

AN EMPIRICAL INVESTIGATION OF THREE PROCEDURES  
FOR MULTIPLE SIGNIFICANCE TESTS  
OF INTERCORRELATIONS

A THESIS

Presented to

The Faculty of the Division of Graduate  
Studies and Research

By

Robert Earl Larzelere


In Partial Fulfillment  
of the Requirements for the Degree  
Master of Science in Psychology


Georgia Institute of Technology


March, 1975

AN EMPIRICAL INVESTIGATION OF THREE PROCEDURES  
FOR MULTIPLE SIGNIFICANCE TESTS  
OF INTERCORRELATIONS

Approved:

  
\_\_\_\_\_  
S. A. Mulaik, Chairman

  
\_\_\_\_\_  
C. V. Riche

  
\_\_\_\_\_  
M. C. Spruill

Date approved by Chairman: 3/7/75

## ACKNOWLEDGMENTS

The author wishes to thank the members of his thesis advisory committee, Dr. Stanley A. Mulaik, Charles V. Riche, and Marcus C. Spruill for their help and advice. Dr. Mulaik, chairman of the committee suggested the problem which this thesis investigated. His efforts, criticisms, suggestions, guidance, and encouragement were especially appreciated. Dr. Paul A. Games was also appreciated for reading and evaluating an earlier draft of this thesis.

The author particularly wishes to acknowledge the help of his wife Rosalie. Without her encouragement, support, and typing this thesis would not have been possible.

Finally, the most indispensable Person to the author was Jesus Christ, who provided everything necessary for the completion of this thesis.

## TABLE OF CONTENTS

	Page
ACKNOWLEDGMENTS . . . . .	ii
SUMMARY . . . . .	iv
<b>Chapter</b>	
I. INTRODUCTION . . . . .	1
Multiple-Test Procedures	
Some Related Aspects of Intercorrelations	
Hypotheses and Objectives of Study	
II. METHOD . . . . .	43
Monte Carlo Method	
Generation of Independent Random Normal	
Deviates	
Generation of Sample Correlation Matrices	
Tests of Hypotheses of This Study	
Summary of Analyses	
III. RESULTS. . . . .	72
Statistical Tests of Pseudorandom Number	
Generator	
Results of Tests of Hypotheses of This	
Study	
IV. DISCUSSION . . . . .	93
Adequacy of Pseudorandom Number Generator	
Control of Type I Error Rate	
Control of Type II Error Rate	
Conclusions	
Recommended Procedure for Controlling	
Type I Error Rates	
Applications to Extant Data	
APPENDICES. . . . .	134
BIBLIOGRAPHY. . . . .	172

## SUMMARY

Multiple  $t$  tests can be used for multiple comparisons in analysis of variance and for multiple significance tests of a set of intercorrelations. In both types of analysis, if a nominal Type I error rate is applied to each single test (e.g., .05), then the probability of observing one or more Type I errors in the set of significance tests increases a great deal above .05. In multiple comparisons in analysis of variance, diverse multiple-test procedures have been recommended to correct for this. However, in multiple significance tests of intercorrelations, the same rationale for more stringent Type I error rates remains largely ignored. In this study a computer simulation was used to determine the effect of using the ordinary single-test procedure on Type I error rates in multiple significance tests of intercorrelations. Population correlation matrices were specified, and random samples were drawn from those populations. The results indicated that when a nominal value of  $\alpha = .05$  is applied to each single significance test, the familywise Type I error rate increases rapidly to undesirable levels as the number of variables increases. Two alternative procedures were also investigated, a Bonferroni  $t$  procedure and an assumed-independent-tests procedure. Both were successful

in keeping the familywise Type I error rate at the nominal value of  $\alpha = .05$  or below. Both over-controlled for Type I errors when at least a small proportion of the null hypotheses of interest were false. It was also found that the mutual dependence of the component significance tests results in high conditional Type I error rates if some correlations between the variables of interest are moderate or large in magnitude. Generally, this increases the probability of a relatively large number of Type I errors occurring simultaneously. A multistage Bonferroni  $\dagger$  procedure is outlined and recommended.

## CHAPTER I

## INTRODUCTION

The purpose of this study is to examine the effects of different procedures for controlling Type I error rates in multiple significance tests of intercorrelations. When only one correlation coefficient is involved, the  $t$  statistic is an appropriate statistical test for a hypothesis of the form

$$\begin{aligned} H_0 &: \rho = 0 \\ H_1 &: \rho \neq 0 \end{aligned} \quad (1)$$

But when the  $t$  statistic is used with an unadjusted alpha level (i.e., a critical value appropriate for testing a single correlation coefficient at a given alpha level) to test each of the  $\frac{k(k-1)}{2}$  intercorrelations in a correlation matrix of  $k$  variables, the expected number of Type I errors increases rapidly as  $k$  increases. Hays (1973, p. 712) states that

The resulting significance levels are largely meaningless, for reasons much like those making  $t$  tests for all differences among a set of means a dubious procedure. In the first place, even for independent tests of significance, when so many tests are carried out the probability that some Type I errors are being made may be very high. Even worse, the  $t$  tests for correlations are quite redundant and are not statistically in-

dependent when carried out on a table of inter-correlations. Consequently, the set of results can be grossly misleading.

If such significance tests are used, he adds, the experimenter should interpret the significance levels with considerable latitude.

As the above quote from Hays indicates, the problem of controlling Type I error in multiple tests of inter-correlations is similar to the problem of controlling Type I error in multiple comparisons in analysis of variance. Multiple-test procedures are fairly widely used in psychological research for multiple comparisons in analysis of variance. Although the rationale is very similar for the use of multiple-test procedures in multiple tests of inter-correlations, ordinary, single-test procedures continue to be widely (if not universally) used with multiple tests of correlations. This contradictory state of affairs probably has resulted from the historical fact that the major multiple-test procedures were first published in textbook form in Analysis of Variance (Scheffé, 1959), and from the mathematical fact that multiple tests of intercorrelations do not generally meet the assumptions of the better-known multiple-test procedures. As early as 1959 Ryan specifically mentioned multiple tests of intercorrelations as one case where multiple-test procedures should be applied. However, following historical precedent, the bulk of his article was about multiple comparisons in analysis



of variance. Later he presented a general multiple-test method for statistical analyses other than comparisons among means. His method can be applied to any set of two-sample significance tests with hypotheses of the form

$$\begin{aligned} H_0 &: P_1 = P_2 \\ H_1 &: P_1 \neq P_2 \end{aligned} \quad (2)$$

where  $P_i$  represents a population parameter for the  $i^{\text{th}}$  population (Ryan, 1960). This is only a short step away from a general method which can be applied to hypotheses of the form

$$\begin{aligned} H_0 &: P_1 = a \\ H_1 &: P_1 \neq a \end{aligned} \quad (3)$$

where  $a$  is the hypothesized parameter value. However, there have apparently been few further developments in the application of multiple-test procedures to intercorrelations in a correlation matrix since Ryan's (1959, 1960) papers, at least not in sources used much by psychological researchers.

Marascuilo (1966) did consider multiple-test procedures for hypotheses of the form

$$\begin{aligned} H_0 &: \rho_{12} = \rho_{34} = \rho_{56} = \dots = \rho_{(n-1)n} \\ H_1 &: \rho_{(i-1)i} \neq \rho_{(j-1)j} ; \text{ for some } i, j \end{aligned} \quad (4)$$

His method assumes mutually independent tests and has unknown adequacy for small samples. Harris (1967) investigated the effects of the non-independence of significance tests of correlations in correlation matrices. He concluded that the Type I error rate was greatly distorted by non-independence and that there may be no feasible correction for such distortion. He did not consider any multiple-test procedures to correct this distortion. Currently, psychological researchers continue to ignore questions about the distortion of Type I error rates in ordinary single-test procedures for multiple significance tests of intercorrelations.

This thesis examines the problem of controlling Type I error rate in multiple tests of intercorrelations by first reviewing the literature on multiple-test techniques and then reviewing those mathematical characteristics of correlation matrices which must be taken into account in applying multiple-test procedures to them. Thirdly, an empirical investigation of Type I and Type II errors under several procedures when the population correlation matrix is known is reported. Finally, the implications of the results of the literature review and the empirical findings are discussed.

#### Multiple-Test Procedures

#### Widely Divergent Opinions and Procedures

Several authors (Carmer & Swanson, 1973; Dunnett,

1970; O'Neill & Wetherill, 1971; Waller & Duncan, 1969) have commented recently on the controversy among statisticians about multiple-test procedures and their disagreement concerning the basic principles involved. Controversy seems to exist concerning the nature of the basic problems, relevant criteria for multiple-test procedures, and properties of currently proposed procedures. Such disagreement was illustrated in a meeting of the Royal Statistical Society on this very topic. Following O'Neill and Wetherill's (1971) paper, the first discussant mentioned that it was good to have a meeting on these issues since the problems their paper discussed still existed after 30 years and some 200 odd papers on the topic (Plackett, 1971). The next discussant stated that "multiple comparison methods have no place at all in the interpretation of data (Nelder, 1971, p. 244)", adding that their principle purpose was to lend an air of respectability to otherwise uninteresting data.

Disagreements among statisticians are also evident in their recommendations for multiple-test procedures. Currently available techniques for multiple comparisons lead to very different results in many cases. For example, the error rate per experiment varied from essentially zero for Scheffé's procedure to over 1.00 for Duncan's Multiple Range Test in one case simulated in Petrinovich and Hardyck's (1969) study. Yet both procedures have their pro-

ponents among applied statisticians. For example, Petrinovich & Hardyck (1969) recommended Scheffé's or Tukey's procedures as vastly superior to Duncan's method, whereas Carmer & Swanson (1973) concluded that Scheffé's & Tukey's procedures were both clearly inferior to Duncan's procedure.

Aside from such major issues, opinions differ greatly on other matters, such as the importance of distinguishing between a priori and a posteriori tests and of distinguishing between the cases of independent tests and non-independent tests.

#### Greater Complexities Than Single-Test Procedures

Probably a primary factor behind such widely divergent opinions is the greater complexity involved in multiple tests as compared with a single significance test. At the single-test level, confidence limits also indicate the result of a significance test; this is not always true with multiple tests. Distinctions can be made between different kinds of Type I error rates for multiple tests; all these are the same at the single-test level. At the single-test level, a decision is made between the null hypothesis and an alternative hypothesis; with multiple-test situations there are more distinct decisions possible, involving various combinations of decisions on the component significance tests. This also

complicates the relationship between Type I error rate and power.

Probably the central issue of multiple-test procedures is the appropriateness of various generalizations from a single-test procedure to a multiple-test procedure. For example, what kind of Type I error rate is an appropriate multiple-test-procedure generalization from the usual single-test Type I error rate? Does it indicate the same dependability of results as a reported Type I error rate for a single test? This issue is considered in more detail in the next section.

#### Type I Error Rates

As already indicated, expressing Type I error rates for multiple tests in a way that is directly analogous to single tests is not a simple problem. Waller and Duncan (1969) call this issue a major source of disparities in multiple-test procedures.

Three Type I error rates have been distinguished for multiple tests, error rate per individual test ( $\alpha_T$ ), error rate per family ( $\alpha_{PF}$ ), and error rate familywise ( $\alpha_{FW}$ ). The simplest multiple-test situation will be used to illustrate the differences among these Type I error rates. This multiple-test situation involves a set of two statistically independent significance tests. In any multiple-test procedure, the set of individual significance tests (in this case 2) is called a family

(This concept of a family is considered in more detail later in this chapter). The individual significance tests are called component tests.

Table 1 gives the probabilities of observing zero,

Table 1. Probability of a Given Number of Type I Errors in a Family of Two Independent Tests<sup>a</sup>

	Number of Type I Errors		
	0	1	2
Probability	.9025	.095	.0025

<sup>a</sup>  $\alpha_T = .05$

one, or two Type I errors in this particular family of tests, given  $\alpha_T = .05$  when the null hypothesis is true for both tests (i.e., two true component null hypotheses). The Type I error rate per test ( $\alpha_T$ ) is simply the probability of a Type I error on a single statistical test, in this example, .05. The Type I error rate per family is the expected number of Type I errors in the entire family of tests, i.e.,

$$\alpha_{PF} = E(\text{total Type I errors}) \quad . \quad (5)$$

Note that  $\alpha_{PF}$  is actually not a probability but an expectation. In the example in Table 1, the error rate per

family is

$$\begin{aligned}\alpha_{PF} &= 0 (.9025) + 1 (.095) + 2 (.0025) \\ &= .010\end{aligned}\quad (6)$$

The familywise Type I error rate is the probability of observing one or more Type I errors in a family of tests, i.e.,

$$\alpha_{FW} = 1 - \text{Pr}(\text{zero total Type I errors}) \quad (7)$$

In Table 1, the familywise error rate is

$$\begin{aligned}\alpha_{FW} &= 1 - .9025 \\ &= .0975\end{aligned}\quad (8)$$

Both the error rate per test and the familywise error rate are used in current psychological literature. The overall F test in analysis of variance is an example of the use of familywise error rate, whereas tests of inter-correlations in a correlation matrix (e.g., Kolb, 1973; Paige, 1973) or between predictor and criterion variables (e.g., Brooks, 1973; Jessor & Jessor, 1974; Pedersen, 1973a, 1973b; Siess, 1973) are examples of the use of error rate per test.

When a statistic exceeds the critical values as

determined by all three error rates, the null hypothesis is clearly rejected. When it is smaller than the critical values for all three error rates, the null hypothesis is always accepted. When only one significance test is

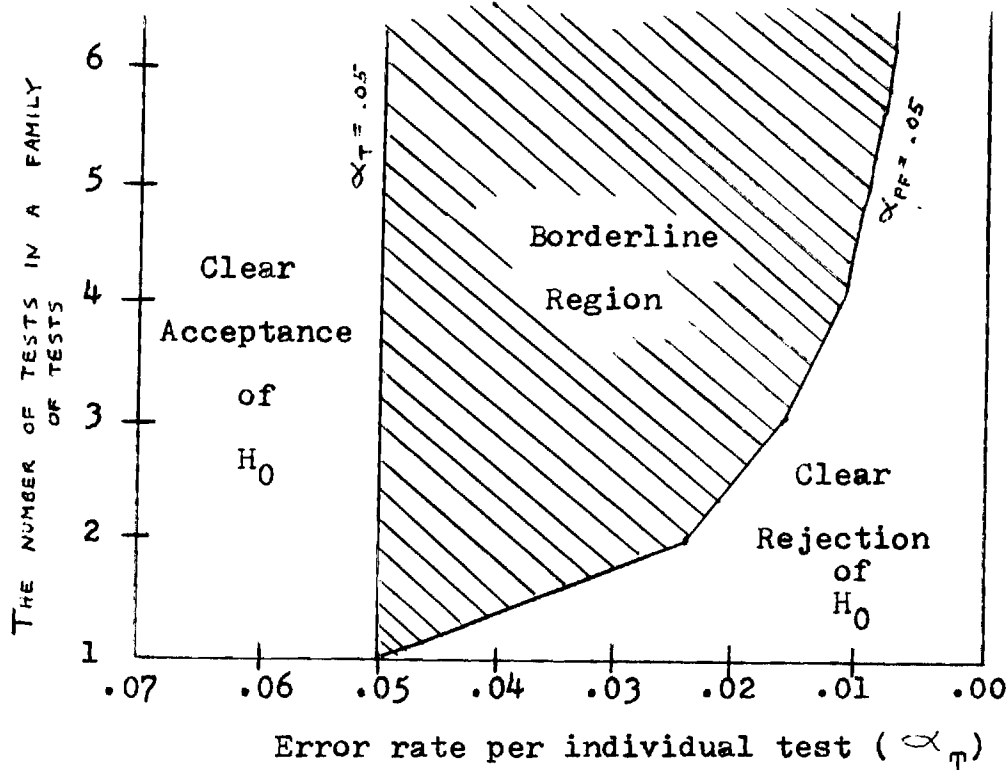


Figure 1. Borderline Region for Type I Error in Multiple Tests

under consideration, this exhausts all the possibilities. However, when a family of two or more tests is under consideration, some statistical values are possible which exceed the critical value for  $\alpha_T = 0.05$  (for example) but are less than the critical value for  $\alpha_{PF} = 0.05$ . Such



statistical values would fall in the shaded region of Figure 1. A central issue in multiple-test procedures is what conclusion to make when the statistical value falls in this borderline region. This issue is equivalent to the issue of which kind of Type I error rate to control for. If  $\alpha_T$  is controlled at .05, then the borderline region is included in the rejection region. If  $\alpha_{PF}$  is controlled at .05, then the borderline region is included in the acceptance region. Notice that the borderline region gets larger as the number of component tests increases.

What Type I error rate should be controlled for? The literature on multiple-test procedures and current practices in statistical analysis of psychological research suggests that no general clear-cut answer can be given. Two major issues, however, should be kept in mind to answer this for a particular situation. Ryan (1959) has pinpointed perhaps the most important issue: Which Type I error rate is the best representation of our results? Most psychological researchers associate a strong degree of dependability with experimental "facts" at the .05 level, and a greater degree of dependability with "facts" accepted at the .01 level. When .05 is the reported significance level associated with "facts" presented in a multiple-test situation, it should represent the same degree of dependability as a .05 significance level in a

single-test situation.

Some statisticians view this issue as an attempt to formalize into a Neyman-Pearson statistical model something which should be much more flexible (e.g., O'Neill & Wetherill, 1971). Such a statement is not without merit, but the fact of the matter is that psychologists are much more familiar with Neyman-Pearson statistical concepts than with Bayesian concepts. A Neyman-Pearson model is useful for reporting the dependability of experimental results, which psychologists use, in turn, to distinguish between findings that must be integrated into their theories and findings that may be ignored until their replicability can be demonstrated. If findings were reported in terms of Bayesian concepts, most psychologists would have difficulty in interpreting the dependability of the reported findings. So the issue of a known degree of protection for a family of tests against Type I errors seems to be most crucial, at least for current psychological research.

A second issue is the consistency of treatment of research results regardless of the type of analysis. Currently, a familywise Type I error rate is commonly used in analysis of variance (e.g., an overall  $F$  test), while a Type I error rate per individual test is commonly used in statistical tests of a set of intercorrelations. This inconsistency is unreasonable. One researcher should

not be penalized simply because an analysis of variance is applicable to his research while another researcher can use tests with more power simply because he has correlational data.

The advantages and disadvantages of the three Type I error rates are considered next. The Type I error rate per test ( $\alpha_T$ ) is the easiest to use and results in more powerful tests than the other two more conservative alternatives. Although generally favoring  $\alpha_{FW}$ , Miller (1966) indicates that  $\alpha_T$  can be appropriate if the consulting statistician and the researcher are both aware of the implications of using  $\alpha_T$  in a multiple-test situation. However, the logical extension of this is that the average reader should be aware of these implications if  $\alpha_T$  is used in publishable research. And one implication is that the probability of observing at least one Type I error in a family of tests increases rapidly as the number of component tests increases. For example, in a family of five independent component tests with  $\alpha_T = .05$ , there would be at least one Type I error over 22% of the time; in a family of 10 tests, over 40% of the time; and in a family of 20 tests, over 64% of the time. Ryan (1959) has pointed out that if  $\alpha_T$  is used, then the more variations of experimental conditions a researcher investigates, the better the chance of finding some apparently significant

results. This leads to a greater reward for working harder on irrelevant variables.

Most of the literature on multiple-test procedures favors a more conservative Type I error rate (e.g., Miller, 1966; Ryan, 1959; Scheffé, 1959). The Type I error rate familywise ( $\alpha_{FW}$ ) and the Type I error rate per family ( $\alpha_{PF}$ ) are usually almost equivalent, particularly when the desired alpha level is small and when  $\alpha_{FW}$  can be calculated accurately. Of these two,  $\alpha_{FW}$  is generally preferred. The most common use of  $\alpha_{FW}$  is the overall F test in analysis of variance. The Type I error rate familywise is a probability whereas  $\alpha_{PF}$  is an expectation and not a probability. The Type I error rate familywise gives a known probability of protection to the family of tests against any Type I error. These may be the main reasons for the preference for  $\alpha_{FW}$ . Ryan (1959), however, concluded that  $\alpha_{FW}$  and  $\alpha_{PF}$  simply represent different viewpoints. If one Type I error is viewed as nearly as costly as several Type I errors, then  $\alpha_{FW}$  is to be controlled. If two errors in one family of tests is considered as bad as one error in each of two families, then  $\alpha_{PF}$  is to be controlled. But in general the literature favors  $\alpha_{FW}$  over the other two Type I error rates, although  $\alpha_{FW}$  is usually practically equivalent to  $\alpha_{PF}$ .

If we decide to control  $\alpha_{FW}$ , then most statistical values in the borderline region (Figure 1) result

in acceptance of the null hypothesis, since  $\alpha_{FW}$  is usually nearly equivalent to  $\alpha_{PF}$ . But at the single-test level, this borderline region becomes the single critical value itself. So, analogously, it should be remembered that the borderline region represents statistical values on the fence between the acceptance and rejection regions. As such they would represent more dependable results than single test results reported as "tending toward significance". The implication of this is that results in the borderline region might be especially worthy of follow-up studies. So, as recommended elsewhere (Miller, 1966; Petrinovich & Hardyck, 1969),  $\alpha_T$  is useful for exploratory research, whereas definitive, publishable research should use  $\alpha_{FW}$ .

The distinction here is similar to the distinction made by Fisher (1935) between results which can suffice in themselves to establish the point at issue and results which are of less value except insofar as they confirm or are confirmed by other experiments of like nature. Statistical effects which are large enough to reject the null hypothesis even with  $\alpha_{FW} = .05$  can be considered to stand alone in establishing the point at issue, whereas other statistical effects which fall in the borderline region need to be confirmed by other research.

One necessary consequence of a more stringent Type I error rate, such as  $\alpha_{FW}$ , is the accompanying loss in power. Ryan (1962) points out that this decrease in power

will result in Type II errors only for small, and therefore, generally less important, effects (empirical demonstration of this: Carmer & Swanson, 1973). It is difficult to evaluate this because the knowledge of power functions for multiple test statistics is very limited except for the  $F$  and  $t$  statistics (Miller, 1966; O'Neill & Wetherill, 1971). However, as the number of tests per family increases, the power decreases rapidly. Multi-stage procedures have been proposed to lessen the loss of power.

As the name suggests, multistage procedures involve several stages in the analysis. In the first stage the critical value for the component tests is determined so that  $\alpha_{FW}$  is controlled at the desired level. If no component null hypothesis is rejected at this stage, the procedure terminates. However, if at least one component null hypothesis is rejected, less stringent critical values are used in the following stages to test the component null hypotheses which were not rejected in the first stage.

The rationale for this is that the multistage procedure increases power without increasing the familywise Type I error rate for the complete null hypothesis (i.e., the overall hypothesis that all the component null hypotheses are true). Consider once again the example of a family of two statistically independent component tests. Let  $\alpha_{FW} = .05$ . Then  $\alpha_T = .0253$ . In a non-multistage

procedure, the smaller of the two sample statistical values must still exceed the critical value based on  $\alpha_T = .0253$  even if the larger sample statistic results in the rejection of its component null hypothesis. So, in effect, a non-multistage multiple-test procedure is unfair to the smaller statistical value. A multiple-test procedure gives a more stringent critical value to allow for the possibility that all the component null hypotheses may be true, but even if some of those component null hypotheses are rejected, a non-multistage procedure retains the same stringent critical value. So a component significance test is penalized just because it happens to be grouped with  $m-1$  other component tests, even if those other component tests involve strongly significant effects. A multistage procedure, on the other hand, relaxes the critical values after at least one component null hypothesis has been rejected. Since for the complete null hypothesis, the rejection of any component null hypothesis is a Type I error, the probability of having zero Type I errors is the same as for a non-multistage procedure (since the procedure terminates if all component null hypotheses are accepted at the first stage). Therefore, the familywise Type I error rate is unchanged also (see equation (7)).

The fact that  $\alpha_{FW}$  is affected only by the first stage permits a great diversity in multistage procedures after their first stage. Consequently, there exists a

large variety of multistage procedures for multiple comparisons in analysis of variance.

Two criteria have been proposed for evaluating multistage procedures. One is Duncan's (1955) concept of a  $\underline{p}$ -mean significance level. This is designed to represent the Type I error rate at various stages of a multistage procedure in analysis of variance. Miller (1966) provides a good explanation of the details of  $\underline{p}$ -mean significance levels.

A second criterion for multistage tests is the maximum  $\alpha_{FW}$ , maximized over all possible combinations of true component null hypotheses (Tukey, 1953, cited by Ryan, 1959). Most multistage procedures control  $\alpha_{FW}$  only for the complete null hypothesis. For many multistage procedures,  $\alpha_{FW}$  can be much larger for other possible combinations of true component null hypotheses. For multiple comparisons in analysis of variance, keeping the maximum  $\alpha_{FW}$  at .05 is generally more conservative than keeping all  $\underline{p}$ -mean significance levels at .05.

### Families

The question of what constitutes a family of tests is an obviously important issue concerning multiple test procedures. Yet there are no set rules for what constitutes a family (Aitkin, 1971; Miller, 1966). It is on this issue that statisticians must leave mathematics and be guided by subjective judgment (Miller, 1966).



Ryan (1959) and Miller (1966) consider the experiment as the normal choice for a family of tests. Ryan adds that there should be strong specified reasons for any exceptions to this. Miller sees large experiments as an exception since if it were considered to be one family of tests there would be an unjustifiable loss in power. Others (Kirk, 1968; Wilson, 1962) favor the hypothesis as the unit for a family of tests. Researchers who ignore multiple test methods in their analyses are actually regarding each single test as the family. Miller (1966) points out that some justification can be given to this last position from a Bayesian viewpoint if the total loss for a sequence is the sum of the component losses. Different loss structures would yield different results. However, he does not consider any Bayesian approach a practical solution since it is almost impossible to specify a priori probabilities of Type I and Type II errors. Also the decision loss functions become quite unrealistic for practical applications such as data analysis (Plackett, 1971).

It is the opinion of this author that a useful distinction could be made between an alpha family and an analysis family. An alpha family is the set of tests which is being protected from one or more Type I errors at the reported alpha level. An analysis family is the set of tests which is being analyzed as a group. This is

what is called a family in the multiple-test literature. In many multiple-test procedures the alpha family and the analysis family are equivalent. The procedures in which they are different (e.g., Duncan's multiple range test) have been the source of additional confusion in this area.

In multistage procedures, the alpha family may change from stage to stage while the analysis family is the first stage's alpha family. In a non-multistage procedure, the alpha family and the analysis family are the same. The purpose of the concept of an alpha family is to clarify what is being protected from the occurrence of a Type I error at the reported alpha level.

The purpose of a research study affects what set of tests should be considered as an alpha family. For exploratory research it would be good to consider each individual test as an alpha family for discovering leads for future research (Miller, 1966; Petrinovich & Hardyck, 1969). Note that this is equivalent to the previous recommendation of using  $\alpha_T$  for exploratory research. When the results are to be used to support a particular theoretical position or are to be proclaimed to the scientific community as experimental "facts", then a larger alpha family should be used (Miller, 1966, Ryan, 1959). This is also where the issue of a priori vs. a posteriori analysis fits in. As Ryan (1959) points out, the central issue is the number of tests in the alpha family. However,

a priori analyses can be made more powerful by selecting to analyze only some of the possible tests. In a posteriori analyses, the alpha family must include all conceivable tests, not just those that look interesting as a result of the data (Williams, 1973).

### Specific Multiple Test Procedures

Most multiple-test procedures have been proposed for multiple comparisons in analysis of variance. Some of them are useful for calculating confidence intervals but are considered to be unnecessarily conservative for significance tests (Miller, 1966). These include Tukey's Honestly Significant Difference procedure, Scheffé's  $\underline{S}$  method, and a non-multistage Bonferroni  $\underline{t}$  method. Multistage procedures give better power for significance tests, but are not considered applicable to confidence intervals. Although many of them increase  $\alpha_{FW}$  beyond the specified level for some possible combinations of true component null hypotheses, this need not be the case. Such procedures include Ryan's (1960) Method of Adjusted Significance Levels, the Newman-Keuls procedure, Duncan's (1955) New Multiple Range Test, and Fisher's Least Significant Difference procedure, in decreasing order of conservativeness. Dunnett proposed a procedure for the special case in which one group is a control and other groups are to be compared with it but not with each other (Miller, 1966). Other approaches to the problem have included Bayesian

methods (Waller & Duncan, 1969), Simultaneous Test Procedures, which are closely related to Tukey's and Scheffé's methods (Gabriel, 1969), and subset selection procedures (Gupta & Panchapakesan, 1972).

Duncan's New Multiple Range Test and the Bonferroni  $\underline{t}$  method will be discussed further because Duncan's rationale is unique and because the Bonferroni  $\underline{t}$  is applicable to the case of tests of intercorrelations. Duncan's analysis family is different than his alpha family. He advocates increasing the familywise Type I error rate above the reported alpha level. He protects each possible statistically independent test at a .05 level (for example) and computes his overall protection level as  $1-(1-.05)^n$ , where  $n$  is the possible number of statistically independent comparisons. In a comparison among four means there are three statistically independent comparisons possible, so Duncan's  $\alpha_{FW}$  would be

$$1 - (1 - .05)^3 \approx .14 \quad (9)$$

although his reported  $\alpha$  would be .05. As  $n$  increases,  $\alpha_{FW}$  continues to increase rapidly. In general, Duncan's New Multiple Range Test is less conservative than the Newman-Keuls procedure and more conservative than the Least Significant Difference procedure. However, at the first stage of the multistage procedure (which is equivalent to

$\alpha_{FW}$  under the complete null hypothesis), Duncan's procedure is the least conservative of the three by far. His rationale for allowing  $\alpha_{FW}$  to increase seems to be (1) this gives increased power, while affording greater protection than that provided by non-multiple-test procedures, (2) this gives alpha levels consistent with a series of the possible independent tests among the means, and (3) this resembles a Bayesian solution with an additive loss function.

The Bonferroni  $\dagger$  method is apparently an old but little-used statistical tool. The first statistical user of the method is unknown (Miller, 1966). Fisher (1935) recommended its use for a posteriori  $t$ -tests. The name Bonferroni is connected with the probability inequality on which it is based,

$$1 - \alpha_{FW} \geq 1 - m \alpha_T \quad (10)$$

where  $m$  is the number of tests in the family of tests.

This reduces to

$$\frac{\alpha_{FW}}{m} \leq \alpha_T \quad (11)$$

and

$$\alpha_{FW} \leq m \alpha_T \quad (12)$$

If we want to keep  $\alpha_{FW}$  at .05 or less, we can give  $\alpha_{FW}$  a nominal value of .05, calculate  $\alpha_T$  according to the equality in equation (11), and we get the desired upper bound of .05 on  $\alpha_{FW}$  regardless of the dependence of the component tests. Actually this sets  $\alpha_{PF}$  at .05. Since  $\alpha_{PF} \geq \alpha_{FW}$  for a given  $\alpha_T$  (Ryan, 1960), the actual  $\alpha_{FW}$  is no more than .05. For example, for a family of four component significance tests, which need not be mutually independent, we can set  $\alpha_{FW}$  nominally at .05. The  $\alpha_T$  that we would use for each component test would be  $\alpha_T = .05/4 = .0125$ . By the Bonferroni inequality (10), the true  $\alpha_{FW}$  is less than or equal to  $4 \times .0125 = .05$ . Therefore, we can be certain that the true  $\alpha_{FW}$  is not greater than .05 by using  $\alpha_T = .0125$ . The Bonferroni inequality could be applied in the same way to many other statistics, but it is usually applied to the  $t$ -statistic for paired comparisons in analysis of variance, hence the name Bonferroni  $t$  method. For most purposes it has been found to approximate the nominal  $\alpha_{FW}$  very well (Dunn & Massey, 1965).

In order to compare multiple-test procedures, it is necessary to distinguish between the complete null hypothesis and other combinations of true component null hypotheses (Ryan, 1959). The complete null hypothesis occurs when all component null hypotheses are true.

Under the complete null hypothesis for pairwise

comparisons in analysis of variance, the multiple test procedures can be listed in the following order of decreasing conservativeness: Scheffé's  $\underline{S}$  method, the Bonferroni  $\underline{t}$  method, Tukey's HSD procedure, Fisher's protected Least Significant Difference method, the Newman-Keuls procedure, the Waller-Duncan Bayesian procedure, Duncan's New Multiple Range Test, and the use of multiple  $\underline{t}$  tests with unadjusted  $\alpha_T$  values (Boardman & Moffitt, 1971; Carmer & Swanson, 1973; Petrinovich & Hardyck, 1969). Tukey's, Fisher's, and the Newman-Keuls procedures have identical  $\alpha_{FW}$ 's under the complete null hypothesis and could be interchanged in this list. Under other combinations of true component null hypotheses, Fisher's LSD and the Bayesian method may approximate the Type I error rate of multiple  $\underline{t}$ -tests with unadjusted  $\alpha_T$  values (Carmer & Swanson, 1973), and the Newman-Keuls procedure may approach Duncan's Type I error rate (Petrinovich & Hardyck, 1969). Only Scheffé's  $\underline{S}$  Method, Tukey's HSD procedure, and the Bonferroni  $\underline{t}$  method keep  $\alpha_{FW}$  at about the nominal alpha level (usually .05) or below for all possible null hypothesis combinations. Miller (1966) calls these methods unnecessarily conservative for significance testing, but this can be corrected by a multistage modification such as suggested by Tukey (1953, cited by Ryan, 1959). Ryan's (1960) method of adjusted significance levels is actually a multistage version of the Bonferroni  $\underline{t}$  method.

Either of these multistage modifications increases the power over its non-multistage analog, but keeps the actual  $\alpha_{FW}$  at .05 or below for any null hypothesis combinations.

When methods are determinable for controlling  $\alpha_{FW}$  exactly, such methods are generally superior to the Bonferroni  $\underline{t}$  method. However, the Bonferroni  $\underline{t}$  may compete with Tukey's HSD procedure (Aitkin, 1971), at least when robustness is a critical issue (Miller, 1966; O'Neill & Wetherill, 1971). Whenever no exact methods are applicable, the Bonferroni  $\underline{t}$  is definitely a method to consider, and it is usually more powerful than alternative methods (F. B. Alt, personal communication, 1974; Christensen, 1973; Keselman, 1974). It is based on minimal assumptions and consequently can be applied to almost any situation (Miller, 1966). For example, it gives a conservative approximation to  $\alpha_{FW}$  when the component significance tests are not independent. This approximation is not too crude if  $m$  is not too large and if  $\alpha_T$  is small (Miller, 1966). The only complication is the need for critical values of the  $\underline{t}$  statistic at oddball values of  $\alpha_T$ . Miller describes three methods for interpolation from ordinary  $\underline{t}$  tables. Dunn & Massey (1965) have fairly adequate tables for the necessary critical values. Perlmutter & Myers (1973) state that the equation



$$\frac{t}{\frac{\alpha_{FW}}{m}}, \nu = \sum \frac{\alpha_{FW}}{m} + \frac{\sum^3 \frac{\alpha_{FW}}{m} + \sum \frac{\alpha_{FW}}{m}}{4(\nu - 2)} ; \nu, \text{ the degrees of (13) freedom}$$

can be used to calculate the necessary critical value for the  $t$  statistic from ordinary unit normal distribution ( $Z$ ) tables.

### Applications for Multiple Test Procedures

As previously noted, multiple-test procedures have been applied mostly to analysis of variance. Most of the same procedures are apparently applicable to comparisons between regression coefficients in linear multiple regression as long as the regressors are mutually independent (Dunnett, 1970; Williams, 1972). When the regressors are not independent, Scheffé's  $S$  method and the Bonferroni  $t$  method can be applied (Christensen, 1973). Multiple-test procedures have also been applied to other linear regression problems, including choosing among the possible regression functions (Spjøtvål, 1972), setting confidence intervals for points predicted by the regression equation, and setting confidence intervals for the regressor value which would be associated with a known criterion variable (Miller, 1966; O'Neill & Wetherill, 1971). Multiple-test procedures have also been proposed for certain non-parametric and multivariate analysis problems (Miller, 1966). Ryan (1960) presented multiple-test procedures for comparisons among medians, variances, or proportions and

a general method for comparisons among any statistics (the previously mentioned Method of Adjusted Significance Levels). Marascuilo (1966) presented large sample multiple-test procedures for comparisons among independent bivariate correlations, among parameters of independent binomial populations, among interaction measures in contingency tables, and among parameters of normal populations with unequal variances. Ryan (1959) also suggested that multiple-test procedures should be applied to multiple tests of intercorrelations, multiple variables in analysis of variance, replicated tests of a single hypothesis, and overlapping measures relating to a single hypothesis.

#### Some Related Aspects of Intercorrelations

There are two major situations in which psychological researchers are concerned with a set of tests regarding Pearson product-moment correlation coefficients. The first situation involves testing individual hypotheses about each of the intercorrelations in the correlation matrix  $R$  of  $k$  variables. The complete null hypothesis is

$$H_0 : R_p = I \quad . \quad (14)$$

The second situation is represented by testing all the correlations between  $k-1$  predictor variables and one criterion variable. In this second case all the correlations

in a  $(k-1) \times 1$  correlation vector  $\underline{r}$  are tested, which is equivalent to testing only the correlations in the first column of the matrix  $R_{mp}$  in the first situation. The complete null hypothesis in this case is

$$H_0 : \underline{r}_p = \underline{0} \quad (15)$$

Testing the correlations in equation (14) is equivalent to testing whether the covariances are zero. So equation (14) is equivalent to testing the null hypothesis that the covariance matrix is a diagonal matrix, i.e.,

$$H_0 : C = D \quad (16)$$

Testing a single correlation  $\rho_{il}$  in equation (15) is equivalent to testing the null hypothesis that the slope coefficient  $\beta_{il}$  is zero in an analysis in bivariate linear regression, i.e.,

$$H_0 : \beta_{il} = 0 \quad (17)$$

where  $\beta_{il}$  is the slope coefficient for predicting the first variable from the  $i^{\text{th}}$  variable.

As for tests of a single correlation coefficient, the  $t$  statistic for testing the hypothesis

$$\begin{aligned} H_0 &: \rho = 0 \\ H_1 &: \rho \neq 0 \end{aligned} \quad (18)$$

is a uniformly most powerful test among unbiased tests (UMPU; Kendall & Stuart, 1967). Similarly, for testing

$$\begin{aligned} H_0 &: \rho = \rho_0 \\ H_1 &: \rho \neq \rho_0 \end{aligned} \quad (19)$$

Fisher's  $r$  to  $Z$  transformation provides a simply calculated statistic which is a good approximation of the normal distribution even for fairly small samples (Cole, 1969).

Table 2. Type I Error Rates for Various Correlation Matrices

error rate	number of variables								
	2	3	4	5	6	7	8	9	10
$\alpha_{FW}$	.05	.14	.26	.40	.54	.66	.76	.84	.90
$\alpha_{PF}$	.05	.15	.30	.50	.75	1.05	1.40	1.80	2.25

Note. - Independent tests assumed,  $\alpha_T = .05$ .

However, when either or both of these statistics are used to test all the correlations in either a correlation matrix or a correlation vector, the probability of

observing at least one Type I error usually increases above the specified alpha level in a manner similar to multiple  $t$  tests with unadjusted  $\alpha_T$  values in multiple comparisons in analysis of variance. Even if the tests were mutually independent, the probability of observing at least one Type I error ( $\alpha_{FW}$ ) would increase rapidly for an increasing number of tests. Letting  $m$  represent the number of tests of correlation coefficients in either a matrix or a vector,  $\alpha_{FW}$  would exceed .50 for  $m = 14$ . Table 2 illustrates, for example, how  $\alpha_{FW}$  and  $\alpha_{PF}$  increase with increasingly larger correlation matrices.

The true  $\alpha_{FW}$  error rate is further complicated by the fact that the significance tests are not independent in general. For correlation matrices this is evident from the joint probability density function of the sample  $r_{ij}$ 's when the population correlation matrix  $R_p = I$ . This joint probability density function implies a zero probability for all sample  $R$ 's which are not positive definite (i.e., Gramian) (Cramér, 1946). The average of all intercorrelations among  $k$  variables cannot be less than  $-\frac{1}{k-1}$  (Hays, 1973). Therefore, the sample  $r_{ij}$ 's are not independent even when the variables themselves are independent.

The dependencies among the tests of the  $r_{ij}$ 's become more severe when  $R_p \neq I$ , which is precisely the case for which multiple-test procedures are most needed. This can be illustrated by the extreme case where  $R_p$



for any given sample. The direct implication of this is that either a Type I error will be made for each  $\rho_{ij} = 0$  (5% of the time, for  $\alpha_T = .05$ ) or no Type I error will occur for any  $\rho_{ij}$ .

While this exact case would not apply to any research, it is typical for a psychologist investigating the relationship between two concepts to use several highly correlated operational measures for each concept and to investigate the correlations between these two sets of variables. If the two concepts are statistically independent (assuming normality), the probability of simultaneously making Type I errors on most or all of the intercorrelations of interest would approach one in twenty (.05) for the typical  $\alpha_T$ . If it does happen to be that one occurrence in twenty, the researcher will be impressed by all the significant between-set correlations and will likely conclude that he has conclusively demonstrated a relationship between the two concepts. Later a journal editor will probably agree, and some psychological research will be wasted following up the erroneous conclusions. Actually, for this particular case a multivariate procedure such as canonical correlation (Mulaik, 1972) may be more appropriate. However, the main purpose of this example is to highlight some problems in significance tests of intercorrelations which apply also to more general cases for which multivariate procedures may not meet the re-

searcher's needs.

Ryan (1959) suggested that non-independence was not a critical factor in analysis of variance multiple comparisons, but added that such a conclusion would not necessarily apply to other applications of multiple-test procedures. It seems likely that non-independence may be a very critical factor in tests of intercorrelations and that more stringent Type I error rate controls, such as multiple-test procedures, may be necessary to take this into account.

The effect of non-independence seems to be critical also for correlation vectors. There are no restrictions on the sample correlations due to positive-definite-matrix restraints, but all the other sources of dependence previously discussed affect correlation vectors as well. In particular, if the correlation vector under consideration were the first column of the matrix in equation (20), there would be either no Type I errors or all possible Type I errors on any given sample. In realistic cases, with high positive correlations rather than correlations of +1 as in equation (20), this case is closely analogous to the case of multiple tests of slope coefficients in multiple linear regression among highly correlated regressors (Christensen, 1973).

Exact methods for controlling  $\alpha_{FW}$  are not generally determinable for either the correlation matrix or



the correlation vector case. Such methods require knowledge of the joint probabilities of Type I error, which are generally mathematically unobtainable in such statistical problems (Dunn & Massey, 1965; Miller, 1966).

However, some multiple-test procedures can be directly applied to tests of intercorrelations. The rationale for Fisher's Least Significant Difference (LSD) procedure can be applied to a correlation matrix, since overall tests are known for the hypothesis

$$\begin{aligned} H_0 : R_p &= I \\ H_1 : R_p &\neq I \end{aligned} \quad (21)$$

(e.g., Anderson, 1958). Just as the LSD procedure uses a preliminary overall  $F$  test, and when that is significant, proceeds to test all comparisons, a researcher could use a test of equation (21) followed by the usual  $t$ -tests when the overall null hypothesis is rejected. However, such a procedure would reduce Type I error rates very little when there existed a few large effects. For example, in the hypothetical example of equation (20), this procedure would not reduce the Type I error rate at all.

Other multiple test procedures which are applicable to tests of intercorrelations are the Bonferroni  $t$  method and apparently Scheffé's  $S$  method (Christensen, 1973; Miller, 1966). Both are approximate methods, general-

ly providing conservative estimates of  $\alpha_{FW}$ . Scheffé's method is conservative because it controls  $\alpha_{FW}$  for the complete set of linear combinations of the component tests, while only a few of these are of interest. The Bonferroni  $\frac{1}{2}$  method is conservative because it controls for  $\alpha_{PF}$  which is always a conservative estimate of  $\alpha_{FW}$ . Since both methods overcontrol for  $\alpha_{FW}$ , the preferred method would be the more powerful one. Christensen (1973) compared the power of the two methods for the closely related problem of hypothesis tests of the slope coefficients in multiple linear regression when the regressors are correlated. He concluded that the Bonferroni method always resulted in more powerful individual

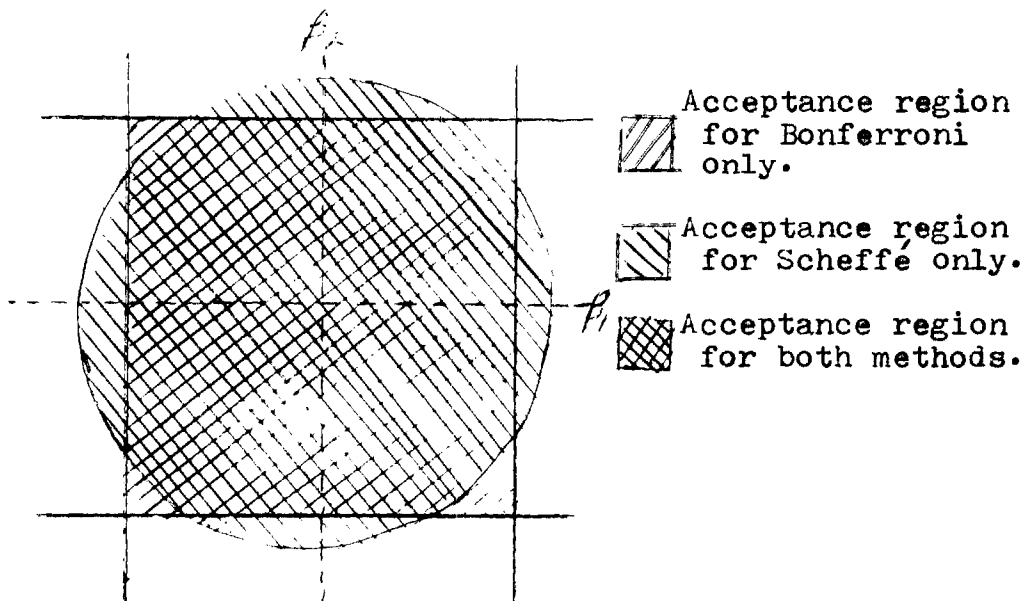



Figure 2. Null Hypothesis Acceptance Regions for the Bonferroni and Scheffé Methods (after Miller, 1966)

tests (i.e., less stringent  $\alpha_T$ 's) with  $\alpha_{FW} = .05$ . Although he showed that the power of the joint hypothesis favored Scheffé's method in most cases, he only considered the case of two regressors. Figure 2 illustrates the null hypothesis acceptance region when Scheffé's method is more powerful for the joint hypothesis (which is sometimes true) and the Bonferroni method is more powerful for individual null hypotheses (which is always true with  $\alpha_{FW} = .05$ ). All statistical values which would result in acceptance of the alternative hypothesis by Scheffé's method but not by the Bonferroni  $\dagger$  method fall in the regions marked . For such statistical values, the researchers would reject  $H_0 : \beta_1 = \beta_2 = 0$ , but would accept  $H_0 : \beta_1 = 0$  and  $H_0 : \beta_2 = 0$ . So even when power could be gained by Scheffé's method for the joint hypothesis, the power gained is only an advantage for results not interpretable in terms of the individual tests. So Scheffé's method is probably not a competitor with the Bonferroni method for multiple tests of intercorrelations of the kind considered in this thesis (i.e., tests of equation (1)).

The Bonferroni  $\dagger$  test can also be modified into a multistage procedure to increase power still further. Ryan (1960) describes such a method for analysis of variance. For testing intercorrelations, the alpha family could be reduced to the number of non-rejected component

null hypotheses after the first stage. In the second stage, a new  $\alpha_T$  could be calculated by using this new number for  $m$  in equation (11). If any more component null hypotheses were rejected, a new value of  $m$  would be used in a similar way for a third stage. This would continue until a stage is reached with no new rejections of component null hypotheses, whereupon the procedure terminates. Such a procedure would increase power but keep  $\alpha_{PF}$  at .05 (and therefore  $\alpha_{FW} \leq .05$ ) for any combination of true component null hypotheses (see Appendix A).

#### Hypotheses and Objectives of Study

The literature review covered so far has emphasized previous studies closely related to the problem of controlling Type I error adequately in multiple significance tests of intercorrelations. It is apparent that few investigations have been done on multiple tests of intercorrelations themselves (Harris, 1967). The present study investigated multiple tests of intercorrelations by means of an empirical study of Type I error rates in such multiple tests for which the population correlation matrices were known.

Three methods were used for controlling Type I error. Method I was the customary procedure of setting  $\alpha_{T(I)} = .05$  for all the tests. Method II calculated a more conservative  $\alpha_{T(II)}$  such that  $\alpha_{FW}$  would be .05 if all the individual significance tests were mutually in-

dependent, which they are not. Method III was the Bonferroni  $t$  test with  $\alpha_{T(III)}$  calculated so that  $\alpha_{FW} \leq .05$ . Dunn & Massey (1965) suggested Methods II and III as approximate methods for controlling familywise Type I error rates for multiple- $t$  tests.

In the empirical investigation, these three methods were examined for various cases of the number of variables ( $k$ ), sample size ( $N$ ), and population correlation parameters ( $R_{n \times p}$ ).

The major dependent variables in the study were familywise Type I error rate and conditional Type I error rate. For familywise Type I error rate, the family of tests was the set of all  $m$  tests in the correlation matrix, i.e.,

$$m = \frac{k(k-1)}{2}, \quad (22)$$

or the set of all  $m$  tests in the first column, i.e.,

$$m = k - 1. \quad (23)$$

Conditional Type I error rate here means the Type I error rate of one component test given that a Type I error has occurred on one other component test in the same family of tests. For example, in the hypothetical example

of equation (20), the conditional Type I error rate would be 1.0 (i.e., 100%). That is, whenever a Type I error occurs on any one individual significance test, Type I errors will always occur on any other test for which  $\rho_{ij}$  is actually zero. The empirical study investigated conditional Type I error rate under more realistic conditions.

The effect of Methods I, II, and III on statistical power was also investigated, but it was expected that this would merely reflect the differences in familywise Type I error rate. As Games (1971) has pointed out, when the same statistic is used for different procedures which vary only the critical value, any "reduction in risk of Type I error is paid for by an increase in the risk of Type II error (p. 101)."

The major hypotheses were as follows:

(1) For Method I,  $\alpha_{FW}$  is significantly larger than the nominal value of .05 for  $k \geq 3$ . Table 2 showed that the value of  $\alpha_{FW}$  increases above .05 for  $k \geq 3$  under the assumption that the component significance tests are mutually independent. Although this assumption does not actually apply to multiple tests of intercorrelations, it can be shown mathematically that this hypothesis holds for the general case. In a sense, this hypothesis is trivial, but its implications have not had any effect on procedures for statistical analyses of intercorrelations. So the primary purpose of this hypothesis is to highlight the effect of Method I on  $\alpha_{FW}$ .

(2) Familywise Type I error rate ( $\alpha_{FW}$ ) is not significantly greater than the nominal value of .05 for Methods II and III. As discussed previously, the Bonferroni  $\frac{\alpha}{k}$  always provides a conservative estimate of the nominal alpha level. Dunn & Massey (1965) conjectured that Method II was also conservative for most families of significance tests.

(3) As the number of variables gets larger,  $\alpha_{FW}$  for Method III becomes significantly less than the nominal value of .05. Miller (1966) reported that the Bonferroni  $\frac{\alpha}{k}$  procedure provides an adequate approximation of the nominal  $\alpha_{FW}$  for small  $\alpha_{FW}$  and a small number of component tests. As the number of variables increases the number of component tests increases and Method III may not provide such a good approximation.

(4) The conditional Type I error rate is greater than  $\alpha_T$ . That is, when a Type I error occurs on one correlation test, then the Type I error rate is increased for other component tests in that family. If the component tests were mutually independent, the conditional Type I error rate would equal  $\alpha_T$ . However, the component tests in this case are not mutually independent and this is expected to result in higher conditional Type I error rates.

(5) The effect hypothesized in hypothesis (4) is especially strong when some of the variables of interest are highly correlated. In the discussion of the hypothe-

tical population correlation matrix equation (20), it was noted that the conditional Type I error rate would be 1.00,, while  $\alpha_{\text{T}}$  would be .05. A similar phenomenon (although less striking) is expected for more realistic correlation matrices with some reasonably high correlations.

(6) In general, the Type II error rates reflect the differences in  $\alpha_{\text{T}}$ . When the critical value is set to allow more Type I errors, then the Type II error rate decreases. Whatever is gained in Type I error rate is gained at the expense of Type II error rate.



## CHAPTER II

## METHOD

Monte Carlo Method

A Monte Carlo method generally involves a computer simulation in which samples are randomly drawn from a hypothetical population to evaluate a particular method of statistical analysis (Halton, 1970). Such a method is especially appropriate when it can aid the researcher in selecting appropriate statistical procedures where the necessary theoretical information is incomplete (Cole, 1969). The theoretical information for multiple tests of intercorrelations is incomplete because the determination of exact critical values for multivariate  $t$  distributions depends on many nuisance parameters which are generally not known beforehand by the researcher (Dunn & Massey, 1965). Therefore, an empirical Monte Carlo investigation of multiple tests of intercorrelations was considered appropriate. Such Monte Carlo investigations have been made for other questions concerning multiple test procedures (Boardman & Moffitt, 1971; Carmer & Swanson, 1973; Keselman, 1974; Keselman & Toothaker, 1973; Petrinovich & Hardyck, 1969; Smith, 1971). In this present study sample correlation matrices were computed from samples of scores randomly drawn from multivariate normal

populations with specified population correlation matrices.

Generation of Independent Random Normal Deviates

Generation of Uniformly Distributed Random Numbers

Review of the Literature. The foundation of the Monte Carlo method is the pseudorandom number generator. The term "pseudorandom" indicates that the numbers generated are not actually random. A pseudorandom number generator gives the same sequence of numbers every time unless one of the starting values is changed. However, such generators have been preferred over random number generators which are not deterministic (such as a set of dice), because the latter generators are nonrepeatable, slower, often unstable, and need to be tested frequently for randomness. Some of these disadvantages may be removed by recently developed non-deterministic generators (e.g., Cohn, 1971; Maddocks, Matthews, Walker, & Vincent, 1972; Murry, 1970), but such methods have not been widely proven and necessitate equipment which is often unavailable.

Many pseudorandom number generators have the advantages of rapid number generation, small computer storage requirements, and repeatable sequences. If the method and the starting values are carefully selected, a pseudorandom number generator can provide an adequate simulation of random numbers for most applications. Halton (1970) states that true randomness cannot be evaluated, anyway.

If a sequence behaves randomly with respect to any number of tests of randomness, it is generally impossible to be sure that it would not miserably fail another test of randomness (Knuth, 1969). This would be quite limiting for any random number generator, except that only a few properties of randomness are usually required. Pseudorandom number generators can be designed so that the generated number sequence will pass most ordinary tests of randomness.

The multiplicative pseudorandom number generator, originally due to Lehmer (1949, cited by Dieter, 1971), computes pseudorandom numbers  $x_1, x_2, \dots, x_i, x_{i+1}, \dots$  successively by the equation

$$x_{i+1} = a x_i \pmod{m} \quad (24)$$

where  $a$  is a multiplier and  $\text{mod } m$  denotes modular arithmetic. Modular arithmetic involves first performing the arithmetic normally (e.g.,  $a x_i$ ) and then subtracting the largest possible integer multiple of  $m$ . For example, in modular arithmetic

$$4 \pmod{3} = 1,$$

$$8 \pmod{3} = 2$$

and

$$2 \times 5(\text{mod } 4) = 2 \quad .$$

The multiplicative congruential pseudorandom number generator (equation (24)) is considered by many to be the most successful pseudorandom number generator (e.g., Coveyou & MacPherson, 1967; Knuth, 1969). It is supported by the literature on number theory (Keuhl, 1969), it passes most tests of statistical performance (Dieter, 1971; Jansson, 1966), and it is fast and easy to program (Dieter, 1971). However, there are some sequences of  $p$  numbers ( $p > 1$ ) which can never be sampled in a pseudorandom number sequence for a given multiplicative generator (Coveyou & MacPherson, 1967; Marsaglia, 1968, 1970). Marsaglia's alternative, the combined congruential method, has not proven any better in some direct comparisons (Brown & Rowland, 1970; Seawright, Larkin, & Locks, 1966) and takes about twice as much computer time. As Knuth (1969) has illustrated, merely designing a more complex pseudorandom number algorithm apart from theoretical considerations often results in a poorer simulation of random numbers. So a multiplicative congruential pseudorandom number generator was chosen for this study. The combined congruential generator may prove to be superior in the future after knowledge about proper selection of parameters becomes more complete. But currently more is known about selection of parameters and advantages of the

multiplicative congruential generator as well as more about its disadvantages.

A good multiplicative generator must have properly chosen parameters (Coveyou & MacPherson, 1967; Jansson, 1966). The multiplier

$$a \approx \frac{m}{8} (\sqrt{5} - 1) \quad (25)$$

recommended by Ahrens, Dieter, & Grube (1970; also Dieter & Ahrens, 1971) is reported to result in a sequence of numbers best approximating independent numbers. In order to provide the longest possible period of the pseudorandom number sequence before it repeats itself,  $a$  must be either

$$a = 3 \pmod{8} \quad (26)$$

or

$$a = 5 \pmod{8} \quad (27)$$

and the starting number  $x_0$  must be odd (Dieter & Ahrens, 1971).

Generator Used. In the present study, a multiplicative pseudorandom number generator

$$x_{i+1} = 5308871541 x_i \pmod{2^{35}} \quad (28)$$

was used. The modulus  $2^{35}$  is the word size on the UNIVAC 1108 computer. The multiplier is the one recommended by Ahrens et al (1970), calculated by equations (25) and (27). A machine-language subroutine written for the UNIVAC 1108 (Math-Pack, 1970) was used to generate the numbers according to equation (28) for this study. This subroutine generated pseudorandom numbers uniformly distributed on the interval  $(0, 2^{35})$ . This was transformed to a uniform distribution on  $(0,1)$  in subroutine NORGEN (on file in the School of Psychology).

$$U_i = \frac{x_i}{2^{35}} \quad . \quad (29)$$

### Transformation to Normally Distributed Random Numbers

Review of the Literature. Box and Muller (1958; Muller, 1959) developed a method of transforming uniform random numbers to random normal deviates which has an accuracy limited only by the accuracy of a few available computer library programs. Letting  $U_i$  and  $U_{i+1}$  be two independent random variables from a uniform distribution on  $(0,1)$ , they showed that

$$X_i = \sqrt{-2 \ln U_1} \cos 2 \pi U_2 \quad (30)$$

and

$$X_{i+1} = \sqrt{-2 \ln U_1} \sin 2 \pi U_2 \quad (31)$$

are independent random deviates. Muller (1959) demonstrated that this method gave better accuracy and comparable speed with respect to other such methods of normal deviate transformations which were known at that time. Later Marsaglia and Bray (1964) improved the Box-Muller method with the equations

$$Z_i = V_j \left[ \frac{-2 \ln (V_j^2 + V_{j+1}^2)}{V_j^2 + V_{j+1}^2} \right]^{\frac{1}{2}} \quad (32)$$

and

$$Z_{i+1} = V_{j+1} \left[ \frac{-2 \ln (V_j^2 + V_{j+1}^2)}{V_j^2 + V_{j+1}^2} \right]^{\frac{1}{2}} \quad (33)$$

where  $V_j$  and  $V_{j+1}$  are uniform on  $(-1, 1)$ , conditioned by  $V_j^2 + V_{j+1}^2 < 1$ . This method is faster on a computer than equations (30) and (31) and just as accurate.

Marsaglia and his associates have also designed methods for the transformation to the normal distribution which are much faster and just as accurate, although they take more computer space and are more difficult to program (Marsaglia & Bray, 1964; Marsaglia, MacLaren, & Bray, 1964).

Neave (1973) reported an unsatisfactory attempt to

use the Box-Muller transformation together with a multiplicative pseudorandom number generator. He reported several local maxima and tails truncated at  $-3.3\sigma$  and  $+3.6\sigma$  in the generated distribution. Although his problem was caused partly by setting the multiplier too small in his pseudorandom number generator, he pointed out that equations (32) and (33) would also correct the problem.

Normal Distribution Transformation Used. This study used equations (32) and (33), Marsaglia and Bray's (1964) form of the Box-Muller (1958) transformation. The interval of the uniform distribution was first changed to  $(-1, 1)$  from the  $(0, 1)$  interval obtained in equation (29) as follows:

$$V_i = 2 (U_i - \frac{1}{2}) \quad (34)$$

Then equations (32) and (33) were applied to successive pairs of  $V_i$  that met the condition  $V_i^2 + V_{i+1}^2 < 1$  until a  $k \times N$  matrix  $\underline{Z}$  of  $N$  observations on  $k$  independent variables was complete.

#### Statistical Tests of Pseudorandom Normal Numbers

The computer generation of normally distributed random numbers was tested in three ways. Two of these tests were statistical tests of the independent pseudorandom normal numbers ( $\underline{Z}_i$ ) and are discussed in this



section. The first test was the Pearson  $\chi^2$  test of fit to an 8-variate mutually independent normal distribution. For this test each marginal univariate normal distribution was divided into three intervals of equal theoretical probability. This resulted in  $3^8 = 6561$  cells in the 8-variate joint probability distribution. The expected frequency for any of these cells in a given computer run was

$$f_{\text{exp}(i)} = \frac{1000 N}{6561} \quad (35)$$

where  $N$  is the sample size for that run. Pearson's chi-square statistic

$$\chi^2 = \sum_{i=1}^{6561} \frac{(f_{\text{exp}(i)} - f_{\text{obs}(i)})^2}{f_{\text{exp}(i)}} \quad (36)$$

was computed. Since the degrees of freedom were so large (6560), a direct reference to a computer library subroutine or to a  $\chi^2$  table was impossible. Therefore, a normal approximation to the  $\chi^2$  distribution,

$$Z = \sqrt{2\chi^2 - \nu} - \sqrt{2\nu - 1} \quad , \quad (37)$$

was used, where  $\nu$  is the degrees of freedom. This normal approximation is considered adequate with  $\nu > 100$  (Hays, 1973).

The second test was a test-of-fit to a bivariate normal distribution of two independent variables. This was similar to the previous test except that the random deviates for only two variables were considered, those corresponding to the first variable and one other specified variable. For this test, each marginal univariate normal distribution was divided into 50 intervals of equal probability. This resulted in  $50^2 = 2500$  cells in the bivariate normal joint probability distribution, each with an expected frequency of

$$f_{\text{exp}(i)} = \frac{1000 N}{2500} \quad . \quad (38)$$

The normal approximation to the chi-square distribution was used as in the first test, except with a summation limit of 2500 in equation (36) and 2499 degrees of freedom in equation (37).

For both of these statistical tests, the probability of obtaining a  $\chi^2$  greater than the one observed was reported. For different runs of the computer program these reported probabilities should vary somewhat over the range (0, 1). If all the  $\chi^2$  values are small this could indicate that the pseudorandom normal numbers are not random enough. A large proportion of large  $\chi^2$  values would indicate nonnormality or non-independence. If the  $\chi^2$  values are neither too large nor too small the pro-

bability of obtaining a  $\chi^2$  greater than that observed should be less than .05 about once in twenty computer runs. If such small probabilities occurred significantly ( $p < .05$ ) more often according to a two-tailed binomial test, then it would have been considered that the normal numbers deviated significantly from a distribution of independent normal numbers. This two-tailed binomial test was based on a binomial distribution with  $p = .05$  and  $n$  equal to the number of tests of fit of each type (bivariate or 8-variate).

#### Generation of Sample Correlation Matrices

The multiplicative pseudorandom number generator and the normal distribution transformation were used to produce a  $k \times N$  matrix  $\underline{Z}$  of  $N$  independent observations on  $k$  independent (uncorrelated) normal variables. For certain cases in this study,  $\underline{Z}$  was transformed to a  $k \times N$  matrix  $\underline{Y}$  of  $N$  independent observations on  $k$  multivariate normal variables with specified population correlations.

#### Review of the Literature

Recently, Barr and Slezak (1972) and Oplinger (1971) both evaluated methods of transforming a matrix  $\underline{Z}$  of uncorrelated multivariate normal scores to a matrix  $\underline{Y}$  of multivariate normal scores with a desired population covariance matrix  $\underline{C}_p$ . Both studies investigated the same

three methods, the conditional density function approach, the similarity transformation technique (also called the rotation method), and the general recursive method (also called the triangular factorization method). Both concluded that the general recursive method was the preferred method. Barr and Slezak demonstrated that it was faster and took less computer space than the other methods.

The general recursive method transforms the matrix  $Z$  to correlated data  $Y$  by

$$Y = A Z \quad (39)$$

where  $A$  is a lower triangular matrix such that

$$A A^T = R_p \quad (40)$$

The matrix  $A$  is calculated by recursive equations which turn out to be identical to the equations for the square root method of linear algebra (Capra & Elster, 1971; Oplinger, 1971; Scheuer & Stoller, 1962) and for the Cholesky method of factor analysis (Harman, 1960, 38-41; Mulaik, 1972, 108-109). Wold (1948) was apparently the first to recognize such an application of this system of equations. Scheuer & Stoller have presented a fairly general discussion of it.

A fourth method, not considered by Barr and Slezak



Therefore, the recursive equations, which are stated generally in terms of the elements of  $\underline{C}_Y$  for the general case could be restated in terms of  $\underline{R}_p$  for this particular case. The resulting equations for the elements of the  $k \times k$  matrix  $\underline{A}$  (in equation (40)) were

$$a_{11} = \rho_{11} = 1 \quad (44)$$

$$a_{i1} = \frac{\rho_{1i}}{a_{11}} = \rho_{1i} ; i = 2, 3, \dots, k \quad (45)$$

$$a_{jj} = \sqrt{\rho_{jj} - \sum_{l=1}^{j-1} a_{jl}^2} ; j = 2, 3, \dots, k \quad (46)$$

and

$$a_{ij} = \frac{\rho_{ji} - \sum_{l=1}^{j-1} a_{jl} a_{il}}{a_{jj}} ; \begin{matrix} i = 3, 4, \dots, k \\ j = 2, 3, \dots, i-1 \end{matrix} . (47)$$

These equations were used in subroutine ACOMP (on file in the School of Psychology) to calculate  $\underline{A}$  from the specified  $\underline{R}_p$  for each computer run. Then for each of the 1000

sample correlation matrices in a computer run, a matrix  $Z$  was generated which was then transformed to a matrix  $Y$  of correlated observations by equation (39). This matrix  $Y$  represented  $N$  subjects' scores on  $k$  variables with the intercorrelations specified in  $R_p$ . The sample correlation matrix  $R$  was computed from  $Y$  in the usual way,

$$R = D_Y^{-\frac{1}{2}} Y Y^T D_Y^{-\frac{1}{2}} \quad (48)$$

where  $D_Y^{-\frac{1}{2}}$  is a diagonal matrix consisting of elements  $d_{ii}$  such that

$$d_{ii} = \frac{1}{s_i} \quad (49)$$

where  $s_i$  is the sample estimate of the standard deviation of the  $i^{\text{th}}$  variable.

#### Additional Statistical Tests of Pseudorandom Numbers

Shreider (1966) has stated that "the quality of pseudorandom numbers may also be investigated by means of a model problem for which the exact solution is known (p. 334)." This seemed to be appropriate in this study since, as previously discussed, no sequence of numbers can definitely be considered random in every sense. Also Coveyou and MacPherson (1967) and Marsaglia (1968, 1970) have shown that the pseudorandom number sequence of the type used in this study is never random in at least one

particular sense. However, Halton (1970) pointed out that only a few properties of randomness are of interest for any particular application. Shreider's suggestion seemed to be the best way to investigate those properties of randomness (some of which are unknown) which are of interest for this application.

Consequently, a method was used to test the pseudo-random number generator by means of a model problem. The model problem was a significance test of one correlation coefficient between two uncorrelated variables. This problem was simulated by specifying two independent variables (i.e.,  $k = 2$  and  $R_p = I$ ) for this test run of the computer program. The program then generated 1000 sample  $2 \times 2$  correlation matrices, which only had one unique inter-correlation coefficient in it,  $r_{12}$ . Since the  $t$ -test is an exact test of this null hypothesis and no multiple test considerations are involved, it is known that Type I errors should occur on about 50 out of 1000 independent trials when  $\alpha_T = .05$ . The observed number of Type I errors was compared against this expected number using a two-tailed Poisson test ( $\lambda = 50$ ). The observed number of test runs with number of Type I errors significantly different than 50 ( $p < .05$ ) was compared with the binomial distribution with  $p = .05$  and  $n$  equal to the total number of test runs. This evaluated whether the number of test runs with a significantly different number of Type I



errors could be reasonably explained by random variation alone.

### Tests of Hypotheses of This Study

#### Populations Studied

All populations in this study from which samples were drawn were multivariate normal populations with mean vector  $\mu_Y$  and covariance matrix  $C_{\mu_Y}$  (identical to  $R_p$ ) as specified in equations (41) and (42). Nine cases were sampled in which the complete null hypothesis was true, i.e.,

$$R_p = I \quad (50)$$

For three cases each, the number of variables ( $k$ ) was 3, 4, or 8. For each  $k$ , sample size ( $N$ ) was 15, 40, or 100 for one case each.

Other cases were selected mostly to provide realistic analogues of equation (20). The following population correlation matrices were used:

$$R_p = \begin{bmatrix} 1 & & \\ 0 & 1 & \\ 0 & .3 & 1 \end{bmatrix} \quad (51)$$

$$R_p = \begin{bmatrix} 1 & & \\ 0 & 1 & \\ 0 & .6 & 1 \end{bmatrix} \quad (52)$$

$$R_p = \begin{bmatrix} 1 & & & \\ 0 & 1 & & \\ 0 & -.6 & 1 & \\ & & & 1 \end{bmatrix} \quad (53)$$

$$R_p = \begin{bmatrix} 1 & & & & \\ 0 & 1 & & & \\ 0 & .6 & 1 & & \\ 0 & .6 & .6 & 1 & \\ & & & & 1 \end{bmatrix} \quad (54)$$

$$R_p = \begin{bmatrix} 1 & & & & \\ .6 & 1 & & & \\ 0 & 0 & 1 & & \\ 0 & 0 & 0 & 1 & \\ 0 & 0 & .6 & 1 & \end{bmatrix} \quad (55)$$

$$R_p = \begin{bmatrix} 1 & & & & & \\ .6 & 1 & & & & \\ 0 & 0 & 1 & & & \\ 0 & 0 & .6 & 1 & & \\ 0 & 0 & 0 & 0 & 1 & \\ 0 & 0 & 0 & 0 & .6 & 1 \end{bmatrix} \quad (56)$$

$$R_p = \begin{bmatrix} 1 & & & & & \\ 0 & 1 & & & & \\ 0 & 0 & 1 & & & \\ 0 & 0 & 0 & 1 & & \\ .6 & 0 & 0 & 0 & 1 & \\ .6 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad (57)$$

$$R_p = \begin{bmatrix} 1 & & & & & \\ .6 & 1 & & & & \\ .6 & .6 & 1 & & & \\ 0 & 0 & 0 & 1 & & \\ 0 & 0 & 0 & .6 & 1 & \\ 0 & 0 & 0 & .6 & .6 & 1 \end{bmatrix} \quad (58)$$

For each 3-variable  $R_p$ , sample correlation matrices were generated using a sample size of 15, 40, or 100 respectively for each of three cases. For larger  $k$ , sample sizes

of 40 and 100 were investigated for each  $R_{sp}$ .

### Flow of the Computer Program

The population correlation matrix and the sample size were specified for each run of the computer program. A series of tests of hypotheses of the form

$$\begin{aligned} H_0 : \rho_{ij} &= 0 \\ H_1 : \rho_{ij} &\neq 0 \end{aligned} \quad (59)$$

were performed on all the sample correlations in 1000 sample correlation matrices.

Three methods were used to calculate the critical value for the hypothesis tests. Method I was the customary procedure of setting  $\alpha_{T(I)} = .05$  for all the tests.

Method II calculated  $\alpha_{T(II)}$  according to the equation

$$\alpha_{T(II)} = 1 - \sqrt[m]{1 - .05} \quad (60)$$

where  $m$  is the number of significance tests in the matrix or vector. This equation sets the  $\alpha_{T(II)}$  rate correctly for  $\alpha_{FW} = .05$  under the assumption that all significance tests in the family are mutually independent. It is realized that this assumption is not tenable in the inter-correlation case, but it provides an easily calculated  $\alpha_{T(II)}$  which seems to control  $\alpha_{FW}$  more appropriately

than Method I.

Method III was the Bonferroni  $\dagger$  method discussed previously. For this method

$$\alpha_{T(III)} = \frac{.05}{m} \quad . \quad (61)$$

This always yields a conservative estimate of  $\alpha_{FW}$ . Dunn and Massey (1965) have previously investigated Methods II and III and found them to be adequate for most cases of equicorrelated multiple  $\dagger$  tests.

The single-test alpha levels  $\alpha_{T(I)}$ ,  $\alpha_{T(II)}$ , and  $\alpha_{T(III)}$  were calculated once considering the entire correlation matrix as the family of tests, with

$$m = \frac{k(k-1)}{2} \quad , \quad (62)$$

and once regarding the first column as the family of tests, with

$$m = k-1 \quad . \quad (63)$$

This resulted in six distinct single-test alpha levels, which were designated  $\alpha_{T(I)M}$ ,  $\alpha_{T(II)M}$ ,  $\alpha_{T(III)M}$ ,  $\alpha_{T(I)V}$ ,  $\alpha_{T(II)V}$ , and  $\alpha_{T(III)V}$  to distinguish between a matrix family of intercorrelation tests and a vector

(first column) family of intercorrelation tests (Note that  $\alpha_{T(I)M} = \alpha_{T(I)V}$ , but the other single-test alpha levels are generally numerically different). Corresponding to the six single-test alpha levels were six distinct critical values for  $r_{ij}$ .

One thousand sample correlation matrices were generated. For each matrix, Type I and Type II errors were counted according to all six critical values for the first column and according to the three applicable critical values for the rest of the matrix. After all matrices were generated and analyzed, tables were printed summarizing the frequency of Type I and Type II errors. An example of these tables is printed in Appendix B.

Options were available in the program to test the pseudorandom normal independent numbers by means of a Pearson  $\chi^2$  test-of-fit to a multivariate normal distribution.

#### Methods of Analysis of Matrices of Correlations

Type I Error Rate: Complete Null Hypothesis. For each computer run with its particular case of a given sample size  $N$ , a given number of variables  $k$ , and  $R_{n \times p} = I_n$ , empirical size was evaluated in several ways. For Hypotheses 1, 2, and 3, the familywise empirical size was evaluated for Methods I, II, and III with the family of tests including a component test for each intercorrelation in the matrix. Familywise empirical size is the proportion of

the 1000 sample correlation matrices in which one or more Type I errors occur.

Hypothesis 1, that the familywise Type I error rate is larger than .05 for Method I with  $k \geq 3$ , was evaluated by testing the statistical hypothesis

$$\begin{aligned} H_0 : \alpha_{FW(I)M} &= .05 \\ H_1 : \alpha_{FW(I)M} &> .05 \end{aligned} \quad (64)$$

where  $\alpha_{FW(I)M}$  is the familywise Type I error rate for the entire matrix according to Method I. The binomial distribution was used for this test since the 1000 trials were considered to be 1000 independent replications of the experiment. In such a case the binomial distribution assigns probabilities with each possible number of observed trials having one or more Type I errors, with  $p = .05$ . Because of the large number of trials (1000), a Poisson approximation to the binomial distribution

$$1 - P(x; \lambda) = 1 - \sum_{i=0}^x \frac{e^{-\lambda} \lambda^i}{i!} \quad (65)$$

$$\lambda = 1000 \times .05 = 50$$

$$x = 1000 \times \text{familywise empirical size}$$

was used to determine the probability of obtaining a larger familywise empirical size than that observed, given an actual  $\alpha_{FW}$  of .05. If the familywise empirical size

was significantly ( $p < .05$ ) more than an  $\alpha_{FW}$  of .05, the null hypothesis of equation (64) would have been rejected, and it would have been concluded that  $\alpha_{FW(I)M}$  is actually greater than .05.

The second hypothesis was that  $\alpha_{FW(II)}$  and  $\alpha_{FW(III)}$  are not greater than the nominal  $\alpha_{FW}$  of .05. This was evaluated by testing the statistical hypotheses

$$\begin{aligned} H_0 : \alpha_{FW(II)M} &= .05 \\ H_1 : \alpha_{FW(II)M} &> .05 \end{aligned} \quad (66)$$

and

$$\begin{aligned} H_0 : \alpha_{FW(III)M} &= .05 \\ H_1 : \alpha_{FW(III)M} &\neq .05 \end{aligned} \quad (67)$$

The Poisson approximation to the binomial distribution (equation (65)) was also used for these statistical analyses as above. A two-tailed alternative hypothesis was used in equation (67) because one tail is of interest for the second hypothesis and the other tail for the third hypothesis, that as the number of variables gets larger,  $\alpha_{FW}$  for Method III becomes significantly less than .05.

The fourth hypothesis was that the conditional Type I error rate is greater than  $\alpha_T$ . Two analyses were

performed concerning this hypothesis. The first analysis was a test of fit to a binomial distribution of the frequencies of each possible number of Type I errors for the 1000 sample correlation matrices. The family of component tests on the  $m$  intercorrelations in one sample correlation matrix may result in 0, 1, 2, 3, ..., or  $m$  Type I errors. If the component tests are mutually independent, a binomial distribution with  $n = m$  and  $p = .05$  should describe the probabilities of each possible number of Type I errors. From this, the expected frequencies of each possible number of Type I errors can be calculated for a total frequency of 1000 sample correlation matrices. If the observed frequencies of each possible number of Type I errors for the 1000 sample correlations differ significantly ( $p < .05$ ) from the expected frequencies, then the conclusion would have been made that the component tests could not be considered mutually independent. If the component tests are not mutually independent, then the conditional Type I error rate differs from the unconditional Type I error rate ( $\alpha_T$ ). A Pearson  $\chi^2$  statistic was used for the test of fit. The cells for the larger numbers of Type I errors were lumped together so that each cell had an expected frequency of five or more for the test of fit.

The second analysis of conditional Type I error rate was conditional empirical size. For this analysis, a component test such as



$$\begin{aligned} H_0 : \rho_{12} &= 0 \\ H_1 : \rho_{12} &\neq 0 \end{aligned} \quad (68)$$

was selected for the conditional empirical size to be conditional on. The conditional empirical size for any other component test was the proportion among the sampling replications with Type I errors on component test (68) which also had Type I errors on the other component test. For example, if 50 of the 1000 sampling replications had a Type I error on component test (68) and there was also a Type I error on the component test of  $\rho_{13}$  on 15 of those 50 replications, then the conditional empirical size for  $\rho_{13}$  would be  $\frac{15}{50} = .30$ . The conditional empirical size was calculated for all component tests (except component test (68)) in this way. A one-tailed Poisson test was used to evaluate whether each conditional empirical size was significantly more than .05. In the above example the expected number of Type I errors on  $\rho_{13}$  (among the 50 replications) would be 2.5. Therefore, the observed number of replications with Type I errors on  $\rho_{13}$  (in this case, 15) would be statistically evaluated against a Poisson distribution with  $\lambda = 2.5$ .

These two analyses of conditional Type I error rate were done only for Method I. Methods II and III resulted in too few cells for the test of fit and in too few Type I errors for conditional empirical size.

Type I Error Rate: Other Null Hypotheses. The analyses mentioned thus far have been only those in which the complete null hypothesis is true (i.e.,  $R_p = \underline{I}$ ). Other population correlation matrices were also examined in this study. In general, the same analyses were used with respect to Hypotheses 1, 2, 3, 4, and 5 as were used for samples from populations of independent variables. Of course, Type I errors are possible only for true component null hypotheses. If a component null hypothesis is actually false (i.e.,  $\rho_{ij} \neq 0$ ) then no Type I errors are possible on that component test. The critical value for the three Methods were the same as in the previous cases, computed with  $\alpha_{T(I)} = .05$  and by equations (60) and (61). For the analysis, however, of conditional empirical size, the binomial distribution of number of Type I error was based on the number of true component null hypotheses. The only other change from the previous analyses was that a two-tailed alternative hypothesis was used in equation (66) for the analysis of the familywise Type I error rate by Method II.

Type II Error Rate. The cases of population correlation matrices  $R_p \neq \underline{I}$  also provided opportunities to evaluate Type II error rates (Hypothesis 6). There were no statistical analyses of Type II error rates. Instead, the empirical power per component test was calculated. This is the proportion of times that a false component null hypothesis was correctly rejected in the 1000 trials.

For some population matrices, there was more than one false component null hypothesis. In these cases, the familywise empirical power was calculated, i.e., the proportion of times that all false component null hypotheses in the family of tests were correctly rejected. The empirical power per test and, when applicable, familywise empirical power were compared for the three methods of determining  $\alpha_T$ .

#### Analyses of Vectors of Correlations

These analyses were based on the data from the first column of each correlation matrix, using the same data generated for the analyses of matrix intercorrelations. Familywise empirical size, conditional empirical size, and empirical power were analyzed similarly to the matrix cases discussed previously. The main difference was that the family of tests included only those in the first column of the matrix rather than those in the entire matrix. Therefore, the value of  $m$  used for calculating  $\alpha_T$  according to the three methods was calculated by equation (63) rather than equation (62) which applied in the matrix cases.

#### Summary of Analyses

Table 3 presents an overview of the analyses of the Monte Carlo empirical data. One thousand sample correlation matrices were generated for each indicated combination of sample size, number of variables, and population correlation matrix. The indicated analyses were performed on those 1000

Table 3. Summary of Analyses of this Study

$R_{sp}$	k	N	Familywise Type I Error Rate		Conditional Type I Error Rate		Type II Error Rate		
			Matrix	Vector	Matrix	Vector	Matrix	Vector	
I	3	15	S	S	B	B			
		40	S	S	B	B			
		100	S	S	B	B			
	4	15	S	S	B	B			
		40	S	S	B	B			
		100	S	S	BC	BC			
	8	15	S	S	B	B			
		40	S	S	B	B			
		100	S	S	B	B			
(51) <sup>a</sup>	3	15	S	S	B	B	F		
		40	S	S	B	B	F		
		100	S	S	B	B	F		
(52) <sup>a</sup>	3	15	S	S	B	B	F		
		40	S	S	B	B	F		
		100	S	S	B	B	F		
(53) <sup>a</sup>	3	15	S	S	B	B	F		
		40	S	S	B	B	F		
		100	S	S	B	B	F		
(54) <sup>a</sup>	4	40	S	S	BC	BC	F	P	
		100	S	S	BC	BC	F	P	
(55) <sup>a</sup>	4	40	S	S	BC	BC	F	P	F
		100	S	S	BC	BC	F	P	F
(56) <sup>a</sup>	6	40	S	S	BC	BC	F	P	F
		100	S	S	BC	BC	F	P	F
(57) <sup>a</sup>	6	40	S	S	BC	BC	F	P	F P
		100	S	S	BC	BC	F	P	F P
(58) <sup>a</sup>	6	40	S	S	BC	BC	F	P	F P
		100	S	S	BC	BC	F	P	F P

Note. - Key: S, familywise empirical size; B, test-of-fit to binomial distribution; C, conditional empirical size; P, empirical power per test; F, familywise empirical power. The left column indicates the population correlation matrices and sample sizes investigated.

<sup>a</sup>Refers to the correlation matrix designated by this equation number in the text.

sample correlation matrices for each of the three methods of determining  $\alpha_T$  and the corresponding critical values for  $r_{ij}$ .

## CHAPTER III

## RESULTS

Statistical Tests of Pseudorandom Number GeneratorTest of Fit to 8-Variate Normal Distribution

The unit normal pseudorandom number generator was used to generate scores for as many as eight variables at a time. To test the hypothesis that these scores are jointly distributed as an 8-variate normal distribution with mean vector  $\underline{\mu} = \underline{0}$  and covariance matrix  $\underline{C} = \underline{I}_8$ , a  $\chi^2$  goodness-of-fit test was performed as described in Chapter II. The results are shown in Table 4.

Table 4. Summary of Tests of Goodness of Fit to An 8-Variate Normal Distribution

k	N	Total 8-Tuples <sup>a</sup>	$\chi^2$ Value <sup>b</sup>	Probability of Larger $\chi^2$
8	40	40,000	6743.06	.056
8	100	100,000	6578.08	.44

Note. - There were 6561 frequency cells with equal expected frequencies for each cell.

<sup>a</sup>An 8-Tuple is one set of observations on the 8 variables.

<sup>b</sup>Degrees of freedom = 6560

The tests of goodness of fit indicate that the generated numbers adequately fit the desired 8-variate normal distribution ( $p < .05$ ). So the null hypothesis was accepted

that the pseudorandom normal numbers did fit an independent 8-variate normal distribution. Additional tests would have been preferable, but limitations of computer time prevented this.

#### Test of Fit to Bivariate Normal Distribution

The test of goodness of fit to an 8-variate normal distribution was limited both by computer time and by having each variable divided into only three equally probable intervals. Tests of goodness of fit to bivariate normal distributions were used because they diminished these limiting factors. Such tests used less computer time and permitted more intervals on each variable. To test the hypothesis that observations on two variables were jointly distributed as a bivariate normal distribution with mean vector  $\underline{\mu} = \underline{0}$  and covariance matrix  $C = I_2$ , a  $\chi^2$  goodness-of-fit test was performed as described in Chapter II. The results (Table 5) indicate that the generated pseudorandom numbers adequately fit the desired bivariate normal distribution. The probabilities of observing a  $\chi^2$  value larger than the one actually observed range from .14 to .83 for the different tests. So the null hypothesis was accepted that the pseudorandom normal numbers did fit an independent bivariate normal distribution.

#### Statistical Test of Type I Error Rate for a Single Correlation Coefficient

As discussed in Chapter II, the expected proportion of Type I errors is known when only one significance test of one correlation coefficient is of interest. This case was simulated six times in this study by specifying  $k = 2$  and  $R_{\rho} = I$  for six runs of the computer program of 1000 trials each. The results of these computer runs are presented in Table 6. It can be seen that for two of the six runs, the number of observed Type I errors was significantly ( $p < .05$ ) different than 50 (the expected number). The probability of observing two or more significant results at the .05 level of significance in a family of six independent tests of fit is .033 according to the binomial distribution. So this test indicated that the method of generating simulated sample correlation coefficients resulted in a different rate of Type I errors than truly random numbers would produce.

The implications of the results of these tests will be discussed further since they suggest that some of the main results of this study must be somewhat qualified. From Table 6 there seems to be a tendency toward fewer Type I errors than should be expected. This will be called the "undergeneration of Type I errors". Note that the observed number of Type I errors is less than expected on each of the six computer runs. It will be shown in the next chapter that there seemed to be a tendency toward undergeneration of Type I errors in the computer runs of



Table 5. Summary of Tests of Goodness of Fit to a Bivariate Normal Distribution

k	N	Two Variables <sup>a</sup>	Total 2-Tuples	$\chi^2$ Value <sup>c</sup>	Probability of Larger $\chi^2$
3	40	1,3	40,000	2503.75	.47
3	40	1,2	40,000	2447.00	.77
4	15	1,4	15,000	2505.32	.46
4	40	1,3	40,000	2576.25	.14
8	15	1,8	15,000	2431.32	.83

Note. - There were 2500 frequency cells with equal expected frequencies for each cell.

<sup>a</sup>Of the k variables, these were selected for the goodness of fit test.

<sup>b</sup>A 2-Tuple is one set of observations on the 2 variables.

<sup>c</sup>Degrees of freedom = 2499

Table 6. Summary of Statistical Tests of Type I Error Rate of a Single Correlation Coefficient

Computer Run	N	Observed Type I Errors <sup>a</sup>
1	15	49
2	15	34*
3	40	49
4	40	36*
5	100	43
6	100	45

<sup>a</sup>The expected number of Type I errors was 50 for each case.

\* $p < .05$ ; two-tailed Poisson probability.

primary interest in this study. Out of 28 such computer runs, only seven had more total Type I errors than the expected number. The undergeneration of Type I errors seemed to be more pronounced for cases with small sample size than for cases with large sample size. A detailed summary of the observed total Type I errors for each computer run is presented in Table C.1 in Appendix C.

### Results of Tests of Hypotheses of This Study

The major hypotheses of this study involve the Type I error rates, the conditional Type I error rates, and the Type II error rates of a family of significance tests of intercorrelations using the three different methods of finding  $\alpha_T$  for each significance test. These hypotheses are listed at the end of Chapter I. In this section, the results of the tests of hypotheses about Type I error rates are discussed first, followed by the results concerning Type II error rates.

#### Type I Error Rates

The results of the tests of hypotheses about Type I error rates are considered in two groups here. The first group consists of the results concerning familywise Type I error rates. The second group consists of the results concerning conditional Type I error rates.

Familywise Type I Error Rates. Three hypotheses concerned familywise Type I error rates. The major kind of data relevant to these hypotheses was familywise

empirical size, which was the proportion of familywise Type I errors that occurred on a given computer run. Table D.1 in Appendix D reports the familywise empirical size, the population correlation matrix, and the sample size for each computer run.

The first major hypothesis was that the familywise Type I error rate ( $\alpha_{FW}$ ) would be significantly larger than .05 when Method I was used with  $\alpha_T = .05$ . Table D.1 indicates that this hypothesis was strongly supported. Furthermore, the familywise empirical size became larger as the number of true component null hypotheses increased. The number of true component null hypotheses is the number of hypothesis tests in the family of tests (tests of all intercorrelations in either the matrix or the first column) for which  $\rho_{ij} = 0$ . Table 7 summarizes the familywise empirical sizes according to the number of true component null hypotheses. Note that as the number of true component null hypotheses increases, the familywise empirical sizes increase.

The second major hypothesis was that  $\alpha_{FW}$  would not be significantly larger than .05 for Methods II and III. This was supported by the data. Using Methods II and III, the familywise empirical size was never larger than .05 except for a few cases that could easily be attributed to chance (see Table D.1). However, these results must be qualified somewhat because the undergeneration of Type

Table 7. Familywise Empirical Sizes by Number of True Component Null Hypotheses in a Family (Method I)

True Component Null Hypotheses <sup>a</sup>	Familywise Empirical Size <sup>b</sup>	
	Mean	Range
2	.087	.062 - .119
3	.125	.094 - .164
4	.162	.139 - .181
6	.239	.209 - .262
7	.276	.224 - .305
9	.244	.239 - .248
12	.396	.393 - .398
13	.454	.452 - .456
21	.725	.669 - .755

<sup>a</sup>The number of significance tests in the family of tests (matrix or vector) for which the null hypothesis was true.

<sup>b</sup>Proportion of test samples containing at least one Type I error.

I errors is a confounding factor. This qualification does not seem too serious, though, since the familywise empirical sizes supported the hypothesis even for sample sizes of 100, for which there was no apparent under-generation of Type I errors.

The third major hypothesis was that the familywise Type I error rate for Method III would be less than .05 for a larger number of variables. This was not sup-

ported by the data. The familywise empirical size was significantly less than .05 for many computer runs ( $p < .05$ ), but this did not seem to be related to the number of variables. The cases for which  $R_p = \underline{I}$  are considered first. In these cases familywise empirical size was significantly less than .05 for three out of 18 cases. Two of these three deviant cases occurred with eight variables. Furthermore, these three deviant cases could be attributed to the undergeneration of Type I errors.

There were more cases with  $R_p \neq \underline{I}$  in which the familywise empirical sizes for Method III were significantly less than .05. However, these were not related to the number of variables in any apparent way. For the cases in which the entire matrix of intercorrelations provided the data for the family of significance tests, the only familywise empirical sizes not significantly less than .05 were for cases with the largest number of variables.

The familywise empirical sizes for Method III seem to be explained best by considering the proportion of true component null hypotheses. The proportion of true component null hypotheses is the ratio of the number of true component null hypotheses in the family of tests to the total number of component null hypotheses in the family of tests. For example, when  $R_p = \underline{I}$ , all component null hypotheses are true, so this proportion is 100%. Table 8

Table 8. Range of Familywise Empirical Size by Proportion of True Component Null Hypotheses in a Family (Method III)

Proportion of True Component Null Hypotheses <sup>a</sup>	Familywise Empirical Size <sup>b</sup>	
	Mean	Range
100%	.042	.026 - .057
87%	.039	.035 - .042
80%	.039	.034 - .042
67%	.027	.018 - .033
60%	.025	.019 - .034
50%	.022	.016 - .023

<sup>a</sup>The proportion of component significance tests (matrix or vector) for which the null hypothesis is true.

<sup>b</sup>Proportion of test samples containing at least one Type I error.

summarizes the trend in familywise empirical size according to the proportion of true component null hypotheses in a family. It can be seen that as the proportion of true component null hypotheses decreases, the familywise empirical sizes decrease also.

It is interesting that Methods II and III led to making decisions that differed very little from each other in this study. In 25 of the 56 comparisons, the familywise empirical size was identical for the two methods. In another 24 comparisons, they differed by only .001.

Conditional Type I Error Rates. Two major hypotheses concerned conditional Type I error rates. The first hypothesis was that the conditional Type I error rate is greater than  $\alpha_T$ . The second hypothesis was that this effect is particularly strong when some of the variables in the sample are highly correlated. Since these hypotheses are similar they are considered here together. The data indicated that the conditional Type I error rate increases above  $\alpha_T$  when  $R_{mp} \neq I_m$  but not when  $R_{mp} = I_m$ . Two types of analysis of conditional Type I error rates were used. The first was a test of fit of the frequencies of each number of Type I errors for the 1000 sampling replications to a binomial distribution, as explained in Chapter II. The results of these tests of fit are reported in Table D.2 in Appendix D. The second type of analysis was conditional empirical size, the observed proportion of Type I errors on one component test, given that a Type I error occurred on the same sampling replication on another previously specified component test in the same family of tests. This is also explained in more detail in Chapter II.

The cases for which  $R_{mp} = I_m$  are considered first. From Table D.2 it can be seen that in five of the 18 tests of fit, the observed frequencies of number of Type I errors deviated from the expected binomial distribution more than could be accounted for by chance ( $p < .05$  for

each test of fit). Taken together, this indicates a significant deviation from the expected binomial distribution ( $p < .05$ ). However, four of the five significant tests of fit could easily be explained by the undergeneration of Type I errors for small sample sizes. In each of these cases, there were a larger-than-expected frequency of zero Type I errors and smaller-than-expected frequencies in most of the other cells. This implies that the deviations from the expected binomial distribution could be attributed to  $p$  not being  $\alpha_T$  rather than the component tests not being mutually independent.

The second type of analysis, conditional empirical size, was performed for only one case ( $k = 4$ ;  $N = 100$ ). It can be seen from Table D.3 in Appendix D that the conditional empirical sizes are close to  $\alpha_T = .05$ .

Now the cases for which  $R_{\substack{p \\ m}} \neq I$  are considered. Conditional Type I error rates were often greater than  $\alpha_T$  for many of these cases. This was evident from the results of many of the tests of fit reported in Table D.2 in Appendix D. For each Pearson  $\chi^2$  test of fit the expected frequency of each number of Type I errors was calculated according to a binomial distribution, as explained in Chapter II. The tail of the distribution was combined so that each cell would have an expected frequency of at least five. Thus, for example, with  $\alpha_T = .05$ , 1000 sampling replications, and three component null hy-



potheses, the expected frequencies would be 857.4 replications with zero Type I errors, 135.4 replications with one Type I error and 7.2 replications with two or more Type I errors.

Of the 20 cases in which  $R_p \neq I$  and there were more than two cells for the test of fit, 19 cases had observed frequencies of number of Type I errors which deviated from the expected binomial distribution ( $p < .05$ ; see Table D.2). Most of these Pearson  $\chi^2$  values are very large.

Most of the cases with  $R_p \neq I$  and two cells for the test of fit are reported as adequate fits to the binomial distribution in Table D.2. However, this is because all replications with any Type I errors were counted in the same cell for the test of fit. Table 9 gives a detailed report of these cases. It can be seen that there were consistently more cases with two Type I errors than would be expected if the component tests were mutually independent. However, because the expected frequency of two Type I errors was only 2.5, the cell for one Type I error was combined with the cell for two Type I errors for the Pearson  $\chi^2$  test of fit. Thus the fact that Type I errors tended to occur together was obscured by the requirements for the Pearson  $\chi^2$  test of fit. If none of the three cells had been combined for the tests of fit, 8 of the 10 cases in Table 9 would have been considered significant

Table 9. Distribution of Number of Type I Errors  
for Cases with Two Component Null Hypotheses  
in the Family of Tests (Method I)

$R_{mp}$	k	N	Matrix or Vector	Observed Frequencies of Number of Type I Errors <sup>a</sup>			Conditional Empirical Size <sup>b</sup>
				0	1	2	
(51) <sup>c</sup>	3	15	Both	921	76	3	.073
		40	Both	910	84	6	.125
		100	Both	909	84	7	.143
(52) <sup>c</sup>	3	40	Both	923	69	8	.188
		100	Both	908	80	12	.231
(53) <sup>c</sup>	3	15	Both	938	55	7	.203
		40	Both	908	79	13	.248
		100	Both	914	79	7	.151
(55) <sup>c</sup>	4	40	Vector	918	78	4	.093
		100	Vector	903	82	15	.268

Note. - Expected frequencies of number of Type I errors: 0 errors, 902.5; 1 error, 95; 2 errors, 215.

<sup>a</sup>By Method I

<sup>b</sup>The observed proportion of Type I errors on one component test, given that a Type I error occurred on the other component test. These values in this column were calculated by assuming that the frequencies of one Type I error was divided evenly among the two component tests.

<sup>c</sup>Refers to the population correlation matrix designated by this equation number in Chapter II.

deviations from the expected binomial distribution (However, this would necessitate an expected frequency of 2.5, which is smaller than is generally acceptable for Pearson  $\chi^2$  tests of fit).

In conclusion, in the cases considered in this study for which  $R_p \neq I_m$ , the observed frequencies of number of Type I errors deviated significantly from the binomial distribution expected if the tests had been mutually independent. The deviation generally followed a similar pattern, with higher-than-expected frequencies of zero Type I errors, higher-than-expected frequencies of a large number of Type I errors, and lower-than-expected frequencies of intermediate numbers of Type I errors. In Table 9, for example, the observed frequencies of zero and two Type I errors are consistently higher than expected, whereas the observed frequencies of one Type I error are consistently lower than expected.

The second type of analysis, conditional empirical size, also showed the effect of the dependence of the significance tests. The conditional empirical size for certain component tests given that a Type I error occurred on a specified test is reported in Tables D.4 through D.13 in Appendix D. These tables report the observed numbers of Type I errors on a test of  $r_{ij}$  given that Type I error has occurred also on another specified test (on  $r_{21}$ ,  $r_{31}$ , or  $r_{41}$ ). Most of the significant results reported in Tables D.4 through D.13 involve what will be called here "strongly-linked" component tests. Such "strongly-linked" component tests are two component tests

$$\begin{aligned}
 H_0 &: \rho_{ij} = 0 \\
 H_1 &: \rho_{ij} \neq 0
 \end{aligned}
 \tag{69}$$

and

$$\begin{aligned}
 H_0 &: \rho_{kj} = 0 \\
 H_1 &: \rho_{kj} \neq 0
 \end{aligned}
 \tag{70}$$

in which the null hypothesis is true for both of them and  $\rho_{ki} = .6$ . In other words, "strongly-linked" component tests involve a pair of correlations with one variable in common (represented by  $j$ ) and the other variables correlated .6 with each other (represented by  $k$  and  $i$ ). On any "strongly-linked" component test the conditional empirical size (conditional on a Type I error occurring on its "strongly-linked" partner) was more than .05 (the unconditional Type I error rate,  $\alpha_T$ ). Table 10 reports the data on conditional empirical size for "strongly-linked" component tests. It can be seen that the number of conditional Type I errors are generally significantly larger ( $p_{FW} < .01$ ) than would be expected if the two component tests were independent. The mean conditional empirical size for these "strongly-linked" component tests is .242. Conditional empirical sizes are also reported in Table 9. Most of these also involve "strongly-linked" component tests (All except those using

Table 10. Conditional Empirical Sizes of  
"Strongly-Linked"<sup>a</sup> Component Tests

$R_p$	From Table No.	Con- ditional on Type I Error on:	"Strongly- Linked" <sup>a</sup> $r_{ij}$	Conditional Type I Errors Ex- pected <sup>b</sup> Actual		Con- ditional Empirical Size <sup>c</sup>
(54) <sup>d</sup>	D.4	$r_{21}$	$r_{31}$ $r_{41}$	2.2	8**	.182
				2.2	12**	.273
	D.5	$r_{21}$	$r_{31}$ $r_{41}$	2.15	12**	.279
				2.15	10**	.233
(55) <sup>d</sup>	D.6	$r_{31}$	$r_{32}$ $r_{41}$	2.3	8**	.174
				2.3	4	.087
	D.7	$r_{31}$	$r_{32}$ $r_{41}$	2.7	16**	.296
				2.7	15**	.278
(56) <sup>d</sup>	D.8	$r_{31}$	$r_{32}$ $r_{41}$	3.4	15**	.221
				3.4	16**	.235
	D.9	$r_{31}$	$r_{32}$ $r_{41}$	2.45	7 <sup>f</sup>	.143
				2.45	14**	.286
(57) <sup>d</sup>	D.10	$r_{21}$	$r_{52}$ $r_{62}$	2.9	11**	.190
				2.9	15**	.259
	D.11	$r_{21}$	$r_{52}$ $r_{62}$	2.4	9** <sup>e</sup>	.188
				2.4	14**	.292
(58) <sup>d</sup>	D.12	$r_{41}$	$r_{42}$	2.7	14**	.259
			$r_{43}$	2.7	14**	.259
			$r_{51}$	2.7	13**	.241
			$r_{61}$	2.7	15**	.278
D.13	$r_{41}$	$r_{42}$	2.05	9**	.220	
		$r_{43}$	2.05	13**	.317	
		$r_{51}$	2.05	14**	.341	
		$r_{61}$	2.05	11**	.268	

Note.- These data are based on Method I, with  $\alpha_T = .05$ .

<sup>a</sup>"Strongly-linked" component tests involve a pair of correlations with one variable in common and the other variables correlated .6 with each other.

<sup>b</sup>Expectation based on assumption of independence of the second test from the first.

<sup>c</sup>The observed proportion of Type I errors on one component test (on  $r_{ij}$  in 4th column) among the sampling replications with Type I errors on the "strongly-linked"

Table 10 Cont'd

partner (on  $r_{ij}$  in 3rd column).

<sup>d</sup>Refers to the population correlation matrix designated by this equation number in Chapter II.

<sup>e</sup>multistage procedure

<sup>f</sup> $p_T < .05$

<sup>\*\*</sup> $p_{FW} < .01$ , Bonferroni Poisson one-tailed test, with the tests for one table (e.g. Table D.4) regarded as a family of tests.

$R_p$  equation (51)). For most of these cases, however, the exact number of Type I errors on each component test is unknown. Therefore, the values for conditional empirical size were calculated by assuming that the Type I errors on sampling replications with one Type I error were divided evenly among the two component tests. For the five cases which were not analyzed directly in Table 10 (the cases using  $R_p$  equation (55) were analyzed both ways), the mean conditional empirical size was .202.

The other significant results and some borderline results in Tables D.4 through D.13 involve what will be called here "moderately-linked" component tests. Such "moderately-linked" component tests are two significance tests

$$\begin{aligned} H_0 : \rho_{ij} &= 0 \\ H_1 : \rho_{ij} &\neq 0 \end{aligned} \quad (71)$$

and

$$\begin{aligned} H_0 : \rho_{k\ell} &= 0 \\ H_1 : \rho_{k\ell} &\neq 0 \end{aligned} \quad (72)$$

in which the null hypothesis is true for both of them and

$\rho_{ik} = .6$  and  $\rho_{j2} = .6$ . In other words, "moderately-linked" component tests involve a pair of correlations

Table 11. Conditional Empirical Size of "Moderately-Linked"<sup>a</sup> Component Tests

$R_p$	From Table No.	Con- ditional Error on: on Type I	"Moder- ately- Linked" <sup>a</sup> $r_{ij}$	Conditional Type I Errors Ex- pected	Actual	Con- ditional Empirical Size <sup>b</sup>
(55) <sup>c</sup>	D.6	$r_{31}$	$r_{42}$	2.3	6 <sup>d</sup>	.130
	D.7	$r_{31}$	$r_{42}$	2.7	5	.093
(56) <sup>c</sup>	D.8	$r_{31}$	$r_{42}$	3.4	7 <sup>d</sup>	.103
	D.9	$r_{31}$	$r_{42}$	2.45	6 <sup>d</sup>	.122
(58) <sup>c</sup>	D.12	$r_{41}$	$r_{52}$	2.7	7*	.130
			$r_{53}$	2.7	14**	.259
			$r_{62}$	2.7	8*	.148
			$r_{63}$	2.7	5	.093
	D.13	$r_{41}$	$r_{52}$	2.05	5	.122
			$r_{53}$	2.05	2 <sup>d</sup>	.049
			$r_{62}$	2.05	6 <sup>d</sup>	.146
			$r_{63}$	2.05	5	.122

Note. - These data are based on Method I, with  $\alpha_T = .05$ .

<sup>a</sup>"Moderately-linked" component tests involve a pair of correlations for which the two variables of one correlation are each correlated .6 with one of the variables of its partner correlation.

<sup>b</sup>The observed proportion of Type I errors on one component test (on  $r_{ij}$  in 4th column) among the sampling replications with Type I errors on the moderately-linked partner (on  $r_{ij}$  in 3rd column).

<sup>c</sup>Refers to the population correlation matrix designated by this equation number in Chapter II.

<sup>d</sup> $P_T < .05$

\* $P_{FW} < .05$ , Multistage Bonferroni one-tailed test, with the tests for one table (e.g., Table D.4) regarded as a family of tests.

\*\* $P_{FW} < .01$ , Multistage Bonferroni as above.

for which the two variables  $(i,j)$  of one correlation are each correlated .6 with one of the variables of its partner ( $i$  with  $k$  and  $j$  with  $l$ ). Table 11 reports the conditional empirical size for such cases. It can be seen that no definite conclusions can be made based on this data alone, but it is suggestive that conditional Type I error rates do increase above .05 for "moderately-linked" component tests. The mean conditional empirical size for the cases in Table 11 is .126.

There were no other clear effects present in the data on conditional empirical size, although Tables D.9 and D.11 include a total of three other instances of results of borderline significance.

#### Type II Error Rates

Power. The power of the different methods of determining  $\alpha_T$  was investigated in terms of the empirical power of component tests and, in some cases, in terms of familywise empirical power. The data concerning empirical power of component tests is summarized in Table 12. When the population correlation coefficient was equal to  $\pm .6$  and the sample size was 100, the empirical power was 1.00 in all cases for all methods. When the population correlation coefficient was equal to  $\pm .6$  and the sample size was 40, the empirical power was about .99 for Method I, which set  $\alpha_T$  equal to .05. For Methods II and III, the empirical power varied also with the number of tests in the family of tests (Methods II and III were nearly



Table 12. Empirical Power on Individual Component Tests

$\rho_{ij}$	N	k	Number of Empirical Power Estimates	Empirical Power <sup>a</sup>			
				Method I		Method III	
				Mean	Range	Mean	Range
	100	3,4,6	18	1.00	1.00	1.00	1.00
±.6	40	3	2	.991	.988-.993	.971	.966-.975
	40	4	5	.991	.988-.994	.951	.943-.956
	40	6	11	.991	.985-.994	.909	.896-.927
	15	3	2	.757	.705-.809	.587	.517-.656
.3	100	3	1	.863	.863	.741	.741
	40	3	1	.472	.472	.301	.301
	15	3	1	.200	.200	.094	.094

Note. - This table includes empirical power based on analyses of correlation matrices only.

<sup>a</sup>The proportion of sampling replications of a component test without a Type II error occurring.

identical in empirical power). For  $\rho_{ij} = \pm.6$  and  $N = 40$ , the power for these methods decreased from about .97 for  $k = 3$  to about .91 for  $k = 6$ . So the difference between Method I and Method III (or II) in terms of power becomes greater as the number of component tests in a family of tests increases.

Only two estimates were made of power given  $\rho_{ij} = \pm.6$  and a sample size of 15. These estimates were both for cases of  $k = 3$  and they varied a good deal from each other. However, empirical power by Method III (about .59) was further below empirical power by Method I (about .76) than for any other cases where  $\rho_{ij} = \pm.6$ .

Empirical power by Method III was also about .10 to .17 less than empirical power by Method I when  $\rho_{ij} = .3$ .

The results of the analysis of power was very similar for correlation vectors. Power for component tests by Method I is exactly the same for correlation vectors as for correlation matrices. Empirical power for component tests by Method III was higher for correlation vectors of  $k$  variables than for correlation matrices of  $k$  variables, but this was only a result of the different number of tests in their respective families of tests (see equations (62) and (63)).

The analysis of familywise empirical power added nothing to the analysis of empirical power per component test. In all computer runs with  $N = 100$  for which familywise empirical power was applicable, the empirical power was 1.00 regardless of method of determining  $\alpha_T$ . In the other applicable computer runs, with  $N = 40$ , familywise empirical power for correlation matrices varied from .952 to .984 for Method I, and from .637 to .900 for Method III. For both methods, familywise empirical power was related to the number of false component null hypotheses in the family of tests. The lowest values of familywise empirical power occurred when there were the most non-zero population correlation coefficients.

## CHAPTER IV

## DISCUSSION

This chapter first discusses the results of the empirical study and then relates the findings to some examples of statistical analyses of intercorrelations from recent journal articles. The discussion of the empirical study begins with a consideration of the adequacy of the pseudorandom number generator which was used in this study. There is some indication that the pseudorandom numbers did not demonstrate some important properties of randomness. The implication of this for the main results of this study are considered.

Secondly, the primary results of this study are discussed. The control of Type I error rate is considered first, followed by a discussion of the control of Type II error rate.

Then the major conclusions from this study are summarized. Following the summary, a multistage Bonferroni procedure is recommended for controlling Type I error rate in multiple significance tests of intercorrelations.

Finally, the major findings of this study are related to two recent journal articles. This final section includes an illustration of the use of a multistage Bon-

ferroni procedure applied to one of the sets of data.

#### Adequacy of Pseudorandom Number Generator

The results of two of the statistical tests of the pseudorandom number generator used in this study indicate an adequate fit to the bivariate normal distribution and to an 8-variate normal distribution. However, the other test indicates that the computer simulation is biased toward producing fewer total Type I errors than expected. If the generated independent normal numbers were truly random and if the transformations were applied correctly in the computer program, no such bias would exist.

There was also some tendency toward an undergeneration of the total number of Type I errors in the computer runs of primary interest in this study. No exact inferential statistical analysis seemed applicable to this data as a whole, so descriptive statistics are emphasized in this presentation rather than inferential statistics.

The observed number of total Type I errors for each computer run is presented in Table C.1. These are compared with the expected number of total Type I errors, which is actually the error rate per family ( $\alpha_{PF}$ ) multiplied by the number of sampling replications (1000). As Miller (1966) and Ryan (1959) have noted,  $\alpha_{PF}$  is not affected by the dependence of component tests in a family of tests. The expected number of Type I errors on one com-

ponent test in 1000 sampling replications is

$$\begin{aligned} & E \text{ (total Type I errors on one component test)} \\ & = 1000 \alpha_T . \end{aligned} \quad (73)$$

If  $R_p = I_k$ , then the expected number of Type I errors on the  $\frac{k(k-1)}{2}$  component tests in this family of tests in 1000 sampling replications is

$$\begin{aligned} & \frac{k(k-1)}{2} E \text{ [(total Type I errors on one component test)]} \\ & = \frac{k(k-1)}{2} (1000 \alpha_T) \\ & = 1000 \alpha_{PF} . \end{aligned} \quad (74)$$

Since no assumption of independence is necessary for this conclusion, equation (74) holds regardless of any dependence among the component tests.

Although the expected number of total Type I errors is known, the variance of the total Type I errors is unknown and therefore, exact statistical inference is impossible. The binomial distribution and its normal approximation would be appropriate if the component tests were mutually independent. Technically the component tests of interest are never mutually independent, but the results of this study indicate that if  $R_p = I_k$ , the component tests do not deviate significantly from mutual independence. If it

is assumed that a mutual independence assumption is justified, then a normal approximation to a binomial distribution can be used to make inferences about the possible undergeneration of Type I errors. Admittedly, this procedure is not entirely justified, but it provides some information needed for evaluating the main results of this study.

Table 13. Total Number of Type I Errors on Each Computer Run ( $R_p = I$  only)

k	N	Total Type I Errors	
		Expected	Observed
3	15	150	147
	40	150	132
	100	150	168
4	15	300	237***
	40	300	284
	100	300	297
8	15	1400	1109***
	40	1400	1327*
	100	1400	1411
Total		5550	5112***

Note. - The observed total Type I errors may be slightly underestimated due to incomplete data. The statistical analyses are based on two-tailed normal approximations to binomial distributions ( $\mu = Np - 1000 m \times .05$ ;  $\sigma = \sqrt{Npq} = \sqrt{1000 m \times .05 \times .95}$ ).

\* $P_T < .05$

\*\*\* $P_T < .001$

Table 13 reports the total Type I errors for cases for which  $R_p = I$ . The total of the total numbers of Type I errors is significantly less than expected ( $Z = -6.03$ ;

$p < .001$ ). Three of the eight computer runs produced significantly less total Type I errors than expected ( $p < .05$ ).

There are four possible explanations for the under-generation of the total number of Type I errors. First, it could have been a case of a large random deviation from the expected value that does occur a small percentage of the time. Second, it could have been due to some misapplication of the appropriate transformations from independent normal numbers to correlated normal numbers. Third, it could have resulted from inaccuracies in counting Type I errors, such as would result from an erroneous critical value. Fourth, it could have been a result of some non-random properties of the pseudorandom number generator.

The first explanation seems unlikely since the two tests which resulted in a conclusion that the Type I error rate was undergenerated were carried out on entirely separate data.

The second explanation appears unlikely because the computer simulation did not undergenerate the Type I error rate for large sample sizes, but only for smaller sample sizes (see Table 14). Of the computer runs with  $N = 100$ , 5 of 11 had more total Type I errors than expected. The range of observed total Type I errors deviated equally in either direction from the expected value. So no undergen-

Table 14. Summary of Total Number of Type I Errors by Sample Size

N	Number of Computer Runs	Runs with Total Type I Errors Greater Than Expected Number	Range <sup>a</sup>
100	11	5	88%-112%
40	11	2	85%-105%
15	6	0	69%-98%
Total	28	7	69%-112%

<sup>a</sup>Range of the observed numbers of total Type I errors expressed as a percentage of the expected number.

eration of Type I errors was evident for cases with  $N = 100$ . However, with  $N = 15$ , all computer runs had fewer total Type I errors than expected. Furthermore, the range of observed total Type I errors deviated further from the expected value than for any other sample size. The computer runs with  $N = 40$  were intermediate between these extreme sample size cases. The transformation that the computer program used to transform a matrix of independent normal pseudorandom deviates to correlated normal deviates and then to a sample correlation matrix were also checked by a hand calculator and found to be accurate.

The third explanation also appears unlikely since no inaccuracies were found in a careful check of the computer program's count of Type I errors. Using the case of two independent variables ( $R_p = I_{m2}$ ), 500 sample correlation matrices were generated and printed. The computer program's count of the number of Type I errors was exactly



what it should have been.

The most plausible explanation seems to be that the fault lies with the pseudorandom number generator. Previously, multiplicative congruential generators have been found sometimes to produce systematic biases when certain transformations involving combinations of pseudorandom numbers have been used (e.g., Marsaglia, 1968; Neave, 1973). Apparently something on the same order occurred with the transformations of this study. Two explanations of this bias seem possible. The first one is that large sequences of these particular pseudorandom numbers may have better statistical properties than short sequences. Jansson (1966) has made a distinction between global randomness and local randomness. Local randomness deals with the statistical properties of relatively small samples whereas global randomness is concerned with asymptotic statistical properties of randomness. The particular pseudorandom number generator used in this empirical study was recommended on the basis of a crucial asymptotic statistical property of the generated number sequence, the serial correlations of the longest possible sequence of numbers (Ahrens, Dieter, & Grube, 1970). It may be that the generator used in this study has adequate global randomness for this type of application but not adequate local randomness for small sample sizes. Table 15 shows that there does seem to be an increasing undergeneration of total Type I

Table 15. Summary of Total Numbers of Type I Errors by Number of Normal Numbers Generated in a Run of the Computer Program

Number of Unit Normal Numbers in Run <sup>a</sup>	Computer Runs	Runs with More Type I Errors Than Expected	Range <sup>b</sup>
320-1,600	8	3	88%-108½%
240-300	7	3	93%-112%
120-160	8	1	79%-105%
45-60	5	0	69%-112%

<sup>a</sup>In thousands

<sup>b</sup>Range of the observed numbers of total Type I errors expressed as a percentage of the expected number.

errors as the total number of normal pseudorandom numbers generated increases.

The second possible explanation for the bias is that the starting number for each run of the computer program was less than  $2^{27}$ . This meant that the starting number was limited to  $\frac{1}{256}$ <sup>th</sup> of the overall possible interval  $(0, 2^{35})$ . Perhaps this caused a systematic undergeneration of Type I error rate in the first group of sample correlation matrices generated, but that this bias was negligible for runs of the computer program which required larger sequences of normal pseudorandom numbers.

In any case the undergeneration of total Type I error rate somewhat qualifies the results in the main part of this study. How serious is this undergeneration of total Type I errors? The 95% confidence interval for  $\alpha_T$  based on all six computer runs in Table 6 is (.0376,

.0478). The 95% confidence interval from the grand total in Table 13 is (.0449, .0473). This would indicate that the simulated  $\alpha_T$  for the computer program was about .046 rather than .050. The simulated  $\alpha_T$  would apparently be lower for small sample sizes (Table 14) and for small sequences of generated normal deviates (Table 15). For larger sample sizes and/or large sequences of generated normal deviates, the simulated  $\alpha_T$  probably approximated .050 very well. So the undergeneration of total Type I errors somewhat qualifies the results of this study, particularly those results based on small sample sizes or small sequences of generated normal deviates.

For future research there are several methods for pseudorandom number generation which may improve on the method here. Knuth (1969) recommends using a congruential pseudorandom number generator with a modulus of  $2^{35} \pm 1$  rather than  $2^{35}$  (for a computer with a word size of  $2^{35}$ ). This makes the right hand digits of the pseudorandom numbers on the interval  $(0, 2^{35})$  more random than using a modulus of  $2^{35}$ .

The generator used in this study could perhaps be improved by allowing  $X_0$  to vary over the entire range  $(0, 2^{35})$ .

Future research might benefit from using one of

MacLaren & Marsaglia's (1965) two alternative methods for pseudorandom number generation, a combined congruential method and a method using a table of uniform random numbers. They claim that such alternative methods produce pseudorandom numbers with better statistically random performance (Marsaglia 1968, 1970). Their methods certainly have the potential of achieving this, although as noted previously, direct comparisons have not shown them to be superior to the multiplicative congruential method used here.

#### Control of Type I Error Rate

This section considers the main results of this study, i.e., those related to the major hypotheses. First the results concerning familywise Type I error rate are considered, then the results concerning conditional Type I error rate.

#### Familywise Type I Error Rate

Method I. Method I for controlling Type I error rates and setting the corresponding critical values was the customary procedure of setting  $\alpha_T = .05$  for each individual significance test of a hypothesis

$$\begin{aligned} H_0 : \rho_{ij} &= 0 \\ H_1 : \rho_{ij} &\neq 0 \end{aligned} \tag{75}$$

Using Method I, the familywise Type I error rate increases rapidly as the number of true component null hypotheses increases. The present empirical study showed this as summarized in Table 7 in Chapter III. In this study whenever there was more than one true component null hypothesis, the familywise Type I error rate was almost always significantly ( $p < .01$ ) greater than .05. In three cases with 21 true component null hypotheses each, familywise empirical size (which is an empirical estimate of familywise Type I error rate) ranged from .669 to .755. So if the intercorrelations among eight variables are being analyzed and if the complete null hypothesis is true ( $R_{mp} = I$ ), then at least one Type I error occurs in the analysis about  $\frac{2}{3}$  to  $\frac{3}{4}$  of the time. Psychologists do not often analyze eight completely unrelated variables. However, it is not uncommon in the literature to analyze a much larger number of intercorrelations which could easily include 21 true component null hypotheses. How many true component null hypotheses actually exist would be unknown to the experimenter. The important point is that the familywise Type I error rate increases above .05 (given  $\alpha_T = .05$ ) if there are even two true component null hypotheses. If more component null hypotheses are true, then the familywise Type I error rate reaches proportions at least as high as .75. In all cases, however, the only reported Type I error rate is usually .05, the Type I

error rate per component test.

Using Method I to control Type I error rates is very similar to using  $t$ -statistics for all pairwise comparisons in analysis of variance with  $\alpha_T = .05$  for each  $t$  - test. Because of similar results in familywise Type I error rates, various methods for multiple comparisons in ANOVA have been proposed to control  $\alpha_{FW}$  more stringently. The rationale in favor of these more conservative procedures in analysis of variance should also be applied against Method I in statistical analyses of intercorrelations. The fact that multiple-test procedures are widely used in analysis of variance but used hardly at all in analyses of intercorrelations indicates that the rationale for simultaneous procedures has been applied inconsistently to analysis of variance and correlational analysis.

Methods II and III. Methods II and III were conservative alternatives to Method I for controlling the Type I error rate. Method II set  $\alpha_T$  such that  $\alpha_{FW}$  would equal .05 if all the component significance tests were mutually independent. Method III used the Bonferroni inequality to set  $\alpha_T$  such that  $\alpha_{FW}$  would be less than or equal to .05 regardless of any dependencies among the component significance tests. The results of this empirical study indicate that in practice these two methods

give almost identical conclusions. In the cases examined in this study Methods II and III led to different conclusions for only about one out of 1000 significance tests. If Methods II and III were calculated using a larger  $\alpha_{FW}$  and if the number of significance tests in the family of tests were large then the two methods would differ more in their results. However, if the experimenter wants to control  $\alpha_{FW}$  at the .05 level, it apparently makes little practical difference whether Method II or III is used.

The alternate methods (Methods II and III) resulted in familywise empirical sizes which generally were near the desired .05 level when the complete null hypothesis was true ( $R_p = \underline{I}$ ). The familywise empirical sizes for Method III were significantly different ( $p < .05$ ) from .05 for 3 out of 18 cases (9 correlation-matrix cases, 9 correlation-vector cases) for which  $R_p = \underline{I}$ . While this is a higher proportion than we would usually expect, it is not significantly higher ( $p < .06$ , binomial distribution probability of 3 or more successes of 18 observations with  $p = .05$ ). Furthermore, the deviant familywise empirical sizes occurred when shorter sequences of pseudorandom numbers were used, suggesting that this may be due to the under-generation of Type I errors rather than due to the Bonferroni method itself.

In cases for which the complete null hypothesis is

not true (i.e.,  $R_{mp} \neq I$ ), Methods II and III are consistently over-conservative in controlling for Type I errors. In 28 out of the 38 cases with  $R_{mp} \neq I$ , familywise empirical size was significantly ( $p < .05$ ) less than .05. The less the number of true component null hypotheses proportionately, the more the Bonferroni method tended to be over-conservative. Table 8 shows that the lower the proportion of true component null hypotheses among the null hypotheses to test, the less the mean familywise empirical size. These findings support Miller's (1966) observation that the Bonferroni  $\underline{t}$  test is unnecessarily conservative unless a multistage procedure is used. Later, a multistage Bonferroni  $\underline{t}$  method will be discussed, which would correct for over-conservativeness.

Note here also that the Bonferroni  $\underline{t}$  method was also over-conservative for the family of significance tests

$$\begin{aligned} H_0 : \rho_{il} &= 0 \\ H_1 : \rho_{il} &\neq 0 \end{aligned} \quad i = 2, 3, \dots, k \quad (76)$$

when all  $\rho_{il} = 0$ , but some  $\rho_{ij} \neq 0$  (i.e., the family of tests involve intercorrelation between  $k-1$  predictor variables and one criterion variable, with some predictors correlated with each other). This was the situation for the population correlation matrices described in equations (51)



through (54). For 5 of the 11 cases using the first column of these matrices, the familywise empirical size was significantly less ( $p < .05$ ) than the nominal  $\alpha_{FW}$  of .05. So we must conclude that  $\alpha_{FW}$  is actually less than .05 for such a family of tests. (A word of caution, however; this finding may again be a result of the undergeneration of total Type I errors for smaller sequences of pseudorandom numbers. The 11 cases included 7 cases based on relatively smaller sequences of pseudorandom numbers (160,000 or less)).

#### Conditional Type I Error Rate

This section focuses on the effects on conditional Type I error rate of using Method I to control for Type I error. Conditional Type I error rate is the Type I error rate on one component test given that a Type I error occurs on another component test. Methods II and III are not considered in this section. Similar effects would occur with Methods II and III, but the effects are counteracted somewhat by controlling  $\alpha_{FW}$  at a given level rather than controlling only  $\alpha_T$  as Method I does. Not enough Type I errors occurred in this study with Methods II and III to make a meaningful analysis of conditional Type I error rate for those methods possible.

Using Method I, the conditional Type I error rate was greater than the unconditional Type I error rate ( $\alpha_T$ ) of .05 for tests of

$$\begin{aligned}
 H_0 &: \rho_{ij} = 0 \\
 H_1 &: \rho_{ij} \neq 0
 \end{aligned}
 \tag{77}$$

and

$$\begin{aligned}
 H_0 &: \rho_{ik} = 0 \\
 H_1 &: \rho_{ik} \neq 0
 \end{aligned}
 \tag{78}$$

when the actual population correlation matrix among the three variables was

$$\begin{array}{c}
 \begin{array}{ccc}
 & i & j & k \\
 i & \left[ \begin{array}{ccc}
 1 & & \\
 0 & & \\
 0 & \pm .6 & 1
 \end{array} \right] & & \\
 j & & & \\
 k & & &
 \end{array}
 \end{array}
 \tag{79}$$

This represents the case in which the first of three variables is actually uncorrelated with the second and third variables, while the second and third variables are correlated  $\pm .6$ . The conditional Type I error rate may be the Type I error rate in testing equation (78) conditional on a Type I error in testing equation (77) or vice versa (the choice between these two is arbitrary). The configuration of equation (79) may be the entire inter-correlation matrix of interest or may be embedded in a

larger intercorrelation matrix. In any case the conditional Type I error rate deviates greatly from the unconditional Type I error rate for such cases of "strongly-linked" component tests. Table 10 summarized all the cases in which this configuration (equation (77)) was embedded in a larger correlation matrix. Almost all of these had significantly larger numbers of conditional Type I errors than expected. As noted previously, the mean conditional empirical size was .242. This indicates that in such a configuration as equation (77), if a Type I error occurs on a test of  $P_{ij}$ , then the probability of a Type I error on a test of  $P_{ik}$  increases to about .242, i.e., to about  $\frac{1}{4}$  of the time. This means that the probability of Type I errors occurring simultaneously on both tests for the same sample is much higher than would be expected if the tests were independent. Both Type I errors would occur simultaneously about 1.21% of the time, rather than .25% of the time, which would be the case if the component tests were independent (given  $\alpha_T = .05$ ).

Equations (52) and (53) were two 3 x 3 correlation matrices that fit the configuration of equation (79). The number of times that two Type I errors occurred on the same sample is reported in Table 9. The five computer runs involving these matrices had a mean percentage of trials with two Type I errors of .94%. Similarly, the estimated mean conditional empirical size was also a little lower than would be expected from the above findings, about .202.

Three of these five computer runs had smaller sequences of pseudorandom numbers, so this difference may have been partially due to undergeneration of total Type I errors. Harris (1967) had six computer runs using the population correlation matrices of equations (52) and (53) with  $\alpha_T = .05$  and sample sizes of 25, 100, and 200. The mean estimated conditional empirical size from his data is .238.

Combining Harris' (1967) results with this study's results, the conditional Type I error rate is about .24 ( $\alpha_T = .05$ ) for the second significance test given a configuration such as equation (79) and a Type I error on the first significance test. This will obviously change when the value of  $\rho_{jk}$  (the nonzero correlation coefficient in equation (79)) is different. If the absolute value of  $\rho_{jk}$  is greater than .6 the conditional Type I error rate will be greater; if  $\rho_{jk}$  is less than .6, the conditional Type I error rate will be smaller. The only empirical estimate of such a change from this study is based on the three computer runs using equation (51). With  $\rho_{jk} = .3$ , the mean percentage of two simultaneous Type I errors was .533% and the mean estimated conditional empirical size was .114. While this is based on a very few cases, it suggests that even such small values of  $\rho_{jk}$  increase conditional Type I error rate to more than twice the stated  $\alpha_T$ . These empirical estimates along with estimates from Harris' (1967) data give the results

Table 16. Estimated Mean Conditional Empirical Size for Various  $\rho_{jk}$

$\rho_{jk}$	Estimated Mean Conditional Empirical Size
.0	.05
.2	.05
.3	.11
.4	.15
±.6	.24
±.9	.52

Note. -  $\rho_{jk}$  is the nonzero correlation coefficient in a 3 x 3 matrix such as equation (79). These estimated mean conditional empirical sizes are based on data from this study and from Harris (1967).

summarized in Table 16. While some of these estimates of conditional Type I error rates are based on limited information, it gives some idea of the effect of the magnitude of  $\rho_{jk}$  on conditional Type I error rates in testing such hypotheses as equations (77) and (78).

Note that these considerations of conditional empirical Type I error rate are applicable not only when an intercorrelation matrix is of interest, but also when a correlation vector is of interest. A correlation vector is of interest, for example, whenever a researcher is interested in the correlations between two or more predictor variables and one or more criterion variables. In such studies there is often no information given concerning correlations between predictor variables or correlations

between criterion variables, both of which could correspond to  $\rho_{jk}$  in equation (79). So it is generally difficult to estimate the probability of observing a group of Type I errors in a given sample, since the relevant  $r_{jk}$  is not reported. But the problem of conditional Type I error is just as relevant in such cases.

The findings of the present study suggest that conditional Type I error rates may be affected for other pairs of correlation coefficient tests, too. For example, Table 11 summarized the conditional empirical size results for "moderately-linked" intercorrelations. Two correlations  $r_{ij}$  and  $r_{kl}$  would be "moderately-linked" if they involve four distinct variables (e.g.,  $i$ ,  $j$ ,  $k$ , and  $l$ ) with each variable (e.g.,  $i$ ) in the first correlation of interest correlated .6 with one variable (e.g.,  $k$ ) in the second correlation of interest. While only a few of the results in Table 11 were significantly different ( $p_{FW} < .05$ ) from the unconditional Type I error rate, the mean conditional empirical size for these cases was .126. This suggests that conditional Type I error rates may be affected by "moderately-linked" intercorrelations, although not conclusively so from this data alone. Harris (1967) also has data that fits this definition of "moderately-linked" intercorrelations. His data yields an estimated mean conditional empirical size of .270. Such a high value appears somewhat dubious since it is not logical for the conditional Type I error rate for "moderately-linked" intercorrelations to be higher than

Table 17. Frequencies of Various Numbers of Type I Errors in Certain Intercorrelation Samples<sup>a</sup>

Number of Type I Errors	Frequency			
	Method I		Method III	
	Expected <sup>b</sup>	Observed	Expected <sup>b</sup>	Observed
0	1261	1513	1941	1957
1	597	301	59	31
2	126	109	0	8
3	15	41	0	3
4	1	13	0	1
5	0	10	0	0
6	0	7	0	0
7	0	4	0	0
8	0	2	0	0

<sup>a</sup>Intercorrelation samples in two computer runs with

$$R_p = \begin{bmatrix} 1 & & & & \\ .6 & & & & \\ .6 & .6 & & & \\ 0 & 0 & 0 & & \\ 0 & 0 & 0 & .6 & \\ 0 & 0 & 0 & .6 & .6 & 1 \end{bmatrix},$$

one with  $N = 40$ , and one with  $N = 100$ .

<sup>b</sup>Expected frequencies for 2000 replications of 9 independent significance tests according to the binomial distribution.

that for "strongly-linked" intercorrelations and since it deviates so much from the results of the present study. However, the possibility remains that conditional Type I error rates for "moderately-linked" intercorrelations may

be much higher than the estimate from this present data.

In actual intercorrelation matrices 4 x 4 and larger, there could be complex relationships of "strongly-linked" and "moderately-linked" intercorrelations and other dependent interrelationships. One such example would be the population correlation matrix of equation (58). Table 17 gives the frequencies of various numbers of Type I errors for the two computer runs using this population correlation matrix. It can be seen that there was an unexpectedly large frequency of three or more Type I errors occurring simultaneously on a sample intercorrelation matrix. This is the result of a two-fold problem: 1) the non-multiple-test procedure of setting  $\alpha_T = .05$  ensures a high probability of at least one Type I error, and 2) the moderately high non-zero population correlation coefficients result in high conditional Type I error rates. Consequently, there were three or more Type I errors (out of nine possible) on 3.85% of the sample correlation matrices despite a reported alpha ( $\alpha_T$ ) of .05.

This result illustrates a need for a method of controlling Type I error that takes conditional Type I error rates into account. It can be seen from the right-hand columns of Table 17 how Method III, the Bonferroni  $\frac{1}{k}$  method, would control Type I error rates in this particular example. The effect of the dependence of the component tests causes this also to deviate sharply from the expected frequencies of number of Type I errors. However,



by keeping the probability of the first Type I error below .05, Method III improves greatly over Method I in controlling against multiple Type I errors.

#### Control of Type II Error Rate

The sample sizes, number of variables, and population correlation matrices for this study were chosen with effects on Type I error rates primarily in mind. This study does not provide an adequate analysis of the relative power of Methods I, II, and III. It is clear that Methods II and III are nearly identical in power for the kinds of intercorrelation matrices examined here. Also it is clear then that whatever is gained in controlling Type I error rate is gained at the expense of Type II error rate. This is to be expected since the three methods differ only in setting the critical value for the rejection of a component null hypothesis.

Although Method I fares poorly in controlling Type I error rates, it is the best of the three methods for controlling Type II error rates. This supports Miller's (1966) contention that some justification can be given to Method I if the total loss for a sequence of hypothesis tests is the sum of the component losses and a Bayesian approach is taken. So the possibility remains that a Bayesian approach would yield a better solution. However, Miller (1966) thinks otherwise, and the major Bayesian multiple-test procedure to date (Waller & Duncan, 1969) resembles Fisher's

protected Least Significant Difference method, a procedure which loses all conservativeness once any significant effect is found.

### Conclusions

The major conclusions will be reviewed at this point: Method I, the customarily used procedure of setting  $\alpha_T = .05$ , results in a large familywise error rate. This familywise Type I error rate increases quickly to undesirable levels as the number of variables increases (and thus the number of true component null hypotheses increases). Method II, which is based on a false assumption of independent significance tests, and Method III, the Bonferroni  $\frac{1}{k}$  procedure, successfully keep the familywise Type I error rate at .05 or below. However, both of these methods over-control for Type I error when even a small proportion of the component null hypotheses in a family of tests are false. The mutual dependence of the component significance tests in an intercorrelation matrix or an intercorrelation vector is an important factor if any correlation between any variables involved is moderate or large in magnitude. This dependence may dramatically increase conditional Type I error rates over unconditional Type I error rate levels.

### Recommended Procedure for Controlling

#### Type I Error Rates

Taking these major results into account, this writer recommends a multistage modification of Method III, the Bonferroni  $\frac{1}{k}$  method. Method III is chosen over Method I because of its superiority in keeping the familywise Type I error rates at .05 or below. As noted previously, this becomes even more crucial when the mutual dependence of the component significance tests is an important factor. The multistage modification is recommended to counter the major drawback of Method III, which is its over-conservativeness under certain conditions. This multistage modification will be described in detail in the next section.

Method II is nearly identical to Method III for all practical purposes. These two methods result in different acceptance versus rejection decisions only about .1% of the time. Furthermore, Method II is slightly more powerful and has been shown to control  $\alpha_{FW}$  to .05 or less for all 3 x 3 matrices and for all matrices without negative population correlation coefficients (Dunn & Massey, 1965). However, Method III is recommended over Method II because it is more widely known (e.g., in multiple comparisons in analysis of variance), it gives a conservative estimate of  $\alpha_{FW}$  regardless of the dependence of the significance tests, and it is easier to use.

Finally, a word about borderline results is here given. A borderline result is one that would result in a rejected null hypothesis by Method I but that results in

an accepted null hypothesis by the multistage modification of Method III. As illustrated in Figure 1, such a borderline result is analogous to a result in a single significance test that falls exactly at the critical point. As such, borderline results should not be lumped together with "non-significant" results but should be placed in a category between "non-significant" results and conclusively "significant" results. Borderline results should be considered as conclusive evidence only when considered together with similar (or more conclusive) results from other samples.

#### Applications to Extant Data

In this section the findings of this study will be related to two examples of correlational analysis taken from Brooks (1973) and Jessor & Jessor (1974). An example will be given of how to use a Bonferroni multistage procedure, using the Jessor & Jessor (1974) data.

Jessor & Jessor's (1974) article dealt with the relationships between maternal ideology and adolescent problem behavior. Table 18, which is reproduced from Jessor and Jessor, summarizes their analysis. Note that the reported intercorrelations are between five predictor variables (maternal ideology and home climate) and three criterion variables (adolescent problem behavior) for two separate samples (males and females). For each sample, the 15 reported correlations are part of the 28 possible

Table 18. Product-Moment Correlations  
between Maternal Socialization Measures and  
Junior and Senior High School Student  
Behavior and Attitudes

Maternal socialization measure	Problem behavior index		Student measure* Total negative functions		Total positive functions	
	Females	Males	Females	Males	Females	Males
Ideology						
Mother's traditional beliefs	-.29**	-.34***	.35***	.03	-.13	-.23**
Mother's religiosity	-.32***	-.20*	.23**	.08	-.18	-.16
Mother's attitude toward deviance	.42****	-.13	.33***	.05	-.23**	-.22**
Home climate						
Mother's controls and regulations	.22**	-.30***	.29***	.18*	-.11	-.17
Mother's affectionate interaction	.28***	-.05	.12	.21**	.22**	-.03

Note. - Reprinted from Jessor & Jessor (1974, p. 251).

- \*Note:  $p$  values are based on two-tailed tests.  
 \* The  $n$  for females ranges from 75 to 91 for the different measures, for males, the  $n$  ranges from 79 to 93.  
 \*  $p < .10$   
 \*\*  $p < .05$   
 \*\*\*  $p < .01$   
 \*\*\*\*  $p < .001$

Table 19. Hypothetical Population  
Correlation Matrix for Maternal Ideology and Control  
and Adolescent Problem Behavior

Total Positive Functions	1						
Total Negative Functions	-.32						
Problem Behavior Index	.25	-.465					
Traditional	0	0	0				
Religiosity	0	0	0	.505			
Attitude Toward Deviance	0	0	0	.505	.505		
Controls	0	0	0	.30	.30	.30	
Affection	0	0	0	.01	.01	.01	.175
							1

Note. - The non-zero correlations are based on estimates primarily from Jessor and Jessor (1974).

intercorrelations among the eight variables.

Now suppose that all the population correlations of interest were actually zero. Then the population correlation matrix might look something like Table 19. If this were the actual population correlation matrix, what would be the probability of observing as many significant results (which in such a case would be Type I errors) as Jessor & Jessor (1974) did? To answer this question partially, a computer simulation was performed using the same program as the main part of this present study, but with the population correlation matrix of Table 19,

Table 20. Frequencies of Number of Type I Errors in Computer Simulation of Samples from the  $R_p$  in Table 19

Number of Type I Errors in a Sampling Replication	Frequencies of Given Number of Type I Errors	
	Observed	Expected <sup>a</sup>
0	263	206
1	329	343
2	211	267
3	116	129
4	38	43
5	32	10
6 or more	11	2

Note. - Total sampling replications = 1000;  $\alpha_T = .10$ .

<sup>a</sup>Assuming mutually independent component tests

with  $N = 85$ , and with  $\alpha_T = .10$ . The sample size chosen was the average of the sample sizes reported in Table 18 and the alpha level of .10 was the one used by Jessor & Jessor (1974) to determine a "significant" result. Table 20 summarizes the results of this computer simulation. Note that there was at least one Type I error on almost three-fourths (73.7%) of the sampling replications. Furthermore, there were five or more Type I errors (out of fifteen possible) on 4.3% of the sampling replications. By comparison, Jessor & Jessor (1974) reported 7 correlations significantly different from zero in their male sample and 11 in their female sample. If all the relationships of interest are actually zero (i.e., Table 19 represents the actual relationships), there is less than a .5% chance of obtaining 7 significant results (as in their male sample) and a negligible chance of obtaining 11 significant results (as in their female sample). So there is no reasonable basis for suggesting that Jessor & Jessor's (1974) results are entirely Type I error artifacts. However, this example is useful to show how this type of analysis compares with others in controlling Type I errors.

This type of analysis is compared first with an analysis involving only one significance test of a correlation coefficient. Such a single-test analysis would result in a Type I error about 10% of the time (given  $\alpha_T = .10$ ;  $\rho_{ij} = 0$ ). As noted previously, the Type I error

rate in a multiple-test situation (such as Jessor & Jessor's) differs from a single-test situation because

- 1) multiple tests greatly increase the likelihood of the occurrence of a Type I error even if the tests are mutually independent and
- 2) the interdependence of the component tests affects the Type I error rate in generally unknown ways.

The right-hand column of Table 20 gives the expected frequencies (out of 1000 sampling replications) of various numbers of Type I errors if the component tests were mutually independent. Note that if the independence assumption were justified, there would still be a .055 probability of observing 4 or more Type I errors out of 15 possible Type I errors. But the independence assumption is not justified, since the observed frequencies in Table 20 do not adequately approximate the expected frequencies ( $\chi^2_{(5)} = 101.97, p < .001$ ). From the observed frequencies, there is approximately a .043 probability of observing five or more Type I errors.

Now compare this with a single-test analysis. If the correlation of interest were zero in a single-test analysis, a Type I error would be obtained 10% of the time (given  $\alpha_T = .10$ ). However, if the 15 correlations of interest to Jessor & Jessor (1974) were all zero, four or more Type I errors would be obtained about 8.1% of the time (given  $\alpha_T = .10$ ). Furthermore, most of the time (about 65.5%) one to three Type I errors would be obtained out



of the 15 possible, whereas there would be no Type I errors 90% of the time in a single-test analysis. The high probability of at least one Type I error in a multiple-test situation and the substantial probability of several simultaneous Type I errors are not obvious from the nominal probability level of .10.

Thus far Jessor & Jessor's (1974) analysis (a Method I analysis) has been compared only with a single-test analysis. At this point Method III, the Bonferroni  $\dagger$  method, is applied to the Jessor & Jessor data. Because of the over-conservativeness of this method, a multistage modification is applied. At the first stage of the multistage procedure, Method III is applied in the usual way. However, if any component null hypotheses are rejected at the first stage, a second stage is then performed with new critical values based on the number of remaining non-rejected component null hypotheses. By this method, the familywise Type I error rate remains at or below the nominal probability level (e.g., .10) for any possible set of true component null hypotheses without being unnecessarily conservative. The significance tests of fifteen correlations of interest on one sample (male or female) is considered a family of tests. The alpha level per test at the first stage of the analysis is calculated by the equality of equation (11),

$$\alpha_T = \frac{\alpha_{FW}}{m} = \frac{.10}{15} = .006\frac{2}{3} \quad (80)$$

Since two-tailed significance tests are appropriate in this analysis,  $\alpha_T$  is divided by two,

$$\frac{\alpha_T}{2} = .003\frac{1}{3} . \quad (81)$$

While the two critical values of the  $t$  distribution cannot be obtained from readily available tables, they can be approximated by equation (13):

$$\begin{aligned} t_{\alpha_T/2, \nu} &= z_{\alpha_T/2} + \frac{z_{\alpha_T/2}^3 + z_{\alpha_T/2}}{4(\nu - 2)} \\ &= \pm 2.71 \pm \frac{19.90 + 2.71}{4(81)} \\ &= \pm 2.78 \end{aligned} \quad (82)$$

Next the formula

$$r_{\alpha_T/2} = \frac{t_{\alpha_T/2, \nu}}{\sqrt{N - 2 + t_{\alpha_T/2}^2}} \quad (83)$$

can be used to obtain a critical value of the sample correlation coefficient. In this case,

$$r_{\alpha_T/2} = \frac{\pm 2.78}{\sqrt{85 - 2 + (2.78)^2}} = \pm .292 \quad (84)$$

Consider the data for the male sample first. Two of the fifteen component tests result in rejection of the component null hypothesis in the first stage of the analysis. They correspond to the two sample correlations which are larger than .292 in absolute value. If no component null hypothesis were rejected at this first stage, the analysis for the male sample would be terminated. However, since two component null hypotheses are rejected at the first stage, the procedure continues on to the second stage.

The analysis at the second stage is done just like the analysis at the first stage except that  $\alpha_T$  is computed with a value of 13 for  $m$  in equation (80). Thirteen is the remaining number of non-rejected component null hypotheses. The computations of equations (80) through (84) are repeated again using this new value for  $m$ . This results in  $\frac{\alpha_T}{2} = .00385$ ,  $t_{\alpha_T/2, \nu} = \pm 2.74$ , and  $r_{\alpha_T/2} = \pm .288$ . None of the sample correlations corresponding to the 13 previously non-rejected component null hypotheses are larger than the new critical value in absolute magnitude. Since no further component null hypotheses were rejected, the multistage procedure is terminated at this point. The conclusion of this analysis is that two of the fifteen sample correlations are considered significantly different from zero at the  $\alpha_{FW} = .10$  level.

The first stage of the multistage Bonferroni procedure for the female sample is identical to the first stage for the male sample. Equation (84) gives the cri-

tical value of  $r_{\alpha_T/2} = \pm .292$ . Using this critical value, four component null hypotheses are rejected for the female sample. Therefore, the multistage procedure proceeds to the second stage, which uses  $m = 11$  in recalculating equations (80) through (84). Two of the sample correlations (.29 and -.29) are larger than the new critical value ( $\pm .281$ ) in absolute value. Therefore, a third stage is performed which in turn results in the null hypothesis being rejected for the sample correlation of  $-.28$ .

Table 21. Summary of a Multistage Bonferroni Analysis Applied to Jessor & Jessor's (1974) Female Sample<sup>a</sup>

Stage	m	$\frac{\alpha_T}{2}$	$r_{\alpha_T/2}$	Continue Multistage Procedure?
1	15	.00333	$\pm .292$	yes
2	11	.00455	$\pm .281$	yes
3	9	.00556	$\pm .274$	yes
4	8	.00625	$\pm .270$	no

$$^a \alpha_{FW} = .10$$

Since an additional null hypothesis was rejected at the third stage, a fourth stage is performed, with  $m = 8$ . At this stage no additional null hypotheses are rejected, so the procedure is terminated. If another null hypothesis would have been rejected at the fourth stage, a fifth stage would have been performed. This would continue until a stage is reached in which no additional null hypo-

theses are rejected. A summary of the four stages of the multistage Bonferroni procedure is presented in Table 21. Seven of the sample correlations are considered to be significantly different from zero at the  $\alpha_{FW} = .10$  level.

The first stage of the multistage Bonferroni procedure is identical to the usual Bonferroni procedure, which has been called Method III in this study. The above examples illustrate how the multistage modification increases power over its non-multistage alternative. The multistage procedure for the female sample terminated with a final critical value of  $\pm .270$  instead of a critical value of  $\pm .292$ , which a non-multistage Bonferroni analysis would give. The multistage procedure increases power without increasing the familywise Type I error rate above the nominal level (.10 in this case), as shown in Appendix A.

Table 22 gives the results of the multistage Bonferroni procedure for the female sample as they would be presented in a publication. A couple of features of this table facilitate comparisons with the more common correlational analysis procedure (Method I). First, in the footnote section of the table, the equivalent Type I error rates per test are given for each familywise Type I error rate. This information tells the reader, for example, that a significant result with  $p_{FW} < .10$  in this analysis is equivalent to a significant result with

Table 22. Product-Moment Correlations Between Maternal Socialization Measures and Junior and Senior High School Student Behavior and Attitudes (Females; Multistage Bonferroni Analysis)<sup>a</sup>

Maternal Socialization Measure	Student Measure		
	Problem Behavior Index	Total Negative Functions	Total Positive Functions
Ideology			
Mother's Traditional Beliefs	-.29*	.35**	-.23 <sup>b</sup>
Mother's Religiosity	-.32**	.23 <sup>b</sup>	-.16
Mother's Attitude Toward Deviance	-.42****	.33**	-.22 <sup>b</sup>
Home Climate			
Mother's Controls and Regulations	-.22 <sup>b</sup>	.29*	-.17
Mother's Affectionate Interaction	-.28*	.12	-.03

Note.- Two-tailed multistage Bonferroni procedure.  $p_{FW}$  is based on familywise Type I error rate;  $p_T$  is based on Type I error rate per test.

<sup>a</sup>The n was assumed to be 85 for this analysis.

<sup>b</sup>Borderline significance;  $p_T < .05$

\* $p_{FW} < .10$ ;  $p_T < .0125$

\*\* $p_{FW} < .05$ ;  $p_T < .0045$

\*\*\* $p_{FW} < .01$ ;  $p_T < .00036$

\*\*\*\* $p_{FW} < .001$ ;  $p_T < .00004$

$p_T < .0125$  in the common correlational analysis procedure (Method I). Secondly, results for which the null hypothesis would be rejected according to the usual non-multiple-test procedure are considered in Table 22 as borderline results. These results would fall in the borderline region of Figure 1. They do not provide strong enough evidence by themselves to conclusively reject the

**Table 23. Correlations between Childhood Ratings of "Satisfactions in Artistic Pursuits" and Adult Q Sort Items (Males).**

Adult Q-Sort Items	Satisfactions in Artistic Pursuits Age in Years		
	8-11 N = 35	11-14 N = 35	14-18 N = 34
1. Is critical, skeptical.	.33*	.21	.27
27. Shows condescending behavior to others.	.32*	.18	.30*
17. Behaves in sympathetic, considerate manner.	-.30*	-.25	-.22
43. Is facially, gesturally expressive.	.30*	.23	.29*
33. Is calm, relaxed in manner.	-.29*	-.29*	-.17
94. Expresses hostile feelings directly.	.29*	.13	.12
29. Is turned to for advice.	-.28*	-.24	-.17
#66. Enjoys esthetic impressions, esthetically reactive.	.21	.45***	.64***
100. Does not vary roles; relates to everyone in same way.	-.27	-.40**	-.12
13. Is thin-skinned; vulnerable to slight.	.23	.31*	.11
24. Prides self on being objective, rational.	-.11	-.31*	-.11
50. Is unpredictable, changeable in behavior, attitude.	.21	.30*	.24
# 3. Has a wide range of interests.	.17	.28*	.36**
47. Tends to feel guilty.	-.23	-.12	-.35**
#63. Judges self, others in conventional terms.	.09	-.19	-.32*
18. Initiates humor.	.12	.15	.30*
84. Is cheerful.	.08	.13	.30*
15. The light touch as compared to the heavy touch.	.05	.11	.29*
41. Is moralistic.	-.04	-.27	-.29*

**Note.** - Reprinted from Brooks (1973, p. 116).

\*Significant at .10 level.

\*\*Significant at .05 level.

\*\*\*Significant at .01 level.

#Items also significantly correlated for females.

**Table 24. Correlations between Childhood Ratings of "Satisfactions in Artistic Pursuits" and Adult Q Sort Items (Females)**

Adult Q-Sort Items	Satisfactions in Artistic Pursuits Age in Years		
	8-11 N = 38	11-14 N = 38	14-18 N = 37
51. Genuinely values intellectual, cognitive matters.	.44***	.51***	.46***
8. Appears to have a high degree of intellectual capacity.	.40**	.39**	.38**
54. Emphasizes being with others; gregarious.	-.40**	-.45***	-.37***
# 3. Has a wide range of interests.	.39**	.47***	.46***
59. Is concerned with own body and adequacy of functioning.	-.35**	-.19	-.24
93. Behaves in a feminine style and manner.	-.35**	-.25	-.18
#66. Enjoys esthetic impressions; is esthetically reactive.	.33**	.20	.32*
90. Is concerned with philosophical problems.	.33**	.42***	.35**
39. Thinks and associates to ideas in unusual ways.	.32**	.41**	.33*
11. Is protective of those close to him.	-.30*	-.10	-.06
5. Behaves in a giving way towards others.	-.29*	-.12	-.03
69. Is battered by demand.	.29*	.15	.05
22. Feels a lack of personal meaning in life.	.28*	.25	.01
#63. Judges self and others in conventional terms.	-.21	-.39**	-.25
60. Has insight into own motives.	.12	.32**	.29*
16. Is introspective.	.25	.31*	.26
7. Favors conservative values in a variety of areas.	-.18	-.30*	-.20
57. Is an interesting, arresting person.	.08	.29*	.30*

**Note.** - Reprinted from Brooks (1973, p. 117).

\*Significant at .10 level.

\*\*Significant at .05 level.

\*\*\*Significant at .01 level.

#Items also significantly correlated for males.

null hypothesis. However, when considered together with similar results from different samples they might constitute just as conclusive evidence. But they are not conclusive from this one investigation alone.

With the commonly-used procedure for correlational analysis, it is possible to obtain a large number of "significant" results which actually could be due to chance alone, simply by examining a large enough number of variables. For example, Brooks (1973) reported the data in Tables 23 and 24. This data is taken from the Berkeley Guidance Study. The correlations are between measures of adult functioning at age 30 and satisfactions in artistic pursuits at each of the three adolescent age periods. Tables 23 and 24 only include the adult functioning variables which showed at least one significant correlation with an artistic interest variable. Actually, 100 adult functioning variables were used in the investigation. Consequently, for each sample (males or females) there were 300 sample correlations of interest (100 adult functioning variables X 3 artistic interest variables). Since an alpha level of .10 was used for each component significance test, the expected number of Type I errors would be 30. For each sample Brooks (1973) found 24 "significant" results for the male sample and 33 for the female sample. This by itself suggests that almost all the "significant" results are actually Type I errors. However, this conclusion is obscured



by leaving out all variables which did not correlate with artistic interest at any age, by the apparent consistencies across ages and by the fact that three variables correlated significantly with artistic interests for both the male and the female samples. The apparent consistencies across ages is especially noticeable for the female sample. Six variables correlated significantly with artistic interest at all three age levels and three others correlated significantly at two age levels. This is probably an artifact, due to conditional Type I error rate. Brooks (1973) reports that the average intercorrelation of the ratings of female artistic interests for the three age periods was .76. So for any one unrelated variable, the population correlation matrix might be

$$R_p = \begin{matrix} \text{unrelated} \\ \text{variable} \\ \text{1st age} \\ \text{2nd age} \\ \text{3rd age} \end{matrix} \begin{bmatrix} 1 & & & & \\ & 0 & & & \\ & & 0 & .76 & \\ & & & 0 & .76 & .76 \\ & & & & & 1 \end{bmatrix} \quad (85)$$

This closely resembles some of the population correlation matrices that have been used in this present study (e.g., equation (54)), except that the nonzero correlations are even higher. Therefore, the effect of the dependence of the component significance tests on conditional Type I error rates would be even more pronounced than in the ex-

amples used in this present study. This means that if a Type I error did occur on one significance test of one intercorrelation in Table 24, a Type I error would be much more likely to occur on another component test that involves the same adult functioning variable. This could very easily account for the apparent consistencies across adolescent age levels.

As for the fact that three adult functioning variables related "significantly" to artistic interest variables for both males and females, this could also have easily occurred by chance. Since 19% of the variables were related to artistic interest for males and 18% for females, 3.42% would be expected to overlap by chance ( $.19 \times .18 = .0342$ ).

If a multistage Bonferroni procedure were used on Brook's (1973) data with  $\alpha_{FW} = .10$ , only one result for either sample would be significant, the .64 correlation between adult functioning variable 66 and artistic interest in males at 14-18 years (see Table 23). Again, the other correlations which were deemed significant by Brooks' (1973) analysis could be classified as borderline results.

These two examples from the literature illustrate the main conclusions of this study. First, as the number of component significance tests increase, the probability of observing one or more Type I errors increases rapidly regardless of any dependence of the component tests.

Secondly, the dependence among significance tests of intercorrelation further complicates the Type I error rates. Generally, the dependence increases the probability of a relatively large proportion of Type I errors occurring simultaneously. Thirdly, the high conditional Type I error rates that can occur with certain intercorrelation patterns can lead to some apparent regularity in results which would otherwise be discarded as probable Type I error.

Multiple-test procedures have been widely recommended for multiple comparisons in analysis of variance for similar reasons. Some multiple-test procedure seems to be the best kind of solution for the problems in the currently-used procedure for intercorrelational analysis (Method I). A multistage Bonferroni procedure has been outlined and recommended.

## APPENDIX A

PROOF ABOUT THE TYPE I ERROR CONTROL OF  
THE MULTISTAGE BONFERRONI PROCEDURE

Let there be  $M$  component significance tests in a family of tests of intercorrelations. Assume each component significance test is of the form

$$\begin{aligned} H_0 : \rho_{ij} &= 0 \\ H_1 : \rho_{ij} &= 0 \quad . \end{aligned} \tag{A.1}$$

Assume further that there are  $n$  true component null hypotheses ( $n \leq M$ ). Specify some nominal value of  $\alpha_{FW(N)}$  (e.g., .05) to be used in the calculation of  $\alpha_T$  by equation (80). We want to show that the actual familywise Type I error rate ( $\alpha_{FW(A)}$ ) does not exceed the nominal familywise Type I error rate ( $\alpha_{FW(N)}$ ) for any value of  $n$ ,  $0 < n \leq M$ .

For the purposes of this proof, let the multistage procedure be restricted by the requirement that only one component null hypothesis may be rejected at any one stage (this would be impractical for using a multistage procedure, but the final results would be no different from the results if this restriction were omitted). Furthermore, let  $m_i$  be the number of previously non-rejected com-

ponent null hypotheses at the beginning of the  $i^{\text{th}}$  stage. Then  $m_i = M$  at the first stage,  $m_i = M-1$  at the second stage, and  $m_i = M-i+1$  at the  $i^{\text{th}}$  stage in general (assuming the  $i^{\text{th}}$  stage is reached before the termination of the procedure).

First it will be shown that only the first  $M-n+1$  stages need to be considered, since the remaining stages cannot affect the familywise Type I error rate. For if the  $(M-n+2)^{\text{th}}$  stage is reached, then the number of component null hypotheses which have been rejected is

$$\begin{aligned}
 M - m_{(M-n+2)} &= M - [M - (M-n+2) + 1] \\
 &= M - (n - 2 + 1) \\
 &= M - n + 1 \quad .
 \end{aligned}
 \tag{A.2}$$

However, there are only  $M-n$  component null hypotheses which can be correctly rejected, so  $M-n+1$  rejected component null hypotheses must include at least one Type I error. Since a Type I error must have occurred if the procedure reaches the  $(M-n+2)^{\text{th}}$  stage, at no stage following the  $(M-n+1)^{\text{th}}$  stage can the first Type I error in the procedure be made. And the first Type I error is the critical one since familywise Type I error is the probability of the occurrence of one or more Type I errors. So only the first  $M-n+1$  stages need to be considered.

For a Type I error to occur in the first  $M-n+1$

stages, a sample statistical value corresponding to one of the  $n$  true component null hypotheses must exceed the critical value at the  $(M-n+1)^{\text{th}}$  stage. For if a Type I error occurs in the first  $M-n$  stages (i.e., the sample statistical value corresponding to a true component null hypothesis exceeds the critical value for one of the first  $M-n$  stages), the corresponding sample statistical value will also exceed the critical value at the  $(M-n+1)^{\text{th}}$  stage, since each successive stage gives a less stringent critical value. By the multistage Bonferroni procedure the critical value for each component test at the  $(M-n+1)^{\text{th}}$  stage is based on

$$\begin{aligned} \alpha_T &= \frac{\alpha_{FW(N)}}{m(M-n+1)} \\ &= \frac{\alpha_{FW(N)}}{M-(M-n+1) + 1} \\ &= \frac{\alpha_{FW(N)}}{n} \end{aligned} \quad (A.3)$$

But by the Bonferroni inequality, the probability of a sample statistical value corresponding to one of the  $n$  true null hypotheses being greater than the critical value for the component test ( $\alpha_{FW(A)}$ ) is

$$\begin{aligned} \alpha_{FW(A)} &\leq n \alpha_T = n \frac{\alpha_{FW(N)}}{n} \\ \alpha_{FW(A)} &\leq \alpha_{FW(N)} \end{aligned} \quad (A.4)$$

So, by the multistage Bonferroni procedure, the actual familywise Type I error rate does not exceed the nominal familywise Type I error rate regardless of how many of the component null hypotheses are actually true.

## APPENDIX B

## SAMPLE COMPUTER PROGRAM OUTPUT

This appendix presents a sample of the output of the computer program used in this study.



DATA FOR 'AN EMPIRICAL INVESTIGATION OF SOME SIGNIFICANCE TEST PROCEDURES'

ROBERT E. LARZELLE

DEPT. OF PSYCHOLOGY

GEORGIA TECH

EMPIRICAL SIZE AND EMPIRICAL POWER

NO. OF VARIABLES = 4

SAMPLE SIZE = 40

FAMILYWISE ALPHA = .05

NO. OF TRIALS = 1000

POPULATION COVARIANCE MATRIX:

	1	2	3	4
1	1.0000			
2	.6000	1.0000		
3	.0000	.0000	1.0000	
4	.0000	.0000	.6000	1.0000



EMPIRICAL SIZE FOR SAMPLE CORRELATIONS IN FIRST COLUMN

```

*****
*          METHOD *          I          *          II          *          III          *
*          ALPHA(T) *          .05000 *          .01005 *          .01667 *
*          *          *          *          *          *          *          *          *          *
*          * ACTUAL *          EXP. *          ACTUAL *          EXP. *          ACTUAL *          EXP. *
*****
*          *          *          *          *          *          *          *          *          *
*          (*          918 *          903 *          975 *          966 *          970 *          967 *
*          FREQ. OF          *          *          *          *          *          *          *          *
*          1*          73 *          95 *          55 *          34 *          25 *          33 *
*          TYPE I ERRORS *          *          *          *          *          *          *          *
*          *          *          *          *          *          *          *          *          *
*          *          *          *          *          *          *          *          *          *
*          *          *          *          *          *          *          *          *          *
*          *          *          *          *          *          *          *          *          *
*          *          *          *          *          *          *          *          *          *
*          *          *          *          *          *          *          *          *          *
*          *          *          *          *          *          *          *          *          *
*          *          *          *          *          *          *          *          *          *
*          *          *          *          *          *          *          *          *          *
*          *          *          *          *          *          *          *          *          *
*          *          *          *          *          *          *          *          *          *
*          *          *          *          *          *          *          *          *          *
*          *          *          *          *          *          *          *          *          *
*****

```

IL = 2

TEST OF FIT TO BINOMIAL DISTRIBUTION (FOR METHOD I)

PEARSON CHI-SQUARE( 1) = 2.73031 P < .09846

CONDITIONAL EMPIRICAL SIZE FOR METHOD I (MATRIX)

```

*****
*           * TYPE I ERRORS *
*           * ACJUAL + EXP. *
*****
*           *           *
* RHO(3,1) * 46 * 46. *
*           *           *
* RHO(3,2) * 8 * 2.3 *
*           *           *
* RHO(4,1) * 4 * 2.3 *
*           *           *
* RHO(4,2) * 6 * 2.3 *
*           *           *
*****

```

EMPIRICAL POWER FOR SINGLE TESTS (MATRIX)

```

*****+*****+*****+*****+*****+*****+*****+*****+*****+*****
*          METHOD *          I          +          II          +          III          *
*          ALPHA(T) *          .35000 *          .3451 *          .3383 *
*          *          *          *          *          *          *          *          *          *
*          *H(0): RHO(I,J)=0 *H(0): RHO(I,J)=0 *H(0): RHO(I,J)=0 *
* * * * * + * + * * * * * + * * * * * + * * * * * + * * * * * + * * * * *
*          +VALUE *ACCEPTED* REJECTED*ACCEPTED* REJECTED*ACCEPTED* REJECTED*
*****+*****+*****+*****+*****+*****+*****+*****+*****+*****
*          *          *          *          *          *          *          *          *
* RHO(2,1)* .6 * 10 * 990 * 46 * 954 * 47 * 953 *
*          *          *          *          *          *          *          *          *
* RHO(4,3)* .6 * 12 * 988 * 56 * 944 * 57 * 943 *
*          *          *          *          *          *          *          *          *
*****+*****+*****+*****+*****+*****+*****+*****+*****+*****

```



EMPIRICAL POWER FAMILYWISE (MATRIX)

```

*****
*                               *
*           METHOD * I * II * III *
*                               *
* NO. OF TYPE II ERRORS *
*                               *
*****
*                               *
*           0           * 979 * 992 * 980 *
*                               *
*           1           * 20 * 94 * 95 *
*                               *
*           2           * 1 * 4 * 4 *
*                               *
*****

```

EMPIRICAL POWER FAMILYWISE (FIRST COLUMN)

```

*****
*                               * I *  I *  I *
*                               * * * * * * * *
* NO. OF TYPE II ERRORS * * * * *
*****
*                               * * * * *
*           0 * 990 * 971 * 971 *
*                               * * * * *
*           1 * 10 * 29 * 29 *
*                               * * * * *
*****

```



EMPIRICAL ACCURACY (TOTAL OF TYPE I AND TYPE II ERRORS) (MATRIX)

```

*****
* METHOD * I * II * III *
* * * * * * * * * *
* ERRORS * * * * *
*****
* * * * *
* 0 * 843 * 878 * 976 *
* * * * *
* 1 * 123 * 115 * 117 *
* * * * *
* 2 * 27 * 6 * 6 *
* * * * *
* 3 * 7 * 1 * 1 *
* * * * *
* 4 * 0 * 0 * 0 *
* * * * *
* 5 * 0 * 0 * 0 *
* * * * *
* 6 * 0 * 0 * 0 *
* * * * *
*****

```

EMPIRICAL ACCURACY (FIRST COLUMN)

```

*****
* METHOD * I * II * III *
* * * * * * * * *
* ERRORS * * * *
*****
* * * * *
* 0 * 909 * 947 * 947 *
* * * * *
* 1 * 86 * 52 * 52 *
* * * * *
* 2 * 5 * 1 * 1 *
* * * * *
* 3 * 0 * 0 * 0 *
* * * * *
*****

```

## APPENDIX C

TABLE CONCERNING PSEUDORANDOM NUMBER GENERATOR

Table C.1. Total Number of Type I Errors

k	N	Total Normal Numbers Generated	Total Type I Errors	
			Expected	Actual
3	15	45,000	150	147 <sup>a</sup>
3	15	45,000	100	82
3	15	45,000	200 <sup>b</sup>	177
3	15	45,000	100	69
3	40	120,000	150	132 <sup>a</sup>
3	40	120,000	100	96
3	40	120,000	100	85
3	40	120,000	100	105 <sup>d</sup>
3	100	300,000	150	168 <sup>ad</sup>
3	100	300,000	100	98
3	100	300,000	100	104 <sup>d</sup>
3	100	300,000	100	93
4	15	60,000	300	237 <sup>a</sup>
4	40	160,000	300	284 <sup>a</sup>
4	40	160,000	150	133
4	40	160,000	200	176

Table C.1 - Continued

k	N	Total Normal Numbers Generated	Total Type I Errors	
			Expected	Actual
4	100	400,000	300	297 <sup>ac</sup>
4	100	400,000	150	148
4	100	400,000	200	217 <sup>d</sup>
6	40	240,000	600	585
6	40	240,000	650	654 <sup>d</sup>
6	40	240,000	450	436
6	100	600,000	600	637 <sup>d</sup>
6	100	600,000	650	645
6	100	600,000	450	394
8	15	120,000	1,400	1,109 <sup>a</sup>
8	40	320,000	1,400	1,327 <sup>a</sup>
8	100	1,600,000	1,400	1,411 <sup>ad</sup>

<sup>a</sup>This number may be slightly underestimated.

<sup>b</sup> $\alpha_T = .10$  for this computer run

<sup>c</sup>This number could possibly have exceeded the expected value.

<sup>d</sup>This actual value was larger than the expected value.

## APPENDIX D

## TABLES CONCERNING TYPE I ERROR RATES

Table D.1. Familywise Empirical Size

$R_p$	k	N	Matrix			Vector				
			$m^a$	Method I	Method II	Method III	$m^a$	Method I	Method II	Method III
I		15	3	.144**	.047	.046	2	.093**	.049	.048
	3	40	3	.122**	.036	.036#	2	.090**	.049	.049
		100	3	.162**	.044	.044	2	.119**	.047	.046
		15	6	.209**	.044	.042	3	.116**	.044	.041
	4	40	6	.246**	.052	.048	3	.126**	.040	.040
		100	6	.262**	.058	.057	3	.137**	.040	.040
		15	21	.669**	.027	.026##	7	.224**	.034	.034#
	8	40	21	.752**	.044	.043	7	.300**	.046	.045
		100	21	.755**	.050	.048	7	.305**	.051	.050
(51) <sup>b</sup>		15	2	.079**	.024##	.023##	2	.079**	.033#	.033#
	3	40	2	.090**	.028##	.027#	2	.090**	.048	.048
		100	2	.091**	.034#	.033#	2	.091**	.049	.048
(52) <sup>b</sup>		15 <sup>c</sup>	2	.150 <sup>c</sup> **	.046 <sup>c</sup> ##	.044 <sup>c</sup> ##	2	.150 <sup>c</sup> **	.079 <sup>c</sup> ##	.078 <sup>c</sup> ##
	3	40	2	.077**	.023##	.022##	2	.077**	.028##	.028##
		100	2	.092**	.033#	.032##	2	.092**	.048	.048
(53) <sup>b</sup>		15	2	.062*	.018##	.018##	2	.062*	.028##	.028##
	3	40	2	.092**	.030##	.030##	2	.092**	.046	.045
		100	2	.086**	.028##	.027##	2	.086**	.048	.047

Table D.1 - Continued

$R_{*p}$	k	N	Matrix			Vector				
			$m^a$	Method I	Method II	Method III	$m^a$	Method I	Method II	Method III
(54) <sup>b</sup>	4	40	3	.110**	.028##	.028##	3	.110**	.042	.041
		100	3	.118**	.016##	.016##	3	.118**	.034#	.033#
(55) <sup>b</sup>	4	40	4	.139**	.025##	.025##	2	.082**	.025##	.025##
		100	4	.160**	.032##	.031##	2	.097**	.032##	.032##
(56) <sup>b</sup>	6	40	12	.393**	.042	.042	4	.167**	.035#	.034#
		100	12	.398**	.039	.037	4	.181**	.043	.042
(57) <sup>b</sup>	6	40	13	.452**	.042	.042	3	.164**	.028##	.028##
		100	13	.456**	.037	.035#	3	.134**	.035#	.034#
(58) <sup>b</sup>	6	40	9	.248**	.019##	.019##	3	.097**	.022##	.022##
		100	9	.239**	.024##	.024##	3	.094**	.022##	.022##

Note. - Familywise empirical size is the proportion of replications out of 1000 sample replications in which there occurred one or more Type I errors in the family of tests. The family of tests includes either the tests of intercorrelations in the entire matrix or the tests of intercorrelations in the first column (vector) only. Method I set  $\alpha_T = .05$ ; Method II was the assumed-independent-tests correction; Method III was the Bonferroni  $t$ .

<sup>a</sup>Number of true component null hypotheses.

<sup>b</sup>Refers to the population correlation matrix designated by this equation number in Chapter II.

<sup>c</sup> $\alpha_{FW} = .10$  for this computer run

\* $p < .05$ ; one-tailed test

\*\* $p < .01$ ; one-tailed test

# $p < .05$ ; two-tailed test

## $p < .01$ ; two-tailed test

Table D.2. Summary of Tests of Fit of Frequencies  
of Number of Type I Errors to a Binomial Distribution

$R_{\rightarrow p}$	k	N	Matrix			Vector		
			$m^{1a}$	df's	$\chi^2$ Value	$m^{1a}$	df's	$\chi^2$ Value
I K	3	15	3	2	2.73	2	1	.23
		40	3	2	5.58	2	1	.64
		100	3	2	3.80	2	1	5.25*
	4	15	6	2	16.05**	3	2	6.44*
		40	6	2	2.91	3	2	2.31
		100	6	2	.06	3	2	.86
	8	15	21	5	72.60**	7	2	28.86**
		40	21	5	7.00	7	2	.47
		100	21	5	3.48	7	2	.75
(51) <sup>b</sup>	3	15	2	1	3.89*	2	1	3.89*
		40	2	1	.64	2	1	.64
		100	2	1	.48	2	1	.48
(52) <sup>b</sup>	3	15	2	2 <sup>c</sup>	48.93**	2	2 <sup>c</sup>	48.93**
		40	2	1	4.78*	2	1	4.78*
		100	2	1	.34	2	1	.34
(53) <sup>b</sup>	3	15	2	1	14.32**	2	1	14.32**
		40	2	1	.34	2	1	.34
		100	2	1	1.50	2	1	1.50
(54) <sup>b</sup>	4	40	3	2	34.83**	3	2	34.83**
		100	3	2	63.10**	3	2	63.10**
(55) <sup>b</sup>	4	40	4	2	46.72**	2	1	2.73
		100	4	2	75.25**	2	1	.003

Table D.2 - Continued

$R_{mp}$	k	N	Matrix			Vector		
			$m^a$	df's	$\chi^2$ Value	$m^a$	df's	$\chi^2$ Value
(56) <sup>b</sup>	6	40	12	3	57.82**	4	2	40.91**
		100	12	3	107.11**	4	2	45.78**
(57) <sup>b</sup>	6	40	13	3	22.90**	3	2	4.14
		100	13	3	23.51**	3	2	6.17*
(58) <sup>b</sup>	6	40	9	3	249.89**	3	2	123.50**
		100	9	3	174.39**	3	2	53.01**

Note. - The characteristics of each case investigated are reported in the left-hand columns. For each particular case the observed frequencies for each number of Type I errors were compared to frequencies expected by the binomial distribution under the assumption that the component significance tests are mutually independent. The expected binomial distribution had  $p = .05$  and  $n = m'$ . The statistic for the test of fit was a Pearson  $\chi^2$ . The upper tail of the distribution was lumped together so that the expected frequency (out of 1000 sample replications) was at least five.

<sup>a</sup>The number of true component null hypotheses in a family of tests

<sup>b</sup>Refers to the population correlation matrix designated by this equation number in Chapter II.

<sup>c</sup> $\alpha = .10$  for this computer run

\* $p < .05$

\*\* $p < .01$



Table D.3. Conditional Empirical Size for Method I<sup>a</sup>

	Type I Errors		Conditional Empirical Size <sup>b</sup>
	Expected	Observed	
$P_{21}$	59.0	59	
$P_{31}$	2.9	4	.068
$P_{32}$	2.9	5	.085
$P_{41}$	2.9	2	.034
$P_{42}$	2.9	6	.102
$P_{43}$	2.9	1	.017

<sup>a</sup> $R_p = I, k = 4, N = 100$

<sup>b</sup>Conditional empirical size is the proportion of sample replications with a Type I error on the component test of  $P_{ij}$  among those sample replications with a Type I error on  $P_{21}$ .

Table D.4. Conditional Empirical Size for Method I<sup>a</sup>

	Type I Errors		Conditional Empirical Size <sup>b</sup>
	Expected	Observed	
$P_{21}$	44.0	44	
$P_{31}$	2.2	8**	.182
$P_{41}$	2.2	12**	.273

Note. - The two tests of this table are considered a family of tests. A one-tailed Bonferroni Poisson test was used.

$$R_{mp} = \begin{bmatrix} 1 & & & \\ 0 & \diagdown & & \\ 0 & .6 & & \\ 0 & .6 & .6 & 1 \end{bmatrix}, N = 40$$

<sup>b</sup>Conditional empirical size is the proportion of sample replications with a Type I error on the component test of  $P_{ij}$  among those sample replications with a Type  $i$  error on  $P_{21}$ .

$$**P_{FW} < .01; P_T < .005$$

Table D.5. Conditional Empirical Size for Method I<sup>a</sup>

	Type I Errors		Conditional Empirical Size <sup>b</sup>
	Expected	Observed	
$P_{21}$	43.0	43	
$P_{31}$	2.1	12**	.279
$P_{41}$	2.1	10**	.233

Note. - The two tests of this table are considered as a family of tests. A one-tailed Bonferroni Poisson test was used.

$$R_p^a = \begin{bmatrix} 1 & & & \\ 0 & \diagdown & & \\ 0 & .6 & & \\ 0 & .6 & .6 & 1 \end{bmatrix}, N = 100$$

<sup>b</sup>Conditional empirical size is the proportion of sample replications with a Type I error on the component test of  $P_{ij}$  among those sample replications with a Type I error on  $P_{21}$ .

$$**P_{FW} < .01; P_T < .005$$

Table D.6. Conditional Empirical Size for Method I<sup>a</sup>

	Type I Errors		Conditional Empirical Size <sup>b</sup>
	Expected	Observed	
$P_{31}$	46.0	46	
$P_{32}$	2.3	8**	.174
$P_{41}$	2.3	4	.087
$P_{42}$	2.3	6 <sup>c</sup>	.130

Note. - The three tests in this table are considered as a family of tests. A one-tailed Bonferroni Poisson test was used.

$${}^a R_{mp} = \begin{bmatrix} 1 & & & \\ .6 & \diagdown & & \\ 0 & 0 & & \\ 0 & 0 & .6 & 1 \end{bmatrix}, N = 40$$

<sup>b</sup> Conditional empirical size is the proportion of sample replications with a Type I error on the component test of  $P_{ij}$  among those sample replications with a Type I error on  $P_{21}$ .

$$c P_T < .05$$

$$** P_{FW} < .01; P_T < .0033$$

Table D.7. Conditional Empirical Size for Method I<sup>a</sup>

	Type I Errors		Conditional Empirical Size <sup>b</sup>
	Expected	Observed	
$P_{31}$	54.0	54	
$P_{32}$	2.7	16**	.296
$P_{41}$	2.7	15**	.278
$P_{42}$	2.7	5	.093

Note. - The three tests on this table are considered a family of tests. A one-tailed Bonferroni Poisson test was used.

$$R_{mp} = \begin{bmatrix} 1 & & & \\ .6 & \diagdown & & \\ 0 & 0 & & \\ 0 & 0 & .6 & 1 \end{bmatrix}, N = 100$$

<sup>b</sup> Conditional empirical size is the proportion of sample replications with a Type I error on the component test of  $P_{ij}$  among those sample replications with a Type I error on  $P_{31}$ .

$$**P_{FW} < .01; P_T < .0033$$



Table D.9. Conditional Empirical Size for Method I<sup>a</sup>

	Type I Errors		Conditional Empirical Size <sup>b</sup>
	Expected	Observed	
$P_{31}$	49.0	49	
$P_{32}$	2.4	7 <sup>c</sup>	.143
$P_{41}$	2.4	14**	.286
$P_{42}$	2.4	6 <sup>c</sup>	.122
$P_{51}$	2.4	4	.082
$P_{52}$	2.4	6 <sup>c</sup>	.122
$P_{53}$	2.4	1	.020
$P_{54}$	2.4	1	.020
$P_{61}$	2.4	2	.041
$P_{62}$	2.4	6 <sup>c</sup>	.122
$P_{63}$	2.4	4	.082
$P_{64}$	2.4	1	.020

Note. - The eleven tests of this table are considered a family of tests. A one-tailed Bonferroni Poisson test was used.

$$R_{mp} = \begin{bmatrix} 1 & & & & & \\ .6 & & & & & \\ 0 & 0 & & & & \\ 0 & 0 & .6 & & & \\ 0 & 0 & 0 & 0 & & \\ 0 & 0 & 0 & 0 & .6 & 1 \end{bmatrix}, N = 100$$

<sup>b</sup> Conditional empirical size is the proportion of sample replications with a Type I error on the component test of  $P_{ij}$  among those sample replications with a Type I error on  $P_{31}$ .

$$c \underline{P}_T < .05$$

$$** \underline{P}_{FW} < .01; \underline{P}_T < .0009$$





Table D.11. Conditional Empirical Size for Method I<sup>a</sup>

	Type I Errors		Conditional Empirical Size <sup>b</sup>
	Expected	Observed	
$P_{21}$	48.0	48	
$P_{31}$	2.4	3	.062
$P_{32}$	2.4	7 <sup>c</sup>	.146
$P_{41}$	2.4	3	.062
$P_{42}$	2.4	4	.083
$P_{43}$	2.4	0	.000
$P_{52}$	2.4	9**	.188
$P_{53}$	2.4	1	.021
$P_{54}$	2.4	0	.000
$P_{62}$	2.4	14**	.292
$P_{63}$	2.4	1	.021
$P_{64}$	2.4	4	.083
$P_{65}$	2.4	3	.062

Note. - The twelve tests of this table are considered as a family of tests. A one-tailed Bonferroni Poisson test was used.

$$R_p = \begin{bmatrix} 1 & & & & & & \\ 0 & & & & & & \\ 0 & 0 & & & & & \\ 0 & 0 & 0 & & & & \\ .6 & 0 & 0 & 0 & & & \\ .6 & 0 & 0 & 0 & 0 & 1 & \end{bmatrix}, N = 100$$

<sup>b</sup>Conditional empirical size is the proportion of sample replications with a Type I error on the component test of  $P_{ij}$  among those sample replications with a Type I error on  $P_{21}$ .

$$c \underline{P}_T < .05$$

$$** \underline{P}_{FW} < .01; \underline{P}_T < .0008$$





## BIBLIOGRAPHY

- Ahrens, J. H., Dieter, U. & Grube, A. Pseudo-random numbers: A new proposal for the choice of multipliers. Computing, 1970, 6, 121-138.
- Aitkin, M. Statistical theory (behavioral science application). Annual Review of Psychology, 1971, 22, 225-250.
- Anderson, T. W. An introduction to multivariate statistical analysis. New York: Wiley & Sons, 1958.
- Barr, D. R., & Slezak, N. L. A comparison of multivariate normal generators. Communications of the Association for Computing Machinery, 1972, 15, 1048-1049.
- Boardman, T. J., & Moffitt, D. R. Graphical Monte Carlo Type I error rates for multiple comparison procedures. Biometrics, 1971, 27, 738-744.
- Box, G. E. P., & Muller, M. E. A note on the generation of random normal deviates. Annals of Mathematical Statistics, 1958, 29, 610-611.
- Brooks, J. B. Familial antecedents and adult correlates of artistic interests in childhood. Journal of Personality, 1973, 41, 110-120.
- Brown, R. J., Jr., & Rowland, J. R. Auto correlation significance in digital pseudo-random number generation. Unpublished manuscript, School of Electrical Engineering, Georgia Institute of Technology, March 2, 1970.
- Capra, J. R., & Elster, R. S. A note on generating multivariate data with desired means, variances, and covariances. Educational and Psychological Measurement, 1971, 31, 749-752.
- Carmer, S. G., & Swanson, M. R. An evaluation of ten pairwise multiple comparison procedures by Monte Carlo methods. Journal of the American Statistical Association, 1973, 68, 66-74.
- Christensen, L. R. Simultaneous statistical inference in the normal multiple linear regression model. Journal of the American Statistical Association, 1973, 68, 457-461.

- Cohn, C. E. The performance of random-bit generators. Simulation, 1971, 17, 234-236.
- Cole, N. S. Statistical tests of the equality of correlation matrices. (Doctoral dissertation, University of North Carolina at Chapel Hill) Ann Arbor, Mich.: University Microfilms, 1969, No. 69-10,143.
- Coveyou, R. R., & MacPherson, R. D. Fourier analysis of uniform random number generators. Journal of the Association for Computing Machinery, 1967, 14, 100-119.
- Cramér, H. Mathematical methods of statistics. Princeton: Princeton University Press, 1946.
- Dieter, U. Pseudo-random numbers: The exact distribution of pairs. Mathematics of Computation, 1971, 25, 855-883.
- Dieter, U., & Ahrens, J. An exact determination of serial correlations of pseudo-random numbers. Numerische Mathematik, 1971, 17, 101-123.
- Duncan, D. B. Multiple range and multiple F tests. Biometrics, 1955, 11, 1-42.
- Dunn, O. J., & Massey, F. J., Jr. Estimation of multiple contrasts using t-distributions. Journal of the American Statistical Association, 1965, 60, 573-583.
- Dunnett, C. W. Multiple comparison tests. Biometrics, 1970, 26, 139-141.
- Fisher, R. A. The design of experiments. Edinburgh: Oliver & Boyd, 1935.
- Fisher, R. A. The simultaneous distribution of correlation coefficients. Sankhyā: The Indian Journal of Statistics: Series A, 1962, 24, 1-8.
- Gabriel, K. R. Simultaneous test procedures - Some theory of multiple comparisons. Annals of Mathematical Statistics, 1969, 40, 224-250.
- Games, P. A. Inverse relation between the risks of Type I and Type II errors and suggestions for the unequal  $n$  case in multiple comparisons. Psychological Bulletin, 1971, 25, 97-102.
- Guenther, W. C. Concepts of statistical inference. New York: McGraw-Hill, 1965.

- Gupta, S. S., & Panchapakesan, S. On multiple decision procedures. Journal of Mathematical and Physical Sciences, 1972, 6, 1-72.
- Halton, J. H. A retrospective and prospective survey of the Monte Carlo method, SIAM Review, 1970, 12, 1-63.
- Harmon, H. H. Modern factor analysis. Chicago: University of Chicago Press, 1960.
- Harris, R. N. Multiple significance tests of correlation coefficients in correlation matrices. (Doctoral dissertation, University of Maryland) Ann Arbor, Mich.: University Microfilms, 1967, No. 67-11,302.
- Hays, W. L. Statistics for the social sciences. (2nd ed.) New York: Holt, Rinehart, and Winston, 1973.
- Jansson, B. Random number generators. Stockholm: Victor Pettersons Bokindustri Aktiebolag, 1966.
- Jessor, S. L., & Jessor, R. Maternal ideology and adolescent problem behavior. Developmental Psychology, 1974, 10, 246-254.
- Kendall, M. G., & Stuart, A. The advanced theory of statistics. Vol. 2. Inference and relationships. (2nd ed.) New York: Hafner Publ. Co., 1967.
- Keselman, H. J. The statistic with the smaller critical value. Psychological Bulletin, 1974, 81, 130-131.
- Keselman, H. L., & Toothaker, L. E. Error rates for multiple comparison methods: Some evidence concerning the misleading conclusions of Petrinovich and Hardyck. Psychological Bulletin, 1973, 80, 31-32.
- Kirk, R. E. Experimental design: Procedures for the behavioral sciences. Belmont, Cal.: Brooks/Cole, 1968.
- Knuth, D. E. The art of computer programming. Vol. 2. Seminumerical algorithms. Reading, Mass.: Addison-Wesley, 1969.
- Kolb, D. A. Achievement motivation training for under-achieving high-school boys. In J. H. Hamsher & H. Sigall (Eds.), Psychology and social issues. New York: MacMillan, 1973. (Originally published: Journal of Personality and Social Psychology, 1965, 2, 783-792.)

- Kuehl, F. W. Evaluation of a multiplicative generator of pseudo-random numbers. Technical Research Note 215, AD 707 374, U.S. Army Behavioral Science Research Laboratory, September, 1969.
- Lehmer, D. H. Mathematical method in large-scale computing units, Proceedings of the Second Symposium on Large-Scale Digital Calculating Machinery, Cambridge, Mass.: Harvard University Press, 1949. Cited by Dieter (1971). P. 883.
- MacLaren, M. D., & Marsaglia, G. Uniform random number generators. Journal of the Association for Computing Machinery, 1965, 12, 83-89.
- Maddocks, R. S., Matthews, S., Walker, E. W., & Vincent, C. H. A compact and accurate generator for truly random binary digits. Journal of Physics E: Scientific Instruments, 1972, 5, 542-544.
- Marascuilo, L. A. Large-sample multiple comparisons. Psychological Bulletin, 1966, 65, 280-290.
- Marsaglia, G. Random numbers fall mainly in the planes. Proceedings of the National Academy of Sciences, 1968, 61, 25-28.
- Marsaglia, G. Regularities in congruential random number generators. Numerische Mathematik, 1970, 16, 8-10.
- Marsaglia, G., & Bray, T. A. A convenient method for generating normal variables. SIAM Review, 1964, 6, 260-264.
- Marsaglia, G., MacLaren, M. D., & Bray, T. A. A fast procedure for generating normal random variables. Communications of the Association for Computing Machinery, 1964, 7, 4-10.
- MATH-PACK programmers reference. UP7542 Rev. 1, Sperry Rand Corporation, 1970.
- Miller, R. G., Jr. Simultaneous statistical inference. New York: McGraw-Hill, 1966.
- Mulaik, S. A. The foundations of factor analysis. New York: McGraw-Hill, 1972.
- Muller, M. E. A comparison of methods for generating normal deviates on digital computers. Journal of the Association for Computing Machinery, 1959, 6, 376-383.

- Murry, H. F. A general approach for generating natural random variables. IEEE Transactions on Computers, 1970, 19, 1210-1213.
- Neave, H. R. On using the Box-Müller transformation with multiplicative congruential pseudo-random number generators. Journal of the Royal Statistical Society (Series C), Applied Statistics, 1973, 22, 92-97.
- Nelder, J. A. Discussion on the papers by Dr. Wynn and Dr. Bloomfield and by Mr. O'Neill and Professor Wetherill. Journal of the Royal Statistical Society: Series B, 1971, 33, 244-246.
- Odell, P. L., & Feiveson, A. L. Corrigenda. Journal of the American Statistical Association, 1966, 61, 1248-1249. (a)
- Odell, P. L., & Feiveson, A. H. A numerical procedure to generate a sample covariance matrix. Journal of the American Statistical Association, 1966, 61, 199-203. (b)
- O'Neill, R. & Wetherill, G. B. The present state of multiple comparison methods. Journal of the Royal Statistical Society: Series B. 1971, 33, 218-241.
- Oplinger, J. L. Generation of correlated pseudo-random numbers for Monte Carlo simulation. Unpublished master's thesis. University of Pennsylvania, 1971.
- Paige, J. M. Changing patterns of anti-white attitudes among blacks. In J. H. Hamsher & H. Sigall (eds.), Psychology and social issues. New York: MacMillan, 1973. (Originally published: Journal of Social Issues, 1970, 26, 69-86.)
- Pedersen, D. M. Correlates of behavioral personal space. Psychological Reports, 1973, 32, 828-830. (a)
- Pedersen, D. M. Personality and demographic correlates of simulated personal space. Journal of Psychology, 1973, 85, 101-108. (b)
- Perlmutter, J., & Myers, J. L. A comparison of two procedures for testing multiple contrasts. Psychological Bulletin, 1973, 79, 181-184.
- Petrinovich, L. F., & Hardyck, C. D. Error rates for multiple comparison methods: Some evidence concerning the frequency of erroneous conclusions, Psychological Bulletin, 1969, 71, 43-54.



- Plackett, R. L. Discussion on the papers by Dr. Wynn and Dr. Bloomfield and by Mr. O'Neill and Professor Wetherill. Journal of the Royal Statistical Society, Series B, 1971, 33, 242-244.
- Ryan, T. A. Multiple comparisons in psychological research. Psychological Bulletin, 1959, 56, 26-47.
- Ryan, T. A. Significance tests for multiple comparison of proportions, variances, and other statistics. Psychological Bulletin, 1960, 57, 318-328.
- Ryan, T. A. The experiment as the unit for computing rates of error. Psychological Bulletin, 1962, 59, 301-305.
- Scheffé, H. The analysis of variance. New York: Wiley, 1959.
- Scheuer, E. M., & Stoller, D. S. On the generation of normal random vectors. Technometrics, 1962, 4, 278-281.
- Seawright, M. A., Larkin, W. D., & Sachs, S. A. Studies on methods of random number generation for computer simulation. Research Memorandum 66-8, U.S. Army Personnel Research Office, November 1966.
- Shreider, Y. A. (Ed.) The Monte Carlo method: The method of statistical trials. Oxford: Pergamon Press, 1966.
- Siess, T. F. Personality correlates of volunteers' experimental preferences. Canadian Journal of Behavioral Science, 1973, 5, 253-263.
- Smith, R. A. The effect of unequal group size on Tukey's HSD procedure. Psychometrika, 1971, 36, 31-34.
- Spjøtvoll, E. Multiple comparison of regression functions. Annals of Mathematical Statistics, 1972, 42, 1076-1088.
- Tukey, J. W. The problem of multiple comparisons. Ditto, Princeton University, 1953. Cited by Ryan (1959). P. 26.
- Waller, R. A., & Duncan, D. B. A Bayes rule for the symmetric multiple comparisons problem. Journal of the American Statistical Association, 1969, 64, 1484-1503.
- Williams, J. D. Multiple comparisons in a regression approach. Psychological Reports, 1972, 30, 639-647.

- Williams, J. D. Note on familywise error rates. Psychological Reports, 1973, 32, 1221-1222.
- Wilson, W. A note on the inconsistency inherent in the necessity to perform multiple comparisons. Psychological Bulletin, 1962, 59, 296-300.
- Wold, H. Random normal deviates. Tracts for Computers, Cambridge, G. B.: University Press, 1948, No. 25.