NASA Contractor Report 165788

METHODS FOR PRESENTATION AND DISPLAY OF MULTIVARIATE DATA

Raymond H. Myers

RAYMOND H. MYERS
206 Fincastle Drive
Blacksburg, Virginia 24060

NASA Purchase Order L-72497A
September 1981

National Aeronautics and
Space Administration

Langley Research Center
Hampton, Virginia 23665

Methods for Presentation and
Display of Multivariate Data

I. Introduction

This report deals with the development of methods for presentation of or display of multivariate data. Multivariate analyses often involve a rather complicated data structure and one often is confronted with the prospect of merely quoting a test statistic and a corresponding significance level as his sole analysis product or output. Multivariate data in which certain factors are varied and several responses are being measured is difficult to interpret but the sheer volume certainly suggests that considerable data display is warranted, not mere data tabulation but certain displays that will aid in the interpretation of the analysis. In this report we attempt to suggest and illustrate various data displays, and we emphasize multivariate analysis of variance problems. Of course, the usual Hotelling's T^2 solution in the two sample case becomes a special case and thus then will receive special emphasis with an illustrative example.

II. The Two Sample Problem

The two sample problem is designed to test

$$H_0: \underline{\mu}_1 = \underline{\mu}_2$$

$$H_1: \underline{\mu}_1 \neq \underline{\mu}_2$$

N82-11833#

where $\underline{\mu}_1$ and $\underline{\mu}_2$ are population mean vectors that represent means of p responses under two conditions. For example in NASA applications, these might be the means of several correlated responses to experiments under two different panel displays or two different conditions of G-seat (on or off). The total data array, of course, involves n_1 p-dimensional vectors under condition 1 and n_2 vectors under condition 2. Mean vectors $\bar{\underline{x}}_1$ and $\bar{\underline{x}}_2$ are computed where

$$\bar{\underline{x}}_1 = [\bar{x}_{11}, \bar{x}_{12}, \ldots, \bar{x}_{1,p}]$$

$$\bar{\underline{x}}_2 = [\bar{x}_{21}, \bar{x}_{22}, \ldots, \bar{x}_{2,p}]$$

where $\bar{x}_{1j}$ represents the average of the n_1 observations under condition 1, response j; $\bar{x}_{2j}$ represents a similar quantity for condition 2. One assumes generally that each of the two multivariate populations has variance-covariance matrix Σ and an empirical estimate is obtained by pooling variance covariance estimates. We shall call this pooled estimate S. ($n_1 + n_2 - 2$ degrees of freedom). The test statistic for testing H_0 is given by

$$T^2 = \frac{n_1 n_2}{(n_1+n_2)} (\bar{\underline{x}}_1 - \bar{\underline{x}}_2)' S^{-1} (\bar{\underline{x}}_1 - \bar{\underline{x}}_2)$$

which is Hotelling's T^2. An F-variate is then used to carry out the mechanics of the test.

(a) Need for Data Display

Generally the user of the Hotelling's T^2 is interested in not only whether the above hypothesis is accepted or rejected but he also is keenly interested in the relative roles of the responses. Tha analyst should also gain some insight concerning the correlation of the responses and what effect this correlation is having on the test procedure. It would also seem reasonable that one should be able to have some light shed on the question of whether or not some of the responses can be totally ignored, i.e., whether or not the total dimensionality of the problem can be substantially reduced.

There are analytic procedures to aid in answering the question mentioned here. However, like the Hotelling's T^2 or F-statistic, they involve the rather unpedagogic computation of certain types of numbers that are somewhat difficult to interpret; however, there is very little in the standard multivariate analysis that leads one to interpretation through plots, pictoral data displays, or informative tables. Here we shall suggest a few procedures and corresponding data displays that will hopefully shed light on interpretation and also complement the conclusion derived from the test statistic. Most of the procedures suggested here center around the procedures of discriminant analysis (stepwise discriminant analysis) and the computation of partial correlation coefficients. Thus we shall proceed to give a brief summary of these two concepts.

(b) Discriminant Analysis

The procedure of Discriminant Analysis is designed to generate the linear combination $\underline{a}'\underline{x}$ which best separates or discriminates between the two conditions or treatments in the two sample problem. The multivariate hypothesis is handled by considering the univariate hypothesis

$$H_0: \underline{a}'\underline{\mu}_1 = \underline{a}'\underline{\mu}_2$$

and a univariate t-test is conducted, and the test is based on the largest $t^2(\underline{a})$. Thus $\underline{a}$ is determined for which the t^2 value

$$t\ (\underline{a}) = \frac{[\underline{a}'(\bar{\underline{x}}_1-\bar{\underline{x}}_2)]\ n_1 n_2}{(n_1+n_2)\underline{a}'S\underline{a}} \qquad (2.1)$$

is maximized. As we outlined in an earlier task, the structure of the $t^2(\underline{a})$ statistic is actually as an F-distributed variate and thus the statistic

$$\frac{(n_1+n_2-p-1)}{(n_1+n_2-2)p} T^2$$

follows F_{p,n_1+n_2-p-1}.

One of the important facets of this analysis is to be able to reduce the dimensionality of the problem by allowing for the elimination of responses that are either redundant or provide little in the way of separation of the two groups. Many statistical packages provide a stepwise or stage wise discriminant analysis that actually allows the user to follow pictorially the reduction in dimensionality. It also allows the user to attain a rather keen insight into not only which responses are relevant, but what is the minimum number of responses one can use to describe the separation between the groups, as well as an indication as to what the correlation structure is among the responses. We shall now proceed to discuss the stepwise algorithm and discuss the "data display" aspect later.

Stepwise discriminant analysis proceeds much like stepwise regression, at least conceptually. One sequentially brings responses into the multivariate model, each time looking at the contribution of each response in terms of its ability to provide separation between the groups or between the treatments. The criterion for including a response is very intriguing.

Suppose, for example, there are four responses and 2 groups (or treatments). The forward stepwise procedure begins by entering the response that provides the largest univariate t separating the two groups. Call this variable x_1. For step 2 the methodology will treat the model as if the new candidate variable is a univariate response and response variable 1 is a covariate in an analysis of covariance with treatment effects representing effects due to the two groups. The response included is one that provides the largest F comparing treatments in this analysis of covariance model which would be written (for response k)

$$y_{ij}^{(k)} = \mu + \tau_i + \beta y_{ij}^{(1)} + \epsilon_{ij}$$

where the superscript denotes the response in question. The response k is chosen which separates the treatments the largest amount, <u>adjusted</u> for the initial response 1. This procedure is continued but each time the responses entered in previous steps become covariates in succeeding steps. In addition, the procedure will continue until the response to be entered at a specific step is not significant at some specified level. In addition, the procedure will eliminate responses that have entered at previous steps if, in light of other responses, they cease to become significant. This would imply that the separation of the two responses provided by that variable is redundant and it is not needed as it was at an earlier stage.

The Backward Elimination Procedure is identical to the forward procedure except the method begins with all responses and eliminates one at a time. This is sometimes preferable to forward stepwise procedures.

III. Possible Data Display

Most data analysts and members of the scientific community can identify with such displays as group means, correlation coefficients, significance levels of tests, etc. Our suggested data displays involve plots and tables that center around these concepts with view toward illustrating the answers to two questions.

(i) Is there an appreciable change in the responses as you go from treatment 1 (say G seat off) to treatment 2 (G seat on)?

(ii) What is the true dimensionality of the problem, i.e., how many responses are truly effected and what responses play the important roles?

(a) Partial Correlation Coefficients

These measures of linear association are quite different from the usual simple correlation coefficients [1] ordinarily observed in a multivariate analysis, in that they measure degree of linear association between two responses, conditional on the others. The interpretation would be that it expresses how much correlation exists between the two responses when all the others are held fixed. This is meant to give, at the outset, some indication to the user which linear associations among the responses might create problems in reducing the dimensionality of the problem.

(b) Plots of Sample Means

In the two sample case it is particularly enlightening and actually requires very little data preparation, to plot the sample means using standardized observations. This is not intended as a direct vehicle for statistical inference but rather, along with the partial correlations, as an initial display. The example in the next section will illustrate this preliminary display.

(c) Plots Resulting from Stepwise Discriminant Analysis

The major data display would be a product of the stepwise discriminant analysis and should illustrate the important responses, the reduction in dimensionality of the problem, as well as the statistical significance associated with difference between the two treatments. Output from the forward stepwise discriminant analysis at each stage includes F-statistics indicating the significance of the incoming response and a corresponding level of significance, and an F-statistic (or Hotelling's T^2) indicating the significance of the difference between the two groups <u>at this stage</u> (degree of separation with the responses present in the current stage). Displays that would be of interest in an illustrative way would be plots showing these significance levels plotted pictorially as a function of the stage of the stepwise procedure, the latter also being the number of responses currently in the multivariate model. With these plots the user (and hence the reader) can see step by step the role of each response as it enters the picture. Displayed will be measures of what responses are critical and at what point do additional responses provide no more separation in the two treatments. This will become clearer with our example in the next section.

IV. Example

The example we use to illustrate the data display features a subset of NASA data in which there are 27 data points in each of two groups (G-seat on and G-seat off) where six responses are being measured. The purpose of the experiment of course is to determine if there is a difference "on the average" between the two groups or treatments, and then to determine if this difference is explained by one, two, three or perhaps more responses, and an indication of what these responses are. Of course, it would be advantageous for the analysis to display illustrations that point out these results. The original data is given in Table I. Table II gives the partial correlation coefficients. Of course, any strong partial correlation would give a clear indication of redundancy between two responses despite the activity of the other responses. Here, of course, the only responses showing a strong partial linear association are responses 2 and 3, while responses 4 and 5 show at least a moderate partial linear association. Incidentally, this partial correlation is taken within the group conditions (i.e., seat off-seat on).

To illustrate the analysis, Figures 1., 2., and 3. should be observed. Figure 1 is a simple plot of the means of the two groups for the six responses, using standardized data. Here, of course, it is clear from the display that variables 2 and 6 supply a goodly portion of the separation between the two seat conditions. The next phase of the data display and analysis deals with illustration of the significance tests in the stepwise procedure. As each variable enters the model, essentially a hypothesis

H_0: variable entered does not provide any increase in separation between the groups

H_1: variable provides an increase in separation

is being tested through the mechanism described earlier. Figure 2 provides a plot of the significance level associated with each variable as it enters sequentially. Small values are evidence in favor of H_1 above. Clearly some subjectivity must be used by the analyst here concerning at what point he must decide that no further responses provide additional separation. Here, it is clear that

(a) Response 2 provides a substantial separation between the two G-seat conditions.

(b) Response 6 significantly increases the separation between the two conditions.

(c) No additional responses provide significant separation beyond these two.

Figure 3 is a bit more difficult to interpret but still provides essential information. The basic analysis at each stage of the stepwise procedure is to conduct the Hotelling's T^2 (F-statistic) to determine if the two seat conditions differ on the average across the stepwise, i.e., the test is of the hypothesis

$$H_0: \underline{\mu}_1 = \underline{\mu}_2$$

$$H_1: \underline{\mu}_1 \neq \underline{\mu}_2$$

as described earlier in this report. Plotted in the figure is the significance level of that test at each stage of the stepwise discriminant analysis. The indication is that at every step the separation between the seat conditions is significant. However, at step 2, with the entry of variable 6 in the presence of 2, the separation is enhanced, whereas in succeeding stages there is an apparent "dilution" of this significance due to the addition of non-discriminating responses. Ideally, one might consider (rather loosely) that the minimum point in the plot would indicate the smallest subset of responses.

The computer software for this work was the BMD package. It provides all the tests described as well as the partial correlation coefficients.

Bibliography

1. Morrison, Donald. *Multivariate Statistical Methods*, McGraw-Hill, 1976.

TABLE I

---------------------------- SEAT=UN ----------------------------

OBS	VAR1	VAR2	VAR3	VAR4	VAR5	VAR6
1	0.1939	0.0029	2.5768	0.1118	0.9962	0.9462
2	0.1671	0.0529	2.1413	0.0820	0.1808	1.3088
3	0.2017	0.0354	1.2471	0.0581	0.2515	1.6986
4	0.1984	0.0605	2.5814	0.0456	1.1566	1.4363
5	0.1780	0.0427	2.5541	-0.0798	0.1912	1.5067
6	0.1677	0.0244	1.2874	0.0142	0.1192	1.5977
7	0.2147	0.0507	2.2030	-0.1493	-0.0170	1.4999
8	0.1970	0.0604	2.7916	-0.0208	0.0364	1.8896
9	0.1768	0.0334	1.5545	0.0584	0.1004	1.9103
10	0.1709	0.0633	2.4836	0.2194	0.6511	0.9350
11	0.1969	0.0578	2.3211	-0.0065	0.2967	1.7287
12	0.1596	0.0366	1.4127	0.0793	0.2079	1.8488
13	0.1994	0.0369	2.4250	0.2047	0.3223	0.9678
14	0.1906	0.0647	2.0130	-0.0120	1.0609	1.8659
15	0.2039	0.0198	1.4273	0.0053	0.1630	1.7675
16	0.2020	0.0585	2.2788	0.0384	0.5486	1.5994
17	0.2193	0.0411	2.4931	-0.0306	0.2889	1.0316
18	0.1994	0.0024	1.0058	0.0214	0.2206	0.8102
19	0.1617	0.0514	2.5384	-0.1532	-0.0084	0.7501
20	0.2164	0.0747	2.0180	0.1375	0.2902	1.2141
21	0.2110	0.0076	1.1183	-0.0465	0.4522	1.9919
22	0.2006	0.0161	2.3327	-0.0352	0.3601	1.3408
23	0.2160	0.0709	2.3508	0.1738	0.5738	1.3765
24	0.1681	0.0130	1.2282	0.0879	0.2958	1.2283
25	0.2018	0.0447	2.5976	0.0983	0.2758	1.2865
26	0.2033	0.0411	2.5953	0.2444	0.7402	1.6271
27	0.1686	0.0243	1.3298	0.0755	1.0853	1.6221

TABLE I. (Continued)

OBS	VAR1	VAR2	VAR3	VAR4	VAR5	VAR6
28	0.1870	0.0456	2.3585	-0.1641	-0.1142	1.3501
29	0.2021	0.0402	2.2250	0.0663	0.1771	1.3796
30	0.1677	0.0084	0.9500	0.0145	0.0657	0.9232
31	0.1631	0.0254	2.2023	0.1513	0.9696	1.0774
32	0.2204	0.0138	1.7848	0.1930	0.4283	0.9713
33	0.1817	0.0107	0.9590	0.0816	0.1265	1.2355
34	0.1602	0.0516	2.3205	0.1518	0.4463	1.2373
35	0.1910	0.0477	2.2924	0.0432	0.2780	1.9285
36	0.1521	0.0327	1.4713	0.0519	0.2333	1.5103
37	0.1690	0.0541	2.2969	-0.0845	0.0752	0.8123
38	0.1857	0.0364	2.1756	0.0716	0.2137	1.9252
39	0.1620	-0.0096	0.8936	-0.0251	0.1577	1.8130
40	0.1992	0.0625	2.2886	-0.1155	0.0412	1.0761
41	0.1728	0.0470	2.0922	0.1607	0.4312	0.9520
42	0.1503	0.0052	1.2119	-0.0731	0.0951	1.6150
43	0.2023	0.0463	2.4906	0.0105	0.0410	0.7504
44	0.2107	0.0031	1.5454	-0.1254	0.0999	1.0211
45	0.1754	-0.0010	0.9477	0.1140	0.4282	0.9390
46	0.2034	0.0620	2.2392	0.1764	1.1067	0.6755
47	0.1765	0.0302	2.3312	0.1904	0.2525	1.2577
48	0.2001	0.0167	1.3027	-0.0068	0.1615	0.7190
49	0.1900	0.0414	2.5904	-0.0119	0.6371	0.7103
50	0.1833	0.0321	2.4759	-0.0985	0.4426	1.6393
51	0.1499	0.0048	1.0838	0.0236	0.1024	1.1201
52	0.2322	0.0181	1.6858	-0.2412	0.2816	1.3172
53	0.1794	0.0206	2.4308	0.0515	0.4937	1.7535
54	0.2440	0.0208	1.4450	0.0769	0.1836	1.3087

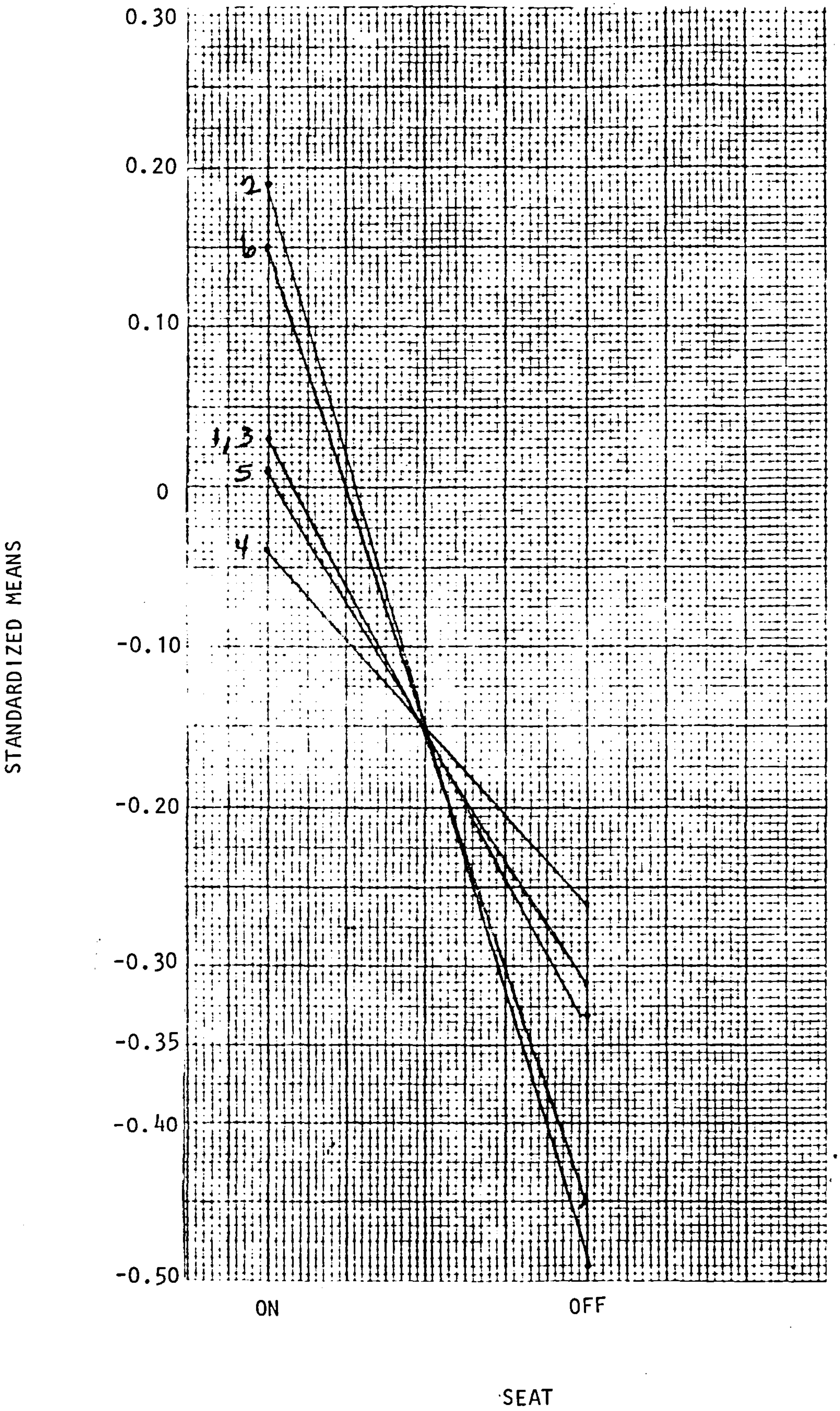
TABLE II

PARTIAL CORRELATIONS OF DEPENDENT VARIABLES REMOVING

LINEAR EFFECTS OF INDEPENDENT VARIABLES

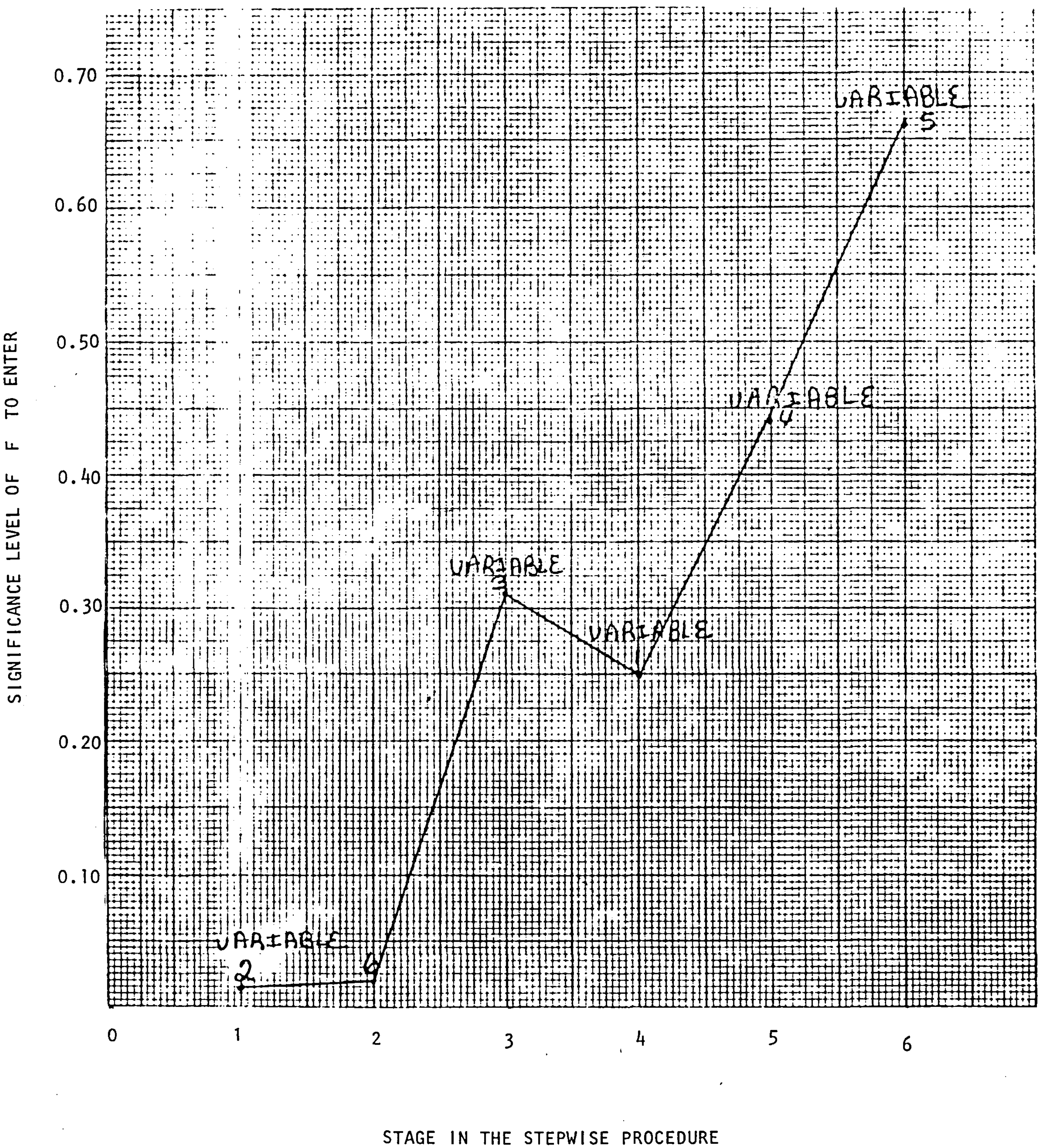
	1	2	3	4	5	6
1	1.000					
2	0.165	1.000				
3	0.264	0.790	1.000			
4	-0.092	0.136	0.029	1.000		
5	0.062	0.256	0.246	0.454	1.000	
6	-0.075	-0.132	-0.104	-0.161	-0.105	1.000

Figure 1



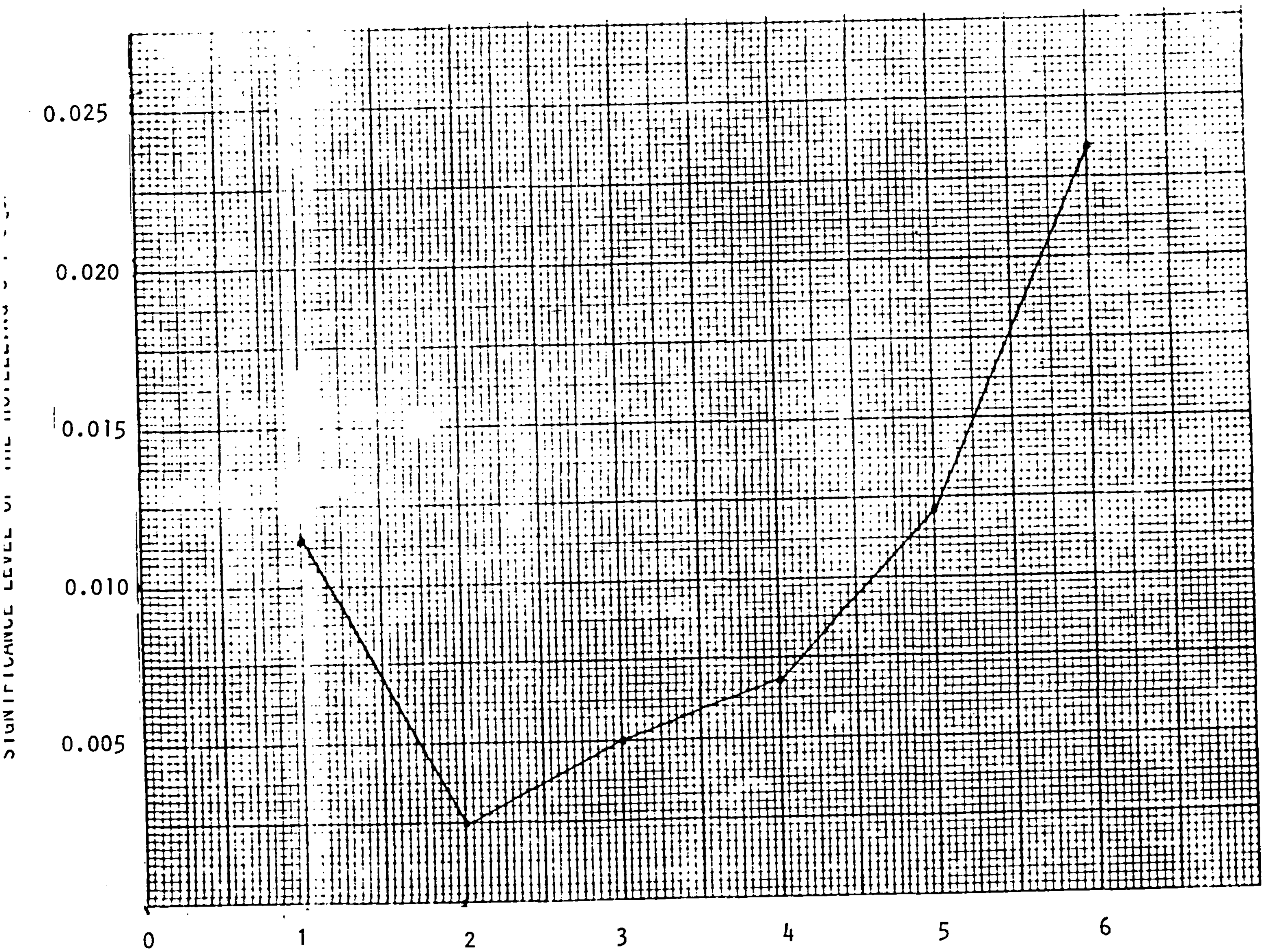
0.30
0.20
0.10
0
-0.10
-0.20
-0.30
-0.35
-0.40
-0.50
STANDARDIZED MEANS
2
6
1,3
5
4
ON
OFF
SEAT

Figure 2



SIGNIFICANCE LEVEL OF F TO ENTER
0.70
0.60
0.50
0.40
0.30
0.20
0.10
VARIABLE 5
VARIABLE 4
VARIABLE 3
VARIABLE 1
VARIABLE 2
6
0
1
2
3
4
5
6
STAGE IN THE STEPWISE PROCEDURE

Figure 3



1. Report No. NASA CR-165788	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Methods for Presentation and Display of Multivariate Data		5. Report Date September 1981
		6. Performing Organization Code
7. Author(s) Raymond H. Myers		8. Performing Organization Report No.
9. Performing Organization Name and Address Raymond H. Myers 206 Fincastle Drive Blacksburg, VA 24060		10. Work Unit No.
		11. Contract or Grant No. L-72497A
12. Sponsoring Agency Name and Address National Aeronautics and Space Adminstration Washington, DC 20546		13. Type of Report and Period Covered Contractor Report
		14. Sponsoring Agency Code

15. Supplementary Notes

Langley Technical Monitor: Burnell T. McKissick
Final Report

16. Abstract

This report deals with the development of methods for presentation of or display of multivariate data. Multivariate analyses often involve a rather complicated data structure and one often is confronted with the prospect of merely quoting a test statistic and a corresponding significance level as his sole analysis product or output. Multivariate data in which certain factors are varied and several responses are being measured is difficult to interpret, but the sheer volume certainly suggests that considerable data display is warranted, not mere data tabulation, but certain displays that will aid in the interpretation of the analysis. In this report, we attempt to suggest and illustrate various data displays, and we emphasize multivariate analysis of variance problems. Of course, the usual Hotelling's T^2 solution in the two sample case becomes a special case and thus then will receive special emphasis with an illustrative example.

17. Key Words (Suggested by Author(s))	18. Distribution Statement
Data displays Multivariate data	Unclassified - Unlimited Subject Category 65

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	17	A02

For sale by the National Technical Information Service, Springfield, Virginia 22161