Methods for Analyzing Data from Computer Simulation Experiments

THOMAS H. NAYLOR, KENNETH WERTZ,* AND THOMAS H. WONNACOTT†
Duke University, Durham, North Carolina

This paper addresses itself to the problem of analyzing data generated by computer simulations of economic systems. We first turn to a hypothetical firm, whose operation is represented by a single-channel, multistation queueing model. The firm seeks to maximize total expected profit for the coming period by selecting one of five operating plans, where each plan incorporates a certain marketing strategy, an allocation of productive inputs, and a total cost.

The results of the simulated activity under each plan are subjected to an *F*-test, two multiple comparison methods, and a multiple ranking method. We illustrate, compare, and evaluate these techniques. The paper adopts the position that the particular technique of analysis (possibly not any one of the above) chosen by the experimenter should be an expression of his experimental objective: The *F*-test tests the homogeneity of the plans; multiple comparison methods quantify their differences; and multiple ranking methods directly identify the one best plan or best plans.

Introduction

The major impetus behind the use of computer simulation by decision makers and policy makers is the possibility of testing and evaluating alternative decision rules, strategies, and policies before they are put into effect on actual business and economic systems. Complete exploitation of simulation experiments implies a thorough analysis of the data so generated. Yet a preoccupation with model building among many experimenters simulating business and economic systems has unduly diverted attention from experimental design and output analysis.

Presented at the national meeting of The Institute of Management Sciences in Boston, April 5, 1967. This research was supported by National Science Foundation Grant GS-1104. W. Earl Sasser of the Econometric System Simulation Program at Duke University contributed a number of helpful comments. David Patterson wrote the computer programs for the IBM 360/75.

* Graduate student at Carnegie-Mellon University
† Associate Professor of Mathematics at Western Ontario University

The aim of this paper is to meet the problem of analyzing data generated by computer simulation experiments especially for business and economic systems. For this task we have selected three alternative forms of the analysis of variance which are particularly well-suited for comparing outputs of computer models, where those outputs represent the simulated results associated with alternative decision rules and policies. These techniques include the *F*-test, multiple comparison methods, and multiple ranking methods. Of course, other techniques exist, notably sequential sampling methods, spectral analysis [18, 40, 41], and response surface techniques.

With the aid of an example model of a firm we shall illustrate, compare, and evaluate the techniques listed above. However, our analysis of these techniques will not be restricted to their application to the example model. Following a brief exposition of the model, we shall present the results of several simulation runs—that is, the data necessary for evaluating five alternative strategies which are available to the firm. To this end, the output of the simulations shall be subjected to an *F*-test, two different multiple comparison methods, and a multiple ranking procedure. Lastly, we shall discuss the relative advantages and shortcomings of each of these techniques as well as the necessary assumptions underlying their application to the analysis of data generated by computer simulation experiments.

An Example Model

We have chosen a relatively simple model of a firm developed by Chu and Naylor [9]. A complete mathematical description of this model and its corresponding computer flowchart may be found elsewhere [9, 38]. The assumptions underlying the model are summarized below:

(1) The firm possesses an n-stage production process capable of manufacturing a single product. Without exception, each unit of final output of the firm must pass through all n stages in a particular order (see Figure 1).

(2) Each process has its own separate production function which is independent of the production functions of the other $n - 1$ processes.

(3) The rate of output (production rate) of the jth

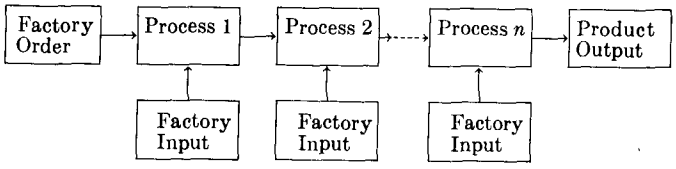

FIG. 1. A flowchart for the model of a firm

TABLE I. Theoretical Values

Theoretical values for expected demand and expected production rates (in units per day) and total cost and approximate expected profit (in dollars) for a computer model of the firm.

Plans	Expected demand rate $E(D)$	Expected production rates				Total cost C	Approximate expected Total profit Π
		Process 1 $E(Q_1)$	Process 2 $E(Q_2)$	Process 3 $E(Q_3)$	Process 4 $E(Q_4)$		
I	3.00	3.33	3.75	4.00	3.50	\$800	\$2918.64
II	3.00	3.50	3.33	6.00	3.50	\$800	\$2918.64
III	3.00	5.00	4.25	6.00	5.00	\$1250	\$2704.00
IV	3.75	5.00	4.25	6.00	5.00	\$1550	\$3285.00
V	3.75	5.00	···	4.50	4.50	\$1720	\$3147.50

process, Q_j ($j = 1, 2, \cdots, n$) during planning period T is a random variable. Its probability density function $f_j(q)$ is completely determined by the level of factor inputs for process j during planning period T—which is to say that by altering its allocation of productive inputs the firm can alter the probability distributions of the Q_j.[1] If $f_j(q)$ is determined then obviously the expected value $E(Q_j)$ and variance Var (Q_j) for process j are also determined.

(4) The number of orders which arrive at the firm per unit time (or the quantity of output which can be sold per unit time at a particular price) is a random variable D with probability density function $f(d)$, expected value $E(D)$, and variance Var (D). Hence the firm cannot ordinarily (Var $(D) \neq 0$) predict with complete certainty the number of units which it can sell at a given price during T. However, it is able to influence $f(d)$, $E(D)$, and Var (D) by adjusting its expenditure strategies for advertising, marketing and promotion.[2]

(5) Once committed to a chosen rate of factor inputs, then the firm accepts all orders which are received throughout planning period T, even though it may not be able to finish production (or possibly begin production) on all such orders in the period.

(6) At the beginning of planning period T, management must make two different types of decisions: (a) those pertaining to levels of expenditure for advertising and marketing, and (b) those pertaining to factor input allocations for the n production processes. Recall that the former completely determine $f(d)$, $E(D)$, and Var (D) over T, while the latter likewise govern $f_j(q)$, $E(Q_j)$, and Var (Q_j) ($j = 1, 2, \cdots, n$).

Having set forth the model, let us now endow the firm with more specific characteristics. The length of the firm's planning horizon is three months ($T = 90$ days) and is assumed to have been determined by the environment in which the firm exists rather than on the basis of statistical considerations. That is, the firm's decision-makers are interested in making plans for the next 90 days—no more, no less.

The *response variable* or dependent variable in our simulation is profit. The *factors* in the experiment are (1) expenditures for productive inputs (labor, raw materials, equipment, etc.) and (2) expenditures for advertising, marketing, and promotion. As previously defined in the description of the model, both of these factors are *quantitative*.[3] That is, in theory there exists a functional relationship between the numerical values of the levels of (1) expenditures for productive inputs and (2) expenditures for advertising, marketing, and promotion and the profitability of the firm. Although the firm's decision makers may choose from among an infinite number of levels for each factor, in practice, due to indivisibilities, institutitional rigidities, incomplete information, and other reasons, the decision makers may restrict their factor level decision to a finite number of levels. In our example model, we assume that the firm has simplified its factor level decision to the point where it is considering only five different operating plans, each one featuring (1) a particular advertising and marketing strategy, (2) a particular allocation of inputs to the various stages of production, which we limit to four in number ($0 < n \leqq 4$), and (3) a total cost, C. (We have already elaborated on points (1) and (2) in the preceding section; total costs appear in Table I). In other words, the firm's controllable quantitative factors have in effect been reduced to five levels of a single qualitative factor, i.e., five operating plans or decision rules.

As a further simplification, we specify $f(d)$ and the $f_j(q)$ to be Poisson distributions (arising from Poisson processes) for all five operating plans. This means that each operating plan consists of the specification (Table I) of a total expenditure C and a set of values for the parameters $E(D)$, $E(Q_1)$, $E(Q_2)$, $E(Q_3)$, and $E(Q_4)$. The purpose of the experiment is to evaluate the profitability of the five plans.

The steady-state properties of a single-channel, multistation queueing model with Poisson arrivals and service rates are available and will serve as a guide to the theoretical values for the expected total profit (Π) associated with each operating plan, which is near steady-state after 90 days. Our model can accommodate without complication any type of probability distribution or empirical distribution for both $f(d)$ and any number of $f_j(q)$, thus extending the reach of investigation into the realm where analytical solutions or approximations are too difficult to obtain.[4]

In any event, with the approximate (steady-state)

[1] See the detailed description of the model in [9, pp. 740–742] or [38, pp. 141–143] for an explanation of how the factor input decision variables are related to the rate of production Q_j and the probability density function $f_j(q)$ for each process.

[2] The way in which advertising, marketing, and promotion expenditures affect demand is outlined in [9, pp. 740, 749–750] and [38, pp. 141, 153].

[3] A factor is quantitative if its levels are numbers which are expected to have a meaningful relationship with the responses. Otherwise a factor is qualitative [7].

[4] See [38, Ch. 4] for a collection of Fortran subroutines for generating stochastic variates on a computer for most of the standard theoretical probability distributions, as well as any empirical distribution.

theoretical values for the expected total profit[5] at hand as a guide (Table I), we may better evaluate the F-test, multiple comparisons, and multiple rankings as techniques for differentiating between the firm's five alternatives when the sample (output) has been generated by a computer simulation experiment. Using a constant price P of \$15 per unit of finished product, Π may be calculated by the following formula [32, 44]:

$$\Pi = \text{expected total revenue} - \text{expected total cost}$$

$$= P\left[E(D)\cdot T - \sum_{j=1}^{n} \frac{E(D)/E(Q_j)}{1 - E(D)/E(Q_j)}\right] - C$$

where

$E(D)\cdot T$ = the expected number of orders which enter the system or expected total demand during the planning period,

$\sum_{j=1}^{n} \frac{E(D)/E(Q_j)}{1 - E(D)/E(Q_j)}$ = the expected number of units remaining in the system either being processed or waiting to be processed at the end of the planning period,

$E(D)\cdot T - \sum_{j=1}^{n} \frac{E(D)/E(Q_j)}{1 - E(D)/E(Q_j)}$ = the expected number of completed orders or expected sales measured in units during the planning period.

P, T, n, and C have already been defined. This formula for expected profit assumes $E(D)/E(Q_j) < 1$, and is merely an approximation, since it assumes that the system has reached a steady-state within 90 days.

The computer simulation which we conducted on this model consisted of 5 runs, one for each operating plan. The parameters used—demand rate (in units per day), production rates (in units per day), and total cost—are tabulated in Table I. Note that plan V consists of 3 processes rather than 4.

Initial Conditions and Sample Size

The *initial conditions* were identical for all replications of each simulation run. The system was assumed to be "empty" at the beginning of each replication for all 5 simulation runs.[6] Activity was simulated for a period of 90 days and total profit was calculated for the period. The simulation was repeated 50 times using the given parameters for plan I. (Repetition was accomplished by altering the starting value of the pseudorandom number generator.[7]) In a similar manner, 90-day runs, each repeated 50 times, were made for strategies II through V. For each sample of 50 observations, the sample mean and standard deviation were calculated and tabulated in Table II.

We now turn our attention to the rationale underlying the sample sizes chosen for this experiment as well as an analysis of some of the effects which these sample sizes have had on the experimental results.

The problem of sample size with computer simulation experiments is indeed complex and has been treated by a number of researchers including Burdick and Naylor [7], Fishman and Kiviat [18], Gafarian and Ancker [19], Geisler [20], and Mechanic and McKay [37]. With computer simulation, sample size may be increased in two different ways: (1) the total length of the simulation run may be increased from, say one month of simulated time to two months of simulated time; (2) runs of a given length may be replicated by using different sets of pseudorandom numbers.

First, consider the length of the simulation run. The length of the firm's planning horizon, 90 days, is assumed to be given. The choice of a suitable planning horizon is assumed to have been made by the firm's policy makers prior to and independent of the decision to use simulation as a mode of analysis.[8] In other words, the length of the simulation run was not determined on the basis of statistical considerations.

Second, we consider the number of replications for each of our five simulation runs. We elected to use the same number of replications for each of the five simulation runs, because inequality of variances over the five operating plans has little effect on inferences about population means in the analysis of variance when the sample size is the same for all five operating plans [45, p. 345].

It is well known that the optimal sample size in analysis of variance depends on the answers one gives to the following three questions: (1) How large a shift in means do you wish to detect? (2) How much variability is present in the population? (3) What size risks are you willing to take? Power function charts for the specification of sample size in analysis of variance are available for determining n, the number of replications per plan for: (1) a given number of plans k; (2) a given population variance σ^2 for each plan; (3) a given level of significance α; and (4) a given power P to detect (5) a specified difference $\Pi_j - \Pi$ between the jth population mean and the grand mean.

Although it may be possible to specify a difference $\Pi_j - \Pi$ which we wish to detect for each plan, a level of significance, and a power for our experiment, meaningful estimates of the unknown parameter σ^2 are not so easy to obtain. Estimates of σ^2 must be based on past experimentation, a pilot study, or familiarity with the system being simulated. Matters are further complicated by the fact that there is reason to believe that the variance is not exactly the same for all five plans in our experiment. However, in order to obtain some idea of what n should

[5] Mathematical expectation—also called long-run average or true profit.

[6] See [9, p. 743] and [38, p. 144].

[7] We used a variation of the "combination method" of generating pseudorandom numbers developed by MacLaren and Marasaglia [36].

[8] The assumption of a given planning horizon is not at all uncommon in the literature in economics. See, for example [38,Ch. 6].

be, we assume that

$$k = 5,$$
$$\sigma = 225,$$
$$\alpha = .05,$$
$$P = .90,$$
$$\Pi_j - \Pi = \begin{cases} 100, & j = 1, 2, \\ 0, & j = 3, \\ -100, & j = 4, 5. \end{cases}$$

Using the power function charts described in [50, p. 104], we obtain a sample size of $n = 20$ for each plan.[9] For $\sigma = 350$, and everything else held constant, we would obtain $n = 50$. To be safe, we have set the sample size at 50 replications per plan.

In the remainder of this paper we shall apply the F-test, multiple comparisons, and multiple rankings to the data generated by the aforementioned experiment. Before turning to these specific data analysis techniques we should inquire about the accuracy of the sample means which appear in Table II. This question can be answered in part by constructing 99% (or any other appropriate level) confidence intervals using the formula [35, p. 175]:

$$\Pi = \bar{X} \pm zs/\sqrt{n}$$

where $\bar{X}$ is the sample mean, s is the sample standard deviation, $n = 50$ is the sample size, z is the percentile of the normal distribution which leaves .5% probability in each tail, and Π is the true profit. (This formula is only an approximation since s is used for σ.) Constructing 99% confidence intervals for each of the five plans we obtain:

Plan I	$2912 < \Pi_1 < 3040$
Plan II	$2918 < \Pi_2 < 3065$
Plan III	$2584 < \Pi_3 < 2766$
Plan IV	$3185 < \Pi_4 < 3345$
Plan V	$3031 < \Pi_5 < 3233.$

The approximate (steady-state) true profits Π_j are, in fact, contained in these confidence intervals. We notice, however, that in plans I and II that the steady-state Π_j come close to missing the confidence interval. This is because these two plans involve the most congested queues (that take the longest to reach the steady state), and therefore have their true Π_j approximated most poorly. A longer planning horizon (greater than 90 days) would have brought us closer to the steady-state and doubtless improved the accuracy of the approximate true profit in Table I.

Analysis of Variance

The analysis of variance is a collection of techniques which are appropriate when the factors affecting the response are *qualitative*. We shall illustrate three different forms of the analysis of variance: the F-test [45], the multiple comparisons of Tukey [45] and Dunnett [14], and the multiple ranking procedure of Bechhofer, Dunnett, and Sobel [4].

[9] Similar power function charts appear in [45, App].

TABLE II. Comparison of Profit

Comparison of approximate expected profit with simulation results for five alternative plans for a computer model of the firm.

Plans	Approximate expected profit (Π)	Sample mean of profit ($\bar{X}$)	Sample standard deviation of profit (s)
I	$2918.64	$2976.40	$175.83
II	$2918.64	$2992.30	$202.20
III	$2704.00	$2675.20	$250.51
IV	$3285.00	$3265.30	$221.81
V	$3147.50	$3131.90	$277.04

Assumptions

All of these procedures were developed on three assumptions: (1) independence of the statistical errors, (2) equality of variance, and (3) normality. The first assumption is satisfied by virtue of the independence of the pseudorandom numbers [36]. We know that the second and third assumptions are not exactly satisfied by our queueing model. The means of the five plans are slightly different (Table II). The variances are doubtless different too, although the sample standard deviations in Table II indicate that the differences are slight. Profit fluctuates according to the number of orders arriving in 90 days, less the number of orders remaining in the system. Both of these numbers are approximately normally distributed (Poisson variates with large means are very nearly normal), so that we can expect the total profit to be approximately normal too. This expectation was borne out by sample histograms and data analysis.

However, all is not lost as a result of the departure from assumptions two and three of the analysis of variance. Certain procedures, such as the F-test, are known to be robust, that is, quite insensitive to departures from assumptions [45, pp. 331–368]. For example, Scheffé argues that, "inequality of variances in the cells of a layout has little effect on inferences about means if the cell numbers are equal, serious effects with unequal cell numbers," [45, p. 345]. It is for this reason that we have chosen equal sample sizes for each of our five simulation runs. With regard to non-normality, Scheffé concludes in chapter 10 that "the effect of violation of the normality assumption is slight on inferences about means but dangerous on inferences about variances." Unfortunately, the robustness properties of multiple comparisons and multiple ranking procedures are not as well known as those of the simple F-test. We can safely hope that our departures from the assumptions of a common variance and normality are small enough not to seriously matter. An interesting extension of this paper might include the use of Monte Carlo sampling techniques to evaluate the robustness of various multiple comparison and multiple ranking procedures. In any event, a methodological paper of this type cannot dwell on a matter that has to be judged in individual cases.

TABLE III. FORMULAS FOR ONE-WAY ANALYSIS OF VARIANCE

Source of variation	Sum of squares	Degrees of freedom	Mean square
Between plans	$SS_{\text{plans}} = n\sum_{j=i}^{k} (\bar{X}_{.j} - \bar{X}_{..})^2$	$k - 1$	$MS_p = SS_{\text{plans}}/k - 1$
Error	$SS_{\text{error}} = \sum_{i=1}^{n} \sum_{j=1}^{k} (X_{ij} - \bar{X}_{.j})^2$	$k(n - 1)$	$MS_e = SS_{\text{error}}/k(n - 1)$
Total	$SS_{\text{total}} = \sum_{i=1}^{n} \sum_{j=1}^{k} (X_{ij} - \bar{X}_{..})^2$	$nk - 1$	

TABLE IV. STATISTICS FOR ONE-WAY ANALYSIS OF VARIANCE

Source of variation	Sum of squares	Degrees of freedom	Mean square
Between plans	9,677,758	4	2,419,440
Error	12,715,825	245	51,901
Total	22,393,583	249	

F-Test

We may wish to test the null hypothesis, H_0, that the expected profits for each of the five operating plans are equal; in symbols:

$$H_0\colon \Pi_1 = \Pi_2 = \cdots = \Pi_5.$$

By employing the F-statistic, the decision rule for accepting or rejecting H_0 becomes

$$\text{If } F \geq F_{\alpha,k-1,k(n-1)}, \qquad \text{reject } H_0;$$

$$\text{otherwise accept } H_0.$$

where F is the appropriate percentile of the F-distribution, α is the significance level, $k = 5$ is the number of operating plans, and $n = 50$ is the number of replications per operating plan. If H_0 is accepted, then one tentatively concludes that the sample differences between plans are attributable to random fluctuations rather than to actual differences in population values (expected profits). On the other hand, if H_0 is rejected, then further analysis, such as multiple comparisons and multiple rankings, is recommended.

Since the pseudorandom numbers generated for the jth operating plan are independent of those for the other four plans, our experiment is analyzed as a single-factor experimental design. Let X_{ij} denote the total profit for the ith replication of plan j. $\bar{X}_{.j}$ is the average profit for plan j over all 50 replications. $\bar{X}_{..}$ is the grand average for all 5 plans over all 50 replications.[10]

Table III contains a summary of the formulas necessary to compute the statistics used in the analysis of a single-factor experiment.

The F-statistic is then computed by the formula:

$$F = MS_p/MS_e.$$

By substituting the results of our experiment for the quantities in Table III, we obtain Table IV. From the data in Table IV, we see that $F = 46.6$, easily exceeding the critical value $F_{.05,4,245} = 2.21$. In this case, F is even much greater than the critical value for $\alpha = .001$. Hence, the data generated by the simulation experiment do not support the null hypothesis that the expected profits are equal for each of the five strategies. One may check the decision to reject H_0 against Tables I and II, which show that the approximate expected profits do indeed vary from plan to plan.

The papers by Box and Andersen [6] and Kruskal and Wallis [33], among others, describe even more robust tests for testing hypotheses about differences in population means.

Multiple Comparisons

Typically, economic policy makers are interested not only in whether alternatives differ but also in *how* they differ. Multiple comparison and multiple ranking procedures often become tools relevant to meeting the latter query, for they have been designed specifically to attack questions of how means of many populations differ.

In contrast with the analysis of variance, multiple comparison methods emphasize the use of confidence intervals rather than the testing of hypotheses. Because our concern in this paper has centered upon differences in population means, it may be tempting at this point to construct a number of 95 % (say) confidence intervals of $\Pi_j - \Pi_J$:

$$(\bar{X}_j - \bar{X}_J) \pm t\cdot\sqrt{2MS_e/n}, \qquad j, J = 1, 2, \cdots, k,$$

by employing the familiar Student's t-statistic. But a problem arises. The intervals developed in this manner are not all *simultaneously* true at the 95 % level; indeed, the confidence level for the aggregate of intervals sinks considerably.

What is needed, therefore, is a way of constructing a *set* of confidence intervals which will all simultaneously be

[10] At this point we shall make our only explicit reference to experimental design. The analysis could have been considerably sharpened by using the same random numbers for all five plans, in each replication. Thus, the numbers X_{11}, X_{12}, X_{13}, X_{14}, X_{15}, for example, in sharing the same random numbers, would share roughly the same statistical fluctuations, so that differences between them could be attributed primarily to the real differences in the underlying Π_j's. All 50 replications would enjoy this property, as would the averages for each of the five plans. This type of experimental design is called "blocking" or "close replication," and requires the use of two-way analysis of variance.

TABLE V. DIFFERENCES OF SAMPLE MEANS $(\bar{X}_{.j} - \bar{X}_{.J})$

j \ J	2	3	4	5
1	−15.9	301.2*	−288.9*	−155.5*
2	···	317.1*	−273.0*	−139.6*
3	···	···	−590.1*	−456.7*
4	···	···	···	−133.4*

TABLE VI. DIFFERENCES OF SAMPLE MEANS $(\bar{X}_{.j} - \bar{X}_{.c})$

j	2	3	4	5
$(\bar{X}_{.j} - \bar{X}_{.c})$	15.9	−301.2*	288.9*	155.5*

true with probability 95%. The May, 1965 issue of *Technometrics* [12, 13, 21, 22, 34] contains a comprehensive review of alternative methods which have been proposed for solving this problem. For illustrative purposes, we shall discuss two of these methods and relate each of them to our simulation experiment: (1) Tukey's method [45, 48, 50], and (2) Dunnett's method [14, 16]. The general form of these methods can be found in the appropriate references. In this paper we shall give the specific form for *one-factor* experiments, although they are equally valid for many-factor experiments.

Tukey's method [45, 48, 50] yields simultaneous confidence intervals (of the type previously described) for the differences between *all* pairs. With 95% probability, *all* of the following confidence intervals for $\Pi_j - \Pi_J$ are true:

$$(\bar{X}_{.j} - \bar{X}_{.J}) \pm q_{k,\nu}\sqrt{MS_e/n}, \qquad j, J = 1, 2, 3, \cdots, k,$$

where $q_{k,\nu}$ is tabulated under the title "Distribution of the Studentized Range Statistic" [50], k is the number of sample means, and ν is the number of degrees of freedom for MS_e, $k(n-1)$ in the case of one-factor experiments. For the actual data generated by our single-factor computer simulation experiment the formula for 95% confidence intervals is given by

$$(\bar{X}_{.j} - \bar{X}_{.J}) \pm q_{5,245}\sqrt{MS_e/n} = (\bar{X}_{.j} - \bar{X}_{.J}) \pm 3.86\sqrt{51{,}901/50} = (\bar{X}_{.j} - \bar{X}_{.J}) \pm 124, \qquad j, J = 1, 2, 3, 4, 5.$$

Table V contains a tabulation of the differences between sample means for all 10 pairs of differences in our experiment. An asterisk (*) indicates that a particular difference exceeds the confidence allowance 124, thus making the difference "statistically significant," if this form of inference is desired. At the same time, and still covered by 95% certainty, we can make more subtle comparisons, technically called linear contrasts. For example, "Does the difference $(\Pi_1 - \Pi_2)$ exceed the difference $(\Pi_2 - \Pi_3)$ and by how much?" "Do the first 3 means exceed the last 2 means on the average, and by how much?" If general linear contrasts are of more interest to the experimenter than the paired comparisons, then Scheffé's method [45] is usually preferred.

Dunnett's [14, 16, 50] method of multiple comparisons compares one specific mean, called the control mean, with all others. In simulations of business and economic systems the control mean is usually the mean associated with the present operating plan, decision rule, or managerial strategy. Dunnett's multiple comparison procedure is summarized as follows: with 95% probability, all of the following confidence intervals for $\Pi_j - \Pi_c$ are true:

$$(\bar{X}_{.j} - \bar{X}_{.c}) \pm d\sqrt{2MS_e/n}, \qquad j = 2, \cdots, k,$$

where Π_c = the control population mean,

$\bar{X}_{.c}$ = the control sample mean,

d = the percentile of Dunnett's t-statistic [14, 16, 50] with degrees of freedom equal to $k(n-1)$ for one-factor experiments.

In our simulation experiment we assume that plan I is the control plan and compare it with all the other plans. The formula for 95% confidence intervals is given by

$$(\bar{X}_{.j} - \bar{X}_{.c}) \pm 2.16\sqrt{\frac{(2)(51,901)}{50}} = (\bar{X}_{.j} - \bar{X}_{.c}) \pm 98.4, \qquad j = 2, 3, 4, 5.$$

Table VI contains a tabulation of the differences between sample means for comparisons between the control mean (plan I) and the means for plans II through V. Again an asterisk (*) indicates that a particular difference exceeds the confidence allowance 98.4 thus making the difference "statistically significant," if this form of inference is of interest.

Multiple Rankings

Frequently, the objective of computer simulation experiments with economic systems is to find the "best," "second best," "third best," etc. plan (or others unlisted). Although multiple comparison methods of estimating the sizes of differences between plans (as measured by population means) are often used as a way of attempting, indirectly, to achieve goals of this type, multiple ranking methods represent a more direct approach to a solution of the ranking problem.

The best estimate of the rank of a set of operating plans is simply the ranking of the sample means associated with the given plans. Because of random error, however, sample rankings may yield incorrect results. With what probability can we say that a ranking of sample means represents the true ranking of the population means? It is basically this question which multiple ranking procedures attempt to answer.

Bechhofer [1] has developed a procedure for selecting a single population and guaranteeing with probability P that the selected population is the "best" provided some other condition on the parameters is satisfied. Like the F-test and multiple comparisons Bechhofer's procedure assumes normality and statistical independence. However, it also assumes *known variances* which may be equal or

unequal. Unfortunately, this procedure is not applicable to our experiment since σ^2 is unknown.

Bechhofer and Sobel [5], Bechhofer, Dunnett, and Sobel [4], Chambers and Jarratt [8], and Huyett and Sobel [31] have considered several variations of multiple ranking procedures. The paper by Bechhofer, Dunnett, and Sobel [4] is of particular interest, since it describes a two-sample multiple decision procedure for ranking the means of normal populations with a common *unknown variance.* Similar problems with various specific probability distributions have been treated by Gupta [22–25], Gupta and Sobel [26–30], Rizvi [43], and Seal [46, 47]. The article by Gupta [22] contains a comprehensive review of multiple ranking procedures.

We now turn to a more detailed description of Bechhofer, Dunnett, and Sobel's two-sample multiple decision procedure for ranking means of normal populations with a common unknown variance [4] and the application of this procedure to our experiment.[11] Using the notation of our experiment, we assume that for a given population (plan) j, X_{ij} is a normally and independently distributed random variable with expected value Π_j and common variance $\sigma_j^2 = \sigma^2$ $(j = 1, 2, \cdots, k)$. We further assume that σ^2 and the Π_j are unknown. Denote the ranked Π_j by

$$\Pi_{[1]} \leqq \Pi_{[2]} \leqq \cdots \leqq \Pi_{[k]}$$

and the differences between the ranked means by

$$\delta_{ij} = \Pi_{[i]} - \Pi_{[j]}, \qquad i, j = 1, 2, \cdots, k.$$

We do not know which population is associated with $\Pi_{[j]}$.

Assume that the experimental goal calls for the selection of the population having the largest expected value. (This is by no means the only goal which may be chosen [2, 4].) Assume also that the experimenter specifies a parameter δ^* which is the smallest value of $\delta_{k,k-1}$ that he is willing to accept. In addition, the experimenter specifies the smallest acceptable value P for the probability of achieving his given goal when $\delta_{k,k-1} \geqq \delta^*$.

Bechhofer, Dunnett, and Sobel's two-sample procedure consists of the following five steps:

1. Take a first sample of N_1 observations from each of the k populations.

2. Calculate the mean square error, MS_e, which is an unbiased estimate of σ^2 having $v = k(n - 1)$ degrees of freedom for $n = N_1$.

3. Take a second sample of $N_2 - N_1$ observations from each of the k populations, $N_2 = \max \{N_1, [2MS_e(h/\delta^*)^2]\}$, where the brackets [ ] denote the smallest integer equal to or greater than the rational number contained within the brackets and h is obtained from Table 3 of Dunnett and Sobel [17] for given values of v and P. If $2MS_e(h/\delta^*)^2 \leqq N_1$, then no second sample is necessary and, therefore, $N_2 = N_1$.

4. For each population calculate the overall sample mean $\bar{X}_j$ where

$$\bar{X}_j = \frac{1}{N_2} \sum_{i=1}^{N_2} X_{ij}, \qquad j = 1, 2, \cdots, k.$$

5. Denote the ranked values of $\bar{X}_j$ by

$$\bar{X}_{[1]} < \bar{X}_{[2]} < \cdots < \bar{X}_{[k]}.$$

Rank the populations according to the ranking of the observed $\bar{X}_j$ and select the population which gives rise to $\bar{X}_{[k]}$ as the population having the largest population mean.

For our experiment, suppose that we want to select the plan having the largest expected profit and to guarantee that the probability of correctly choosing that population will be at least .90 when the difference between the plan with the highest expected profit and the plan with the second highest expected profit is \$100.00. In other words, we are assuming that $P = .90$ and $\delta^* = 100$. We then let $N_1 = n = 50$ and calculate $MS_e = 51{,}901$. For $P = .90$ and $v = k(n - 1) = 245$ we obtain $h = 1.58$ from Table 3 of [17]. Next we determine $\max \{N_1, [2MS_e(h/\delta^*)^2]\} = \max\{50, [2(51{,}901)(1.58/100)^2]\} = \max\{50, 26\} = 50$. Since $26 < 50$ no second sample is required and $N_2 = N_1 = n = 50$. Sample means for $n = 50$ were previously calculated in Table II. On the basis of the ranking of the sample means we would select operating plan IV as the plan with the highest expected profit. If in fact the best operating plan has an expected profit that is \$100.00 larger than the next best, we have at least a probability of 90% of correctly choosing it despite the random statistical fluctuations of sampling. Similar probabilistic statements can be made with this procedure concerning (1) the "best two" plans, (2) the "best three" plans, (3) the "best," "second best," "third best," etc. plans.

Summary

With the aid of a simple example we have attempted to demonstrate the use of three alternative forms of the analysis of variance to analyze data generated by computer simulation experiments with economic systems—the F-test, multiple comparisons, and multiple rankings. The differences in these three types of analysis of variance lie not so much in the assumptions underlying their use, but rather with the types of experimental objectives with which they are most compatible. If one's experimental objective is to test the hypothesis that there is no difference between two or more plans or policies then the F-test is an appropriate analytical tool. If one's objective is to obtain estimates of the sizes of these differences then multiple comparisons are more appropriate. But if the object is to find with a specified degree of certainty the best plan, second best plan, etc., then multiple ranking procedures represent the more direct approach. The reader is cautioned, however, to avoid the indiscriminate use of these techniques without due regard for the assumptions on which they are based. This is particularly true of the latter two techniques.

Finally, we note that although we have limited our

[11] In a forthcoming paper on the use of sequential sampling methods to analyze data from simulation experiments we shall investigate a sequential multiple-decision procedure developed by Bechhofer and Blumenthal [3] for selecting from a group of k normal populations with a common but unknown population variance the one with the largest population mean.

analysis to a single-factor experiment, all of the techniques described in this paper can be extended to experiments with many factors.

RECEIVED OCTOBER 1966; REVISED JUNE, 1967

REFERENCES

1. BECHHOFER, R. E. A single-sample multiple decision procedure for ranking means of normal populations with known variances. *Ann. Math. Stat. 25* (1954), 16-39.
2. ——. A sequential multiple decision procedure for selecting the best one of several normal populations with a common unknown variance, and its use with various experimental designs. *Biometrics 14* (1958), 408-429.
3. ——, AND BLUMENTHAL, S. A sequential multiple-decision procedure for selecting the best one of several normal populations with a common unknown variance, II: Monte Carlo sampling results and new computing formulae. *Biometrics 18* (March, 1962), 52-67.
4. ——, DUNNETT, C. W., AND SOBEL, M. A two-sample multiple decision procedure for ranking means of normal populations with a common unknown variance. *Biometrika 41* (1954), 170-176.
5. ——, AND SOBEL, M. A single-sample multiple decision procedure for ranking variances of normal populations. *Ann. Math. Stat. 25* (1954), 273-289.
6. BOX, G. E. P., AND ANDERSEN, S. L. Permutation theory in the derivation of robust criteria and the study of departures from assumptions. *J. Royal Stat. Soc. Ser. B 17* (1955), 1-34.
7. BURDICK, D. S., AND NAYLOR, T. H. Design of computer simulation experiments for industrial systems. *Comm. ACM 9* (May 1966), 329-339.
8. CHAMBERS, M. L., AND JARRATT, P. Use of double sampling for selecting best population. *Biometrika 51* (1964), 49-64.
9. CHU, K., AND NAYLOR, T. H. A dynamic model of the firm. *Man. Sci. 11* (May 1965), 736-750.
10. CONWAY, R. W. Some tactical problems in digital simulation. *Man. Sci. 10* (Oct. 1963), 47-61.
11. ——, JOHNSON, B. M., AND MAXWELL, W. L. Some problems of digital machine simulation. *Man. Sci. 6* (Oct. 1959), 92-110.
12. COX, D. R. A remark on multiple comparisons. *Technometrics 7* (May 1965), 223-224.
13. DUNCAN, D. B. A Bayesian approach to multiple comparisons. *Technometrics 7* (May 1965), 171-172.
14. DUNNETT, C. W. A multiple comparison procedure for comparing several treatments with a control. *J. Am. Stat. Assoc. 50* (1955), 1096-1121.
15. ——. On selecting the largest of K normal population means. *J. Roy. Stat. Ser. B* (1960), 1-40.
16. ——. New tables for multiple comparisons with a control. *Biometrics 20* (Sept. 1964), 482-491.
17. ——, AND SOBEL, M. A bivariate generalization of Student's T-distribution with tables for certain special cases. *Biometrika 41* (1954), 153.
18. FISHMAN, G. S., AND KIVIAT, P. J. Spectral analysis of time series generated by simulation models. *Man. Sci. 13* (March 1967), 525-557.
19. GAFARIAN, A. V., AND ANCKER, C. J. Mean value estimation from digital computer simulation. *Oper. Res. 14* (Jan. 1966).
20. GEISLER, M. A. The sizes of simulation samples required to compute certain inventory characteristics with stated precision and confidence. *Man. Sci. 10* (Jan. 1964), 261-286.
21. GOODMAN, L. A. On simultaneous confidence intervals for multinomial proportions. *Technometrics 7* (May 1965), 247-254.
22. GUPTA, S. S. On some multiple decision (selection and ranking) rules. *Technometrics 7* (May 1965), 225-246.
23. ——. Probability integrals of the multivariate normal and multivariate T. *Ann. Math. Stat. 34* (1963), 792-828.
24. ——. On a selection and ranking procedure for gamma populations. *Ann. Inst. Math. Tokyo 14* (1963), 199-216.
25. GUPTA, S. S. Selection and ranking procedures and order statistics for the binomial distribution. Proceedings of the International Symposium on Discrete Distributions, August, 1963.
26. ——, AND SOBEL, M. On a statistic which arises in selection and ranking problems. *Ann. Math. Stat. 28* (1957), 957-967.
27. ——, AND SOBEL, M. On selecting a subset which contains all populations better than a standard. *Ann. Math. Stat. 29* (1958), 235-244.
28. ——, AND SOBEL, M. Selecting a subset containing the best of several binomial populations. *Contributions to Probability and Statistics*. Stanford U. Press, Stanford, Calif., 1960.
29. ——, AND SOBEL, M. On selecting a subset containing the population with the smallest variance. *Biometrika 49* (1962a), 495-507.
30. ——, AND SOBEL, M. On the smallest of several correlated F statistics. *Biometrika 49* (1962), 409-523.
31. HUYETT, M. J., AND SOBEL, M. Selecting the best one of several binomial populations. *Bell Sys. Tech. J. 36* (1957), 537-576.
32. JACKSON, R. R. P. Queueing systems with phase type service. *Oper. Res. Quart. 5* (1954).
33. KRUSKAL, W. H., AND WALLIS, W. A. Use of ranks in one-criterion analysis of variance. *J. Am. Stat. Assoc. 47* (1952), 584-621.
34. KURTZ, T. E., LINK, R. F., TUKEY, J. W., AND WALLACE, D. L. Short-cut multiple comparisons for balanced single and double classifications: Pt. 1, Results. *Technometrics 7* (May 1965), 95-162.
35. LINDEGREN, B. W., AND MCELRATH, G. W. *Introduction to Probability and Statistics*. Macmillan, New York, 1966.
36. MACLAREN, M. D., AND MARSAGLIA, G. Uniform random number generators. *J. ACM 12* (1965), 83-89.
37. MECHANIC, H., AND MCKAY, W. Confidence intervals for averages of dependent data in simulations. Tech. Memo 17-7008, IBM-ASDD, March 2, 1964.
38. NAYLOR, T. H., BALINTFY, J. L., BURDICK, D. S., AND CHU, K. *Computer Simulation Techniques*. John Wiley & Sons, New York, 1966.
39. ——, BURDICK, D. S., AND SASSER, W. E. Computer simulation experiments with economic system: The problem of experimental design. *J. Am. Stat. Assoc.* (Dec. 1967).
40. ——, WALLACE, W. H., AND SASSER, W. E. A computer simulation model of the textile industry. *J. Am. Stat. Assoc.* (Dec. 1967).
41. ——, WERTZ, K., AND WONNACOTT, T. Spectral analysis of data generated by simulation experiments with econometric models. *Econometrica* (to appear).
42. ——, WERTZ, K., AND WONNACOTT, T. Some methods for evaluating economic stabilization policies. *Intern. Stat. Inst. Rev.* (Feb. 1967).
43. RIZVI, M. H. Ranking and selection problems of normal populations using the absolute values of their means: Fixed sample size case. Tech. Rep. No. 31, Dept. of Statistics, U. of Minnesota, Minneapolis, Minn.
44. SAATY, T. L. *Elements of Queueing Theory*. McGraw-Hill, New York, 1961.
45. SCHEFFÉ, H. *The Analysis of Variance*. John Wiley & Sons, New York, 1959.
46. SEAL, K. C. On a class of decision procedures for ranking means of normal populations. *Ann. Math. Stat. 26* (1955), 387-398.
47. ——. On ranking parameters of scale in type III populations. *J. Am. Stat. Assoc. 53* (1958), 164-175.
48. TUKEY, J. W. *The Problem of Multiple Comparisons*. Dittoed manuscript, Princeton University, Princeton, N. J., 1953.
49. WALSH, J. E. *Handbook of Nonparametric Statistics, II*. Van Nostrand, Princeton, N. J., 1965.
50. WINER, B. J. *Statistical Principles in Experimental Design*. McGraw-Hill, New York, 1962.