AN EMPIRICAL INVESTIGATION OF TUKEY'S HONESTLY SIGNIFICANT DIFFERENCE TEST WITH VARIANCE HETEROGENEITY AND UNEQUAL SAMPLE SIZES, UTILIZING KRAMER'S PROCEDURE AND THE HARMONIC MEAN

DISSERTATION

Presented to the Graduate Council of the North Texas State University in Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

By

William Lane McKinney, B.S., M.Ed.
Denton, Texas
May, 1976

This study sought to determine the effect upon Tukey's Honestly Significant Difference (HSD) statistic of concurrently violating the assumptions of homogeneity of variance and equal sample sizes. Two forms for the unequal sample size problem were investigated. Kramer's form and the harmonic mean approach were the two unequal sample size procedures studied. The study employed a Monte Carlo simulation procedure which varied sample sizes with a heterogeneity of variance condition. Four thousand experiments were generated.

The simulation procedure involved setting the number of treatment levels (k=4), variance ratio condition (VC=1:2:3:4), and sample sizes (5<n<30). Eight sample size patterns were used and were 5, 5, 30, 30; 5, 10, 20, 30; 5, 10, 25, 30; 5, 15, 25, 30; 30, 30, 5, 5; 30, 20, 10, 5; 30, 25, 10, 5; and 30, 25, 15, 5. Each sample had a mean of zero. The respective samples had a variance magnitude of 1, 2, 3, and 4. For each sample, a set of criterion variables, $X_{ij}$, was
computer-generated, using the model

\[ X_{ij} = \mu + \tau_j + \epsilon_{ij} \]

The \( \epsilon_{ij} \) terms were randomly generated by the computer, using a random number generator consisting of IBM subroutines Randu and Gauss. Data points obtained for the given sample sizes were used to calculate the two unequal sample size forms. Four HSD values were computed for each set of data. Two nominal significance levels, .05 and .01, were used for both HSD forms (HSDK for the Kramer approach and HSDH for the harmonic mean approach), thus yielding four HSD values. This process was repeated 500 times, once for each experiment. With 500 experiments run for each sample size pattern, a total of 4,000 experiments was generated. For each sample size pattern, the familywise Type I (FWI) error rate was computed. The obtained FWI, "actual significance level," was then compared to the expected rate, "nominal significance level," by use of a 95% confidence limit about the nominal level.

Findings of this study were based upon the empirically obtained significance levels. Five conclusions were reached in this study. The first conclusion was that for the conditions of this study the Kramer form of the HSD statistic is not robust at the .05 or .01 nominal level of significance.

A second conclusion was that the harmonic mean form of the HSD statistic is not robust at the .05 and .01 nominal level of significance. A general conclusion reached from
all the findings formed the third conclusion. It was that
the Kramer form of the HSD test is the preferred procedure
under combined assumption violations of variance hetero-
genecity and unequal sample sizes.

Two additional conclusions are based on related findings. The fourth conclusion was that for the combined
assumption violations in this study, the actual significance
levels (probability levels) were less than the nominal sig-
nificance levels when the magnitude of the unequal variances
were positively related to the magnitude of the unequal
sample sizes. The fifth and last conclusion was that for
the concurrent assumption violation of variance hetero-
genecity and unequal sample sizes, the actual significance
levels significantly exceed the nominal significance levels
when the magnitude of the unequal variances are negatively
related to the magnitude of the unequal sample sizes.
# TABLE OF CONTENTS

**LIST OF TABLES** .................................................. iv

**LIST OF ILLUSTRATIONS** ........................................ vi

## Chapter

### I. INTRODUCTION .............................................. 1

- Statement of the Problem
- Purpose of the Study
- Hypotheses
- Structural Model
- Mathematical Models of the Statistic
- Definition of Terms
- Limitations
- Basic Assumptions

### II. SURVEY OF RELATED RESEARCH ............................ 13

- Robustness Studies
- Multiple Comparison Procedures
- The Tukey Test
- Summary

### III. PROCEDURES .............................................. 32

- Model Validation Study
- Simulation Plan
- Specific Procedures
- Random Number Generator
- Data Generation
- Summary

### IV. ANALYSIS OF DATA AND FINDINGS ....................... 43

- Hypotheses Investigation
- Related Findings

### V. SUMMARY, CONCLUSIONS, IMPLICATIONS, AND RECOMMENDATIONS ........................................ 55

### APPENDIX ................................................................ 61

### BIBLIOGRAPHY .................................................. 93

iii
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Sample Size Patterns</td>
<td>35</td>
</tr>
<tr>
<td>2. 95% Confidence Limits for Proportions Corresponding to Nominal Significance Levels</td>
<td>41</td>
</tr>
<tr>
<td>3. Comparison of Actual Significance Levels for Familywise Type I Error Rates to Nominal Significance Levels in Simulated Experiments on the HSDK Test with Variance Condition 1:2:3:4</td>
<td>44</td>
</tr>
<tr>
<td>4. Comparison of Actual Significance Level for Familywise Type I Error Rates to Nominal Significance Levels in Simulated Experiments on the HSDH Test with Variance Condition 1:2:3:4</td>
<td>49</td>
</tr>
<tr>
<td>5. Percentage Points of the Studentized Range for the Study</td>
<td>65</td>
</tr>
<tr>
<td>6. Chi Square Test for the Normal Distribution of a Population Based 100,000 Pseudorandom Numbers</td>
<td>70</td>
</tr>
<tr>
<td>7. Chi Square Goodness of Fit Test for Runs Above and Below the Mean Based on 50,000 Pseudorandom Numbers</td>
<td>72</td>
</tr>
<tr>
<td>8. Kolmogorov-Smirnov Goodness of Fit Test for the Lagged Product Statistic to the Normal Distribution</td>
<td>74</td>
</tr>
<tr>
<td>9. Comparison of Actual Significance Levels for FWI Error Rates to Nominal Significance Levels in Simulated Experiments for the HSDK and HSDH Statistics</td>
<td>84</td>
</tr>
<tr>
<td>10. Frequencies of Significant HSDK Tests and F-ratios under the Same Simulated Conditions of Variance Condition 1:2:3:4 and Unequal Sample Sizes for the .05 Nominal Level of Significance</td>
<td>88</td>
</tr>
<tr>
<td>Table</td>
<td>Page</td>
</tr>
<tr>
<td>-------</td>
<td>------</td>
</tr>
<tr>
<td>11. Frequencies of Significant HSDK Tests and F-ratios under the Same Simulated Conditions of Variance Condition 1:2:3:4 and Unequal Sample Sizes for the .01 Nominal Significance Level</td>
<td>89</td>
</tr>
<tr>
<td>12. Frequencies of Significant HSDH Tests and F-ratios under the Same Simulated Conditions of Variance Condition 1:2:3:4 and Unequal Sample Sizes for the .05 Nominal Significance Level</td>
<td>91</td>
</tr>
<tr>
<td>13. Frequencies of Significant HSDH Tests and F-ratios under the Same Simulated Conditions of Variance Condition 1:2:3:4 and Unequal Sample Sizes for the .01 Nominal Significance Level</td>
<td>92</td>
</tr>
</tbody>
</table>
LIST OF ILLUSTRATIONS

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Overall Experimental Design (All Sample Sizes in a Variance Ratio of 1:2:3:4)</td>
<td>37</td>
</tr>
<tr>
<td>2. Actual Significance Levels for FWI Error Rates Corresponding to .05 Nominal Significance Level for HSDK and HSDH Forms</td>
<td>52</td>
</tr>
<tr>
<td>3. Actual Significance Levels for FWI Error Rates Corresponding to .01 Nominal Significance Level for HSDK and HSDH Forms</td>
<td>53</td>
</tr>
</tbody>
</table>
CHAPTER I

INTRODUCTION

Educational researchers frequently have a need for a multiple comparison procedure to compare groups two at a time. Although a multiple comparison may be applied on either an a priori or an a posteriori basis, of particular interest is the a posteriori situation in which a significant $F$-ratio in an analysis of variance indicates a significant difference among $k$ groups with $k > 2$. One of several multiple comparison procedures considered appropriate in determining the location of differences among groups is Tukey's Honestly Significant Difference (HSD) procedure (Kirk, 1968).

The HSD procedure specified the familywise Type I (FWI) error rate at $\alpha$ for a family of tests on all possible pairs of means, permitting the error rate per comparison to decrease as $k$ increases. Petrinovich and Hardych (1969) indicated that little had been published on the characteristics and properties of Tukey's HSD. Although Petrinovich and Hardych attempted to provide this information, Games (1971) stated they "provide limited evidence that it [HSD test] is robust to non-normality with n's of 5 or 18, but further studies are needed" (p. 543).
This study was designed to provide additional evidence of the HSD test's robustness or lack of it. When populations differ with respect to variances and it may not be possible to sample an equal number of subjects from each population, yet the means are equal, statistical tests designed to determine the mean difference may be influenced by the difference in variances and unequal sample sizes. That influence may cause the statistical test to yield more or less significant results by chance than would be expected. Significant departures from expected results determined by the FWI error rate at $\alpha$ when the means are equal is evidence that the results are being influenced by the unequal variances and unequal sample sizes.

A study of robustness involves systematically creating differences in parameters other than the parameter for which the statistical test was designed to test a difference. The two variables of concern in this study were those relating to the assumptions of homogeneity of variance and equal sample sizes. This research was performed to determine the robustness of the HSD test in the presence of variance heterogeneity with unequal sample sizes. Unequal sample sizes were handled with the harmonic mean approach and Kramer's form for unequal sample sizes.
Statement of the Problem

The problem of this study was the effects of concurrently violating the assumptions of homogeneity of variance and equal sample sizes upon Tukey's Honestly Significant Difference procedure.

Purpose of the Study

The purpose of this study was to empirically analyze Tukey's Honestly Significant Difference test under concurrent assumption violations of heterogeneity of variance and unequal sample size, utilizing two proposed solutions for unequal samples sizes: (a) Kramer's solution and (b) harmonic mean approach.

Hypotheses

The following hypotheses were formulated to carry out the purpose of this study. Variance condition (VC) was given as ratios and corresponded respectively to the group sample size (n).

1. Utilizing Kramer's procedure, actual significance levels will not differ significantly from nominal significance levels at the .05 level of significance for

   (a) VC = 1:2:3:4 and n's = n_1, n_2, n_3, n_4 such that 5 \leq n_1 \leq n_2 \leq n_3 \leq n_4 \leq 30, and

   (b) VC = 1:2:3:4 and n's = n_1, n_2, n_3, n_4 such that 30 \geq n_1 \geq n_2 \geq n_3 \geq n_4 \geq 5.
2. Utilizing Kramer's procedure, actual significance levels will not differ significantly from nominal significance levels at the .01 level of significance for

(a) VC = 1:2:3:4 and n's = n₁, n₂, n₃, n₄ such that

\[ 5 \leq n₁ \leq n₂ \leq n₃ \leq n₄ \leq 30, \] and

(b) VC = 1:2:3:4 and n's = n₁, n₂, n₃, n₄ such that

\[ 30 \geq n₁ \geq n₂ > n₃ \geq n₄ \geq 5. \]

3. Utilizing the harmonic mean, actual significance levels will not differ significantly from nominal significance levels at the .05 level of significance for

(a) VC = 1:2:3:4 and n's = n₁, n₂, n₃, n₄ such that

\[ 5 \leq n₁ \leq n₂ < n₃ \leq n₄ \leq 30, \] and

(b) VC = 1:2:3:4 and n's = n₁, n₂, n₃, n₄ such that

\[ 30 \geq n₁ \geq n₂ > n₃ \geq n₄ \geq 5. \]

4. Utilizing the harmonic mean, actual significance levels will not differ significantly from nominal significance levels at the .01 level of significance for

(a) VC = 1:2:3:4 and n's = n₁, n₂, n₃, n₄ such that

\[ 5 \leq n₁ \leq n₂ < n₃ \leq n₄ \leq 30, \] and

(b) VC = 1:2:3:4 and n's = n₁, n₂, n₃, n₄ such that

\[ 30 \geq n₁ \geq n₂ > n₃ \geq n₄ \geq 5. \]

The subdivision of the hypotheses into sub-hypothesis (a) and sub-hypothesis (b) allowed an investigation when the sample sizes were positively related to the variance condition and negatively related to the variance condition, respectively.
Structural Model

The structural model for this study was presented by Winer (1971, pp. 160–167) and was the structural model for a single-factor experiment. The population of measures (scores) was normal in form, with parameters $\mu_j$ and $\sigma_j^2$. The mean of each treatment group (sample) was designated as $\mu_j$, where the subscript $j$ denoted the treatment group. In this case with a fixed model design there were four treatment groups ($k = 4$). The variance $\sigma_j^2$ corresponded to the appropriate group mean $\mu_j$.

The effect of treatment $j$, $\tau_j$, was defined as the difference between the mean for treatment $j$ and the grand mean of the population means. Mathematically this was stated,

$$\tau_j = \mu_j - \mu.$$  

(1)

where $\tau_j = \text{effect of treatment } j$, $\mu_j = \text{mean for group } j$, $\mu = \text{grand mean for the population means}$.

If the criterion measure on a randomly selected element $i$ in a treatment population $j$ was designated as $X_{ij}$, then the structural model assumed for this study was

$$X_{ij} = \mu + \tau_j + \varepsilon_{ij}$$

(2)

where $\mu = \text{grand mean of treatment population}$, $\tau_j = \text{effect of treatment } j$, $\varepsilon_{ij} = \text{experimental error}$.
The element \( \mu \) of the model was constant for all measures in all treatment populations, while the effect \( \tau_j \) was constant for all measures within population \( j \). However, the true null condition was programmed into the simulation procedure such that \( \tau_1 = \tau_2 = \tau_3 = \tau_4 \) (i.e., no significant difference among the treatments). The experimental error \( \varepsilon_{ij} \) represented all the uncontrolled sources of variance affecting individual measurements. This effect was independent of \( \tau_j \) and unique for each element \( i \). Experimental error, \( \varepsilon_{ij} \), was normally distributed with a mean of zero and a variance of \( \sigma^2_\varepsilon \).

In summary, the simulated data for a single experiment were generated from the structural model:

\[
X_{ij} = \mu + \tau_j + \varepsilon_{ij} : i = 1, 2, \ldots, n \text{ and } j = 1, 2, \ldots, k
\]

where \( X_{ij} \) = simulated observation for the \( i \)th replication of the \( j \)th treatment,

\( \mu \) = grand mean of the \( k \) treatments,

\( \tau_j \) = effect of treatment \( j \) with the restrictions that \( \Sigma \tau_j = 0 \),

\( \varepsilon_{ij} \) = experimental error which is distributed normally with mean zero and finite variance \( \sigma^2_\varepsilon \).

The values of \( \varepsilon_{ij} \) were selected from a normal distribution with a zero mean and a unit variance. These values were generated by a random number generator discussed later in both Chapter III and Appendix C.
Mathematical Models of the HSD Statistic

The HSD statistic was mathematically defined by Kirk (1968, p. 88) as

\[ HSD = q_{\alpha, \nu} \sqrt{\frac{MS_{\text{error}}}{n}} \]  

where \( HSD \) = the value to be exceeded for a comparison involving two means to be declared significant,
\( q_{\alpha, \nu} \) = the value determined by entering a table for the percentage points of the studentized range with \( \nu \) degrees of freedom corresponding to the \( MS_{\text{error}} \) term degrees of freedom, \( \alpha \) level of significance, and the number of treatment groups (levels) in the experiment or range of levels in the experiment,
\( MS_{\text{error}} \) = an estimate taken from the one-way analysis of variance mean square within group of the experiment,
\( n \) = the sample size of each group.

If the difference between two groups exceeded the HSD value, then the results were declared significant at the given \( \alpha \) level.

Alternative models based on the above model have been proposed for conditions of unequal sample sizes. One of two models of special interest for unequal sample sizes has been
developed by Kramer (1956) for use with Duncan's Multiple Range procedure. Steel and Torrie (1960) suggested it as a solution for the unequal sample size problem with the HSD statistic. Adapting Kramer's approach to Kirk's model yielded

\[
\text{HSD}_K = q_{\alpha, \nu} \sqrt{\frac{\text{MS}_{\text{error}}}{2}} \left( \frac{1}{n_i} + \frac{1}{n_j} \right)
\]  

(5)

where HSD\(_K\) = the value to be exceeded for a comparison involving two means to be declared significant using Kramer's approach for unequal sample sizes,

\( q_{\alpha, \nu} \) = the value determined by entering a table for the percentage points of the studentized range with \( \nu \) degrees of freedom corresponding to the MS\(_{\text{error}}\) term degrees of freedom, \( \alpha \) level of significance, and the range of levels in the experiment,

MS\(_{\text{error}}\) = the mean square within group value of the experiment based on a one-way analysis of variance,

\( n_i \) = the sample size of the \( i \)th treatment group involved in the comparison, and

\( n_j \) = the sample size of the \( j \)th treatment group involved in the comparison.
The subscripts $i$ and $j$ are variables corresponding to the two levels being contrasted such that $i \neq j$.

The harmonic mean procedure for unequal groups has been discussed by Kirk (1968, p. 90) and Winer (1971, p. 216). The harmonic mean, $\tilde{n}$, was defined as

$$\tilde{n} = \frac{k}{\frac{1}{n_1} + \frac{1}{n_2} + \ldots + \frac{1}{n_k}}$$

where $\tilde{n}$ = the harmonic mean,

$k$ = the number of levels in the experiment,

$n_1, n_2, \ldots, n_k$ = the sample sizes of the respective treatment levels.

Algebraic substitution of $\tilde{n}$ for $n$ in the HSD statistical model yielded an HSDH statistical model for the harmonic mean. Algebraic manipulation yielded the parsimonious form

$$HSDH = q_{\alpha, v} \sqrt{\frac{MS_{error}}{k} \left( \frac{1}{n_1} + \frac{1}{n_2} + \ldots + \frac{1}{n_k} \right)}$$

where $HSDH$ = the value to be exceeded for a comparison involving two means to be declared significant using the harmonic mean unequal $n$ procedure,

$q_{\alpha, v}$ = the value determined by entering a table for the percentage points of the studentized range with $v$ degrees of freedom corresponding to that of the $MS_{error}$ term, number of treatment levels, and $\alpha$ level of significance,
\( k = \) the number of treatment levels in the experiment, \\
\( n_1, n_2, \ldots, n_k = \) the sample sizes of the respective treatment levels.

All three statistics--HSD, HSDK, and HSDH--yield identical results when the sample sizes of the treatment levels of the experiment are equal. However, unequal results were produced when unequal sample sizes occurred in the experiment.

In summary, only the two procedures for unequal sample sizes, HSDK and HSDH, were employed in this study. In both cases, the \( q_{\alpha,\nu} \) term was constant. Only the radicand was unique to the respective procedures. This resulted from the two distinct approaches of computing an estimate for the sample size.

**Definition of Terms**

**Actual significance level**--The actual significance level is the percentage of the computed statistical values obtained which exceed the tabled value of the statistic in an empirical distribution.

**A posteriori comparison**--An a posteriori comparison is applied following a significant F-ratio.

**A priori comparison**--A comparison formulated prior to inspection of the data is an a priori comparison.
Error rate--An error rate is a probability that erroneous conclusions will be reached for one of three units: per comparison, per experiment, and familywise.

Family--A family is a unit which corresponds to all comparisons among means associated with a single factor.

Familywise error rate--A familywise error rate is an error rate that is the ratio of the number of families with at least one statement (comparison) falsely declared significant to the total number of families.

Monte Carlo simulation--A procedure in which random samples are drawn from populations having specified parameters and then a given statistic is computed. This process is repeated until an empirical sampling distribution of the statistic is obtained. This distribution is then compared to the theoretical distribution to determine possible deviations from the theoretical distribution.

Nominal significance level--The nominal significance level is that percentage of the computed statistical values obtained which exceed the tabled value of the statistic for the theoretical distribution.

Robust--A statistical model is robust to an assumption violation when that violation does not seriously affect the results of the statistic.

Significant difference between nominal and actual significance levels--An actual significance level is said to be statistically different from a nominal significance level
when the actual significance level fails to fall within a 95% confidence interval about the nominal significance level.

Type I error—Error that occurs when the experimenter rejects the null hypothesis when it is true is a Type I error.

Limitations

This study was limited to experimental conditions simulated with the following conditions:

1. Only four treatment groups were considered.
2. Selected unequal sample sizes ranging from 5 to 30 were employed.
3. Heterogeneous variance magnitudes were selected in the ratio, 1:2:3:4.

Basic Assumptions

It was assumed that data produced by the random number generator utilized in this study and sampled through a Monte Carlo simulation procedure were not different from data that might be collected by an educational researcher.
CHAPTER II

SURVEY OF RELATED RESEARCH

The purpose of this chapter is to present the background and significance of robustness studies and, in particular, the HSD statistic. An overview of studies of robustness form the first section of this chapter. Second, the Tukey test and knowledge of its utilization is presented. Finally, a summary to provide the necessary continuity concludes this chapter.

Robustness Studies

The question of a statistic's robustness surfaced about 1930. Two of the earlier works on robustness were those of Pearson (1931) and Bartlett (1935). They studied the effect of non-normality on the F-test and t-test, respectively.

Scheffé (1959) noted robustness studies "cannot be exhaustive, . . . because assumptions . . . can be violated in many more ways than they can be satisfied" (p. 361). This is verified by the quantity of robustness studies subsequent to the 1930's and continuing at present. Each study has attempted to determine the effect of a specific assumption violation or a set of assumption violations.

Approaches to a robustness study have been either a mathematical analysis or an empirical investigation. In the
empirical studies, which were generally Monte Carlo type simulations, an empirical sampling distribution was obtained for comparison with a theoretical sampling distribution to determine whether or not there was any degree of departure. The mammoth chore of hand calculating was a deterrent to conducting empirical studies until the advent of computers. Lindquist (1956) cites one monumental study of robustness among empirical investigations by Norton, who computed 43,332 F-tests before the availability of computers.

The F-test with the analysis of variance has been investigated (Atiqullah, 1962; Norton, 1952 [found in Lindquist]; Pearson, 1931; and Scheffé, 1959); this has led to a conclusion of a remarkable robustness. Determination of robustness for the analysis of variance (ANOVA) has had a marked effect upon the procedures followed by researchers using ANOVA in the analysis of their respective data. This result of robustness for the analysis of variance has led to similar questions concerning the assumptions for multiple comparison procedures.

**Multiple Comparison Procedures**

Hypotheses about mean differences from a set of $k$ means ($k \geq 3$) may produce a multiple comparison problem. The analysis of variance provided results of a possible significant difference but did not demonstrate which contrast was significantly different. Kirk (1968) stated the analysis of
variance

is equivalent to a simultaneous test of the hypothesis that all possible comparisons among means are equal to zero. . . . If an over-all test of significance using an F-ratio is significant, an experimenter can be certain that some set of orthogonal comparisons contains at least one significant comparison among means. . . . It remains for an experimenter to carry out follow-up tests [multiple comparisons] to determine what has happened (p. 73).

Originally, the solution to the problem of determining the location of a significant difference was to use a multiple t-test (Games, 1971, p. 537). Although this procedure was extremely powerful, it allowed FWI error rate to increase as the number of t-tests increased, usually resulting in an unacceptable rate.

To counterbalance the high FWI rate, alternative methods were developed. Games (1971) reported twelve different procedures for considering multiple comparisons among means. The more popular simultaneous multiple comparison procedures Games discussed were the multiple t-test, Tukey's procedure, Scheffé's least significant difference test, Bonferroni's t statistic, and Dunnett's test. Games also included sequential multiple comparison procedures, among which were the Newman-Keuls test and Duncan's multiple range test.
That multiple comparison techniques are one of the more controversial areas of statistics was confirmed by Petrinovich and Hardyck (1969) when they stated,

Textbook authors--at least in the area of psychological statistics--have not been particularly helpful. Authors such as Edwards (1960), Federer (1955), Hays (1963), McNemar (1952), and Winer (1962) either offer no evaluation as to which method is preferable, or preface their remarks with a cautionary statement to the effect that these methods are still under study and that mathematical statisticians are not entirely in agreement concerning the preferred method. Similarly, disagreement exists as to when these methods may be used. Some discussions state that a significant $F$ ratio over all conditions must be obtained before multiple comparison methods can be used; other discussions make no mention of such a requirement, or deny that it is necessary at all (p. 44).

Perhaps a most thorough presentation of multiple comparison techniques was found by combining the works of Games (1971), Kirk (1968), and Miller (1966).
Since Tukey's procedure (a) controls FWI error rate, (b) allows a precise determination of error rate, and (c) is frequently used by researchers, additional information gleaned from an empirical investigation would be relevant. Also, additional information might affect the use of the HSD procedure in educational research.

The Tukey Test

Tukey has two popular multiple comparison procedures. Winer (1971, p. 198) refers to the more popular procedure as Tukey A. Tukey A has had several labels. It has been known as the T-Method (Glass & Stanley, 1970; Scheffé, 1959), the HSD test (Winer, 1971; Kirk, 1968; Games, 1971), and the WSD test (Games, 1971). It appears that Tukey's A procedure is best known as the honestly significant difference test (HSD). Games (1971) generally referred to the HSD as the Wholly Significant Difference procedure (WSD), quoting Miller as his source. Preferring WSD as opposed to HSD, Games perceives "Whole" of mnemonic value in reminding the researcher that the Type I error rate of this procedure is familywise. A possible point of disagreement was observed between Games and Kirk in Kirk's statement, "The WSD test merits consideration but is more complex than the HSD test" (1968, p. 90). Kirk's statement yielded an interpretation that the WSD test and HSD test are two unique procedures. The statistic considered in this study is that found in Kirk (1968, p. 88).
Tukey's HSD has been used for making all possible pairwise comparisons that involve three or means. This test utilized the studentized range statistic, q. Tukey extended an approach originally suggested by Fisher to control FWE error rate (Winer, 1971). This procedure is what has been called the HSD test. It was first proposed in a privately circulated monograph by J. W. Tukey in 1953. In "The Problem of Multiple Comparison,"* Tukey presented the most extensive discussion of the logical problems of multiple comparison at that time (Ryan, 1959). Basic assumptions of the HSD are normality, homogeneity of variance, randomization, and equal sample sizes.

Games (1971) indicated that the effect of violating the assumptions of the HSD test is relatively unknown. The studies presented below, in chronological order, provide information regarding the HSD statistic.

To better appreciate and understand the problems involved in utilizing a multiple comparison procedure, Ryan (1959) presented two general issues in multiple comparisons. These were (a) a priori versus a posteriori comparisons and (b) the concept of error rate.

Ryan defined the a priori test as one in which "the experimenter states in advance all possible conclusions and

* A letter from J. W. Tukey (see Appendix F) stated this 407-page monograph is essentially unavailable.
the rules by which these conclusions will be drawn." The a posteriori test, according to Ryan, is one in which comparisons are suggested by data. This type of comparison is also known as a post-mortem comparison. Ryan concluded that the differences between the a priori and the a posteriori comparisons are slight or even nonexistent.

There are several kinds of error rates involved in multiple comparisons, according to Ryan. Kirk (1968), with a more thorough presentation, defined six kinds of error rates: (a) error rate per comparison, (b) error rate per hypothesis, (c) error rate per experiment, (d) error rate experimentwise, (e) error rate per family, and (f) error rate familywise. Kirk noted "that the various error rates are all identical for an experiment involving a single comparison. The error rates become more divergent as the number of comparisons and hypotheses evaluated in an experiment are [sic.] increased" (p. 83). The error rates per family and familywise "are applicable to complex analyses of variance" (Kirk, p. 85). Tukey conceptualized "familywise" as the error rate unit for the HSD statistic.

Ryan (1959), in his second footnote, stated that in the one-dimensional case, "per family" and "per experiment" and "familywise" and "experimentwise" are equivalent terms. The consequence of this equivalency was that in a one-way analysis of variance Tukey's term "family" and "familywise" took on the more simple definition of "experiment" and
"experimentwise," respectively. Kirk defined error rate per experiment (family in the one-dimension case) as the ratio of the "number of comparisons falsely declared significant to the total number of experiments" (p. 84) and error rate experimentwise (familywise in the one dimensional case) as the ratio of the "number of experiments with at least one statement falsely declared significant to the total number of experiments" (p. 84). Also, Kirk defined error rate per comparison as the ratio of the "number of comparisons falsely declared significant to the total number of comparisons" (p. 82).

Ryan's conclusion to the error rate problem has been updated by recent developments. For example, Kirk concluded that

once an experimenter has specified an error rate and has decided on an appropriate conceptual unit for error rate, he can compute the corresponding rate for any other conceptual unit. Basically, the problem facing an experimenter is that of choosing, prior to the conduct of an experiment, a test statistic that provides the kind of protection desired (p. 86).

Kirk's definitions, developed into an illustration, yielded the following: Suppose there were 100 experiments each with 10 statements of significance, 1,000 statements in all. Of all possible statements, 9 were false among 7 of
the experiments. Three different error rates were

1. Error rate per comparison: \(\frac{9}{1,000} = .009\).

2. Error rate per experiment (family): \(\frac{9}{100} = .09\).

3. Error rate experimentwise (familywise): \(\frac{7}{100} = .07\).

These error rates become more divergent as the number of statements (comparisons) per experiment increase.

Petrinovich and Hardyck (1969) studied seven multiple comparison procedures for Type I and Type II error rates with a computer simulation procedure. In their study they violated assumptions of equal sample size, normality, and homogeneity of variance. Five conclusions reached were (a) sample sizes reduced from \(n = 30\) to \(n = 15\) or \(n = 5\), under equal sample size condition have no change in error rates; (b) unequal sample size of \(n = 5, 10, \) and \(15\) (\(k = 3\)) have no appreciable change in error rates; (c) use of an exponential distribution did not change the error rates; (d) combining unequal sample sizes and unequal variances produced drastically incorrect values for all seven procedures; (e) regardless of the number of sample, the Tukey A, Tukey B, and Newman-Keuls maintained a stability at the .01 and .05 levels.

Games (1971) disputed Petrinovich and Hardyck's conclusions on the effects of unequal variances and unequal sample sizes. In particular, Games was critical that Petrinovich and Hardyck only used two sample sizes with three groups.
Smith (1971) conducted an empirical analysis of Tukey's HSD procedure using three different techniques for unequal sample sizes. The three techniques compared were (a) Kramer's method, (b) harmonic mean, and (c) a median value of group size. These three approaches were analyzed to determine their effect upon Type I error. Smith used a 99% confidence band for the nominal levels of significance to compare the actual levels of significance.

Smith concluded that Kramer's and the harmonic mean methods were in the acceptable range, with the Kramer method being the more conservative. He concluded that the Tukey HSD test was insensitive to violation of the equal sample size assumption and that the Kramer method should be used for unequal group sizes. His closing statement was, "This study indicates that even under severe conditions of unequal sample sizes, the test is robust." Unfortunately, this statement generalizes to conditions not included in his study. Other variables need to be studied, such as various combinations of assumption violations of normality, equal sample sizes, and homogeneity of variance.

Howell and Games (1973a) conducted a computer simulation study to determine the effects of variance heterogeneity on three multiple comparison procedures, the HSD, Scheffé, and multiple $t$ tests. There were four levels of treatment ($k = 4$) with equal sample sizes of five. Six pairwise contrasts ($+1, -1$) were considered.
For the HSD statistic they concluded that it was as robust to heterogeneous variances as was the analysis of variance. Under the conditions of heterogeneous variances and equal sample sizes, the HSD statistic produced inflated familywise Type I error rates, limited to a maximum of 2α or 3α.

Howell and Games (1973b) investigated the HSD statistic with a computer simulation of four level (k = 4) experiments at the .05 level of significance. Two factors in this study were variance condition and sample size. The harmonic mean was used with unequal sample sizes. The variance ratios were (a) 4:4:4:4, (b) 1:3:5:7, (c) 1.76:1.76:6.24:6.24, and (d) 2.71:2.71:2.71:7.87, while the sample sizes were three in number at (a) 5,5,5,5, (b) 3,4,8,11, and (c) 11,8,4,3.

They concluded that the HSD statistic behaved very similarly to the F test. Also, they stated that their study "clearly suggests that the use of the traditional WSD [HSD] with unequal sample sizes [harmonic mean approach] is dubious when the assumption of homogeneous population variances is violated." Finally, they posed the question, would Kramer's method produce an increase in robustness?

Keselman and Toothaker (1973) evaluated four multiple comparison techniques. Two were nonparametric multiple comparison tests, Marascuilo's test and the normal scores test. The other two were parametric procedures, the Scheffé and Tukey statistics. They investigated these four procedures
for the empirical probability of a Type I error at the .05 and .10 level, using a Monte Carlo procedure. The independent variables of their study were (a) unequal variances, (b) unequal sample sizes, (c) varying number of treatment levels, and (d) different distributions, namely, normal and exponential.

In Keselman and Toothaker's (1973) study variances were in the ratios of (a) 1:2:3:4 (k = 4), (b) 1:2:3:4:5:6 (k = 6), and (c) 1:2:3:4:5:6:7:8 (k = 8). Five combinations of variance-sample size relationships were evaluated for each population distribution. The five combinations were (a) equal observations per treatment level-equal variances, (b) equal observations per treatment level-unequal variances, (c) unequal observations per treatment level-equal variances, (d) unequal observations per treatment level-unequal variances (positively related), and (e) unequal observations per treatment level-unequal variances (negatively related). When unequal observations per treatment level was a variable, Kramer's method was used with the Tukey statistic.

Results of Keselman and Toothaker's (1973) study were (a) assumption violations of nonnormality and heterogeneity on the multiple-comparison statistics were similar to the consequences of violating assumptions of the ANOVA F test; (b) the HSD test was particularly sensitive to the assumption violation of negatively relating unequal sample sizes and unequal variances. (Type I error rates were considerably larger than theoretical \( \alpha \).)
Weaknesses in this study were that only one unequal sample size method was utilized in the study and, according to Toothaker (1975), that no statistical procedures were utilized to determine statistical differences in the Type I error rates.

Keselman, Toothaker, and Shooter (1973) investigated the harmonic mean and Kramer unequal sample size forms of the HSD statistic for Monte Carlo Type I and II errors at .05 and .10 under assumption violation conditions of unequal sample size, non-normality, and variance heterogeneity. Three experimental level conditions were used, \( k = 4, k = 6, \) and \( k = 8 \). Unequal variances were in the ratio of 1:2:3:4, 1:2:3:4:5:6, and 1:2:3:4:5:6:7:8. Non-normality was violated by an exponential distribution. Sample sizes ranged from 4 to 26.

They generated five combinations of unequal variances and unequal sample sizes for both a normal and an exponential distribution. The five combinations were (a) equal observations per treatment level--equal variances, (b) equal observations per treatment level--unequal variances, (c) unequal observations per treatment level--equal variances, (d) unequal observations per treatment level--unequal variances (positively related), and (e) unequal observations per treatment level--unequal variances (negatively related).

Keselman, Toothaker, and Shooter concluded that "the effect of non-normality and heterogeneity of variance is similar to the consequences reported for the ANOVA F test"
and that "increasing the number of treatment levels under conditions of assumption violations does not generally affect the Type I error probabilities" (p. 8). They further concluded that the Kramer form was the more appropriate when sample sizes and variances are negatively related and that the harmonic mean form was the more appropriate for a positive related variance and sample size condition.

Ramseyer and Tcheng (1973) studied three multiple comparison techniques which utilize the studentized range statistic, $q$. They were the Tukey HSD test, the Newman-Keuls test, and the Duncan multiple range test. They directed their study toward determining the effect on the Type I error rate of assumption violations on the above three statistics.

In their study, homogeneity of variance was violated with variance ratios of (a) $1:1:2$ ($k = 3$), (b) $1:1:4$ ($k = 3$), (c) $1:1:1:2:2$ ($k = 5$), and (d) $1:1:1:4:4$ ($k = 5$). Normality was violated with populations which were (a) positively skewed exponential, (b) negatively skewed exponential, and (c) rectangular. The last assumption violation was a combination of the two above, heterogeneity of variance and non-normality simultaneously.

Conclusions reached by Ramseyer and Tcheng were (a) $q$ like $t$ and $F$ withstands the violation of homogeneity of variance and normality singularly and (b) only the
violation of normality produced Type I error rates systematically lower than nominal levels. Suggestions made by Ramseyer and Tcheng were that additional research must be conducted on these popular statistics and that the effect of unequal group sizes on Type I error must be examined.

Keselman, Toothaker, and Shooter (1975) provided an evaluation of the harmonic mean and Kramer unequal sample size forms of the Tukey HSD statistic. In their study unequal sample sizes and unequal variances were combined in five relationships, which were (a) all $n_k$ equal, all $\sigma_k$ equal, (b) all $n_k$ equal, not all $\sigma_k$ equal, (c) no two $n_k$ equal, all $\sigma_k$ equal, (d) no two $n_k$ equal, not all $\sigma_k$ equal (and positively related to sample sizes), and (e) no two $n_k$ equal, not all $\sigma_k$ equal (and negatively related to sample sizes). Variance conditions were generated in the ratios of 1:1:4:4, 1:1:1:2, 1:2.5:2.5:4, and 1:2:3:4 for four sample sizes. This study was conducted only at the .05 level of significance.

Keselman, Toothaker, and Shooter (1975) found no differences due to the four patterns of variance heterogeneity and a close agreement among the unequal sample size estimates and the normal and exponential estimates. They concluded that inversely pairing unequal variances with unequal sample sizes caused the observed Type I error rates to exceed true alpha for the harmonic mean approach.
and Kramer's method; therefore, a transformation or another multiple comparison test may be in order with comparable data.

Summary

Several conclusions have been reached concerning assumption violations for the HSD statistic. The first robustness study on the HSD procedure was conducted by Petrinovich and Hardyck in 1969. The consideration of robustness for a multiple comparison procedure has been a recently developed problem for researchers. During this short time, major findings with conclusions have been proposed. Assumption violations of the HSD test which have been studied are normality, homogeneity of variances, and equal sample sizes. The violation of equal sample sizes and its combination with heterogeneity of variance have produced the most interest.

The assumption of normality was studied by Petrinovich and Hardyck (1969), Keselman and Toothaker (1973), Keselman, Toothaker, and Shooter (1973; 1975). Petrinovich and Hardyck concluded that the use of an exponential distribution did not change the error rates. Keselman and Toothaker (1973) concluded that violating the normality assumption with an exponential distribution was no more serious than a similar violation in analysis of variance F-test. Keselman, Toothaker, and Shooter (1973) confirmed the findings of Keselman and Toothaker. Both studies were conducted
at .05 and .10 error rates. However, it was observed that no statistical technique was used to determine any significant difference in the nominal versus actual error rates.

Homogeneity of variance was another assumption receiving considerable attention (Howell and Games, 1973a; Keselman and Toothaker, 1973). Howell and Games concluded the HSD statistic was as robust to this violation as was the analysis of variance F-test with FWI = .05. Keselman and Toothaker supported Howell and Games' conclusion but with an added dimension of FWI = .10.

The equal sample size assumption violation has had to depend upon a source other than Tukey for a mathematical model. Robustness to this assumption violation has been studied by Petrinovich and Hardyck (1969) and Smith (1971). Petrinovich and Hardyck concluded that unequal sample sizes of n = 5, 10, 15 have no appreciable change in error rates. Smith presented the most thorough study at that time, using three unequal forms. He concluded that the HSD test was robust under the most severe conditions and that the Kramer form was the superior of the three studied. Problems with his generalizations were presented in the section "The Tukey Test" earlier in this chapter.

Assumptions, equal sample size, and homogeneity of variance have been studied under a condition of simultaneous violation by Games (1971), Howell and Games (1973b), and Petrinovich and Hardyck (1969). Petrinovich and Hardyck
concluded that combining heterogeneity of variance and unequal sample sizes produced drastically incorrect Type I error rates. Games' conclusion disagreed with conclusions reached by Petrinovich and Hardyck. Howell and Games concluded that unequal sample sizes (harmonic mean form) and variance heterogeneity led to equivocal results.

The simultaneous violation of assumptions normality, homogeneity of variance, and equal sample sizes has been studied by Keselman, Toothaker, and Shooter (1973, 1975). Five combinations were studied under a normal and an exponential distribution. The five combinations were (a) all \( n_k \) equal, all \( \sigma_k^2 \) equal; (b) all \( n_k \) equal, unequal \( \sigma_k^2 \); (c) unequal \( n_k \), equal \( \sigma_k^2 \); (d) unequal \( n_k \), unequal \( \sigma_k^2 \) (positively related); and (e) unequal \( n_k \), unequal \( \sigma_k^2 \) (negatively related). The difference between these two articles was observed to be different variance ratios, sample sizes, and Type I error rates. The 1973 article had a conclusion that the Kramer form was appropriate for condition (e) and the harmonic mean form was appropriate for condition (d). The 1975 article concluded Type I error rates for both forms (Kramer and harmonic mean) exceeded true alpha. Consequently, a transformation or another multiple comparison procedure should be considered. Toothaker (1975) reported that no statistical procedures to determine significant deviations from the theoretical Type I error rate distribution were used.
In conclusion, the HSD statistic has been frequently utilized by researchers and additional information with regard to assumption violations, since assumptions are seldom met in total, would be relevant to and effect the continued use of the HSD procedure. Factors proposed for this study were heterogeneity of variance and unequal sample sizes with the Kramer and harmonic mean forms. Actual levels of significance were investigated at .05 and .01 nominal levels of significance, using a 95% confidence limit for proportions. This study investigated actual levels of significance for FWI error rates at .05 and .01 nominal level of significance. Contrasted to other HSD robustness studies, different levels of sample size and error rates were present. In addition, a statistical procedure of the 95% confidence limit, present only in Smith's study, was utilized. The combination of these factors yielded additional relevant information.
CHAPTER III

PROCEDURES

In order to conduct this investigation a Monte Carlo simulation procedure was employed. This investigation was concerned with two assumptions of the HSD statistic, homogeneity of variances and equal sample sizes. Kirk (1968, p. 88) listed two additional assumptions, normality and randomization, which were met in this study.

Homogeneity of variance defined by Winer (1971) was the assumption "that the sources of variance within each of the samples are essentially the same and that the variances in the corresponding populations are equal" (p. 27). The equal sample size assumption was defined as each sample selected having the same number of members. These two assumptions were concurrently violated. Therefore, conditions of heterogeneity of variances and unequal sample sizes were built into the simulation procedure to meet the purpose of this study.

Model Validation Study

Before the simulation procedure was engaged with the two assumption violations, the statistical models for the HSDK and HSDH procedures, presented in Chapter I, were
evaluated under conditions of the assumptions being met. This provided a validation study on the performance of the simulation process since no assumptions were violated.

The HSDK and HSDH statistics were computed with true null hypothesis situations, equal sample sizes, and equal variances at the .05 and .01 levels of significance. Four samples were generated, each with a mean of zero and variance of one. This process was reiterated 500 times (experiments) with equal sample sizes of 10.

Both Kramer's solution for unequal sample sizes and the harmonic mean approach for unequal sample sizes were investigated in the validation study. As no assumptions were violated in the validation study, Kramer's solution for unequal sample sizes and the harmonic mean approach for unequal sample sizes yielded identical sample size values for each experiment computed in the validation study. Results of the validation study yielded actual levels of significance for FWI error rates which were within the 95% confidence interval about the nominal significance levels for .01 and .05. These results are reported in Appendix E.

Conclusively, the mathematical models for the HSDK and HSDH statistics, presented in Chapter I, performed consistently with the theoretical expectation for the familywise Type I error rate at the .05 and .01 level. Following the determination that the mathematical models functioned as they should, data were then generated to conduct this study.
Simulation Plan

The simulation consisted of a series of experiments, 500 experiments per experimental condition, each conducted with four samples \([k = 4]\) of varying sample sizes ranging from 5 to 30. The four levels (samples) of the experiment were compared using the means \((\bar{X}_j)\) of the samples. Every pairwise contrast was made among the means, resulting in six \([k(k-1)/2]\) comparisons. The pairwise contrasts were performed separately under the two (HSDK and HSDH) comparison methods with the appropriate critical value \((q_{\alpha}, \nu)\) extracted from a table for the percentage points of the studentized range for the .05 and .01 levels of significance, making 24 rates of rejection of the null hypothesis \((Ho)\).

Eight experimental levels were designed. Under all conditions variance heterogeneity was in the ratio 1:2:3:4 for the respective levels (samples) of the experiment. Sample sizes were generated according to the eight sample patterns reported in Table 1.

These sample size patterns were developed to meet two conditions. First, in sample patterns 1, 2, 3, and 4 a positive relationship between the sample sizes and the variance condition existed. Specifically, the variances in the variance condition ratio, 1:2:3:4, increased as sample sizes \((n_j)\) increased. For instance, in Sample Pattern 2, \(n_1\) had a variance of 1 and 5 members, \(n_2\) had a
Table 1

Sample Size Patterns

<table>
<thead>
<tr>
<th>Sample Patterns</th>
<th>Sample Sizes&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n₁</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>30</td>
</tr>
<tr>
<td>6</td>
<td>30</td>
</tr>
<tr>
<td>7</td>
<td>30</td>
</tr>
<tr>
<td>8</td>
<td>30</td>
</tr>
</tbody>
</table>

<sup>a</sup>Variances are 1, 2, 3, and 4 for the respective samples.

variance of 2 and 10 members, n₃ had a variance of 3 and 20 members, and n₄ had a variance of 4 and 30 members.

Second, a negative relationship between the sample sizes and the variance condition existed with Sample Patterns 5, 6, 7, and 8. That is, the variances in the variance condition ratio, 1:2:3:4, increased as sample sizes decreased.

In summary, this study consisted of 4,000 experiments. Four groups were evaluated at two levels of significance,
.05 and .01, for both HSD statistics, HSDK and HSDH, with a variance condition ratio of 1:2:3:4 and eight unequal sample size patterns. Under all conditions sample means (\(\bar{X}_j\)) were equal, \(\bar{X}_1 = \bar{X}_2 = \bar{X}_3 = \bar{X}_4 = 0.0\). A true null hypothesis situation was generated for each experiment. For each sample size pattern 500 experiments were repeated in blocks of four, one block per HSD statistic per level of significance. A three-dimensional illustration of this study is provided in Figure 1. All probability values varied from 0.0 to 1.0.

All four hypotheses of this study were concerned with actual significance levels. To eliminate confusion associated in the notion of significance level, Kirk (1968) stated,

If an experiment contains two treatment levels with one comparison among the means, the probability of committing a Type I error if the null hypothesis is true is determined by the significance level adopted. The interpretation of a significance level is unambiguous for treatments with two treatment levels but becomes confusing for multi-treatment experiments involving several simultaneous comparisons. The confusion arises in the case of multiple comparisons because a significance level can be specified for a number of different conceptual units. For example,
Figure 1. Overall experimental design. (All sample sizes in a variance ratio of 1:2:3:4.)
the conceptual unit can be the individual comparison, hypothesis, family of comparisons, or experiment. An "error rate" can be defined for each of these conceptual units (p. 82).

The significance level for the HSD test was the conceptual unit familywise. Within this framework the familywise Type I (FWI) error rate was the significance level of concern. Therefore, the contrast rate recorded was the familywise rate (FWI) defined as the number of times at least one of the six contrasts was declared to be significant. Also, record was kept of the number of significant F-ratios in an ANOVA. This was an ANOVA in that subsequent multiple comparisons were not conditional on the ANOVA results. The ANOVA provided the mean square within term.

Specific Procedures

This investigation used a Monte Carlo simulation procedure that selected samples from a computer population, with each experiment performed through simulation. The North Texas State University Computing Center wrote for this study a specifically designed program that included the main program and two sub-routines. The program (Appendix D) was written in Fortran IV for use on the IBM 360 Model 50 Computer System installed at the North Texas State University Computing Center.
Random Number Generator

The population from which the error terms were selected was generated by a random number generator (Appendix C) which employed IBM sub-routines Randu and Gauss. The term "random number" should be understood to be "pseudorandom number." This random number generator utilized a seed number from which each pseudorandom number was generated. A seed number was introduced into the random number generator and pseudorandom numbers were yielded. Different seed numbers produced different series of pseudorandom numbers. However, the same seed number produced identical pseudorandom number sequences every time it was used. The consequence of this was that the sequences produced internally by the computer were not random in the true sense because they were completely determined by a given seed number. Naylor (1966, p. 46) stated, "a sequence may be considered random if it satisfies some predetermined set of statistical tests of randomness."

For generating random numbers, Naylor (p. 46) stated that an acceptable method "must yield sequences of numbers which are (1) uniformly distributed, (2) statistically independent, (3) reproducible, and (4) nonrepeating for any desirable length." Property (3) is intuitive from the above discussion of pseudorandom numbers. Three statistical tests were run to determine compliance with Properties (1), (2), and (4), and are reported in Appendix C. The frequency test
produced acceptable results; therefore, Property (1) was met. Property (2), tested by the lagged product test, was met. The runs test above and below the mean produced acceptable results, indicating that Property (4) was met. Therefore, according to Naylor, "these pseudorandom numbers can be treated as 'truly' random numbers even though they are not" (p. 57).

**Data Generation**

Data were generated under specified conditions. The variance ratio was set at 1:2:3:4. Each sample had a zero mean, which provided a true $H_0$ situation. Sample sizes were generated according to the sample size patterns reported in Table 1. Both HSD statistics, HSDK and HSDH, were computed for every pairwise comparison. Also, an F-ratio for an ANOVA was computed. Familywise Type I error rates were tabulated for each sample size pattern at .05 and .01 levels of significance. If the actual levels of significance for FWI error rate fell within the 95% confidence limits for proportions corresponding to nominal level of significance (Table 2), then the given statistic was declared robust. If the actual significance levels for FWI error rate were not in the 95% confidence limit, then the statistic was interpreted as not being robust under the given experimental conditions.
Table 2

95% Confidence Limits for Proportions Corresponding to Nominal Significance Levels

<table>
<thead>
<tr>
<th>Number of Sample Points</th>
<th>Limits</th>
<th>Proportions</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>Upper</td>
<td>.074</td>
<td>.025</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lower</td>
<td>.033</td>
<td>.004</td>
<td></td>
</tr>
</tbody>
</table>

aSee Appendix A for confidence limits development.

Summary

In summary, a Monte Carlo simulation procedure was utilized, where

(a) there were four samples;
(b) each sample had a mean of zero;
(c) sample variances were in the ratio of 1:2:3:4;
(d) sample sizes ranged between 5 and 30 (Table 1);
(e) there were eight sample size patterns;
(f) for each sample size pattern 500 experiments were generated;

(g) for each possible pairwise comparison within the experiment (1) the HSDK statistic was computed at the .05 and .01 levels of significance, (2) the HSDH statistic was computed at the .05 and .01 levels of significance, and (3) an ANOVA F-ratio was computed;
(h) tabulation of the familywise error rate was made and recorded; and

(i) the actual level of significance determined from FWI was compared to the nominal level of significance utilizing 95% confidence limits for proportions corresponding to the nominal significance levels.
CHAPTER IV

ANALYSIS OF DATA AND FINDINGS

The results of this investigation into the effect of combining variance heterogeneity with unequal sample sizes upon the HSD statistic is presented in two parts. The first part is the analysis of data directly related to the four hypotheses presented in Chapter I. The second part presents the analysis of related data.

Hypotheses Investigation

Data produced in this study are reported in Table 3 for the HSDK test and Table 4 for the HSDH test. Both tables were constructed to allow a comparison of the actual significance level for familywise Type I error rates to nominal significance levels of .05 and .01. The sample size pattern number (Table 1) and sample sizes are reported to facilitate interpretation of the sub-hypotheses. In all experimental levels, eight in total, the variance condition maintained was a constant ratio of 1:2:3:4. The interpretation of this was that the standard deviation of (a) the first sample in all experiments was one, (b) the second sample in all experiments was $\sqrt{2}$, (c) the third sample in all experiments was $\sqrt{3}$, and (d) the fourth sample in all experiments was $\sqrt{4}$.

43
Table 3
Comparison of Actual Significance Levels for Familywise Type I Error Rates to Nominal Significance Levels in Simulated Experiments\textsuperscript{a} on the HSDK Test with Variance Condition 1:2:3:4

<table>
<thead>
<tr>
<th>Pattern Number</th>
<th>Sample Size ((n_1, n_2, n_3, n_4))</th>
<th>Actual Significance Levels (FWI) Corresponding to Nominal Significance Levels of .05</th>
<th>.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5, 5, 30, 30</td>
<td>.016\textsuperscript{b}</td>
<td>.000\textsuperscript{b}</td>
</tr>
<tr>
<td>2</td>
<td>5, 10, 20, 30</td>
<td>.032\textsuperscript{b}</td>
<td>.002\textsuperscript{b}</td>
</tr>
<tr>
<td>3</td>
<td>5, 10, 25, 30</td>
<td>.028\textsuperscript{b}</td>
<td>.006</td>
</tr>
<tr>
<td>4</td>
<td>5, 15, 25, 30</td>
<td>.038</td>
<td>.008</td>
</tr>
<tr>
<td>5</td>
<td>30, 30, 5, 5</td>
<td>.180\textsuperscript{c}</td>
<td>.076\textsuperscript{c}</td>
</tr>
<tr>
<td>6</td>
<td>30, 20, 10, 5</td>
<td>.160\textsuperscript{c}</td>
<td>.050\textsuperscript{c}</td>
</tr>
<tr>
<td>7</td>
<td>30, 25, 10, 5</td>
<td>.150\textsuperscript{c}</td>
<td>.054\textsuperscript{c}</td>
</tr>
<tr>
<td>8</td>
<td>30, 25, 15, 5</td>
<td>.142\textsuperscript{c}</td>
<td>.042\textsuperscript{c}</td>
</tr>
</tbody>
</table>

\textsuperscript{a}500 experiments were generated for each pattern, yielding a total of 4,000 experiments.

\textsuperscript{b}Less than the lower limit of the 95\% confidence limit of the nominal significance level (see Table 2).

\textsuperscript{c}Greater than the upper limit of the 95\% confidence limit of the nominal significance level (see Table 2).
The four hypotheses developed provided direction for this study. Each hypothesis was subdivided into a sub-hypothesis (a) and (b). This subdivision contributed to the investigation of the effect of a positive or negative relationship between sample size and variance condition upon actual significance levels for FWI rates.

The first hypothesis considered the Kramer form at the .05 nominal level of significance. Hypothesis 1(a) stated that in utilizing Kramer's procedure, actual significance levels would not differ significantly from the nominal significance level at the .05 level of significance for a variance condition ratio of 1:2:3:4 and sample sizes varying between 5 and 30 such that an ascending ordinal relationship existed respectively from the first sample to the fourth sample with respect to size. This hypothesis in the general case was rejected, as evidenced in Table 3. Patterns numbered 1, 2, 3, and 4 at the .05 nominal level of significance satisfied the conditions of Hypothesis 1(a). For patterns 1, 2, and 3 the actual significance level (actual probability level) was less than the lower limit of the 95% confidence limits for the proportion corresponding to the .05 nominal significance level. The lower limit .033 was greater than .016, .032, and .028 actual significance levels for FWI rates for patterns 1, 2, and 3, respectively. Only pattern 4, with an actual significance level of .038, fell within the 95% confidence interval.
Hypothesis 1(b) stated that, utilizing Kramer's procedure, actual significance levels would not differ significantly from nominal significance levels at the .05 level of significance for variance condition 1:2:3:4 and sample sizes varying from 30 to 5 such that a decreasing ordinal relationship existed respectively from the first sample to the fourth sample with respect to size. As evidenced in Table 3, Hypothesis 1(b) was rejected. At the .05 nominal significance level patterns 5, 6, 7, and 8 fulfilled the conditions of Hypothesis 1(b). The actual significance levels for FWI rates reported in Table 3 for patterns 5, 6, 7, and 8 were .180, .160, .150, and .142, respectively. These four levels exceeded the upper limit, .074, for the 95% confidence limits for the proportion corresponding to the .05 nominal significance level.

The second hypothesis probed Kramer's form at the .01 nominal significance level. The (a) part of the second hypothesis stated that, utilizing Kramer's procedure, actual significance levels would not differ significantly from the nominal significance level at the .01 level of significance for variance condition ratio 1:2:3:4 and sample size ranging from 5 to 30 such that the size relationship among the samples was an ascending ordinal relationship from the first sample to the fourth sample. Results reported in Table 3 at the .01 nominal significance level for patterns 1, 2, 3, and 4 led to the rejection in the general case of Hypothesis 2(a). Patterns 1 and 2 with actual significance levels for
FWI rates of .000 and .002, respectively, were less than the lower limit, .004, of the 95% confidence limits for the proportions corresponding to the .01 nominal significance level. Patterns 3 and 4 yielded .006 and .008 actual significance levels, respectively. Both were within the 95% confidence limits.

Hypothesis 2(b) stated that, utilizing Kramer's procedure, actual significance levels would not differ significantly from nominal significance levels at the .01 level of significance for variance condition ratio 1:2:3:4 and sample size ranging from 30 to 5 such that the samples were in descending order from the first to the fourth sample with respect to size. This hypothesis was rejected. In Table 3, for the .01 nominal significance level with patterns 5, 6, 7, and 8, all results exceeded the upper limit, .025, for the 95% confidence limits for the proportion corresponding to the .01 nominal significance level. Patterns 5, 6, 7, and 8 yielded actual significance levels for FWI rates of .076, .050, .054, and .042, respectively.

Hypotheses 3 and 4 were both concerned with the harmonic mean form of the HSD statistic. The third hypothesis considered the .05 level of significance. Hypothesis 3(a) stated that, utilizing the harmonic mean, actual significance levels would not differ significantly from nominal significance levels at the .05 level of significance for variance condition ratio 1:2:3:4 and sample sizes ranging from 5 to
such that the samples were arranged in an ascending order with respect to size. Results reported in Table 4 apply to this hypothesis. Patterns 1, 2, 3, and 4 for the .05 nominal level of significance fulfilled the conditions of Hypothesis 3(a). Actual significance levels for FWI rates of .006, .004, .004, .010 for patterns 1, 2, 3, and 4 at the .05 nominal significance level were below the lower limit, .033, of the 95% confidence limits for the proportion corresponding to the .05 nominal significance level. Consequently, Hypothesis 3(a) was rejected.

The second part of the third hypothesis, 3(b), stated that, utilizing the harmonic mean, actual significance levels will not differ significantly from nominal significance levels at the .05 level of significance for variance condition 1:2:3:4 and sample sizes ranging from 30 to 5 such that the samples were arranged in descending order with respect to sample size from the first sample to the fourth sample. This hypothesis was rejected as evidenced in Table 4 for patterns 5, 6, 7, and 8 at the .05 nominal significance level. Patterns 5, 6, 7, and 8 yielded actual significance levels of .242, .228, .214, and .226, respectively. Each was greater than the upper limit, .074, of the 95% confidence limits for the proportion corresponding to the .05 nominal significance level.
Table 4
Comparison of Actual Significance Levels for Familywise Type I Error Rates to Nominal Significance Levels in Simulated Experiments\(^a\) on the HSDH Test with Variance Condition 1:2:3:4

<table>
<thead>
<tr>
<th>Pattern Number</th>
<th>Sample Size ((n_1, n_2, n_3, n_4))</th>
<th>Actual Significance Levels (FWI) Corresponding to Nominal Significance Levels of .05</th>
<th>.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5, 5, 30, 30</td>
<td>.006(^b)</td>
<td>.002(^b)</td>
</tr>
<tr>
<td>2</td>
<td>5, 10, 20, 30</td>
<td>.004(^b)</td>
<td>.000(^b)</td>
</tr>
<tr>
<td>3</td>
<td>5, 10, 25, 30</td>
<td>.004(^b)</td>
<td>.000(^b)</td>
</tr>
<tr>
<td>4</td>
<td>5, 15, 25, 30</td>
<td>.010(^b)</td>
<td>.002(^b)</td>
</tr>
<tr>
<td>5</td>
<td>30, 30, 5, 5</td>
<td>.242(^c)</td>
<td>.122(^c)</td>
</tr>
<tr>
<td>6</td>
<td>30, 20, 10, 5</td>
<td>.228(^c)</td>
<td>.098(^c)</td>
</tr>
<tr>
<td>7</td>
<td>30, 25, 10, 5</td>
<td>.214(^c)</td>
<td>.112(^c)</td>
</tr>
<tr>
<td>8</td>
<td>30, 25, 15, 5</td>
<td>.226(^c)</td>
<td>.106(^c)</td>
</tr>
</tbody>
</table>

\(^a\)500 experiments were generated for each pattern yielding a total of 4,000 experiments.

\(^b\)Less than the lower limit of the 95% confidence limit of the nominal significance level (see Table 2).

\(^c\)Greater than the upper limit of the 95% confidence limit of the nominal significance level (see Table 2).
The fourth hypothesis was constructed to consider the harmonic mean form at the .01 nominal significance level. Hypothesis 4(a) stated that, utilizing the harmonic mean, actual significance levels would not differ significantly from nominal significance levels at the .01 level of significance for variance condition 1:2:3:4 and sample sizes ranging from 5 to 30 such that the sample sizes were in an ascending order from the first sample to the fourth sample. Table 4 evidenced that this hypothesis should be rejected. The actual significance levels .002, .000, .000, and .002 for patterns 1, 2, 3, and 4, respectively, were less than the lower limit, .004, of the 95% confidence limits for the proportion corresponding to the .01 nominal significance level.

The second part, (b), of the fourth hypothesis stated that, utilizing the harmonic mean, actual significance levels would not differ significantly from nominal significance levels at the .01 level of significance for variance condition 1:2:3:4 and sample sizes ranging from 30 to 5 such that the sample sizes are arranged in descending order from sample one to sample four. The .01 nominal significance level for patterns 5, 6, 7, and 8 in Table 4 reported results of the conditions specified in Hypothesis 4(b). Actual significance levels of .122, .098, .112, and .106 for patterns 5, 6, 7, and 8, respectively, were outside the 95% confidence limits for the proportion corresponding to
the .01 nominal level of significance. In every case the actual levels of significance for FWI rate were greater than the upper limit, .074.

In summary, the four hypotheses, or eight sub-hypotheses, developed to direct this investigation were all rejected. For each sub-hypothesis two or more sample patterns yielded actual significance levels significantly different from the nominal significance levels. The actual significance levels were compared to the nominal significance level by the use of a 95% confidence limit corresponding to the proportion for the given nominal significance level. If the actual significance level did not fall within the 95% confidence limit, then results were declared significantly different.

**Related Findings**

Figures 2 and 3 illustrate some interesting patterns. The first observation was that for the .05 nominal significance level the discrepancy between the actual significance level and nominal significance level was always less for the HSDK form. At the .01 nominal significance level, the case was the same, with the exception of sample size pattern 1. Under this condition the discrepancy between the actual and nominal levels of significance was reversed, with the HSDH form producing results closer to the nominal significance level. However, it should be noted that the
Figure 2. Actual significance levels for FWI error rates corresponding to .05 nominal significance level for HSDK and HSDH forms.
Figure 3. Actual significance levels for FWI error rates corresponding to .01 nominal significance level for HSDK and HSDH forms.
difference at this point between the HSDK and HSDH results was extremely small.

A second observable pattern was that the discrepancy between the actual and nominal significance levels was less under conditions of a positively related sample size and variance condition than the negatively related sample size and variance condition.

The third observation was in two parts. First, when the relationship between sample sizes and variance condition ratio was positive, as in sub-hypotheses (a), the actual levels of significance were always less than the nominal levels of significance. Second, if there existed a negative relationship between sample sizes and variance condition as the sub-hypotheses (b) for the four hypotheses allowed, then the actual levels of significance were inflated, greater in every case than the nominal significance level.

An evaluation of the F-ratio from an ANOVA is presented in Appendix G. Since the HSD procedure set the risk of finding at least one significant pairwise comparison at $\alpha$ and the F-ratio set the risk of a Type I error at $\alpha$, the actual significance levels should be identical. However, discrepant results occurred.
A Monte Carlo simulation procedure was conducted to determine the effect of variance heterogeneity and unequal sample sizes upon two unequal forms of the HSD statistic. The two unequal forms were the harmonic mean (HSDH) and Kramer (HSDK) forms. Actual significance levels for family-wise Type I error rates were contrasted to corresponding .05 and .01 nominal significance levels. In each simulation the true null condition was assumed. The HSD form for a given condition was declared robust if the actual significance level fell within the 95% confidence limits for the proportion of the corresponding nominal significance level. An incidental ANOVA F-ratio was also computed to compare error rates between the F-ratio and HSD forms.

Five conclusions were reached in this study. They were direct products of the hypotheses employed in providing direction for this study.

Considering the variance condition ratio 1:2:3:4 combined with unequal sample sizes ranging from 5 to 30, Tukey's honestly significant difference procedure utilizing
the Kramer form was not robust at the .05 or .01 level of significance. This first conclusion was based on the findings from testing Hypotheses 1 and 2.

The second conclusion was based on the findings from testing Hypotheses 3 and 4. Conclusion two was that Tukey's honestly significant difference procedure utilizing the harmonic mean form with combined assumption violations of heterogeneity of variance at a variance ratio of 1:2:3:4 and unequal sample sizes ranging from 5 to 30 was not robust at the .05 and .01 levels of significance.

Smith (1971) and Howell and Games (1973a) reached conclusions different from the conclusions in this study. Smith concluded the HSD procedure was robust to unequal sample sizes. The reason conclusions one and two contradict Smith's conclusion was that his investigation did not include a combination of assumption violations. Howell and Games (1973a) reached a conclusion that the HSD statistic was robust to variance heterogeneity. Reasons for this apparent contradiction were that they (a) compared the HSD error rates to past ANOVA error rate studies, (b) limited their study to equal sample sizes, and (c) were willing to accept error rates as great as $2\alpha$ or $3\alpha$; in another perspective they did not use confidence limits.

A conclusion reached by Howell and Games (1973b) supported the first two conclusions of this study. They concluded that the use of the HSD statistic employing the
harmonic mean approach for unequal sample sizes was dubious when the assumption of homogeneous population variances was violated. The first two conclusions stated in a single summative form were that for a variance condition ratio 1:2:3:4 and unequal sample sizes ranging from 5 to 30, the Kramer and harmonic mean forms of the honestly significant difference procedure were not robust when using 95% confidence limits about a corresponding proportion for the .05 and .01 nominal significance levels.

A general conclusion derived from considering findings related to all four hypotheses was that the Kramer form (HSDK) was the preferred form to employ with the honestly significant difference procedure under combined assumption violations of variance heterogeneity and unequal sample sizes. Although neither approach was determined robust, the HSDK form yielded fewer discrepant actual significant levels from nominal significant levels across all experimental conditions of this study. Also, when a discrepancy occurred between the actual and nominal significance levels, the magnitude of the discrepancy with the HSDK form was less than the magnitude of the discrepancy with the HSDH form.

This third conclusion was consistent with one of the conclusions of Smith (1971) in that he recommended the Kramer form. Smith violated only one assumption, equal
sample sizes, whereas this study was an extension from Smith's investigation to a combination of assumption violations. Keselman, Toothaker, and Shooter (1973) reached a conclusion which both agreed and disagreed with the third conclusion in this study. They agreed when they concluded that the Kramer form was the appropriate form to be used with negatively related sample sizes and variances. However, a disagreement was observed in that they concluded that the harmonic mean form was the appropriate form to be used with positively related sample sizes and variances. The third conclusion reached in this study was that the Kramer form was the desirable alternative to be used with the honestly significant difference procedure when unequal sample sizes were present.

Conclusions number four and five were obtained from the related findings observed in the sub-hypotheses (a) and (b) of each of the four hypotheses proposed in this study. When Hypotheses 1(a), 2(a), 3(a), and 4(a) were analyzed, the fourth conclusion was reached, which was that the actual significance levels were less than the nominal significance levels when the magnitude of the unequal variances were positively correlated to the magnitude of the unequal sample sizes. The fifth conclusion derived from Hypotheses 1(b), 2(b), 3(b), and 4(b) was that the actual significance levels were greater than the nominal significance levels when the magnitudes of the unequal variances were negatively correlated
to the magnitude of the unequal sample sizes. Conclusion five was congruent with conclusions reached by Keselman and Toothaker (1973) and Keselman, Toothaker, and Shooter (1975).

There are several implications. The first implication of which an educational researcher should be aware is that if conditions of combined assumption violations exist in the form of unequal sample sizes and unequal variances, then another multiple comparison procedure may be preferred. A second implication is that if only the two HSD forms were considered with these conditions of heterogeneity of variance and unequal sample sizes, the HSDK form should be the better performing statistic. Another implication is that a researcher should be aware of the relationship which exists between the magnitude of an unequal variance condition and the magnitude of an unequal sample size condition, as a positively related pairing produced deflated actual significance levels and a negatively related pairing produced inflated actual significance levels.

Recommendations for further study are many, as this investigation posed numerous questions with the potential of developing into good research problems. The need for additional research became evident for the following areas. A direct extension of this study should be conducted which would investigate the combined assumption violations of this study at different levels. For example, additional variance condition ratios should be studied. Another study might
investigate the power of the HSDK and HSDH forms under combined assumption violations. A related area to be investigated might be the power of the sequential nature of the HSD procedure following a significant ANOVA F-ratio. Finally, the combined assumption violations of this study should be joined with a third assumption of non-normality to determine the effect upon Type I familywise error rates with the simultaneous violation of three assumptions.

In summary, five conclusions were reached with regard to the two assumption violations of this investigation for the HSDK and HSDH forms. Three implications have been presented which should affect the use of the HSD procedure when the two assumption violations in this study exist. Finally, four recommendations for further study have been made.
APPENDIX A

DEVELOPMENT OF THE 95% CONFIDENCE LIMITS FOR PROPORTIONS CORRESPONDING TO NOMINAL SIGNIFICANCE LEVELS

It was desirable to obtain maximum accuracy consistent with economic considerations. As a result of this consideration and one for accuracy, the 95% confidence limits with 500 experiments were computed by following a procedure for large numbers recommended by Brownlee (1965, p. 150). The 95% confidence limits for each of the significance levels are given in Table 2 (Chapter III). It should be noted that these limits were constructed around arbitrary values corresponding to the nominal significance levels.

The following is an example of how the 95% confidence limits about a .05 nominal significance level for 500 experiments were computed. To compute the upper limit of a proportion, Brownlee gave the formula

$$
\hat{p} = \frac{1}{n+u^2} \left\{ x + \frac{u^2 p_1}{2} - u_1 \left[ \frac{(x+\frac{1}{2})(n-x-\frac{1}{2})}{n} + \frac{u^2 p_1}{4} \right]^{1/2} \right\},
$$

where

- $n$ = the number of experiments,
- $x$ = the product of $n$ and the proportion about which the confidence limit is established,
\( p_1 = \) the proportion divided by two, \\
\( u_{p_1} = \) the normal score for \( p_1 \).

Therefore, for the upper limit of the 95% confidence limits about .05 error rate,

\[
\begin{align*}
n & = 500 \\
x & = 25
\end{align*}
\]

For .05 nominal significance level (proportion)

\[
p_1 = .025 \\
u_{p_1} = -1.96, \quad u_{p_1}^2 = 3.8416
\]

Substituting into Equation (8) yields

\[
\theta = \frac{1}{500 + 3.8416} \left\{ 25 + 0.5 + \frac{3.8416}{2} - 1.96 \left[ \frac{(25 + 0.5)(500 - 25 - 0.5)}{500} + \frac{3.8416}{4} \right]^{1/2} \right\} (9)
\]

\( \theta = 0.0739361. \)

The upper limit then was rounded off to .074.

Brownlee gave the following for the lower limit formula:

\[
\theta = \frac{1}{n + u^2} \left\{ X - \frac{1}{2} + \frac{u^2 p_2}{2} - u_{p_2} \left[ \frac{X - \frac{1}{2}}{n} \left( \frac{n - X + \frac{1}{2}}{n} + \frac{u^2 p_2}{4} \right) \right]^{1/2} \right\}, (10)
\]

where \( n = \) the number of experiments,

\( X = \) the product of \( n \) and the proportion about which the confidence limit is established,
\[ p_2 = 1 - p_1, \]

\[ u_{p_2} = \text{the normal score for } p_2. \]

Therefore, for the lower limit of the 95% confidence limits about .05 error rate,

\[ n = 500 \]
\[ X = 25. \]

For .05 nominal significance level (proportion)

\[ p_2 = .975 \]

\[ u_{p_2} = 1.96, u^2_{p_2} = 3.8416. \]

Substituting into Equation (10) yields

\[ \theta = \frac{1}{500 + 3.8416} \left\{ \frac{25 - .5 + 3.8416}{2} - 1.96 \left[ \frac{(25-.5)(500-25+.5) + 3.8416}{n} \right]^{1/2} \right\} \]  \hspace{1cm} (11)

\[ \theta = 0.0332782. \]

The lower limit was rounded off to .033.

The results were a boundary of .033 and .074 for a 95% confidence limit about a .05 error rate for 500 experiments.
APPENDIX B

Table 5 contains percentage points of the studentized range for this study. Percentage points were for conditions of a range of four at the .05 and .01 levels of significance. Degrees of freedom for the mean square error term of 30, 40, 60, and 120 were taken from Harter, Clemm, and Guthrie (1959). Degrees of freedom needed in this investigation but not available in the table of Harter et al. (1959) were interpolated from the tabular values. The interpolation precedent was performed in Kirk (1968, p. 88) and Winer (1971, p. 192). Values which resulted from interpolation were for degrees of freedom 36, 61, 66, 71, 76, and 116.
Table 5
Percentage Points of the Studentized Range for this Study

<table>
<thead>
<tr>
<th>Error df</th>
<th>( \alpha )</th>
<th>( q ) (range = 4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>.05</td>
<td>3.845</td>
</tr>
<tr>
<td></td>
<td>.01</td>
<td>4.799</td>
</tr>
<tr>
<td>36(^a)</td>
<td>.05</td>
<td>3.813</td>
</tr>
<tr>
<td></td>
<td>.01</td>
<td>4.737</td>
</tr>
<tr>
<td>40</td>
<td>.05</td>
<td>3.791</td>
</tr>
<tr>
<td></td>
<td>.01</td>
<td>4.696</td>
</tr>
<tr>
<td>60</td>
<td>.05</td>
<td>3.737</td>
</tr>
<tr>
<td></td>
<td>.01</td>
<td>4.595</td>
</tr>
<tr>
<td>61(^a)</td>
<td>.05</td>
<td>3.736</td>
</tr>
<tr>
<td></td>
<td>.01</td>
<td>4.593</td>
</tr>
<tr>
<td>66(^a)</td>
<td>.05</td>
<td>3.732</td>
</tr>
<tr>
<td></td>
<td>.01</td>
<td>4.585</td>
</tr>
<tr>
<td>71(^a)</td>
<td>.05</td>
<td>3.727</td>
</tr>
<tr>
<td></td>
<td>.01</td>
<td>4.577</td>
</tr>
<tr>
<td>76(^a)</td>
<td>.05</td>
<td>3.723</td>
</tr>
<tr>
<td></td>
<td>.01</td>
<td>4.569</td>
</tr>
<tr>
<td>116(^a)</td>
<td>.05</td>
<td>3.688</td>
</tr>
<tr>
<td></td>
<td>.01</td>
<td>4.504</td>
</tr>
<tr>
<td>120</td>
<td>.05</td>
<td>3.685</td>
</tr>
<tr>
<td></td>
<td>.01</td>
<td>4.497</td>
</tr>
</tbody>
</table>

\(^a\)Interpolated values.
APPENDIX C

STATISTICAL TESTS ON THE RANDOM NUMBER GENERATOR

The random number generator used in this study consisted of a sequence of two IBM subroutines Gauss and Randu. Gauss computed a normally distributed random number with a given mean and standard deviation. Randu computed uniformly distributed random real numbers between zero and one and random integers between zero and $2^{31}$. In order to produce one normal random observation, Gauss utilized Randu to generate twelve uniform random numbers.

Gauss used the formula

$$y = \frac{\sum_{i=1}^{k} \left( x_i - \frac{k}{2} \right)}{\sqrt{\frac{k}{12}}}$$

(12)

where $x_i$ was a uniformly distributed random number, $0 < x_i < 1$, which was supplied by the subroutine Randu. $k$ was the number of values $x_i$ utilized in computing one normally distributed random number. $Y$ approached a true normal distribution asymptotically as $k$ approached infinity. However, satisfactory results were produced when $k$ equals 12. If $k$ equals 12, Equation (12) simplified to
\[ Y = \sum_{i=1}^{12} (X_i - 6.0). \]  

The results of Equation (13) was a normal random number. This normal random value was then adjusted to meet the heterogeneous variance conditions of this study. An adjusted value was computed by

\[ Y' = Y \cdot S \]  

where \( Y' \) was the required normally distributed random number, \( Y \) was the value produced by Equation (13) and \( S \) was the standard deviation required by the given variance heterogeneity.

Three statistical tests, as outlined by Naylor (1966), were computed on the results of the random number generator discussed above. The frequency test was used to check the uniformity of the values generated. Results of this test indicated whether the numbers were "truly" random numbers. The test of runs above and below the mean was employed to test the oscillatory nature of sequences. Satisfactory results of these two statistical tests on Gauss implied that Randu, employed by Gauss, functioned acceptably. The third statistical test used was the lagged product test computed on Randu. This test served as a measure of the independence of the pseudorandom numbers.
Frequency Test

The frequency test was used on 100 sets of 1,000 random numbers. Each set was generated by Gauss with a normal distribution, a mean of zero, and a standard deviation of one. Results of each set of 1,000 numbers were placed in equi-probable intervals. The expected number of random numbers in each interval was 100. Intervals were determined by using the unit normal distribution with a mean of zero. A chi-square goodness of fit was completed on the actual frequencies in the equi-probable intervals. The formula utilized to compute the chi-square was

$$\chi^2 = \sum_{j=1}^{10} \frac{(f_j - 100)^2}{100},$$  \hspace{1cm} (15)

where \( f_j \) was the observed frequency in the equi-probable interval \( j \) and 100 was the expected frequency in each interval.

Naylor (1966) indicated that if the entire sequence of 100,000 random numbers is composed of "truly" random observations on a variable normally distributed with a mean of zero and a standard deviation of one, then the 100 \( \chi^2 \)'s would have approximately a chi-square distribution with nine degrees of freedom. Therefore, to test for a "truly" random distribution another chi-square goodness of fit was computed using the formula
where $F_j$ was the observed frequency of $\chi_1^2$'s in equi-probable interval; and $10$ was the expected frequency in each interval. These equi-probable intervals were determined from a chi-square distribution with nine degrees of freedom. Each of the $\chi_1^2$'s from Equation (15) was placed in one of the equi-probable intervals. Results of the 100 $\chi_1^2$'s yielded the observed frequency $F_j$, noted in Equation (16). The results of this test are reported in Table 6.

A chi-square value of 13.00 was obtained. A critical value of 16.92 at the .05 level of significance with nine degrees of freedom was necessary for the rejection of the null hypothesis. Therefore, no significant difference was found. The interpretation of this finding was that the 100,000 pseudorandom numbers could be from the unit normal distribution with a mean of zero. The frequency test of Gauss yielded satisfactory results for the utilization of Gauss and Gauss' use of Randu for this simulation study.
Table 6
Chi-Square Test for the Normal Distribution of a Population
Based 100,000 Pseudorandom Numbers

<table>
<thead>
<tr>
<th>Equi-probable Interval</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Observed</td>
</tr>
<tr>
<td>0.00 - 4.17</td>
<td>5</td>
</tr>
<tr>
<td>4.18 - 5.38</td>
<td>10</td>
</tr>
<tr>
<td>5.39 - 6.39</td>
<td>7</td>
</tr>
<tr>
<td>6.40 - 7.36</td>
<td>12</td>
</tr>
<tr>
<td>7.37 - 8.34</td>
<td>5</td>
</tr>
<tr>
<td>8.35 - 9.41</td>
<td>17</td>
</tr>
<tr>
<td>9.42 - 10.66</td>
<td>9</td>
</tr>
<tr>
<td>10.67 - 12.24</td>
<td>10</td>
</tr>
<tr>
<td>12.25 - 14.68</td>
<td>11</td>
</tr>
<tr>
<td>14.69 - greater</td>
<td>14</td>
</tr>
</tbody>
</table>

\[ \chi^2_F (9) = 13.00, \ p = .16. \]

**Runs Test**

Naylor (1966) stated, "the random oscillatory nature of sequences of pseudorandom numbers can be tested by 'tests of runs'" (p. 60). A test of runs above and below the mean was computed as outlined by Naylor. The subroutine Gauss, with a unit normal distribution and mean of zero, generated 100 samples (sequences) of 500 random numbers. For each of the 100 sequences, runs were counted.
A run was defined as a subsequence of numbers in which each term was either above or each term was below the mean. For the purpose of this runs test, the mean was zero. Consequently, a run was a subsequence of consecutive numbers less than zero or consecutive numbers greater than zero. To complete definitions needed for this runs test, the length of the run was defined as the number of elements (numbers) in a run.

Runs of length 1, 2, 3, \ldots, 7, and 8 or greater were counted for the 100 sequences generated. Naylor (p. 61) gave the formula for the expected frequency of runs for each length as \((N - k + 3)2^{-k-1}\), where \(k\) equaled the length of the run and \(N\) equaled the number of terms of the sequence.

On each of the 100 sequences a chi-square goodness of fit statistic was computed on the observed frequency and the expected frequency of runs of the eight specified lengths. Then, the 100 chi-square values were tested by a chi-square goodness of fit utilizing the 10 equi-probable intervals noted in Table 7.

The \(\chi^2\) of 10.40 with nine degrees of freedom was not significant. As noted earlier, with nine degrees of freedom at .05 level of significance, a \(\chi^2\) of 16.92 or greater was needed to declare significance. Therefore, the interpretation of the results was that no significant difference existed between the observed frequency distribution and the expected frequency distribution. This test indicated a satisfactory
Table 7
Chi-Square Goodness of Fit Test for Runs Above and Below the Mean Based on 50,000 Pseudorandom Numbers

<table>
<thead>
<tr>
<th>Equi-probable Interval</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Observed</td>
</tr>
<tr>
<td>0.00 - 4.17</td>
<td>14</td>
</tr>
<tr>
<td>4.18 - 5.38</td>
<td>14</td>
</tr>
<tr>
<td>5.39 - 6.39</td>
<td>6</td>
</tr>
<tr>
<td>6.40 - 7.36</td>
<td>13</td>
</tr>
<tr>
<td>7.37 - 8.34</td>
<td>8</td>
</tr>
<tr>
<td>8.35 - 9.41</td>
<td>7</td>
</tr>
<tr>
<td>9.42 - 10.66</td>
<td>7</td>
</tr>
<tr>
<td>10.67 - 12.24</td>
<td>10</td>
</tr>
<tr>
<td>12.25 - 14.68</td>
<td>7</td>
</tr>
<tr>
<td>14.69 - greater</td>
<td>14</td>
</tr>
</tbody>
</table>

\[ \chi^2 (9) = 10.400, \ p = .3203. \]

random oscillatory nature of the sequence of pseudorandom numbers generated by Gauss.

**Lagged Product Test**

The third and final statistic on the random number generator was the lagged product test on the subroutine Randu. This test is a measure of the independence of the pseudorandom numbers. Randu generated 500 sets (sequences)
of 100 random numbers uniformly distributed with a mean of zero and a standard deviation equal to one.

For each sequence of 100 random numbers a lagged product statistic $C_k$ was computed for $k = 1, 2, 3, \ldots, 10$, with

$$C_k = \frac{1}{N-k} \sum_{i=1}^{N-k} r_i (r_{i+k})$$

(17)

where $k$ was the length of lag, $N$ was the number of random numbers per sequence, $r_i$ was the $i$th term of the sequence, and $r_{i+k}$ was the term $k$ terms of the sequence beyond the $i$th term. Naylor (p. 59) stated, "It can be shown that if there is no correlation between $r_i$ and $r_{i+k}$, the values of $C_k$ will be approximately distributed normally with expected value equal to 0.25 and standard deviation equal to $\sqrt{\frac{13N - 19k}{12(N-k)}}$ for $k > 0$." Each $C_k$ value was converted to $G_k$, a standardized unit normal value, with a mean of zero; therefore,

$$G_k = \frac{C_k - 0.25}{\sqrt{\frac{1300 - 19k}{12(100 - k)}}}$$

(18)

Consequently, if the $G_k$'s for a given $k$ were normally distributed with a mean of zero and a standard deviation of one, then there was no correlation between $r_i$ and $r_{i+k}$. For each lag ($k$) the 500 values of $G_k$ were tested for a unit normal distribution with a mean of zero by the Kolmogorov-Smirnov goodness of fit test. Table 8 displays the results.
Table 8
Kolmogorov-Smirnov Goodness of Fit Test for the Lagged Product Statistic to the Normal Distribution

<table>
<thead>
<tr>
<th>Lag (k)</th>
<th>D</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.037872</td>
<td>.4701</td>
</tr>
<tr>
<td>2</td>
<td>.025038</td>
<td>.9126</td>
</tr>
<tr>
<td>3</td>
<td>.023953</td>
<td>.9365</td>
</tr>
<tr>
<td>4</td>
<td>.026586</td>
<td>.8715</td>
</tr>
<tr>
<td>5</td>
<td>.029897</td>
<td>.7928</td>
</tr>
<tr>
<td>6</td>
<td>.020401</td>
<td>.9854</td>
</tr>
<tr>
<td>7</td>
<td>.028487</td>
<td>.8119</td>
</tr>
<tr>
<td>8</td>
<td>.023106</td>
<td>.9523</td>
</tr>
<tr>
<td>9</td>
<td>.023144</td>
<td>.9516</td>
</tr>
<tr>
<td>10</td>
<td>.024732</td>
<td>.9198</td>
</tr>
</tbody>
</table>

Using the formula Siegel (1956, p. 48) gave to determine the critical value for the Kolmogorov-Smirnov test with a sample size of 500 at the .05 level of significance, a maximum deviation, D, of .0608209 was needed to reject the null hypothesis. The null hypothesis in this situation was that no significant difference existed between the observed frequency and the normal distribution. The results of the goodness of fit tests on all 10 lags indicated that retention of the null hypothesis in each situation was in order.
Therefore, the interpretation was that the subroutine Randu yielded independent pseudorandom numbers.
APPENDIX D

THE COMPUTER PROGRAM
REAL*8 Pgef, F
INTEGER N(4,8), NN(4)
INTEGER GP(4)
REAL R(4), XM(4), SD(4)
REAL DIFF(6)
REAL XT(31)
DATA R (1, 2, 3, 4)
DATA N/ 5, 5, 5, 5, 10, 10, 30, 30, 5, 5, 5, 5, 30, 20, 10, 20, 30, 30, 30, 25, 10, 5, 30, 25, 15, 15/7
NG = 4
NS = 500
IX = 99920343
KKK = 1
NP = 8
DO 9999 LP = KKK, NP
IT105 = 0
IC105 = 0
IF105 = 0
ID105 = 0
IE105 = 0
IT205 = 0
IC205 = 0
IF205 = 0
ID205 = 0
IE205 = 0
IT101 = 0
IC101 = 0
IF101 = 0
ID101 = 0
IE101 = 0
IT201 = 0
IC201 = 0
IF201 = 0
ID201 = 0
IE201 = 0
HM = 0
DO 7 I = 1, NG
NN(I) = N(I, LP)
HM = HM + 1.0 / FLOAT(NN(I))
7 CONTINUE
HM = 4.0 / HM
WRITE (6, 700) LP, NN, HM, IX
700 FORMAT('1', PATTERN, 'R2.5X', 'S = ', '4 R2.5X', 'IX = ', 'I11'/)
ICT = 0
DO 999 LS = 1, NS
ICT = ICT + 1
IF (ICT .LE. 5) GO TO 8
WRITE (6, 3000)
3000 FORMAT('1')
ICT = 1
8 CONTINUE
SW = 0
SW2 = 0
SSWB = 0
TN = 0
IX1 = IX
DO 3 I=1,NG
M = N(I,LP)
NN(I) = M
XN = M
RB = SQRT(R(I))
GP(I) = I
SX = 0
SX2 = 0
TN = TN + XN
DO 5 J=1,M
A = 0
DO 6 K=1,12
6 A = A + RANDME(Ix, Y)
X = (A - 6.) * BB
XT(J) = X
XX = X * X
SW2 = SW2 + XX
SX = SX + X
SX2 = SX2 + XX
SW = SW + X
5 CONTINUE
MM = M + 1
XT(MM) = 99999.
C WRITE (8,2000) (XT(KT),KT=1,MM)
C2000 FORMAT(F12.6)
C WRITE (6,1000) (XT(KT),KT=1,M)
C1000 FORMAT(10',15F8.4,' ',15F8.4)
TSX = (SX * SX) / XN
SSWR = SSWB + TSX
XM(I) = SX / XN
SD(I) = SQRT((SX2 - TSX) / (XN - 1.))
3 CONTINUE
SW = (SW * SW) / TN
SSR = SSWR - SW
SSW = SW2 - SSWR
SST = SW2 - SW
DFW = TN - 40.
DFT = TN - 10.
XMSR = SSR / 30.
XMSW = SSW / DFW.
F = XMSR / XMSW
P = PPRF(3, [FIX(DFW), F])
WRITE (6,20) XM, SO, IX1
20 FORMAT(' MEANS',4F10.6,'3X', 'STDEV',4F10.6,5X,'IX=',I11)
WRITE (6,30) SSR, XMSR, F, P
30 FORMAT('0BETWEEN',5X,'3o',4F10.4)
WRITE (6,40) DFW, SSW, XMSW
40 FORMAT(' WITHIN',F8.0,2F10.4)
WRITE (6,50) DFT, SST
50 FORMAT(' TOTAL',F9.0,F10.4)
CALL SORTEM (XM, GP, NN, NG, 1, 3, NG)
L = 0
DO 10 I=1,3
II = 5 - I
TXM = XM(II)
K = II - 1
DO 10 I=1,K
GAIN
DATE = T76019

FIFF(L) = ABS(TXM - XM(J))

ALPHA = .05
Q = QTAB(ALPHA, NFW)
CALL TUKFY1(XMSW, Q, DIFF, NN, GP, ICOUNT)
PS = .05
IF (ICOUNT .EQ. 0) GO TO 201
IT105 = IT105 + 1
IC105 = IC105 + 1
IF (P .LE. PS) GO TO 301
IF105 = IF105 + 1
GO TO 401

301 CONTINUE
IE105 = IF105 + 1
GO TO 401

201 CONTINUE
IF (P .LE. PS) ID105 = ID105 + 1

401 CONTINUE
CALL TUKFY2(XMSW, Q, DIFF, HM, GP, ICOUNT)
IF (ICOUNT .EQ. 0) GO TO 202
IT205 = IT205 + 1
IC205 = IC205 + 1
IF (P .LE. PS) GO TO 302
IF205 = IF205 + 1
GO TO 402

302 CONTINUE
IC205 = IC205 + 1
GO TO 402

202 CONTINUE
IF (P .LE. PS) ID205 = ID205 + 1

402 CONTINUE
ALPHA = .01
PS = .01
Q = QTAB(ALPHA, NFW)
CALL TUKFY1(XMSW, Q, DIFF, NN, GP, ICOUNT)
IF (ICOUNT .EQ. 0) GO TO 203
IT101 = IT101 + 1
IC101 = IC101 + 1
IF (P .LE. PS) GO TO 303
IF101 = IF101 + 1
GO TO 403

303 CONTINUE
IE101 = IF101 + 1
GO TO 403

203 CONTINUE
IF (P .LE. PS) ID101 = ID101 + 1

403 CONTINUE
CALL TUKFY2(XMSW, Q, DIFF, HM, GP, ICOUNT)
IF (ICOUNT .EQ. 0) GO TO 204
IT201 = IT201 + 1
IC201 = IC201 + 1
IF (P .LE. PS) GO TO 304
IF201 = IF201 + 1
GO TO 404

304 CONTINUE
IF201 = IF201 + 1
GO TO 404
204 CONTINUE
   IF (DF.LT. PS) IN201 = IC201 + 1
404 CONTINUE
   WRITE (6,500)
500 FORMAT(' ',100(*----------*'))
999 CONTINUE
   EXP = 1000/N5
   COMP = 1000/(6 *N5)
* C
   PA105 = IT105 * COMP
   PA205 = IT205 * COMP
   PA101 = IT101 * COMP
   PA201 = IT201 * COMP
* C
   PX105 = IT105 * FXP
   PX205 = IT205 * FXP
   PX101 = IT101 * FXP
   PX201 = IT201 * FXP
* C
   PB105 = IC105 * FXP
   PB205 = IC205 * FXP
   PB101 = IC101 * FXP
   PB201 = IC201 * FXP
   WRITE (6,200) IT105, IC105, IF105,
    * ID105, IF105, PA105, PX105, PB105,
    * IT205, IC205, IF205,
    * ID205, IF205, PA205, PX205, PB205,
    * IT101, IC101, IF101,
    * ID101, IF101, PA101, PX101, PB101,
    * IT201, IC201, IF201,
    * IC201, IF201, PA201, PX201, PB201
200 FORMAT(1G0000.5F00.4/,1G0000.5F00.4/,1G0000.5F00.4/,1G0000.5F00.4/)
9999 CONTINUE
   STOP
   END
SUBROUTINE TUKEY
EXTERNAL XM(1), HSDF(6)
INTEGER NN(1), GP(1), JGP(6), JGP(6)
INTEGER * 2 AST, BLNK, FLAG(6)
DATA AST, BLNK /' * ', '*'/'
ENTRY TUKEY1(XMSW, Q, XM, NN, GP, ICOUNT)
ICOUNT = 0
L = 0
XMSWT = 0.5 * XMSW
DO 3 I = 1, 3
II = 5 - I
TNI = NN(I)
K = II - 1
DO 3 J = 1, K
TNJ = NN(J)
L = L + 1
HSDF(J) = Q * SQRT(XMSWT * ((1.0 / TNI) + (1.0 / TNJ)))
FLAG( ) = BLNK
IF(XM(L) .LE. HSDF(L)) GO TO 13
FLAG( ) = AST
ICOUNT = ICOUNT + 1
3 CONTINUE
GO TO 99
ENTRY TUKEY2(XMSW, Q, XM, HM, GP, ICOUNT)
ICOUNT = 0
HSDD = Q * SQRT(XMSW / HM)
L = 0
DO 4 I = 1, 3
II = 5 - I
K = II - 1
DO 4 J = 1, K
L = L + 1
HSDF(J) = HSDD
FLAG( ) = BLNK
IF(XM(L) .LE. HSDD) GO TO 14
FLAG( ) = AST
ICOUNT = ICOUNT + 1
14 JGP(L) = GP(I)
JGP(L) = GP(J)
4 CONTINUE
99 CONTINUE
WRITE (6,10) (JGP(L), JGP(L), HSDF(L), FLAG(L), I = L + 1, 6), Q
10 FORMAT( ' ', 6(2I2, F9.4, A2, 5X), 'Q= ', F7.4)
RETURN
END
FUNCTION QTAB(ALPHA, DFW)
C
SOURCE FOR TABLE VALUES - KIRK
REAL TABLE(7,6), XN(6)
REAL LOW, HIGH, QTAB
DATA TABLE/ 32.82, 164.3, 3.845, 4.799, 3.791, 4.656, 3.737, 4.595, 3.685, 4.497, 3.633, 4.433 /
DATA XN/ 10, 30, 40, 60, 120, 99999 /
I = 1
IF(ALPHA .LT. .05) I = 2
DO 1 J = 2, 5
IF(DFW .LE. XN(J)) GO TO 2
1 CONTINUE
J = 6
2 CONTINUE
C
USE J-1 AND J
J1 = J - 1
LOW = TABLE(I, J1)
HIGH = TABLE(I, J)
QTAB = LOW + (HIGH - LOW) / (XN(J) - XN(J1)) * (DFW - XN(J1))
RETURN
END
The two HSD models, HSDK (Kramer form) and HSDH (harmonic mean form), presented in Chapter I were evaluated with model assumptions met. The purpose of this evaluation was to empirically analyze the HSDK and HSDH under conditions when assumptions were satisfied. The actual significance levels were compared to the nominal significance levels by observing the discrepancy between the observed proportions and the theoretically expected proportions for the nominal values of .05 and .01.

Assumptions of normality and randomization were programmed into the simulation model. The two assumptions of interest in this study, homogeneity of variance and equal sample sizes, were met with the variance condition programmed in the ratio of 1:1:1:1 for the four samples per experiment and equal samples of 10 yielding a condition of 10, 10, 10, 10 for the four samples. In this evaluation 500 experiments were generated.

Ryan (1968) concluded that the notion of significance level was clarified for multiple comparison procedures by the concept of error rate. Since the appropriate error rate for the HSD procedure was familywise, actual level of
significance was tabulated for familywise Type I (FWI) error rates at the nominal levels of significance of .05 and .01. Results of the validation study are reported in Table 9.

Table 9
Comparison of Actual Significance Levels for FWI Error Rates to Nominal Significance Levels in Simulated Experiments for the HSDK and HSDH Statistics

<table>
<thead>
<tr>
<th>Variance Condition Ratio</th>
<th>Sample Sizes ( (n_1, n_2, n_3, n_4) )</th>
<th>HSD Statistic</th>
<th>Actual Significance Levels (FWI) Corresponding to Nominal Significance Levels of .05</th>
<th>.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:1:1:1</td>
<td>10, 10, 10, 10</td>
<td>HSDK</td>
<td>.060</td>
<td>.014</td>
</tr>
<tr>
<td>1:1:1:1</td>
<td>10, 10, 10, 10</td>
<td>HSDH</td>
<td>.060</td>
<td>.014</td>
</tr>
</tbody>
</table>

A consequence of equal sample sizes was that the sample size estimates in the Kramer and harmonic mean forms were equal, in this case 10. All results were within the 95% confidence interval for the given nominal significance level. The interpretation was that when assumptions for the HSD statistic were met the unequal forms, HSDK and HSDH, behaved as expected as there was no significant departure of actual significance levels for FWI rates from nominal significance levels.
This study served as a verification of the simulation model and simulation process as the discrepancies between the actual and nominal significance levels were well within the 95% confidence limits. This indicated that the simulation process performed accurately.
Mr. William McKinney  
3615 San Rafael Drive  
Arlington, Texas 76013  

Dear Mr. McKinney:  

Copies of "The problem of multiple comparisons" are essentially unavailable. A xerox copy of multiple comparisons can be had at a cost of 10¢/page (407) plus postage (book rate).  

If you so desire, upon receipt of your check or money order in the amount of $42.58, payable to John Tukey, the xerox copy of multiple comparisons will be made and then mailed to you.  

Thank you for your interest.  

Sincerely,  

John W. Tukey
APPENDIX G

COMPARISON OF F-RATIO AND HSD RESULTS

An evaluation of the F-ratio from an ANOVA provided results presented in Tables 10, 11, 12, and 13. As noted earlier (p. 16) in a quotation from Petrinovich and Hardyck there exist contradictory opinions among statisticians as to whether or not multiple comparisons should follow an omnibus F-ratio. In an a priori investigation the omnibus F-ratio of ANOVA is not necessarily computed prior to making multiple comparisons. Hypotheses stated prior to data collection provide direction for contrasts. Whereas, in an a posteriori investigation multiple comparisons are made only after a significant F-ratio is found. In this situation statistical findings rather than hypotheses provide direction for study.

Tables 10 and 11 report findings for the comparison of the HSDK test and ANOVA F-ratios. Table 10 presents results at the .05 nominal level of significance, while Table 11 is for the .01 nominal significance. At both nominal levels of significance, significant F-ratios were found without a significant HSDK contrast. Also, significant HSDK contrasts were found in an experiment not yielding a significant F-ratio. A conclusion from these findings is that with
Table 10

Frequencies of Significant HSDK Tests and F-ratios Under the Same Simulated Conditions of Variance Condition 1:2:3:4 and Unequal Sample Sizes for the .05 Nominal Level of Significance

<table>
<thead>
<tr>
<th>Sample Size (n₁, n₂, n₃, n₄)</th>
<th>Frequency of Significant F-ratios</th>
<th>HSDK Tests</th>
<th>$F \notin HSDK^a$</th>
<th>$HSDK \notin F^b$</th>
<th>$F \cap HSDK^c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5, 5, 30, 30</td>
<td>8</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>5, 10, 20, 30</td>
<td>10</td>
<td>16</td>
<td>3</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>5, 10, 25, 30</td>
<td>11</td>
<td>14</td>
<td>1</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>5, 15, 25, 30</td>
<td>20</td>
<td>19</td>
<td>3</td>
<td>2</td>
<td>17</td>
</tr>
<tr>
<td>30, 30, 5, 5</td>
<td>96</td>
<td>90</td>
<td>12</td>
<td>6</td>
<td>84</td>
</tr>
<tr>
<td>30, 20, 10, 5</td>
<td>94</td>
<td>80</td>
<td>17</td>
<td>3</td>
<td>77</td>
</tr>
<tr>
<td>30, 25, 10, 5</td>
<td>74</td>
<td>75</td>
<td>6</td>
<td>7</td>
<td>68</td>
</tr>
<tr>
<td>30, 25, 15, 5</td>
<td>76</td>
<td>71</td>
<td>9</td>
<td>4</td>
<td>67</td>
</tr>
</tbody>
</table>

$^a_F \notin HSDK$ notes frequency of significant F-ratios without a significant HSDK result.

$^b_HSDK \notin F$ notes frequency of at least one significant HSDK test without a significant F-ratio.

$^c_F \cap HSDK$ notes frequency of significant F-ratios when at least one significant HSDK test occurred.
Table 11
Frequencies of Significant HSDK Tests and F-ratios Under the Same Simulated Conditions of Variance Condition 1:2:3:4 and Unequal Sample Sizes for the .01 Nominal Level of Significance

<table>
<thead>
<tr>
<th>Sample Size ( (n_1, n_2, n_3, n_4) )</th>
<th>Frequency of Significant</th>
<th>( F \cap ) HSDK ( ^a )</th>
<th>HSDK ( \cap ) F</th>
<th>( F \cap ) HSDK ( \cap ) HSDK</th>
</tr>
</thead>
<tbody>
<tr>
<td>5, 5, 30, 30</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5, 10, 20, 30</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>5, 10, 25, 30</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>5, 15, 25, 30</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>30, 30, 5, 5</td>
<td>37</td>
<td>38</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>30, 20, 10, 5</td>
<td>23</td>
<td>25</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>30, 25, 10, 5</td>
<td>27</td>
<td>27</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>30, 25, 15, 5</td>
<td>22</td>
<td>21</td>
<td>6</td>
<td>5</td>
</tr>
</tbody>
</table>

\(^a\) F \cap \) HSDK notes frequency of significant F-ratios without a significant HSDK result.

\(^b\) HSDK \( \cap \) F notes frequency of at least one significant HSDK test without a significant F-ratio.

\(^c\) F \( \cap \) HSDK notes frequency of significant F-ratios when at least one significant HSDK test occurred.
variance heterogeneity (variance ratio = 1:2:3:4) and unequal sample sizes (ranging from 5 to 30) the a priori and a posteriori approaches will not yield identical results for the HSDK form.

Tables 12 and 13 report results from the HSDH form and ANOVA F-ratio. The .05 nominal significance level is presented in Table 12. Table 13 contains data for the .01 nominal significance level. For the HSDH form with variance heterogeneity (variance ratio = 1:2:3:4) and unequal sample sizes (ranging from 5 to 30) significant F-ratios in ANOVA were produced without a significant HSDH contrast. In addition, significant HSDH values were found in the absence of a significant ANOVA F-ratio. It is concluded that under these conditions (variance heterogeneity and unequal sample sizes) the HSDH form will not produce identical a priori and a posteriori results.

Games (1971, p. 556) presented graphically the regions of retention of the omnibus F-ratio and HSD procedure. There was not a true correspondence between the two regions as nonidentical regions of retention existed. The tables reported in this Appendix G are generally consistent with Games' observations.
Table 12
Frequencies of Significant HSDH Tests and F-ratios Under the
Same Simulated Conditions of Variance Condition 1:2:3:4
and Unequal Sample Sizes for the .05 Nominal
Significance Level

<table>
<thead>
<tr>
<th>Sample Size ( (n_1, n_2, n_3, n_4) )</th>
<th>Frequency of Significant</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F-ratios</td>
</tr>
<tr>
<td>5, 5, 30, 30</td>
<td>8</td>
</tr>
<tr>
<td>5, 10, 20, 30</td>
<td>10</td>
</tr>
<tr>
<td>5, 10, 25, 30</td>
<td>11</td>
</tr>
<tr>
<td>5, 15, 25, 30</td>
<td>20</td>
</tr>
<tr>
<td>30, 30, 5, 5</td>
<td>96</td>
</tr>
<tr>
<td>30, 20, 10, 5</td>
<td>94</td>
</tr>
<tr>
<td>30, 25, 10, 5</td>
<td>74</td>
</tr>
<tr>
<td>30, 25, 15, 5</td>
<td>76</td>
</tr>
</tbody>
</table>

\(^a\) _F \neq HSDK_ notes frequency of significant F-ratios without a significant HSDK result.

\(^b\) _HSDK \neq F_ notes frequency of at least one significant HSDK test without a significant F-ratio.

\(^c\) _F \cap HSDK_ notes frequency of significant F-ratios when at least one significant HSDK test occurred.
Table 13
Frequencies of Significant HSDH Tests and F-ratios Under the Same Simulated Conditions of Variance Condition 1:2:3:4 and Unequal Sample Sizes for the .01 Nominal Significance Level

<table>
<thead>
<tr>
<th>Sample Size ((n_1, n_2, n_3, n_4))</th>
<th>Frequency of Significant</th>
<th>F-ratios</th>
<th>HSDH Test</th>
<th>(F \notin \text{HSDH})^a</th>
<th>HSDH (\notin F)</th>
<th>(F \cap \text{HSDH})</th>
</tr>
</thead>
<tbody>
<tr>
<td>5, 5, 30, 30</td>
<td></td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>5, 10, 20, 30</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5, 10, 25, 30</td>
<td></td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5, 15, 25, 30</td>
<td></td>
<td>4</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>30, 30, 5, 5</td>
<td></td>
<td>37</td>
<td>61</td>
<td>2</td>
<td>26</td>
<td>35</td>
</tr>
<tr>
<td>30, 20, 10, 5</td>
<td></td>
<td>23</td>
<td>49</td>
<td>4</td>
<td>30</td>
<td>19</td>
</tr>
<tr>
<td>30, 25, 10, 5</td>
<td></td>
<td>27</td>
<td>56</td>
<td>5</td>
<td>34</td>
<td>22</td>
</tr>
<tr>
<td>30, 25, 15, 5</td>
<td></td>
<td>25</td>
<td>53</td>
<td>2</td>
<td>33</td>
<td>20</td>
</tr>
</tbody>
</table>

^aF \notin \text{HSDH} notes frequency of significant F-ratios without a significant HSDH result.

^bHSDH \notin F notes frequency of at least one significant HSDH test without a significant F-ratio.

^cF \cap HSDH notes frequency of significant F-ratios when at least one significant HSDH test occurred.
BIBLIOGRAPHY


Pearson, E. S. Analysis of variance in cases of non-normal variation. Biometrika, 1931, 23, 114-133.


