Technical Report TR-66-28
NsG-398 May 1966

# AN EMPIRICAL EVALUATION OF THE EFFECTS OF VIOLATING THE ASSUMPTION OF HOMOGENEITY OF COVARIANCE FOR THE REPEATED MEASURES DESIGN OF THE ANALYSIS OF VARIANCE

by
Peter H. Stoloff

This research was performed at the Computer Science Center of the University of Maryland in partial fulfillment of the requirements for the degree of Master of Arts. Support from the National Aeronautics and Space Administration under Grant NsG-398 is greatfully acknowledged.

# ABSTRACT


Samples were generated from populations with known mean, standard deviation and covariance matrix for the repeated measures design for specified values of the $\nu_1$ and $\nu_2$ parameters.

The results of this study indicate that as d.f. and heterogeneity of covariance increased in the population, the approximations to the central F distribution and the $T^2$ test become increasingly more effective in accounting for bias in the F statistic resulting from violations of the model.

When the power of the various tests was investigated over the various values of noncentrality, it was found that the overall level of power for the exact F remained relatively invariant when the model was violated. When the approximate procedures using the F distribution were used when violations of the model for the exact F was present in the population, and hence "logically" appropriate, the power of these tests approached that of the exact F.

Descriptive measures of the distribution of epsilon revealed the sample estimates of the parameter to be best when sampling from the extreme ranges of the domain of the statistic. As the d.f. in the covariance matrix increased, differential prediction of the parameter, given the sample mean, improved.

ACKNOWLEDGEMENTS

This project was made possible by support for computer time from the University of Maryland and the National Aeronautics and Space Administration under grant NsG 398; the facilities of the University's Computer Science Center were used for all computations. The author wishes to express deep appreciation to his thesis director, Dr. N. S. Anderson, who so kindly assumed the responsibility of the direction of this project from Dr. E. F. Heermann, who was the initiator and previous director of this thesis; and to Drs. C. J. Bartlett and L. O. Walder, Associate Professors of Psychology, for serving on this thesis committee.

TABLE OF CONTENTS


| Chapter | | Page |
|---|---|---|
| ACKNOWLEDGEMENTS | | ii |
| I. | INTRODUCTION ........................ | 1 |
| II. | METHOD AND PROCEDURE ............... | 11 |
| | 1. Simulation of Data ............ | 11 |
| | 2. The Simplex ................... | 12 |
| | 3. Evaluation of the Magnitude of the Type I Error ........... | 14 |
| | 4. Power of the Tests ............ | 15 |
| | 5. Sampling Characteristics of Epsilon ...................... | 16 |
| | 6. Selection of Parameters ....... | 16 |
| | a. Average universe correlation | 16 |
| | b. R matrices and epsilon .... | 17 |
| | c. k, the number of treatment classifications ........... | 18 |
| | d. n, the number of subjects | 18 |
| | e. Values of the treatment means ..................... | 22 |
| | f. Number of sample covariance matrices .................. | 23 |
| III. | RESULTS AND DISCUSSION .............. | 25 |
| | 1. The Type I Error .............. | 25 |

2. The Power of the Tests ........ 35

3. Sampling Characteristics of Epsilon ...................... 38

4. Summary ...................... 43

APPENDIX I: Probabilities of the Type I Error ......... 45

APPENDIX II: Selected Power Curves .................... 50

FOOTNOTES ........................................... 63

REFERENCES .......................................... 64

## LIST OF TABLES


| Table | | Page |
|---|---|---|
| 2.1 | Matrix of Intercorrelations Having Perfect Simplicial Form | 13 |
| 2.2 | Population Covariance Matrices: $k = 3$ | 20 |
| 2.3 | Population Covariance Matrices: $k = 5$ | 21 |
| 2.4 | Values of the Noncentrality Parameter | 24 |
| 3.1 | Probabilities of the Type I Error: Averaged across n and k | 31 |
| 3.2 | Probabilities of the Type I Error: at $\alpha = .01$ Averaged across n | 32 |
| 3.3 | Probabilities of the Type I Error: at $\alpha = .05$ Averaged across n | 32 |
| 3.4 | Probabilities of the Type I Error: at $\alpha = .01$ Averaged across k | 33 |
| 3.5 | Probabilities of the Type I Error: at $\alpha = .05$ Averaged across k | 34 |
| 3.6 | Descriptive Measures of Samples of Box's Epsilon ($k = 5$) | 41 |
| 3.7 | Descriptive Measures of Samples of Box's Epsilon ($k = 3$) | 42 |

# CHAPTER I

## INTRODUCTION

An experimental study can be classified as a repeated measures design whenever the total degree of freedom (d.f.) is greater than the total number of subjects (n) in an analysis of variance (Anova) paradigm. In order to have more d.f. than experimental subjects, more than one observation per subject must have been made; consequently, there is a departure from the usual Anova assumption of independence of measures.

Repeated measures designs are quite popular in the psychological literature. Lana & Lubin (1961, 1963) made a survey of three journals: Journal of Comparative and physiological Psychology, Journal of Experimental Psychology and, Journal of Abnormal and Social Psychology, for the years 1957 through 1959 and found that over 35% of the experiments appearing in these journals used repeated measures designs.

The lay-out of a typical repeated measures design, assuming a single classification, fixed constants univariate model, would consist of an n x k score matrix; i.e., $\underline{n}$ subjects, each measured across $\underline{k}$ classifications, and each of the measures between subjects would be independent, but measures along the k classifications would be dependent since the measures are made on the same experimental subjects.[1]

Under such a model, deviations about the grand mean; e.g., total sum of squares (S.S.), are broken down into between subjects and within subject variation. The within subject variation is again partitioned into between treatment and "residual" variation. The residual variation is the difference between total within subject variability and between treatment variation. This residual variation, when divided by the appropriate d.f. (in the general case, (n-1)(k-1)d.f.), becomes the error term or denominator of the F ratio and the between treatment variation divided by its appropriate d.f. (k-1) becomes the numerator for the F ratio.

Thus, it should be observed that the sampling unit is subjects within groups and that the total d.f., nk-1, is greater than the number of subjects used in the experiment. This is one of the great advantages of the repeated measures design--economy of subjects; and where the assumption of homogeneity of covariance holds up, it has been contended (Winer, 1962) that the individual difference factor has been controlled. (As we shall see later, such a contention results in a paradox; to minimize the error variance, to increase the power of the F test, maximum and constant covariation is needed ... which implies maximization of individual differences.)

A basic assumption of the univariate Anova model is zero or homogeneous covariation, between treatments, in the universe. Since repeated measures involves measuring subjects across time, the usual outcome of such experimentation does not result in zero correlation and oftentime, the correlation matrix does not show homogeneity (Gaito & Wiley, 1963; Lubin, 1962; Lana & Lubin, 1963). Non-homogeneous correlation results when such factors as carry-over and sequence effects, fatigue, warm-up and transfer of training occur during collection of data. Suggestions have been made (Box & Mueller, 1958; Winer, 1962) to administer treatments randomly within subjects--thus, sequence and carry-over effects should cancel out. However, when the independent variables lie on a time continumn; i.e., measurements are made from one time interval to the next, as in learning experiments, such randomization is not possible.

Having established that violations of the repeated measures model do occur, the effects of such violations, especially violation of the assumption of homogeneity of covariance, should be examined.

As far back as 1948, Kogan had suggested that non-homogeneous correlations result in a biased F statistic. He contended that where the correlations are positive, but unequal, a positive bias in the F test results; i.e., the test yields significant results too often. This view is also held by Box

(1954), Gaito & Wiley (1963), Geisser & Greenhouse (1958), Lana & Lubin (1963) and Winer (1962), just to mention a few.

The reason the resultant F, arising out of non-homogeneously correlated data, is biased can be seen from the fact that the central F distribution is based on the null case of no treatment differences, where the numerator and denominator of the F ratio estimate the same quantity. Thus, when the null hypothesis ($H_o$) is true, both numerator and denominator are estimates of experimental error. Box (1954) has shown that when the assumption of homogeneity of covariance is violated, the numerator and denominator of the F ratio do not estimate the same quantity when there is no treatment effect.

Various attempts have been made to contend with this bias. Three approaches have been predominant in attempting to deal with this problem of the inflated F; they are: (1) where possible, more careful design and control of the experiment, (2) use of multivariate analysis and, (3) correction of the biased F by adjusting the d.f. to have F approximate a central F distribution.

As mentioned before, Box & Mueller (1958) have recommended randomization of treatment assignment within subjects. They have shown that if treatments are assigned by the fully randomized design of randomized blocks design (i.e., matched subjects), the expected value (E) for the covariance is zero. Where economy of subjects is necessary and constant or zero covariance is not possible by any randomization procedure, one of the other two approaches would have to be used.

When there is a lack of homogeneity in the population variance-covariance matrix ($\Sigma$), Bock (1963), Gaito & Wiley (1963), and Lana & Lubin (1963) have recommended the use of Hotelling's $T^2$ statistic, which when modified by Rao's method (1952, pp. 239-244), provides a test for homogeneity of means --by an exact test for differences between correlated means.

There are, however, certain disadvantages in the use of multivariate procedures. In order to calculate $T^2$, it is necessary to obtain the inverse of the sample variance-covariance matrix ($V_k$). Without the use of a high-speed data processing system and corresponding computer programs, the entire operation is both laborious and time consuming. Danford, Hughes & McNee (1960) have shown that when the assumption of equal covariances is fulfilled, the usual univariate procedures lead to a more powerful test than Hotelling's $T^2$. However, as _n_ becomes large; i.e., as the d.f. increases, both procedures lead to equally powerful tests. They also found that when the assumptions of homogeneity of variance and covariance did not hold, and both multivariate and univariate Anova tests were made on the same data, it was noted that, "... asymptotically, the univariate and multivariate tests are identical". The conclusions drawn from their study are, "... essentially, the same inferences are made from the univariate and multivariate analyses".

The final approach, where the d.f. for a univariate F are adjusted to account for bias, has been the most popular with respect to the solutions suggested and derivations of correction factors.

The initial work done with correction of the univariate F distribution by adjustment of the d.f. was due to Box (1954). Box assumes a multivariate normal distribution and shows that under the null hypothesis the true distribution of the univariate F can be approximated by adjusting the d.f. for the biased F. The d.f. for numerator and denominator are both multiplied by a fraction, epsilon ($\epsilon$). ($\hat{\epsilon}$ is used to designate the sample estimate of epsilon). Here, an attempt is made to account for the positive bias by a reduction of the d.f.; consequently, a larger value for F is needed to reject $H_o$ when the statistic is biased.

The test ratio is

$$F_{(k-1)\epsilon\ ,\ (k-1)(n-1)\epsilon}$$

... where $\epsilon$ can be estimated from $V_k$, where $V_k$ is the sample variance covariance matrix.

Epsilon is equal to

$$k^2(\bar{v}_{tt} - \bar{v})^2/(k-1)\ (\sum_{st}^{kk}\sum v_{ts}^2 - 2k\sum_{t=1}^{k}\bar{v}_t^2 + k^2\bar{v}_{..}^2).$$

Then, $v_{ts}$ are the elements of $V_k$, $\bar{v}_{tt}$ is the mean of the diagonal terms; i.e., the variances, $\bar{v}_t$ is the mean of $t^{th}$ row or column, and $\bar{v}_{..}$ is the grand mean of V.

Box developed this correction for the case of the single classification. Geisser & Greenhouse (1958) and Bhat (1958) have extended Box's results to include multiple groups. Geisser & Greenhouse, working with a two factor model, analogous to Lindquist's Type I design (1953), demonstrated that the lower bound on $\varepsilon$ is $(k-1)^{-1}$ and the upper bound is unity.

In a later article, Greenhouse & Geisser (1959) recommend the following steps in testing an F ratio, from a repeated measures model, for significance. First, the univariate F is computed and the regular test @ (k-1), (n-1)(k-1) d.f. is made. If however, the regular test results in a rejection of $H_o$, the next step would be a (conservative) test @ 1, (n-1) d.f... But if this test shows no evidence of significance, a problem arises: should Box's epsilon statisitic be used to make an approximate test of the null hypothesis?

Lana & Lubin (1963) have interpreted Geisser & Greenhouse's rationale in the following manner

> They (Geisser & Greenhouse) argue that since no one has shown what sample estimate of epsilon is most appropriate, and the robustness of epsilon has not been investigated, it is best to use a conservative test. (Lana & Lubin, 1963, pp.733)

Lana & Lubin also comment on Geisser & Greenhouse's position with respect to the situation where $H_o$ is rejected when the regular test is used, but not with the conservative test. "... Geisser & Greenhouse apparently would next try Box's approximate test, using a sample estimate of epsilon. We would recommend an exact multivariate test such as Rao's (the modification of Hotelling's $T^2$)."

It would seem that Box's approach would be one of last resort. The implication here is that either the sample estimate of epsilon is a poor one, for some undisclosed reason, or that the experimenter has no idea as to the structure of the population variance-covariance matrix. When no adequate a priori estimates of $\epsilon$ can be made, one would assume maximum bias to operating and uses a conservative test; i.e., @ 1, (n-1) d.f. In their definitive article, Geisser & Greenhouse's final recommendation is to use the conservative test although the correction may be too conservative. Thus, we now have a negative bias in the F test rather than a positive one.

Another problem related to the test for a significant F is the Type II error. The fact that the Type I error is underestimated when the population variance-covariance matrix shows a lack of homogeneity, the probability of a Type II error may not be the same as for the unbiased test. The effect of not reducing the d.f. to compensate for an inflated F is to raise $\alpha$ level, e.g.; an $\alpha$ of .05 is really .07: an increase in the probability of the Type I error. All other things being equal, this will result in a corresponding decrease of the probability of a Type II error. Thus, the power of the F test is artificially inflated. Conversely, applying a conservative or overcorrection to the d.f. will result in a decrease of the probability of a Type I error, and a corresponding <u>loss</u> in power. Just how much the probability of the Type II error is effected by a lack of homogeneity of covariance is not known.

In order that the power of the F test may be evaluated when these correction procedures are applied to limit the bias due to heterogeneous covariances, the power functions of the repeated measures design needs to be known. There is no indication in the literature of adequate power curves to fit repeated measures models. However, for those models where scores are independent of one-another, we see that power is dependent upon three parameters: degrees of freedom in the numerator ($\nu_1$), degrees of freedom in the denominator of the F ratio ($\nu_2$) and a noncentrality parameter, designated by delta ($\delta$).

Dixon & Massey (1957) and Scheffé (1959) have shown that delta is dependent upon: the number of cases sampled per treatment or condition, the variation between population means and an estimate of the error variance in the population. For repeated measures designs, especially where Σ shows heterogeneity, an estimate of the error variance is difficult to make. Since the error variance is a direct function of the average covariance, the noncentrality parameter will differ from the independent measures model where the expected value for the covariance is zero.

The noncentrality parameter mentioned by Scheffé (1959) has no term to account for non-zero covariance. Consequently, the problem arises: are the three (aforementioned) parameters of Scheffé sufficient to account for the power of the repeated measures model? However, the problem becomes critical only when the power of the independent and dependent measures models are to be compared along a continuum of noncentrality. For the case where the average correlation in the population is the same and homogeneity of variance exists between treatment conditions, comparisons of different power levels of sample F's drawn under varying sample sizes, degree of heterogeneity in Σ, or number of treatment conditions is possible using Scheffé measure of noncentrality, where the error variance term is equal to $(1-\bar{r}^2)$, where $\bar{r}$ is the average correlation between the k treatments.

## CHAPTER II

## METHOD AND PROCEDURE

The procedure used in the present research involved calculating a large number of statistics, each based upon samples which were drawn at random from a universe having specified characteristics. On the basis of the statistical tests made on these samples, the probabilities of the Type I and II errors were determined.

### 1. Simulation of Data

Let $Z_s$ equal a k x 1 vector drawn from a _k_ variate multivariate universe, k denoting the number of treatment classifications for which a given subject produces a measureable response, where

$$E\ (\ Z_s\ Z_s'\ ) = I_r$$

This vector, $Z_s$, contains _k_ independent scores for any one individual in a universe of _n_ people and n such vectors may be represented as a sample of n x k scores.

Let $R_k$ equal a universe of intercorrelations among errors, where

$$R = F\ F'$$

To generate correlated sample scores from the universe of errors, we let

$$F\ Z_s = Z_r, \text{ where}$$

$$E\ (\ Z_r\ Z_r'\ ) = E\ (\ F\ Z_s\ Z_s'\ F\ ) = F\ (\ E\ Z_s\ Z_s'\ )\ F'$$

$$= F\ F' \ = R.$$

Thus, to impose intercorrelation matrix, R on the errors, the vector $Z_s$ is sampled from a universe of scores and premultiplied by factor matrix, F.

Let $u_i$ ($i = 1,2, \ldots, k$) be a vector of means, where $\Sigma u_i = 0$. In order to simulate raw scores with universe means given by $u_i$, the $i^{\underline{th}}$ treatment mean is added to the $i^{th}$ element of the vector of error scores. The result, $S_r$ is a k x 1 vector of _k_ raw scores for a given individual; _n_ such vectors are drawn from the universe with the resultant raw scores having appropriate u and sigma in expectation.

2. The Simplex

It was previously mentioned that R will not be homogeneous when there is a practice or carry over effect across treatments. In experimentation where repeated measures models are appropriate, this is the usual case. The simplex is one of the more common forms of sampling covariance matrices encountered in psychological research, especially where "learning" is involved. Anderson (1958) and Jones (1960) describe the patterning of correlations in the simplex as a decrease in magnitude of the correlations in successive diagonals away from the main diagonal.

The principle of the simplex is like that of a folding telescope, in that the different stages or components of the simplex follow the principle of inclusion (Jones, 1960). That is to say, $R_1$ is included in $R_2$, which is included in $R_3$, ..., which is included in $R_j$. Algebraically,

$$R_1 = f_1$$

$$R_2 = f_1 + f_2$$

$$R_j = f_1 + f_2 + \ldots + f_j \text{ (after Jones, 1960).}$$

The perfect simplex is so structured as to have $R_1$ and $F_1$ subject to the same conditions--except, the structured variables don't all have equivalent variances (Jones, 1960).

Table 2.1

| VARIABLE | 2 | 3 | 4 |
|---|---|---|---|
| 1 | .60 | .40 | .30 |
| 2 | | .80 | .50 |
| 3 | | | .90 |

Matrix of Intercorrelations Having Perfect Simplicial Form.

In addition, the matrix could be Grammian in form. In the manipulation of psychological variables or the measurement of psychological traits, one would not expect to find the intercorrelation matrix to have negative roots.

3. Evaluation of the Magnitude of the Type I Error

To recapitulate, various procedures have been recommended to test the F statistic resulting from a repeated measures design where the assumption of homogeneity of covariance has been violated. These procedures have been suggested because F, sampled under conditions of heterogeneity in $\Sigma$ will become inflated, and the use of the F distribution @ k-1 and (n-1)(k-1) d.f. to test $H_o$ results in a positive bias in F when $H_o$ is true. Assuming that the experiment from which the sample covariance matrix was constructed can't be altered, three alternatives for testing $H_o$ against some alternative, $H_o$, are open to the experimenter:

a. perform the usual test @ k-1 and (n-1)(k-1) d.f.;

b. use an approximation of the central F distribution by adjusting the d.f.;

c. or use an exact, multivariate statistic such as $T^2$.

By sampling from a universe where the parameters $\nu_1, \nu_2$ (d.f.'s) as well as the degree of bias in $\Sigma$ (as indexed by Box's epsilon) are known, just how much of a departure from the assumptions of the "usual" F test results in a Type I error as different from the alpha level chosen to test F, could be determined empirically. Testing these same data using sample estimates of epsilon to adjust $\nu_1$ and $\nu_2$ would enable us to measure the degree to which this correction procedure removes the bias in F. As a control, $T^2$ could also be calculated on these same data. This would provide a control in the sense that $T^2$ would be the appropriate <u>exact</u> statistic to use when $\Sigma$ is not homogeneous.

Lana & Lubin (1963) suggest that the sample estimate of epsilon may not be the best estimate of the parameter, epsilon. In the present study, since the value of the parameter epsilon is known, adjusting the $\nu_1$ and $\nu_2$ parameters using the population value of epsilon, and comparing the resulting probability of the Type I error with those determined using the "biased" F test, F- adjusted by the statistic epsilon and $T^2$, the relative merit of the use of the actual parameter, epsilon, could be determined.

### 4. Power of the Tests Under Conditions of Heterogeneity in $\Sigma$

Testing $H_o$ against some alternate hypothesis, $H_1$, when we can specify that $H_o$ is false, we are confronted with the probability of committing a Type II error. Using the various approximations of the F distribution, and the $T^2$ statistic, knowing the values of the $\nu_1$, $\nu_2$, $\epsilon$, and $\delta$, parameters, the power of these tests may be determined by the relative number of times $H_o$ may be rejected. Unlike the Type I error, there is no readily available index, such as alpha, which we may use to compare the empirical values of the Type II error. Power curves for the F-test are presented in Dixon & Massey (1957). Since the present model departs from the fixed constants model as given by Dixon & Massey, a comparison of the empirical value of the power of these tests and any expected value, is not possible. However, the relative power of the various test statistics may be compared.

### 5. Sampling Characteristics of Epsilon

Since the sampling distribution of epsilon is unknown, any attempts to associate a probability of getting a sample value differing from its expected value, by some magnitude, would be beyond the scope of this study. Selected descriptive measures of how well a limited sample of epsilon statistics from a specified population, compares with the parameter, epsilon, are used.

### 6. Selection of Parameters

A. Average universe correlation.

The R matrices were selected such that the average correlation, between treatments, in the universe was constant, throughout, but the resultant variance-covariance matrices differed in degree of heterogeneity. The basic axis of difference was in terms of the magnitude of the parameter, epsilon, i.e., the "severity" of Box's correction: the smaller the value of epsilon, the more d.f. lost.

The average correlation[2] was determined such that the reliability of the data was both constant for the entire study and uniformly high from matrix to matrix; e.g., r = .64, where

$$r = 1 - \frac{\sigma_e^2}{\sigma_t^2} \quad \text{and } \sigma_e^2 = \sigma_t^2 - \bar{\rho}\,\sigma_t^2 \text{, thus...}$$

$$r = 1 - \frac{(\sigma_t^2 - \bar{\rho}\,\sigma_t^2)}{\sigma_t^2} = \bar{\rho}.$$

($\sigma_t^2$ = total variance)

B. R matrices and the parameter epsilon.

The index used to define the degree of heterogeneity in $\Sigma$, from which sample covariance matrices were drawn, is the parametric value of Box's epsilon. By specifing a series of such matrices, where the value of $\epsilon$ is manipulated along the 1 to $(k-1)^{-1}$ continuum, effects of the change in heterogeneity in $\Sigma$ may be observed in a distribution of sample Fs from $\Sigma$. As heterogeneity increases what is the effect on the F statistic? The values of epsilon were ordinally chosen such that a high, medium and low bias matrix $\Sigma$, could be defined. Since the range of sample values of epsilon is dependent upon the $\nu_1$ parameter, a different series of epsilons was used for each $\nu_1$ parameter used in the study. Tables 2.2 through2.3 contain a complete layout of the values of epsilon chosen for each $\nu_1$ value used.

C. k, the number of treatment classifications.

One advantage of the analysis of variance is that k means are simultaneously compared. Any additional comparisons, using the t test, results in a change in the alpha level, as given by the following equation...

$1 - (1- \text{alpha})^h$, where h specifies the number of comparisons being made among means.

Consequently, if one wants to take advantage of the stability of alpha using a proper exact test and the central F distribution, at least three means of treatment should be compared. For the present study, at least three treatments, are necessary if a simplex patterning of correlations in R is desired. Two values for k were used for this study. A $k = 3$ was chosen in order that a minimum value in the acceptable range of $\nu_1$ could be had and $k = 5$ was arbitrarily chosen.

D. n, the number of subjects or sampling units per element of $\Sigma$.

The values of n in this study, as in most models for the use of testing statistical hypotheses, have an effect on the sampling error the sample covariance matrix will contain. Generally speaking, as n increases, the sampling error decreases. In terms of the power of the F test, the sampling error in the covariance matrix must be estimated in order that the noncentrality parameter may be defined. As the sampling error decreases, all other factors being constant, power of the test increases.

It was mentioned that both an approximate and exact test was used to test $H_o$ for significance. The exact test uses the F distribution to determine the p level for a sample value of $T^2$, however, the d.f. parameters are not determined in the same way for a univariate F as for $T^2$. The d.f. for the covariance matrix from which the value of $T^2$ is determined must be at least n-k-2. Therefore, any value of $T^2$ calculated from a $\Sigma$ not having the required minimal d.f. will not be meaningful, but in some cases, e.g.; where n-k-1 is non-negative, and F statistic will be meaningful.

These facts have important implications when selecting values for _n_. If $T^2$ is to be used and the data are appropriate for the model, sufficient d.f. must be present for the $T^2$ test. However, as in the present study, where it is of interest to know the empirical values of the Type I and Type II errors of the F test when $T^2$ can't be used, this comparison cannot be made.

The values of _n_ were chosen such that: 1) a minimum d.f. was present for an F but not a $T^2$ test; 2) a minimum d.f. was present for the $T^2$ test; 3) and two other values, n= 10 and n= 15 were also chosen. The first two values for _n_, just above and just below the required number for the $T^2$ test, were dif-

Table 2.2

Population Covariance Matrices

(k= 3; n= 4, 6, 10, 15)

| | | | Heterogeneity Index |
|---|---|---|---|
| 1.00 | .66 | .64 | .99 |
| | 1.00 | .66 | |
| | | 1.00 | |
| 1.00 | .59 | .44 | .74 |
| | 1.00 | .83 | |
| | | 1.00 | |
| 1.00 | .48 | .17 | .54 |
| | 1.00 | .92 | |
| | | 1.00 | |

Table 2.3

Population Covariance Matrices

(k= 5; n= 6, 7, 10, 15)

| | | | | | Heterogeneity Index |
|---|---|---|---|---|---|
| 1.00 | .63 | .63 | .63 | .62 | .99 |
| | 1.00 | .66 | .65 | .64 | |
| | | 1.00 | .66 | .67 | |
| | | | 1.00 | .68 | |
| | | | | 1.00 | |

| | | | | | Heterogeneity Index |
|---|---|---|---|---|---|
| 1.00 | .58 | .52 | .48 | .42 | .78 |
| | 1.00 | .73 | .70 | .58 | |
| | | 1.00 | .78 | .75 | |
| | | | 1.00 | .79 | |
| | | | | 1.00 | |

| | | | | | Heterogeneity Index |
|---|---|---|---|---|---|
| 1.00 | .36 | .31 | .17 | .14 | .50 |
| | 1.00 | .83 | .75 | .52 | |
| | | 1.00 | .88 | .84 | |
| | | | 1.00 | .89 | |
| | | | | 1.00 | |

ferent for the various values of k, as should be evident by the fact that $\nu_2$ is dependent upon k as well as n. Tables 2.2 through 2.3 contain a complete description of the values of n chosen for the study.

E. Values of the treatment means.

When the probability of the Type I error was investigated, the values of the constants $\alpha$, were equal across treatments. Since the Type I error is relevant only when population means are equal, the constants were set equal to zero when the null hypothesis was true.

Investigation of the Type II error and its complement, power, involves another parameter, the noncentrality parameter. The power of the F test is dependent upon three parameters, d.f. for the numerator, d.f. for the denominator and the non-centrality parameter designated, delta, ($\delta$).

The dimensions of the ${}_{n}Z_{R}^{k}$ matrix determines the two d.f. parameters and the values of the fixed constants ($\alpha$'s), relative to the population error variance and sample size, determines the non-centrality parameter.

Scheffé (1959) indicates the non-centrality parameter is equal to

$$\left[ \frac{n\,\Sigma(u-u..)^2}{k} \times \frac{1}{\sigma_e^2} \right]^{\frac{1}{2}}$$

... for the fixed constants model with <u>independent</u> observations, viz.; non-repeated measures.

Since assumptions of the model were violated no appropriate non-centrality parameter is indicated. $\delta^2$ was set equal to the variance of the constants $\alpha_j$ weighted by n, the sample size for each of the treatment observations, e.g.;

$$\delta^2 = \frac{n \Sigma (\alpha_j - \alpha_{..})^2}{k}$$

The restriction, $n_1 = n_2 = n_j$, was used such that each treatment received equal weighting.

Ten values of noncentrality, illustrated in Table 2.4 ranging from .01 to 20.00 were chosen such that a reasonably continuous function for power was obtained. The usual comparisons of the power of two tests having the same d.f. is not possible, although the F and $T^2$ statistics were both the results of samples from the same variance-covariance matrix. Power comparisons were made between the various approximate F tests and $T^2$ in terms of <u>k</u> and <u>n</u> rather than $\nu_1$ and $\nu_2$.

F. Number of sample covariance matrices.

For the investigation of the Type I error 1000 sample covariance matrices were drawn for each n, k, and epsilon combination, resulting in 12,000 sample covariance matrices. For the power comparisons 500 sample covariance matrices were drawn for each n, k, epsilon and noncentrality combination, resulting in 60,000 covariance matrices.

Table 2.4

Values for $\delta$ - the Noncentrality Parameter

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| .01 | .050 | .150 | .500 | 1.00 | 1.25 | 2.00 | 2.50 | 5.00 | 20.0 |

# CHAPTER III

## RESULTS AND DISCUSSION

Samples were generated and evaluated for the various combinations of the n, k, heterogeneity and noncentrality parameters. When noncentrality was zero, the Type I errors for the various tests of $H_0$ were determined by the percentage of times $H_0$ was rejected. When noncentrality was greater than zero, power of the test was calculated in a similar manner.

### 1. The Type I Error.

The results of the probabilities of the Type I error are shown in Tables 3.1 through 3.5. Appendix I contains tables of the probabilities for each of the conditions. The results of Table 3.1 show a general pattern for each of the tests on $H_0$ for both alpha = .01 and .05.[3]

Results using the exact F test ($\nu_1, \nu_2$) indicate an increase in positive bias in the statistic as heterogeneity in the population covariance matrix is increased. A positive bias is also observed even when epsilon (the value of which is used to index heterogeneity) approaches unity. This would indicate that the exact F test is sensitive to slight departures from the assumption of homogeneity of covariance.

The general pattern of results for tests on $H_0$ using the Box statistic approximation to the central F distribution, F@$\nu_1$e, $\nu_2$e, indicates that as heterogeneity was increased bias in the statistic was slightly negative for minimal heterogeneity and becomes slightly positively biased for departures from homogeneity of covariance. It should be noted that the degree of departure from homogeneity of covariance had a negligable effect on bias. When the Anova model was appropriate, i.e.; epsilon approaches unity, the approximate F test produced relatively less bias than the exact F test. Using the Box statistic, therefore, will serve to limit bias due to sampling fluctuations in $V_k$ when the population covariance matrix indicates homogeneity.

The results of tests on $H_0$ using the population value of epsilon to adjust $\nu_1$ and $\nu_2$ of F were similar to those obtained when $H_0$ was tested by adjustment of d.f. with the sample estimate of epsilon. It is observed that this procedure is ineffective when the covariance matrix is homogeneous because this test is the same as the exact F. It would seem that the sample estimate of epsilon is a good one judging by the similar, (unbiased) F statistics obtained with both the population and sample estimates of the statistic.

The test on $H_0$, as suggested by Geisser & Greenhouse, whereby the $\nu_1$ and $\nu_2$ parameters assume the values of 1 and n-1, respectively, produced rather conservative Type I errors. That is to say, the use of this procedure, in all cases, produced a negatively biased F statistic. When the model was appropriate, a severe loss of d.f. results in near zero probabilities. Increasing heterogeneity resulted in only a slight decrease in the bias, but the resulting Type I error remained highly biased. The results just reported are based on the probabilities averaged across the n and k parameters. The data as a whole, however, reveals an interesting fact about the use of this "conservative" procedure. Only when heterogeneity is maximal, in a statistical sense, and the Grammian structure of R is demanded, as in the case of epsilon approaching .50 for $k = 3$, does the Geisser & Greenhouse procedure yield a relatively unbiased test. Failure to obtain similar results when k was 5 is explained by the fact that for larger order R matrices, maximal heterogeneity of $1/k-1$ is <u>not</u> obtainable if R is to have Grammian properties and positive intercorrelations. Therefore, the value of $\epsilon = .50$ when $k = 5$ does not fulfill the criterion of being the lower bound for the domain of epsilon. Only when data are sampled from covariance matrices with maximal heterogeneity is such a severe curtailment of d.f. warranted.

Results obtained using the exact, multivariate $T^2$ test show Type I errors similar to those produced by the approximate F test using the Box procedure to adjust d.f. In general, when there was a departure from the assumption of homogeneity of covariance, $T^2$ produced Type I errors converging on alpha.

Table 3.2 and Table 3.3 show the results of the probabilities of the Type I error for different k for various degrees of heterogeneity in the covariance matrix. The results of the exact F test ($\nu_1$, $\nu_2$) indicate that as k increased from 3 to 5, the positive bias in the statistic also increased. The over-all effect becomes more pronounced as heterogeneity increases; for the case where heterogeneity was moderate, the bias was somewhat less for k = 3, than k = 5.

The test on $H_O$ using the sample estimate of epsilon to adjust d.f. shows an overall effect of a shift in bias, from negative to slightly positive, as k increases. As heterogeneity increased, a stability of the Type I error was noted when k = 3, but when k = 5, bias was slightly negative for low heterogeneity and became slightly positive when heterogeneity was maximal. Tests using the parametric value of epsilon to approximate F indicate that an increase in k results in an overall increase in the Type I error. Increasing heterogeneity didn't result in an appreciable change in the Type I error when k = 5; the results for k = 3 are not as clear. Tests using the Geisser & Greenhouse procedure show that an increase in k results in an increase in the degree of negative bias. Increasing heterogeneity serves to decrease the bias, but only slightly.

As departures from homogeneity of covariance were evidenced, the bias in the $T^2$ test became negligible, regardless of the value of k.

Tables 3.4 and 3.5 show the effect, on the Type I error, of an increase in n for varying degrees of heterogeneity in the covariance matrix. When the covariance matrix was homogeneous, we were sampling from a population which was congruent with the assumptions of the model. As we increase n and consequently, the $\nu_2$ parameter, the sampling error for any given (exact) statistic should decrease and the Type I error should approach alpha when the statistic is free of bias. If the statistic in question has relatively less sampling error as n increases, it is said to be consistent.

For all tests employed on $H_0$, with the exception of the exact F test @ $\nu_1$, $\nu_2$, the overall effect of increasing n was negligable. For increasing n, the consistent properties of the exact F statistic were observed. For large n, the probability of the Type I error and alpha tend to converge, indicating the unbiased characteristic of the statistic. Increasing heterogeneity in the covariance matrix, the effect of increasing n showed a relatively consistant decrease in bias, though the bias was still largely positive.

The effect of increasing n for tests on $H_0$ using Box's sample estimate to adjust d.f. of F, was negligable. This statistic remained relatively stable even as heterogeneity was increased. Only a slight positive bias was evidenced for maximal heterogeneity. Similiar results were obtained when the parametric value of epsilon was used to approximate the F distribution.

The Geisser & Greenhouse procedure produced a negatively biased statistic (close to zero in some cases); as heterogeneity increased, the bias decreased somewhat. The results reported in tables 3.4 and 3.5 indicates the $T^2$ test to be robust to heterogeneity of covariance and produces a relatively unbiased test for all n's across the heterogeneity continuum. Only when heterogeneity was evidenced in the covariance matrix did the test produce a negative bias which tends to increase with increases in n.

Conclusions about the Type I error.

Increases in the heterogeneity of the population covariance matrix produced an increasing degree of positive bias in the exact F statistic @ $\nu_1, \nu_2$. As departures from homogeneity of covariance was evidenced, approximate tests using the F distribution, adjusted by the sample estimate and parametric values of Box's epsilon, as well as the exact multivariate $T^2$ statistic, tended to produce relatively unbiased tests of $H_0$. Only when heterogeneity, approached the maximum value (as indexed by $\epsilon = 1/k-1$) was the Geisser & Greenhouse procedure of testing $H_0$, using an approximation to the central F distribution, effective in producing a relatively unbiased test. Increasing k for various degrees of heterogeneity produced slight increases in the Type I error for all tests on $H_O$. The exact F showed a decrease in bias with an increase in n, illustrating the consistency of the statistic. For the other tests on $H_O$ employed in this study, the effect of changes in n were negligable over the range employed.

Table 3.1

Probabilities of the Type I Error:
Averaged across n and k

| Alpha Level | Heterogeneity | TEST F | $\epsilon$ | $\epsilon'$ | 1,n-1 | $T^2$ |
|---|---|---|---|---|---|---|
| .01 | Min. | .01813 | .00750 | .01813 | .00063 | .00505 |
| | Mod. | .02963 | .01225 | .01088 | .00231 | .00967 |
| | Max. | .05088 | .01350 | .01300 | .00525 | .01260 |
| .05 | Min. | .07488 | .04863 | .07488 | .01575 | .03583 |
| | Mod. | .09286 | .05638 | .06025 | .02475 | .04117 |
| | Max. | .11838 | .06238 | .06413 | .04125 | .06083 |

Table 3.2

Probabilities of the Type I Error: at $\alpha$ = .01
Averaged across n

| k | Heterogeneity Index | F | $\epsilon$ | $\epsilon'$ | 1,n-1 | $T^2$ |
|---|---|---|---|---|---|---|
| 3 | .99 | .01775 | .01100 | .01775 | .00125 | .00433 |
| | .74 | .03250 | .01525 | .00675 | .00450 | .00800 |
| | .54 | .04325 | .01200 | .00900 | .00825 | .01366 |
| 5 | .99 | .01850 | .00400 | .01850 | .00000 | .00566 |
| | .78 | .02675 | .00925 | .01500 | .00012 | .01133 |
| | .50 | .05850 | .01500 | .01700 | .00225 | .01066 |

Table 3.3

Probabilities of the Type I Error: at $\alpha$ = .05
Averaged across n

| k | Heterogeneity Index | F | $\epsilon$ | $\epsilon'$ | 1,n-1 | $T^2$ |
|---|---|---|---|---|---|---|
| 3 | .99 | .07275 | .05575 | .07275 | .02750 | .03300 |
| | .74 | .09475 | .06450 | .05425 | .04000 | .03800 |
| | .54 | .11050 | .06325 | .05900 | .05800 | .06833 |
| 5 | .99 | .07700 | .04150 | .07700 | .00400 | .03866 |
| | .78 | .09100 | .04825 | .06625 | .00950 | .04433 |
| | .50 | .12625 | .06150 | .06925 | .02450 | .05333 |

Table 3.4

Probabilities of the Type I Error: at $\alpha = .01$
Averaged across $k$

| $n$ | Heterogeneity | F | $\epsilon$ | $\epsilon'$ | 1,n-1 | $T^2$ |
|---|---|---|---|---|---|---|
| 4 (or) 6 | Min. | .0200 | .0050 | .0200 | .0000 | |
| | Mod. | .0295 | .0105 | .0110 | .0030 | |
| | Max. | .0550 | .0115 | .0110 | .0045 | |
| 6 (or) 7 | Min. | .0225 | .0080 | .0225 | .0015 | .0060 |
| | Mod. | .0290 | .0080 | .0105 | .0015 | .0090 |
| | Max. | .0550 | .0140 | .0125 | .0060 | .0110 |
| 10 | Min. | .0205 | .0105 | .0205 | .0000 | .0060 |
| | Mod. | .0245 | .0125 | .0085 | .0025 | .0090 |
| | Max. | .0490 | .0125 | .0135 | .0050 | .0145 |
| 15 | Min. | .0095 | .0065 | .0095 | .0010 | .0030 |
| | Mod. | .0310 | .0180 | .0135 | .0030 | .0110 |
| | Max. | .0440 | .0160 | .0150 | .0050 | .0110 |

Table 3.5

Probabilities of the Type I Error: at $\alpha$ = .05
Averaged across k

| n | Heterogeneity | F | $\epsilon$ | $\epsilon'$ | 1,n-1 | $T^2$ |
|---|---|---|---|---|---|---|
| 4 (or) 6 | Min. | .0770 | .0395 | .0770 | .0135 | |
| | Mod. | .1055 | .0530 | .0635 | .0235 | |
| | Max. | .1430 | .0645 | .0650 | .0440 | |
| 6 (or) 7 | Min. | .0900 | .0560 | .0900 | .0170 | .0360 |
| | Mod. | .0890 | .0515 | .0590 | .0195 | .0430 |
| | Max. | .1125 | .0605 | .0625 | .0410 | .0580 |
| 10 | Min. | .0715 | .0510 | .0715 | .0180 | .0395 |
| | Mod. | .0825 | .0565 | .0540 | .0255 | .0415 |
| | Max. | .1120 | .0670 | .0710 | .0430 | .0680 |
| 15 | Min. | .0610 | .0480 | .0610 | .0145 | .0320 |
| | Mod. | .0945 | .0645 | .0645 | .0305 | .0390 |
| | Max. | .1060 | .0575 | .0580 | .0370 | .0565 |

## 2. The Power of the Tests

The data in Appendix II shows the power of the different F tests and that of $T^2$ when these statistics were tested for significance at the .05 and .01 levels for the various combinations of n, k, and index of heterogeneity in the covariance matrix. For each point on these graphs, 500 statistics were sampled.

In general, when the model was said to hold, i.e.; the covariance matrix was homogeneous, the exact F and the Box procedure for approximating the F distribution using the parameter $\epsilon$ to adjust d.f., were the most powerful, with the other Box procedure using the parameter, $\epsilon$, the Geisser & Greenhouse, and $T^2$ tests showing a decreasing degree of power, in that order.

The criterion chosen to provide a basis of comparison of the obtained power curves is twofold.

1. Which test is most powerful, and does this power hierarchy among tests remain consistent as noncentrality increases?
2. How divergent are the power curves for a given comparison?

Comparing power curves for all n, k, and epsilon values, when $H_0$ is tested at .01 and .05, we observe that as the $\alpha$ level increases (from .05 to .01), the curves for the five tests show a greater degree of divergence and tend to have the expected level of power.

As the degree of heterogeneity in the population increased, the exact F statistic remained consistently the most powerful test and maintains the same level of power. The divergence among the power curves for the other tests decreased and approached that of the exact F as both heterogeneity in the covariance matrix and n, increased to their maximum values for this study. As the $\nu_1$ parameter increases from 2 to 4, a greater divergence of the power curves for the five different tests on $H_0$ for the data drawn from the same population was found. As expected, the overall level of power for all five tests is higher as $\nu_1$ and $\nu_2$ increases.

With respect to comparisons among the five tests on $H_0$, where the statistics were sampled from the same population, in all but one case, the exact F was most powerful. Tests using the population value of epsilon and sample value of the statistic, were second and third in the power hierarchy. Although the differences in power of the aforementioned Box tests and the exact F was slight, they were considerably more powerful than the Geisser & Greenhouse and $T^2$ procedures. The $T^2$ and Geisser & Greenhouse tests showed a consistant reversal of power across the noncentrality continuum. As a rule, $T^2$ was more powerful for low noncentrality but as noncentrality increased, the Geisser & Greenhouse procedure became more powerful. Apparently $T^2$ is more sensitive to differences between means when these differences are slight. As these differences became relatively larger, the Geisser & Greenhouse procedure produced the more sensitive test.

Only one departure from these generalizations was observed. When $\nu_1 = 2$ for $n = 10$, and the degree of heterogeneity was moderate, $T^2$ produced the most powerful test. For various levels of non-centrality, $T^2$ was the least powerful test, relatively speaking, for all other parameters studied.

## General Conclusions Concerning Power Of The Tests

Although power comparisons were made between statistics which showed a degree of bias of type I error when $H_0$ was true, the exact F test for the current model tends to remain robust with respect to power. When the assumption of homogeneity of covariance was violated, as indicated by a value of the parameter, epsilon <1, the power of the approximate tests using the F distribution increased and approached that of the exact F test. This indicates that the various correction procedures are effective in maintaining relatively high levels of power for approximations of the F distribution, even though a curtailment of the d.f. results from these procedures, when the assumption of homogeneity of covariance has been violated.

Whether or not the power of these tests is inflated was not determinable from this study. However, if the power of the exact F test was inflated, any increase of heterogeneity in the covariance matrix did not tend to further inflate the power of the statistic. Again we see that the F statistic is fairly robust, powerwise, to violations of the assumption of homogeneity of covariance.

### 3. Sampling Characteristics of Box's $\epsilon$

Table 3.7 contains the means and standard deviations of the epsilon statistics which were used to adjust the d.f. of the F statistic for data collected to evaluate the Type I error.

The distributions of $\epsilon$ show marked skewness when the number of treatment classifications of the covariance matrix is three. When k is increased to five, the distribution becomes more symetrical.

As the d.f. of $\Sigma$ increased (an increase of n, relative to k), the variability in epsilon decreased, as indicated by a reduction in the standard deviation of $\epsilon$. An increase in the d.f. in $\Sigma$ resulted in a greater disparityof the sample mean of the distribution and the value of the parameter, epsilon, when k was increased from three to five. Therefore, as the number of elements which compose $\Sigma$ increased, the greater was the departure of the sample mean of $\epsilon$ from the parametric value of the statistic. It was also observed that as the index of heterogeneity of the covariance matrix, given by the parameter, epsilon, showed greatest heterogeneity, when k= 3, the relation of the sample mean and the parameter, $\epsilon$, remained invariant with an increase in d.f. in $\Sigma$. Under this condition, a convergence of the sample mean and the parametric value of $\epsilon$ was evidenced. The frequency distributions of these epsilon statistics tended to be skewed positively, i.e.; away from the end of the continuum indicating extreme heterogeneity, and were leptokurtic. Why this happened when heterogeneity was maximal, but not when minimal, is not clear.

The point biserial correlation coefficients shown in tables 3.6 and 3.7 were used as descriptive measures to determine how well the sample estimates of epsilon could predict the value of the parameter. If, for example, we draw samples of $\epsilon$ from two distinct covariance matrices, differing only in the value of the parameter, epsilon, a point biserial correlation coefficient between the samples drawn from these different populations can provide an index of the degree to which we can specify the population values (parameters) of epsilon from knowledge of the means and standard deviations of the sampling distribution of the statistic.

When samples were drawn from the extremes of the range of the domain of epsilon, prediction was superior to when samples were drawn from the midrange. As the d.f. of $\Sigma$ increased, so did the predictability of the parameter, epsilon. When the order of $\Sigma$ was maximal, for the conditions studied, prediction of $\varepsilon$ from the middle portion of the range of the statistic was slightly superior to the case when $k = 3$.

Table 3.6

Descriptive Measures of Samples of Box's Epsilon
(k = 5)

| $\underline{n}$ | Heterogeneity Index | Mean | $\sigma$ | $r_{pt}$ |
|---|---|---|---|---|
| 7 | .99 | .609 | .099 | .466 |
| | .50 | .426 | .081 | .605 |
| | .78 | .553 | .099 | .139 |
| | (overall) | .529 | .121 | |
| 10 | .99 | .690 | .086 | .567 |
| | .50 | .454 | .075 | .707 |
| | .78 | .611 | .098 | .140 |
| | (overall) | .585 | .131 | |
| 15 | .99 | .770 | .076 | .651 |
| | .50 | .464 | .070 | .789 |
| | .78 | .661 | .095 | .138 |
| | (overall) | .632 | .150 | |

Table 3.7

Descriptive Measures of Samples of Box's Epsilon
(k = 3)

| n | Heterogeneity Index | Mean | $\sigma$ | $r_{pt}$ |
|---|---|---|---|---|
| 6 | .99 | .774 | .126 | .508 |
| | .54 | .547 | .051 | .619 |
| | .74 | .694 | .126 | .111 |
| | (overall) | .672 | .142 | |
| 10 | .99 | .851 | .103 | .668 |
| | .54 | .543 | .033 | .738 |
| | .74 | .720 | .112 | .070 |
| | (overall) | .705 | .155 | |
| 15 | .99 | .890 | .082 | .736 |
| | .54 | .543 | .026 | .791 |
| | .74 | .735 | .098 | .054 |
| | (overall) | | | |

# 4. SUMMARY

Samples were generated from populations with known mean $\sigma$ and $\Sigma$ (covariance matrix) for the repeated measures design for specified values of the $\nu_1$ and $\nu_2$ parameters. The values of n for a k of three were 4, 6, 10 and 15, and for k = 5 were 6, 7, 10, and 15. Heterogeneity in $\Sigma$ was defined by the severity of the loss of d.f., as indexed by the parametric value of Box's epsilon, $\epsilon$. Minimum, moderate and near maximal conditions of heterogeneity were established. When the value of the noncentrality parameter was zero, the Type I error was studied; when noncentrality was greater than zero, the power of the tests was investigated.

The results of this study indicate that as d.f. and heterogeneity of covariance increased in the population, the approximations to the central F distribution, as suggested by Box, and the $T^2$ test, become increasingly more effective in accounting for bias in the F statistic resulting from violations of the model. When heterogeneity in $\Sigma$ was maximal for the present study, only then was the Geisser and Greenhouse procedure effective in achieving a relatively unbiased test of $H_0$.

When the power of the various tests was investigated over the various values of noncentrality, it was found that the overall level of power for the exact F remained relatively invariant when homogeneity of covariance was violated. When the approximate procedures using the F distribution were used when violations of the model for the exact F was present in $\Sigma$, and hence "logically" appropriate, the power of these tests approached that of the exact F.

Descriptive measures of the distribution of $\epsilon$ revealed the sample estimates of the parameter to be best when sampling from the extreme ranges of the domain of the statistic. As the d.f. in $\Sigma$ increased, differential prediction of the parameter, given the sample mean, improved.

# APPENDIX I

## PROBABILITIES OF THE TYPE I ERROR

Probabilities of the Type I error are reported for each of the n, k, epsilon or heterogeneity conditions for statistics tested at the .01 and .05 levels.

Probability of the Type I Error.
(k= 5; alpha= .01)

| n | Heterogeneity Index | TEST F | TEST ε | TEST ε' | TEST 1,n-1 | TEST $T^2$ |
|---|---|---|---|---|---|---|
| 6 | .99 | .021 | .004 | .021 | .000 | |
| 7 | | .023 | .002 | .023 | .000 | .006 |
| 10 | | .023 | .007 | .023 | .000 | .008 |
| 15 | | .007 | .003 | .007 | .000 | .003 |
| 6 | .78 | .025 | .006 | .013 | .000 | |
| 7 | | .032 | .006 | .016 | .000 | .009 |
| 10 | | .021 | .011 | .012 | .000 | .012 |
| 15 | | .029 | .014 | .019 | .001 | .012 |
| 6 | .50 | .051 | .009 | .014 | .001 | |
| 7 | | .063 | .015 | .017 | .004 | .011 |
| 10 | | .066 | .016 | .018 | .002 | .010 |
| 15 | | .054 | .020 | .019 | .002 | .011 |

Probability of the Type I Error.
($k = 5$; alpha $= .05$)

| $\underline{n}$ | Heterogeneity Index | TEST: F | $\epsilon$ | $\epsilon'$ | 1,n-1 | $T^2$ |
|---|---|---|---|---|---|---|
| 6 | .99 | .086 | .036 | .086 | .004 | |
| 7 | | .091 | .047 | .091 | .005 | .040 |
| 10 | | .071 | .044 | .071 | .005 | .049 |
| 15 | | .060 | .039 | .060 | .002 | .027 |
| 6 | .78 | .094 | .035 | .064 | .009 | |
| 7 | | .093 | .047 | .068 | .007 | .044 |
| 10 | | .072 | .046 | .057 | .006 | .046 |
| 15 | | .105 | .065 | .076 | .016 | .043 |
| 6 | .50 | .133 | .050 | .062 | .020 | |
| 7 | | .125 | .058 | .067 | .024 | .049 |
| 10 | | .124 | .073 | .081 | .028 | .059 |
| 15 | | .123 | .065 | .067 | .026 | .052 |

Probability of the Type I Error.
(k= 3; alpha= .01)

| $\underline{n}$ | Heterogeneity Index | F | $\varepsilon$ | $\varepsilon'$ | 1,n-1 | $T^2$ |
|---|---|---|---|---|---|---|
| 4 | .99 | .019 | .006 | .019 | .000 | |
| 6 | | .022 | .014 | .022 | .003 | .006 |
| 10 | | .018 | .014 | .018 | .000 | .004 |
| 15 | | .012 | .010 | .012 | .002 | .003 |
| 4 | .74 | .034 | .015 | .009 | .006 | |
| 6 | | .026 | .010 | .005 | .003 | .009 |
| 10 | | .028 | .014 | .005 | .005 | .005 |
| 15 | | .033 | .022 | .008 | .005 | .010 |
| 4 | .54 | .060 | .014 | .008 | .008 | |
| 6 | | .047 | .013 | .008 | .008 | .011 |
| 10 | | .032 | .009 | .009 | .008 | .019 |
| 15 | | .034 | .012 | .011 | .009 | .011 |

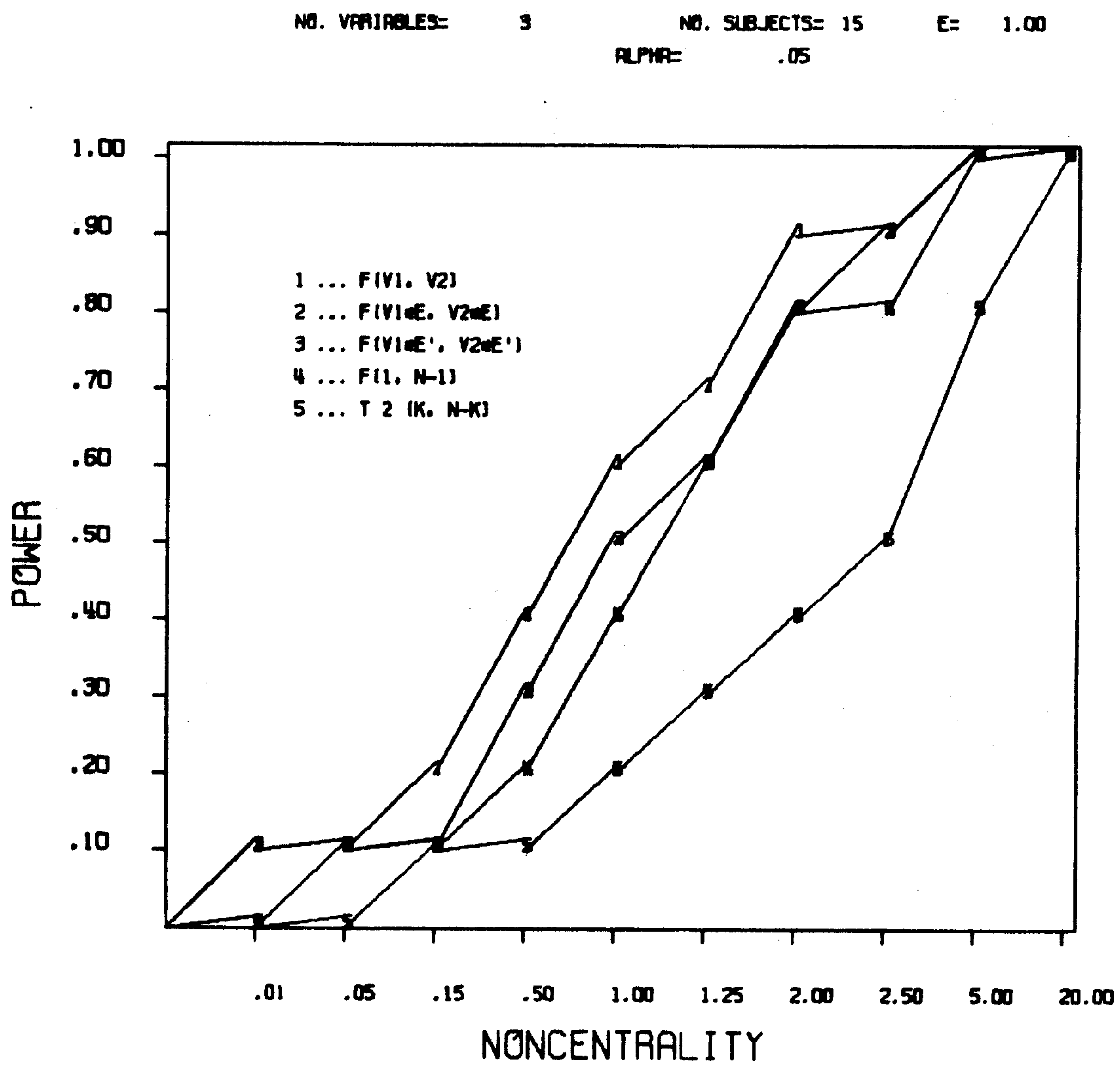
Probability of the Type I Error.
(k= 3; alpha= .05)

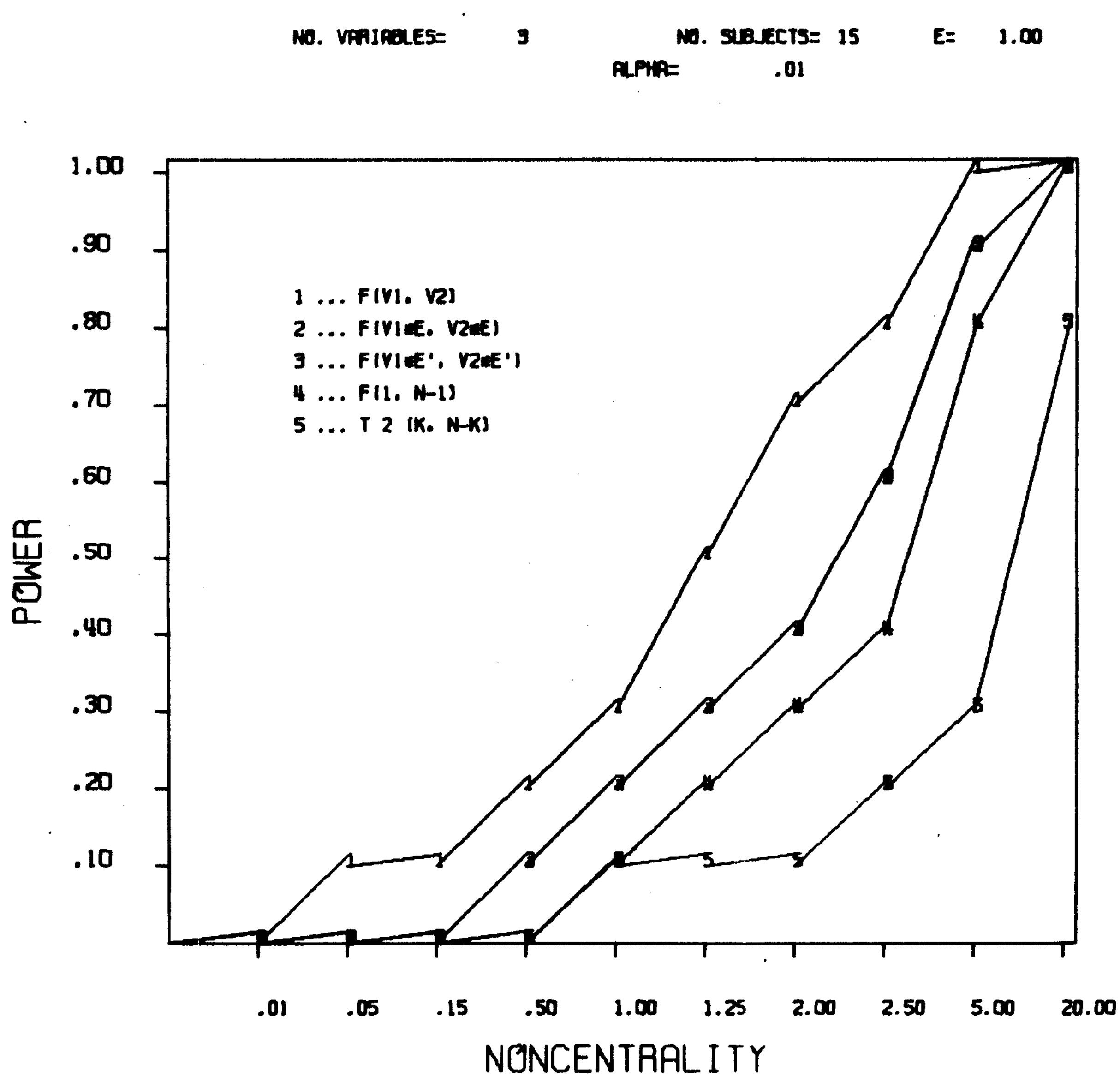
| n | Heterogeneity Index | TEST | | | | |
|---|---|---|---|---|---|---|
| | | F | $\epsilon$ | $\epsilon'$ | 1,n-1 | $T^2$ |
| 4 | .99 | .068 | .043 | .068 | .023 | |
| 6 | | .089 | .065 | .089 | .029 | .023 |
| 10 | | .072 | .058 | .072 | .031 | .030 |
| 15 | | .062 | .057 | .062 | .027 | .037 |
| 4 | .74 | .117 | .071 | .063 | .038 | |
| 6 | | .085 | .056 | .050 | .032 | .042 |
| 10 | | .093 | .067 | .051 | .045 | .037 |
| 15 | | .084 | .064 | .053 | .045 | .035 |
| 4 | .54 | .153 | .079 | .068 | .068 | |
| 6 | | .100 | .063 | .058 | .058 | .067 |
| 10 | | .100 | .061 | .061 | .058 | .077 |
| 15 | | .089 | .050 | .049 | .048 | .061 |

# APPENDIX II

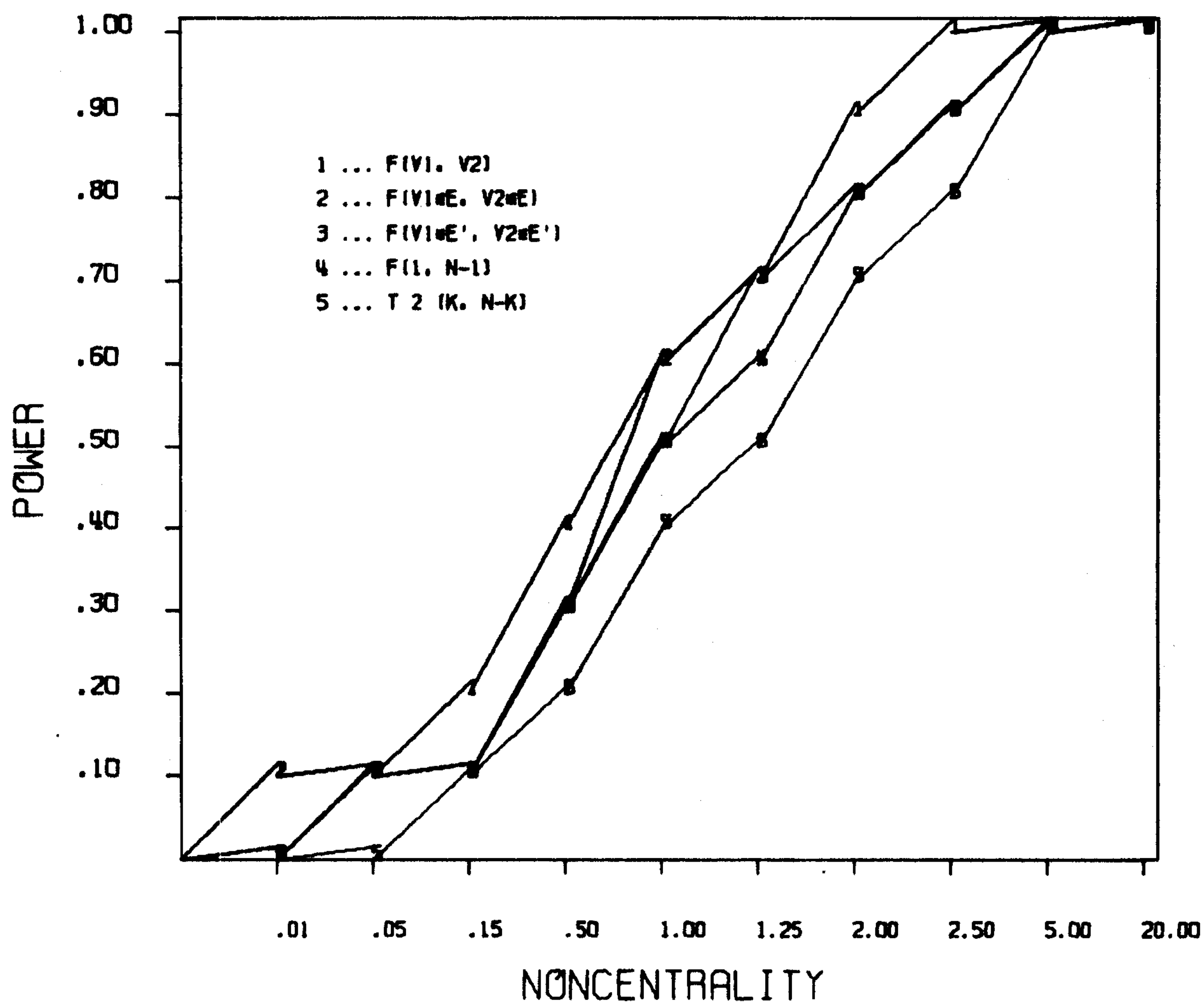
## SELECTED POWER CURVES

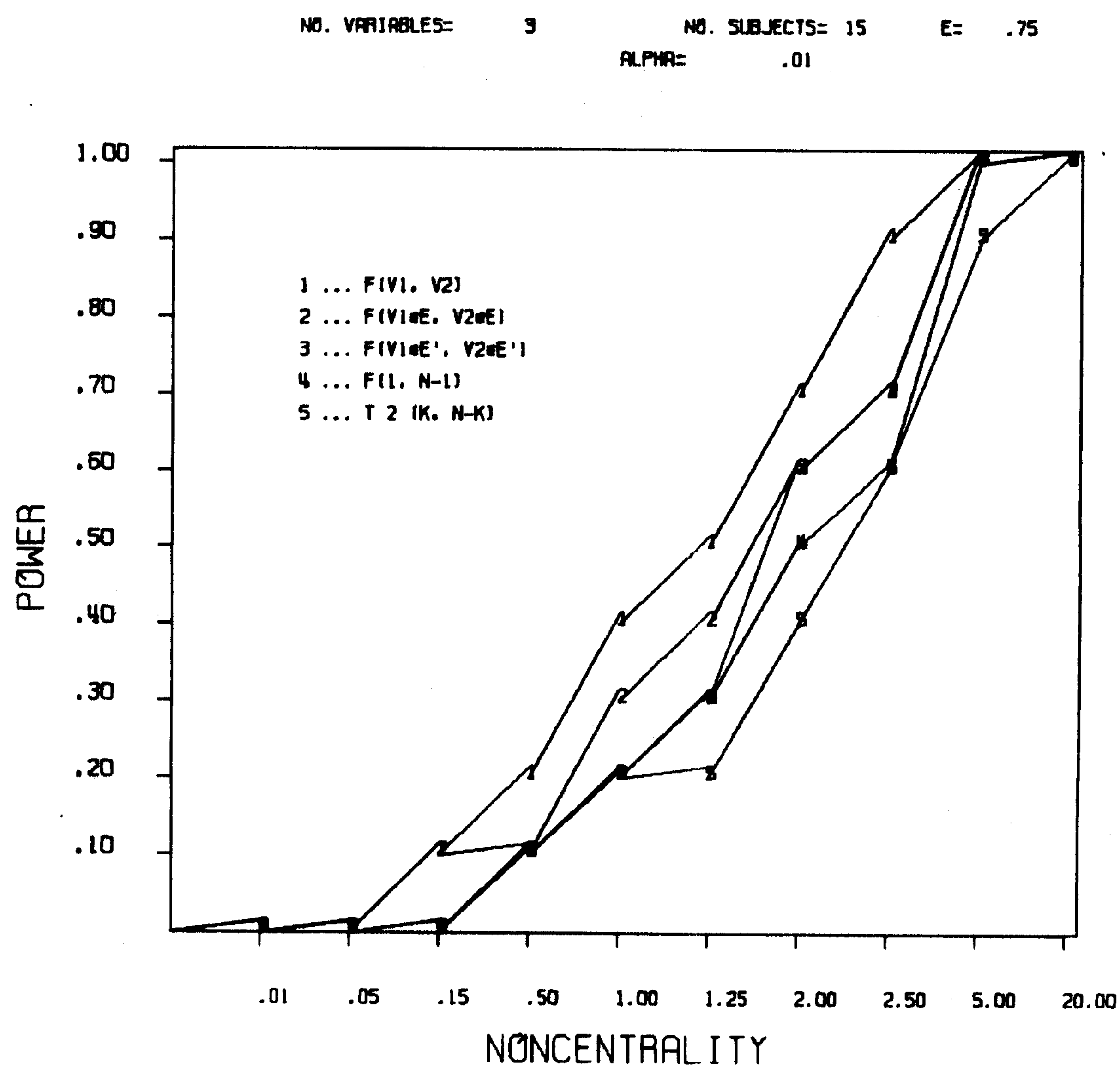
Twelve power curves are presented in this section which are concidered to be representative of the power functions obtained in this study.


NO. VARIABLES= 3
NO. SUBJECTS= 15
E= 1.00
ALPHA= .05
1.00
.90
.80
.70
.60
.50
.40
.30
.20
.10
POWER
1 ... F(V1, V2)
2 ... F(V1*E, V2*E)
3 ... F(V1*E', V2*E')
4 ... F(1, N-1)
5 ... T 2 (K, N-K)
.01
.05
.15
.50
1.00
1.25
2.00
2.50
5.00
20.00
NONCENTRALITY

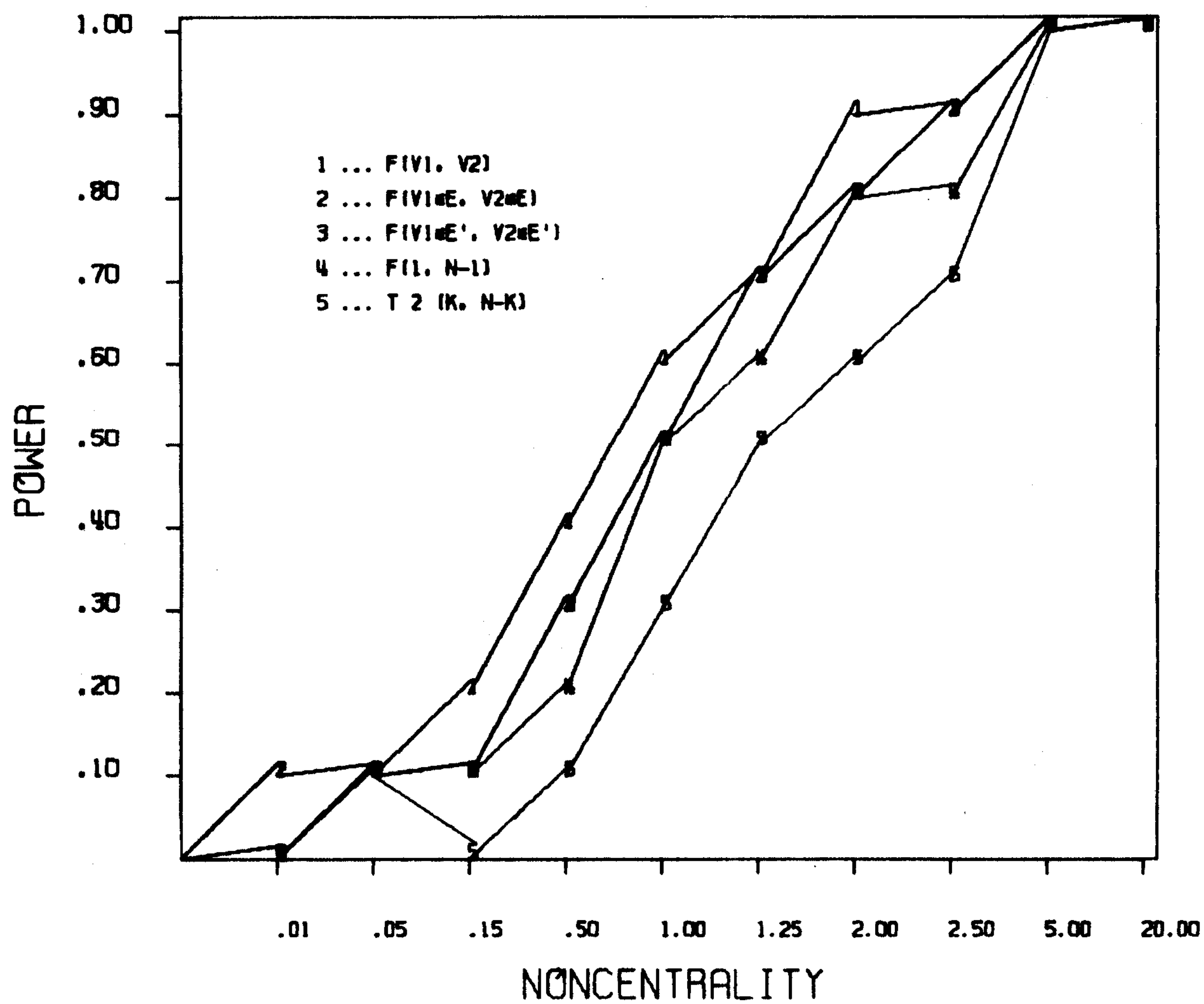

NO. VARIABLES= 3
NO. SUBJECTS= 15
E= 1.00
ALPHA= .01
1 ... F(V1, V2)
2 ... F(V1*E, V2*E)
3 ... F(V1*E', V2*E')
4 ... F(1, N-1)
5 ... T 2 (K, N-K)
POWER
1.00
.90
.80
.70
.60
.50
.40
.30
.20
.10
.01
.05
.15
.50
1.00
1.25
2.00
2.50
5.00
20.00
NONCENTRALITY

NO. VARIABLES= 3 NO. SUBJECTS= 15 E= .75
ALPHA= .05



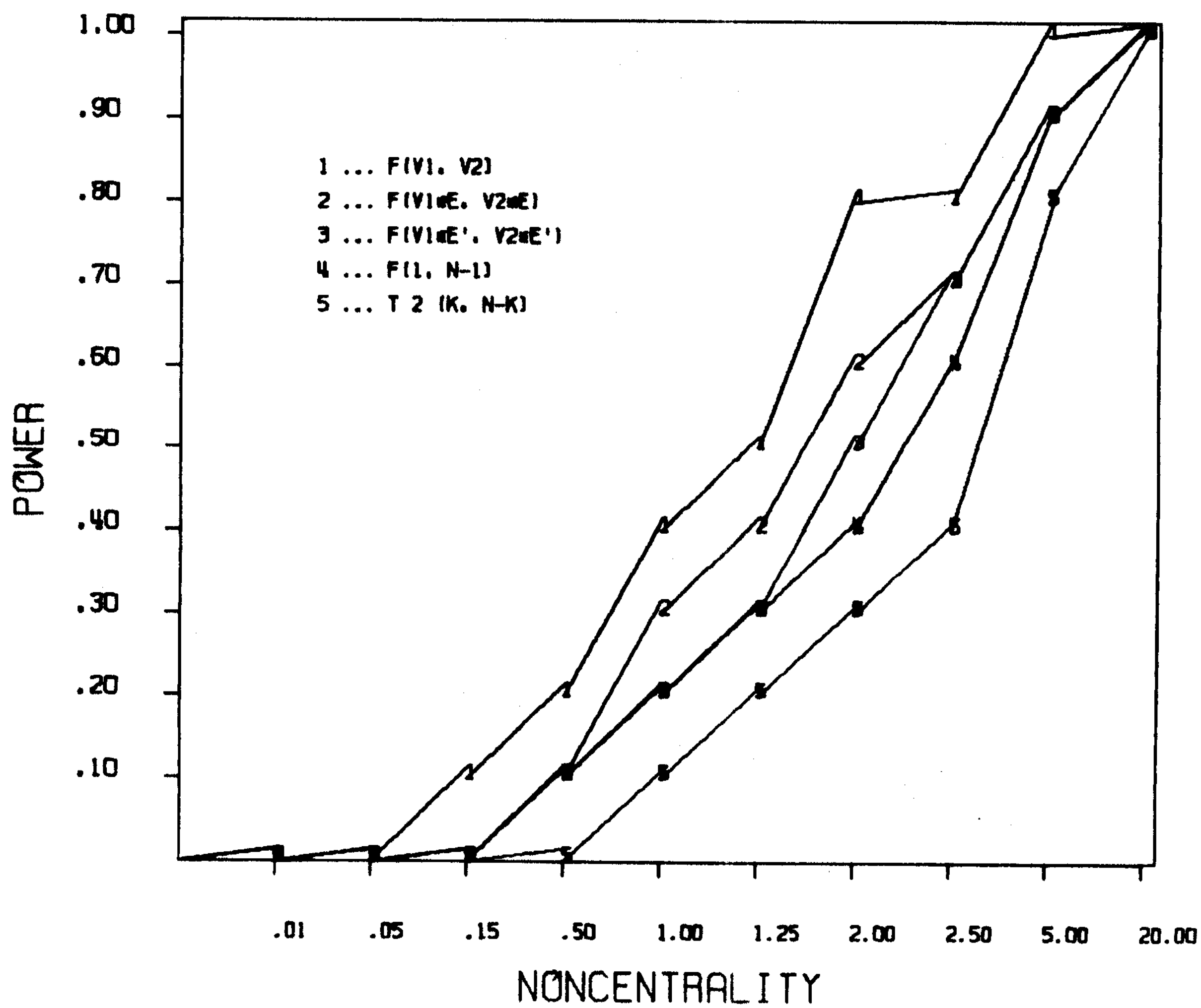

NO. VARIABLES= 3
NO. SUBJECTS= 15
E= .75
ALPHA= .01
1 ... F(V1, V2)
2 ... F(V1*E, V2*E)
3 ... F(V1*E', V2*E')
4 ... F(1, N-1)
5 ... T 2 (K, N-K)
POWER
1.00
.90
.80
.70
.60
.50
.40
.30
.20
.10
.01
.05
.15
.50
1.00
1.25
2.00
2.50
5.00
20.00
NONCENTRALITY

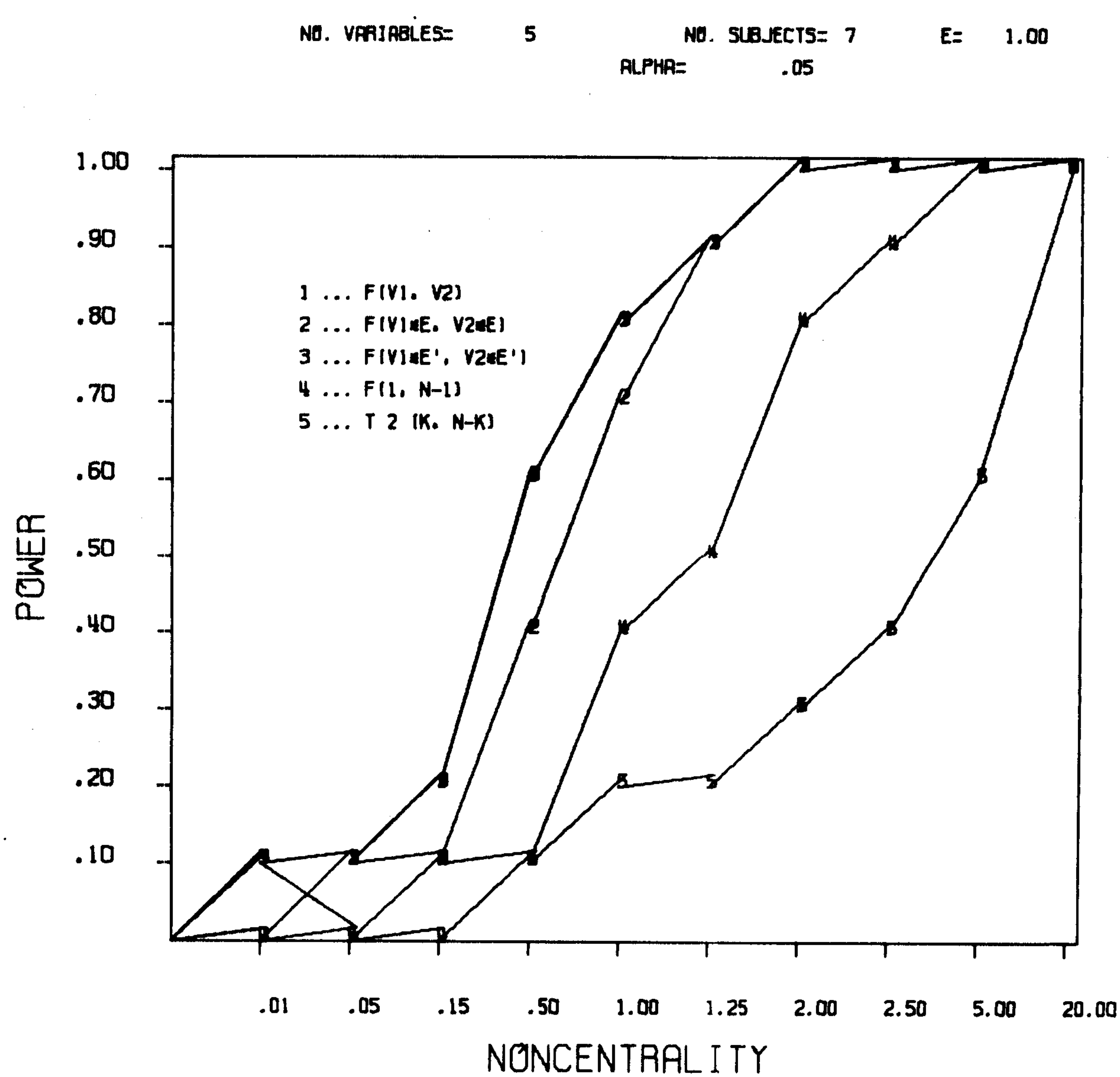
NO. VARIABLES= 3 NO. SUBJECTS= 15 E= .50
ALPHA= .05



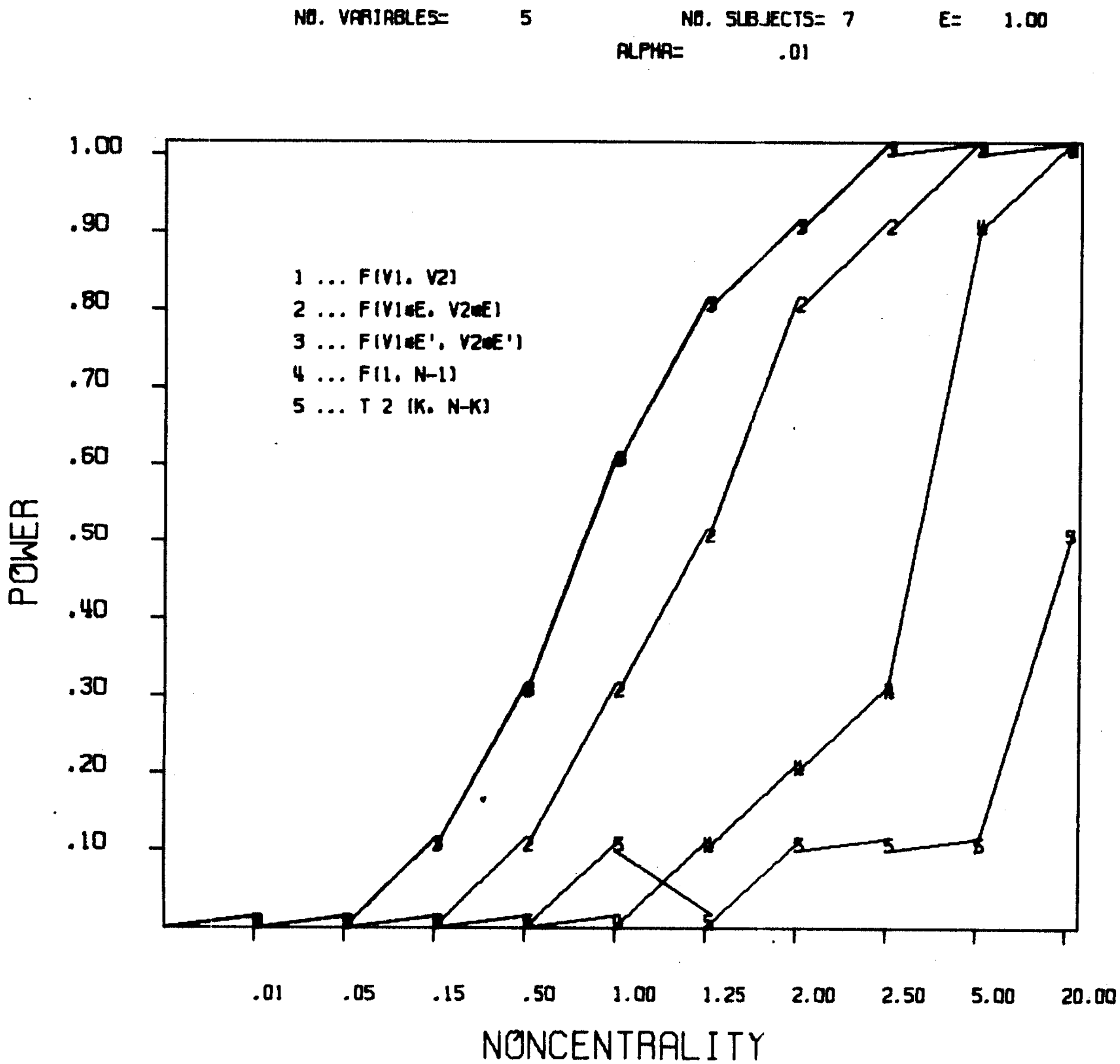
NO. VARIABLES= 3 NO. SUBJECTS= 15 E= .50

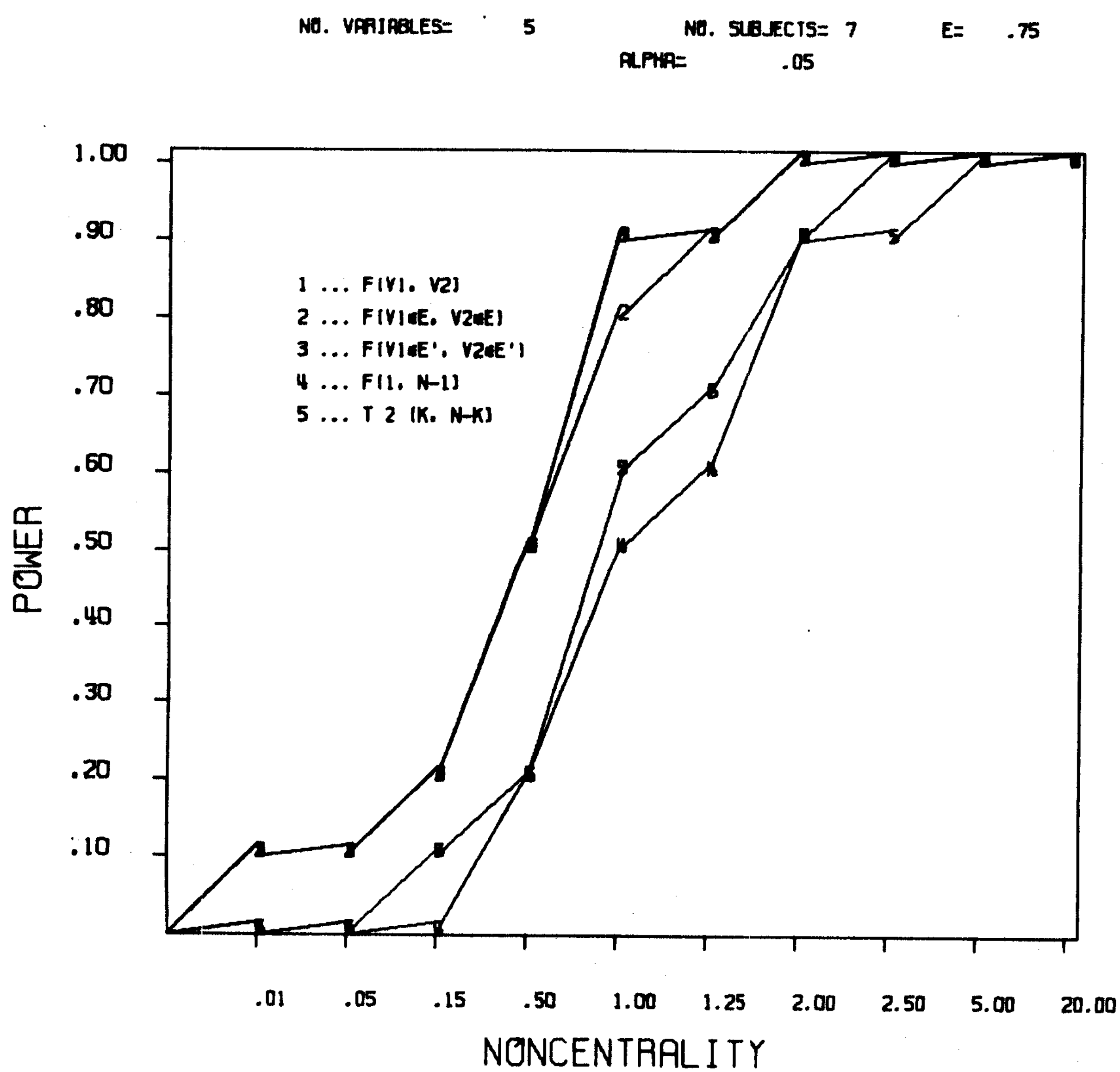
ALPHA= .01

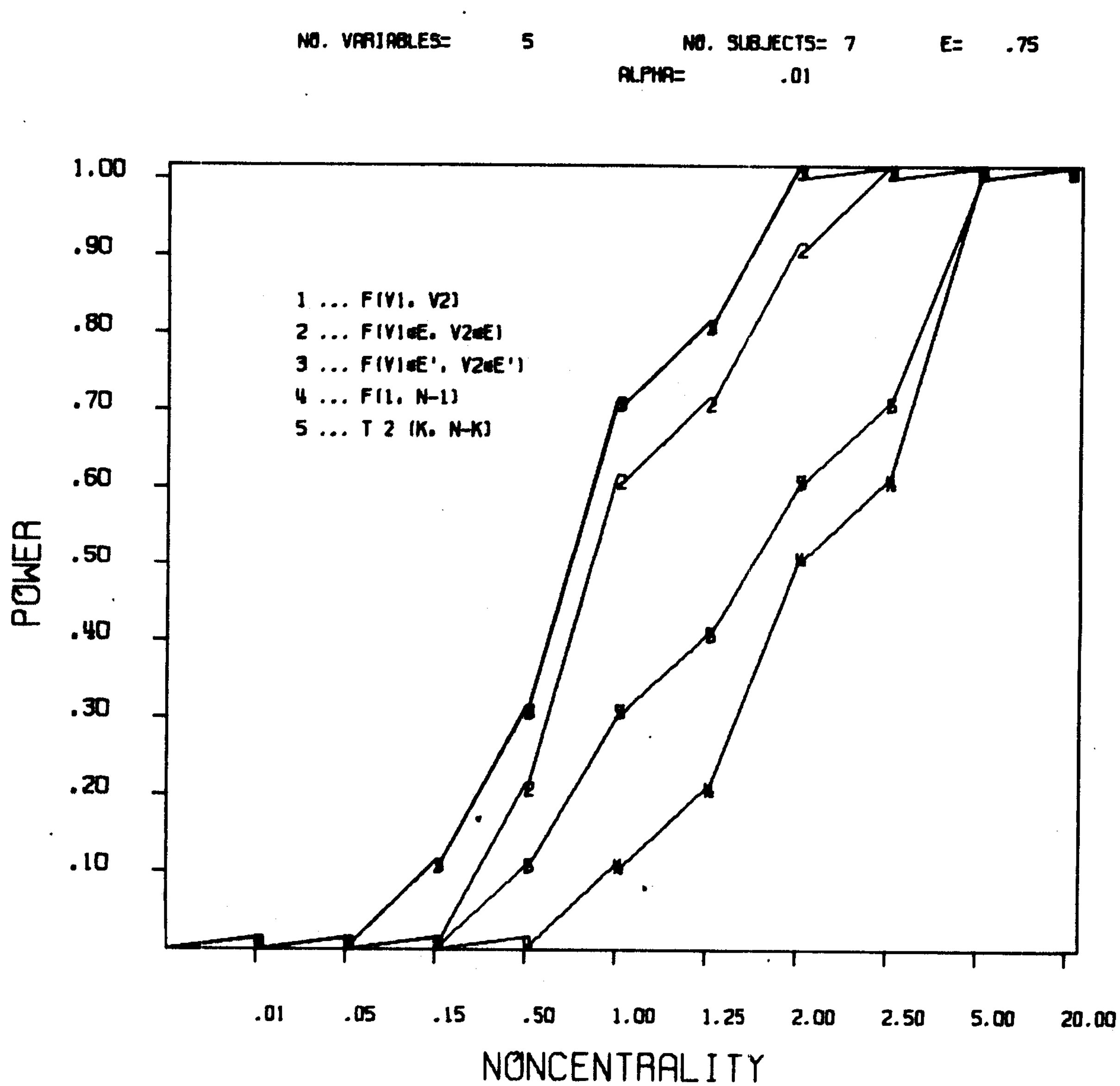




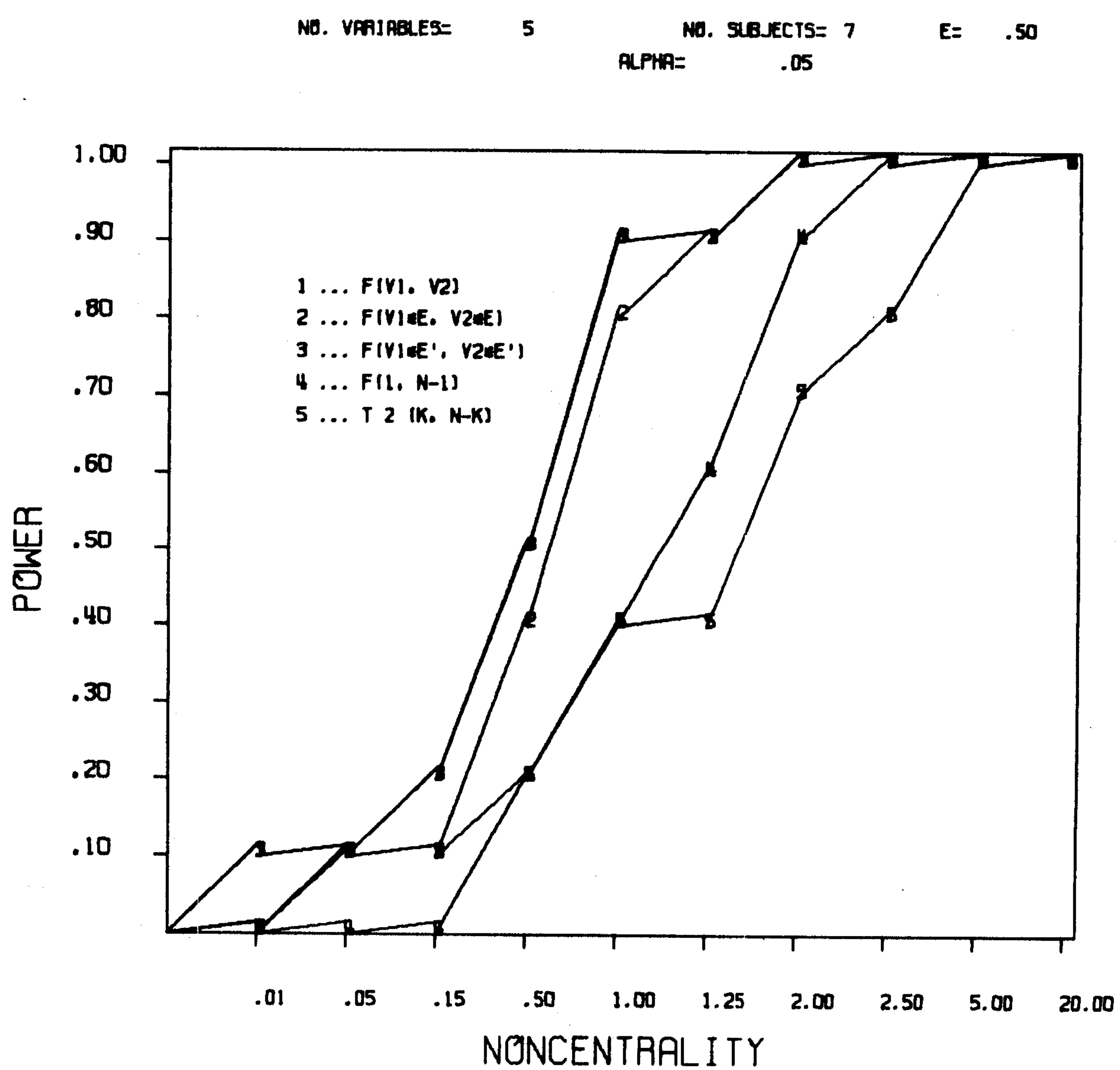
NØ. VARIABLES= 5
NØ. SUBJECTS= 7
E= 1.00
ALPHA= .05
1.00
.90
.80
.70
.60
.50
.40
.30
.20
.10
POWER
1 ... F(V1, V2)
2 ... F(V1*E, V2*E)
3 ... F(V1*E', V2*E')
4 ... F(1, N-1)
5 ... T 2 (K, N-K)
.01
.05
.15
.50
1.00
1.25
2.00
2.50
5.00
20.00
NØNCENTRALITY


NO. VARIABLES= 5
NO. SUBJECTS= 7
E= 1.00
ALPHA= .01
1 ... F(V1, V2)
2 ... F(V1*E, V2*E)
3 ... F(V1*E', V2*E')
4 ... F(1, N-1)
5 ... T 2 (K, N-K)
POWER
1.00
.90
.80
.70
.60
.50
.40
.30
.20
.10
.01
.05
.15
.50
1.00
1.25
2.00
2.50
5.00
20.00
NONCENTRALITY

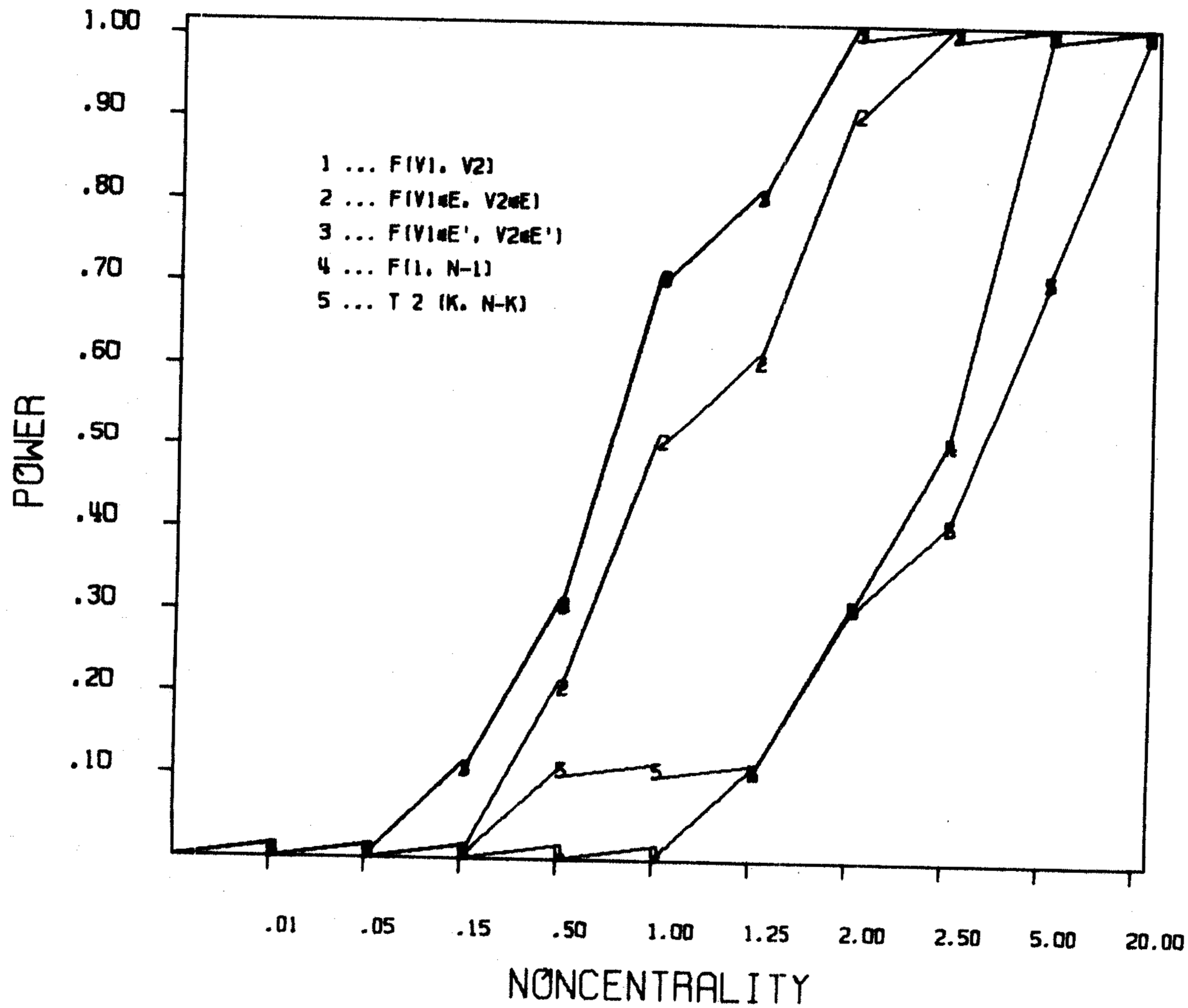

NO. VARIABLES= 5
NO. SUBJECTS= 7
E= .75
ALPHA= .05
1.00
.90
.80
.70
.60
.50
.40
.30
.20
.10
POWER
1 ... F(V1, V2)
2 ... F(V1*E, V2*E)
3 ... F(V1*E', V2*E')
4 ... F(1, N-1)
5 ... T 2 (K, N-K)
.01
.05
.15
.50
1.00
1.25
2.00
2.50
5.00
20.00
NONCENTRALITY



NO. VARIABLES= 5
NO. SUBJECTS= 7
E= .75
ALPHA= .01
1 ... F(V1, V2)
2 ... F(V1*E, V2*E)
3 ... F(V1*E', V2*E')
4 ... F(1, N-1)
5 ... T 2 (K, N-K)
POWER
1.00
.90
.80
.70
.60
.50
.40
.30
.20
.10
.01
.05
.15
.50
1.00
1.25
2.00
2.50
5.00
20.00
NONCENTRALITY


NO. VARIABLES= 5
NO. SUBJECTS= 7
E= .50
ALPHA= .05
1 ... F(V1, V2)
2 ... F(V1*E, V2*E)
3 ... F(V1*E', V2*E')
4 ... F(1, N-1)
5 ... T 2 (K, N-K)
POWER
1.00
.90
.80
.70
.60
.50
.40
.30
.20
.10
.01
.05
.15
.50
1.00
1.25
2.00
2.50
5.00
20.00
NONCENTRALITY

NO. VARIABLES= 5 NO. SUBJECTS= 7 E= .50
ALPHA= .01



## FOOTNOTES

[1]Some authors have refered to this as a "mixed" model because the _n_ dimension, the subjects, is the result of a random sample, but the _k_ dimension, the treatments, is composed of a discrete or finite variable.

[2]Correlation coefficients were transformed by a hyperbolic $\arctan^{-1}$ conversion for the averaging procedure.

[3]The tests of the null hypothesis, as presented in the tables indicated are coded in the following manner: $F_{\nu_1 \nu_2}$ as F; $F_{\nu_1 \epsilon \ \nu_2 \epsilon}$ as $\epsilon$ (sample estimate of Box's statistic); $F_{\nu_1 \epsilon' \ \nu_2 \epsilon'}$ as $\epsilon'$ (parametric value of espilon); $F_{1,\ n-1}$ as 1, n-1 (Geisser & Greenhouse's procedure); and $T^2_{\nu_1 \nu_2}$ as $T^2$.

## REFERENCES


Anderson, T. W. Introduction to Multivariate Statistical Analysis. N.Y.: Wiley, 1960.

Bhat, B. R. On the distribution of various sums of squares in an analysis of variance table for different classifications with correlated and non-homogeneous errors. J. Roy. Statist. Soc. Ser. B., 1960, 21, 114-119.

Bock, D. R. Multivariate anova procedures for repeated measures. In Harris, C. Problems In Measuring Change. Madison: University of Wisconson Press, 1963.

Boneau, C. A. The effects of violations of the assumptions underlying the t test. Psychol. Bull., 1960, 57, 49-63.

Box, G. E. P. Some theorems on quadratic forms applied in the study of analysis of variance problems. I: Effect of inequality of variance in the one-way classification. Ann. Math. Statist., 1954, 25, 290-302. (a)

Box, G. E. P. Some theorems on quadratic forms applied in the study of analysis of variance problems. II: Effects of inequality of variance and correlation between errors in the two-way classification. Ann. Math. Statist., 1954, 25, 484-498. (b)

Box, G. E. P. & Mueller, M. E. Randomization and least squares estimates. J. Am. Statist. Assoc., 1959, 54, 489-502.

Danford, M. B., Hughes, H. M., & McNee, R. C. On the analysis of repeated-measurements experiments. Biometrics, 1960, 16, 547-565.

Dixon, W. J. & Massey, F. J. Introduction to Statistical Analysis. N. Y.: McGraw-Hill, 1957.

Gaito, J. & Wiley, D. Univariate anova procedures in the measurement of change. In Harris, C. Problems in Measuring Change. Madison: University of Wisconson Press, 1963.

Geisser, S. & Greenhouse, S. S. Extension of Box's results on the use of the F distribution in multivariate analysis. Ann. Math. Statist., 1958, 29, 885-891.

Geisser, S. & Greenhouse, S. S. On methods in the analysis of profile data. Psychometrika, 1959, 24, 95-112.

Jones, M. B. Molar correlational analysis. U. S. Naval School of Aviation Medicine, U. S. Naval Aviation Medical Center, Pensacola, Fla., Monograph No. 4, 1960.

Kogan, L. S. Analysis of variance -- repeated measurements. Psychol. Bull., 1948, 45, 131-143.

Lana, R. E. & Lubin, A. The effect of correlation on the repeated measures design. Educat. and Psychol. Meas., 1963, 23, 729-739.

Lubin, A. Statistics. Annual Review of Psychol., 1962, 13, 345-370.

Lindquist, E. F. Design and Analysis of Experiments. Boston: Houghton Mifflin, 1953.

Rao, C. R. Advanced Statistical Methods in Biometric Research. N. Y.: Wiley, 1952.

Scheffe, H. The Analysis of Variance. N. Y.: Wiley, 1959.

Winer, B. J. Statistical Principles in Experimental Design. N. Y.: Mc Graw-Hill, 1962.