Hindawi Publishing Corporation
Advances in Decision Sciences
Volume 2010, Article ID 948359, 12 pages
doi:10.1155/2010/948359

Research Article

A New Approach to Estimate the Critical Constant of Selection Procedures

E. Jack Chen[1] and Min Li[2]

[1] *Business Systems, BASF Corporation, 333 Mount Hope Avenue, Rockaway, NJ 07866-0909, USA*
[2] *College of Business Administration, California State University, Sacramento, 6000 J Street, Sacramento, CA 95819-6088, USA*

Correspondence should be addressed to E. Jack Chen, e.jack.chen@basf.com

Received 12 March 2009; Revised 26 September 2009; Accepted 8 January 2010

Academic Editor: Eric J. Beh

A solution to the ranking and selection problem of determining a subset of size m containing at least c of the v best from k normal distributions has been developed. The best distributions are those having, for example, (i) the smallest means, or (ii) the smallest variances. This paper reviews various applicable algorithms and supplies the operating constants needed to apply these solutions. The constants are computed using a histogram approximation algorithm and Monte Carlo integration.

1. Introduction

Discrete-event simulation has been widely used to compare alternative system designs or operating policies. When evaluating k alternative system designs, we select one or more systems as the best and control the probability that the selected systems really are the best. This goal is achieved by using a class of ranking and selection (R&S) procedures in simulation (see [1] for a detailed description). Chen [2, 3] considered a general version of this problem to select a subset of size m containing at least c of the v best of k normally distributed systems with the lth smallest θ_{i_l} (mean or variance), where $\theta_{i_1} \leq \theta_{i_2} \leq \cdots \leq \theta_{i_k}$ represent proposed sampling solutions. Moreover, in practice, if the difference between θ_{i_v} and $\theta_{i_{v+1}}$ is very small, we might not care if we mistakenly choose system i_{v+1}, whose expected response is $\theta_{i_{v+1}}$. The "practically significant" difference d^* (a positive real number) between a desired and a satisfactory system is called the indifference zone in the statistical literature, and it represents the smallest difference about which we care.

If the absolute difference is $\theta_{i_{v+1}} - \theta_{i_v} < d^*$ or the relative difference is $\theta_{i_{v+1}}/\theta_{i_v} < d^*$, we say that the systems are in the indifference zone for correct selection. Note that the absolute difference and the relative difference are two completely separate problems. On the other

hand, if the absolute difference is $\theta_{i_{v+1}} - \theta_{i_v} \geq d^*$ or the relative difference is $\theta_{i_{v+1}}/\theta_{i_v} \geq d^*$, we say that the systems are in the preference zone for correct selection.

Let P^* denote the required minimal probability of correct selection. The goal is to make a correct selection (CS) with probability at least P^*, where

$$P^* \geq P(c, v, m, k) = \binom{k}{m}^{-1} \sum_{i=c}^{\min(m,v)} \binom{v}{i} \binom{k-v}{m-i}. \tag{1.1}$$

If $P^* < P(c, v, m, k)$, the precision requirement is satisfied by choosing the subset at random. The minimal correct selection probability P^* and the "indifference" amount d^* are specified by the user. If $c = v = m = 1$, the problem is to choose the best system. When $m > c = v = 1$, we are interested in choosing a subset of size m that contains the best. If $c = v = m > 1$, we are interested in choosing the m best systems.

This paper proposes a new approach to estimate the operating constants needed to apply the solutions in [2, 3]. We first review these procedures in Sections 2 and 3. We then describe the histogram approximation algorithm of Chen and Kelton [4] and the Monte Carlo integration technique used to compute tables of operating constants in Section 4. A brief illustration demonstrating how to apply the tables is provided in Section 5.

2. Selection Problems with Respect to Means

Consider k independent normal distributions having means μ_i and variances σ_i^2, $i = 1, 2, \ldots, k$. It is assumed that μ_i and σ_i^2 are unknown. Selection procedures generally sample certain observations from each alternative (at the initial stage) and select the systems having the best sample means as the best systems. The question that arises is whether enough observations have been sampled and if not, the number of additional observations that are needed. Hence, at the second stage of the procedure, the required number of observations is generally estimated based on the available sample variances and sample means.

Extending the work of Dudewicz and Dalal [5] and Mahamunulu [6], Chen [3] proposed a two-stage solution when the parameter of interest θ_i is μ_i. Chen's procedure is as follows. Let X_{ij}, $i = 1, 2, \ldots, k$; $j = 1, 2, \ldots, n_i$ be the jth observation from the ith population. Randomly sample n_0 observations from each of the k populations. Let $\overline{X}_i^{(1)} = (\Sigma_{j=1}^{n_0} X_{ij})/n_0$ be the usual unbiased estimate of μ_i and let $s_i^2(n_0) = \Sigma_{j=1}^{n_0}(X_{ij} - \overline{X}_i^{(1)})/(n_0 - 1)$ be the usual unbiased estimate of σ_i^2. Compute

$$n_i = \max\left(n_0 + 1, \left\lceil \left(\frac{h s_i(n_0)}{d^*}\right)^2 \right\rceil\right), \quad \text{for } i = 1, 2, \ldots, k, \tag{2.1}$$

where h is a constant to be described later and $\lceil z \rceil$ denotes the integer ceiling (round-up) of the real number z. Randomly sample an additional $(n_i - n_0)$ observations from the ith population.

We then compute the second-stage sample means:

$$\overline{X}_i^{(2)} = \frac{1}{n_i - n_0} \sum_{j=n_0+1}^{n_i} X_{ij}, \quad \text{for } i = 1, 2, \ldots, k. \tag{2.2}$$

Define the weights

$$W_{i1} = \frac{n_0}{n_i}\left[1 + \sqrt{1 - \frac{n_i}{n_0}\left(1 - \frac{(n_i - n_0)(d^*)^2}{h^2 s_i^2(n_0)}\right)}\right] \tag{2.3}$$

and $W_{i2} = 1 - W_{i1}$, for $i = 1, 2, \ldots, k$. Compute the weighted sample means

$$\widetilde{X}_i = W_{i1}\overline{X}_i^{(1)} + W_{i2}\overline{X}_i^{(2)}, \quad \text{for } i = 1, 2, \ldots, k \tag{2.4}$$

and select the m systems with the smallest $\widetilde{X}_i$ values. Note that the expression for W_{i1} was chosen to guarantee that $(\widetilde{X}_i - \mu_i)/(d^*/h)$ has a t distribution with $(n_0 - 1)$ degrees of freedom (d.f., see [5]).

The derivation is based on the fact that for $i = 1, 2, \ldots, k$,

$$T_i = \frac{\widetilde{X}_i - \mu_i}{d^*/h} \tag{2.5}$$

has a t distribution with $(n_0 - 1)$ d.f., where h depends on k, m, v, c, n_0, and P^*. Note that T_i's are independent. Furthermore, correct selection occurs if and only if the cth smallest μ_{i_l}'s of systems i_l for $l = 1, 2, \ldots, v$ is less than the $(m - c + 1)$th smallest μ_{i_l}'s of systems i_l for $l = v + 1, v + 2, \ldots, k$.

Let f and F, respectively, denote the probability density function (pdf) and the cumulative distribution function (cdf) of the random variable Y. Hogg and Craig ([7], page 198) show that the pdf of the uth order statistic out of n observations of Y is

$$g_{n,u}\left(y_{[u]}\right) = \beta\left(F\left(y_{[u]}\right); u, n - u + 1\right) f\left(y_{[u]}\right), \tag{2.6}$$

where $\beta(x; a, b) = (\Gamma(a + b)/\Gamma(a)\Gamma(b))x^{a-1}(1 - x)^{b-1}$ is the beta distribution with shape parameters a and b. In our case, f and F are, respectively, the pdf and cdf of the t distribution with $(n_0 - 1)$ d.f. For selection problems with respect to means, the *least favorable configuration*

(LFC) occurs when $\mu_{i_1} = \mu_{i_2} = \cdots = \mu_{i_v}$ and $\mu_{i_v} + d^* = \mu_{i_{v+1}} = \cdots = \mu_{i_k}$ (Mahamunulu [6]). Let $\tilde{X}_{[c]}$ be the cth smallest weighted sample mean from $\tilde{X}_{i_l}$ for $l = 1, 2, \ldots, v$ and $\mu_{[c]}$ be its unknown true mean. Let $\tilde{X}_{[u]}$ be the uth $(u = m - c + 1)$ smallest weighted sample mean from $\tilde{X}_{i_l}$ for $l = v+1, v+2, \ldots, k$ and $\mu_{[u]}$ be its unknown true mean. We can write the probability of correct selection as

$$\begin{aligned} \mathrm{P(CS)} &= \mathrm{P}\left[\tilde{X}_{[c]} < \tilde{X}_{[u]}\right] \\ &= \mathrm{P}\left[\frac{\tilde{X}_{[c]} - \mu_{[c]}}{d^*/h} \leq \frac{\tilde{X}_{[u]} - \mu_{[u]}}{d^*/h} + \frac{\mu_{[u]} - \mu_{[c]}}{d^*/h}\right] \\ &= \mathrm{P}\left[T_{[c]} \leq T_{[u]} + \frac{\mu_{[u]} - \mu_{[c]}}{d^*/h}\right] \\ &\geq \mathrm{P}\left[T_{[c]} \leq T_{[u]} + h\right]. \end{aligned} \tag{2.7}$$

The inequality follows because $\mu_{[u]} - \mu_{[c]} \geq \mu_{i_{v+1}} - \mu_{i_v} \geq d^*$. Furthermore, if $d_a(\mu_{[u]}, \mu_{[c]}) = \mu_{[u]} - \mu_{[c]}$ is used instead of d^* in the above equations, we obtain strict equality.

Note that $T_{[c]} \sim \beta(F(T_{[c]}); c, v - c + 1) f(T_{[c]})$ and $T_{[u]} \sim \beta(F(T_{[u]}); m - c + 1, k - v - m + c) f(T_{[u]})$. Here "$\sim$" denotes "is distributed as." Hence,

$$\mathrm{P(CS)} \geq \int_{-\infty}^{\infty} \int_{-\infty}^{y+h} \beta(F(x); c, v - c + 1) f(x) \beta\big(F(y); m - c + 1, k - v - m + c\big) f(y)\, dx\, dy. \tag{2.8}$$

We equate the right-hand side to P^* to solve for h. The value of h is determined by $\mathrm{P}[T_{[c]} \leq T_{[u]} + h] = P^*$. Let $\tau = T_{[c]} - T_{[u]}$. Then $\mathrm{P}[\tau \leq h] = P^*$. That is, under the LFC, the value of h is the P^* quantile of the distribution of τ.

For example, if we are interested in the probability of correctly selecting a subset of size 5 containing 3 of the first 3 best from 10 alternatives, then $T_{[c]} \sim g_{3,3}(t_{[c]})$ and $T_{[u]} \sim g_{7,3}(t_{[u]})$. Furthermore, if the initial sample size is $n_0 = 20$, then f and F are, respectively, the pdf and cdf of the t-distribution with 19 d.f.

Let $w_{[c][u]}$ denote the one-tailed P^* confidence interval (c.i.) half-width of $(\mu_{[u]} - \mu_{[c]})$. We conclude that the sample sizes allocated by (2.1) achieve $w_{[c][u]} \leq d^*$. That is, the indifference amount d^* in (2.1) corresponds to the upper bound of the desired c.i. half-width $w_{[c][u]}$. Hence, under the LFC $\mu_{[c]} + d^* = \mu_{[u]}$, the sample sizes allocated by (2.1) ensure that

$$\begin{aligned} \mathrm{P(CS)} &= \mathrm{P}\left[\tilde{X}_{[c]} < \tilde{X}_{[u]}\right] \\ &= \mathrm{P}\left[\tilde{X}_{[c]} - \tilde{X}_{[u]} - d^* < \mu_{[c]} - \mu_{[u]}\right] \\ &\geq \mathrm{P}\left[\tilde{X}_{[c]} - \tilde{X}_{[u]} - w_{[c][u]} < \mu_{[c]} - \mu_{[u]}\right] \\ &= P^*. \end{aligned} \tag{2.9}$$

The last equality follows from the definition of the c.i. Note that if $w_{[c][u]} > d^*$, then P(CS) < P^*. It can be shown that

$$w_{[c,u]} = \frac{h}{\sqrt{2}}\sqrt{\frac{s^2_{[c]}(n_0)}{n_{[c]}} + \frac{s^2_{[u]}(n_0)}{n_{[u]}}}. \tag{2.10}$$

Koenig and Law [8] provide some h values for the case that $c = v = 1$ or $c = v = m$. This paper supplies a table of the h values with selected c, v, m, and k; where c, v, and m may be different.

3. Selection Problems with Respect to Variances

Extending the work of Bechhofer and Sobel [9] and Mahamunulu [6], Chen [2] proposed a single-stage solution when the parameter of interest θ_i is σ_i^2. Chen's procedure is as follows. Let $n_i = n_0$ for $i = 1, 2, \ldots, k$. Thus, we use the notation s_i^2 instead of $s_i^2(n_i)$ in the rest of this section. Randomly sample n_0 observations from each of the k populations and compute $s_i^2 = \Sigma_{j=1}^{n_0}(X_{ij} - \overline{X}_i^{(1)})/(n_0 - 1)$. Select the m systems with the smallest s_i^2.

For selection problems with respect to variances, the LFC occurs when $\sigma^2_{i_1} = \sigma^2_{i_2} = \cdots = \sigma^2_{i_v}$ and $\sigma^2_{i_v} d^* = \sigma^2_{i_{v+1}} = \cdots = \sigma^2_{i_k}$ (Mahamunulu [6]). The derivation is based on the fact that $(n_0 - 1)s_i^2/\sigma_i^2$ has a χ^2 distribution with $(n_0 - 1)$ d.f. Let $s^2_{[c]}$ be the cth smallest sample variance from $s^2_{i_l}$ for $l = 1, 2, \ldots, v$ and $\sigma^2_{[c]}$ be its unknown true variance. Let $s^2_{[u]}$ be the uth $(u = m - c + 1)$ smallest sample variance from $s^2_{i_l}$ for $l = v + 1, v + 2, \ldots, k$ and $\sigma^2_{[u]}$ be its unknown true variance. Then

$$\begin{aligned}
\mathrm{P(CS)} &= \mathrm{P}\left[s^2_{[c]} < s^2_{[u]}\right] \\
&= \mathrm{P}\left[(n_0 - 1)\frac{s^2_{[c]}}{\sigma^2_{[c]}} \le (n_0 - 1)\frac{s^2_{[u]}}{\sigma^2_{[u]}}\frac{\sigma^2_{[u]}}{\sigma^2_{[c]}}\right] \\
&= \mathrm{P}\left[X_{[c]} \le X_{[u]}\frac{\sigma^2_{[u]}}{\sigma^2_{[c]}}\right] \\
&\ge \mathrm{P}\left[X_{[c]} \le X_{[u]} d^*\right].
\end{aligned} \tag{3.1}$$

The third equality follows because $X_{[c]} = (n_0 - 1)s^2_{[c]}/\sigma^2_{[c]}$, $X_{[u]} = (n_0 - 1)s^2_{[u]}/\sigma^2_{[u]}$, and

$$1 < d^* \le \frac{\sigma^2_{i_{v+1}}}{\sigma^2_{i_v}} \le \frac{\sigma^2_{[u]}}{\sigma^2_{[c]}}. \tag{3.2}$$

Furthermore, if $d_r(\sigma^2_{[u]}, \sigma^2_{[c]}) = \sigma^2_{[u]}/\sigma^2_{[c]}$ is used instead of d^* in the above equation, we obtain strict equality. Note that under the LFC, $d_r(\sigma^2_{[u]}, \sigma^2_{[c]}) = d^*$.

Let ω and Ω, respectively, denote the pdf and cdf of the χ^2 distribution with (n_0-1) d.f. Then $X_{[c]} \sim \beta(\Omega(X_{[c]}); c, v-c+1)\omega(X_{[c]})$ and $X_{[u]} \sim \beta(\Omega(X_{[u]}); m-c+1, k-v-m+c)\omega(X_{[u]})$. Hence,

$$\begin{aligned}
\mathrm{P(CS)} &\geq \int_0^\infty \int_0^{yd^*} \beta(\Omega(x); c, v-c+1)\omega(x)\beta\big(\Omega(y); m-c+1, k-v-m+c\big)\omega(y)\,dx\,dy \\
&= \frac{v!}{(c-1)!(v-c)!}\,\frac{(k-v)!}{(m-c)!(k-v-m+c-1)!} \\
&\quad \times \int_0^\infty \int_0^{yd^*} [\Omega(x)]^{c-1}[1-\Omega(x)]^{v-c}\omega(x)\big[\Omega(y)\big]^{m-c}\big[1-\Omega(y)\big]^{k-v-m+c-1}\omega(y)\,dx\,dy.
\end{aligned} \tag{3.3}$$

We can compute P(CS) values under the LFC given k, m, v, c, d^*, and n_0. Bechhofer and Sobel [9] provide some P(CS) for the cases that $k \leq 4$. We provide additional P(CS) values for some selected parameters in Section 4.2.

Let $\gamma = X_{[c]}/X_{[u]}$. There exists some $0 \leq p \leq 1$ such that $\mathrm{P}[\gamma \leq d^*] = p$. Hence, under the LFC, the value of d^* is the p quantile of the distribution of the random variable γ. For example, if we are interested in the probability of correctly selecting a subset of size 5 containing all 3 of the first 3 best from 10 alternatives, then $X_{[c]} \sim g_{3,3}(x_{[c]})$ and $X_{[u]} \sim g_{7,3}(x_{[u]})$, with f and F replaced by ω and Ω, respectively.

If users prefer to specify the indifference amount d^* in the absolute form, $d_a(\theta, \theta_0) = \theta - \theta_0$, instead of the relative form, $d_r(\theta, \theta_0) = \theta/\theta_0$, when the parameter of interest is a scale parameter, we can transform the absolute indifference amount into the relative indifference, $d_r(\theta, \theta_0) = 1 + d_a(\theta, \theta_0)/\theta_0$. Since θ_0 is unknown, the estimator $\widehat{\theta}_0$ needs to be used and $d_r(\theta, \theta_0) \approx 1 + d^*/\widehat{\theta}_0$. Moreover, a conservative adjustment can be used. Rank the sample variances such that $s^2_{b_1} < s^2_{b_2} < \cdots < s^2_{b_v} < s^2_{b_{v+1}} < \cdots < s^2_{b_k}$. Let y_q be the q quantile of the χ^2 distribution with $(n_{b_v} - 1)$ d.f., where $0 < q < 1$. We can conservatively set $d_r(\sigma^2_{i_{v+1}}, \sigma^2_{i_v}) \approx 1 + d^* y_q/((n_{b_v} - 1)s^2_{b_v}(n_{b_v}))$ (see [2]). Conversely, if users prefer to specify the indifference amount in the relative form instead of the absolute form when the parameter of interest is the location parameter, we can set $d_a(\theta, \theta_0) \approx (d^* - 1)\widehat{\theta}_0$.

4. Method of Computation

Analytical solutions to multidimensional integration problems in the previous section are difficult to obtain. Below we show our approaches to find h and P(CS).

4.1. Computing the Value of h

Recall that under the LFC the value of h is the P^* quantile of the distribution of τ. Consequently, we can use any quantile-estimation procedures to estimate the P^* quantile of the variable τ given k, m, v, c, and n_0. In this section, we briefly review quantile estimates and the histogram-approximation procedure of Chen and Kelton [4].

Let $X_1, X_2, \ldots, X_n$ be a sequence of i.i.d. (independent and identically distributed) random variables from a continuous cdf $F(x)$ with pdf $f(x)$. Let x_p $(0 < p < 1)$ denote the $100p$th percentile or the p quantile, which has the property that $F(x_p) = \Pr(X \leq x_p) = p$. Thus, $x_p = \inf\{x : F(x) \geq p\}$. If $Y_1, Y_2, \ldots, Y_n$ are the order statistics corresponding to the X_i's from n independent observations (i.e., Y_i is the ith smallest of $X_1, X_2, \ldots, X_n$), then a point estimator for x_p based on the order statistics is the sample p quantile:

$$\widehat{x}_p = y_{\lceil np \rceil}. \tag{4.1}$$

Chen and Kelton [4] control the precision of quantile estimates by ensuring that the p quantile estimator $\widehat{x}_p$ satisfies the following:

$$\mathrm{P}\left[x_p \in \widehat{x}_{p\pm\epsilon}\right] \geq 1 - \alpha_1, \quad \text{or equivalently} \quad \mathrm{P}\left[\left|F(\widehat{x}_p) - p\right| \leq \epsilon\right] \geq 1 - \alpha_1. \tag{4.2}$$

Using this precision requirement (i.e., (4.2)), the required sample size n_p for a fixed-sample-size procedure of estimating the p quantile of an i.i.d. sequence is the minimum n_p that satisfies

$$n_p \geq \frac{z^2_{1-\alpha_1/2} p(1-p)}{\epsilon^2}, \tag{4.3}$$

where $z_{1-\alpha_1/2}$ is the $(1-\alpha_1/2)$ quantile of the standard normal distribution, ϵ is the maximum proportional half-width of the c.i., and $(1-\alpha_1)$ is the confidence level. For example, if the data are independent and we would like to have 95% confidence that the coverage of the 0.9 quantile estimator has no more than $\epsilon = 0.0005$ deviation from the true but unknown quantile, the required sample size is $n_p \geq 1382976$ $(=1.960^2 0.9(1-0.9)/0.0005^2)$. Consequently, we are 97.5% confident that the quantile estimate will cover at least $p - 0.0005$ (for $p \geq 0.9$), with a sample size of 1382976.

The histogram-approximation procedure sets up a series of grid points based on a pilot run. New samples are then stored in the corresponding grids according to their observed value. A histogram is created at the end of the procedure when it has processed the required sample size. The p quantile estimator is obtained by interpolating among grid points. Interested readers can see [4] for the detailed steps of the histogram-approximation procedure.

In the appendix, we show how to generate order statistics random variates without storing and sorting the entire sequence. In order to use this algorithm, we need to be able to perform an inverse transformation of the cdf of the random variable. Unfortunately, the inverse transformation of the cdf of the t-distribution and (2.8) are not available. Nevertheless, numerical methods are available to compute the inverse of the cdf of the t-distribution; see [10]. Hence, the variates $T_{[c]}$ and $T_{[u]}$ can be generated efficiently without sorting a series of t-distributed variables.

Table 1 shows the resulting h values for several chosen k, m, v, c, n_0, and P^*. Four significant digits are retained. Negative values indicate that P^* can be achieved with a sample size of $(n_0 + 1)$ and are set to 0.

Table 1: Values of h for the subset selection procedure.

				$P^* = 0.90$				$P^* = 0.95$			
k	m	v	c	15	20	25	30	15	20	25	30
7	3	1	1	1.715	1.693	1.680	1.672	2.166	2.132	2.112	2.099
7	3	2	1	0.710	0.704	0.701	0.699	1.069	1.060	1.054	1.051
			2	2.495	2.454	2.431	2.416	2.939	2.880	2.846	2.825
7	3	3	1	0.072	0.072	0.071	0.071	0.413	0.410	0.408	0.407
			2	1.511	1.495	1.486	1.480	1.865	1.843	1.830	1.822
			3	3.495	3.414	3.368	3.339	3.998	3.887	3.825	3.786
8	3	1	1	1.853	1.830	1.817	1.808	2.305	2.268	2.247	2.234
8	3	2	1	0.889	0.882	0.878	0.875	1.243	1.232	1.225	1.221
			2	2.630	2.586	2.561	2.545	3.070	3.007	2.972	2.949
8	3	3	1	0.329	0.326	0.324	0.323	0.657	0.652	0.648	0.646
			2	1.686	1.667	1.656	1.649	2.037	2.011	1.996	1.987
			3	3.621	3.533	3.484	3.453	4.122	4.004	3.939	3.897
9	3	1	1	1.968	1.943	1.929	1.920	2.417	2.380	2.358	2.344
9	3	2	1	1.027	1.019	1.013	1.010	1.375	1.362	1.355	1.350
			2	2.740	2.694	2.668	2.651	3.180	3.115	3.078	3.055
9	3	3	1	0.508	0.504	0.501	0.500	0.828	0.821	0.816	0.814
			2	1.818	1.796	1.784	1.776	2.163	2.133	2.116	2.106
			3	3.723	3.630	3.579	3.546	4.221	4.096	4.028	3.984
10	4	1	1	1.740	1.719	1.707	1.699	2.181	2.146	2.126	2.114
10	4	2	1	0.796	0.790	0.787	0.784	1.135	1.126	1.120	1.117
			2	2.388	2.350	2.329	2.315	2.810	2.754	2.723	2.703
10	4	3	1	0.270	0.268	0.267	0.266	0.578	0.574	0.571	0.570
			2	1.453	1.439	1.431	1.426	1.770	1.751	1.741	1.733
			3	2.977	2.920	2.889	2.869	3.3403	3.325	3.282	3.255
10	4	4	1	0	0	0	0	0.150	0.149	0.148	0.148
			2	0.929	0.922	0.917	0.914	1.223	1.212	1.206	1.202
			3	2.064	2.039	2.024	2.015	2.396	2.362	2.343	2.331
			4	3.888	3.788	3.732	3.696	4.381	4.249	4.175	4.129
10	5	1	1	1.453	1.434	1.423	1.417	1.894	1.862	1.844	1.832
10	5	2	1	0.479	0.475	0.474	0.472	0.815	0.808	0.804	0.802
			2	2.041	2.008	1.989	1.977	2.459	2.408	2.379	2.361
10	5	3	1	0	0	0	0	0.221	0.219	0.218	0.218
			2	1.076	1.067	1.062	1.058	1.392	1.378	1.370	1.365
			3	2.506	2.460	2.434	2.418	2.921	2.855	2.818	2.794
10	5	4	1	0	0	0	0	0	0	0	0
			2	0.510	0.506	0.504	0.503	0.797	0.791	0.788	0.785
			3	1.560	1.544	1.535	1.529	1.871	1.849	1.837	1.829
			4	3.027	2.966	2.932	2.910	3.448	3.364	3.318	3.289

4.2. Computing the Probability of Correct Selection P(CS)

Monte Carlo integration can be used to approximately evaluate the integrals. Let hypercube V be the integration volume and hypercube $V' \subseteq V$. Monte Carlo integration picks random uniformly distributed points over some simple domain V, which contains V', checks whether

each point is within V', and estimates the area of V' as the area of V multiplied by the fraction of points falling within V'. Suppose that we pick randomly distributed points $X_1, \ldots, X_n$ in d-dimensional volume V to determine the integral of a function f in this volume:

$$\int f dV \approx V\langle f\rangle \pm V\sqrt{\frac{\langle f^2\rangle - \langle f\rangle^2}{n}}, \tag{4.4}$$

where

$$\langle f\rangle \equiv \frac{1}{n}\sum_{i=1}^{n} g(\mathrm{X}_i), \qquad \left\langle f^2\right\rangle \equiv \frac{1}{n}\sum_{i=1}^{n} g^2(\mathrm{X}_i) \tag{4.5}$$

(see Press et al. [11]). Note that $V\sqrt{(\langle f^2\rangle - \langle f\rangle)/n}$ is a one standard deviation error estimate of the integral and g is a function to be specified depending on the problem at hand.

In our case V is the unit volume and g will be the indicator function of whether a correct selection was made. Let r_i be the index of the ith simulation and

$$I(r_i) = \begin{cases} 1, & \text{correct selection was made in simulation } r_i, \\ 0, & \text{otherwise.} \end{cases} \tag{4.6}$$

If we perform n independent simulation replications and the observed P(CS) is $\widehat{p}$, then

$$\langle f\rangle \equiv \frac{1}{n}\sum_{i=1}^{n} I(r_i) =, \qquad \left\langle f^2\right\rangle \equiv \frac{1}{n}\sum_{i=1}^{n} I^2(r_i) = \widehat{p}. \tag{4.7}$$

Let p denote the true P(CS) with given parameters, that is, $p = \int f dV$. Then

$$p \approx \widehat{p} \pm \sqrt{\frac{\widehat{p} - \widehat{p}^2}{n}}. \tag{4.8}$$

Note that the number of times that the best design is selected from n simulation runs has a binomial distribution $\mathrm{B}(n, p)$, where $n \geq 0$ is the number of trials and $0 \leq p \leq 1$ is the success probability. Furthermore, when n is large, $\mathrm{B}(n, p)$ can be approximated by the normal distribution $\mathrm{N}(np, np(1-p))$ with mean np and variance $np(1-p)$ [7]. Consequently,

$$\mathrm{P}\left[p \geq \widehat{p} - \sqrt{\frac{\widehat{p} - \widehat{p}^2}{n}}\right] \geq 0.84. \tag{4.9}$$

If the target $p = 0.9$ and $n = 1000000$, then $\mathrm{P}[p \geq \widehat{p} - 0.0003] \geq 0.84$.

We perform simulation experiments to estimate the value of the integrals. Table 2 shows the resulting probability of correct selection (with four significant digits) for several chosen k, m, v, c, and n_0.

Table 2: Values of P(CS) when $n_0 = 20$.

				d^*					
k	m	v	c	1.2	1.4	1.6	1.8	2.0	2.2
7	3	1	1	0.6251	0.7743	0.8725	0.9310	0.9638	0.9814
7	3	2	1	0.8732	0.9495	0.9809	0.9931	0.9976	0.9991
			2	0.3090	0.4942	0.6562	0.7791	0.8630	0.9172
7	3	3	1	0.9650	0.9900	0.9971	0.9992	0.9997	0.9999
			2	0.6110	0.7872	0.8924	0.9477	0.9755	0.9885
			3	0.0933	0.2009	0.3336	0.4693	0.5910	0.6924
8	3	1	1	0.5698	0.7289	0.8399	0.9102	0.9510	0.9740
8	3	2	1	0.9134	0.9626	0.9849	0.9943	0.9978	0.9991
			2	0.2515	0.4283	0.5955	0.7303	0.8275	0.8929
8	3	3	1	0.9376	0.9804	0.9941	0.9983	0.9995	0.9998
			2	0.5221	0.7201	0.8501	0.9243	0.9630	0.9822
			3	0.0656	0.1558	0.2773	0.4096	0.5346	0.6430
9	3	1	1	0.5250	0.6897	0.8107	0.8901	0.9389	0.9667
9	3	2	1	0.7844	0.9016	0.9587	0.9837	0.9938	0.9976
			2	0.2105	0.3772	0.5451	0.6877	0.7950	0.8705
9	3	3	1	0.9088	0.9693	0.9903	0.9970	0.9991	0.9997
			2	0.4518	0.6610	0.8102	0.9007	0.9498	0.9752
			3	0.0491	0.1245	0.2347	0.3622	0.4876	0.6012
10	4	1	1	0.6014	0.7593	0.8651	0.9286	0.9636	0.9822
10	4	2	1	0.8497	0.9416	0.9792	0.9931	0.9977	0.9992
			2	0.3052	0.5018	0.6748	0.8026	0.8862	0.9368
10	4	3	1	0.9481	0.9860	0.9965	0.9991	0.9997	0.9999
			2	0.5944	0.7920	0.9051	0.9600	0.9839	0.9937
			3	0.1180	0.2612	0.4318	0.5937	0.7253	0.8221
10	4	4	1	0.9841	0.9968	0.9994	0.9998	0.9999	0.9999
			2	0.7943	0.9220	0.9733	0.9914	0.9972	0.9991
			3	0.3131	0.5372	0.7225	0.8466	0.9195	0.9591
			4	0.0265	0.0815	0.1744	0.2931	0.4197	0.5394
10	5	1	1	0.7003	0.8378	0.9187	0.9616	0.9825	0.9924
10	5	2	1	0.9182	0.9739	0.9925	0.9979	0.9994	0.9998
			2	0.4401	0.6469	0.7989	0.8934	0.9464	0.9740
10	5	3	1	0.9806	0.9960	0.9992	0.9998	0.9999	0.9999
			2	0.7553	0.9006	0.9642	0.9881	0.9962	0.9988
			3	0.2370	0.4389	0.6299	0.7755	0.8711	0.9294
10	5	4	1	0.9963	0.9994	0.9999	0.9999	1.0000	1.0000
			2	0.9136	0.9763	0.9941	0.9986	0.9996	0.9999
			3	0.5328	0.7553	0.8874	0.9526	0.9812	0.9926
			4	0.0986	0.2371	0.4108	0.5786	0.7157	0.8171

5. An Illustration

As a brief illustration of how to use Tables 1 and 2, consider 10 systems with θ_i being the expected performance of the ith system, $i = 1, 2, \ldots, 10$. It is desired to select 4 systems such

that they include at least 2 of the 3 best systems, the systems that have the smallest θ_i's. Suppose that for each system the performance of $n_0 = 20$ sampled observations is measured.

If the performance measure is the mean, the question that arises is whether enough observations have been sampled and if not, the number of additional observations that are needed. If the required minimum probability of correct selection is to be at least $P^* = 0.95$ when the difference between μ_{i_4} and μ_{i_3} is 0.5, then from Table 1, $h = 1.748$. Suppose that the sample variance of system 1 is $s_1^2(n_0) = 3^2$. In this case, the required sample size of system 1 is $n_1 = \max(20 + 1, \lceil (1.748 \times 3/0.5)^2 \rceil) = 110$.

If the performance measure is the variance, the question that arises is what the probability guarantee with the chosen parameters will be. If the specified indifference amount is 1.4, that is, the ratio between $\sigma_{i_4}^2$ and $\sigma_{i_3}^2$ is at least 1.4, then from Table 2 the probability guarantee is approximately 0.79.

Since the binomial distribution $B(n, p)$ can be approximated by the normal distribution $N(np, np(1 - p))$, the algorithms discussed in the paper can also be applied when the underlying processes have a binomial distribution, provided that users agree that the approximation is acceptable. Furthermore, it is known that order statistics quantile estimates are asymptotically normal [12]. Consequently, the algorithms are also applicable when the parameter of interest is a quantile; see, for example, [13].

Appendix

Generating Order Statistics Random Variates

For completeness, we list the algorithms needed to generate order statistics random variates.

(i) Generate $X \sim \gamma(\alpha, 1)$ (see [14]). The prespecified constants are $a = 1/\sqrt{2\alpha - 1}$, $b = \alpha - \ln 4$, $q = \alpha + 1/a$, $\theta = 4.5$, and $d = 1 + \ln \theta$. The steps are as follows.

(1) Generate U_1 and U_2 as independent and identically distributed $U(0, 1)$.
(2) Let $V = a \ln[U_1/(1 - U_1)]$, $Y = \alpha e^V$, $Z = U_1^2 U_2$, and $W = b + qV - Y$.
(3) If $W + d - \theta Z \geq 0$, return $X = Y$. Otherwise, proceed to step 4.
(4) If $W \geq \ln Z$, return $X = Y$. Otherwise, go back to step 1.

(ii) Generate $X \sim \text{beta}(\alpha_1, \alpha_2)$ (see [15]).

(1) Generate $Y_1 \sim \text{gamma}(\alpha_1, 1)$ and $Y_2 \sim \text{gamma}(\alpha_2, 1)$ independent of Y_1.
(2) Return $X = Y_1/(Y_1 + Y_2)$.

(iii) Let Y_i be the ith order statistic from n random variables with cdf F. Generate $X \sim Y_i$ (see [15]).

(1) Generate $V \sim \text{beta}(i, n - i + 1)$.
(2) Return $X = F^{-1}(V)$.

Acknowledgment

The authors thank the anonymous referees for their valuable comments.

References

[1] R. E. Bechhofer, T. J. Santner, and D. M. Goldsman, *Design and Analysis of Experiments for Statistical Selection, Screening and Multiple Comparisons*, Wiley Series in Probability and Statistics: Applied Probability and Statistics, John Wiley & Sons, New York, NY, USA, 1995.

[2] E. J. Chen, "Selecting designs with the smallest variance of normal populations," *Journal of Simulation*, vol. 2, no. 3, pp. 186–194, 2008.

[3] E. J. Chen, "Subset selection procedures," *Journal of Simulation*, vol. 3, pp. 202–210, 2009.

[4] E. J. Chen and W. D. Kelton, "Estimating steady-state distributions via simulation-generated histograms," *Computers & Operations Research*, vol. 35, no. 4, pp. 1003–1016, 2008.

[5] E. J. Dudewicz and S. R. Dalal, "Allocation of observations in ranking and selection with unequal variances," *Sankhyā*, vol. 37, no. 1, pp. 28–78, 1975.

[6] D. M. Mahamunulu, "Some fixed-sample ranking and selection problems," *Annals of Mathematical Statistics*, vol. 38, pp. 1079–1091, 1967.

[7] R. V. Hogg and A. T. Craig, *Introduction to Mathematical Statics*, Prentice Hall, Upper Saddle River, NJ, USA, 5th edition, 1995.

[8] L. W. Koenig and A. M. Law, "A procedure for selecting a subset of size m containing the l best of k independent normal populations," *Communications in Statistics: Simulation and Computation*, vol. B14, pp. 719–734, 1985.

[9] R. E. Bechhofer and M. Sobel, "A single-sample multiple decision procedure for ranking variances of normal populations," *Annals of Mathematical Statistics*, vol. 25, pp. 273–289, 1954.

[10] C. Hastings Jr., *Approximations for Digital Computers*, Princeton University Press, Princeton, NJ, USA, 1955.

[11] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C: The Art of Scientific Computing*, Cambridge University Press, Cambridge, UK, 2nd edition, 1992.

[12] H. A. David, *Order Statistics*, Wiley Series in Probability and Mathematical Statistic, John Wiley & Sons, New York, NY, USA, 2nd edition, 1981.

[13] E. J. Chen, "Some procedures of selecting the best designs with respect to quantile," *Simulation*, vol. 84, no. 6, pp. 275–284, 2008.

[14] R. C. H. Cheng, "The generation of gamma variables with non-integral shape parameter," *Applied Statistics*, vol. 26, pp. 71–75, 1977.

[15] A. M. Law, *Simulation Modeling and Analysis*, McGraw-Hill, New York, NY, USA, 4th edition, 2007.