A quantile-based approach to system selection

D. Batur, F. Choobineh

Industrial and Management Systems Engineering
University of Nebraska-Lincoln
Lincoln, NE 68588

Abstract

We propose a quantile-based ranking and selection (R&S) procedure for comparing a finite set of stochastic systems via simulation. Our R&S procedure uses a quantile set of the simulated probability distribution of a performance characteristic of interest that best represents the most appropriate selection criterion as the basis for comparison. Since this quantile set may represent either the downside risk, upside risk, or central tendency of the performance characteristic, the proposed approach is more flexible than the traditional mean-based approach to R&S. We first present a procedure that selects the best system from among \( K \) systems, and then we modified that procedure for the case where \( K - 1 \) systems are compared against a standard system. We present a set of experiments to highlight the flexibility of the proposed procedures.

Keywords: simulation, ranking and selection, risk, quantile

1 Introduction

Selecting the best system out of a set of stochastic systems may be done with the aid of discrete-event simulation. Often the selection is based on a performance metric that represents a system characteristic of interest. Typically, simulation experiments are conducted to estimate a location measure for the performance metric. The most popular location measure is the mean, and the system with the largest (or smallest) mean is deemed the best system.

The primary focus of the simulation literature has been refinement of the mean-based ranking and selection (R&S) procedures, a collection of statisti-
cal procedures for comparing a finite set of stochastic systems with the goal of finding the best among them. Recent developments in the mean-based R&S procedures when the samples are independent and identically (IID), normally distributed are by Kim and Nelson (2001), Pichitlamken et al. (2006), Chick and Inoue (2001), and Chen et al. (2000). Kim and Nelson (2001) propose a fully-sequential procedure that generates observations from each system one at a time until either there is enough evidence that a system’s mean is dominated by one of the others, or a certain precision is obtained. The ultimate objective is to select a system with a guarantee on the probability of correct selection. 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. A comparison of the mean-based R&S procedures is given in Branke et al. (2007). Refer to Kim and Nelson (2006) for a comprehensive review of the R&S procedures in simulation.

![Graph of distribution functions](image)

Fig. 1. Possible distribution functions for the production lead time of two different production control strategies.

Use of the mean for selecting the best system explicitly assumes that the decision maker is primarily interested in the average behavior of the performance metric. Selecting the best system based on the mean location measure also implies that the decision maker views consideration of other behaviors of the performance metric, such as the downside risk or upside risk, as being less important or unimportant. 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. In certain applications, consideration of the
downside or upside risks is a more appropriate selection criterion than the expected performance.

For example, when a production system is being designed or modified, a number of production control strategies are compared to select the best strategy. The metric of interest for selecting the best control strategy is often the production lead time, the amount of time it takes to process a part, and the control strategy with the shortest lead-time is deemed to be the best. Since the on-time availability of materials and resources is often uncertain, the lead time metric becomes uncertain, and it is treated as a random variable. If we are interested in utilizing a control strategy that minimizes the average lead time, then we select the control strategy that results in the shortest expected lead time. However, if customer service and on-time delivery are important, the control strategies need to be compared based on the upper section of their production lead-time distributions (upside risk). For instance, suppose the probability distribution of lead time for two competing control strategies are as depicted in Figure 1. The control strategy with the Expo(μ = 45) distribution has a smaller mean than the control strategy with the N(μ = 48, σ = 10) distribution, and therefore it should be the strategy of choice when the mean lead time is the selection criterion. However, the control strategy with the N(μ = 48, σ = 10) distribution should be the choice when the upside risk is considered. In order to have the choice of using a selection criterion other than the mean, we propose the use of a quantile measure.

![Figure 1](image1.png)

**Fig. 2.** Difference between the quantiles when the means are the same.

Quantiles, as a measure for system selection, provide the flexibility to design an R&S procedure that is more discerning than an R&S procedure that uses the mean. For example, suppose the underlying distributions of a performance metric for two alternative systems are N(μ=40, σ=10) and Expo(μ=40) as shown in Figure 2. Since both distributions have the same mean, a mean-based R&S procedure would not detect any difference between the two systems.
However, quantiles help detect the difference between the two systems. For instance, for the small and central quantiles, the $N(\mu=40, \sigma=10)$ system has larger value than the $\text{Expo}(\mu=40)$ system; however, the $\text{Expo}(\mu=40)$ system has larger value than the $N(\mu=40, \sigma=10)$ system for the large quantiles. This flexibility allows the decision maker to compare and select systems based on the criterion that is the most appropriate for the problem under consideration. The basis for comparison may be the central tendency, downside risk, or upside risk associated with a performance metric. When the basis for comparison is the central tendency and the distributions tested are symmetric, the proposed quantile-based procedure will be akin to a mean-based procedure.

Measures that characterize a distribution, such as the measures of location and scale, are said to be robust if slight changes in their distribution have a relatively small effect on their value. The theory of robustness was developed by Tukey (1960), Huber (1964), and Hampel (1968). They have shown that sample mean and sample standard deviation are not robust measures. For smaller sample sizes, sample mean and variance may provide an unsatisfactory summary of data. Hence, R&S procedures that use mean as a location measure may have low testing power or low coverage probability caused by a single unusual value. A location measure that is robust is the quantile measure (see Section 2).

We propose a flexible, quantile-based R&S procedure for comparing a finite set of stochastic systems. The flexibility associated with the use of quantiles allows the decision maker to compare systems based on the behavior of a performance metric that is the most appropriate for the problem under consideration. This flexibility transcends to the following two cases: (i) the decision maker carefully chooses a set of quantiles representing the section of the distribution that is the most appropriate, e.g., the downside risk, for system comparison; or (ii) one threshold value for the metric of interest is specified, and systems are compared based on their behavior on one side of the threshold value.

In Section 2, the robustness of quantile as a measure of location is discussed and the properties of different quantile estimators are introduced. In Section 3, quantile-based system comparison is introduced. The procedures for selecting the best system or comparing systems with a standard system given a performance metric quantile set are presented in Section 3.1. The case where the quantile set is defined by a threshold value is presented in Section 3.2. Finally, experimental results are presented in Section 4, followed by a conclusion in Section 5.
2 Quantile properties and estimation

In Section 2.1, we introduce the influence function as a tool to contrast the robustness of the mean and variance measures with that of the quantile measure and provide a formal definition of quantile. In Section 2.2, we address the issues related to the estimation of quantiles.

2.1 Robust estimation

A measure of location is said to be robust if slight changes in a distribution have a relatively small effect on its value. The robustness of an estimator may be demonstrated by examining its influence function (Hampel, 1974). The influence function of an estimator $T$ measures the effect of a new observation $x$ on the estimator. If this effect is bounded for any $x$, then the estimator is said to be a robust estimator.

Let $\mathbb{R}$ be the real line, $F$ be a probability distribution function on $\mathbb{R}$, $T$ be an estimator of a measure of the distribution $F$, and $\Delta_x$ be a probability distribution where $x \in \mathbb{R}$ occurs with probability 1. If an observation $x$ is generated from distribution $\Delta_x$ with probability $\epsilon$, where $0 < \epsilon < 1$, then the value of the estimator becomes $T((1-\epsilon)F + \epsilon \Delta_x)$. Hence, the effect of the new observation on $T$ is the difference $T((1-\epsilon)F + \epsilon \Delta_x) - T(F)$. The influence function is defined by dividing the difference caused by the observation $x$ by the probability of the observation $x$ and taking the limit with respect to the probability $\epsilon$ going to zero. Therefore, the influence function of $T$ at $F$ is defined asymptotically by

$$IF_{T,F}(x) = \lim_{\epsilon \downarrow 0} \frac{T((1-\epsilon)F + \epsilon \Delta_x) - T(F)}{\epsilon}$$ (1)

assuming the limit exists for every point $x \in \mathbb{R}$. For the finite-sample version of the influence function, refer to Hampel et al. (1986), p. 93.

Let $\mu$ and $\sigma^2$ be the mean and variance of the distribution function $F$, respectively. If the mean and variance exist, then the influence function of the mean is

$$IF_{\mu,F}(x) = \lim_{\epsilon \downarrow 0} \frac{(1-\epsilon)\mu + \epsilon x - \mu}{\epsilon} = x - \mu, \ x \in \mathbb{R}.$$ (2)
The influence function of the variance is

\[ IF_{\sigma^2, F}(x) = \lim_{\epsilon \to 0} \frac{(1 - \epsilon)\sigma^2 + \epsilon(x - \mu)^2 - \sigma^2}{\epsilon} = (x - \mu)^2 - \sigma^2, \quad x \in \mathbb{R}. \] (3)

As seen in Equations 2 and 3, both influence functions are unbounded in \( x \); therefore, neither the mean nor the variance are robust measures.

For a random sample, \( X_1, X_2, \ldots, X_N \), let \( \bar{X} = \sum_{n=1}^{N} X_n/N \) be the sample mean and \( S^2 = \sum_{n=1}^{N} (X_n - \bar{X})^2/(N - 1) \) be the sample variance. Since \( \mu \) and \( \sigma^2 \) are the asymptotic values of \( \bar{X} \) and \( S^2 \), we can claim that sample mean and sample variance also have unbounded influence functions.

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 influence function for the quantile (Staudte and Sheather (1990), pp. 59-61) is

\[ IF_{x[q], F}(x) = \begin{cases} 
\frac{q-1}{f(x[q])} & \text{if } x < x[q], \\
0 & \text{if } x = x[q], \\
\frac{q}{f(x[q])} & \text{if } x > x[q],
\end{cases} \] (4)

where \( f \) is the probability density function. As seen in Equation 4, \( x[q] \) has a bounded influence function. Therefore, unlike the mean and variance, the quantile measure is not susceptible to the value of an outlier.

2.2 Quantile estimation

Let \( X_1, X_2, \ldots, X_N \) be a random sample from an absolutely continuous distribution function \( F(x) \) with a probability density function \( f(x) \). Let \( X_{[1]} \leq X_{[2]} \leq \ldots \leq X_{[N]} \) be the corresponding order statistics. The simplest nonparametric point estimator of \( x[q] \) is the \( q \)th sample quantile \( \hat{x}[q] = X_{[\lfloor Nq \rfloor + 1]} \) 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), pp. 84):

\[
\begin{align*}
E(\hat{x}[q]) &= x[q] - \frac{q(1 - q)f'(x[q])}{2(N + 2)f^3(x[q])} + O(1/N^2), \\
\text{Var}(\hat{x}[q]) &= \frac{q(1 - q)}{(N + 2)f^2(x[q])} + O(1/N^2).
\end{align*}
\]
For the sample quantile estimator, it follows from the Central Limit Theorem (CLT) that

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

When the sample size is large, the sample quantile 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 weighting of the 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.,

\[ \sum_{n=1}^{N} w_{N,n} X_{[n]} \text{ where } \sum_{n=1}^{N} w_{N,n} = 1. \]

Bickel and Lehmann (1975) show that L-estimators offer the best compromise between efficiency and robustness among the estimators for estimating location in the case of skewed distributions. For a thorough discussion of the properties of L-estimators, see Serfling (1980).

Harrell and Davis (1982) proposed the following L-estimator for a quantile:

\[ L_q = \sum_{n=1}^{N} w_{N,n} X_{[n]}, \]

where

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

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

Harrell and Davis (1982) claimed that the HD estimator is asymptotically normally distributed under mild assumptions on \( F(\cdot) \). Recently, Brodin (2007) provided a correct proof for this claim. Monte Carlo studies have shown that for the 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 80-100 may be required for \( q = 0.9 \) or above. The HD estimator is compared with nine different nonparametric estimators for a broad range
of quantiles (from 0.02 to 0.98) in Dielman et al. (1994). They analyzed the performance of the estimators using nine different distributions. Their results indicate that there is no best performer across all distributions. The best could only be identified in the context of the distribution, the quantile, and sample size. The HD estimator generally performed the best for a wide range of cases except for high percentiles, e.g., 0.02 and 0.98, and very skewed distributions.

The simulation literature on the quantile-based 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 (2006) propose two heuristic algorithms to determine the simulation run length for estimating quantiles that satisfy a prespecified precision requirement. Chen and Kelton (2008) propose obtaining quantile estimates at certain grid points and then using the Lagrange interpolation to estimate any qth quantile. The grid point approach reduces the storage requirement but introduces bias.

Examples of the 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 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. (2008) 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 CLT, as the basic observations in the R&S procedures. However, they do not generalize their idea in a way that is suitable for making comparisons based on a section of the distribution function.
3 Quantile-based system comparison

The decision circumstances compel the decision maker to first establish the basis for comparison. The basis for comparison may be a section $Q$ of the probability distribution that corresponds to the behavior of interest. This 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, and c. The cardinality of the $Q$ set represents the coarseness or granularity of comparison.

The decision circumstances may also compel the decision maker to compare systems based on a specific set of values of the performance metric of interest. In this case, the set of values $Q^{-1}$ is specified by the decision maker as the basis for comparison. Then the corresponding $Q$ quantile sets are determined for each system for the quantile-based system selection.

The behavior of a performance metric is fully captured and characterized by its probability distribution. The $Q$ set may represent the whole distribution or only a section of it. The cardinality and the range of the $Q$ set is an indication of the amount of information about the behavior of the metric that is deemed important to make the selection decision. When the $Q$ set does not represent the whole distribution, it signals that the decision maker is interested in a partial behavior of the metric, albeit the most important behavior. In the extreme case that the $Q$ set consists of only one element, the selection is based on the least amount of information about the behavior of the metric, i.e., one-point comparison of distributions. The mean-based selection, which is basically the comparison of systems based on the $Q^{-1}$ set that has only one element, is an example of this extreme case. In general, the larger the cardinality and the range of the $Q$ set, the more information about the behavior of the metric is in play in the selection process. Obviously, the amount of information should be tempered by the importance of the information that ultimately specifies the $Q$ set.

In Section 3.1, we present procedures for comparing systems when the $Q$ set is first specified. In Section 3.2, we present procedures for comparing systems when the $Q^{-1}$ set is first identified.

3.1 System comparison given the $Q$ set

Suppose the decision maker has specified a discrete $Q$ set as the basis for comparison. We first present a quantile-based selection ($QBS$) procedure for comparing $K$ alternative systems to select the best given the $Q$ set. In the following subsection, we present the $QBS^S$ procedure to accommodate a situ-
Procedure QBS

Step 1: Set the confidence level $1 - \alpha$, indifference-zone parameter $\delta$, sample size $N$, and initial number of samples $r_0 \geq 2$.

Step 2: Generate $r_0$ IID samples of size $N$ of the metric from each system $i = 1, 2, \ldots, K$. Calculate the HD quantile estimates $\hat{x}_{i,1}, \hat{x}_{i,2}, \ldots, \hat{x}_{i,r_0}$ and the standard error

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

where

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

for every $q \in Q$ and $i = 1, 2, \ldots, K$.

Step 3: Let $\max_{q \in Q} S_i^{[q]} = S_{i,\text{max}}$ for $i = 1, 2, \ldots, K$. Determine the total number of samples needed for each system $i$ by

$$R_i = \max \left\{ r_0, \left\lceil \left( \frac{h(K, 1 - \alpha/|Q|, r_0) S_{i,\text{max}}}{\delta} \right)^2 \right\rceil \right\}$$

where $\lceil \cdot \rceil$ indicates truncation of any fractional part.

Step 4: If $R_i > r_0$, generate $R_i - r_0$ additional samples and calculate the HD quantile estimates $\hat{x}_{i,r_0+1}^{[q]}, \hat{x}_{i,r_0+2}^{[q]}, \ldots, \hat{x}_{i,R_i}$ for every $q \in Q$ and $i = 1, 2, \ldots, K$. Then calculate the overall quantile estimates

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

for every $q \in Q$ and $i = 1, 2, \ldots, K$.

Step 5: If $\hat{x}_i^{[q]} > \hat{x}_j^{[q]}$ for every $q \in Q$ and $j \neq i$ where $i = 1, 2, \ldots, K$, then select system $i$ as the best. Otherwise go to Step 6.

Step 6: If $\hat{x}_i^{[q]} < \hat{x}_j^{[q]}$ for every $q \in Q$ and any $j \neq i$ where $i = 1, 2, \ldots, K$, then eliminate system $i$. Based on the specified performance metric, the remaining set of systems are non-dominated; a secondary performance metric needs to be used to select the best from among the non-dominated systems.

Fig. 3. Algorithmic statement for selection of the best system.
Let the quantile values of systems $i$ and $j$ corresponding to the $Q$ set be represented by the set $Q_i^{-1} = \{x_i^{[q]}, q \in Q\}$ and set $Q_j^{-1} = \{x_j^{[q]}, q \in Q\}$, respectively. Without loss of generality, we assume that larger is better. 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 larger.

Our proposed solution approach is a two-stage, indifference-zone (IZ) R&S procedure. Since the variances of systems are unknown and most likely unequal, we employ a two-stage sampling procedure to satisfy the specified confidence level of $1 - \alpha$. In a two-stage procedure, the sample variances are calculated in the first stage. Then the 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 IZ parameter $\delta$ 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 $\delta$. In the selection procedure, the required sample sizes are determined using $\delta$, as well as the sample variances, thereby preventing an unnecessary amount of sampling to determine insignificant differences between the two systems.

The $QBS$ procedure shown in Figure 3 outlines the proposed two-stage, quantile-based R&S procedure for selecting the best system from a system set $A = \{a_1, a_2, \ldots, a_K\}$ given the $Q$ quantile set. This is an adaptation of Rinott’s mean-based R&S procedure (Rinott, 1978) which is well-known and easy to implement.

In the $QBS$ procedure, the HD estimate of the $q$th quantile of system $i$ calculated from sample $j$ is denoted by $\hat{x}_{i,j}^{[q]}$. In Step 1 of the procedure, the user-specified parameters are set. In Step 2, $r_0$ independent samples of size $N$ are obtained from each system. Then the HD quantile estimates and the sample variances of these quantile estimates are calculated for each system. Using the IZ value and the maximum sample variance per system, Step 3 identifies the approximate number of samples needed for each system so that the procedure can select one system over another with $1 - \alpha$ confidence. A FORTRAN program to calculate the value of the constant $h(K, 1 - \alpha/|Q|, r_0)$ is given in Appendix C of Bechhofer et al. (1995). On pages 62-63 of the same reference, a limited table for the constant is generated. The constant $h$ increases in $K$ and $|Q|$ and decreases in $r_0$ and $\alpha$. In Step 4, the prescribed number of additional independent samples from each system are collected, and the overall sample quantile estimates are calculated. In Step 5, the system with the largest overall quantile estimates for every $q \in Q$ is selected as the best system. If one system (the best) is not identified in Step 5, then the procedure goes to Step 6. In
Step 6, 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 between 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 between the systems in the non-dominated set, there are worthy differences that 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 could 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.

The QBS procedure requires the input observations to be IID which is satisfied if multiple replications of the simulation model are performed. However, if a single long replication of the simulation model—in the case of steady-state simulations—is performed, then the IID data assumption can only be satisfied approximately by way of batching within the replication. The batch sample averages are assumed to be IID if the stochastic process is stationary and the batches are sufficiently large (Law and Kelton, 2000).

A correct selection in quantile-based selection is defined as the event that the procedure selects system \( i \) as the best when \( x_i^{[q]} \geq x_j^{[q]} + \delta, \forall q \in Q \) and \( j \neq i \).

A statistically valid quantile-based selection procedure should guarantee the following probability of correct selection statement:

\[
\text{PCS} \equiv \Pr\{\text{select system } i \mid x_i^{[q]} \geq x_j^{[q]} + \delta, \forall q \in Q \text{ and } j \neq i\} \geq 1 - \alpha.
\]

**Theorem 1** The QBS procedure satisfies PCS \( \geq 1 - \alpha \).

**Proof:** Let \( \text{CS}^{[q]} \) and \( \text{ICS}^{[q]} \) be the correct and incorrect selection events, respectively, when only one quantile \( q \) is considered. Rinott’s procedure is shown to be a statistically valid procedure (selects the system with the best mean with a confidence level \( 1 - \alpha \)) when the samples from each system are IID normal and systems are simulated independently (Rinott, 1978). This directly implies that \( \Pr\{\text{CS}^{[q]}\} \geq 1 - \alpha \) in the QBS procedure because the quantile estimates \( \hat{x}_i^{[q]}, \hat{x}_{i,1}, \hat{x}_{i,2}, \ldots \) for each system \( i \) are approximately IID normal for large enough sample size \( N \) (see Section 2.2) and systems are simulated independently. However, if \( |Q| > 1 \) quantiles are considered, the dependence of the quantile estimates from the same system (they originate from the same
sample) require certain adjustments in the $h$ function: The $\alpha$ parameter in the $h$ function is divided by $|Q|$ so that $\Pr\{CS^{[q]}\} \geq 1 - (\alpha/|Q|)$. Then

$$\text{PCS} = \Pr\{\bigcap_{q \in Q} CS^{[q]}\}$$

$$\geq 1 - \sum_{q \in Q} \Pr\{ICS^{[q]}\}$$

$$\geq 1 - |Q|\frac{\alpha}{|Q|}$$

$$= 1 - \alpha,$$

where the first inequality follows from the Bonferroni inequality. □

3.1.2 Comparison with a standard system

**Procedure QBS$^S$**

**Step 1-4:** Same as Steps 1-4 in the QBS procedure.

**Step 5:** For any $i, j = 2, \ldots, K$ if $\hat{x}_i^{[q]} > \hat{x}_1^{[q]} + \delta$ and $\hat{x}_i^{[q]} > \hat{x}_j^{[q]}$ for every $q \in Q$ and $j \neq i$, then select system $i$ as the best. Otherwise retain system 1, the standard, as the best system.

Fig. 4. Algorithmic statement for comparison with a standard system.

It is often the case that a standard system exists and the performance of the alternative systems need to be compared with the standard based on a performance metric. The standard system may be the existing (base-case) system or an ideal system. In order for a system to be deemed better than the standard system, it needs to outperform the standard system by at least a predetermined margin $\delta$. Otherwise the standard system is retained. If there are more than one system that are better than the standard system, then the objective is to select the best among those systems. Two recent mean-based comparison with a standard procedures are by Nelson and Goldsman (2001) and Kim (2005).

The QBS$^S$ procedure shown in Figure 4 is the quantile-based comparison with a standard procedure. In the procedure, $K$ systems–system 1 is designated as the standard system–are compared based on the $Q$ quantile set of a performance metric. It is assumed that the standard system is also simulated. The QBS$^S$ procedure consists of five steps of which the first four are the same as the first four steps of the QBS procedure; however, the decision criterion in Step 5 is changed reflecting the minimum margin $\delta$ that is needed to eliminate the standard system.
3.2 System comparison given the $Q^{-1}$ set

In this type of selection problem, a threshold value, $x_t$, for the performance metric of interest is specified. The threshold value $x_t$ represents a demarcation above or below which the systems are to be compared, i.e., $Q^{-1} = \{x : x^{[b]} \leq (\text{or} \geq)x_t\}$. The objective here is to choose the system that performs the best based on the $Q$ quantile set defined by $Q^{-1}$. For example, when comparing alternative production control strategies, the best control strategy may be defined as the one with the largest production rate below the threshold value of expected maximum demand.

Without loss of generality, we assume that the interest is on the comparison of systems below the threshold value $x_t$ and larger is better. Then the value $q(t)$ that corresponds to $x_t$ is the probability of falling below the threshold value, and it gives the upper quantile for evaluation of the downside risk. When comparing two systems, the system selection cannot be made solely on the respective $q(t)$ values since the two distributions may be crossing below the $q(t)$s. Therefore, for the downside risk, a quantile interval bounded by zero and $q(t)$ needs to be considered.

To establish a quantile set $Q_i$ for each system $i$, we first establish a $100(1-\alpha)\%$ confidence interval for $q_i(t)$. To establish that confidence interval, we assume that falling below the threshold is a success and get a point estimate for $q_i(t)$ by $\hat{q}_i(t) = s/N$, where $s$ represents the number of successes in $N$ trials. By assuming that $\hat{q}_i(t) = s/N$ is approximately normally distributed by the CLT, the upper one-sided $100(1-\alpha)\%$ confidence bound for the binomial parameter $q_i(t)$ is given by

$$\hat{q}_i^{u}(t) = \hat{q}_i(t) + z_\alpha \sqrt{\frac{\hat{q}_i(t)(1-\hat{q}_i(t))}{N}},$$

where $z_\alpha$ is the quantile that leaves an area of $\alpha$ to the right of the standard normal distribution and $N$ is the sample size.

The normal distribution is a good approximation to the binomial distribution when $N$ is sufficiently large and $q_i(t)$ is not too close to 0 or 1. How large $N$ needs to be depends on the value of $q_i(t)$. A good rule of thumb is to use the normal approximation and hence the confidence-interval procedure established above only if $N\hat{q}_i(t) \geq 5$ and $N(1-\hat{q}_i(t)) \geq 5$ (Walpole et al., 2007).

We use the interval between zero and the upper bound $\hat{q}_i^{u}(t)$ as the interval quantile set of system $i$. Let $Q_i = [0,q_i(t)]$ and $Q_j = [0,q_j(t)]$ be the downside-risk quantile intervals of systems $i$ and $j$, respectively. The quantile-based system comparison requires a common quantile set for every pairwise...
comparison. Two options for forming a common quantile interval for systems \( i \) and \( j \) are either \( Q_{ijL} = Q_i \cap Q_j \) or \( Q_{ijU} = Q_i \cup Q_j \). The proposition below shows that, for a given \( x_t \), the two distribution functions \( F_i \) and \( F_j \) do not cross between \( F_i(x_t) = q_i(t) \) and \( F_j(x_t) = q_j(t) \). Hence, we should be indifferent between the choice of \( Q_{ijL} \) or \( Q_{ijU} \). However, we choose the common quantile set \( Q_{ij} \) as \( Q_{ijL} \) because it is the shorter of the two intervals.

**Lemma:** Two distributions \( F_i \) and \( F_j \) cannot cross in the interval \( Q_{ijU} \setminus Q_{ijL} \).

**Proof:** Assume that \( q_j(t) < q_i(t) \) without loss of generality. Take any \( q \) such that \( q_j(t) < q < q_i(t) \).

Since distribution functions are nondecreasing functions, the following relationship holds:

\[
F_i^{-1}(q_j(t)) < F_i^{-1}(q) < F_i^{-1}(q_i(t)).
\]

Since \( F_i^{-1}(q_i(t)) = x_t \),

\[
F_i^{-1}(q) < x_t. \quad (5)
\]

Similarly, the following relationship holds:

\[
F_j^{-1}(q_j(t)) < F_j^{-1}(q) < F_j^{-1}(q_i(t)).
\]

Again, since \( F_j^{-1}(q_j(t)) = x_t \),

\[
x_t < F_j^{-1}(q). \quad (6)
\]

Thus, Equations 5 and 6 show that the two distribution functions cannot cross between the minimum and maximum quantiles. \( \square \)

For the system selection procedure, the quantile intervals are converted into a discrete set. The required coarseness of the analysis determines the cardinality of the quantile sets.

The \( QBS^T \) procedure shown in Figure 5 outlines the two-stage R&S procedure for selecting the best system from a system set \( A = \{a_1, a_2, \ldots, a_K \} \) given the target value \( x_t \). The \( \alpha \) value in the estimation of the confidence bound is adjusted so that the confidence level of the simultaneous \( K \) one-sided confidence intervals is \( 1 - \alpha_0 \). From the \( QBS \) procedure, we know that once the quantile sets are determined, the confidence level for selecting the best system is \( 1 - \alpha_1 \). Since \( \alpha_0 + \alpha_1 = \alpha \), the overall PCS for the \( QBS^T \) procedure is at least \( 1 - \alpha \).
Procedure \( \text{QBS}^T \)

**Step 1:** Set \( \alpha_0 \) and \( \alpha_1 \) such that \( \alpha = \alpha_0 + \alpha_1 \) and the overall confidence level is \( 1 - \alpha \). Specify the indifference-zone parameter \( \delta \), sample size \( N \), initial number of samples \( r_0 \geq 2 \), and target value \( x_t \).

**Step 2:** Generate an IID sample of size \( N \) of the metric from each system \( i = 1, 2, \ldots, K \). Estimate the proportion of observations \( \hat{q}_i(t) \) that fall below \( x_t \), and form the quantile set \( Q_i = [0, \hat{q}_i(t)] \) where

\[
\hat{q}_i(t) = \hat{q}_i(t) + z_{1-(1-\alpha_0)^{1/K}} \sqrt{\frac{\hat{q}_i(t)(1-\hat{q}_i(t))}{N}},
\]

for \( i = 1, 2, \ldots, K \). Discretize the \( Q_i \) sets by applying the same discretization rule to all systems.

**Step 3:** Generate \( r_0 \) IID samples of size \( N \) of the metric from each system \( i = 1, 2, \ldots, K \). Calculate the HD quantile estimates \( \hat{x}_{i[q]} \), standard error \( S_i[q] \) for every \( q \in Q_i \) and \( i = 1, 2, \ldots, K \).

**Step 4:** Let \( \max_{q \in Q_i} S_i[q] = S_{i,\max} \) for \( i = 1, 2, \ldots, K \). Determine the total number of samples needed for each system \( i \) by

\[
R_i = \max\left\{ r_0, \left\lceil \left( \frac{h(K, 1-\alpha_1/|Q_i|, r_0) S_{i,\max}}{\delta} \right)^2 \right\rceil \right\}.
\]

**Step 5:** If \( R_i > r_0 \), generate \( R_i - r_0 \) additional samples and calculate the HD quantile estimates \( \hat{x}_{i[q]} \) for every \( q \in Q_i \) and \( i = 1, 2, \ldots, K \). Then calculate the overall quantile estimates

\[
\hat{x}_{i[q]} = \frac{1}{R_i} \sum_{r=1}^{R_i} \hat{x}_{i[r]} \text{ for every } q \in Q_i \text{ and } i = 1, 2, \ldots, K.
\]

**Step 6:** Let \( Q_{ij} = Q_i \cap Q_j \) where \( i, j = 1, 2, \ldots, K \) and \( j \neq i \). If \( \hat{x}_{i[q]} > \hat{x}_{j[q]} \) for every \( q \in Q_{ij} \) and \( j \neq i \) where \( i = 1, 2, \ldots, K \), then select system \( i \) as the best. Otherwise go to Step 7.

**Step 7:** If \( \hat{x}_{i[q]} < \hat{x}_{j[q]} \) for every \( q \in Q_{ij} \) and any \( j \neq i \) where \( i = 1, 2, \ldots, K \), then eliminate system \( i \). Based on the specified performance metric, the remaining set of systems are non-dominated; a secondary performance metric needs to be used to select the best from among the non-dominated systems.

Fig. 5. Algorithmic statement for selection of the best system considering a target.
4 Numerical examples

Table 1
Experiments.

<table>
<thead>
<tr>
<th>Exp.</th>
<th>Alternative Systems</th>
<th>Basis for Comparison</th>
<th>Best system</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>QBS Quantile Set</td>
<td>$\kappa N$</td>
</tr>
<tr>
<td>1</td>
<td>$N(\mu = 0, \sigma = 1)$ vs.</td>
<td>{0.1, 0.2, 0.3}</td>
<td>$\mu$</td>
</tr>
<tr>
<td>2</td>
<td>$N(\mu = 0, \sigma = 4)$</td>
<td>{0.7, 0.8, 0.9}</td>
<td>$\mu$</td>
</tr>
<tr>
<td>3</td>
<td>$N(\mu = 48, \sigma = 10)$ vs.</td>
<td>{0.1, 0.2, 0.3}</td>
<td>$\mu$</td>
</tr>
<tr>
<td>4</td>
<td>Expo($\mu = 45$)</td>
<td>{0.5, 0.6, 0.7}</td>
<td>$\mu$</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>{0.7, 0.8, 0.9}</td>
<td></td>
</tr>
</tbody>
</table>

Table 2
Experiment results.

<table>
<thead>
<tr>
<th>Exp.</th>
<th>QBS</th>
<th>$\kappa N$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PCS</td>
<td>SATO</td>
</tr>
<tr>
<td>1</td>
<td>1.00</td>
<td>$211.8 \times 10^3$</td>
</tr>
<tr>
<td>2</td>
<td>1.00</td>
<td>$215.3 \times 10^3$</td>
</tr>
<tr>
<td>3</td>
<td>1.00</td>
<td>$50.7 \times 10^3$</td>
</tr>
<tr>
<td>4</td>
<td>1.00</td>
<td>$213.1 \times 10^3$</td>
</tr>
<tr>
<td>5</td>
<td>1.00</td>
<td>$799.5 \times 10^3$</td>
</tr>
</tbody>
</table>

Fig. 6. Systems with the same mean.

We present five numerical examples (experiments) to demonstrate the flexibility of the proposed quantile-based R&S procedure QBS and compare its selection recommendation with that of the mean-based R&S procedure $\kappa N$. 
(Kim and Nelson, 2001) when appropriate. The $\mathcal{KN}$ procedure requires the observations to be IID normal and is also shown to be robust to non-normality (Kim and Nelson, 2001).

We replicate each experiment 1,000 times. The IZ value is set to $\delta = 0.05$ in the first two experiments and $\delta = 0.5$ in the last three experiments. The sample size $N$ for the HD quantile estimator is set to the recommended value of 100 considering that the distributions are unknown and may possibly be skewed. We set the nominal confidence level to $1 - \alpha = 0.95$ and the initial number of samples to $r_0 = 10$. We report the estimated probability of correct selection (PCS) and the sample average of total number of observations (SATO) that is used to make a decision. The SATO value provides a relative measure of the efficiency of each procedure. In all the experiments, we assume larger is better.

Table 1 lists the configurations of the experiments followed by the results in Table 2. In Experiments 1 and 2, two systems are compared based on one performance metric. The distributions of the performance metric of interest are assumed to be normal with the same mean but different variance values, i.e., $N(\mu = 0, \sigma = 1)$ and $N(\mu = 0, \sigma = 4)$. Figure 6 depicts these two distributions. The $\mathcal{KN}$ procedure compares and selects systems based on the mean performance metric and is indifferent between two systems if there is less than $\delta$ difference between the mean values of the performance metric of interest. Therefore, it can select either of these two systems as the best, and either selection is considered as a correct selection. The result of 1,000 replications shows that the $\mathcal{KN}$ procedure selects the $N(\mu = 0, \sigma = 1)$ system as the best 508 times and the $N(\mu = 0, \sigma = 4)$ system as the best 492 times. The SATO value of the $\mathcal{KN}$ procedure is $39.4 \times 10^3$ observations.

For the $QBS$ procedure, we use the set of lower quantiles $Q = \{0.1, 0.2, 0.3\}$ in Experiment 1 and upper quantiles $Q = \{0.7, 0.8, 0.9\}$ in Experiment 2 as the basis for comparison. In Experiment 1, the $QBS$ procedure correctly identifies the $N(\mu = 0, \sigma = 1)$ system as the best with the estimated PCS of 1.00 and a SATO value of $211.8 \times 10^3$ observations. In Experiment 2, the $QBS$ procedure correctly identifies the $N(\mu = 0, \sigma = 4)$ system as the best with the estimated PCS of 1.00 and a SATO value of $215.3 \times 10^3$ observations.

As seen by the results, the $QBS$ procedure under differing bases for comparison is able to detect the difference between the two systems and provide the best selection recommendation. However, for the $\mathcal{KN}$ procedure, the two systems perform the same. Hence, the procedure randomly selects one of the two systems as the best.

In Experiments 3-5, we consider two systems whose performance metrics have the symmetric distribution $N(\mu = 48, \sigma = 10)$ and the skewed distribution
As expected, the KN procedure selects the $N(\mu = 48, \sigma = 10)$ system as the best with the estimated PCS of 1.00 and a SATO value of $7.8 \times 10^3$ observations. In Experiment 3, the set of lower quantiles $Q = \{0.1, 0.2, 0.3\}$ is used as the basis for comparison. The QBS procedure selects the $N(\mu = 48, \sigma = 10)$ system as the best with the estimated PCS of 1.00 and a SATO value of $50.7 \times 10^3$ observations. In Experiment 4, the set of central quantiles $Q = \{0.5, 0.6, 0.7\}$ is considered as the basis for comparison. In this experiment, the QBS procedure concludes that the two systems form a non-dominated set based on the performance metric of interest with a SATO value of $213.1 \times 10^3$ observations. A secondary performance metric is needed to select the best from the two systems. For Experiment 5, the set of upper quantiles $Q = \{0.7, 0.8, 0.9\}$ is considered as the basis for comparison. The QBS procedure selects the Expo($\mu = 45$) system as the best with the estimated PCS of 1.00 and a SATO value of $799.5 \times 10^3$ observations.

Since the distributions and the quantile sets tested are all symmetric in Experiments 1 and 2, the difference between the SATO values of the QBS procedure is due to randomness. However, when we compare the SATO values of the QBS procedure in Experiments 3-5, we observe an increasing trend. This trend is the result of the changes in the quantile set of interest $Q$. For a fixed sample size, the estimates for the tail quantiles have larger variability than those of the non-tail quantiles, and larger $S_{\max}^{[q]}$ results in a larger $R_i$ when the same $\delta$ value is used in Step 3 of the QBS procedure. The increasing effect of the tail quantiles in the SATO value is more significant when the system tested has a large variability. In Experiments 3-5, the Expo($\mu=45$) system has a significantly larger variability than the $N(\mu = 48, \sigma =10)$ system. This makes the Expo($\mu=45$) system the dominant factor in determining how the SATO values change as the $Q$ set changes. Since the upper quantiles are at the tail of the exponential distribution, the SATO values of the QBS procedure increase from Experiment 3 to 5.

Since the variability of the quantile estimates is significantly affected by the sample size used in Step 2, we ran a few experiments to detect the effect of the sample size on the overall performance of the procedure. We ran Experiment 5 for different sample sizes ranging from the recommended value 100 to 1,000. The sample average of the total number of samples (SATS), the resulting SATO values needed to make a selection decision, and the average computational times are shown in Table 3. As the sample size increases from
Table 3
Effect of the change in the sample size.

<table>
<thead>
<tr>
<th>Sample size (N)</th>
<th>SATS</th>
<th>SATO</th>
<th>Elapsed time (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>7,994.7</td>
<td>799,470</td>
<td>229</td>
</tr>
<tr>
<td>200</td>
<td>4,077.7</td>
<td>815,540</td>
<td>246</td>
</tr>
<tr>
<td>300</td>
<td>2,741.7</td>
<td>822,510</td>
<td>257</td>
</tr>
<tr>
<td>400</td>
<td>2,081.2</td>
<td>832,480</td>
<td>267</td>
</tr>
<tr>
<td>500</td>
<td>1,654.3</td>
<td>827,150</td>
<td>276</td>
</tr>
<tr>
<td>1,000</td>
<td>828.4</td>
<td>828,400</td>
<td>292</td>
</tr>
</tbody>
</table>

the recommended value 100, the SATS needed to make a selection decision decreases significantly. This is a result of the reduction in the sample variances of the quantile estimates. However, if we compare the computational times and the SATO values, which are the multiplication of the sample size and SATS, we do not observe the same decreasing trend but rather an increasing trend. These experimental results show that increasing the sample size from the recommended value 100 does not always improve the overall performance of the procedure.

When we compare the SATO values of the \( \mathcal{KN} \) and \( \mathcal{QBS} \) procedures, we see that the \( \mathcal{QBS} \) procedure requires more observations than the \( \mathcal{KN} \) procedure for any \( Q \) set. This is mainly because each quantile estimate is obtained from a sample of size \( N \). Hence, for the \( \mathcal{QBS} \) procedure, the SATO values are expected to be at least \( N \) times more than that for the \( \mathcal{KN} \) procedure. However, the ratio is generally less than \( N \) because of the reduction effect of the quantile being a robust measure.

Although the \( \mathcal{QBS} \) procedure is computationally less efficient than the \( \mathcal{KN} \) procedure, we recommend the \( \mathcal{QBS} \) 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 risk or the upside risk. In this case, the increased sample size requirements are justified as a trade-off for selecting the correct system.

5 Conclusion

The current state-of-the-art ranking and selection procedures are mean based and define the best system as the one with the largest or smallest expected performance metric. However, when other statistics or attributes of the performance metric are more appropriate for system selection, the mean-based selection procedures are not flexible to accommodate the selection needs and
the decision maker is faced with the potential of using a wrong tool. We propose use of quantiles, as an alternative to the mean, as the basis for comparison in system selection procedures. When systems are compared using the quantiles of a performance metric, different sections (downside, upside, central) of the distribution function of the performance metric may be used as the basis for comparison. This flexibility allows the decision maker to design the experiment suitable for the needs of the problem under consideration. Furthermore, the quantile is a robust measure with respect to outliers while the mean lacks this robustness.

We presented a flexible, two-stage quantile-based ranking and selection procedure ($QBS$) capable of comparing $K$ systems by using a section of the distribution of a metric of interest. We further modified the proposed procedure for the case of comparing systems with a standard system ($QBS^S$) and comparing systems in the presence of a specified threshold value for the performance metric ($QBS^T$). We conducted numerical experiments to highlight the efficacy and efficiency of the proposed quantile-based selection procedure. The results show that the comparison of systems based on multiple quantiles is computationally more difficult than the comparison of systems based on a single mean measure. However, the quantile-based selection approach is the only way to detect the differences between two systems when the comparison of systems is particularly based on the downside or upside risks related to a performance metric. Since the $QBS$ procedure requires comparing competing systems based on multiple quantiles, additional research is needed to develop procedures that are more efficient in terms of the sample size requirements. For example, systems could be compared based on an aggregate of quantiles, and the efficacy and efficiency of such an approach should be investigated.

Acknowledgment

This material was supported by the National Science Foundation under Grant Number EPS-0701892. We would like to thank Dr. Seong-Hee Kim for providing us with the code for the $KN$ procedure. We also would like to thank the anonymous referees for their helpful comments.

References

Bekki JM, Fowler JW, Mackulak GT, Kulahci M. Estimation of cycle-time


Chen HC, Chen CH, Yücesan E. Computing efforts allocation for ordinal optimization and discrete event simulation. IEEE Transactions on Automatic Control 2000; 45; 960-964.


Chick SE, Inoue K. New two-stage and sequential procedures for selecting the best simulated system. Operations Research 2001; 49(5); 732-743.


Jain R, Chlamtac I. The $p^2$ algorithm for dynamic calculation of quantiles and histograms without storing observations. Communications of the ACM


