

DOCUMENT RESUME

ED 339 741

TM 017 665

AUTHOR Kromrey, Jeffrey D.; Blair, R. Clifford
TITLE Power Properties of Multivariate Permutation Tests
Relative to Hotelling's T-Square Test in Small
Samples.
PUB DATE Nov 91
NOTE 34p.; Paper presented at the Annual Meeting of the
Florida Educational Research Association (Clearwater,
FL, November 13-16, 1991).
PUB TYPE Reports - Evaluative/Feasibility (142) --
Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Educational Research; Equations (Mathematics);
*Hypothesis Testing; *Mathematical Models;
*Multivariate Analysis; *Sample Size; Statistical
Distributions
IDENTIFIERS *Hotellings t; Permutations (Mathematics); *Power
(Statistics)

ABSTRACT

New multivariate permutation tests are proposed that may be effectively substituted for Hotelling's T-Square test in situations commonly arising in educational research. The new tests: (1) are distribution-free; (2) provide tests of directional as well as non-directional hypotheses; (3) may be tailored for sensitivity to specific treatment effects; and (4) may be computed when the number of variables is larger than the number of subjects. Comparisons of the power of the permutation tests to that of Hotelling's test suggest substantial advantages in several situations. Results are interpreted in terms of applications to educational research in which multivariate research questions are posed but the number of units for analysis are small. A 20-item list of references and 8 graphs are included. (SLD)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

ED 339741

**Power Properties of Multivariate Permutation Tests
Relative to Hotelling's T^2 in Small Samples**

Jeffrey D. Kromrey

**Department of Educational Measurement and Research
University of South Florida**

R. Clifford Blair

Department of Pediatrics

and

**Department of Epidemiology and Biostatistics
University of South Florida**

**U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)**

This document has been reproduced as
received from the person or organization
originating it.

Minor changes have been made to improve
reproduction quality

Points of view or opinions stated in this docu-
ment do not necessarily represent official
OERI position or policy

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

JEFFREY D. KROMREY

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

**Paper presented at the annual meeting of the Florida
Educational Research Association, November 13-16, 1991,
Clearwater Beach, FL**

Abstract

New multivariate permutation tests are proposed which may be effectively substituted for Hotelling's T^2 test in situations commonly arising in educational research. The new tests (a) are distribution-free, (b) provide tests of directional as well as nondirectional hypotheses, (c) may be tailored for sensitivity to specific treatment effects, and (d) may be computed when the number of variables is larger than the number of subjects. Comparisons of the power of the permutation tests to that of Hotelling's T^2 suggest substantial advantages in a number of situations. Results are interpreted in terms of applications to educational research in which multivariate research questions are posed but the number of units for analysis are small.

**Power Properties of Multivariate Permutation Tests
Relative to Hotelling's T^2 in Small Samples**

The advantages of multivariate statistical tests, relative to univariate hypothesis tests have been well documented in the methodological literature in education and related fields of study (e.g., Stevens, 1986; Huberty & Morris, 1989; and Ottenbacher, 1989). However, limitations of popular multivariate tests have also been recognized.

Practical problems associated with multivariate tests arise in many research applications when the number of observations is limited. A commonly encountered small sample situation is when classrooms or schools are used as the unit of analysis in applied research or evaluation studies. Issues and strategies related to units of analysis have been described in Blair and Higgins (1986), Hopkins (1982), and Barcikowski (1981). The first problem that arises is the fact that the power of multivariate tests in small sample research is often limited (Stevens, 1980), and in extreme circumstances (i.e., when the number of observations is less than the number of variables) common multivariate test statistics cannot be computed. Secondly, the assumption of multivariate normality, which underlies most multivariate test statistics, is often unjustified with educational data (Micceri, 1989). Although the test statistics may be robust to violations of this assumption, the number of subjects required to be certain of this robustness is of little reassurance to researchers dealing with small samples (Everitt, 1979; Olson, 1974). Thirdly, multivariate procedures are formulated to detect any departures from the null hypothesis, and may therefore lack power to detect

specific departures (Meier, 1975; O'Brien, 1984; and Pocock, Geller & Tsiatis, 1987). Finally, multivariate tests are inherently nondirectional (two-tailed) and do not provide the power advantages obtained through the specification of a directional hypothesis test when the researcher can formulate such a directional hypothesis.

The objectives of this research are to present an alternative statistical methodology to popular multivariate testing procedures and to investigate the power properties of this method relative to a popular multivariate test (the paired samples Hotelling's T^2). The alternative method, based upon permutation tests, has the potential to overcome the limitations described above. Moreover, the general methods described in this research are easily extended to the independent samples Hotelling's T^2 . The remainder of this paper consists of a description of the proposed permutation tests, a presentation of the results of a study designed to compare the power of the new tests to that of Hotelling's test, and a brief consideration of the implications of these tests for educational researchers.

Proposed Tests

The theoretical bases of permutation tests (also known as the method of randomization) were developed by Pitman (1937) and Fisher (1966). Univariate permutation tests are relatively well-known and have been described in detail by Bradley (1968) and Noreen (1989). In contrast, extensions of these procedures to multivariate data analysis have been limited (Boyett & Shuster, 1977).

In general, the sampling distribution of a multivariate permutation test statistic is obtained by computing the desired statistic on all possible permutations of the data

vectors obtained from the units of analysis in the research study. All such permutations are equally likely under the null hypothesis of no treatment effect. The probability, under a true null hypothesis, of obtaining the value of the test statistic calculated from the sample is computed by counting the number of such statistics that exceed or equal that obtained value, and dividing this count by the total number of permutations.

More formally, $x_i = (x_{i1}, \dots, x_{ip})$ and $y_i = (y_{i1}, \dots, y_{ip})$ are p-dimensional vectors denoting observed values from the i^{th} subject under control and treatment conditions, respectively, and $d_i = (x_{i1} - y_{i1}, \dots, x_{ip} - y_{ip})$ denotes the p-dimensional vector of differences that represents change between the treatment and control conditions. Finally, $-d_i$ represents the negative of vector d_i (for example, if $+d_i = (-1, 2, 4)$, then $-d_i = (1, -2, -4)$).

The probability level associated with a test statistic t , based on the permutation principle, is computed as follows. For each of the 2^n possible assignments of + or - to the n vectors d_i , $i = 1, \dots, n$, which are equally likely to occur under the null hypothesis, compute the value of the test statistic. If t_0 is the value of the statistic computed on the original data, and $N(t_0)$ is the number of permutations in which the value of t is greater than or equal to t_0 , then the observed (one-tailed) significance level of the test is

$$p = N(t_0) / 2^n$$

Computations of the 2^n test statistics may be prohibitive with moderate sample sizes and modern computers. For this reason, approximate permutation tests (Edgington, 1969) are used in the power study to be described. The

approximate test differs from the exact test in that rather than computing all possible 2^n test statistics, a large random sample of such statistics are computed. In the power study to be described, 1000 such statistics were computed for each permutation test. The associated probability for the approximate permutation test is computed as

$$p = N(t_0) / M$$

where M is the number of random permutations. The difference between the exact and approximate permutation methods is small when the number of random samples, M , is large. The specific multivariate test statistics examined in this study are described below.

The first statistic, t_{sum} , is defined as

$$t_{sum} = \sum_{j=1}^p t_j$$

where t_j denotes the usual one sample t statistic computed on the j th element of d . This statistic was examined in one-tailed and two-tailed versions that will be referred to as t_{sum1} and t_{sum2} , respectively.

The second test statistic, $t|sum|$, is defined as

$$t|sum| = \sum_{j=1}^p |t_j|$$

where $|t_j|$ denotes the absolute value of the one sample t statistic computed on the j th element of d . In contrast to the t_{sum} statistic, $t|sum|$ yields only a two-tailed test.

The final test statistic proposed, t_{max} , is defined as

$$t_{max} = t_j'$$

where t_j' is equal to the t_j ($j = 1, \dots, p$) that is greatest in absolute value. This statistic was examined in one-tailed and two-tailed versions that will be referred to as t_{max1} and t_{max2} , respectively.

The test statistics described above are designed to be sensitive to different forms of departure from the null hypothesis. Because t_{sum} is the summation of the individual univariate t statistics, it should be most efficient in detecting treatment effects that bring about general increases or decreases across all p variables. Note, however, that t_{sum} would not be sensitive to effects that bring about increases in some variables and decreases in others, because the differences in algebraic signs of the univariate t statistics would tend to cancel. For this type of treatment effect, the test statistic $|t_{sum}|$ should be notably more sensitive. Finally, t_{max} is designed to detect treatment effects that impact only a small subset of dependent variables, such as might be seen when student attitudes are affected by a treatment but student achievement is not affected. The relative success of these strategies is assessed in the sections that follow.

Method

The Monte Carlo study described in this section was designed to compare the power of the five multivariate permutation tests to that of Hotelling's T^2 under four treatment effect models. Data were generated by sampling from a multivariate normal distribution with correlations

between any two variables j and j' given by

$$r_{jj'} = 1 - (j-j')(1/p) \quad j' = 1, \dots, p; j \geq j'$$

where p represents the number of variables in the data set. In this study, p took the values of 4, 8, 16, 21, 32, and 48. Data were generated to simulate the d_i defined above with n taking the values of 10 or 25. The code for this Monte Carlo study was written in FORTRAN making use of a number of subroutines from the International Mathematical and Statistical Libraries (IMSL, 1987).

Four treatment effect models were examined in the study. In the first treatment model a constant treatment effect ($+.5\sigma$) was added to all variables, where σ represents the standard deviation of the marginal distributions. This simulates an effect in which all dependent variables are increased by the treatment. From the point of view of an ANOVA design, this effect represents a main effect due to treatment.

The second treatment model was obtained by adding $.5\sigma$ to half of the dependent variables and subtracting $.5\sigma$ from the other half. This represents an effect in which some dependent variables are affected by the treatment in a positive direction, while others are affected in a negative direction. For example, a hypothetical treatment may yield an increase in student achievement, but a decrease in student attitudes. From the perspective of ANOVA, this represents a disordinal interaction.

The third treatment model was obtained by adding $.5\sigma$ to one-fifth of the dependent variables and leaving the other four-fifths unchanged. This represents an effect in which only a small proportion of the dependent variables are affected by the treatment. An example in educational

research is a hypothetical treatment that affects only some measures of student achievement (perhaps students' acquisition of basic skills) but does not affect students' higher-order thinking skills or attitudes.

The last treatment model examined was obtained by adding $(j)(.5/p)\sigma$ to the jth dependent variable, with j taking the values 1 to p. This represents an effect in which all of the dependent variables are affected by the treatment, but the magnitude of the effect is variable. In educational research, a hypothetical treatment may strongly affect students' acquisition of basic skills, but affect students' higher-order thinking skills and attitudes to a much lesser extent. From the perspective of ANOVA, this represents an ordinal interaction.

In addition to the four treatment effects studied, a null model was investigated. Because the permutation tests are distribution-free and the assumption of population normality that underlies Hotelling's T^2 test was met, this model served to verify the FORTRAN program used to carry out the simulations.

Simulations were carried out for situations in which the sample sizes were 10 and 25, and the number of dependent variables ranged from 4 to 48. For this study, the sampling distributions of the permutation statistics (and, hence, the decisions to reject or fail to reject the null hypothesis) were based on 1,000 random permutations of each sample. The Type I error and power estimates were based on 5,000 samples from each experimental condition. Two-tailed tests of significance were carried out at .10, .05, and .01 levels for all of the test statistics. In addition, one-tailed tests were conducted at the same levels for the test statistics t_{sum} and t_{max} .

Results

As expected, the Type I error rates (obtained under the null model) were near nominal levels and are, therefore, not shown. Because of similarities in the patterns of results, only power results obtained for $\alpha = .05$ are presented. Figures 1 through 8 show power estimates plotted as a function of the number of dependent variables for the various tests obtained under the four treatment effect models described above, and for sample sizes of 10 and 25.

Figures 1 and 2 show estimates obtained under the first treatment model, in which the treatment effect was constant for all dependent variables. All of the permutation tests were more powerful than Hotelling's T^2 test across all numbers of dependent variables investigated. Note, particularly, that the power of Hotelling's test declines sharply as the number of variables approaches the number of subjects. When the number of variables was 8 in Figure 1 or 21 in Figure 2, the power of Hotelling's test was only slightly above α , demonstrating the near absence of sensitivity to the treatment effect in this condition. Also important is the fact that because $n = 10$ in Figure 1, Hotelling's T^2 could not be computed when the number of variables was greater than 9. The corresponding effect occurs in Figure 2 ($n = 25$), when the number of variables is greater than 24. There is, of course, no such constraint on the permutation tests. In contrast to the pattern seen for Hotelling's T^2 , the permutation tests show relatively stable power across all numbers of variables. This stability, which is seen in all of the figures, is attributable to the fairly constant effect sizes used in modeling alternatives. Finally, as would be expected in this treatment effect, the one-sided permutation tests were more powerful than their

two-sided counterparts.

Figure 3 shows that, for the disordinal interaction type of treatment model, Hotelling's test was more powerful than all of the permutation tests for the four dependent variable analysis, but fell below $t|sum|$, $tmax1$, and $tmax2$ when the number of variables was increased. This pattern is also seen in Figure 4 ($n = 25$) except that T^2 remained more powerful than $tmax1$ at its upper variable limit. Note also in these two figures that $tsum1$ and $tsum2$ have almost no sensitivity to this treatment effect, because of the canceling effect of opposite signs in the univariate t statistics referred to above.

Figure 5 shows the results for the third treatment model, in which only 20% of the dependent variables are affected by the treatment. In this figure, T^2 and $tmax1$ have similar power and are most powerful for the four dependent variable situation. When n is increased to 25 (Figure 6), Hotelling's test is the most powerful in the analyses with 4, 8, and 16 dependent variables. The decline in the power of T^2 at its upper variable limit leaves $tmax1$ and $tmax2$ as the most powerful tests in this condition.

Figures 7 and 8 show that all of the permutation tests are more powerful than Hotelling's T^2 test for all situations examined in this treatment effect (the model analogous to an ordinal interaction in ANOVA). The power differences are especially substantial for the $n = 25$ situation (Figure 8).

Discussion

The multivariate permutation tests described and investigated in this research are not advocated as general substitutes for Hotelling's T^2 . Hotelling's test is familiar

to researchers, is easily calculated with available software, and has power advantages under certain conditions. However, multivariate permutation tests should be considered as valuable procedures that can be employed in situations where the T^2 test is suspect or is not calculable.

Particularly notable is the characteristic decline in the power of T^2 as the number of dependent variables approaches the number of subjects. The implications for researchers using small samples is obvious. Small sample situations that are encountered in education research include those in which the appropriate unit of analysis is the classroom or school, projects in which resources for data collection are limited, and studies of relatively rare populations (such as autistic children or teachers of German).

Also of note is the fact that the proposed permutation tests are distribution-free under the same condition that the Wilcoxon signed-rank test is distribution-free (population symmetry about zero). This condition is always met if, under a true null hypothesis, pretest and posttest samples are drawn from a common population (Bradley, 1968). The distribution-free property is especially important in small sample situations, where the reliance on the central limit theorem is questionable.

Finally, the permutation tests are constructed to be especially sensitive to specific types of treatment effects. For research situations in which the nature of the expected effects can be specified a priori, this aspect of the multivariate permutation tests provides a surprising level of statistical power with small samples and large numbers of dependent variables.

Educational researchers face a variety of constraints in the conduct of empirical investigations (e.g., obtaining

the cooperation of research subjects and appropriate authorities, applying experimental treatments consistently and for a sufficient duration to obtain reliable outcome estimates). Statistical constraints are recognized as vitally important in careful research design, because many conclusions drawn from the research results are based upon the outcomes of appropriate hypothesis tests.

Researchers should choose research questions and variables for investigation on the basis of substantive theory and not on the basis of constraints imposed by statistical models. The method of permutation tests is proposed as a feasible alternative to common multivariate statistical testing procedures which relaxes some of the statistical constraints faced by researchers.

References

- Barcikowski, R. S. (1981). Statistical power with group mean as the unit of analysis. Journal of Educational Statistics, 6, 267-285.
- Blair, R. C. & Higgins, J. J. (1986). Comment on "Statistical power with group mean as the unit of analysis." Journal of Educational Statistics, 11, 161-169.
- Boyett, J. M. & Shuster, J. J. (1977). Nonparametric one-sided tests in multivariate analysis with medical applications. Journal of the American Statistical Association, 72, 665-668.
- Bradley, J. V. (1968). Distribution-free Statistical Tests. Englewood Cliffs, NJ: Prentice-Hall.
- Edgington, E. S. (1969). Approximate randomization tests. Journal of Psychology, 72, 143-149.
- Everitt, B. S. (1979). A Monte Carlo investigation of the robustness of Hotelling's one- and two-sample T^2 tests. Journal of the American Statistical Association, 74, 48-51.
- Fisher, R. A. (1966). The Design of Experiments (8th ed.). New York: Hafner.
- Hopkins, K. D. (1982). The unit of analysis: Group means versus individual observations. American Educational Research Journal, 19, 5-18.

Huberty, C. J. & Morris, J. D. (1989). Multivariate analysis versus multiple univariate analyses. Psychological Bulletin, 105, 302-308.

International Mathematical and Statistical Libraries (1987). IMSL Version 1.0 Math/Library Fortran Subroutines for Mathematical Applications. Houston, TX: Author.

Meier, P. (1975). Statistics and medical experimentation. Biometrics, 31, 511-529.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. Psychological Bulletin, 105, 156-166.

Noreen, E. W. (1989). Computer Intensive Methods for Testing Hypotheses. New York: John Wiley & Sons.

O'Brien, P. C. (1984). Procedures for comparing samples with multiple endpoints. Biometrics, 40, 1079-1087.

Olson, C. L. (1974). Comparative robustness of six tests in multivariate analysis of variance. Journal of the American Statistical Association, 69, 894-908.

Ottenbacher, K. J. (1989). Multiple testing of dependent data educational research: A quantitative analysis. Journal of Research and Development in Education, 22, 43-46.

Pitman, E. J. G. (1937). Significance tests which may be applied to samples from any population I and II. Journal of the Royal Statistical Society, Supplement 4, 119-130, 225-232.

Pocock, S. J., Geller, N. L., & Tsiatis, A. A. (1987). The analysis of multiple endpoints in clinical trials. Biometrics, 31, 511-529.

Stevens, J. P. (1980). Power of the multivariate analysis of variance tests. Psychological Bulletin, 88, 728-737.

Stevens, J. (1986). Applied Multivariate Statistics for the Social Sciences. Hillsdale, NJ: Lawrence Erlbaum.

Figure 1
Power of Permutation Tests and Hotelling's T-square Test

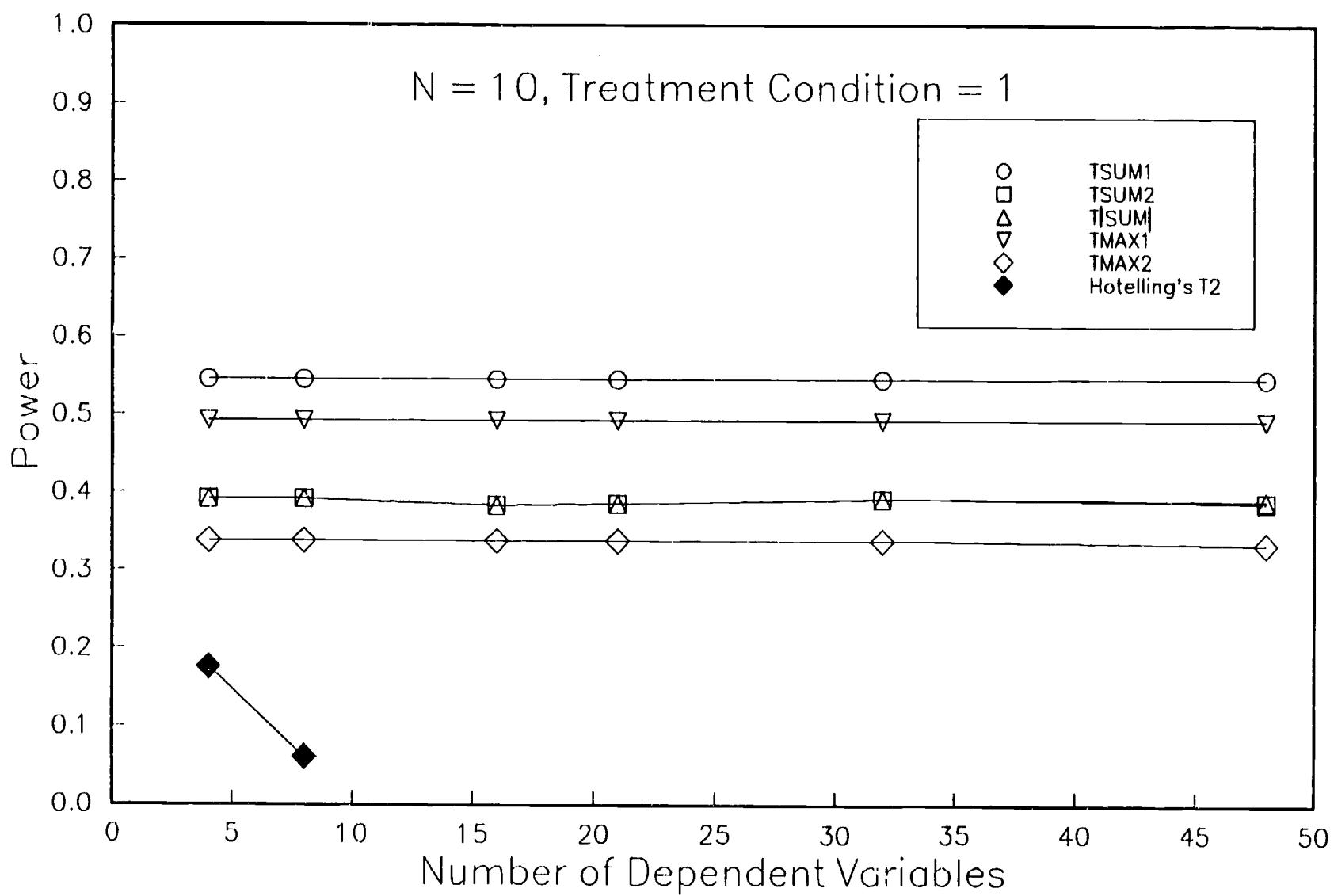


Figure 2
Power of Permutation Tests and Hotelling's T-square Test

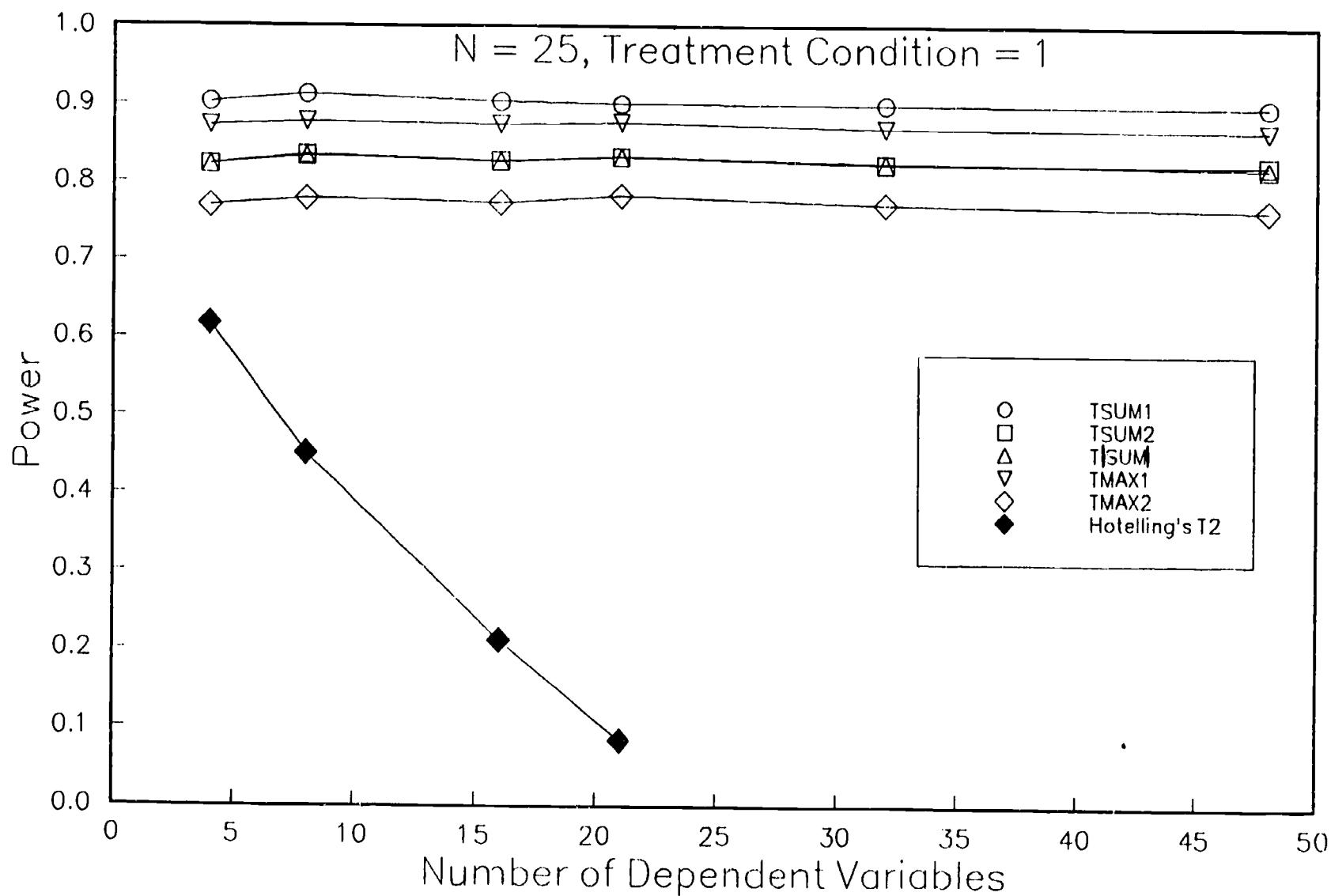


Figure 3
Power of Permutation Tests and Hotelling's T-square Test

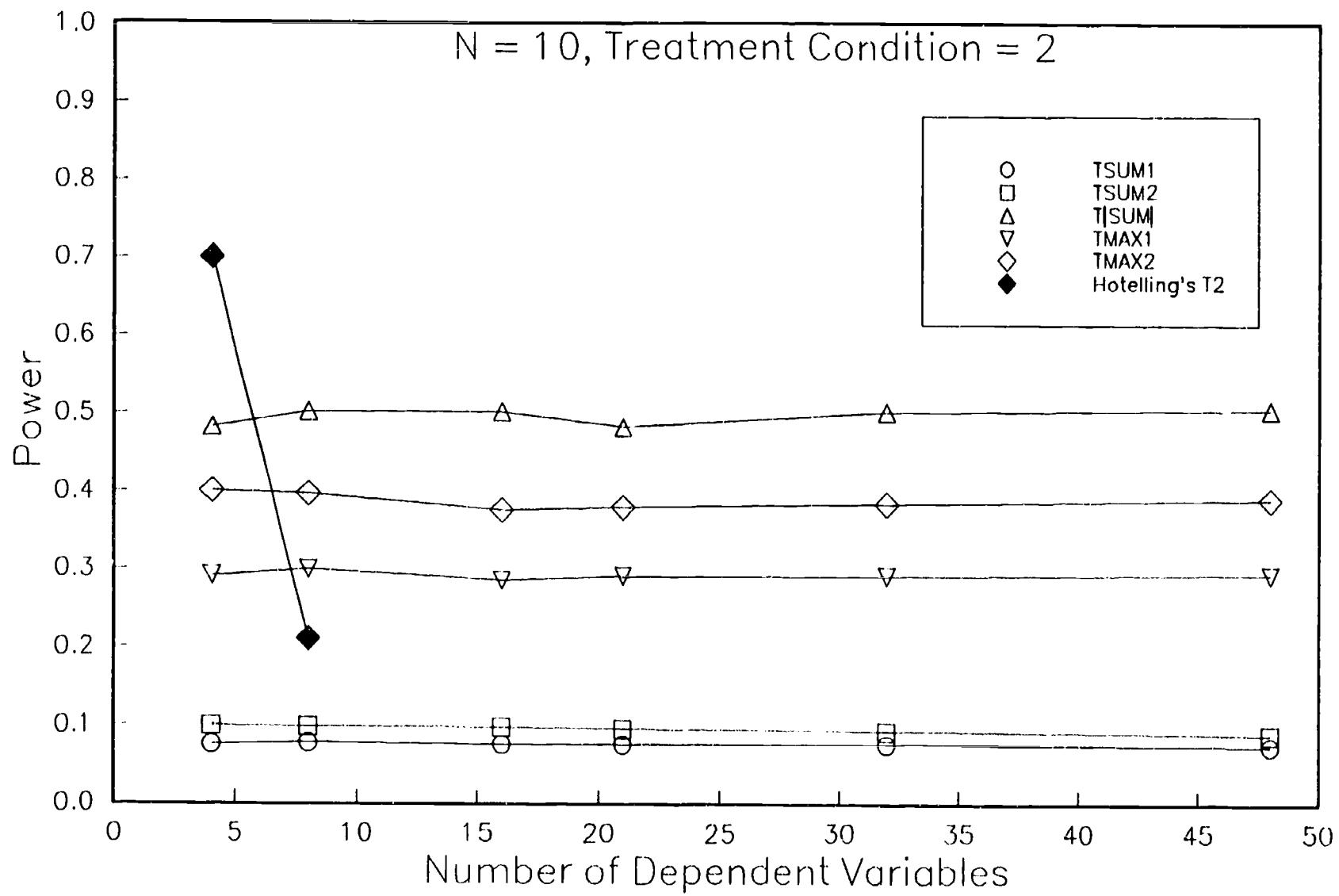


Figure 4
Power of Permutation Tests and Hotelling's T-square Test

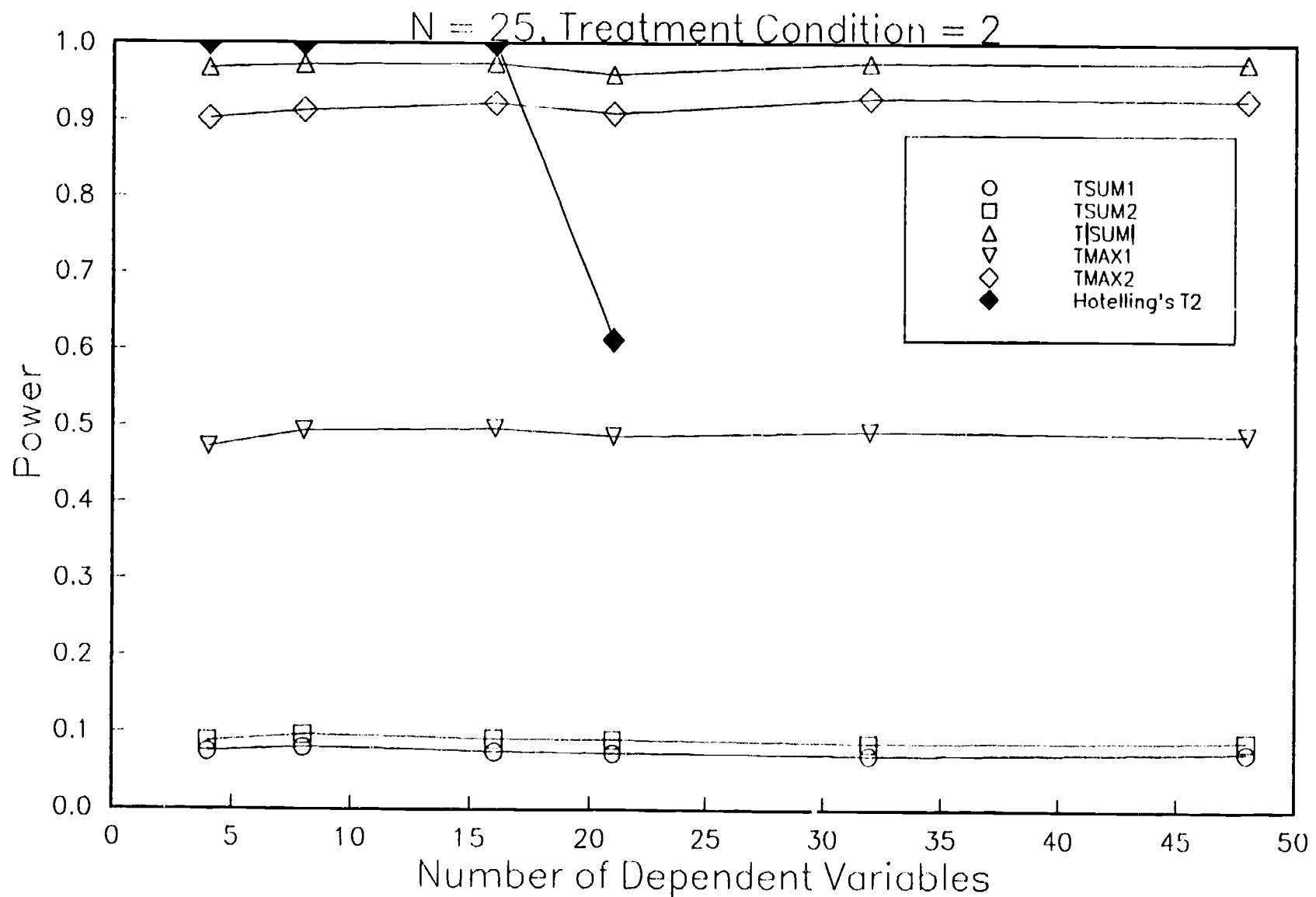


Figure 5
Power of Permutation Tests and Hotelling's T-square Test

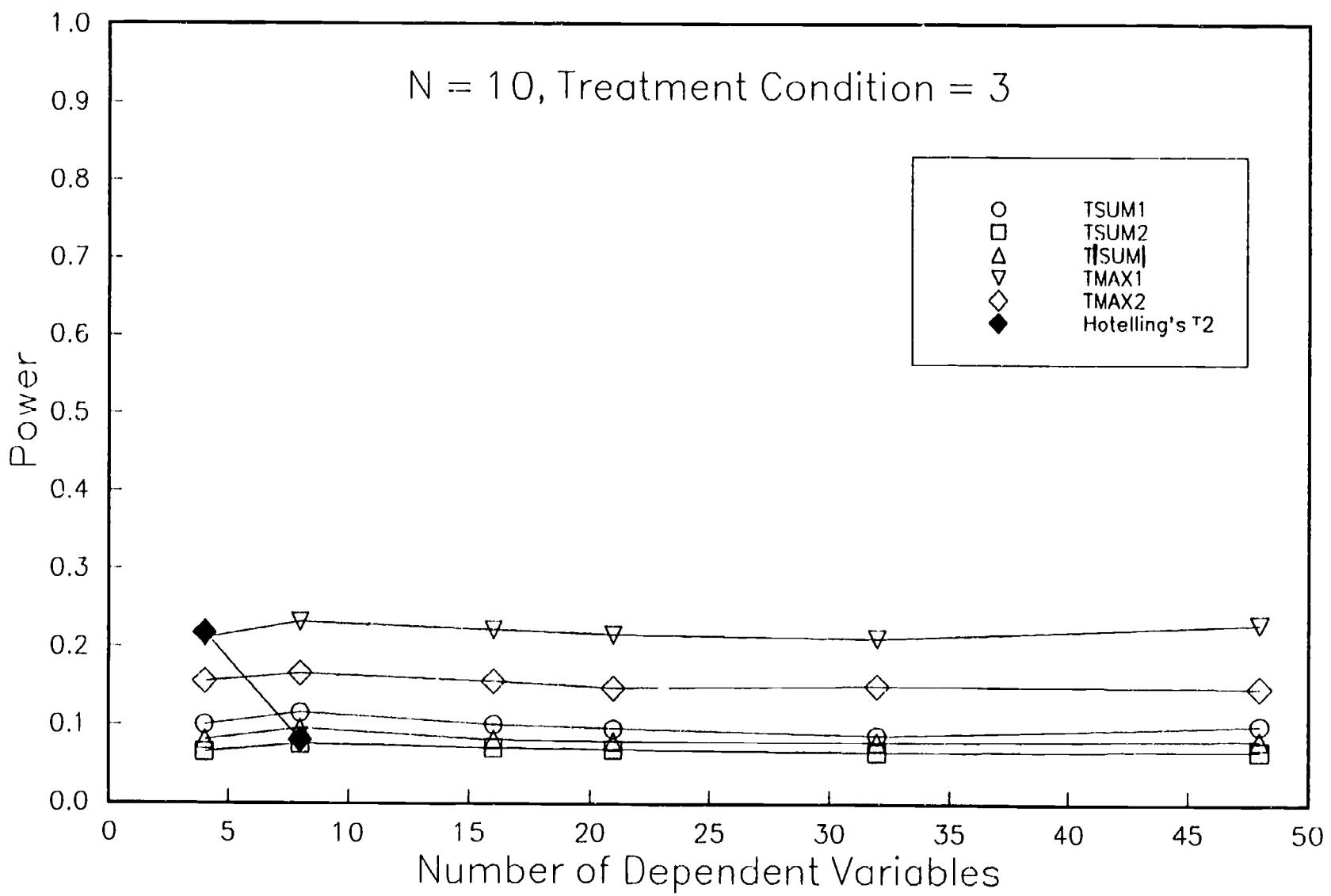


Figure 6
Power of Permutation Tests and Hotelling's T-square Test

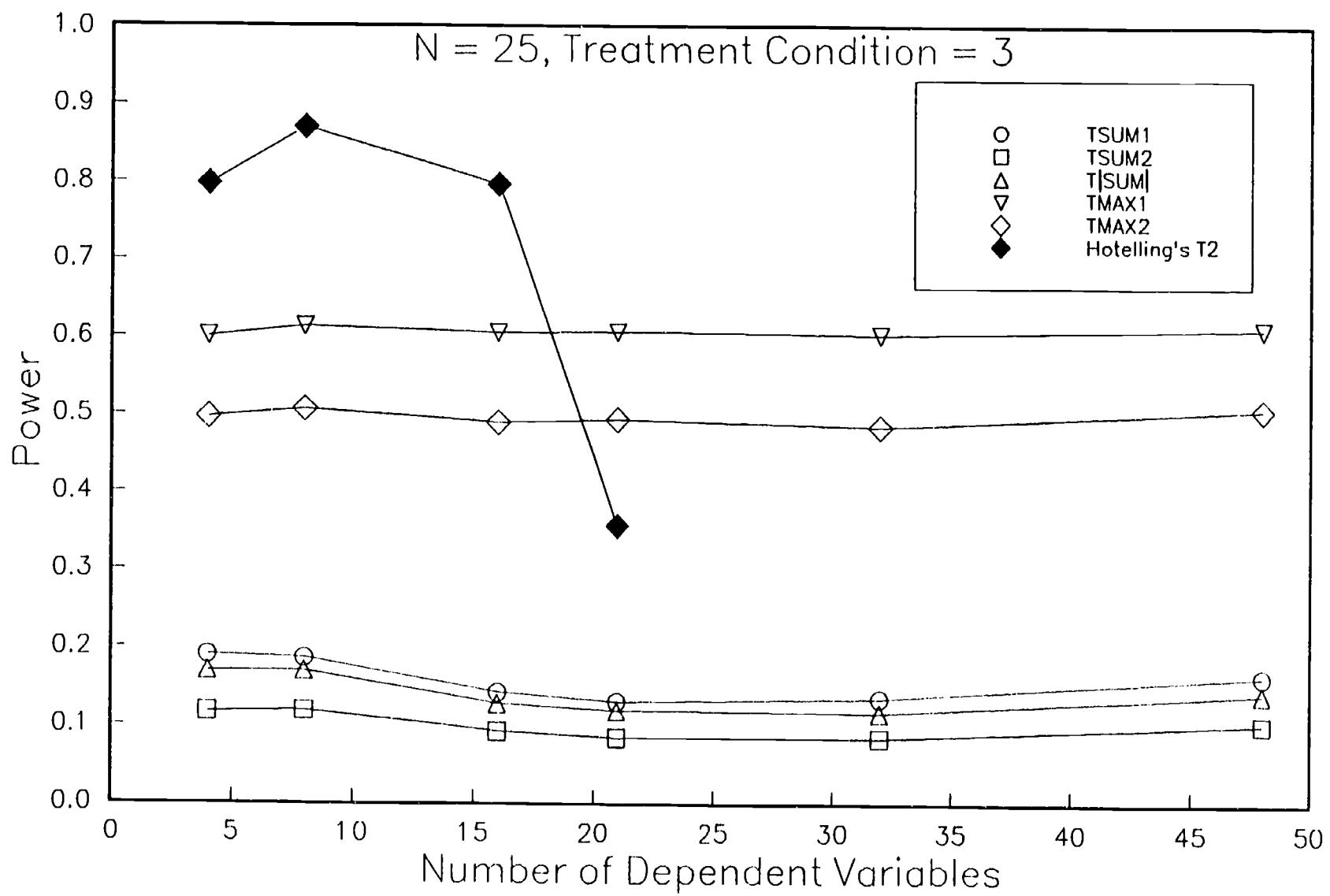


Figure 7
Power of Permutation Tests and Hotelling's T-square Test

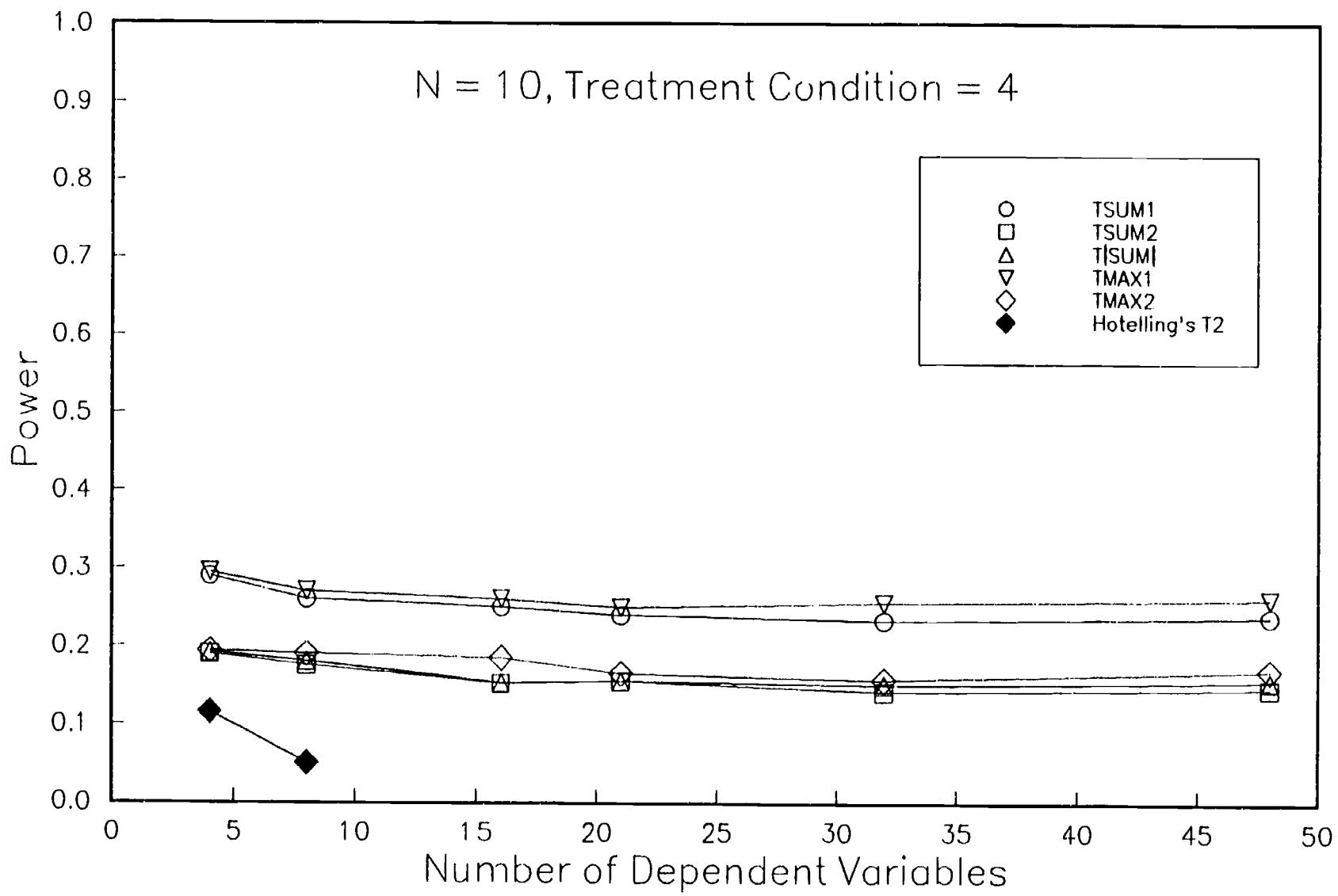
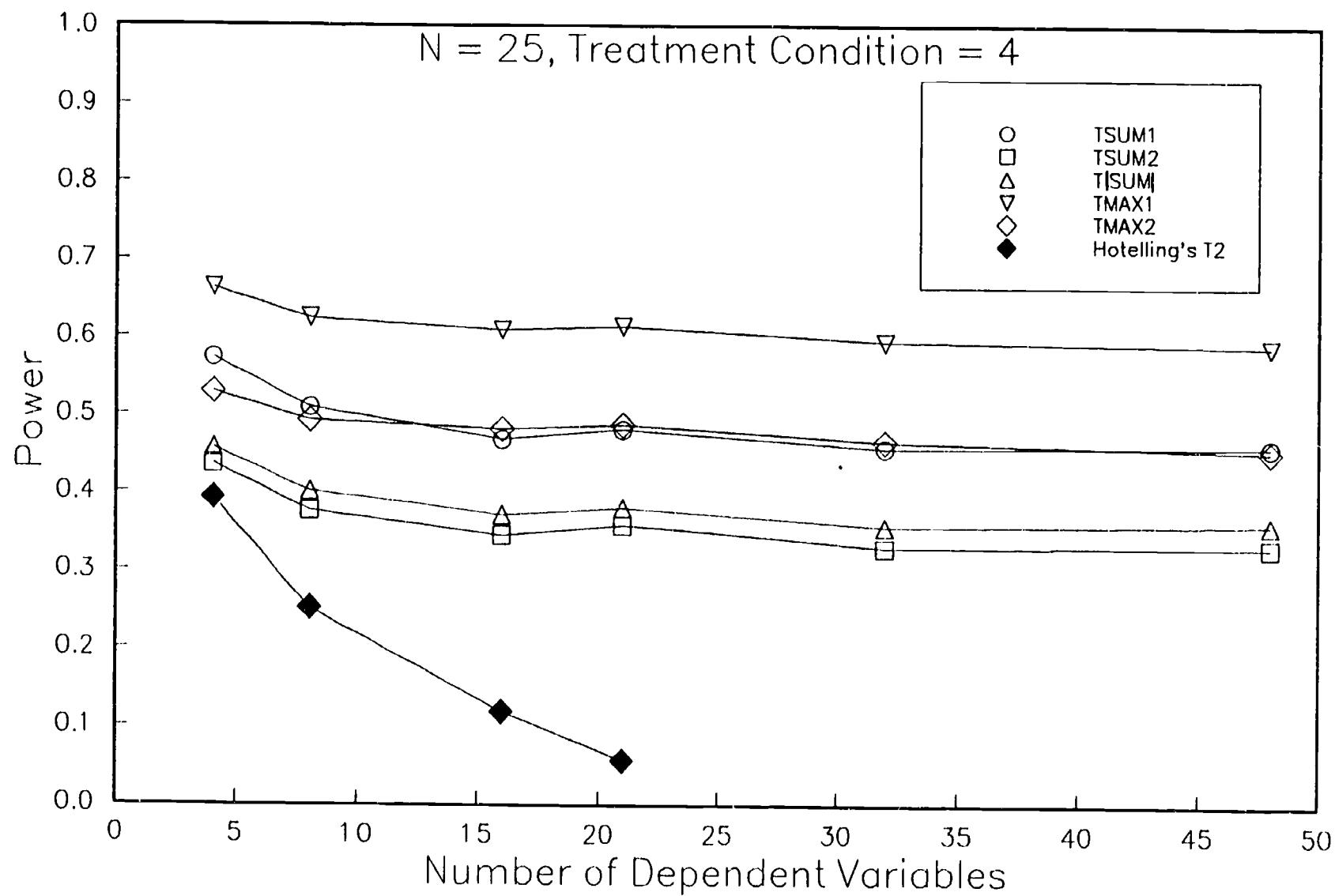


Figure 8
Power of Permutation Tests and Hotelling's T-square Test



Author's Notes

The research reported herein was supported in part by funding for Numerically Intensive Computing, provided by the Central Florida Regional Data Center (CFRDC) at the University of South Florida (USF). However, the opinions expressed are those of the authors and do not reflect the position or policy of the CFRDC or USF.