Proceedings of the 2010 Winter Simulation Conference
B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yücesan, eds.

SELECTING SMALL QUANTILES

Raghu Pasupathy

Industrial and Systems Engineering
Virginia Tech
Blacksburg, VA 24061, USA

Roberto Szechtman

Operations Research Department
Naval Postgraduate School
Monterey, CA 93943, USA

Enver Yücesan

Technology and Operations Management Area
INSEAD
77305 Fontainebleau, FRANCE

ABSTRACT

Ranking and selection (R&S) techniques are statistical methods developed to select the best system, or a subset of systems from among a set of alternative system designs. R&S via simulation is particularly appealing as it combines modeling flexibility of simulation with the efficiency of statistical techniques for effective decision making. The overwhelming majority of the R&S research, however, focuses on the expected performance of competing designs. Alternatively, quantiles, which provide additional information about the distribution of the performance measure of interest, may serve as better risk measures than the usual expected value. In stochastic systems, quantiles indicate the level of system performance that can be delivered with a specified probability. In this paper, we address the problem of ranking and selection based on quantiles. In particular, we formulate the problem and characterize the optimal budget allocation scheme using the large deviations theory.

1 INTRODUCTION

Ranking and selection (R&S) techniques are statistical methods developed to select the best system, or a subset of systems from among a set of alternative system designs. R&S via simulation is particularly appealing as it combines the modeling flexibility of simulation with the efficiency of statistical techniques for effective decision making. Furthermore, simulation experiments also allow for multi-stage sampling as required by some R&S methods. Due to randomness in output data, however, comparing a number of simulated systems requires care. If the precision requirement is high and if the total number of designs in a decision problem is large, then the total simulation cost may be prohibitively high, limiting the utility of simulation for R&S problems. The effective deployment of the simulation budget in R&S is therefore crucial.

The overwhelming majority of the R&S research focuses on the expected performance of competing designs. Alternatively, quantiles, which provide additional information about the distribution of the performance measure of interest, may serve as better risk measures than the usual expected value. In stochastic systems, quantiles indicate the level of system performance that can be delivered with a specified probability. For example, in the financial services industry, Value at Risk (VaR), a quantile of a portfolio's profit or loss over a period of time, is a standard tool to assess the risk of that portfolio. Similarly, in the service industry (e.g., health care or telecommunications), quantiles are used as an indicator for the quality of service. In project management, stochastic activity networks are used to represent complex projects. In such an environment, planners may wish to compute an upper bound on the completion time of the project that would hold with high probability. Similarly, in a newsvendor setting, where a procurement or production quantity must be determined before the market uncertainties are resolved, the optimal quantity, the one that maximizes expected profit, is given by the quantile driven by the demand-supply mismatch costs. Finally, in simulation analysis (or, more generally, in statistics), the critical values for test statistics, confidence intervals, and sequential sampling procedures are expressed as quantiles.

The estimation of quantiles, however, differs considerably from that of expectations. A thorough review of quantile estimation for independent and identically distributed (IID) data is given by Serfling (1980). To improve quantile estimation, authors such as Hsu and Nelson (1990) and Hesterberg and Nelson (1998) apply control variates, while Glynn (1996) uses importance sampling, and Avramidis and Wilson (1998) deploy correlation-induction strategies to obtain variance reduction in simulation-based quantile estimation. Closer to our work, Jin, Fu, and Xiong (2003) provide probabilistic error bounds for simulation quantile estimators using large deviations techniques. Hong (2009) develops an estimator based on infinitesimal perturbation analysis while Liu and Hong (2007) develop kernel estimators for assessing quantile sensitivities. Batur and Choobineh (2009) have recently introduced approaches for quantile-based system selection.

In this paper, we address the problem of identifying the populations that correspond to the m smallest quantiles by sampling independently from d populations. By using the large deviations framework, we characterize the optimal sampling (or budget allocation) scheme that minimizes the probability of incorrect selection given a fixed sampling budget. The remainder of the paper is organized as follows: in the next section, we formally define the problem. We then characterize the budget allocation scheme. As this characterization leads to a difficult, nested optimization problem, we turn our focus to a special case, where we wish to identify those populations whose quantiles exceed a threshold value. We conclude the paper with a number of simple illustrations.

2 PROBLEM DEFINITION

Suppose we have d populations from which we can independently sample. Let X_i be a random variable sampled from population i with distribution function $F_i(\cdot)$. Let q_i be the α_i-quantile of population i; that is

$$q_i = \inf\{k : F_i(k) \geq \alpha_i\}.$$

Throughout we assume that $(F_i(\cdot) : i = 1,\ldots,d)$ and $(q_i : i = 1,\ldots,d)$ are unknown, and that $0 < \alpha_i < 1$. The goal is to determine the populations that correspond to the m smallest quantiles, where the m'th smallest quantile is different than the $m+1$'st smallest quantile. Hence, without loss of generality, we suppose that

$$q_1 \leq q_2 \leq \cdots \leq q_m < q_{m+1} \leq \cdots \leq q_d.$$

The simulation budget is n, $\mathbf{p} = (p_1,\ldots,p_d)$ is the vector of fractional allocations, and $n_i = [np_i]$ is the sample size of population i. Let $(X_{i,k} : k = 1,\ldots,n_i)$ be a collection of IID random samples drawn from F_i, and $X_{i,1:n_i} \leq \cdots \leq X_{i,n_i:n_i}$ the ordered samples of population i. The α_i-quantile estimator is $X_{i,\lceil \alpha_i n_i \rceil:n_i}$, where $\lceil\cdot\rceil$ is the ceiling operator.

To simplify the notation define the sets $\mathscr{A} = \{1,\ldots,m\}$ and $\mathscr{B} = \{m+1,\ldots,d\}$. An incorrect selection (IS) occurs when $\max_{i\in\mathscr{A}} X_{i,\lceil \alpha_i n_i \rceil:n_i} \geq \min_{j\in\mathscr{B}} X_{j,\lceil \alpha_j n_j \rceil:n_j}$. A lower bound for $P(IS)$ is

$$\max_{i\in\mathscr{A},j\in\mathscr{B}} P(X_{i,\lceil \alpha_i n_i \rceil:n_i} \geq X_{j,\lceil \alpha_j n_j \rceil:n_j}) \leq P(IS),$$

and an upper bound for $P(IS)$ is

$$P(IS) = P(\cup_{i\in\mathscr{A},j\in\mathscr{B}} X_{i,\lceil\alpha_i n_i\rceil:n_i} \geq X_{j,\lceil\alpha_j n_j\rceil:n_j}) \leq |\mathscr{A}| \times |\mathscr{B}| \max_{i\in\mathscr{A},j\in\mathscr{B}} P(X_{i,\lceil\alpha_i n_i\rceil:n_i} \geq X_{j,\lceil\alpha_j n_j\rceil:n_j}).$$

Hence, if

$$\frac{1}{n}\log P(X_{i,\lceil\alpha_i n_i\rceil:n_i} \geq X_{j,\lceil\alpha_j n_j\rceil:n_j}) \to -G_{i,j}(p_i,p_j)$$

as $n\to\infty$ for some rate function $G_{i,j}$, we have that

$$\frac{1}{n}\log P(IS) \to -\min_{i\in\mathscr{A},j\in\mathscr{B}} G_{i,j}(p_i,p_j).$$

as $n\to\infty$. The rate functions $G_{i,j}(p_i,p_j)$ depend on the large deviations of $X_{i,\lceil\alpha_i n_i\rceil:n_i}$, which are treated next.

In preparation,

3 CONTINUOUS CASE

If X_i has density $f_i(\cdot)$, then it can be shown (Serfling (1980), pp.85) that $X_{i,\lceil\alpha_i n_i\rceil:n_i}$ has the density

$$f_{i,n_i}(t) = n_i \binom{n_i-1}{\lceil\alpha_i n_i\rceil-1} [F_i(t)]^{\lceil\alpha_i n_i\rceil-1}[1-F_i(t)]^{n_i-\lceil\alpha_i n_i\rceil} f_i(t).$$

For $-\infty<\theta<\infty$ and t in the support of X_i define

$$g_i(t) = \theta t + \alpha_i \log\left(\frac{F_i(t)}{\alpha_i}\right) + (1-\alpha_i)\log\left(\frac{1-F_i(t)}{1-\alpha_i}\right), \tag{1}$$

and

$$\Lambda_{i,n_i}(\theta) = \log E\exp(\theta X_{i,\lceil\alpha_i n_i\rceil:n_i}).$$

When $g_i(\cdot)$ is strictly concave and twice differentiable, it has a unique global maximizer $\tau_i(\theta)$ satisfying $g_i'(t)=0$. Observe that if $0<\alpha_i<1$ then $0<F(\tau_i(\theta))<1$, for otherwise $g(\tau_i(\theta))=-\infty$ and we know that $g_i(q_i)=\theta q_i$ is feasible and greater than $-\infty$. Let

$$\Lambda_i(\theta) = g_i(\tau_i(\theta)). \tag{2}$$

Proposition 1. *If population i has a density $f_i(t)$ with bounded first derivative and the function $g_i(\cdot)$ is twice differentiable with* $\sup g''(t)<0$*, then*

$$\lim_{n_i\to\infty}\frac{1}{n_i}\Lambda_{i,n_i}(n_i\theta) = \Lambda_i(\theta).$$

Proof. From the definition of f_{i,n_i} we have

$$\begin{aligned}\Lambda_{i,n_i}(n_i\theta) & \\ = & \log\left(n_i\binom{n_i-1}{\lceil\alpha_i n_i\rceil-1}\right)+\log\int\exp(n_i\theta t)[F_i(t)]^{\lceil\alpha_i n_i\rceil-1}[1-F_i(t)]^{n_i-\lceil\alpha_i n_i\rceil}f_i(t)\,dt \\ = & \log\left(n_i\binom{n_i-1}{\lceil\alpha_i n_i\rceil-1}\right)+\log\int\exp\left(n_i(g_i(t)+\alpha_i\log(\alpha_i)+(1-\alpha_i)\log(1-\alpha_i))\right)R_i(t)f_i(t)\,dt,\end{aligned} \tag{3}$$

where $R_i(t)=[F_i(t)]^{\lceil\alpha_i n_i\rceil-\alpha_i n_i-1}[1-F_i(t)]^{n_i\alpha_i-\lceil\alpha_i n_i\rceil}$. Since $g_i(t)$ is twice differentiable everywhere, Taylor's Theorem (see Serfling (1980)) for $g_i(t)$ around its global minimum $\tau_i(\theta)$ yields

$$g_i(t)=g_i(\tau_i(\theta))+\frac{(t-\tau_i(\theta))^2}{2}g_i''(\xi), \tag{4}$$

where ξ lies between $\tau_i(\theta)$ and t. Plugging (4) in (3) and dividing through by n_i, we have

$$\begin{aligned}\frac{1}{n_i}\Lambda_{i,n_i}(n_i\theta) = & \frac{1}{n_i}\log\left(n_i\binom{n_i-1}{\lceil\alpha_i n_i\rceil-1}\right)+g_i(\tau_i(\theta))+\alpha_i\log(\alpha_i)+(1-\alpha_i)\log(1-\alpha_i)+ \\ & \frac{1}{n_i}\log\int\exp\left(\frac{n_i}{2}(t-\tau_i(\theta))^2g''(\xi)\right)R_i(t)f_i(t)\,dt.\end{aligned} \tag{5}$$

The binomial term on the right-hand side of (5) becomes

$$n_i\binom{n_i-1}{\lceil\alpha_i n_i\rceil-1}=\frac{n_i!}{(n_i-\lceil\alpha_i n_i\rceil)!(\lceil\alpha_i n_i\rceil-1)!},$$

and Stirling's formula leads to

$$\lim_{n_i\to\infty}\frac{1}{n_i}\log\left(\frac{n_i!}{(n_i-\lceil\alpha_i n_i\rceil)!(\lceil\alpha_i n_i\rceil-1)!}\right)=-(1-\alpha_i)\log(1-\alpha_i)-\alpha_i\log(\alpha_i). \tag{6}$$

Changing variables yields

$$\frac{1}{n_i}\log\int\exp\left(\frac{n_i}{2}(t-\tau_i(\theta))^2g''(\xi)\right)R(t)f_i(t)\,dt=\frac{1}{n_i^{3/2}}\log\int\exp\left(\frac{t^2}{2}g''(\xi)\right)R(\tau_i(\theta)+tn^{-1/2})f_i(\tau_i(\theta)+tn^{-1/2})\,d$$

Expanding $R(\cdot)$ and $f_i(\cdot)$ about $\tau_i(\theta)$ results in

$$R(\tau_i(\theta)+tn^{-1/2})=R(\tau_i(\theta))+\frac{t}{\sqrt{n}}R'(\eta_1) \tag{7}$$

and

$$f_i(\tau_i(\theta)+tn^{-1/2})=f_i(\tau_i(\theta))+\frac{t}{\sqrt{n}}f_i'(\eta_2),$$

for η_1 and η_2 between $\tau_i(\theta)$ and $\tau_i(\theta)+tn^{-1/2}$.

We saw earlier that $0<F(\tau_i(\theta))<1$, which results in $R(\tau_i(\theta))$ and $R'(\tau_i(\theta))$ finite in (7). The two assumptions then lead to

$$
\begin{aligned}
&\frac{1}{n_i^{3/2}}\log\int\exp\left(\frac{t^2}{2}g''(\xi)\right)R(\tau_i(\theta)+tn^{-1/2})f_i(\tau_i(\theta)+tn^{-1/2})\,dt \\
&\quad = \quad \frac{1}{n_i^{3/2}}\log\int\exp\left(\frac{t^2}{2}g''(\xi)\right)\left(R(\tau_i(\theta))f_i(\tau_i(\theta))+O\left(\frac{t}{n_i^{1/2}}\right)\right)dt \\
&\quad \to \quad 0, \qquad (8)
\end{aligned}
$$

as $n_i \to \infty$.

We conclude from (5), (6), and (8) that $\lim_{n\to\infty} n_i^{-1}\Lambda_{i,n_i}(n_i\theta) = \Lambda_i(\theta)$. □

4 DISCRETE CASE

Towards stating an analogous result for the discrete case, let us use a narrower definition of a quantile. Let q_i be the α_i-quantile of population i, meaning that

$$F_i(q_i) = \alpha_i.$$

Suppose X_i is supported on the countable set $\mathscr{L}$. Then (ignoring issues due to non-integral $n_i\alpha_i$), it is seen that $X_{i,\lceil\alpha_i n_i\rceil:n_i}$ has the probability mass function

$$\Pr\{X_{i,\lceil\alpha_i n_i\rceil:n_i} = t\} = \binom{n_i}{\lceil\alpha_i n_i\rceil}([F_i(t)]^{\lceil\alpha_i n_i\rceil} - [F_i(t^-)]^{\lceil\alpha_i n_i\rceil})[1-F_i(t)]^{n_i-\lceil\alpha_i n_i\rceil}, \quad t\in\mathscr{L}$$

where $F_i(t^-) = \Pr\{X_i < t\}$.

Before we state the main result for the discrete context, we note the following simple proposition without proof.

Proposition 2. *Let $\{a_{1,n}\},\{a_{2,n}\},\ldots$ be a finite number of positive-valued sequences with*

$$\lim_{n\to\infty}\frac{1}{n}\log a_{j,n} = a_j.$$

Then,

$$\lim_{n\to\infty}\frac{1}{n}\log\left(\sum_j a_{j,n}\right) = \mathrm{Max}_j\{a_j\}.$$

We are now ready to state the main result in the discrete context.

Proposition 3. *Suppose X_i has finite support $\mathscr{L}$, and satisfies $Pr\{X_i = t\} > 0$ for each $t\in\mathscr{L}$. Furthermore, suppose that the function $g_i(\cdot)$ has a unique maximum at $\tau_i(\theta)$ and that $g_i(t)$ is strictly increasing (decreasing) for $t < \tau_i(\theta)$ (resp., $t > \tau_i(\theta)$). Then*

$$\lim_{n_i\to\infty}\frac{1}{n_i}\Lambda_{i,n_i}(n_i\theta) = \Lambda_i(\theta).$$

Proof. Denote $p_i(t) = \Pr\{X_{i,\lceil \alpha_i n_i \rceil : n_i} = t\}$, $t_m = \max\{t : t \in \mathscr{L}\}$, and $\mathscr{L}' = \mathscr{L} \setminus \{t_m\}$. (Since $p_i(t_m) = 0$, we see that $g_i(t_m) = -\infty$ and hence $\tau_i(\theta) \neq t_m$.) We have

$$\lim_{n_i \to \infty} \frac{1}{n_i} \Lambda_{i,n_i}(n_i \theta) = \lim_{n_i \to \infty} \frac{1}{n_i} \log \left(\sum_{t \in \mathscr{L}'} p_i(t) \exp\{n_i \theta t\} \right). \tag{9}$$

We will show that

$$\lim_{n_i \to \infty} \frac{1}{n_i} \log \left(p_i(t) \exp\{n_i \theta t\} \right) = g_i(t) \quad \forall t \in \mathscr{L}'. \tag{10}$$

The assertion of the theorem then follows from applying Proposition 2 to (9) and (10).

To show (10), we notice that

$$\begin{aligned}
&\frac{1}{n_i} \log \left(p_i(t) \exp\{n_i \theta t\} \right) \\
&\quad = \theta t + \frac{1}{n_i} \log p_i(t) \\
&\quad = \theta t + \frac{1}{n_i} \log \binom{n_i}{\lceil n_i \alpha_i \rceil} + \frac{1}{n_i} \log(F_i(t)^{\lceil n_i \alpha_i \rceil} - F_i(t^-)^{\lceil n_i \alpha_i \rceil}) + \frac{(n_i - \lceil n_i \alpha_i \rceil)}{n_i} \log(1 - F_i(t)) \\
&\quad = \theta t + \frac{1}{n_i} \log \binom{n_i}{\lceil n_i \alpha_i \rceil} + \frac{1}{n_i} \log F_i(t)^{\lceil n_i \alpha \rceil} + \\
&\qquad\qquad \frac{1}{n_i} \log\left(\frac{F_i(t)^{\lceil n_i \alpha_i \rceil} - F_i(t^-)^{\lceil n_i \alpha_i \rceil}}{F_i(t)^{\lceil n_i \alpha_i \rceil}}\right) + \frac{n_i - \lceil n_i \alpha_i \rceil}{n_i} \log(1 - F_i(t)).
\end{aligned} \tag{11}$$

Now, through an application of Stirling's formula we see that the second term appearing on the right-hand side of (11) satisfies

$$\begin{aligned}
\lim_{n_i \to \infty} \frac{1}{n_i} \log \binom{n_i}{\lceil n_i \alpha_i \rceil} &= \lim_{n_i \to \infty} \frac{1}{n_i} \log \left(\frac{n_i!}{(n_i - \lceil \alpha_i n_i \rceil)! \lceil \alpha_i n_i \rceil !} \right) \\
&= -(1 - \alpha_i) \log(1 - \alpha_i) - \alpha_i \log(\alpha_i) = -H(\alpha_i).
\end{aligned} \tag{12}$$

Next, we see that since $\frac{F_i(t)^{\lceil n_i \alpha_i \rceil} - F_i(t^-)^{\lceil n_i \alpha_i \rceil}}{F_i(t)^{\lceil n_i \alpha_i \rceil}}$ is arbitrarily close to 1 for large enough n_i, the fourth term appearing on the right-hand side of (11) satisfies

$$\lim_{n_i \to \infty} \frac{1}{n_i} \frac{F_i(t)^{\lceil n_i \alpha_i \rceil} - F_i(t^-)^{\lceil n_i \alpha_i \rceil}}{F_i(t)^{\lceil n_i \alpha_i \rceil}} = 0. \tag{13}$$

Finally, the third and fifth terms appearing on the right-hand side of (11) satisfy

$$\lim_{n_i \to \infty} \frac{1}{n_i} \log F_i(t)^{\lceil n_i \alpha \rceil} + \lim_{n_i \to \infty} \frac{n_i - \lceil n_i \alpha_i \rceil}{n_i} \log(1 - F_i(t)) = H(\alpha_i) + \alpha_i \log\left(\frac{F_i(t)}{\alpha_i}\right) + (1 - \alpha_i) \log\left(\frac{1 - F_i(t)}{1 - \alpha_i}\right). \tag{14}$$

Using (12), (13), and (14) in (11), we get

$$\begin{aligned}
\lim_{n_i \to \infty} \frac{1}{n_i} \log \left(p_i(t) \exp\{n_i \theta t\} \right) &= \theta t - H(\alpha_i) + H(\alpha_i) + \alpha_i \log\left(\frac{F_i(t)}{\alpha_i}\right) + (1 - \alpha_i) \log\left(\frac{1 - F_i(t)}{1 - \alpha_i}\right) \\
&= g_i(t),
\end{aligned}$$

and thus (10) holds. □

5 QUANTILE SELECTION

Propositions 1 and 3 can be used to obtain an expression for the exponential decay rate of the incorrect selection probability, in terms of the sampling budget allocation. Let $I_k(x) = \sup_\theta \{\theta x - \Lambda_k(\theta)\}$ be the rate function corresponding to population k. In the continuous setting, for x such that $0 < F_k(x) < 1$, Proposition 1 leads to

$$
\begin{aligned}
x &= \Lambda_k'(\theta) \\
&= \tau_k(\theta) + \tau_k'(\theta)\left[\theta + \frac{\alpha_k}{F_k(\tau_k(\theta))} f_k(\tau_k(\theta)) - \frac{1-\alpha_k}{1-F_k(\tau_k(\theta))} f_k(\tau_k(\theta))\right] \\
&= \tau_k(\theta),
\end{aligned}
$$

so that

$$
I_k(x) = \alpha_k \log\left(\frac{\alpha_k}{F_k(x)}\right) + (1-\alpha_k)\log\left(\frac{1-\alpha_k}{1-F_k(x)}\right). \tag{15}
$$

A similar argument for the discrete case shows that Eq. (15) is valid there as well.

Let $Z_n = (X_{i,\lceil \alpha_i n_i \rceil : n_i}, X_{j,\lceil \alpha_j n_j \rceil : n_j})$. Then, as shown in Glynn and Juneja (2004), the rate function of $(Z_n : n \geq 0)$ is given by $p_i I_i(x_i) + p_j I_j(x_j)$, and applying the Gärtner-Ellis Theorem results in

$$
G_{i,j}(p_i, p_j) = \inf_{x_i \geq x_j} \{p_i I_i(x_i) + p_j I_j(x_j)\}.
$$

If $F_i(q_m) < 1, \forall i \in \mathscr{A}$ and $F_j(q_1) > 0, \forall j \in \mathscr{B}$ then the rate functions $G_{i,j}(p_i, p_j)$ are finite for any feasible allocation p_i, p_j. Furthermore, since $I_k(x)$ is strictly decreasing for $x < q_k$ and strictly increasing for $x > q_k$, we must have $G_{i,j}(p_i, p_j) = \inf_x \{p_i I_i(x) + p_j I_j(x)\}$.

An optimal allocation $\mathbf{p}$ maximizes $\min_{i \in \mathscr{A}, j \in \mathscr{B}} G_{i,j}(p_i, p_j)$, which is the same as

$$
\max \zeta
$$

s.t.

$$
\zeta - G_{i,j}(p_i, p_j) \leq 0, \quad \forall i \in \mathscr{A} \text{ and } \forall j \in \mathscr{B}
$$

and

$$
\sum_{i=1}^{d} p_i \leq 1, \quad p_i \geq 0.
$$

The first-order conditions are necessary for optimality (same argument as in Glynn and Juneja (2004)). They are

$$
\sum_{i \in \mathscr{A}} \frac{\partial G_{i,j}(p_i^*, p_j^*)}{\partial p_j} \lambda_{i,j} = \beta \quad \forall i \in \mathscr{A},
$$

$$
\sum_{j \in \mathscr{B}} \frac{\partial G_{i,j}(p_i^*, p_j^*)}{\partial p_i} \lambda_{i,j} = \beta \quad \forall j \in \mathscr{B},
$$

$$\sum_{i\in\mathscr{A},j\in\mathscr{B}} \lambda_{i,j} = 1,$$

and

$$\lambda_{i,j}(\zeta - G_{i,j}(p_i^*, p_j^*)) = 0 \quad \forall i \in \mathscr{A}, \forall j \in \mathscr{B}$$

where $\lambda_{i,j} \geq 0$ for all $i \in \mathscr{A}$, $j \in \mathscr{B}$, and $\beta \geq 0$.

It can be shown that

$$\min_{j\in\mathscr{B}} G_{i,j}(p_i^*, p_j^*) = \zeta^* \quad \forall i \in \mathscr{A},$$

and that

$$\min_{i\in\mathscr{A}} G_{i,j}(p_i^*, p_j^*) = \zeta^* \quad \forall j \in \mathscr{B}.$$

for some $\zeta^* > 0$.

5.1 Crossing a threshold

Getting insights about the optimal allocation appears very difficult because we have a nested optimization problem. It is easier, however, to characterize the allocation that minimizes the probability of crossing a threshold $c \in [q_m, q_{m+1}]$. Let IS_c be the event $(\cup_{i\in\mathscr{A}} X_{i,\lceil\alpha_i n_i\rceil:n_i} > c) \cup (\cup_{j\in\mathscr{B}} X_{j,\lceil\alpha_j n_j\rceil:n_j} < c)$. Then we have

$$P(IS) \leq P(IS_c) \leq \sum_{i\in\mathscr{A}} P(X_{i,\lceil\alpha_i n_i\rceil:n_i} > c) + \sum_{j\in\mathscr{B}} P(X_{j,\lceil\alpha_j n_j\rceil:n_j} < c).$$

Using an argument similar to the one presented in Szechtman and Yücesan (2008), we get

$$\frac{1}{n}\log(\sum_{i\in\mathscr{A}} P(X_{i,\lceil\alpha_i n_i\rceil:n_i} > c) + \sum_{j\in\mathscr{B}} P(X_{j,\lceil\alpha_j n_j\rceil:n_j} < c)) \to -\min\{p_1 I_1(c), \ldots, p_d I_d(c)\}$$

as $n \to \infty$. Following Szechtman and Yücesan (2008), the optimal allocations are

$$p_k = \frac{I_k^{-1}(c)}{\sum_{i\in\mathscr{A}} I_i^{-1}(c) + \sum_{j\in\mathscr{B}} I_j^{-1}(c)},$$

leading to

$$\limsup_{n\to\infty} \frac{1}{n}\log P(IS) \leq -(\sum_{i\in\mathscr{A}} I_i^{-1}(c) + \sum_{j\in\mathscr{B}} I_j^{-1}(c))^{-1}.$$

That is, the optimal threshold is the one that minimizes $\sum_{i\in\mathscr{A}} I_i^{-1}(c) + \sum_{j\in\mathscr{B}} I_j^{-1}(c)$ over $c \in [q_m, q_{m+1}]$.

6 CONCLUDING REMARKS

In this paper, we addressed the problem of identifying the populations that correspond to the m smallest quantiles by sampling independently from d populations. Using a large deviations framework, we characterized the optimal sampling (or budget allocation) scheme that minimizes the probability of incorrect selection given a sampling budget that grows to infinity. In particular, the optimal budget allocation arises as the solution of a 3-layer nested optimization problem. The threshold crossing problem, where we wish to identify those populations whose quantiles exceed a threshold value, leads to more tractable budget allocations.

REFERENCES

Avramidis, A. N., and J. R. Wilson. 1998. Correlation induction techniques for estimating quantiles in simulation experiments. *Operations Research* 46:574–591.

Batur, D., and F. Choobineh. 2009. A quantile-based approach to system selection. *European Journal of Operational Research* 99:574–591.

Glynn, P. W. 1996. Importance sampling for Monte Carlo estimation of quantiles. In *Proceedings of the Second International Workshop on Mathematical Methods in Simulation and Experimental Design*, 180–185. Saint Petersburg, Florida.

Glynn, P. W., and S. Juneja. 2004. A large deviations perspective on ordinal optimization. In *Proceedings of the Winter Simulation Conference*, ed. R. G. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, 192–199. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Hesterberg, T. C., and B. L. Nelson. 1998. Control variates for probability and quantile estimation. *Management Science* 44:1295–1312.

Hong, L. J. 2009. Estimating quantile sensitivities. *Operations Research* 57:118–130.

Hsu, J. C., and B. L. Nelson. 1990. Control variates for quantile estimation. *Management Science* 36:835–851.

Jin, X., M. C. Fu, and X. Xiong. 2003. Probabilistic error bounds for simulation quantile estimators. *Management Science* 49:230–246.

Liu, G., and L. J. Hong. 2007. Kernel estimation for quantile sensitivities. 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, 941–948. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Serfling, R. J. 1980. *Approximation theorems of mathematical statistics*. New York: Wiley.

Szechtman, R., and E. Yücesan. 2008. A new perspective on feasibility determination. In *Proceedings of the Winter Simulation Conference*, ed. R. G. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, 192–199. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

AUTHOR BIOGRAPHIES

RAGHU PASUPATHY is an Assistant Professor in the Industrial and Systems Engineering Department at Virginia Tech. His research interests lie broadly in Monte Carlo methods with a specific focus on simulation optimization and stochastic root finding. He is a member of INFORMS, IIE, and ASA, and serves as an Associate Editor for ACM TOMACS and *INFORMS Journal on Computing*. His web and e-mail addresses are <https://filebox.vt.edu/users/pasupath/pasupath.htm> and <pasupath@vt.edu>.

ROBERTO SZECHTMAN is an Associate Professor in the Operations Research Department at the Naval Postgraduate School. He received his Ph.D. from Stanford University, after which he joined Oracle Corp.(2001-03). His research interests include applied probability and military operations research. He is Associate Editor for *IIE Transactions*, *Naval Research Logistics*, and *Operations Research*. His email address is <rszechtm@nps.edu>.

ENVER YÜCESAN is a Professor of Operations Management at INSEAD in Fontainebleau, France. His research interests include ranking and selection as well as optimization via simulation. He is an Area Editor for *IIE Transactions*. He is also serving as the Program Chair for WSC 2010. His e-mail address is <enver.yucesan@insead.edu>.