COMMUN. STATIST.-SIMULA. COMPUTA., B6(3), 235-262 (1977)

ROBUSTNESS OF A MULTIPLE RANKING PROCEDURE: A MONTE CARLO EXPERIMENT ILLUSTRATING DESIGN AND ANALYSIS TECHNIQUES

Jack P.C. Kleijnen

School of Business and Economics,
Tilburg, The Netherlands

Key Words & Phrases: simulation; regression analysis; analysis of variance; antithetic variables; sample size; simultaneous inference.

ABSTRACT

This case study demonstrates statistical design and analysis techniques applicable to any Monte Carlo or simulation experiment, namely a 2^{7-3} experimental design, antithetic variates, sample size determination, analysis of variance, regression analysis, and simultaneous inference. The example is a Monte Carlo investigation of the robustness of Bechhofer and Blumenthal's multiple ranking procedure (MRP). The investigation shows that their procedure works often, but not always. Factors that make it break down, are identified.

1. INTRODUCTION

We first define some terminology. Simulation is an experiment with an abstract model over time. Stochastic or Monte Carlo simulation is simulation including random variables which are generated from random numbers. Random variables will be underlined in this paper. The Monte Carlo method is the collection

of techniques that use random numbers to exercise a model. Hence stochastic simulation is a form of Monte Carlo applied to dynamic systems.

Stochastic simulation is widely used in the management sciences for the solution of, e.g., waiting, inventory and scheduling problems. However, their statistical design and analysis is not well developed. These statistical aspects concern questions like (i) How many runs should be made for each system variant in order to select the best variant? (A system variant is defined by the settings of the model's parameters.) (ii) How can the variability of the response be reduced so that less computer time is needed? (iii) Which system variants among the many conceivably interesting variants should be simulated? (iv) How can the response be analyzed such that stochastic fluctuations do not lead to false conclusions? These questions are answered by sound statistical methodology as we shall see in later sections.

In statistical jargon each simulation "run" gives one "observation", e.g. the average waiting time during a simulated day. Each "system variant" corresponds to one statistical "population". We assume that k ($\geq$ 2) system variants are distinguished, i.e. the number of variants is greater than just 2, but is not indefinitely great; compare a scheduling system with k priority rules. The "best" system variant may be defined as the population with the largest mean.

The Bechhofer-Blumenthal MRP determines how many observations should be taken from each of k populations in order to guarantee that the population with the largest mean is selected with prescribed probability at least P^*. The probability of correct selection (CS) should be at least P^* provided it is

worthwhile to select the best population, i.e. provided the best population mean $\mu_{(k)}$ is at least δ^* better than the next best mean $\mu_{(k-1)}$, the so-called indifference zone approach. (Indices in parentheses refer to the ranked populations, i.e., $\mu_{(1)} \leq \ldots \leq \mu_{(k)}$.) Hence

$$P(CS) \geq P^* \text{ if } \mu_{(k)} - \mu_{(k-1)} \geq \delta^* \qquad (1)$$

where P^* and δ^* are specified by the experimenter. Their procedure is sequential, i.e. in each stage one observation is taken from each of the k populations and the stopping statistic z is calculated (this complicated statistic is defined in Bechhofer and Blumenthal (1962, p. 55)); as soon as $z \leq (1-P^*)/P^*$, sampling is terminated and the best population is selected as the one yielding the highest sample mean. The procedure assumes that all observations are normally distributed independent observations with a common variance. We want to determine whether the procedure is robust. The procedure may be insensitive to small deviations from normality and homogeneous variances. Therefore we shall consider both small and large deviations. Moreover the procedure may be robust only for particular values of P^*, δ^* and k.

2. FACTORS IN THE EXPERIMENT

To limit the number of factor-level combinations we consider only two "levels" (or values) for each factor except P^*. The following lists the factors and their levels.

Factor (1) Nonnormality: nontruncated versus truncated distributions. One important characteristic might be the "truncation" of the distribution, a distribution

being truncated if it cannot yield values smaller than a particular constant. This occurs in some models. For instance, waiting times cannot be negative.

Factor (2) Nonnormality: (nearly) symmetric distributions with near-normal tails versus skew distributions with heavy tails. Combining the factors (1) and (2) we decided to use the distributions of Table I. Note that we devised linear combinations of exponential variates $\underline{x}$, each $\underline{x}$ being independent with a common parameter λ.

Factor (3) Heterogeneity of variance: $\sigma_{(k)}$ lower or higher than $\sigma_{(j)}$. The probability of a wrong selection increases if the population means are not far apart or if the observations show large fluctuations. We restrict attention to the least favorable configuration (LFC) of the means, defined as:

$$\mu_{(1)} = \cdots = \mu_{(k-1)} = \mu_{(k)} - \delta^{*}. \qquad (2)$$

TABLE I

Types of Distributions Used to Specify Nonnormality

	Truncated	Nontruncated
Skew and heavy tails	Exponential $\underline{x}$: $\lambda e^{-\lambda}$	Weighted difference of 2 exponentials: $\underline{x}_1 - 0.1\,\underline{x}_2$
(Nearly) symmetric and less heavy tails	Gamma (p=4): $\frac{\lambda^p}{\Gamma(p)} x^{p-1} e^{-\lambda x}$	Sum of differences between exponentials: $(\underline{x}_1-\underline{x}_2) + (\underline{x}_3-\underline{x}_4) =$ $(\underline{x}_1+\underline{x}_3) - (\underline{x}_2+\underline{x}_4)$

We investigate the possible effect of the best population having the highest variance. Its variance is denoted by $\sigma^2_{(k)}$.

Factor (4) Heterogeneity of variance: $\sigma_{(k)}$ and max $\sigma_{(j)}$ approximately equal or much different. The factors (3) and (4) are combined and specified in Table II.

Factor (5) Heterogeneity of variance: all $\sigma_{(j)}$ equal or different. We examine two configurations of the variances of the remaining (k-2) inferior populations. If all $\sigma_{(j)}$ are different we have them increase by the same amount, starting with $\sigma_{(1)} = 0.25\sigma_{(k-1)}$. To fix the absolute values we put max $\sigma_{(j)} = 1$.

Factor (6) The number of populations: large (k=7) or few (k=3). Large k requires much computer time in our experiment since the number of observations per stage is proportional to k (take 1 observation from each population) and the number of stages increases with k (more competing populations).

TABLE II

Standard Deviation of Best Population $\sigma_{(k)}$ Versus Highest Standard Deviation of the Inferior Populations max $\sigma_{(j)}$

	$\sigma_{(k)}/\max \sigma_{(j)} < 1$	$\sigma_{(k)}/\max \sigma_{(j)} > 1$
$\sigma_k/\max \sigma_{(j)} \approx 1$	4/5	5/4
$\sigma_k/\max \sigma_{(j)} \neq 1$	1/5	5

Factor (7) Distance between the best mean and the inferior means: $\delta^* = 0.20 \{\sigma_{(k)} + \sigma_{(k-1)}\}/2$ versus $\delta^* = 1.25 \{\sigma_{(k)} + \sigma_{(k-1)}\}/2$. We study only the LFC of the means, defined in eq. (2). Whether δ^* is small depends on the spread of the populations. We relate δ^* to the average standard deviation $\{\sigma_{(k)} + \sigma_{(k-1)}\}/2$.

Factor (8) The guaranteed probability P^*: 22 P^* levels. As highest value we take $P^* = 0.99$; higher values would require excessive computer time. Any level of $P^* < 0.99$ can be tested at the same time since it requires only part of the observations that must be generated anyhow for $P^* = 0.99$. The P^* levels studied are: 0.35(0.05)0.85, 0.89(0.01)0.99. The corresponding 22 responses are dependent since they partly use the same observations.

The factors 3, 4, 5 and 6 fix the exponential parameters $\lambda_i (i = 1,\ldots,k)$. These λ_i may yield means conflicting with the LFC (see factor 7). Therefore we add suitable constants, say a_i, to the observations from population i. These a_i do not influence the variance, skewness and tails.

3. THE EXPERIMENTAL DESIGN

Even if we consider 7 factors with only two levels we have as many as $2^7 = 128$ combinations. We might assume that the 7 factors have no important interactions and estimate the 7 main (first order) effects from a 2^{7-4} design, i.e., from only 8 combinations. However, we prefer a 2^{7-3} design. As we shall see its 16 combinations yield both estimators of the main effects not biased by possible two-factor interactions, and estimators of the sums of particular two-factor

interactions. The latter estimators enable us to test whether two-factor interactions are important (assuming interactions among 3 or more factors are negligible). Before we proceed we note that the reader not familiar with experimental design may consult Box and Hunter (1961). Alternatively he can proceed directly to the text below eq. (4), observing that $\vec{1}2$ corresponds with the interaction between the factors 1 and 2, etc. Vectors and matrices are indicated by $\rightarrow$.

Several 2^{7-3} designs are possible, each having a different alias pattern, i.e. estimators of main effects and interactions are confounded (and hence possibly biased) in different ways. In a 2^{7-3} design all two-factor interactions involving only 4 different factors can be kept unconfounded. We conjectured that the first 4 factors listed in section 2 might be the most important ones, and therefore their two-factor interactions were kept unconfounded with each other. This alias structure is realized by the generators:

$$\vec{5} = 1\vec{2}3 \quad \vec{6} = 1\vec{2}4 \quad \vec{7} = 2\vec{3}4 \tag{3}$$

as they yield two-factor interactions confounded as follows:

$$\begin{aligned} &1\vec{2} + 3\vec{5} + 4\vec{6},\ 1\vec{3} + \vec{2}5 + \vec{6}7,\ \vec{1}4 + \vec{2}6 + \vec{5}7, \\ &2\vec{3} + 1\vec{5} + \vec{4}7,\ \vec{2}4 + \vec{1}6 + \vec{3}7,\ 3\vec{4} + \vec{2}7 + \vec{5}6, \\ &1\vec{7} + \vec{4}5 + \vec{3}6 \end{aligned} \tag{4}$$

If the factors 5, 6 and 7 also have two-factor interactions then eq. (4) shows that no unbiased estimators of the individual interactions are possible.

The generators in eq. (3) yield the design shown in Table III. (The last column will be used in section 7.) We associate the + and - levels in Table III with the levels of the factors in section 2 randomly. In section 2 we first mentioned the resulting - level, then the + level, e.g., the - level of factor 6 is k = 7. It is simple to determine the design in the original factors (not reproduced here), e.g., in combination 16 we sample from $(\underline{x}_1-\underline{x}_2) + (\underline{x}_3-\underline{x}_4)$, k=7, $\sigma_{(k)} = 4/5$, $\sigma_j = 1$ (j=1,...,6), $\delta^* = 0.20\ \{\sigma_{(k)} + \sigma_{(k-1)}\}/2$.

4. THE NUMBER OF REPLICATIONS

In the Monte Carlo experiment the computer generates stochastic variables. To these variables we apply the MRP. The MRP determines when to stop sampling and selects the population with the largest sample mean. We actually know, which population has the highest population mean. So we score 0 if the MRP gives a wrong selection, and 1 if it yields a CS. Assuming n independent replications these scores, say $\underline{v}$, are

TABLE III

The 2^{7-3} Design Generated by Eq. (3)

Factor	Combination																Variable
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
1	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	x_2
2	+	+	-	-	+	+	-	-	+	+	-	-	+	+	-	-	x_3
3	+	+	+	+	-	-	-	-	+	+	+	+	-	-	-	-	x_4
4	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	x_5
5 (=123)	+	-	-	+	-	+	+	-	+	-	-	+	-	+	+	-	x_6
6 (=124)	+	-	-	+	+	-	-	+	-	+	+	-	-	+	+	-	x_7
7 (=234)	+	+	-	-	-	-	+	+	-	-	+	+	+	+	-	-	x_8

binomial variates with mean p and variance $p(1-p)/n$. We want to determine n such that the α and β errors are controlled. Our null-hypothesis is that the MRP does work, or $p \geq P^*$, and the alternative hypothesis is $p = p_1 < P^*$. Using basic statistics, we derived in Kleijnen (1975, p. 699) that both error types are controlled when:

$$n = (1+c)^2 (z_\beta)^2 p_1 (1-p_1)/(P^*-p_1)^2 , \qquad (5)$$

where

$$c = \{(z_\alpha)^2 P^* (1-P^*)/(z_\beta)^2 p_1 (1-p_1)\}^{\frac{1}{2}} \qquad (6)$$

and z_α and z_β are the upper α and β points of the standard normal variable z, which is used to approximate the averaged binomial variable. We choose $\alpha = 0.01$, $\beta = 0.10$ and the combinations of P^* and p_1 shown in Table IV. The number n determines the firmness of the conclusions of our experiment. We decided to replicate each factor combination 800 times. In the next section we shall see that the replications per combination are not independent, but are negatively correlated. These correlations decrease the variance so that the α and β errors are further reduced.

TABLE IV

Number of Replications n

P^*(x100)	50	60	70	75	80	85	90	95	97.5	99	99.5	99.9
p_1(x100)	42	55	65	70	75	80	85	92.5	96.0	97.5	98.5	99.5
n	499	1251	1114	1007	873	714	528	1128	1658	816	1008	1651

5. VARIANCE REDUCTION TECHNIQUES: ANTITHETICS

The reliability of a Monte Carlo experiment can be increased through variance reduction techniques. Kleijnen (1975) discusses a number of such techniques, but unfortunately - except for the following two techniques - most of them are quite complicated. Common random numbers imply that corresponding replications in each factor combination use the same sequence of random numbers. Antithetic variates imply that replication 1 (of the 800 replications) uses the random numbers r_1, r_2,... and replication 2 uses $(1-r_1)$, $(1-r_2)$ .... We expect antithetics to create negative correlation, i.e. the score $\underline{v}=0$ for replication 1 tends to be accompanied by $\underline{v} = 1$ for replication 2. (The reader may imagine a replication with all random numbers generated for the best population, being 0.) Such negative correlation was indeed found in the experiment. This correlation reduces the variance of the response averaged over all 800 replications for a specific factor combination. We decided not to use common random numbers since they complicate the statistical analysis of the experiment as they make the responses of the 16 combinations dependent. Antithetics give no such problems since we can take the average response, say $\tilde{\underline{v}}$, e.g. $\tilde{\underline{v}}_1 = (\underline{v}_1 + \underline{v}_2)/2$, $\tilde{\underline{v}}_2 = (\underline{v}_3 + \underline{v}_4)/2$, etc.; also see eq. (7) below.

For each combination i and P^*-level ℓ the output consists of the estimated fraction of correct selections:

$$\underline{y}_{i\ell} = \sum_{r=1}^{800} \underline{v}_{i\ell r}/800 = \sum_{r'=1}^{400} \tilde{\underline{v}}_{i\ell r'}/400 ,$$

$$i = 1,\ldots,16, \quad \ell = 1,\ldots,22 , \qquad (7)$$

and the unbiased estimator of its standard deviation

$$\underline{s}_{i\ell}=\{\hat{\text{var}}(\tilde{\underline{v}}_{i\ell r'})/400\}^{\frac{1}{2}}=\{\sum_{r'=1}^{400}(\underline{v}_{i\ell r'}-\underline{y}_{i\ell})^2/(399\times400)\}^{\frac{1}{2}} \quad , \quad (8)$$

where we use the symbol y instead of p, since y is customary in regression analysis applied later on.

We used the multiplicative random number generator with multiplier 7^5 and modulo $2^{31}-1$ which was extensively tested by Lewis et al. (1969). These authors also give an assembler subroutine suitable for the IBM System/360.

6. TESTING THE ROBUSTNESS OF THE MRP

The first step in the analysis of the experiment is to test whether the MRP does guarantee the probability requirement in eq. (1), i.e. whether the estimated fraction of correct selections is not significantly lower than P^* (one-sided test), for all P^*-levels and combinations in the 2^{7-3} design. We use the t-statistic since $\underline{y}_{i\ell}$ is a proportion whose distribution should be well approximated by a normal distribution; see also Gross (1976). So

$$\underline{t}_{d_{i\ell}} = \frac{\underline{y}_{i\ell}-P^*_\ell}{\underline{s}_{i\ell}} \qquad (9)$$

with degrees of freedom $d_{i\ell} = 399$.

When applying a significance level α with eq. (9), then even if

$$H_0 : E(\underline{y}_{i\ell}) \geq P^*_\ell \text{ for all } i \text{ and all } \ell \qquad (10)$$

holds, the expected number of (false) rejections is $\alpha \times 16 \times 22$, e.g. for $\alpha = 0.05$ we expect 17.6 false significances. Instead of fixing this per comparison

error rate (α_C) we would prefer to fix the experiment-wise error rate (α_E). The latter means that we control the probability that all statements on the experiment are correct, provided H_0 in eq. (10) holds. A disadvantage of α_E is that the individual tests based on eq. (9) have small probabilities of detecting deviations from the null-hypothesis. We remedy this low power by fixing the family-wise error rates (α_F), i.e., the probability of all statements being correct is fixed per family of statements. (See Kleijnen (1975, pp. 526-531) for more comments on error rates.) Each family consists of the 22 statements for a particular combination in the 2^{7-3} design (m=22). The Bonferroni inequality, see Kleijnen (1975, p. 532), implies:

$$\alpha_F \leq m\, \alpha_C. \tag{11}$$

We may improve the power of the individual tests by limiting attention to, say, the 10 highest levels of P^* so that m=10. Eq. (11) yields Table V, (columns 2 and 3), where we fixed α_F to some reasonable values, which may be much higher than traditional values like 5%. To

TABLE V

Per Comparison Error Rates α_C
and Critical Points z_{α_C}
for Fixed Familywise Error Rates α_F

α_F	α_C		z_{α_C}	
	m = 22	m = 10	m = 22	m = 10
0.20	0.0090909	0.0200	2.362	2.054
0.10	0.0045455	0.0100	2.610	2.327
0.05	0.0022727	0.0050	2.838	2.576

calculate the critical points corresponding with α_C we use the normal approximation z since d = 399 in eq. (9); see columns 4 and 5. The reader may choose his own value for α_F; we prefer $\alpha_F = 0.20$ to protect the power of the individual tests.

Next we compare z_{α_C} with the t-statistics in eq. (9). A fraction is significantly low if its t-value is (algebraically) smaller than $-z_{\alpha_C}$ (one-sided test for eq. 10). A whole family (here a factor combination) is rejected if one or more t-statistics are significant, i.e. if its minimal t-statistic is significant; see Table VI (only the negative t are shown).

Tables V and VI give the following results:

(i) The factor combinations 2, 6 and 9 give significantly low fractions for any of the 3 values of α_F in Table V.

(ii) Combination 15 is rejected at m = 10 and $\alpha_F = 0.10$ (and therefore also at $\alpha_F = 0.20$), or at m = 22 and $\alpha_F = 0.20$.

TABLE VI

Smallest Negative t-value and Corresponding P^*-value

Factor Combination	m = 22		m = 10	
	t_{min}	P^*	t_{min}	$P^*(\geq 0.90)$
2	-4.6875	0.98	-4.6875	0.98
6	-5.2620	0.35	-3.7261	0.97
9	-6.1965	0.93	-6.1965	0.93
10	-2.4492	0.80	-2.3209	0.96
12	-1.4318	0.99	-1.4318	0.99
13	-0.7169	0.98	-0.7169	0.98
15	-2.4038	0.99	-2.4038	0.99
16	-1.9503	0.35	-0.3636	0.99

(iii) Combination 10 is rejected at $\alpha_F = 0.20$ and m = 22, and at $\alpha_F = 0.20$ and m = 10.

(iv) All other 11 combinations are not rejected for any of the values of α_F and m in Table V.

Because of (i) the conclusion is, that the MRP does not work under all circumstances; because of (iv) we conclude, that nevertheless the MRP guarantees the probability requirement in many situations. The former conclusion specifies the robustness more precisely than it was done in Bechhofer (1958, p. 426). The latter conclusion is reassuring after the doubts Bechhofer (1970) expressed about the correctness of the MRP. Once we know that the MRP does not work always, we proceed to the next step.

7. DETECTION OF IMPORTANT FACTORS

Several techniques may be considered for the detection of important factors, i.e. factors that cause the MRP to fail; see Kleijnen (1975, pp. 713-716). To save space we concentrate on the most powerful technique, which is also standard in the analysis of experimental designs, namely analysis of variance (ANOVA). Actually ANOVA is a subset of regression analysis which is known to a wide audience so that we present our analysis in regression terminology, as far as possible. For the 2^{7-3} design we may distinguish the grand mean β_0, the 7 main effects β_1 through β_7 and the 21 two-factor interactions β_{12} through β_{67} confounded in 7 sets of 3 as shown in eq. (4). Denote these ANOVA effects by the regression coefficients γ_1 through γ_{15}:

$$\gamma_1 = \beta_0; \quad \gamma_2 = \beta_1, \ldots, \gamma_8 = \beta_7;$$

$$\gamma_9 = \beta_{12} + \beta_{35} + \beta_{46}, \ldots \quad \gamma_{15} = \beta_{17} + \beta_{45} + \beta_{36}. \qquad (12)$$

With γ_1 through γ_{15} correspond 15 independent variables, say x_1 through x_{15}, which have only the values -1 or +1:

(i) For x_2 through x_8 see Table III.

(ii) For x_9 through x_{15} multiply the appropriate rows of Table III. For instance, for x_9 multiply row 1 with row 2, or row 3 with row 5, etc.; see eq. (4).

(iii) x_1 is identically +1.

In the 2^{7-3} design all independent variables x are orthogonal.

Each combination yields 22 dependent (correlated) variables $\underline{y}$ corresponding with P^*; see eq. (7). For each of the 22 P^* levels we hypothesize:

$$\underline{y}_i = \sum_{j=1}^{15} \gamma_j x_{ij} + \underline{e}_i \quad (i = 1,\ldots,16) \tag{13}$$

where $\underline{e}_i$ is an error term. Remember that $\underline{y}$ is the average of 400 independent observations $\tilde{\underline{y}}$ so that each of the 22 regression equations with 15 parameters is based on 6400 observations. The ANOVA (and ordinary regression) assumptions are univariate responses with normal, independent and homoscedastic error terms, or:

$$\underline{e}_i : \text{NID } (0,\sigma^2). \tag{14}$$

Actually our observations are nonnormal, heteroscedastic and multivariate. For instance, for $P^* = 0.99$ the estimated standard deviations range between 0.0000 and 0.0074, and for $P^* = 0.35$ between 0.0105 and 0.0160. We shall see later on that we can cope with these problems when performing various tests. Note that the point estimators remain unbiased.

ANOVA, or regression, yields the estimated effects or regression coefficients:

$$\hat{\underline{y}}_{j\ell} = \sum_{i=1}^{16} x_{ij}\, \underline{y}_{i\ell}/16 ,$$

$$j = 1,\ldots,15, \quad \ell = 1,\ldots,22 , \tag{15}$$

with estimated standard deviations

$$\underline{s}(\hat{\underline{y}}_{j\ell}) = \{ \sum_{i=1}^{16} \underline{s}_{i\ell}^2/16^2 \}^{\frac{1}{2}}. \tag{16}$$

Observe that contrary to the familiar ANOVA formulas, eq. (16) allows for unequal variances. Table VII displays the effects and their standard errors for $P^* \geq 0.90$. For $P^* < 0.90$ we show only $P^* = 0.35$ since we shall see that the effects for $P^* < 0.90$ are of little value. We examine Table VII in several steps.

(i) Preliminary Analysis

A small number of populations (factor 6) has a favorable effect since with only 3 populations even a random choice procedure gives $P(CS) = \frac{1}{3}$. With a large distance δ^* (factor 7) any procedure gives a CS most times. As P^* increases the factor effects become smaller, for then more observations are required by the MRP so that the sample means converge to μ_i and the 7 factors may become less important. Moreover, a deviation between $\underline{y}$ and P^* of, say, 0.05 is less important at $P^* = 0.35$ than at $P^* = 0.99$. (Perhaps we should have measured the response not by $\underline{y}$ but by, say $(\underline{y}-P^*)/(1-P^*)$.)

(ii) Lack-of-fit Tests for Regression Models

We can test whether the fitted regression equations are adequate approximations, by comparing 2 "spreads" of the observations $\underline{y}$. It is well known that

TABLE VII

Effects in the 2^{7-3} Design for Various Levels of P^*

P^*	Grand mean β_0	Main Effects (x 1000)							Two-factor Interactions (x 1000)							Standard Deviations (x 1000)
		1	2	3	4	5	6	7	12= 35= 46	13= 25= 67	14= 26= 57	23= 15= 47	24= 16= 37	34= 27= 56	17= 45= 36	
.35	60.0	-1	-22	53	8	3	54	175	-8	13	20	-21	-4	28	-30	3.4
.90	92.9	-3	-18	3	11	-2	20	15	9	7	15	-16	-2	15	6	2.1
.91	93.5	-3	-17	0	11	-2	20	14	7	5	15	-14	-1	14	6	2.1
.92	93.9	-3	-17	0	10	-2	19	12	8	5	15	-13	-1	12	6	2.0
.93	94.1	-2	-15	-1	10	-2	18	11	8	4	15	-13	-2	10	6	1.9
.94	95.0	-1	-13	-2	10	-3	16	9	9	4	14	-12	-1	9	6	1.8
.95	95.7	-0	-12	-3	8	-2	15	8	8	3	12	-11	-1	8	6	1.7
.96	96.3	1	-11	-4	5	-3	13	7	7	2	11	-10	-1	7	5	1.6
.97	97.1	0	-10	-3	5	-3	12	6	5	1	11	-9	-2	6	4	1.4
.98	97.8	-1	-8	-4	4	-2	10	4	4	2	9	-7	-2	4	4	1.3
.99	98.6	0	-4	-3	3	-3	7	3	2	-0	6	-4	-2	3	3	1.0

the residual mean squares $\underline{S}_R^2$ in eq. (17) has expected value σ^2 provided the regression equation is correctly specified:

$$_{\ell}\underline{S}_R^2 = \sum_{i=1}^{16} (\underline{Y}_{i\ell} - \hat{\underline{Y}}_{i\ell})^2/(16-15) \qquad (\ell = 1,\ldots,22)$$

$$(\ell=1,\ldots,22) \tag{17}$$

with predicted observations $\hat{\underline{Y}}_{i\ell} = \sum_{j=1}^{15} \hat{\underline{\gamma}}_{j\ell} x_{ij}$. Then $_{\ell}\underline{S}_R^2/\sigma^2$ is a χ^2-variable with 1 degree of freedom. A second estimator of σ^2 is based on the 400 independent duplications for each factor combination regardless of the model in eq. (13); see eq. (8). Because eq. (14) implies a common σ^2, we take the "pooled" estimator, say, $\underline{s}_D^2$:

$$_{\ell}\underline{s}_D^2 = \sum_{i=1}^{16} \underline{s}_{i\ell}^2/16 \qquad (\ell=1,\ldots,22). \tag{18}$$

Hence, $_{\ell}\underline{s}_D^2$ d/σ^2 is a χ^2-variable with $d = 16\times 399 \approx \infty$ degrees of freedom, and is independent of $_{\ell}\underline{S}_R^2$. To compare the 2 estimators of σ^2 we calculate:

$$_{\ell}\underline{F}_{1,d} = \frac{_{\ell}\underline{S}_R^2}{_{\ell}\underline{s}_D^2} \qquad (\ell=1,\ldots,22). \tag{19}$$

High values of $\underline{F}$ indicate lack of fit since an incorrect regression equation inflates $\underline{S}_R^2$; see eq. (17). In Kleijnen (1975, pp. 367, 724-725) this F-test and its insensitivity to the assumptions in eq. (14) are discussed in more detail.

Table VIII gives the "critical" levels α_L, i.e. the smallest value of α for which the calculated F in eq. (19) exceeds $F_{1,\infty}^{\alpha}$. (This upper F-point is tabulated for standard α values only.)

TABLE VIII

Critical Levels α_L for Lack-of-Fit Tests

P^*	.35	.40	.45	.50	.55	.60	.65	.70	.75	.80	.85
α_L	.25	.005	.0005	.0005	.0005	.0005	.001	.0005	.005	.0005	.02
P^*	.89	.90	.91	.92	.93	.94	.95	.96	.97	.98	.99
α_L	.10	>.25	>.25	>.25	.25	.25	>.25	>.25	>.25	>.25	>.25

If we find one or more α_L smaller than the α_C of Table V (under m=22), then we reject the null-hypothesis that all regression equations give good fit. Tables V and VIII do show that we have to reject this hypothesis for any α_F. However, we may decide to restrict our study to the 10 highest levels of P^* (m=10 in Table V, and the lower part of Table VIII). Then we arrive at the important conclusion that for $P^* \geq 0.90$ the fitted regression equations are good approximations. In the sequel we shall concentrate on these regressions.

(iii) Significance Tests for Effects

Per equation we can test whether the regression coefficients are 0 through an F-test per effect, or equivalently a t-test since the numerator degrees of freedom is 1:

$$\ell' \underline{F}_{1,d} = \ell' \underline{t}_d^2 = \frac{\ell' \underline{SS}_j}{\ell' \underline{s}_D^2} = \frac{16(\hat{\underline{\gamma}}_{j\ell'})^2}{\ell' \underline{s}_D^2}$$

$$(\ell'=13,\ldots,22) \qquad (20)$$

where $\underline{SS}_j$ is the sum of squares for effect j; see Johnston (1963, p. 135). These F-tests are quite conservative in the face of heterogeneity of variance and nonnormality, especially with large equal numbers of observations per combination; Scheffé (1964, p. 331-369).

We also test two-factor interactions jointly, i.e. we hypothesize that first order regression equations give adequate approximations, by pooling the sums of squares for the two-factor interactions:

$$\ell' \underline{F}_{7,d} = \frac{\frac{1}{7} \sum_{j'=9}^{15} \ell' \underline{SS}_{j'}}{\ell' \underline{s}_D^2} . \qquad (21)$$

The critical levels α_L for eqs. (20) and (21) are shown in Table IX, which yields the following conclusions:

(1) Even at the small α_C of Table V all effects are significant, except the main effect 1 (truncation) and the confounded two-factor interactions 16+24+37.

(2) The main effects 2 (skewness), 6 (number of populations k) and their interaction give the most significant results (together with the confounded interactions 23+15+47). But also very significant are factor 7 (δ^*) and factor 4 ($\sigma_{(k)}/\max \sigma_{(j)} \neq 1$) (and the interactions 34+27+56 and 12+35+46).

(3) We might claim that the factors 1 and 5 (variance configuration) are unimportant provided we take α_F as low as 0.01.

(4) We may not state that the factors 1, 5 and 3 are unimportant, since in that case $\alpha_F < 0.01$.

(5) We should realize that an effect may be statistically significant, yet unimportant! For, if a regression coefficient is nonzero, but the overall mean β_0 is high, then it is quite well possible that the expected fraction of CS is not smaller than the required probability P^*.

We conjecture that such a phenomenon indeed exists since all factors except one (or two) are significant so that nearly all 16 factor combinations would otherwise yield estimated fractions significantly below P^*. Actually we saw in section 6, that the MRP works most times. So it seems incorrect to conclude from the regression

TABLE IX

Critical Levels α_L of Significance Tests for Effects

P^*	Grand Mean	Main Effects							Two-factor Interactions							Joint Inter-
	β_0	1	2	3	4	5	6	7	12=35 =47	13=25 =67	14=26 =57	23=15 =47	24=16 =37	34=27 =56	17=45 =36	actions
.90	.0005	.25	.0005	.25	.0005	>.25	.0005	.0005	.0005	.005	.0005	.0005	>.25	.0005	.01	.0005
.91	"	"	"	>.25	"	"	"	"	"	.025	"	"	"	"	.005	"
.92	"	"	"	"	"	"	"	"	"	.01	"	"	"	"	"	"
.93	"	>.25	"	"	"	"	"	"	"	.05	"	"	"	"	"	"
.94	"	"	"	.25	"	.25	"	"	"	"	"	"	"	"	.001	"
.95	"	"	"	.05	"	"	"	"	"	.10	"	"	"	"	.0005	"
.96	"	"	"	.025	.001	"	"	"	"	.25	"	"	"	"	.001	"
.97	"	"	"	"	.0005	.01	"	"	"	>.25	"	"	"	"	.005	"
.98	"	"	"	.005	.005	.25	"	.005	.005	.10	"	"	.25	.005	.001	"
.99	"	"	"	.001	"	.005	"	.001	.025	>.25	"	"	.05	"	"	"

coefficients that all factors except factor 1 (and 5) are "important". We conjecture that primarily factors 2 and 6 make the MRP break down. Note that a high β_0 may exist when the MRP yields "overprotection", i.e. even in the LFC $P(CS) > P^*$ instead of $P(CS) = P^*$.

Summarizing, the preceding section showed that the MRP does not always work so that some factors must be important. Hence we like to know which factor combinations make the procedure break down. The analysis in the present section could not reveal exactly which factors are really important. For the regression analysis showed that at the usual error rate values all effects are significant, except β_1 (truncation) and $\beta_{16} + \beta_{24} + \beta_{37}$. However, statistically significant effects may be unimportant. The factors 2 (symmetry) and 6 (k) and their interaction seem most important, but definitive conclusions were not possible.

8. A SECOND EXPERIMENT

In a second experiment we tried to arrive at definitive conclusions. Since we had only a rather slow computer (ICL 1903A) available, we reduced the scale of our experiment quite drastically. Therefore, besides eliminating factor 1 taking only truncated distributions, we also eliminated factor 5 making all "inferior" variances equal. The factors k and δ^* were taken at their unfavorable levels only, i.e. $k = 7$ and $\delta^* = 0.20 \{\sigma_{(k)} + \sigma_{(j)}\}/2$. We restricted P^* to 0.90 and (since the programming was done by students) the program was further simplified eliminating antithetic variates. The remaining 3 factors are: factor 1' (exponential versus Erlang distributions), factor 2'

($\sigma_{(k)}$ lower versus higher than $\sigma_{(j)}$) and factor 3' ($\sigma_{(k)}$ and $\sigma_{(j)}$ approximately equal or not); see Tables I and II above. These three factors are investigated in a full 2^3 design with 800 replications. The resulting distributions are pictured in FIG. 1, which also shows the resulting $\hat{P}$.

Since no antithetics are used, $\hat{P}$ is a binomial variable with $\hat{var}(\hat{P}) = \hat{P}(1-\hat{P})/(n-1)$. As we apply a one-sided test, we are interested in negative t-values only. For combination 3, FIG. 1 shows $\hat{P} = 0.864$ resulting in $t = -2.99$ and for combination 7, $\hat{P} = 0.879$ with $t = -1.84$. For an experiment with 8 tests the experimentwise error rates $\alpha_F = 0.20, 0.10, 0.05$ correspond with $\alpha_C = 0.025, 0.0125, 0.00625$, i.e. $z_{\alpha_C} = 1.96, 2.24, 2.50$. So combination 3 is significant but not combination 7, for any α_F-value. How can this result be interpreted?

From the experimental design it follows that if only 1 factor were important, then 4 of the 8 combinations would be rejected. Likewise, if only a particular two-factor interaction were important, then 2 combinations would be rejected. For instance, combinations 3 and 7 would both give significant results if the interaction between factors 2' and 3' were important, more particulary, if $\sigma_{(k)} > \sigma_{(j)}$ combined with $\sigma_{(k)} \approx \sigma_{(j)}$ would make the MRP break down. Actually, only combination 3 is significant so that a <u>three-factor interaction</u> must be blamed, viz. the combination of $\sigma_{(k)} > \sigma_{(j)}$, $\sigma_{(k)} \approx \sigma_{(j)}$ and exponential distributions.

Does the importance of this particular three-factor interaction not conflict with the <u>first</u> 2^{7-3} experiment? It can easily be checked using Table III that the combination with $\sigma_{(k)} > \sigma_{(j)}$, $\sigma_{(k)} \approx \sigma_{(j)}$ and exponentiality, was combination 9. And indeed Table VI showed that combination 9 gave the most significant

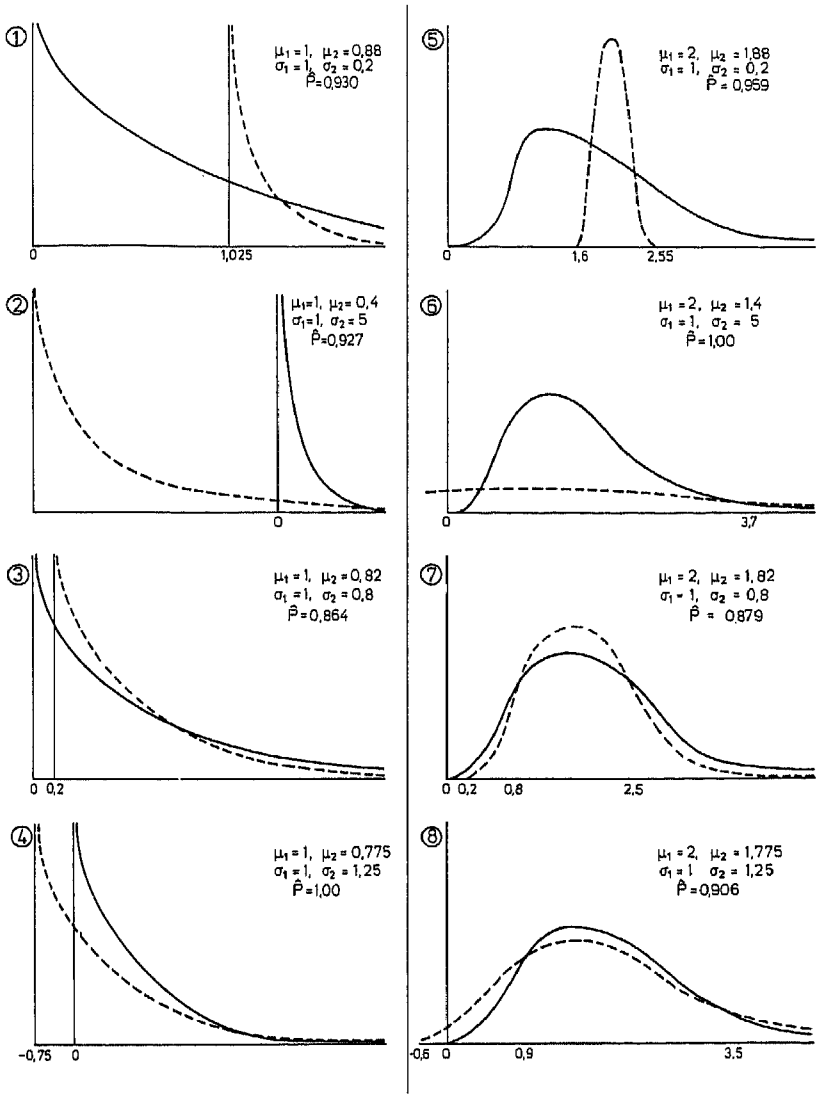

Figure 1. Second experiment

result! Unfortunately the combinations 2 and 6 were also significant. If we accept the unimportance of factor 1 (truncation), then these 2 combinations might be classified as "exponentials" with $\sigma_{(k)} > \sigma_{(j)}$, but (in conflict with the second experiment's results) $\sigma_{(k)} \neq \sigma_{(j)}$, and "exponentials" with $\sigma_{(k)} \approx \sigma_{(j)}$ but $\sigma_{(k)} < \sigma_{(j)}$, respectively. Other factors interact, e.g. δ^* was large in combination 2. If we accept that factor 1 (truncation) has no effect, then combination 10 also becomes relevant. Indeed this combination was significant at $\alpha_F = 0.20$. Its smaller significance might be explained by the small value of k, namely k = 3. Anyhow, the second experiment does not explain all results of the first experiment.

We tried one more explanation of the results of the second experiment. Looking at combination 3 in FIG. 1 one might conjecture that significance is caused by exponentiality combined with equal shape, i.e. the form parameters $1/\lambda$ are nearly equal, or $\sigma_{(k)} \approx \sigma_{(j)}$. This interpretation, however, would imply that both combination 3 and combination 4 would fail. We executed 1 additional run, where combination 4 was made to give even more equal distributions, since we shifted the inferior populations to the right ($\mu_2 = 0.82$). However, again $\hat{P} > P^*$ (namely $\hat{P} = 0.902$) so that combination 4 does not fail.

Observe that the wrong selection in combination 3 was not accompanied by a small number of stages. In Table X we give $\bar{m}$, the average number of stages at which the stopping statistic $z \leq (1-P^*)/P^*$ (see section 1) and the standard deviation of m if available.

Summarizing, the second experiment suggests that the MRP fails if we have exponential distributions with

TABLE X

Number of Stages m

Combination	1	2	3	4	5	6	7	8
$\bar{m}$	57	324	105	16	61	8	105	240
$\hat{\sigma}_m$	47	119	57	1.5	88.9	7.85	122	.

$\sigma_{(k)} > \sigma_{(j)}$ and $\sigma_{(k)} \approx \sigma_{(j)}$. It is reassuring, however, that even if the MRP fails, the probability of correct selection is still high, namely $\hat{P} = 0.86$ for $P^* = 0.90$. For near-normal distributions like gamma distributions, the MRP does give $\hat{P} > P^*$ even for unequal variances. So in general, this MRP seems quite robust.

ACKNOWLEDGEMENTS

Prof. D. Burdick (Duke University) and Prof. G. Shorack (University of Washington) commented on the design and analysis for the first experiment. Computer time was made available by Prof. T. Naylor (Duke University; IBM 360/75) for the data generation; the analysis was done at the Katholieke Hogeschool (IBM 1620). Prof. H. Lombaers (Technische Hogeschool Delft; IBM 360/65) made additional data generation possible. B. Fitzgerald (Duke University) wrote the FORTRAN program for the data generation. The second experiment was performed by J. Geurts and P. Heesters (Katholieke Hogeschool).

BIBLIOGRAPHY

Bechhofer, R.E. (1958). 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, 408-29.

Bechhofer, R.E. (1970). Correction note: An undesirable feature of a sequential multiple-decision procedure for selecting the best one of several normal populations with a common unknown variance. Biometrics 26 (2), 347-49.

Bechhofer, R.E. and S. Blumenthal (1972). 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, 52-67.

Box, G.E.P. and J.S. Hunter (1961). The 2^{k-p} fractional factorial designs, Part I. Technometrics 3 (3), 311-51.

Gross, A.M. (1976). Confidence interval robustness with long-tailed symmetric distributions. J. Amer. Statist. Assoc. 71, 409-16.

Johnston, J. (1963). Econometric Methods. New York: McGraw-Hill Book Company, Inc.

Kleijnen, J.P.C. (1974/1975). Statistical Techniques in Simulation, Vols I + II. New York: Marcel Dekker Inc.

Lewis, P.A.W., A.S. Goodman and J.M. Miller (1969). A pseudo-random number generator for the system/360. IBM systems J. 8 (2), 136-47.

Scheffé, H. (1969). The Analysis of Variance. New York: John Wiley & Sons, Inc.

Received July 1977; revised November 1977; corrected February 1977.

Refereed by Alan M. Gross, Bell Laboratories, Murray Hill, NJ.