STATISTICS IN SIMULATION: HOW TO DESIGN FOR SELECTING THE BEST ALTERNATIVE

Edward J. Dudewicz
Department of Statistics
The Ohio State University

ABSTRACT

In many simulation studies the experimenter (the person running the simulation) has under consideration several (two or more) proposed procedures (e.g., for running a real-world system), and is simulating in order to determine which is the best procedure (with regard to certain specified criteria of "goodness"). Such an experimenter does not wish basically to test hypotheses, or construct confidence intervals, or perform regression analyses (though these may be appropriate minor parts of his analysis); he does wish basically to select the best of several procedures, and the major part of his analysis should therefore be directed towards this goal.

It is precisely for this problem that ranking-and-selection procedures were developed. These procedures set sample size (in simulation this means run-length) explicitly so as to guarantee that the probability that "the procedure actually selected by the experimenter is the best procedure" is suitably large.

In this paper we first review the background ideas of ranking-and-selection and contrast them to other approaches to multi-population problems (which, while sometimes appropriate in such areas as social science experimentation, are almost wholly inappropriate for use in statistical design and analysis of simulation experiments). Recommended procedures for several common situations are then outlined in detail. References where further theoretical details may be obtained are provided, along with information on current developments in the area. It is intended that the motivation and technical detail given be sufficient for intelligent application in many common situations (though other situations will still require supplementary consultation).

I. BACKGROUND OF MULTI-POPULATION PROBLEMS

Statistics for many years concerned itself to a large extent with problems in which the basic observations came from one source or population (one-population problems). Two-population problems were well-known (if unsolved, for example the Behrens-Fisher problem), but for the most part it was a one-population world until some time in the 1950's when R. E. Bechhofer, by pioneering work (see reference (1) for the first published account of this work, a major event in statistical thought) in ranking-and-selection, brought the subject to full light of day with a context other than the type described by saying (as in classical ANOVA) "We have k populations, but would like to test the hypothesis that we really only have one." The relevance of the pioneering ranking-and-selection work to statistical design and analysis of simulation experiments was soon recognized by workers in the field. For example, on p. 53 of (3), Conway stated in 1963 that "...the analysis of variance seems a completely inappropriate approach to these problems. It is centered upon the test of the hypothesis that all of the alternatives are equivalent. Yet the alternatives are actually different and it is reasonable to expect some difference in performance, however slight. Thus, the failure to reject the null hypothesis only indicates that the test was not sufficiently powerful to detect the difference - e.g., a longer run would have to be employed. Moreover, even when the investigator rejects the hypothesis, it is highly likely that he is more interested in identifying the best alternative than in simply concluding that the alternatives are not equivalent. Recently proposed ranking[-and-selection] procedures...seem more appropriate to the problem than the conventional analysis of variance techniques...." This recognition has continued to the present day, as is exemplified by the fact that in Kleijnen's (9) treatise on statistical aspects of simulation 77 pages (out of 390 which are non-introductory) are devoted to ranking-and-selection procedures (which are also often called "multiple ranking procedures" by Kleijnen and others). Nevertheless it was true, as pointed out by Conway (3) in 1963, that "...the investigator is still going to have difficulty satisfying the assumptions (normality, common variance, independence) that the statistician will require." However in recent years this difficulty has also largely been removed. While in Section II below we will introduce the ranking-and-selection area with an example using the traditional assumptions (normality, common variance, and independence of observations) for simplicity, in Section III work of recent years allows us to recommend procedures given recently for much more general situations.

II. RANKING-AND-SELECTION

In order to introduce the area of ranking-and-selection, let us talk in terms of a simple explicit problem, that of choosing (i.e., selecting) the job shop precedence rule which yields the highest output on the average. (This particular example is chosen only for ease of reference, and we could as easily talk of selecting the queue discipline which yields the highest output on the average, or the investment strategy which yields the highest return on the average. The reader is encouraged is think of an example pertinent to his field and rephrase the considerations given below in terms of that example.)

To be specific, suppose that it is desired to select that one of 10 job shop precedence rules which has the highest average output per period. If we run the shop with rule one for one period, we will observe some output, say X_{11}. However, in a second period, still using rule one, we will observe a different output, say X_{12}. Similarly we obtain output X_{13} in a third period,...,output X_{1N} in an Nth period. Thus, in each period the output using rule one differs.

However, it is reasonable to assume that it varies about some value, say μ_1, in the sense that if we average the output per period using rule one over many periods the number so obtained will be close to μ_1. To take into account the variability in output, assume that X_{11} obeys a normal probability distribution, has mean value μ_1, and has variance σ^2. Similarly for $X_{12},\ldots,X_{1N}$. Then, if one period's output doesn't affect another's,

$$\bar{X}_1 = \frac{X_{11}+X_{12}+\ldots+X_{1N}}{N}$$

will obey a normal probability distribution with mean value μ_1 and variance σ^2/N, i.e. its variability from μ_1 is decreased and (if N is large) we expect $\bar{X}_1 \approx \mu_1$.

Now, considerations like the above hold for each of the 10 rules. Thus, we may observe the outputs over N periods of each of the 10 rules and obtain average outputs $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_{10}$. Since we expect these to be close to the mean values $\mu_1,\mu_2,\ldots,\mu_{10}$ of the 10 rules, we select the rule yielding the largest average as having the highest output (see Table 1).

TABLE 1

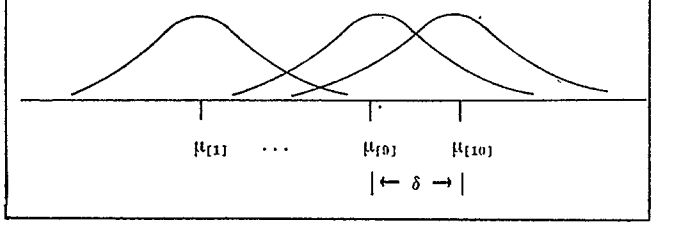
Rule 1		Rule 2		...	Rule 10	
Period 1	X_{11}	Period 1	X_{21}		Period 1	$X_{10,1}$
Period 2	X_{12}	Period 2	X_{22}		Period 2	$X_{10,2}$
.		.			.	
.		.			.	
.		.			.	
Period N	X_{1N}	Period N	X_{2N}		Period N	$X_{10,N}$
	$\bar{X}_1$		$\bar{X}_2$			$\bar{X}_{10}$

X_{ij} = output in period j using rule i. (1)

Select rule yielding $\max(\bar{X}_1,\bar{X}_2,\ldots,\bar{X}_{10})$.

However, it will be hard to distinguish the best rule (i.e., the one with the highest mean output) when the mean outputs of the other rules are very close to the largest one, since although $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_{10}$ will be close to $\mu_1, \mu_2, \ldots, \mu_{10}$ the values $\mu_1, \mu_2, \ldots, \mu_{10}$ are also close to each other. (E.g., although we may be 95% sure that $|\bar{X}_1-\mu_1| \le 0.5$, $|\bar{X}_2-\mu_2| \le 0.5$, ..., $|\bar{X}_{10}-\mu_{10}| \le 0.5$, if in reality $\mu_1=\mu_2=\ldots=\mu_9=233.40$ and $\mu_{10}=233.41$ we may fail to pick the best rule, that with mean output μ_{10}, since $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_9$ as well as $\bar{X}_{10}$ will be close to μ_{10}. This can be remedied by raising the sample size N so that we are, e.g., 95% sure that $|\bar{X}_1-\mu_1| \le 0.01$, $|\bar{X}_2-\mu_2| \le 0.01$, ..., $|\bar{X}_{10}-\mu_{10}| \le 0.01$.) Thus, if N isn't large enough, the probability of selecting the best rule may be unacceptably small; here "large enough" depends on the closeness of $\mu_1, \mu_2, \ldots, \mu_{10}$ (see Illustration 1, where $\mu_{[1]} \le \ldots \le \mu_{[10]}$ denote the mean outputs in numerical order from smallest to largest).

ILLUSTRATION 1



If it is the case, however, that $\mu_{[9]}$ is very close to $\mu_{[10]}$ then we may not care whether we select the rule with mean output $\mu_{[10]}$ or the rule with mean output $\mu_{[9]}$ (which is almost as good). In some cases we may only care about our chances of selecting the best rule when $\mu_{[10]} - \mu_{[9]} \ge 0.1$, in which case we wish to be 90% sure that we make a "Correct Selection" (abbreviated "CS"); i.e. we desire

$$\text{Prob}\{CS\} \geq 0.90 \text{ if } \mu_{[10]} - \mu_{[9]} \geq 0.1.$$

In general, this desired statement is of the form

$$\text{Prob}\{CS\} \geq P^* \text{ if } \delta \equiv \mu_{[10]} - \mu_{[9]} \geq \delta^*, \qquad (2)$$

where δ^* and P^* are pre-set by the experimenter (e.g., $\delta^* = 0.1$ and $P^* = 0.90$). Note that δ^* must be positive (a requirement with $\delta^* \leq 0$ is meaningless since $\mu_{[10]} - \mu_{[9]} \geq 0$ always, by the definition of $\mu_{[10]}$ as the largest of $\mu_1, \mu_2, \ldots, \mu_{10}$) and that $0.10 < P^* < 1$ ($P^* < 1$ since we can never be absolutely certain that the best rule yielded the largest of $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_{10}$ -- there is always a chance, however small, that another rule did; $P^* > 0.10$, since we can be assured of a 10% chance of correct rule selection by picking one of the rules at random). It is clear that the Prob{CS} is minimized when rules other than the best have mean outputs as large as possible. When we "care" (i.e., when $\mu_{[10]} - \mu_{[9]} \geq \delta^*$) this means the Prob{CS} is a minimum when $\mu_{[1]} = \ldots = \mu_{[9]} = \mu_{[10]} - \delta^*$, which is therefore called the least-favorable configuration (LFC). Since available tables (1) allow us to choose N so that Prob{CS} is at least P^* when the LFC $\mu_{[1]} = \ldots = \mu_{[9]} = \mu_{[10]} - \delta^*$ is the case, we can choose N so as to achieve (2).

This example suggests certain conclusions about selection procedures. First, they are precise; that is, the selection approach can give us a rational basis for choosing N (the number of periods to be observed) and tell us (e.g.) how the Prob{CS} varies as we change N, and how large the Prob{CS} is if in fact $\mu_{[1]} = \ldots = \mu_{[9]} = \mu_{[10]} - \delta$ for some value δ other than δ^*. Contrast this to a typical old-style approach: testing the hypothesis that the 10 mean outputs are equal (perhaps by running an Analysis of Variance on an elaborately-designed experiment) and then selecting the rule yielding the largest sample mean as having the largest mean output if the test accepts the hypothesis that they're unequal (while saying "there's no difference" if the test accepts the hypothesis they're equal). Not only does such an approach make little sense because we know they're not equal and should thus always select, but it offers no rational (with regard to the problem for which it is being used), precise grounds for choice of N. Note that this does not mean that one should neglect to use careful design choice if the selection approach is appropriate to his problem (see p. 25 of (1) for further details).

Second, selection procedures are practical in two ways. First, they are applicable to problems often arising in practice, and second, they are feasible because quantities such as necessary sample sizes N have been tabled or can be computed. This may be contrasted to the situation in other branches of statistics where some quantities are almost impossible to compute.

The essential problem formulation of 1954 is thus that we have:

> populations (sources of observations) $\pi_1, \ldots, \pi_k$ ($k \geq 2$) with respective unknown means $\mu_1, \ldots, \mu_k$ for their observations, and whose observations obey a normal probability distribution with a common known variance σ^2 about their respective means; a goal of selecting the population associated with $\mu_{[k]} = \max(\mu_1, \ldots, \mu_k)$; a probability requirement that $\text{Prob}\{CS\} \geq P^*$ ($1/k < P^* < 1$) if $\mu_{[k]} - \mu_{[k-1]} \geq \delta^*$ ($\delta^* > 0$); and a procedure of selecting the population yielding $\bar{X}_{MAX} = \max(\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k)$.

While the above example assumed normality of observations, common variance of output per period for each rule, and independence of observations across periods (all somewhat difficult to justify), the procedures given in Section III weaken these assumptions to an extent sufficient to make this approach feasible in a significant number of simulation studies.

As a numerical example, if we have $k = 10$ rules with $\sigma = 5$ units per period and wish to have probability of correct selection at least $P^* = 0.95$ whenever the average output of the best rule is at least $\delta^* = 2$ units larger than the other outputs, then we will need a sample size of at least

$$N = \frac{h_k^2(P^*)\sigma^2}{(\delta^*)^2} \approx \frac{(3.4182)^2(5)^2}{(2)^2} = 73.03 \qquad (3)$$

periods, and so would take $N = 74$ periods in the simulation. The factor $h_k(P^*)$ needed is obtained from tables originally given by Bechhofer (1) which are reprinted on p. 347 of (5). Note that it is known (4) that a good approximation to the required N is given by

$$N_1 = \frac{-4\sigma^2 \ell n(1-P^*)}{(\delta^*)^2}, \qquad (4)$$

which in our example is $N_1 = 74.89$. Also note that as long as we have a common variance σ^2, if we have correlations within periods (i.e., $X_{11}, X_{21}, \ldots, X_{10,1}$ are correlated) but not across periods (i.e. X_{i1} and X_{i2} are uncorrelated), and if all correlations are positive or zero, then the N of equation (3) is conservative. This was shown in (4) and partially justifies the traditional wisdom that "positive correlation within a block is helpful in simulation".

III. RECOMMENDED SELECTION PROCEDURES

If we are faced with $k \geq 2$ normal populations with unknown means $\mu_1, \ldots, \mu_k$ and a common known variance σ^2, then the sample size N calculated from equation (3) will suffice when the observations are independent (and, in fact, is sufficient and con-

servative even if one has positive correlations within periods).

More often in simulation the rules to be evaluated will have unequal variances $\sigma_1^2,\ldots,\sigma_k^2$ which are also unknown. This problem was solved recently by Dudewicz and Dalal (6), and the solution has been applied in simulations for selecting water resource system alternatives by Vicens and Schaake (11) and in accounting system simulations by Lin (10). This solution is very appropriate if no correlations are present. Since correlations within a population are often present (e.g., due to lack of frequent regeneration points), a heuristic procedure recently developed and studied by Dudewicz and Zaino (8) will be given for this more general problem.

The recommended Procedure $A(\hat{\rho}_i, s_i^2)$ is as follows. Take an initial sample of $N_0 = 30$ observations using each rule. Calculate

$$M_i = \max\left(N_0, \left\lceil \frac{s_i^2 h^2}{(\delta^*)^2} \right\rceil\right) \qquad (5)$$

(which is the number of observations which would be needed if we had all correlations $\rho_i = 0$ (zero correlation), where h depends on k and P* and is given in Table 2 below, extracted from (7)). Calculate

$$\hat{\rho}_i = \frac{\sum_{n=2}^{N_0}(X_{in}-\bar{X}_i)(X_{i,n-1}-\bar{X}_i)}{\sum_{n=1}^{N_0}(X_{in}-\bar{X}_i)^2} \qquad (6)$$

and form the $100(1-\alpha)\%$ confidence interval for ρ_i from

$$(\rho_i-\hat{\rho}_i)^2 \le \frac{N_0-1}{N_0(N_0-3)}(1-\hat{\rho}_i^2)t^2_{N_0-3}(1-\alpha/2) \qquad (7)$$

with $\alpha = .05$. (Here $t_r(q)$ is the 100q percent point of Student's-t distribution with r degrees of freedom.) If this 95% confidence interval contains $\rho_i = 0$, judge the sample size N_0 as being adequate for population i. Otherwise calculate

$$N_{2i} = \left\lceil M_i\left(\frac{1+\hat{\rho}_i}{1-\hat{\rho}_i}\right)\right\rceil \qquad (8)$$

and continue the run until we have N_{2i} observations from π_i. Finally calculate $\bar{X}_1,\ldots,\bar{X}_k$ based on all available observations and select (as being best) that population which produced the largest of $\bar{X}_1,\ldots,\bar{X}_k$.

While Procedure $A(\hat{\rho}_i, s_i^2)$ is heuristic (unlike the procedure of Dudewicz and Dalal for $\rho_i = 0$, which is entirely rigorously derived), studies show it should be sufficient to preclude gross errors due to significant correlations. (Work in progress presently studies properties of Procedure $A(\hat{\rho}_i, s_i^2)$ in further detail by Monte Carlo, compares $A(\hat{\rho}_i, s_i^2)$ with procedures based on the regenerative methods of Iglehart, and develops a corresponding fully rigorous mathematical procedure by utilizing the Heteroscedastic Method recently developed by Dudewicz and Bishop (see (2)).)

TABLE 2

Quantity h Needed in Equation (5)

	P* = .95	P* = .99
k = 2	2.41	3.45
k = 3	2.81	3.81
k = 4	3.03	4.01
k = 5	3.18	4.14
k = 6	3.30	4.25
k = 7	3.39	4.33
k = 8	3.46	4.40
k = 9	3.53	4.46
k = 10	3.58	4.51
k = 15	3.79	4.71
k = 20	3.92	4.84
k = 25	4.03	4.94

BIBLIOGRAPHY

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

2. Bishop, Thomas A. Heteroscedastic ANOVA, MANOVA, and Multiple-Comparisons. Unpublished Ph. D. Dissertation, Department of Statistics, The Ohio State University, Columbus, Ohio, 1976.

3. Conway, R. W. "Some Tactical Problems in Digital Simulation," in Management Science, Vol. 10, No. 1 (October 1963), pp. 47-61.

4. Dudewicz, Edward J. "An Approximation to the Sample Size in Selection Problems," in Annals of Mathematical Statistics, Vol. 40, No. 2 (1969), pp. 492-497.

5. Dudewicz, Edward J. Introduction to Statistics and Probability. Holt, Rinehart and Winston, New York, 1976.

6. Dudewicz, Edward J. and Dalal, Siddhartha R. "Allocation of Observations in Ranking and Selection with Unequal Variances," in Sankhyā, Series B, Vol. 37, Part 1 (1975), pp. 28-78.

7. Dudewicz, E. J., Ramberg, J. S., and Chen, H. J. "New Tables for Multiple Comparisons With a Control (Unknown Variances)," in Biometrische

Zeitschrift, Vol. 17, Part 1 (1975), pp. 13-26.

8. Dudewicz, Edward J. and Zaino, Nicholas A., Jr. "Allowance for Correlation in Setting Simulation Run-length via Ranking-and-Selection Procedures," Technical Report, Department of Statistics, Stanford University, Stanford, California, 1976. To appear in Management Science.

9. Kleijnen, Jack P. C. Statistical Techniques in Simulation, Part II. Marcel Dekker, Inc., New York, 1975.

10. Lin, W. T. "Multiple Objective Budgeting Models: A Simulation," USC Working Paper #12-01-75, Department of Accounting, Graduate School of Business Administration, University of Southern California, Los Angeles, California.

11. Vicéns, Guillermo J. and Schaake, John C., Jr. "Simulation Criteria for Selecting Water Resource System Alternatives," Report No. 154, Ralph M. Parsons Laboratory for Water Resources and Hydrodynamics, Department of Civil Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, September 1972.