DEPARTURE FROM ASSUMPTIONS OF ANALYSIS OF VARIANCE AND BEHAVIOUR OF MULTIPLE COMPARISON PROCEDURES (POWER AND ROBUSTNESS) IN REAL DATA

by

IVANKA BRANISLAV KRSTIĆ

B.S., Mechanical Engineering,
University of Belgrade, Yugoslavia, 1954

M.S., Mechanical Engineering,
University of California, Berkeley, 1972

A MASTER'S THESIS

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1977

Approved by:

K. E. Kemp
Major Professor

TABLE OF CONTENTS

DEPARTURE FROM ASSUMPTIONS OF ANALYSIS OF VARIANCE AND BEHAVIOUR OF MULTIPLE COMPARISON PROCEDURES (POWER AND ROBUSTNESS) IN REAL DATA

Introduction

The main objective of this work is by using real data:

1. To determine which kinds of non-normality are most common as well as to what extent heterogeneity of variance is present and whether departure from assumptions differs between disciplines.
2. To investigate whether performance of selected multiple comparison tests in real application is similar to results based on theoretical considerations and simulation results.
3. To examine whether or not departure from assumptions affected the performance of selected multiple comparison procedures.

The Statistical Laboratory at Kansas State University provides statistical consulting to several departments in the Kansas Agriculture Experimental Station at Manhattan, Kansas. As a result of this work there is easy access to many different data sets processed by Statistical Laboratory statistical programs. Over a period of two years, September 1974 through July 1976, data was collected from each data set processed by the Statistical Laboratory analysis of variance program AARDVARK (Kemp, 1976). In total 1765 different data sets from several departments were used.

Four multiple comparison procedures, because of their different properties concerning protection against Type I error and power, were considered: Fisher's LSD, Duncan's New Multiple Range Test, Tukey's HSD (honestly significant differences) for 20 or fewer levels and Waller-Duncan

Bayes t procedure for more than two levels of main effects only.

Each data set analyzed by AARDVARK was subjected to three basic tests for normality: Shapiro-Wilk W-test for samples of size 50 or less, skewness for samples of size 25 or more and kurtosis for samples of size 11 or more. In addition, the simple correlation between the subclass mean and variance was computed to determine if the cell means and variances were independent as they should be if assumptions hold.
To test homogeneity of variance Bartlett's test was applied.

Anticipated significance of this study:

1. To obtain information about the type and frequency of most common departure from assumptions in real data which should instigate statisticians to develop more robust tests to that particular kind of departure from assumptions. At the same time this information should warn researchers in various fields of agriculture, biology, social science, etc. to be more selective concerning the tests used in the analysis of their data.
2. Several valuable simulation studies concerning the performance of multiple comparison tests are available. However, the magnitude of true differences and the level of homogeneity among the true treatment means depend on the structure of real data and cannot be fully predicted by a simulation study. Therefore, our study should be useful in showing to what extent the generalization of simulation results can be accepted.
3. Finally, on the basis of this study while selecting a multiple comparison procedure, in addition to the criteria concerning the protection against Type I error and power, robustness of the particular procedure should also be considered by researchers.

DEPARTURE FROM ASSUMPTIONS UNDERLYING ANALYSIS OF VARIANCE IN REAL DATA

Introduction

Analysis of variance is based on three basic assumptions: First, and perhaps most important, is the assumption of normality; second, is that the variances of the distributions from which the samples have been taken are the same; third is the statistical independence of error deviations. The last assumption is usually not restrictive because researchers can generally perform most research such that this requirement is met. However, the other two assumptions can be violated in more ways that they can be satisfied.

Many different tests for normality as well as tests for homogeneity of variance have been developed. However, the information concerning these tests, as well as simulation results on the effects of departure from normality and heterogeneity of variance, are scattered throughout the literature. To facilitate understanding the limitations imposed by different methods of testing for normality and homogeneity of variance a brief review of the most significant tests as well as simulation results is given.

The results of applying some of the tests for normality and homogeneous variance to 1765 sets of real data from several different disciplines are presented and discussed. This information should prove valuable to both researchers and statisticians. It should enable either to minimize the chance of applying a nonrobust test to data where it is unlikely that the basic assumptions are true.

TESTING FOR NORMALITY

Limitations and Suitability of Existing Test Statistics

Testing distributional forms in general, and for normality in particular, has been an important area of continuing research in statistics. The main reason of such great interest is the fact that many important statistical procedure have properties that are based on the assumption of normality.

Many researchers have developed, or improved, tests which can detect departures from normality. The result of these efforts is the existence of several tests for normality. The intention of this review is not to go into detail on how to perform these tests, but to summerize what has been done until now.

Probably the oldest, and certainly the best known tests are the tests for skewness and kurtosis. The first is a measure of asymmetry, and the coefficient of skewness $g_1 = (b_1)^{\frac{1}{2}} = 0$ (where $(b_1)^{\frac{1}{2}} = m_3/m_2^{3/2}$, see 33, p 86) for symmetric distribution. When $g_1 \neq 0$ it can be concluded that there is some departure from normality, but the converse is not necessarily true; that is, when $g_1 = 0$ the distribution may be normal, but it may be any other symmetrical distribution as well. Therefore, some authors raise the question whether g_1 should be regarded as a test of normality (18, 38). However, this test, because of its simplicity, is recommended in many text books (7, 30, 33). The magnitude of Kurtosis is given by $g_2 = b_2 - 3$ (where $b_2 = m_4/m_2^2$, see 33, p 87).. For normal distribution $b_2 = 3$ and hence $g_2 = 0$. Positive values of g_2 indicate a density more peaked around its center than the density of a normal curve, while a negative value characterizes a density which is flatter at its center

than the normal curve. Both tests can be readily used, because tables for $(b_1)^{\frac{1}{2}}$ and b_2 at 1%, 5% and 10% significance levels are available.

Geary (1935) proposed the ratio of the mean deviation to the standard deviation to be used as a test of normality. However, his method of testing normality is not very reliable for samples smaller than 50 and, therefore, has limited value.

Some tests for normality are appropriate only as tests of simple hypotheses. The better known tests include: Chi-square goodness of fit test (33), Cramer-Von Mises test (1928) with test statistic CM, Kolmogorov-Smirnov test (1933) with test statistic KS, Weighted Cramer-Von Mises test (1954, see 32) with test statistic WCM and Durbin's test (1961) with test statistic D. However, when used as tests for normality, the mean and standard deviation of the hypothesized distribution must usually be specified. In most cases, where a test for normality is of interest, prior information regarding the parameters of the supposed normal distribution are not available and must be estimated. Unfortunately, these tests are generally quite sensitive even to relatively small misspecification of the parameters, and because of that their usefulness as practical statistical procedures is questionable.

David (1954) developed a test for normality (test statistic U) based on the ratio of range to standard deviation. The U statistic has particularly good properties against symmetric, especially short-tailed (uniform) distributions, but seems to be inefficient with respect to asymmetry, which is often the case regarding a departure from normality.

Shapiro and Wilk (1965) developed an excellent test for normality (test statistic W). The computation requires an extensive table of constants

because a different set of n/2 constants is required for each sample size n. The authors have been provided a table of these constants for sample size up to 50, as well as the table of critical values of W. This test seems to be the most powerful test for normality presently available. However, when sample size exceeds 50 the computation of W for normality testing becomes very cumbersome.

D'Agostino (1971 a) presented an alternative test for normality (test statistic D). The test is very powerful for detecting departures from normality for moderate and large sample sizes, and tables for D statistic are available for sample sizes 10 to 2000, (10, 38). The only inconvenience in using this test is that, because of very small range of D values, at least 5 decimal place accuracy should be used in its computation. This test seems to be an ideal complement to Shapiro-Wilk's test (1965) for sample sizes larger than 50.

Finally, there are some tests for normality, which are useful only in special cases. One such test was presented by Uthoff (1970) with test statistic W. This test is the most powerful test of normality against a distribution which is uniform. Unfortunately, W has relatively poor performance for heavy-tailed distributions.

Testing for Normality in Simulation Experiments

A few Monte Carlo simulation studies have been undertaken to investigate the performance of different tests for normality. The most extensive study of this kind was done by Shapiro and Wilk (1968). They presented results from a simulation study regarding the sensitivity of nine statistical procedures for evaluating the normality of a complete sample. The nine statistics were: W (Shapiro-Wilk, 1965), $(b_1)^{\frac{1}{2}}$ (skewness), b_2 (Kurtosis),

KS (Kolmogorov-Smirnov), CM (Cramer-Von-Mises), WCM (Weighted CM), D (modified KS), CS (chi-square) and U (studentized range). Their conclusion was: "(i) The W statistic provides generally superior omnibus measure of non-normality; (ii) the distance tests (KS, CM, WCM, D) are typically insensitive; (iii) the U statistic is excellent against symmetric, especially short-tailed, distributions, but has virtually no sensitivity to asymmetry; (iv) a combination of both $(b_1)^{\frac{1}{2}}$ and b_2 usually provides a sensitive judgment but even their combined performance is usually dominated by W."

Another interesting Monte Carlo simulation study was presented by Hogg (1972). Four statistics were employed: K, kurtosis, V, the ratio of one-half of the range to the mean deviation from the sample median; U, the ratio of standard deviation to the mean deviation from the sample median, and W, the ratio of one-half of the range to the standard deviation. A random sample of size 21 was taken from each of four symmetric distributions: uniform, normal, logistic and double exponential. For each sample K, U, V and W were computed. This was repeated 1000 times providing empirical distribution functions for these statistics with four different underlying distributions. The statistics were ranked from worst to best within each cell by assigning ranks 1, 2, 3, 4. The following averages of the rank of each statistic were obtained: 2.58 for K, 2.13 for U, 3.04 for V and 2.25 for W. V showed superior performance in this study. By another investigation Davenport (1971) supported this result, but only for sample size less than 30. Thus, K seems to be more suitable for testing normality over a wide range of sample sizes.

Normality in Real Experimental Data

Although testing for normality has been a well established procedure for some time, very little is known about the extent and types of non-normality encountered in real data.

The Statistical Laboratory at Kansas State University provides statistical consulting to several departments in the Kansas Agriculture Experimental Station at Manhattan, Kansas. As a result of this work there is easy access to many different data sets processed by Statistical Laboratory statistical programs. Over a period of two years, September, 1974 through July 1976, data was collected from each data set processed by the Statistical Laboratory analysis of variance program AARDVARK (Kemp, 1976). In total 1765 different data sets from several departments were used.

The primary interest was to determine what, if any, kinds of non-normality exist in real data and how frequently they occur. This information should be helpful to those interested in developing new tests or modifying existing tests to make them more robust to violations of the normality assumption. If one knows what kind of non-normality is most frequent, efforts can be directed toward developing of tests that are robust to that particular type of non-normality.

The secondary concern was to break the data down into subsets, by discipline, so that researchers in various fields of agriculture, biology, social science, etc., may have some idea as to the kinds of non-normality that are most common in their data. Given such information, they may then select statistical tests which are the most robust to that particular type of non-normality, thereby improving their chances of valid analyses.

In analysis of data sets for particular discipline we were obliged to eliminate departments and/or colleges from which we obtained insufficient number of data sets for a valid analysis. However, there were 11 departments and/or colleges that submitted 25 or more different data sets for which we were able to compute normality tests. The 11 departments for which such data were available were: Pathology (College of Veterinary Medicine), Plant Pathology (Agriculture), Dairy and Poultry Science (Agriculture), Agronomy (Agriculture), Horticulture and Forestry (Agriculture), Grain Science (Agriculture), Entomology, Adult and Occupational Education (College of Education), Industrial Engineering (College of Engineering), Foods and Nutrition (College of Home Economics), plus a general category of agriculture. Thus all together we had 11 different sources of data sets.

As is evident from previous discussions the number of normality tests available for general application is fairly limited. The three most commonly used tests for normality were chosen to be run on each of the real data sets. They were: W-test (Shapiro-Wilk, 1965) for samples of size 50 or less, skewness for samples of size 25 or more, kurtosis for samples of size 11 or more. These limitations were based on the availability of tables of critical values. In addition, the simple correlation between the subclass mean and variance was also computed to determine if the cell means and variances were independent. Each data set was analyzed for departure from normality by computing sum of squares using the deviation within the highest order interaction subclass fit in the model.

Table 1 shows that skewness is the least frequent form of non-normality of those computed in this study. In general the other tests show similar

results among themselves. Nearly one third of the data sets processed showed at least one kind of non-normality at the 0.05 level of significance. This certainly makes clear the need for tests robust to the violation of assumptions. Apparently, it is more important to have tests robust to kurtosis than to skewness. However, almost one-fourth of the data sets showed significant skewness at the 0.05 level of significance.

In reviewing Tables 2 through 12 we find that data from the Department of Pathology displays the least amount of non-normality with most of the percentages of significance close to the nominal type I error rate. The W-test was significant at the 0.05 level of significance for nearly half or more of the data sets from Industrial Engineering (Table 9), Grain Science (Table 10) and Dairy and Poultry Science (Table 3). The W-test is not a test for testing a specific type of non-normality. However, researchers in these disciplines should be careful about using tests based on normal distribution theory in view of the large proportion of data sets displaying some form of non-normality as indicated by the W-test. Other departments showed results quite similar to the combined data for all departments.

Data from the College of Agriculture (Table 4) and Grain Science (Table 10) showed very little skewness. Most of the other departments displayed skewness in a magnitude similar to the combined data except Plant Pathology (Table 6) which showed considerable skewness at the 5% and 10% levels of significance.

The Department of Plant Pathology (Table 6), Industrial Engineering (Table 9) and Grain Science (Table 10) showed high incidence of significant kurtosis with more than half the data sets having kurtosis significant at

10% level. Data from the College of Agriculture (Table 4), Department of Agronomy (Table 7), Horticulture and Forestry (Table 8), and Foods and Nutrition (Table 11) indicate that these departments have less kurtosis than most departments, but still have significant kurtosis at the 10% level of significance in about one-fourth of the data sets analysed.

The simple correlation between mean and variance was computed to determine whether or not the subclass mean and variance were independent, as they should be for a normal distribution with homogeneous variance among the subclass cells. The Department of Pathology (Table 2), Dairy and Poultry Science (Table 3), Adult and Occupational Education (Table 5), Plant Pathologh (Table 6), and Foods and Nutrition (Table 11) showed a lower incidence of significant correlation than the combined data (Table 1). The College of Agriculture (Table 4) and Department of Grain Science (Table 10) and Entomology (Table 12) displayed higher percentages of significant correlation than the combined data (Table 1), especially at the 10% level of significance.

To compare the frequency of the joint occurence of significant non-normality of different forms, a set of 2 x 2 contingency tables was constructed (Tables 13 to 16). These tables were based on data sets which had sample sizes in the range of 25 to 50. Only those data sets included all four tests as a result of the restrictions previously mentioned. While all four chi-square statistics were highly significant ($P \leq .1^{-4}$) the reader should note the strong relationship between the W-test and kurtosis. These results show that significant ($P \leq 0.05$) W-test and significant ($P \leq 0.05$) kurtosis very often occur together indicating that the W-test is quite sensitive to kurtosis. From Table 16 it can be seen that while there is a

relationship between kurtosis and skewness it is not nearly as strong as the relationship either has with the W-test.

Conclusion

A total of 1765 different data sets were subjected to three basic tests for normality: Shapiro-Wilk's W-test (1965), skewness and kurtosis. In addition the simple correlation between the subclass mean and variance was computed for the highest order interaction cell in an analysis of variance model. The results showed that about one-third of all data sets displayed a significant ($P \leq 0.05$) W-test, kurtosis or correlation between the subclass mean and subclass variance. About one-fourth of the data sets had significant ($P \leq 0.05$) skewness. Detailed information is given concerning a more specific break down of type of non-normality by disciplines for 11 major research areas. Chi-square contingency analysis showed a very strong relationship between the performance of the W-test and the incidence of kurtosis.

Table 1 Percent of Data Sets which Showed Non-normality for All Disciplines Combined

Test		W-test	Skew.	Kurt.	Corr.
No Ran		1153	1379	1765	1755
sig. level	.01	27.49	5.37	4.19	21.14
	.05	34.78	22.84	34.90	33.33
	.10	41.46	30.09	44.65	39.72

Table 2 Percent of Data Sets which Showed Non-normality for Department of Pathology

Test		W-test	Skew.	Kurt.	Corr.
No Ran		26	10	28	28
sig. level	.01	0.0	0.0	0.0	0.0
	.05	0.0	0.0	7.14	7.14
	.10	0.0	60.0	7.14	14.29

Table 3 Percent of Data Sets which Showed Non-normality for Department of Dairy and Poultry Science

Test		W-test	Skew.	Kurt.	Corr.
No Ran		114	120	134	134
sig. level	.01	29.82	0.0	0.0	0.0
	.05	43.86	31.67	29.85	22.39
	.10	54.39	40.00	31.34	29.85

Table 4 Percent of Data Sets which Showed Non-normality for College of Agriculture

Test		W-test	Skew.	Kurt.	Corr.
No Ran		94	86	142	142
sig. level	.01	21.28	0.0	0.0	19.72
	.05	31.91	6.98	16.90	36.62
	.10	46.81	11.63	25.35	45.07

Table 5 Percent of Data Sets which Showed Non-normality for Department of Adult and Occupational Education

Test		W-test	Skew.	Kurt.	Corr.
No Ran		25	0	25	25
sig. level	.01	12.00	0.0	0.0	0.0
	.05	28.00	0.0	32.00	24.00
	.10	28.00	0.0	40.00	24.00

Table 6 Percent of Data Sets which Showed Non-normality for Department of Plant Pathology

Test		W-test	Skew.	Kurt.	Corr.
No Ran		53	64	105	95
sig. level	.01	28.30	0.0	0.0	7.37
	.05	33.96	39.06	50.48	10.53
	.10	41.51	53.13	58.10	27.37

Table 7 Percent of Data Sets which Showed Non-normality for Department of Agronomy

Test		W-test	Skew.	Kurt.	Corr.
No Ran		168	163	255	255
sig. level	.01	19.64	14.72	9.41	14.12
	.05	27.98	22.09	21.96	30.20
	.10	30.95	31.29	35.69	33.33

Table 8 Percent of Data Sets which Showed Non-normality for Department of Horticulture and Forestry

Test		W-test	Skew.	Kurt.	Corr.
No Ran		144	220	331	331
sig. level	.01	10.42	0.00	0.00	25.68
	.05	16.67	30.91	25.08	33.84
	.10	26.39	41.82	36.86	39.27

Table 9 Percent of Data Sets which Showed Non-normality for Department of Industrial Engineering

Test		W-test	Skew.	Kurt.	Corr.
No Ran		19	88	97	97
sig. level	.01	36.84	2.27	2.06	24.74
	.05	47.37	36.36	41.24	30.93
	.10	47.37	43.18	53.61	38.14

Table 10 Percent of Data Sets which Showed Non-normality for Department of Grain Science

Test		W-test	Skew.	Kurt.	Corr.
No Ran		350	398	414	414
sig. level	.01	46.29	0.0	0.0	28.99
	.05	49.14	4.52	53.62	41.06
	.10	54.86	7.04	60.39	46.38

Table 11 Percent of Data Sets which Showed Non-normality for Department of Foods and Nutrition

Test		W-test	Skew.	Kurt.	Corr.
No Ran		64	82	82	82
sig. level	.01	15.63	0.00	0.00	4.88
	.05	34.88	24.39	14.63	14.63
	.10	43.75	36.59	39.02	19.51

Table 12 Percent of Data Sets which Showed Non-normality for Department of Entomology

Test		W-test	Skew.	Kurt.	Corr.
No Ran		46	84	84	84
sig. level	.01	30.43	0.00	0.00	16.67
	.05	39.13	26.19	33.33	30.95
	.10	43.48	33.33	47.62	45.24

Table 13 Contingency Table for Significant-Nonsignificant W-test versus Significant-Nonsignificant Skewness at Alpha = .05

Test		Skewness*		
		S	NS	TOT
W-test	S	88	201	289
	NS	2	412	414
	TOT	90	613	703

*Test statistic $\chi^2 = 134$; $P \leq 0.0001$

Table 14 Contingency Table for Significant-Nonsignificant W-test versus Significant-Nonsignificant Kurtosis at Alpha = .05

Test		Kurtosis*		
		S	NS	TOT
W-test	S	236	53	289
	NS	32	382	414
	TOT	268	435	703

* Test statistic $\chi^2 = 391$; $P \leq 0.0001$

Table 15 Contingency Table for Significant-Nonsignificant W-test versus Significant-Nonsignificant Correlation at Alpha = .05

Test		Correlation*		
		S	NS	TOT
W-test	S	162	127	289
	NS	56	358	414
	TOT	218	485	703

* Test statistic $\chi^2 = 140$; $P \leq 0.0001$

Table 16 Contingency Table for Significant-Nonsignificant Kurtosis versus Significant-Nonsignificant Skewness at Alpha = .05

Test		Skewness*		
		S	NS	TOT
Kurtosis	S	62	206	268
	NS	28	407	435
	TOT	90	613	703

* Test statistic $\chi^2 = 40$; $P \leq 0.0001$

TESTING FOR HOMOGENEITY OF VARIANCE

Discussion of Existing Tests

One of the basic assumptions underlying analysis of variance is the equality of the subclass variances. Moderate departures from this assumption do not, however, seriously affect the sampling distribution of the resulting F statistics when the equal cell numbers are used (30). However, if extreme inequality of variance is suspected and if cell numbers are not equal a test for homogeneity of variances is needed.

Several statistics are available for testing homogeneity of variances. Probably the most widely used test is Bartlett's test (1937) for homogeneity of variances with test statistic B.B is approximately distributed as chi-square with k-1 degrees of freedom where k is the number of subclasses. This test is very powerful if the assumption of normality holds. However, Box (1953) has shown that this test is extremely sensitive to non-normality and in some cases tends to give significant results when variances are equal. He illustrated this effect by using two extreme degrees of kurtosis with equal variances (n=30). When $b_2 = 2$ the probability of rejecting the hypothesis at the nominal 0.05 level was actually 0.849, while with $b_2 = -1$ the probability was only 0.00001.

Two other statistics commonly used for testing homogeneity of variances are Fmax proposed by Hartley (1950) and C developed by Cochran (1941). These tests are simpler computationally than Bartlett's test, but they show the same lack of robustness to non-normality.

As a remedy for this situation, Box (1953) proposed an approximate test, based on subdividing each population sample into c subsamples of size m (n=cm). However, there are no firm rules for the selection of c and m

and, therefore, this procedure is highly dependent on the skill of the user.

Finally, Miller (1968) recommended the Tukey's (1962) jackknife procedure for testing homogeneity of variances. This test is robust to violations of the normality assumption and seems to have fairly sufficient power.

Performance in Simulation Experiments

Several Monte Carlo simulation studies have been performed to investigate the robustness of different tests for homogeneity of variances to deviations from normality.

Miller (1968) applied seven tests to each of the 1000 pairs of samples of size 25: Fisher's F (1935), Box-Andersen (1955), jackknife (k=1, m=n), jackknife (k=5, m=5), Levene (1960) S, Box (1953) and Moses (1963) (k=5, m=5). The tests were run on data from uniform, normal, double exponential, skew double exponential and sixth power distributions. On the basis of this study the following conclusions were made:

i) The F test is extremely sensitive to non-normality.

ii) The jackknife with k=5 is not as powerful as the jackknife with k=1.

iii) The Box-Andersen test and the jackknife with k=1 have about the same power, and generally are the most powerful.

iv) The Levene S and the Box tests are robust, but they are less powerful then the tests under ii) and iii).

v) The Moses test (k=5) is the least powerful of all the tests.

Games (1972) presented two Monte Carlo simulation studies. In the first study the power curves of the Fmax test, the Cochran test, the two Levene

(1960) tests (L - X^2 and L - A) and Bartlett's test were compared using six different populations for samples of size n = 6. The second study contrasted Bartlett's test with Box and Andersen (1955) M' test, the Bartlett and Kendal (1946) tests (LEV-2 for k=2 and LEV-3 for k=3) and the Foster and Barr (1964) Q test using the same distributions as in the previous study, but for samples of size n = 18.

From the results obtained in the first study it is clear that Bartlett's, Fmax and Cochran's tests show the trends Box (1953) indicated. For the normal population when variances differ, Bartlett's and Fmax tests have superior power over all other tests. The Levene tests (L - X^2 and L - A) do not show the hoped-for robustness to violations of normality and generally have low power. From the results in the second study, it can be seen that the Foster and Barr Q-test performed similarly to the Bartlett's test. The LEV-2 test had unsatisfactory power. LEV-3 test should be used if there is reason to expect lepto-kurtosis (density curve more peaked around its center) in the populations or if there is no a priori information on the form. However, to compensate for lower power of this method the sample size must be increased.

Layard (1973) conducted two Monte Carlo simulation studies which differ only in the sample sized (n=25), n=10). Five hundred sets of four samples were generated in both cases. For each of 500 sets of samples Bartlett's, chi-square (24), Jackknife, and Box (1953) (k=5) test statistics were computed. The tests were run on data from uniform, normal and double exponential distributions. These results agree fairly well with those of Miller (1968). The author recommends the use of either jackknife or chi-square for samples of size greater than ten (minimum).

From the previous discussion it can be seen that a wide variety of tests for homogeneity of variances exist. However, most of the robust tests lack power. The Box-Andersen and jackknife tests seem to be the best choice. Unfortunately, these tests are based on subdivision of population samples into subsamples and for a wide range of sample sizes the number of possible subsamples is very large. Because of a lack of firm rules concerning subdivision into subclasses, these tests have little practical value. Therefore, we chose Bartlett's test because it is the most commonly used test.

Results and Discussion

Bartlett's test was run on 1765 different real data sets (see Normality in Real Experimental Data section of this paper). The results showed that 35% of these data sets had a significant Bartlett's statistic at the 0.01 type I error rate and almost 46% at the 0.10 significance level. These results indicate that unequal variances may well be very common in real data. Therefore, researchers and statisticians, working with such data, should be careful in using tests that are based on the assumption of equal variances.

Among the 1765 data sets there were 11 disciplines which had 25 or more data sets. The results for each discipline are given in Table 17. Most departments seem to have a large proportion of data sets with unequal variances. Only the Department of Pathology had low incidence of unequal variance.

In view of information in the literature regarding the effect of non-normality on Bartlett's test we decided to run 2 x 2 contingency tests on the 703 data sets for which all normality tests were performed. The

results are presented in Tables 18 through 21. These tables indicate that a significant Bartlett's test is associated with all the normality tests, however, it was most related to significant kurtosis. In view of Box's (1953) finding that Bartlett's test is not reliable in the presence of kurtosis, our results showing a high incidence of unequal variance may be invalidated.

Conclusion

One-thousand-seven-hundred and sixty-five different data sets were analysed for homogeneity of variance using Bartlett's test. The lowest heterogeneity of variance was found in Pathology and the highest in Grain Science. Most disciplines had a large proportion of data sets which showed a significant Bartlett's test. Seven hundred and three data sets were analyzed by means of 2 x 2 contingency tables for Bartlett's tests versus four normality tests. Bartlett's test was highly related to all departures from normality, but especially to kurtosis.

Table 17 Percent of Data Sets Showing a Significant Bartlett's Test

Department	No Ran	sig. level		
		0.01	0.05	0.10
Combined	1765	35.41	41.08	45.67
Pathology	28	7.14	7.14	14.29
Dairy and Poultry Science	134	28.36	37.31	43.28
Agriculture	142	16.90	18.31	19.72
Adult and Occupational Education	25	12.00	24.00	32.00
Plant Pathology	105	43.81	49.52	52.38
Agronomy	255	21.18	28.24	34.90
Horticulture and Forestry	331	32.02	36.25	39.88
Industrial Engineering	97	41.24	46.39	49.48
Grain Science	414	48.31	53.14	58.94
Foods and Nutrition	82	29.27	43.90	48.78
Entomology	84	38.10	42.86	47.62

Table 18 Contingency Table for Significant-Nonsignificant Bartlett's Test versus Significant-Nonsignificant W-test at Alpha = .05

Test		W-test*		
		S	NS	TOT
Bartlett's	S	206	85	291
	NS	83	329	412
	TOT	289	414	703

* Test statistic $\chi^2 = 179$; $P \leq 0.0001$

Table 19 Contingency Table for Significant-Nonsignificant Bartlett's Test versus Significant-Nonsignificant Skewness at Alpha = .05

Test		Skewness*		
		S	NS	TOT
Bartlett's	S	56	235	291
	NS	34	378	412
	TOT	90	613	703

* Test statistic $\chi^2 = 17$; $P \leq 0.0001$

Table 20 Contingency Table for Significant-Nonsignificant Bartlett's Test versus Significant-Nonsignificant Kurtosis at Alpha = .05

Test		Kurtosis*		
		S	NS	TOT
Bartlett's	S	224	67	291
	NS	44	368	412
	TOT	268	435	703

* Test statistic $\chi^2 = 318 \leq 0.0001$

Table 21 Contingency Table for Significant-Nonsignificant Bartlett's Test versus Significant-Nonsignificant Correlation at Alpha = .05

Test		Correlation*		
		S	NS	TOT
Bartlett's	S	158	133	291
	NS	60	352	412
	TOT	218	485	703

* Test statistic $\chi^2 = 124$; $P \leq 0.0001$

REFERENCES

1. Bartlett, M. S. 1937. Properties of Sufficiency and Statistical Tests. Proceedings of the Royal Society, A901, Vol. 160, 268.

2. Bartlett, M. S. and D. G. Kendal 1946. The Statistical Analysis of Variance-heterogeneity and the Logarithmic Transformation. J. Roy. Statist. Soc., Supplement 8, 128.

3. Boneau, C. A. 1960. The Effects of Violations of Assumptions Underlying the t-test. Psychological Bulletin, Vol. 57, No. 1, 49.

4. Box, G. E. P. 1953. Non-normality and Tests on Variances. Biometrika, Vol. 40, 318.

5. Box, G. E. P. and S. L. Andersen 1955. Permutation Theory in the Derivation of Robust Criteria and the Study of Departures from Assumptions. J. Roy. Statist. Soc., Vol. 27, 1.

6. Cochran, W. G. 1941. The Distribution of the Largest of a Set of Estimated Variances as a Fraction of Their Total. Annals of Eugenics, Vol. 11, 47.

7. Cramer, H. 1928. "On the Composition of Elementary Errors." Skand. Aktuar, Vol. 11, 482.

8. Croxton, P. E., D. J. Cowden and S. Klein 1967. Applied General Statistics. Prentice-Hall, Inc., Englewood Cliffs, N. J.

9. D'Agostino, R. B. 1971a.. An Omnibus Test of Normality for Moderate and Large Samples. Biometrika, Vol. 58, 341.

10. D'Agostino, R. B. 1971b. Tables for the D Test of Normality. Department of Mathematics, Boston University Research Report.

11. Davenport, A. 1971. Unpublished Ph.D. Dissertation. University of Iowa.

12. David, H. A., H.O. Hartley and E. S. Pearson 1953. The Distribution of the Ratio in a Single Normal Sample of Range to Standard Deviation. Biometrika, Vol. 40, 318.

13. Durbin, J. 1961. "Some Methods of Constructing Exact Tests." Biometrika Vol. 48, 41.

14. Eisenhart, Churchill 1947. The Assumption Underlying the Analysis of Variance. Biometriks, Vol. 3, No. 1, 1.

15. Fisher, R. A. 1935. The Design of Experiments. Oliver and Boyd, London.

16. Foster, L. A. 1964. The Q-test for Equality of Variances. (Unpublished Doctoral Dissertation, Purdue University). Ann Arbor, Michigen; University Microfilms, No. 65-5008.

17. Games, P. A. 1972. Robust Tests for Homogeneity of Variance. Educ. and Psych. Measurement, Vol. 32, 887.

18. Geary, R. C. 1935. The Ratio of the Mean Deviation to the Standard Deviation as a Test of Normality. Biometrika, Vol. 27, 310.

19. Hartley, H. O. 1950. The Maximum F-ratio as a Short-cut Test for Heterogeneity of Variance. Biometrika, Vol. 37, 308.

20. Hogg, R. V. 1972. More Light on the Kurtosis and related Statistics. JASA, Vol. 67, No. 338, 422.

21. Kemp, K. E. 1976. AARDVARK Reference Manual. Kansas Agricultural Experiment Station. Contribution 209.

22. Kirk, K. E. 1969. Experimental Design: Procedures for the Behavioral Science. Brooks/Cole Publishing Co., Belmont, California.

23. Kolmogorov, A. N. 1933. "Sulla Determinazione Empirica di Una Legge di Distribuzione." G. Inst. Ital. Attuari, Vol. 4, 83.

24. Layard, M. W. J. 1973. Large-Sample Tests for Homogeneity of Variances. JASA, Vol. 68, No. 341, 195.

25. Levene, H. 1960. Robust Tests for Equality of Variances, in Contribution to Probability and Statistics. Ed. Olkin, I., Stanford Univ. Press, Stanford, California.

26. Miller, R. G. 1966. Simultaneous Statistical Inference. McGraw-Hill Book Co., New York.

27. Miller, R. G. 1968. Jackknifing Variances. The Annals of Math. Stat., Vol. 39, No. 2, 567.

28. Moses, L. E. 1963. Rank Tests of Dispersion. Ann. Math. Statist. Vol. 35, 1594.

29. Neyman, J. 1926. On Correlation of the Mean and the Variance in Samples Drawn from an "Infinite" Population. Biometrics, Vol. 18,401.

30. Scheffe, H. 1959. The Analysis of Variance. John Wiley & Sons, Inc. New York.

31. Shapiro, S. S. and M. B. Wilk 1965. An Analysis of Variance Test for Normality (Complete Samples). Biometrika, Vol. 52, 591.

32. Shapiro, S. S., M. B. Wilk and Mrs. H. J. Chen 1968. A Comparative Study of Various Tests for Normality. Am. Stat. Journal, Dec., 1343.

33. Snedecor, G. W. and W. G. Cochran 1967. Statistical Methods. The Iowa State University Press.

34. Splawa-Neyman J. 1925. Contribution to the Theory of Small Samples Drawn from a Finite Population. Biometrika Vol. 17, Biometrika, Vol. 17, 472.

35. Tukey, J. W. 1962. Data Analysis and Behavioral Science. Unpublished Manuscript.

36. Uthoff, V. A. 1970. "An Optimum Test Property of Two Well Known Statistics." JASA, Vol. 65, 1597.

37. Winer, B. J. 1962. Statistical Principles in Experimental Design. McGraw-Hill Book Co. Inc., New York.

38. Zar, J. H. 1974. Biostatistical Analysis. Prentice-Hall, Inc. Englewood Cliffs, N. J.

POWER OF MULTIPLE COMPARISON PROCEDURES IN REAL DATA

Introduction

Numerous procedures are available for the performance of multiple comparisons in the analysis of variance. The basic rule is to compare the observed difference between any two means to the critical value corresponding to the multiple comparison test used. If the observed difference is larger than the critical value, the difference is declared significant and vice versa. However, the magnitude of critical values for different multiple comparison test may vary considerably and so will the number of significant differences declared. Because of that, large controversy has arisen among statisticians concerning the reliability of different multiple comparison procedures.

Due to this situation, the area of multiple comparison methods has become one of the most confusing area in statistics. The choice of the proper procedure for a particular set of data from the existing arsenal of multiple comparison techniques has become the problem of how to untie the Gordian knot. A good discussion of the pith of the problem was given by Kemp (1973).

Another stumbling block in the use of multiple comparison tests is whether or not a significant F-test must first be obtained before multiple comparisons can be performed. Although, some authors claim that their procedures do not require an a priori F-test, for example, Duncan (1955), many researchers prefer to apply even those procedures only after a significant overall F-test has been found. However, the purpose of this paper is not to recommend the best multiple comparison procedure, but to present a study of actual behaviour (power) of chosen multiple comparison tests in real data.

To facilitate the reader's understanding of the results obtained, theoretical considerations as well as simulation results concerning the most studied multiple comparison procedures will first be discussed.

Theoretical Considerations Concerning Multiple Comparison Procedures

Up to now, the best known multiple comparison procedures are the following: Multiple t-test or unprotected LSD (least significant difference), Fisher's (1935) LSD, or protected t-test. Scheffe's (1959) S method, Student-Newman-Keuls (1927, 1939, 1952) SNK (also known as Newman-Keuls test), Duncan's (1955) DNMRT (new multiple range test), Tukey's (1953) HSD (honestly significant difference) and the Waller-Duncan (1969) BLSD (Bayes t procedure). However, there is common agreement that Scheffe's S method is the most conservative test of all multiple comparison procedures and, therefore, will be excluded from further consideration.

The critical value for unprotected LSD (multiple t-test) is computed as:

$$LSD = t(\alpha, f)s_d$$

where $t(\alpha, f)$ is the table value of Student's (1908) t-statistic for significance level α and f degrees of freedom of the standard error of the difference between two means (s_d). If the difference between any two means exceeds the LSD value it is declared to be significant. This procedure was criticized by many statisticians because of its low protection against Type I error, especially if all means are homogeneous.

A modification of unprotected LSD was proposed by Fisher (1935). He suggested that LSD test should be performed only if an overall significant F-test is obtained, otherwise the differences should be declared nonsignificant. In other words, if F-test is significant

$$LSD = t(\alpha, f)s_d$$

as previously, but if F-test is not significant

$$LSD = \infty$$

This test is very powerful in detecting differences among the means and has more protection against Type I error when means are homogeneous, than does the unprotected LSD. However, some authors think that protection against Type I error is too low and, therefore, do not recommend its use.

To provide added protection against Type I error, Tukey (1953) developed HSD test, which has critical value:

$$HSD = q(\alpha, p, f)s_d(2)^{-\frac{1}{2}},$$

where $q(\alpha, p, f)$ is studentized range statistic at α significance level, for p treatments and f degrees of freedom for error (s_d). HSD is equal to LSD for two means, but when there are more than two means HSD is considerably larger and, therefore, the power of the test is lower than LSD. A priori F-test is not required for HSD.

While LSD and HSD procedures each require the computation of a single critical value, the Student-Newman-Keuls test requires calculation of $p = (n-1)$ critical values; thus the critical values are:

$$SNK_p = q(\alpha, p, f)s_d(2)^{-\frac{1}{2}},$$

and vary depending on the number (p) of means in the set, where $q(\alpha, p, f)$ and s_d are the same as for HSD. This procedure performs similarly to Tukey's HSD test concerning the power and protection against Type I error.

Duncan (1955) developed a multiple comparison test for which the rate of Type I error is somewhere between LSD and HSD. A preliminary F-test is not required. The critical value is given by:

$$DNMRT_p = q_d(\alpha, p, f)s_d(2)^{-\frac{1}{2}},$$

where q_d can be obtained from special tables computed for this test by Duncan. The critical value $DNMRT_p$ varies as does the SNK test, with the number (p) of means in the set. This test is more powerful than SNK and HSD, but statisticians have divided opinion concerning its protection against Type I error.

Finally, a test based on a Bayesian approach to the multiple comparison problem was proposed by Waller and Duncan (1969). Its critical value is computed as:

$$BLSD = t(k,F,f,q)s_d,$$

where t(k,F,f,q) is the minimum average risk t value for the chosen value of k (Type I to Type II error weighted ratio), F is the value of ordinary F-test, f is degrees of freedom of error, and q is the degrees of freedom among treatments. No preliminary F-test is necessary, but t becomes infinite when F = 1. When F is large the performance of BLSD is similar to LSD, while for F less than 2.5, the BLSD is more conservative and closer to HSD. Such properties of this test should result in a better power for large F and smaller Type I error rate for small F.

Simulation Results

Several Monte Carlo simulation studies concerning the behaviour of most used MRT's (multiple comparison tests) are available.

Balaam (1963) examined the behaviour of LSD, SNK and DNMRT under two conditions, that is, when a preliminary F-test was not performed (method I) and after a significant F-test was obtained (method II). He concluded that under both conditions NSK has considerably lower power than LSD and DNMRT, while LSD had the greatest power with sufficient protection against Type I error.

Waller (1970) in a Monte Carlo simulation study examined four procedures: Multiple-t, Fisher's LSD, Tukey's HSD and Bayes BLSD. He found that HSD had a low power, while BLSD had a good power and sufficient protection against Type I error.

Boardman (1971) studied five MRT's: LSD, DNMRT, SNK, HSD and Scheffe's S procedure to compare their Type I error rates (comparisonwise and experimentwise). He concluded that LSD and DNMRT are too liberal in terms of experimentwise error rates, while Scheffe is extremely conservative, therefore, the choice should be SNK or HSD.

Carmer and Swanson (1971, 1973) presented two Monte Carlo simulation studies. In the first study five MRT's were examined: LSD (unprotected), FLSD (protected), TSD, DNMRT and BLSD. They found remarkably high degree of similarity in the sensitivity of LSD, FLSD and BLSD, DNMRT was consistently slightly less powerful than LSD, while TSD had a very poor sensitivity. They concluded that experimenters should use FLSD, DNMRT or BLSD. TSD should be avoided unless experimenter is extremely concerned about Type I error.

The second study represents the most extensive study of its kind. They compared 10 MRT's with respect to their power and Type I and Type II error rates. The results obtained in the previous study concerning FLSD, DNMRT and BLSD were confirmed. They concluded that Scheffe's S method and TSD should not be used because they are too conservative.

Thomas (1973) compared seven methods of pairwise comparisons. He concluded that LSD should not be used because of its high Type I error rate, while DNMRT seems to be the best choice, because of its acceptably low Type I error rate and uniformly high power.

Einot and Gabriel (1975) compared the power of several multiple comparison methods. They stated: "No Monte Carlo study was needed to realize the alleged inferiority of the Scheffe, Tukey and Newman-Keuls procedures for detecting real differences". They recommend the use of Rayan's (1962) method.

Discussion of Results Obtained in Real Data

As previously stated a large controversy has arisen among statisticians concerning the reliability of different MRT's. Several researchers conducted valuable Monte Carlo simulation studies, which were undertaken to measure the differences in sensitivity of the MRT's. However, there are some factors which depend on the structure of real data and cannot be fully predicted in a simulation study. These factors are: the magnitude of true differences and the level of homogeneity among the true treatment means. These factors might have a big effect on the performance of MRT's, thus, if in a real data set there are many big differences between treatment means, the difference in the power of the procedures may be small and vice versa.

To investigate the sensitivity of MRT's in real data, 1765 different data sets were analyzed by four MRT's: Fisher's (1935) LSD, Duncan's (1955) DNMRT, Waller-Duncan (1969) Bayes-t procedure BLSD for main effects with more than two levels and Tukey's (1953) HSD for treatments with 20 or fewer levels. For the reasons given in the introduction all MRT's were applied in two ways:

1. Comparisons of means was performed irrespective of F value;
2. Comparison of means was performed only if F-test was significant at 5% significance level. When F-test was not significant it was declared that there were no differences among the means.

Percent rejection of MRT's and possible number of comparisons for two methods are presented in Tables 1 and 2 respectively. When all effects and levels were present only LSD and Duncan were compared, because of the restrictions imposed by HSD and BLSD. F-test was significant in 39.28% of the cases. LSD was more powerful than DNMRT as expected. For all effects and 20 or fewer levels HSD was, also, included. LSD had superior power, but this time DNMRT was only slightly less powerful than LSD, while HSD had considerably lower power than the other tests. F-test had slightly lower percentage of significance. All four MRT's were compared for main effects only and twenty treatment levels or fewer. The difference in power of LSD, DNMRT and BLSD was small with LSD again leading, while HSD was again considerably lower. The F-test was significant in 48.36% of the cases, which indicated that there was less homogeneity among the means when only main effects were considered.

By examination of Table 2 it can be seen that the power of all MRT's was increased, when they were applied after a significant F-test was obtained. For all levels and effects as well as for all effects and 20 or fewer treatment levels, LSD and Duncan showed slightly smaller differences in power than in the previous case; LSD having superior power. When all MRT's were compared LSD remained most powerful, but this time BLSD showed slightly higher power than DNMRT. HSD was in all cases even more conservative.

In Table 3 percent of MRT's as well as possible number of multiple comparisons when F ratio was not significant at 0.05 level are given. In all cases LSD had the highest percentage of rejection. DNMRT was slightly lower than LSD. HSD, as expected, was by far lower than LSD and DNMRT.

BLSD showed the best performance being lower even than HSD.

Taking in account the results shown in Tables 2 and 3 we can conclude that BLSD has high power when a preliminary F ratio was significant and the lowest percentage of rejection when the F ratio was not significant. Therefore, this test seems to be the best choice. However, this test can be applied only for two or more levels of main effects and, therefore, cannot be generally used.

To further investigate whether BLSD has the property to be conservative for small F values and to have a high power for large F, all MRT's were applied to the data sets with treatment levels greater than two and 20 or fewer of main effects only. An a priori F-test was not performed. The limiting value of F was 2.5. The results are presented in Table 4. When the F value was smaller than 2.5, BLSD was more powerful than DNMRT and close to LSD, thus the expectation based on theoretical considerations was fully confirmed by its performance in real data.

The results in Tables 1, 2, 3 and 4 seem to be in close agreement with theoretical considerations and simulation results previously discussed. The conclusion made by Carmer and Swanson (1971, 1973) on the basis of their simulation results, that LSD (protected), DNMRT and BLSD should be used when power is the criterion of interest is fully confirmed by their performance on real data.

Conclusion

Four multiple comparison procedures were applied to analyze 1765 different data sets. MRT's were performed with and without a preliminary significant F-test. In both cases LSD showed superior power, while HSD was by far the most conservative procedure. Duncan's test showed slightly

lower power than LSD, while BLSD performed similarly to DNMRT, having slightly lower power than DNMRT when a preliminary F-test was not performed (Table 1) and slightly higher when a significant F-test was required (Table 2). When F-test was not significant (Table 3) LSD and DNMRT showed considerably higher percentage of rejection than HSD and BLSD; LSD being the highest and BLSD the lowest. It was confirmed that BLSD has high protection against Type I error for small F-values and a high power for large F. LSD (protected) and DNMRT can be recommended when power is the criterion of main concern. HSD can be recommended only when avoidance of Type I error is very important. For data sets with two or more levels of main effects only, BLSD is the best choice because of its high protection against Type I error for small F-values and high power for large F ratios.

Table 1 Percent Rejection of MRT's When F Ratio was Ignored

Effects	Levels-L		LSD	DNMRT	HSD	BLSD	F
All	All	%	48.59	42.46	-	-	39.28
		No	1,140,751	1,140,751			
	$L \leq 20$	%	41.77	38.99	26.97		38.72
		No	226,739	226,739	226,739	-	
Main	$2 < L \leq 20$	%	39.88	36.77	25.49	35.96	48.36
		No	48,514	48,514	48,514	45,856	

Table 2 Percent Rejection of MRT's When F Ratio was Significant at 0.05 Level

Effects	Levels-L		LSD	DNMRT	HSD	BLSD	F
All	All	%	59.47	52.47	-	-	100
		No	631,742	631,742			
	$L \leq 20$	%	59.62	56.98	42.72	-	100
		No	76,314	76,314	76,314		
Main	$2 < L \leq 20$	%	54.97	51.67	36.52	53.62	100
		No	75,529	75,529	75,529	30,209	

Table 3 Percent Rejection of MRT's When F Ratio was not Significant at 0.05 Level

Effects	Levels-L		LSD	DNMRT	HSD	BLSD	F
All	All	%	35.10	29.59	-	-	0
		No	500,009	500,009			
	L ≤ 20	%	32.72	29.86	18.96	-	0
		No	150,425	150,425	150,425		
Main	2 < L ≤ 20	%	8.63	5.91	2.66	1.87	0
		No	15,799	15,799	15,799	15,647	

Table 4 Percent Rejection of MRT's for Large and Small Values of F

F	LSD	DNMRT	HSD	BLSD	F
≤ 2.5	8.62	5.82	2.58	1.93	1.56
> 2.5	56.69	53.42	37.81	55.26	89.99

REFERENCES

1. Anderson, D. A. 1972. Overall Confidence Levels of the Least Significant Difference Procedure. The American Statistician, Vol. 26, No. 4, 30.

2. Balaam, L. N. 1963. Multiple Comparisons - A Sampling Experiment. The Australian Journal of Statistics, Vol. 5, No. 2, 62.

3. Boardman, T. J. and D. R. Moffitt 1971. Graphical Monte Carlo Type I Error Rates for Multiple Comparison Procedure. Biometrics, Vol. 27, 738.

4. Carmer, S. G. and M. R. Swanson 1971. Detection of Differences Between Means: A Monte Carlo Study of Five Pairwise Multiple Comparison Procedures. Agronomy Journal, Vol. 63, Nov-Dec, 940.

5. Carmer, S. G. and M. R. Swanson 1973. An Evaluation of Ten Pairwise Multiple Comparison Procedures by Monte Carlo Methods. Journal of the Am. Stat. Assoc., Vol, 68, No. 341, 66.

6. Cox, D. R. 1965. A Remark on Multiple Comparison Methods. Technometrics, Vol. 7, No. 2, 233.

7. Duncan, D. B. 1951. A Significance Test for Differences Between Ranked Treatments in an Analysis of Variance. The Virginia Journal of Science, Vol. 2, New Series, No. 3, 171.

8. Duncan, D. B. 1955. Multiple Range and Multiple F Tests. Biometrics Vol. 11, 1.

9. Duncan, D. B. 1965. A Bayesian Approach to Multiple Comparisons. Technometrics, Vol. 7, No.2, 171.

10. Dunnett, C. W. 1970. 271 Query: Multiple Comparison Tests. Biometrics, Vol. 26, 139.

11. Einot, I. and K. R. Gabriel 1975. A Study of the Powers of Several Methods of Multiple Comparisons. Journal of the Am. Stat. Assoc., Vol. 70, No. 351, 574.

12. Fisher, R. A. 1935. The Design of Experiments. Oliver and Boyd, London.

13. Games, P. A. 1971. Multiple Comparisons of Means. American Educational Research Journal, Vol. 8, No. 3, 531.

14. Gill, J. L. 1971. Current Status of Multiple Comparisons of Means in Designed Experiments. Journal of Dairy Science, Vol. 56, 973.

15. Haas, K. 1970. Multiple Comparison Methods, American Psychologist, Vol. 25, 365.

16. Harter, H. L. 1957. Error Rates and Sample Sizes for Range Tests in Multiple Comparisons. Biometrics, Vol. 13, 511.

17. Harter, H. L. 1970. Multiple Comparison Procedures for Interactions. The American Statistician, Vol. 24, No. 5, 30.

18. Kemp, K. E. 1973. Multiple Comparisons: Comparisonwise Versus Experimentwise Type I Error Rates and Their Relationship to Power. Journal of Dairy Science, Vol. 58, No. 9, 1374.

19. Keselman, H. J. 1976. A Power Investigation of the Tukey Multiple Comparison Statistic. Educational and Psychological Measurement, Vol. 36, 97.

20. Keuls, M. 1952. The Use of the "Studentized Range" in Connection with an Analysis of Variance. Euphytica, Vol. 1, 112.

21. Kirk, R. E. 1969. Experimental Design: Procedures for the Behavioral Sciences. Brooks/Cole Publishing Co., Belmont, California.

22. Miller, R. G. 1966. Simultaneous Statistical Inference. McGraw-Hill Book Co., N. Y.

23. Newman, D. 1939. The Distribution of the Range in Samples from a Normal Population, Expressed in Terms of an Independent Estimate of Standard Deviation. Biometrika, Vol. 31, 20.

24. O'Neill, R. and G. B. Wetherill 1971. The Present State of Multiple Comparison Methods. Royal Stat., Soc. (series B), 218.

25. Petrinovich, L. P. and C. D. Hardyck 1969. Error Rates for Multiple Comparison Methods: Some Evidence Concerning the Frequency of Erroneous Conclusions. Psychological Bulletin, Vol. 71, No. 1, 43.

26. Rayan, A. Thomas 1962. The Experiment as the Unit for Comparing Rates of Error. Psychological Bulletin, Vol. 59, No.4, 301.

27. Scheffe, H. 1953. A Method for Judging all Contrasts in the Analysis of Variance. Biometrika, Vol. 40, 87.

28. Scheffe, H. 1959. The Analysis of Variance. John Wiley & Sons, Inc. New York.

29. Smawley, R. R. 1969. After a Significant F in ANOVA, The Journal of Experimental Education, Vol. 37, No. 3, 75.

30. Snedecor, G. W. and W. G. Cochran 1967. Statistical Methods. The Iowa State University Press.

31. Spjtvoll, E. 1974. Multiple Testing in the Analysis of Variance. Scandinawien Journal of Statistics, Vol. 1, No. 3, 97.

32. Student, 1908. The Probable Error of a Mean, Biometrika. Vol. 6, 1.

33. Student, 1927. "Error of Routine Analysis", Biometrika. Vol. 19, 151.

34. Thomas, D. A. H. 1973. Multiple Comparisons among Means - a Review. The Statistician, Vol. 22, No. 1, 17.

35. Thomas, D. A. H. 1974. Error Rates in Multiple Comparisons among Means - Results of a Simulation Exercize. Applied Statistics, Vol. 23, No. 3, 284.

36. Ury, K. H. and A. D. Wiggins 1975. A Comparison of Three Procedures for Multiple Comparisons among Means. Br. J. Math. Statist. Psychol., Vol. 28, 88.

37. Tukey, J. W. 1953. The Problem of Multiple Comparisons. Ditto, Princeton University.

38. Waldo, D. R. 1976. An Evaluation of Multiple Comparisons Procedures. Journal of Animal Science, Vol. 42, No. 2, 539.

39. Waller, R. A. and D. B. Duncan 1969. A Bayes Rule for the Symmetric Multiple Comparisons Problem. Journal of American Statistical Assoc., Vol. 64, 1484.

40. Waller, R. A. 1970. On the Bayes Rule for the Symmetric Multiple Comparisons Problem. Unpublished Notes. Kansas State University, Manhattan 66506.

41. Waller, R. A. and K. E. Kemp 1976. Computations of Bayesian t-values for Multiple Comparisons. Journal of Statistical Computation and Simulation, Vol. 4, 169.

42. Wine, R. L. A. 1955. A Power Study of Multiple Range and Multiple F Tests. Virginia Polytechnic Institute, Blacksburg, Technical Report No. 12.

43. Winer, J. H. 1962. Statistical Principles in Experimental Design. McGraw-Hill Book Co. Inc., New York.

44. Zar, J. H. 1974. Biostatistical Analysis. Prentice-Hall Inc. Englewood Cliffs, N. J.

ROBUSTNESS OF MULTIPLE COMPARISON PROCEDURES TO DEPARTURES FROM NORMALITY AND HETEROGENEITY OF VARIANCE IN REAL DATA

Multiple comparison procedures in recent years have become an important tool in the analysis and interpretation of experimental results in many sciences. However, very little is known about the robustness of these procedures to violations of the assumptions underlying analysis of variance. On the basis of theoretical and empirical studies it can be concluded that both F and t-tests are little affected by the violation of assumptions if equal cell numbers exist and there is not extreme heterogeneity of variances. In these cases testing the equality of only two means does not present much of a problem.

Unfortunately, with the exception of the studentized range statistic q, the behaviour of the statistics which are used in multiple comparison tests have been explored very little. Ramsayer (1973), in a Monte Carlo simulation study, investigated robustness of the q statistic under various pattern of heterogeneous variances and departures from normality. He concluded that q statistic withstands violations of the homogeneity of variance assumption remarkably well as well as violations of the normality assumption, when Type I error rate is the criterion of interest. However, more work is needed regarding the robustness of q when power is the criterion of interest. As for the other statistics used for multiple comparisons there is little information about their robustness to the violation of assumptions.

To throw more light on this question real data were subjected to three multiple comparison procedures: Fisher's LSD (1935), Duncan's New Multiple Range test (1955), and Tukey's HSD (1953) for treatments with 20 or fewer

levels. Three tests for normality were also computed: Shapiro-Wilk (1965) W-test, skewness and kurtosis (see 6 pp 86, 87). To test whether subclass means and variances were independent, as they should be if all assumptions hold, a simple correlation between the subclass means and variances was computed. Finally Bartelett's test (1937) was used for testing homogeneity of variances.

In Figures 1 through 5 percent differences found by MRT's (multiple range tests) at 0.05 level at various significance levels of normality tests as well as Bartlett's test is given. There were 5 significance levels for W-test, correlation and Bartlett's test, while for kurtosis and skewness the last two levels are pooled. Only data sets for 20 or fewer levels of main effects were analyzed.

Figure 1 shows percent of differences declared significant by the three MRT's for various significance levels of the W-test. All the MRT's showed the same trend. The highest percentage of significant differences occured between the 0 and 0.01 significance level, while the lowest was for alpha hat between 0.01 and 0.05. The percentage was somewhere around the mean of previous two for other three levels.

Figure 2 shows the percent differences found by the three MRT's at various levels of significance of skewness. All tests had their maximum and minimum at the same alpha hat as previously, but, this time the percentage continued to increase at other significance levels.

In Figure 3 the percent differences found by the three MRT's at various significance levels of kurtosis is presented. All tests showed again, the same pattern with maximum percentage for alpha hat between the 0 and 0.01, but this time minimum occured for alpha hat between 0.05 and 0.1.

From Figure 4 the percent differences detected by the MRT's at different significance levels of correlation can be seen. For all tests the maximum percentage occured for alpha hat between the 0 and 0.01, while the minimum for all MRT's was at alpha hat between 0.05 and 0.1.

From Figures 1 to 4 it is evident that LSD and Duncan's tests performed similarly at all levels of significance for the non-normality tests, LSD having slightly higher percentage of detected differences. Tukey's test was much more conservative in all cases, especially for alpha hats in the ranges of 0.01 to 0.05 and 0.05 to 0.1.

Figure 5 shows percent differences detected by three MRT's at five significance levels of Bartlett's test. Duncan's, LSD and Tukey's tests had the same trend at all significance levels of Bartlett's test except at alpha hat between 0.1 and 0.5 where Duncan's was closer to Tukey's than to LSD. This is somewhat surprising.

Conclusion

A study of robustness of four multiple comparison procedures to violations of the assumptions in real data was presented. Percentage of differences found by MRT's for various alpha hats of normality tests as well as Bartlett's test was given. All MRT's showed similar trends at different significance levels of normality tests as well as Bartlett's test; Tukey's test was the most conservative. From Figure 1 through 5 it can be seen that all MRT's had a maximum at alpha hat between the 0 and 0.01 significance levels of the normality tests as well as Bartlett's test. For all MRT's the minimum occured at alpha hat between 0.01 and 0.05 significance levels of W-test, skewness and Bartlett's test, while for kurtosis and correlation it was shifted to alpha hat between 0.05 and 0.1.

FIG. 1 PERCENT DIFFERENCES FOUND BY MULTIPLE RANGE TESTS AT VARIOUS ALPHA HATS FOR W-TEST

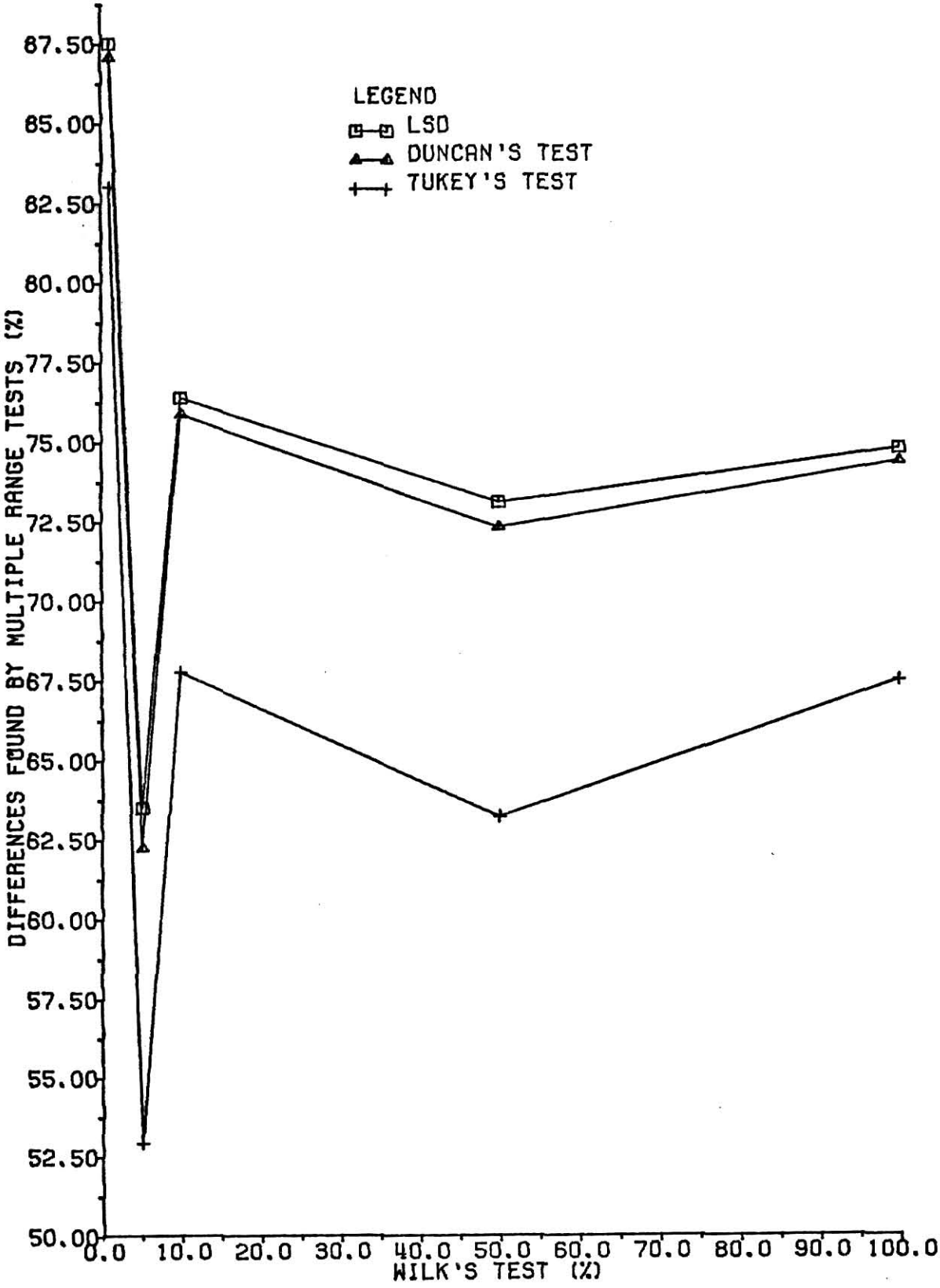


FIG. 2 PERCENT DIFFERENCES FOUND BY MULTIPLE RANGE TESTS AT VARIOUS ALPHA HATS FOR SKEWNESS

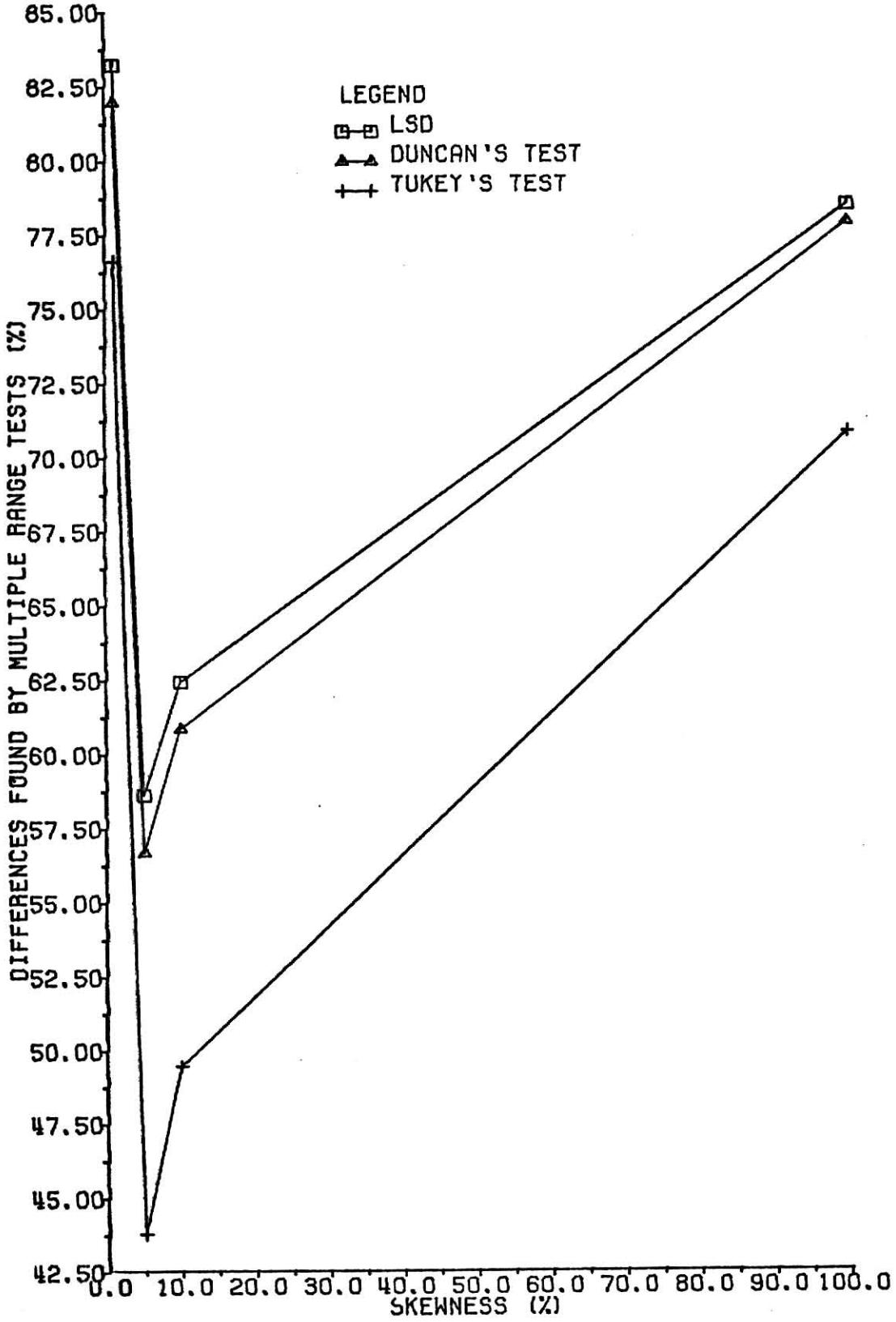


FIG. 3 PERCENT DIFFERENCES FOUND BY MULTIPLE RANGE TESTS AT VARIOUS ALPHA HATS FOR KURTOSIS

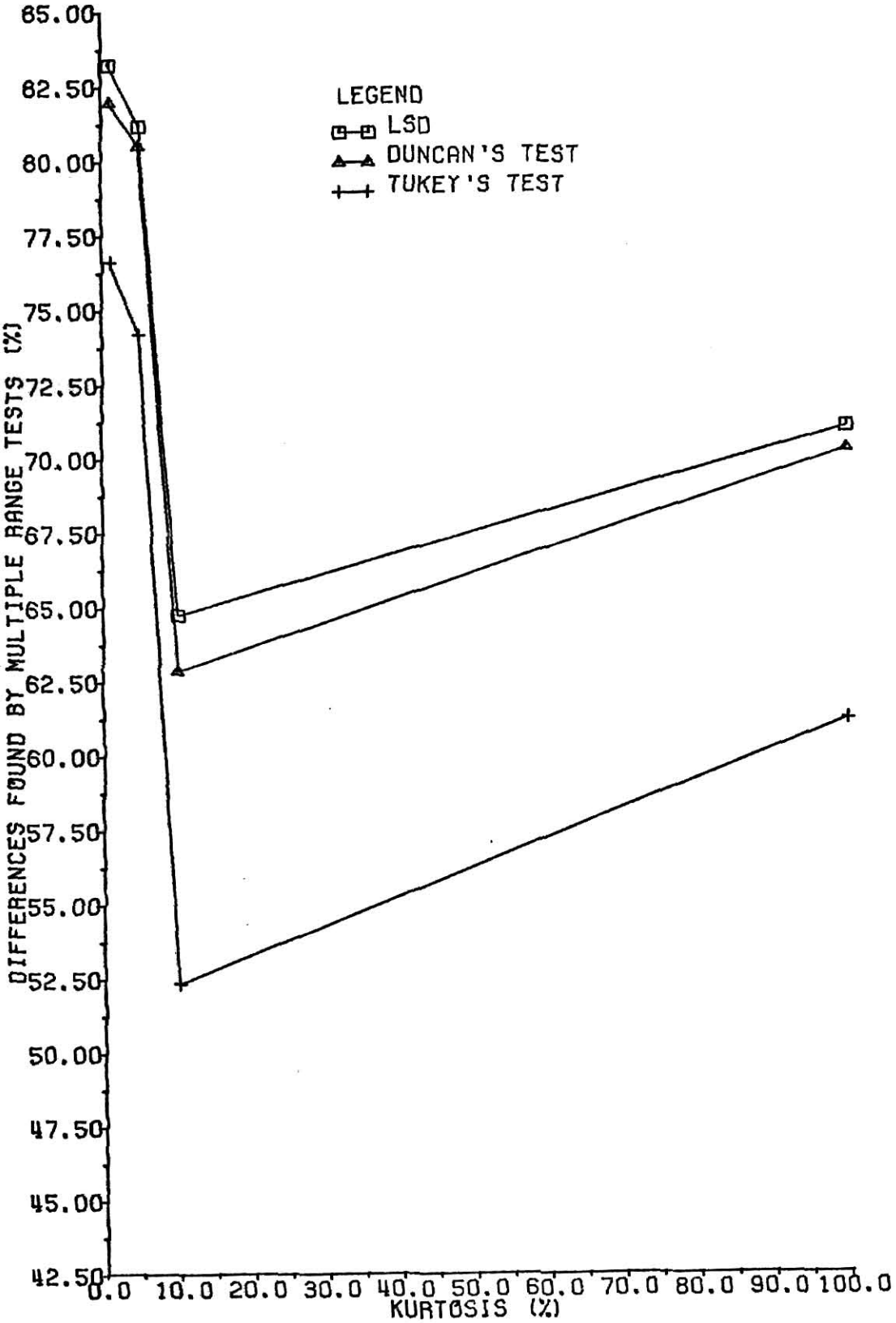


FIG. 4 PERCENT DIFFERENCES FOUND BY MULTIPLE RANGE TESTS AT VARIOUS ALPHA HATS FOR CORRELATION

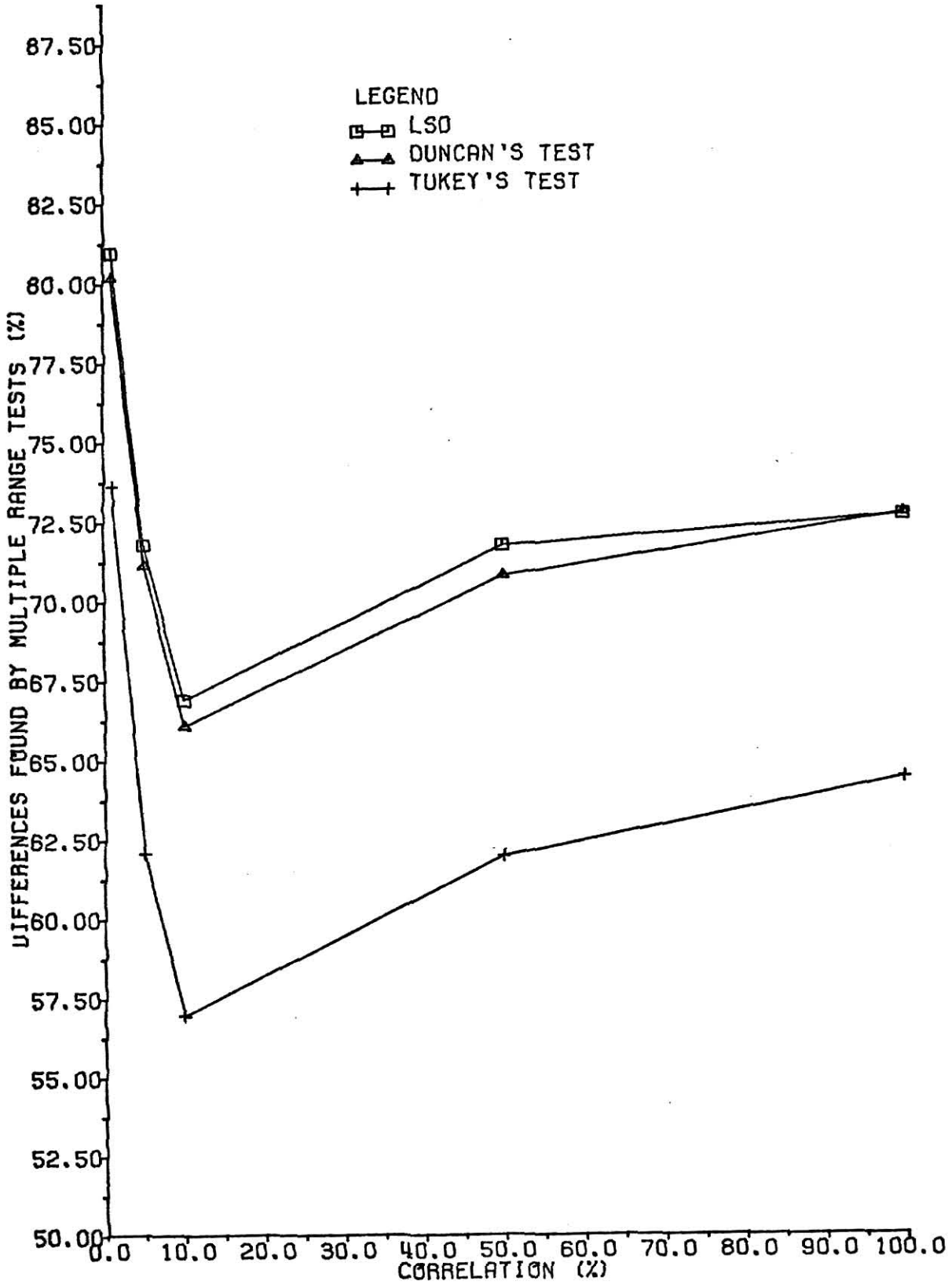
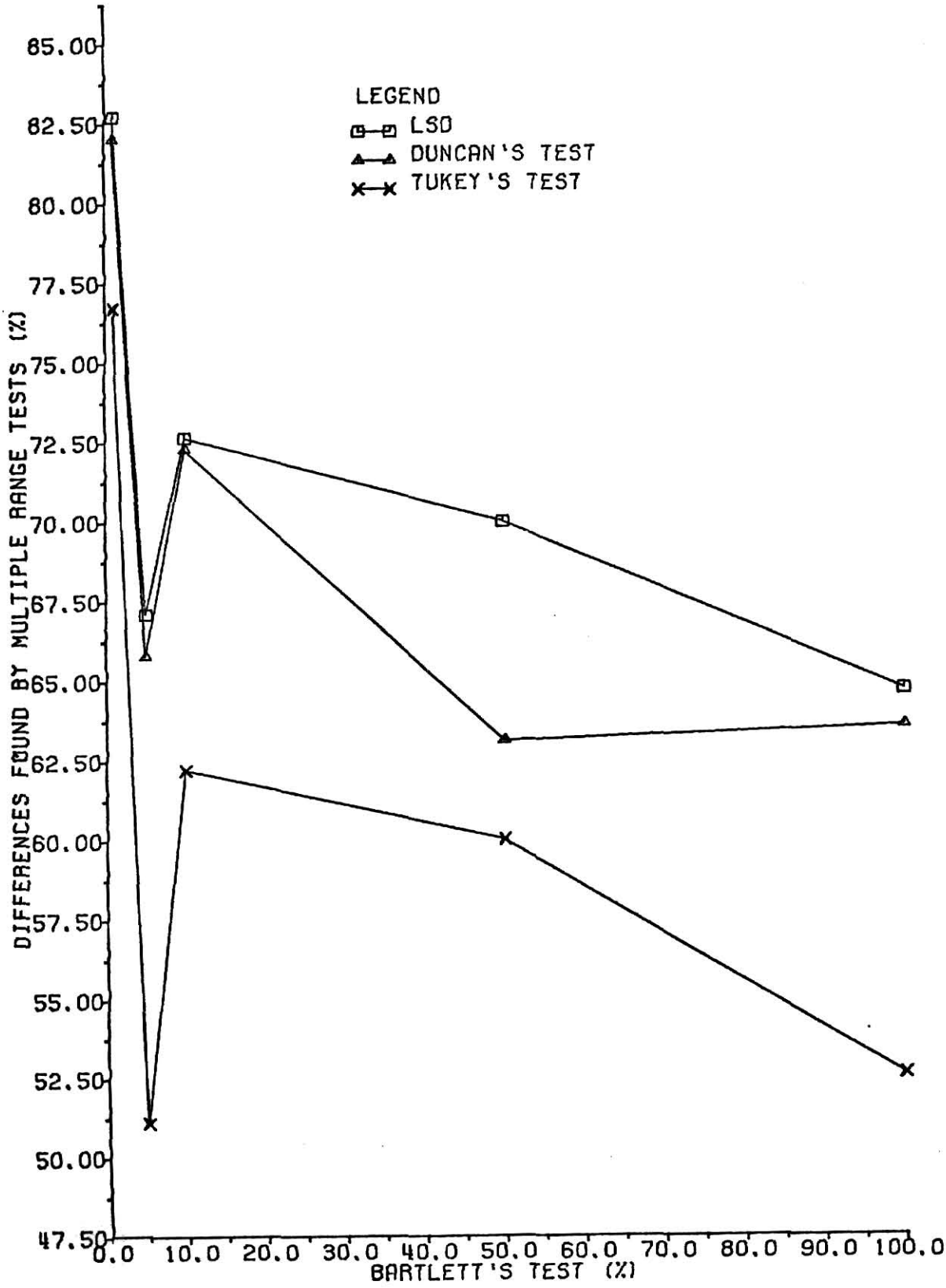


FIG. 5 PERCENT DIFFERENCES FOUND BY MULTIPLE RANGE TESTS AT VARIOUS ALPHA HATS FOR BARTLETT'S TEST



REFERENCES

1. Bartlett, M. S. 1937. Properties of Sufficiency and Statistical Tests. Proceedings of the Royal Society, A901, Vol. 160, 268.

2. Duncan, D. B. 1955. Multiple Range and Multiple F Tests. Biometrics, Vol. 11, 1.

3. Fisher, R. A. 1935. The Design of Experiments. Cliver and Boyd, London.

4. Ramseyer, G. C. 1973. The Robustness of the Studentized Range Statistic to Violation of the Normality and Homogeneity of Variance Assumption. American Educational Research Journal, Vol. 10, No. 3, 235.

5. Shapiro, S. S. and M. B. Wilk 1965. An Analysis of Variance Test for Normality (Complete Samples). Biometrika, Vol. 52, 591.

6. Snedecor, G. W. and W. G. Cochran 1967. Statistical Methods. The Iowa State University Press.

7. Tukey, J. W. 1953. The Problem of Multiple Comparisons. Ditto, Princeton University.

8. Waller, R. A. and D. B. Duncan 1969. A Bayes Rule for the Symmetric Multiple Comparison Problem. Journal of American Statistical Association, Vol. 64, 1484.

ACKNOWLEDGEMENTS

"We learn only from
those we love."
Goethe

I have no words to express my respect and gratitude to my major professor Dr Kenneth E. Kemp. He collected data for this study and gave me the idea for this topic. Under his inspiring guidance and because of his human approach, friendliness, tact and broad tolerance, the work on this thesis was a big joy.

My sincere thanks to the members of my committee Dr Shian-Koong Perng and Dr George Milliken for their constructive criticism of this work.

Thanks are also to Mrs Sumi Marasinghe for her excellent typing of this thesis.

DEPARTURE FROM ASSUMPTIONS OF ANALYSIS OF VARIANCE AND BEHAVIOUR OF MULTIPLE COMPARISON PROCEDURES (POWER AND ROBUSTNESS) IN REAL DATA

by

IVANKA BRANISLAV KRSTIĆ

B.S., Mechanical Engineering,
University of Belgrade, Yugoslavia, 1954

M.S., Mechanical Engineering,
University of California, Berkeley, 1972

AN ABSTRACT OF A MASTER'S THESIS

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1977

ABSTRACT

In the first part departure from assumptions underlying analysis of variance in real data was examined. Each data set (in total 1765 different data sets from eleven disciplines) analyzed by AARDVARK was subjected to three basic tests for normality: Shapiro-Wilk W-test for samples of size 50 or less, skewness for samples of size 25 or more and kurtosis for samples of size 11 or more. In addition, the simple corelation between the subclass mean and variance was also computed to determine if the cell means and variances were independent. To test homogeneity of variance Bartlett's test was applied. The main interest was to determine which kinds of non-normality are most common as well as to what extent heterogeneity of variance is present and whether departure from assumptions differs between disciplines.

In the second part actual behaviour of four multiple comparison procedures: Fisher's LSD, Duncan's New Multiple Range Test, Tukey's HSD for 20 or fewer levels and the Waller-Duncan Bayes t procedure for more than two levels of main effects only was investigated. The main objective was to determine whether performance of multiple comparison tests in real application is similar to results based on theoretical considerations and simulation results.

In the third part robustness of the four (same as in part two) multiple comparison procedures to departure from normality and heterogeneity of variance in real data (same data sets as in parts one and two) was investigated. The main interest was to determine whether or not departure from assumptions affected the performance of the multiple comparison procedures.