ and $E_B(X)$ are the mean values and $\text{Var}_A(X)$ and $\text{Var}_B(X)$ are the variance values of the performance metric of interest for Systems A and B, respectively. System A has *mean-variance* dominance over System B if $E_A(X) \geq$ (or \leq) $E_B(X)$ and $\text{Var}_A(X) \leq \text{Var}_B(X)$ in the case of larger (or smaller) is better and at least one inequality holds strictly. The limitation of the mean-variance dominance approach is that the variance measure includes the deviations both above and below the expected value; hence, the approach takes both of these deviations equally undesirable. However, in the case of larger is better, deviations above the expected value are desirable, while deviations below the expected value are undesirable. The reverse is true in the case of smaller is better. Hence, in some cases, the mean-variance dominance approach is unable to correctly select the best system.

A more comprehensive approach to R&S problem is to utilize the probability distribution of the performance metric of interest and strive to establish stochastic dominance. Suppose $F_A(x)$ and $F_B(x)$ are the cumulative distribution functions of the performance metric of interest for Systems A and B, respectively. System A stochastically dominates System B if $F_A(x) \leq$ (or \geq) $F_B(x)$ for any outcome x in the case of larger (or smaller) is better and if there is at least one x_0 for which a strong inequality holds. This means that for System A, the probability of having an outcome larger than x is more than that for System B. The mean-variance dominance and the stochastic dominance recommendations coincide precisely only in the case of normal distributions. Refer to Levy (1998) for the properties of the dominance rules and their relationships. In general, dominance established by stochastic dominance (to be precise, first-order stochastic dominance) implies dominance by the mean-variance and by the mean, in that order. However, the reverse does not hold. Therefore, the legacy mean dominance selection criterion of the R&S procedures is weaker than both the mean-variance and stochastic dominance selection criteria and ignores the inherent system's risk.

Here we will focus on the stochastic dominance based selection because it is the strongest selection criterion. Depending on the basis for comparison, the stochastic dominance may be full or partial. If the stochastic dominance between two distributions is defined for any $x \in \mathfrak{R}$, then it is called the *full* stochastic dominance. If the stochastic dominance is defined for any $x \in C$ where C is a convex subset of \mathfrak{R} , then it is called the *partial* stochastic dominance. In most problems, comparing systems based on the full dominance may not be needed; rather, a partial comparison based on either the downside risk, upside risk, or central tendency of the performance metric of interest may be required. The downside risk involves the probability of obtaining outcomes smaller than the expected value or a target value, while the upside risk involves the probability of obtaining outcomes larger than the expected value or a target value. The central tendency involves the mid-section of the probability distribution. Therefore, depending on the specific features of the R&S problem, an appropriate section of the distribution function of the performance metric of interest forms the basis for comparison.

For example, in supply chain management, the primary objective is to fulfill customer demands through the most efficient use of resources. One way to achieve this is to maintain a distribution network with the right mix of products and location of factories and warehouses to serve the market. When a company plans to start selling in a new region, they analyze different distribution network models for that region and try to determine the one with the optimal order-delivery times. Since these network models are quite complex and market conditions are uncertain, simulation is often used to study these models. If a mean-based dominance criterion is used in analyzing the simulated order-delivery data, then the best network

model is the one with the shortest expected order-delivery time. However, since on-time delivery and high level of customer service are important characteristics of a distribution network, a more comprehensive approach is to compare models based on the distribution of the order-delivery times. Considering that the upside risk related to the order-delivery times is more important to the decision maker in this case, it is more appropriate to compare the systems based on the upper sections of the order-delivery time distribution functions and thus using a partial stochastic dominance selection criterion.

To illustrate, suppose the probability distributions of the order-delivery times for two competing distribution network models, Models A and B, are as depicted in Figure 1. The performance metric of interest of Model A has $N(\mu = 48, \sigma = 10)$ distribution while that of Model B has $\text{Expo}(\mu = 45)$ distribution. If the models are compared based on the mean performance metric, then Model B is the best because it has a smaller mean performance metric than Model A. However, since the primary goal is to select a distribution network that provides high level of customer service, it is better to focus on the upper sections of the distribution functions. Hence, the model selected as the best will have lower chance of having a very long delivery time. When only the upper sections of the distribution functions are compared, we observe that Model A is the best. As seen in the example, different selection criteria lead to different selection decisions. Hence, it is important to have the flexibility to define the selection criterion based on the specific features of the selection problem and not be constrained by the legacy mean selection criterion.

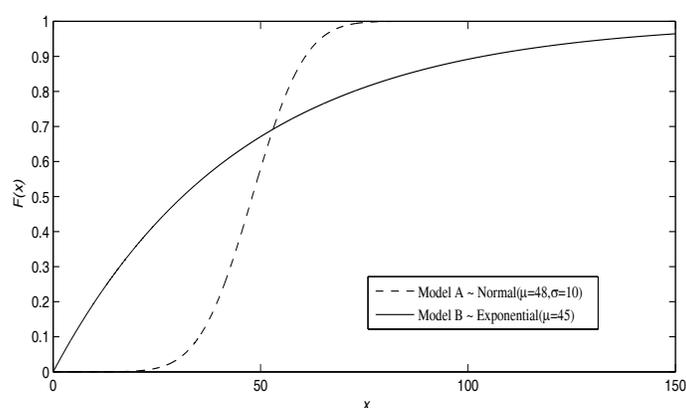


Figure 1: Possible distribution functions for the order-delivery times of two different distribution network models.

The classical goodness-of-fit tests, e.g., Chi-Square and Kolmogorov-Smirnov, are not suitable for establishing stochastic dominance. The Chi-Square test statistic measures the distance between two distributions, but it cannot be used for measuring the directed distance that is needed to establish stochastic dominance. The Kolmogorov-Smirnov test is based on the significance of the maximum distance between two empirical distributions that are built using the same intervals. [McFadden \(1989\)](#) presents a Kolmogorov-Smirnov statistic based stochastic dominance test where the statistic is based on the the maximum directed distance. However, since the dominance decision is based on only one interval where the maximum deviation is observed, the test is not powerful enough to detect if there is any crossover in other intervals. Hence, other techniques need to be considered.

Comparing systems using the stochastic dominance selection criterion requires the estimation of the probability distribution function of X . This estimation could be in terms of quantiles. But rather than using quantiles, it may be easier to use quantile values since the quantile value function is the inverse of the cumulative distribution function. In establishing the partial stochastic dominance, the experimenter determines a quantile set of interest Q and compares systems based on the estimated quantile values in the Q set. The cardinality of the Q set and the spacing between the elements of the set represents the length of the distribution and the granularity of the comparison. Being able to specify the Q set gives the decision maker the flexibility to choose the basis for comparison that is the most appropriate for the problem under consideration. Also, quantile is a more robust measure than the mean because slight changes in the distribution function have less effect on the quantile values than on the mean. Refer to [Tukey \(1960\)](#), [Huber \(1964\)](#), and [Hampel \(1968\)](#) for the theory of robustness.

The simulation literature on quantile-based selection procedures is not very rich, and the focus of the new quantile-estimation techniques has been primarily on reducing the data storage requirement. [Heidelberger and Lewis \(1984\)](#) use the maximum transformation method, which requires storing only a sequence of maxima, to obtain a point estimate of an extreme quantile. They propose using the spectral analysis or the extensions of the batch means method to estimate the confidence intervals. [Jain and Chlamtac \(1985\)](#) use the piecewise-parabolic interpolation of the statistical counts of observations to

estimate specific quantiles. [Chen and Kelton \(2008\)](#) propose obtaining quantile estimates at certain grid points and then using the Lagrange interpolation to estimate any q th quantile. The grid point approach reduces the storage requirement but introduces bias.

The earliest work in the quantile-based comparison is by [Sobel \(1967\)](#). Sobel presents a non-parametric procedure for selecting those t of the k populations which have the largest α -quantile based on n independent observations from each of the k populations. Examples of the recent quantile-based R&S procedures follow. [McNeill et al. \(2003\)](#) and [Bekki et al. \(2006\)](#) propose a heuristic technique for estimating the cycle-time quantiles from the discrete-event simulation models of manufacturing systems. Their quantile estimation technique utilizes the first four terms of the Cornish-Fisher expansion and works well only for high quantiles (> 0.75) which are of particular interest in the manufacturing environments. Moreover, the technique is applicable only when all workstations are operating under the first-in-first-out (FIFO) dispatching rule so that the cycle-time distribution is very close to the normal distribution. If non-FIFO dispatching rules are employed, [Bekki et al. \(2009\)](#) recommend combining the Cornish-Fisher expansion-based quantile estimation technique with a normalizing data transformation to address the issue of the non-normality of the cycle-time distribution. Since the Cornish-Fisher expansion-based quantile estimates are not normally distributed, they cannot be directly used with the R&S procedures that depend on the normality assumption. [Bekki et al. \(2007\)](#) recommend grouping these estimates into batches and using the batch means, which are approximately normally distributed by the Central Limit Theorem (CLT), as the basic observations in the R&S procedures. However, these studies do not generalize the quantile-based comparison idea in a way that is suitable for making comparisons based on a section of the distribution function.

In Section 2, the properties of different quantile estimators are introduced. In Section 3, two new quantile-based selection procedures are discussed. In Section 3.1, the multiple quantile-based selection procedure, where systems are compared based on the comparison of the respective quantiles, is discussed. In Section 3.2, the average quantile-based selection procedure, where the comparison of systems is performed based on the average of the quantile values, is discussed. In Section 4, some experimental results are provided to compare the performance of these two quantile-based selection procedures and to show how the mean-based and quantile-based selection decisions differ. In Section 5, open future research problems are presented followed by a conclusion in Section 6.

2 QUANTILE ESTIMATION

Let X_1, X_2, \dots, X_N be a random sample from an absolutely continuous distribution function $F(x)$ with a probability density function $f(x)$ and $X_{[1]} \leq X_{[2]} \leq \dots \leq X_{[N]}$ be the corresponding order statistics. The q th quantile, say $x^{[q]}$, $0 < q < 1$, of a random variable X with distribution F is defined as $x^{[q]} = \inf\{x : F(x) \geq q\}$, where *inf* indicates the infimum or the greatest lower bound. The simplest nonparametric point estimator of $x^{[q]}$ is the q th sample quantile (SQ) $\hat{x}^{[q]} = X_{[\lfloor Nq+1 \rfloor]}$, where $\lfloor z \rfloor$ denotes the integral part of the real number z . If $f(x)$ is differentiable in the neighborhood of $x^{[q]}$ and $f(x^{[q]}) \neq 0$, then the mean and variance of $\hat{x}^{[q]}$ are ([David 1981](#)) as follows:

$$\begin{aligned} E[\hat{x}^{[q]}] &= x^{[q]} - \frac{q(1-q)f'(x^{[q]})}{2(N+2)f^3(x^{[q]})} + O(1/N^2) \text{ and} \\ \text{Var}(\hat{x}^{[q]}) &= \frac{q(1-q)}{(N+2)f^2(x^{[q]})} + O(1/N^2). \end{aligned}$$

For the SQ estimator, it follows from the CLT that

$$\frac{\hat{x}^{[q]} - x^{[q]}}{\sqrt{\text{Var}(\hat{x}^{[q]})}} \xrightarrow{D} N(0, 1) \text{ as } N \rightarrow \infty.$$

When the sample size is large, the SQ estimator $\hat{x}^{[q]} = X_{[\lfloor Nq+1 \rfloor]}$ provides a very accurate and precise result. However, it is an inefficient estimator since it has a large variance. An approach to remedy this shortcoming is weighing of quantile values. The L-estimator for quantiles is an example of this approach. It takes a convex combination of the order statistics using an appropriate weight function, i.e., $\sum_{n=1}^N w_{N,n} X_{[n]}$, where $\sum_{n=1}^N w_{N,n} = 1$. [Harrell and Davis \(1982\)](#) proposed the L-estimator $L_q = \sum_{n=1}^N w_{N,n} X_{[n]}$ for a quantile where

$$w_{N,n} = \frac{I_n\{q(N+1), (1-q)(N+1)\} - I_{n-1}\{q(N+1), (1-q)(N+1)\}}{N}$$

and $I_x(a, b)$ denotes the incomplete beta function. The Harrell-Davis (HD) estimator is not usable when incomplete beta function is undefined, i.e., for very small N or q .

Harrell and Davis claimed that the HD estimator is asymptotically normally distributed under mild assumptions on $F(\cdot)$; and, recently, Brodin (2007) provided a correct proof for this claim. Monte Carlo studies using the Kolmogorov-Smirnov statistic have shown that, for uniform and normal distributions, the normal approximation of the HD estimator is adequate for $N \geq 20$ when $q = 0.5$ or $30 \leq N \leq 50$ when $q = 0.95$. For skewed distributions such as the exponential, sample sizes as large as 80 to 100 may be required for $q = 0.9$ or above.

3 NEW QUANTILE-BASED R&S PROCEDURES

In R&S, a typical problem is to select the best system from among K stochastic systems based on a performance metric of interest. In the quantile-based selection approach, the decision maker determines a quantile set Q corresponding to a section of the distribution function of the performance metric of interest as the basis for comparison. The section may represent: a) downside risk (lower quantiles); b) upside risk (upper quantiles); c) central tendency (central quantiles); or d) a combination of a, b, or c. The cardinality of the Q quantile set represents the coarseness or granularity of comparison. The higher the cardinality, the finer the comparison. The selection criterion for the quantile-based approach is either full or partial stochastic dominance.

In the following sections, two new quantile-based R&S procedures are discussed. Without loss of generality, it is assumed that *larger* is better.

3.1 Multiple Quantile-based Selection

In the multiple quantile-based selection approach, systems are compared based on the estimated respective quantile values in the Q quantile set. For example, let the quantile values of systems i and j corresponding to the Q set be represented by the sets $Q_i^{-1} = \{x_i^{[q]}, q \in Q\}$ and $Q_j^{-1} = \{x_j^{[q]}, q \in Q\}$, respectively. System i is better than system j if all of its quantile values corresponding to the Q set are larger than or equal to the respective quantile values of system j with at least one quantile value being strictly larger.

The two-stage quantile-based selection (QBS) procedure (Batur and Choobineh 2009) compares systems based on the quantiles in the Q set. Since the variances of systems are unknown and most likely unequal, a two-stage sampling procedure is employed to satisfy the specified confidence level of $1 - \alpha$. In a two-stage procedure, sample variances are calculated in the first stage. Then these sample variances are used to determine the number of observations needed to make a decision for each system in the second stage. The user-defined indifference-zone (IZ) parameter δ specifies the practically significant difference between the respective quantiles of two systems. The decision maker is indifferent between two systems if the difference between all respective quantile values is less than δ . In the selection procedure, the required sample sizes are determined using δ , as well as the sample variances, thereby preventing unnecessary amount of sampling to determine insignificant differences between the two systems.

The QBS procedure is an adaptation of Rinott's mean-based R&S procedure (Rinott 1978), which is well-known and easy to implement. The adaptation is straightforward since the quantile estimates (HD) satisfy the normality condition for large enough sample size N and the dominance criterion is based on the average of quantiles. In the first stage of the procedure, r_0 IID samples of size N of the metric are generated from each system, and the HD quantile estimates $\hat{x}_{i,r}^{[q]}$ for every quantile $q \in Q$ and system $i = 1, 2, \dots, K$ from each sample $r = 1, 2, \dots, r_0$ are calculated. Then the average of the quantile estimates

$$\bar{\hat{x}}_{i,r_0}^{[q]} = \frac{1}{r_0} \sum_{r=1}^{r_0} \hat{x}_{i,r}^{[q]}$$

and the standard errors

$$S_i^{[q]} = \sqrt{\frac{1}{r_0 - 1} \sum_{r=1}^{r_0} (\hat{x}_{i,r}^{[q]} - \bar{\hat{x}}_{i,r_0}^{[q]})^2}$$

for every $q \in Q$ and system i are calculated. In the second stage of the procedure, the number of samples needed for each system to select the best system with $1 - \alpha$ confidence are calculated from the IZ value δ , the sample variances, and a constant value h that depends on K , α , $|Q|$, and r_0 . The prescribed number of additional independent samples from each system are collected, and the HD quantile estimates are calculated. The pairwise comparisons are performed based on the

overall quantile estimates. If there is a system that dominates all the other systems based on every quantile in Q , then that system is selected as the best system, and the procedure stops. Otherwise, all dominated systems are eliminated, and the remaining two or more systems form the non-dominated set. In this situation, the implication is that no distinction could be made among the remaining systems using the performance metric of interest. Thus, the problem reduces to selecting a system from the non-dominated set based on some other criterion, i.e., a secondary performance metric.

Although no distinction could be made among the systems in the non-dominated set, worthy differences could be detected by a secondary performance metric. The selection of a secondary performance metric should be guided by the characteristics of the selection problem. For example, in selecting the optimal production control strategy, if the best system cannot be selected based on the minimization of the primary performance metric of *lead time*, then a secondary performance metric such as *cost* minimization may be used. If the cost function is deterministic, then the best system could easily be selected from the remaining systems. However, if the cost function is stochastic, then the quantile-based selection procedure should be reapplied on the remaining systems using this new performance metric. For the details of the *QBS* procedure, refer to [Batur and Choobineh \(2009\)](#).

The condition of dominance under the *QBS* procedure is very stringent. Comparison of each quantile in the Q quantile set may be computationally more intensive compared to a mean-based selection. But the quantile-based selection approach is the only way to detect the differences between two systems based on the downside or upside risks related to a performance metric.

3.2 Average Quantile-based Selection

A more efficient approach than the multiple quantile-based selection is to compare systems based on an aggregate of quantiles which could increase the efficiency without diminishing the flexibility of the quantile-based approach to system selection. In this approach, systems are compared based on the average quantile value in the Q quantile set of interest rather than on the individual quantiles, i.e., system i is better than system j if $\frac{1}{|Q|} \sum_{q \in Q} x_i^{[q]}$ is larger than $\frac{1}{|Q|} \sum_{q \in Q} x_j^{[q]}$. The average quantile-based selection is computationally more efficient because it requires only $|K|$ comparisons while the multiple quantile-based selection requires $|K| \times Q$ comparisons for comparing K systems to select the best system based on a Q quantile set.

The dominance decision of the multiple and average quantile-based selection approaches are equivalent when the distribution functions of two alternative systems do not cross within the basis for comparison Q . That is, the same recommendation is to be rendered by both approaches within the prespecified confidence level. However, when the distribution functions of the two alternative systems cross, the multiple quantile-based selection approach delivers a non-dominated set of two systems and requires the use of a secondary performance metric while the average quantile-based approach does not require a secondary performance metric to make a recommendation. Therefore, it can be concluded that the dominance criterion under the average quantile-based selection approach is weaker than that of the multiple quantile-based selection approach; however, the average quantile-based selection is computationally more efficient than the multiple quantile-based selection in implementation.

To illustrate, consider two systems with the distribution functions as shown in Figure 2. If the basis for comparison is $Q = \{0.1, 0.2, 0.3, 0.4\}$, the distribution functions of the two systems do not cross within the Q set. In this case, the multiple quantile-based selection approach selects System 1 as the best because the quantile values of System 1 in the Q set are all larger than the respective quantile values of System 2. Also, the average quantile-based selection approach selects System 1 as the best because the average quantile value of System 1 in the Q set is larger than that of System 2. However, if the basis for comparison is $Q = \{0.3, 0.4, 0.5, 0.6\}$, the distribution functions of the two systems cross within the Q set. In this case, the average quantile-based selection approach concludes that System 1 is the best because on average its quantile values are larger than the quantile values of System 2. However, the multiple quantile-based selection approach delivers a non-dominated set of two systems and needs a secondary performance metric to select the best system.

We propose a fully-sequential procedure based on the average quantiles (*AQ*) to select the best system from among K stochastic systems with a confidence level of $1 - \alpha$. The *AQ* procedure is an adaptation of the *KN* procedure ([Kim and Nelson 2001](#)) which is also well-known and easy to implement. The adaptation is straightforward since the quantile estimates (SQ or HD) satisfy the normality condition for large enough sample size N and the dominance criterion is based on the average of quantiles.

In a fully-sequential procedure, pairwise comparisons are performed at each replication of computation, and systems detected to be inferior are eliminated from further considerations. This approach reduces the overall computational effort in terms of the number of observations required to make a selection decision. The IZ parameter δ is the minimum difference worth detecting between the average quantile values in the Q set. When a system's average quantile is within δ of the best

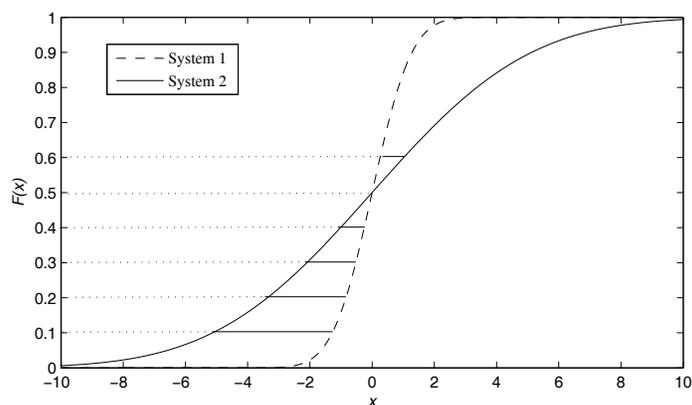


Figure 2: Two possible systems.

system’s average quantile, then that system is deemed as good as the best system, and its selection is assumed to be a correct selection.

The *AQ* procedure starts with generating r_0 IID samples of size N from each system $i = 1, 2, \dots, K$. The quantile estimates $\hat{x}_{i,r}^{[q]}$ for every quantile $q \in Q$ and system $i = 1, 2, \dots, K$ from each sample $r = 1, 2, \dots, r_0$ are calculated. Then the average of the quantile estimates

$$\bar{\hat{x}}_{i,r}(q) = \frac{1}{|Q|} \sum_{q \in Q} \hat{x}_{i,r}^{[q]} \text{ for every } i = 1, 2, \dots, K \text{ and } r = 1, 2, \dots, r_0$$

and the overall average quantile estimates

$$\bar{\hat{x}}_i(q, r_0) = \frac{1}{r_0} \sum_{r=1}^{r_0} \bar{\hat{x}}_{i,r}(q) \text{ for } i = 1, 2, \dots, K$$

are calculated. For all $i \neq \ell$, the sample variance of the differences between systems i and ℓ

$$S_{i\ell}^2 = \frac{1}{r_0 - 1} \sum_{r=1}^{r_0} (\bar{\hat{x}}_{i,r}(q) - \bar{\hat{x}}_{\ell,r}(q) - [\bar{\hat{x}}_i(q, r_0) - \bar{\hat{x}}_{\ell}(q, r_0)])^2$$

are calculated. Then

$$W_{i\ell}(r) = \max \left\{ 0, \frac{\delta}{2r} \left(\frac{h^2 S_{i\ell}^2}{\delta^2} - r \right) \right\}$$

are calculated for each pair of systems when $r = r_0$. It determines how far the average quantile estimate from system i may drop below the average quantile estimate from system ℓ without being eliminated. The constant h^2 depends on the confidence level α , the number of systems K , and the initial number of samples r_0 . All pairwise comparisons are performed and systems detected to be inferior are eliminated in the screening step. As a stopping rule, if the number of surviving systems is one, then that surviving system is selected as the best. Otherwise, one additional sample is generated from each surviving system, the related statistics are updated, and the procedure goes back to the screening step.

The recommended estimator for the *QBS* procedure is the HD estimator; however, the benefit of the HD estimator is not always recognizable in the case of average quantile estimation because of the reducing effect of pooling on the variance of the average quantile estimator. The HD estimator shows better performance than the SQ estimator only when the Q quantile set includes tail quantiles whose estimates are more variable in general. Since the SQ estimator is computationally more efficient than the HD estimator, the SQ estimator is the choice of estimator to be used in the *AQ* procedure when the basis for comparison does not include tail quantiles, and the HD estimator is the choice of estimator in the case of tail quantiles.

The experimental results show that the *AQ* procedure is computationally more efficient than the *QBS* procedure. If the best system dominates all the other systems based on all the quantiles in the *Q* set of the performance metric of interest, then the two procedures lead to the same best system selection decision within the prespecified confidence level. In this case, it is definitely an advantage to use the *AQ* procedure. However, if the best system's distribution function crosses with one of the system's distribution function in the *Q* set of interest, then the *AQ* procedure selects the best system based on the average quantiles of the performance metric while the *QBS* procedure delivers a non-dominated set of systems and provides the experimenter the flexibility to choose a secondary performance metric for the best system selection decision.

4 EXPERIMENTS

We conduct experiments to compare the performance of the *AQ* and *QBS* procedures. We also compare these two procedures to the mean-based selection procedure *KN* to show how the best system selection decisions of the mean-based and quantile-based selection procedures differ. The *AQ* procedure is run using both the *SQ* and *HD* quantile estimators. For all experiments, we report the probability of correct selection (*PCS*) and the average number of observations (*ANO*) required to achieve that probability of correct selection. The *ANO* statistic is calculated as the average of the total number of observations required by each procedure to make the best system selection decision over a number of macro-replications (complete repetition of the procedure). It gives a measure of the computational efficiency of the procedure. In all experiments, we assume *larger* is better.

In the experiments, two alternative systems are considered. The performance metric of interest of System 1 has the symmetric distribution normal with mean 48 and variance 100, and that of System 2 has the skewed distribution exponential with mean 45, as shown in Figure 3. The *IZ* value is set to $\delta = 0.5$. Since the distributions of the systems are assumed to be unknown, the sample size *N* is set to the recommended value 100 considering the worst possible case of skewed distribution (see Section 2). The results are provided over 10,000 macro-replications.

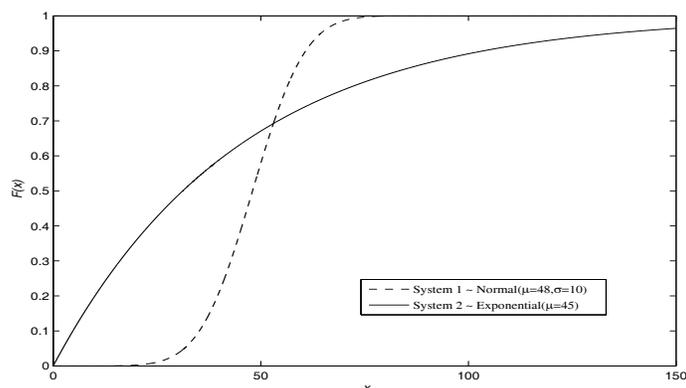


Figure 3: Distributions of the performance metric of the systems in the experiments.

Since the basis for comparison for the *KN* procedure is always the mean, we expect the *KN* procedure to select the system with the largest mean (System 1) as the best. We use different bases for comparing the *AQ* and *QBS* procedures as shown in Table 1. We expect the selection decision of the two procedures to change based on the basis for comparison.

Table 1: Experiments.

Exp.	Alternative Systems	Basis for Comparison		Best System		
		<i>AQ</i> and <i>QBS</i>	<i>KN</i>	<i>AQ</i>	<i>QBS</i>	<i>KN</i>
1	$N(\mu = 48, \sigma = 10)$ vs. $\text{Expo}(\mu=45)$	$Q = \{0.1, 0.2, 0.3\}$	μ	1	1	
2		$Q = \{0.5, 0.6, 0.7\}$		1	Either	1
3		$Q = \{0.7, 0.8, 0.9\}$		2	2	

The experimental results are presented in Table 2. In Experiments 1 and 3, the lower quantiles $Q = \{0.1, 0.2, 0.3\}$ and the upper quantiles $Q = \{0.7, 0.8, 0.9\}$ are used as the basis for comparison, respectively. In these two experiments, the *AQ* and *QBS* procedures are expected to give the same best system selection decision because their distribution functions do not cross within the Q set. In Experiment 1, the *AQ* procedure selects System 1 as the best with an ANO value of 2,000 observations when both the SQ and HD estimators are used. The *QBS* procedure selects the same system as the best with an ANO value of 49,859 observations. In Experiment 3, the *AQ* procedure selects System 2 as the best with an ANO value of 4,706 observations when the SQ estimator is used and 4,281 observations when the HD estimator is used. The *QBS* procedure selects the same system as the best with an ANO value of 780,690 observations.

Table 2: Results of the experiments.

Exp.	<i>AQ</i> (SQ)		<i>AQ</i> (HD)		<i>QBS</i>		<i>KN</i>	
	PCS	ANO	PCS	ANO	PCS	ANO	PCS	ANO
1	1.00	2,000	1.00	2,000	1.00	49,859		
2	1.00	3,950	1.00	4,268	1.00	213,742	1.00	7,876
3	1.00	4,706	1.00	4,281	1.00	780,690		

In Experiment 2, the central quantiles $Q = \{0.5, 0.6, 0.7\}$ are considered as the basis for comparison. In this case, the distribution functions of the two systems cross within the Q set. The *AQ* procedure selects System 1 as the best with an ANO value of 3,950 observations when the SQ estimator is used and 4,268 observations when the HD estimator is used. However, since the distribution functions cross, the *QBS* procedure cannot make any distinction between the two systems and delivers a non-dominated set of two systems using 213,742 observations; some other criterion, i.e., a secondary performance metric is needed to select the best system.

In all three experiments, the estimated PCS values are 1.0. When we compare the ANO values, we observe that, as expected, the *AQ* procedure is significantly more efficient than the *QBS* procedure. When we compare the performance of different quantile estimators for the *AQ* procedure, we observe that the HD estimator requires less ANO than the SQ estimator for the upper quantile set (tail quantiles for the exponential distribution) in Experiment 3; however, for the other two quantile sets, we do not observe that same reduction from the SQ estimator to the HD estimator. Hence, we recommend the HD estimator to be used in the *AQ* procedure when the Q set includes the tail quantiles; otherwise, the SQ estimator should be used.

The change in the ANO value of the *AQ* procedure is mostly under the influence of the actual differences between the average quantiles of the systems tested. As the actual difference increases, the selection problem becomes easier, and the ANO value drops. However, the high variability in the estimation of the tail quantiles compared to non-tail quantiles increases the ANO values. For example, in Experiments 1, 2, and 3, the actual differences between the average quantiles are 28.90, 8.39, and 19.91, respectively. Consequently, Experiment 2 is expected to have the largest ANO value and Experiment 1 the smallest ANO value. However, since the tail quantiles of the exponential distribution in Experiment 3 have an increasing effect on the ANO value, Experiment 3 has the largest ANO value. This increasing effect is less severe in the *AQ* procedure than the *QBS* procedure because of the pooling effect of averaging of quantiles in the *AQ* procedure.

As expected, the *KN* procedure selects System 1 as the best with an ANO value of 7,876 observations. We observe that the *AQ* procedure requires fewer observations to make a selection than the *KN* procedure because the actual difference between the means is smaller than the actual difference between the average quantile values. We also see how the quantile-based selection procedures detect the differences between systems based on the downside or upside risks while the mean-based selection procedure does not provide the experimenter the flexibility to determine a selection criterion other than the mean measure. We recommend a quantile-based selection procedure to be used when the comparison of systems based on the mean is not warranted and it is more appropriate to compare systems based on the downside or upside risks.

5 OPEN RESEARCH PROBLEMS

In addition to the procedures discussed above, the following are open research problems:

- Mean-variance dominance has the disadvantage of being a weaker selection criterion than stochastic dominance. Furthermore, the variance is not a good measure for risk when deviation above or below the expected value is of interest. However, mean-variance dominance is computationally more efficient to test than the stochastic dominance

selection criterion. A remedy to the shortcoming of the variance measure is to use semi-variance (Choobineh 1998). The semi-variance is a measure of the dispersion of all observations that fall below or above the mean performance metric. The *upside* semi-variance is a measure of the deviations above the expected value, while the *downside* semi-variance is a measure of the deviations below the expected value. In the case of larger is better, a larger upside semi-variance and a smaller downside semi-variance are desirable. The reverse is true in the case of smaller is better.

- One-sided hypotheses testing procedures for stochastic ordering based on the likelihood ratio tests are presented in Shaked and Shanthikumar (1994). In these tests, the distribution probabilities are estimated by the method of maximum likelihood estimation. Developing ranking and selection procedures based on the likelihood ratio test is an open research problem.
- The distribution functions of stochastic processes may be estimated using Kernel density estimation techniques. Refer to Alexopoulos (2006) for different Kernel density estimation techniques. There is a great potential in using Kernel density estimation to improve the performance of stochastic dominance comparisons. Developing ranking and selection procedures based on the Kernel density estimators is also an open research problem.
- Sobel (1967) present a non-parametric procedure for selecting those t of the k populations which have the largest α -quantile based on n independent observations from each of the k populations. This procedure is shown to be statistically valid. Extension of this procedure to the comparison of systems based on a quantile set is also an open research problem.

6 CONCLUSION

The legacy approach in ranking and selection procedures compares systems based on a mean performance metric. The best system is most often deemed as the one with the largest (or smallest) mean performance metric. However, mean as a measure only represents the average behavior of the performance metric, and it is a good measure only if consideration of the system's inherent risk is not important. One way to compare systems based on the risk associated with the performance metric of interest is to compare systems based on stochastic dominance. We discussed two procedures based on the stochastic dominance selection criterion. Finally, we presented some open research problems.

REFERENCES

- Alexopoulos, C. 2006. Statistical estimation in computer simulation. In *Handbooks in operations research and management science:simulation*, ed. S. G. Henderson and B. L. Nelson, 212–220. Oxford: Elsevier.
- Batur, D., and F. Choobineh. 2009. A quantile-based approach to system selection. To appear in *European Journal of Operational Research*. Available via <<http://dx.doi.org/10.1016/j.ejor.2009.05.039>> [accessed July 07, 2009]
- Bekki, J. M., J. W. Fowler, G. T. Mackulak, and M. Kulahci. 2009. Simulation-based cycle-time quantile estimation in manufacturing settings employing non-FIFO dispatching policies. *Journal of Simulation* 3:69–83.
- Bekki, J. M., J. W. Fowler, G. T. Mackulak, and B. L. Nelson. 2006. Indirect cycle-time quantile estimation using the Cornish-Fisher expansion. Under review. Available via <<http://users.iems.northwestern.edu/~nelsonb/Publications/McneillMackulakFowlerNelson.pdf>> [accessed July 07, 2009].
- Bekki, J. M., J. W. Fowler, G. T. Mackulak, and B. L. Nelson. 2007. Using quantiles in ranking and selection procedures. In *Proceedings of the 2007 Winter Simulation Conference*, ed. S. G. Henderson, B. Biller, M. H. Hsieh, J. Shortle, J. D. Tew, and R. R. Barton, 1722–1728. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.
- Branke, J., S. E. Chick, and C. Schmidt. 2007. Selecting a selection procedure. *Management Science* 53:1916–1932.
- Brodin, E. 2007. *Extreme value statistics and quantile estimation with applications in finance and insurance*. Ph.D. thesis, Department of Mathematical Sciences and Mathematical Statistics, Chalmers University of Technology and University of Göteborg, Göteborg, Sweden.
- Chen, E. J., and W. D. Kelton. 2008. Estimating steady-state distributions via simulation-generated histograms. *Computers and Operations Research* 35:1003–1016.
- Chen, H.-C., C.-H. Chen, and E. Yücesan. 2000. Computing efforts allocation for ordinal optimization and discrete event simulation. *IEEE Transactions on Automatic Control* 45:960–964.
- Chick, S. E., and K. Inoue. 2001. New two-stage and sequential procedures for selecting the best simulated system. *Operations Research* 49 (5):732–743.
- Choobineh, F. 1998. Semivariance. In *Encyclopedia of statistical sciences*, ed. S. Kotz, C. B. Read, and D. L. Banks, 616–617. New York: John Wiley & Sons, Inc.

- David, H. A. 1981. *Order statistics*. 2nd ed. New York: John Wiley & Sons, Inc.
- Hampel, F. R. 1968. *Contributions to the theory of robust estimation*. Ph.D. thesis, Department of Statistics, University of California, Berkeley, California.
- Harrell, F. E., and C. E. Davis. 1982. A new distribution-free quantile estimator. *Biometrika* 69:635–640.
- Heidelberger, P., and P. A. W. Lewis. 1984. Quantile estimation in dependent sequences. *Operations Research* 32 (1):185–209.
- Henderson, S. G., and B. L. Nelson. 2006. Stochastic computer simulation. In *Handbooks in operations research and management science:simulation*, ed. S. G. Henderson and B. L. Nelson, 1–18. Oxford: Elsevier.
- Hong, L. J. 2006. Fully sequential indifference-zone selection procedures with variance-dependent sampling. *Naval Research Logistics* 53:464–476.
- Huber, P. J. 1964. Robust estimation of location parameters. *Annals of Mathematical Statistics* 35:73–101.
- Jain, R., and I. Chlamtac. 1985. The p^2 algorithm for dynamic calculation of quantiles and histograms without storing observations. *Communications of the ACM* 28:1076–1085.
- 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* 11:251–273.
- Kim, S.-H., and B. L. Nelson. 2006. Selecting the best system. In *Handbooks in operations research and management science:simulation*, ed. S. G. Henderson and B. L. Nelson, 501–534. Oxford: Elsevier.
- Levy, H. 1998. *Stochastic dominance: investment decision making under uncertainty*. Boston: Kluwer Academic Publishers.
- McFadden, D. 1989. Testing for stochastic dominance. In *Studies in the economics of uncertainty*, ed. T. B. Fomby and T. K. Seo, 113–134. New York: Springer-Verlag.
- McNeill, J. E., G. T. Mackulak, and J. W. Fowler. 2003. Indirect estimation of cycle time quantiles from discrete event simulation models using the cornish-fisher expansion. In *Proceedings of the 2003 Winter Simulation Conference*, ed. S. Chick, P. J. Sánchez, D. Ferrin, and D. J. Morrice, 1377–1382. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.
- Pichitlamken, J., B. Nelson, and L. Hong. 2006. A sequential procedure for neighborhood selection-of-the-best in optimization via simulation. *European Journal of Operational Research* 173:283–298.
- Rinott, Y. 1978. On two-stage selection procedures and related probability-inequalities. *Communications in Statistics-Theory and Methods* 7:799–811.
- Shaked, M., and J. G. Shanthikumar. 1994. *Stochastic orders and their applications*. San Diego, CA: Academic Press, Inc.
- Sobel, M. 1967. Nonparametric procedures for selecting the t populations with the largest α -quantiles. *The Annals of Mathematical Statistics* 38 (6):1804–1816.
- Tukey, J. W. 1960. A survey of sampling from contaminated distributions. In *Contributions to probability and statistics: essays in honor of Harold Hotelling*, ed. I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, and H. B. Mann, 448–485. Stanford, CA: Stanford University Press.

AUTHOR BIOGRAPHIES

DEMET BATUR is a postdoctoral research associate in the Industrial and Management Systems Engineering Department at the University of Nebraska-Lincoln. She received her Ph.D. in Industrial and Systems Engineering from the Georgia Institute of Technology. Her research interests are in simulation output analysis and ranking and selection procedures. She is a member of INFORMS and IIE. Her email address is <dbatur2@unl.edu>.

F. FRED CHOOBINEH is a professor of Industrial and Management Systems Engineering and Milton E. Mohr Distinguished Professor of Engineering at the University of Nebraska-Lincoln where he also holds a courtesy appointment as a professor of Management. He received his B.S. in Electrical Engineering, Masters of Engineering, and Ph.D. in Industrial Engineering from Iowa State University in Ames, Iowa. He is a Fellow of the Institute of Industrial Engineers and a member of IEEE and INFORMS. His research interests include design and control of manufacturing systems and decision analysis. His research has been funded by NSF and industry. He is a registered Professional Engineer in Nebraska. His email address is <fchoobineh1@unl.edu>.