TESTS OF SIGNIFICANCE FOR EXPERIMENTS INVOLVING PAIRED COMPARISONS

by

Thomas H. Harold Starks B.A., M.S.

Thesis submitted to the Graduate Faculty of the
Virginia Polytechnic Institute
in candidacy for the degree of

DOCTOR OF PHILOSOPHY
in
STATISTICS

APPROVED:

_____
Chairman, Advisory Committee

_____ _____

_____ _____

December, 1958

Blacksburg, Virginia

TABLE OF CONTENTS

INDEX OF TABLES

I. INTRODUCTION

1.1 Definition

A paired comparison is the process of comparing two treatments and choosing the treatment that is considered better on the basis of some common characteristic. For example, the treatments may be two blends of coffee and the comparison consists of selecting the blend with the better flavor. A paired-comparison experiment will involve the pairwise comparisons of two or more treatments. This thesis is confined to balanced experiments in which all possible pairs of treatments are compared one or more times.

The paired-comparison experiment was introduced by L. L. Thurstone (1927) for the purpose of estimating the relative strengths of treatment stimuli (e.g., the flavors of the blends of coffee) through subjective testing. Subjective experimentation is usually performed with treatments whose effects can be measured only on an ordinal scale through the reaction of individuals to the treatment stimuli. Since experimentation of this type involves one or more of the five senses, it is necessary to restrict the number of treatments per comparison in order to avoid tiring the senses, or overtaxing the memory of the individual, or both. Hence, a paired comparison experiment is the natural design to use.

Although paired-comparison experiments were introduced for estimation purposes, they are now also extensively used in the testing of hypotheses.

1.2 Previous Work

The notation used in this section will, in general, be that of the paper cited. However, in every case, the paired-comparison experiment will be one in which there are t (≥ 2)

treatments and n (≥ 1) repetitions, where a repetition is a set of paired comparisons containing the comparison of each treatment with every other treatment once and only once.

(i) The Thurstone-Mosteller Method

As previously mentioned, Thurstone (1927) introduced the idea of paired-comparison experiments. Mosteller (1951a) refined and extended Thurstone's method through the use of the theory of least squares. The resulting procedure, generally known as the Thurstone-Mosteller method, is based on the following model.

1. The stimulus produced by each treatment gives rise to a response or sensation when presented to an individual.

2. The stimuli of the treatments can be located on a subjective continuum (a response scale, usually not having a measurable physical characteristic).

3. For a population of individuals, the distribution of responses to the stimulus of treatment i ($i = 1,\ldots,t$) is normal with mean S_i and variance σ^2.

4. When treatments are presented in pairs it is possible for the responses in each pair to be correlated with correlation coefficient ρ.

If the scale of the subjective continuum is chosen in such a way that $2\sigma^2(1 - \rho) = 1$, the probability that a paired comparison of treatments i and j ($i \neq j$) will declare response X_i larger than response X_j (i.e., treatment i is preferred to j) is

$$(1.2.1) \qquad \pi_{ij} = \Pr(X_i > X_j) = \frac{1}{\sqrt{(2\pi)}} \int_{-(S_i - S_j)}^{\infty} e^{-\frac{1}{2}y^2} dy .$$

Now if p_{ij} is the proportion of times treatment i is preferred in comparisons with treatment j, a corresponding

estimate D'_{ij} of $D_{ij} = S_i - S_j$ may be obtained from a table of the normal probability integral. To locate the S_i on the subjective continuum, S_1 is taken to be the origin. Those S'_i which minimize

$$(1.2.2) \qquad \sum = \sum_{i,j} \left[D'_{ij} - (S'_i - S'_j) \right]^2, \qquad (S'_1 = 0),$$

are the least squares estimators of the "true" or mean responses.

Mosteller (1951b) has also proposed a χ^2 goodness-of-fit test of the Thurstone-Mosteller model.

Gulliksen and Tukey (1958) propose a variance components analysis to determine if it is reasonable to say that random variation accounts for the difference between the above model and the data. They also use the same technique to find how large the difference is relative to the variation that is accounted for by the theory.

(ii) The Bradley-Terry Method

Bradley and Terry (1952) have suggested another method of analysis for paired-comparison experiments. Their method is based on the following mathematical model:

1. The t treatments in a paired-comparison experiment have true "ratings" $\pi_1, \ldots, \pi_t$ ($\pi_i \geq 0$, $\sum \pi_i = 1$).

2. The probability π_{ij} that treatment i will be preferred to treatment j in a paired comparison is $\pi_i / (\pi_i + \pi_j)$.

Bradley (1954b) has developed a χ^2 goodness-of-fit test of the above model. In comparing this model with the Thurstone-Mosteller model, Bradley (1953,p.33) states, "If we redefine

$$(1.2.3) \qquad \pi_{ij} = \tfrac{1}{4} \int_{-(\ln \pi_i - \ln \pi_j)}^{\infty} \operatorname{sech}^2 \tfrac{1}{2} y \, dy \quad ,$$

it is easy to show that then $\pi_{ij} = \pi_i/(\pi_i + \pi_j)$. Thus, the substitution of the 'squared hyperbolic secant' density for the normal density of Thurstone's model yields the second method of analysis. The squared hyperbolic secant density is very similar to the normal. The substitution in Thurstone's model is a sufficient condition for the application of the model by Terry and the author (Bradley). It would not appear to be necessary. Values $\ln \pi_i$ correspond to values S_i on a subjective continuum. Methods developed for the estimation of these parameters will differ".

Bradley and Terry (1952) employed the method of maximum likelihood, which in this case involves an iterative procedure, to obtain estimates p_i of π_i. They also presented likelihood ratio tests of the following three sets of hypotheses:

(1) $H_{o1} : \pi_i = 1/t \quad (i = 1,\ldots,t)$, the hypothesis of treatment equality,
H_{a1} : No π_i assumed equal to any π_j $(i \neq j)$;

(2) $H_{o2} : \pi_i = 1/t \quad (i = 1,\ldots,t)$,
$H_{a2} : \pi_i = \pi \; (i = 1,\ldots,s)$; $\pi_i = (1 - s\pi)/(t - s)$ $(i = s + 1,\ldots,t)$,

(3) $H_{o3} : \pi_{i\gamma} = \pi_i \; (i = 1,\ldots,t;\ \gamma = 1,\ldots,r;\ 1 \leq r \leq n)$,
$H_{a3} : \pi_{i\gamma}$;

where $\pi_{i\gamma}$ is the true treatment rating of treatment i in the γth group of repetitions. (Note that the $\pi_{i\gamma}$ given here and the π_{ij} given in the model have entirely different meanings).

Tests of H_{o1} against H_{a1} are given for two different assumptions. One test, called "pooled analysis", is for the case in which the experimenter is confident that the true treatment ratings will not change between repetitions. The other test, called "combined analysis", is used when the experimenter thinks there is a possibility of a change in

treatment ratings between repetitions. Set (2) of hypotheses is for the case in which the experimenter believes that if the treatment ratings are not all equal, the treatments will be split into two groups, any two treatments in the same group will have equal ratings, any two not in the same group will have different ratings. A test of H_{03} against H_{a3} is made to determine whether the true treatment ratings change between groups of repetitions.

The cumulative probability distribution of the possible outcomes for experiments of sizes

$$t = 3, \quad n = 1,2,\ldots,10;$$
$$t = 4, \quad n = 1,2,\ldots,6;$$

with the arrangement of possible outcomes in the order of the corresponding values of the likelihood ratio statistic for hypotheses (1), were tabled by Bradley and Terry (1952). Bradley (1954a) extended the tables to include experiments of sizes

$$t = 4, \quad n = 7,8;$$
$$t = 5, \quad n = 1,\ldots,5.$$

Since there are $2^{\frac{1}{2}nt(t-1)}$ possible outcomes to the paired-comparison experiment, it is not too difficult to understand why any extension of these tables would involve considerable labor.

For experiments outside the range of their tables, Bradley and Terry (1952) recommend that the two tests of (1) and the test of (3) be performed by use of the χ^2-approximation to the distribution of $-2\ln\lambda$, where λ is the likelihood ratio statistic corresponding to the hypotheses being tested. Because λ is a function of the p_i's, an iterative procedure is required to obtain its value.

(iii) The Scheffé Method

Scheffé (1952) has introduced a method of paired comparisons which differs from the preceding methods in that it employs a scoring system and the analysis of variance. The following seven-point scoring system illustrates the scoring method when treatment i is compared with treatment j in the fixed order (i,j):

3: i is strongly preferred to j,
2: i is moderately preferred to j,
1: i is weakly preferred to j,
0: there is no preference,
-1: j is weakly preferred to i,
-2: j is moderately preferred to i,
-3: j is strongly preferred to i.

The experiment consists of 2r repetitions with the ordered pair (i,j) occurring in half of them and (j,i) in the other half. This scheme makes possible a test of the effect of the order of presentation. The score given the ordered pair (i,j) by the k th judge (or repetition) is denoted by x_{ijk}. The true response α_i (similar to S_i in the Thurstone-Mosteller method) to treatment i is estimated by

$$(1.2.4) \qquad \hat{\alpha}_i = \frac{1}{2rt}\left[\sum_{k=1}^{r}\sum_{j\neq i}^{t} x_{ijk} - \sum_{m=1}^{r}\sum_{j\neq i}^{t} x_{jim}\right].$$

The assumptions of the model for the Scheffé procedure are as follows:

1. All the x_{ijk} are independent random variables, and for a fixed ordered pair (i,j) all r variables x_{ijk} have the same mean μ_{ij} and the same variance σ^2 which does not depend on (i,j).
2. The x_{ijk}'s may be treated as normal variates.

The second assumption is necessary for testing purposes only.

To test the null hypothesis that all treatments produce equivalent stimuli, the Tukey test based on allowances is used. If the test rejects the null hypothesis, it also separates the treatments with significantly different α_i's.

The relative merits of the three methods given above are discussed in a paper by Jackson and Fleckenstein (1957).

(iv) Other Paired-Comparison Test-Methods

Kendall and Babington Smith (1940) consider experiments with exactly one repetition per judge. They define a circular triad to be an event in which three paired comparisons involving any three treatments i, j, and k yield results of the form $X_i > X_j$, $X_j > X_k$, and $X_k > X_i$ (using the notation of the Thurstone-Mosteller method). The number of circular triads occurring in a repetition is used as a test statistic to test the consistency of the corresponding judge. A test of agreement of the judges was also proposed based on the number of agreements between pairs of judges. (This test of agreement may also be considered a test of the third set of hypotheses in the Bradley-Terry method for the case $r = n$.)

Papers by Durbin (1951) and Benard and Van Elteren (1953), concerning ranking methods, consider tests of the null hypothesis that all treatments are equivalent against the general alternative. In the special case of paired-comparison experiments, they recommend the use of a function of the corrected sum of squares of the treatment scores as a test statistic, where the score of a treatment is the number of times it is preferred in the experiment. For reasonably large experiments, their test statistic is distributed approximately as a χ^2-variate with (t-1) degrees of freedom. This method will be discussed further in Chapter II.

David (1959) has considered paired-comparison experiments in terms of Round Robin and Knock-Out tournaments. The Round Robin tournament with t players is equivalent to the usual paired-comparison experiment with t treatments and one repetition. In the Round Robin tournament, he tests the null hypothesis that all players are of equal strength against three specific alternatives as opposed to the general alternatives tested by the preceding methods. The specific alternative hypotheses are (1) the ability of a specified player is superior to the average of the participant abilities, (2) two specified players differ in ability, and (3) the player winning the most games is better than the average participant. His tests are based essentially on the binomial distribution of each treatment score.

David (1959) also tables the distribution of the outcomes of paired-comparison experiments of sizes $n = 1$, $t = 3,\ldots,8$.

(v) Testing Factorials with Paired Comparisons

Abelson and Bradley (1954) have considered paired-comparison experiments in which the treatments are factorial combinations. A maximum likelihood approach similar to that of Bradley and Terry (1952) is employed. The resulting equations are unwieldy for anything larger than a 2 x 2 factorial.

Bliss, Greenwood, and White (1956) have proposed considering the number of times treatment i is preferred to treatment j as a rank in a series of $(n + 1)$ and replacing it with the corresponding expected normal deviate (rankit). After replacement, an analysis of variance technique is applied to the rankits for testing purposes. They present an example involving treatments that are factorial combinations in a 2 x 2 factorial. Their method may be based on either the Thurstone-Mosteller model or the Scheffé model.

Dykstra (1958) has recently published a procedure for factorials based on the Scheffé method. Through blocking, he decreases the number of required paired comparisons needed to test large factorials. The method is not difficult, but does depend on the applicability of the Scheffé model.

1.3 Objectives of this Dissertation

This dissertation has several objectives, all dealing with tests of hypotheses in paired-comparison experiments. In every case, an attempt will be made to reduce the number of assumptions needed in the model and to provide a test procedure that may be easily and speedily followed by the experimenter.

In Chapter II, approximate test procedures will be presented for testing

(1) H_{o1} : $\pi_{i.} = \sum_{k \neq i}^{t} \pi_{ik}/(t-1) = \frac{1}{2}$ $(i=1,\ldots,t)$,

H_{a1} : $\pi_{i.} \neq \frac{1}{2}$ for some i ;

and

(2) H_{o2} : the π_{ij}'s do not change between groups of repetitions,

H_{a2} : the π_{ij}'s do change between groups of repetitions;

when the experiments lie outside the range of the tabled distributions of experiment outcomes. These tests will be easy to perform and will be based on a very general model.

In Chapter III, extensions will be given of the various tests proposed by David (1959) for Round Robin tournaments. Also, methods will be developed for separating significantly different treatment scores and for judging linear contrasts of the treatment scores.

In Chapter IV, a method of testing factorial effects in paired-comparison experiments will be presented based on a very general model.

In Chapter V, the data obtained from two experiments will be analyzed with the methods developed in Chapters II, III, and IV.

II. APPROXIMATE OVERALL TESTS OF SIGNIFICANCE

2.1 Model and Objectives

In a paired-comparison experiment with t treatments, we shall consider that there is a true probability π_{ij} $(=1 - \pi_{ji})$ that treatment i will be preferred to treatment j associated with each of the $\binom{t}{2}$ possible pairs (i,j). These probabilities will be called "preference probabilities". They may change in value between repetitions of the experiment. It will be assumed that when two treatments, say i and j, have stimuli that are <u>identical</u>, we have $\pi_{ij} = \frac{1}{2}$, and $\pi_{ir} = \pi_{jr}$, for all treatments $r \neq i,j$. The stimuli of two treatments, i and j, will be defined as <u>equal</u> if $\pi_{i.} = \pi_{j.}$, where

$$\pi_{k.} = \sum_{m \neq k}^{t} \pi_{km}/(t-1) \ . \tag{2.1.1}$$

This model includes the Thurstone-Mosteller and Bradley-Terry models and all other linear models as special cases. However, it is not restricted to the linear case as the circular triad situation

$$\pi_{ij} > \tfrac{1}{2}, \quad \pi_{jk} > \tfrac{1}{2}, \quad \pi_{ki} > \tfrac{1}{2}, \tag{2.1.2}$$

is not excluded.

When a linear model is appropriate, the null hypothesis of equality of treatment stimuli,

$$H_0 : \pi_{i.} = \tfrac{1}{2} \qquad (i = 1,\ldots,t), \tag{2.1.3}$$

implies that all the treatment stimuli correspond to the

same location on the response scale and, hence, H_o is equivalent to

$$(2.1.4) \qquad H_o' : \pi_{ij} = \tfrac{1}{2} \qquad (i,j = 1,\ldots,t;\ i \neq j).$$

However, for the non-linear case, H_o and H_o' are not equivalent [e.g., $\pi_{12} = 1$, $\pi_{23} = 1$, $\pi_{31} = 1$, implies $\pi_{i.} = \tfrac{1}{2}$ $(i = 1,2,3)$].

In this chapter, approximate tests of the null hypothesis H_o against its general alternative are presented for the case in which it can be assumed prior to the experiment that there will be no interaction between repetitions and preference probabilities, and also for the case in which the previous "no interaction" assumption cannot be made. Further, an approximate test of the null hypothesis that there is no interaction between repetitions and preference probabilities is developed.

Comparisons are made between the proposed tests and other existing approximate tests on the basis of accuracy, model, and ease of computation.

2.2 Test of Equality of Treatment Stimuli.

Consider a paired-comparison experiment with t treatments and n repetitions. Each treatment occurs in n(t-1) paired comparisons. Let a_i $(i = 1,\ldots,t)$ denote the number of comparisons in which treatment i is the preferred treatment. It is convenient to refer to a_i as the score of the i th treatment. The total number of paired comparisons in the experiment is $n\binom{t}{2}$, and, since one preference is associated with each comparison, we have

$$(2.2.1) \qquad \sum_{i=1}^{t} a_i = n\binom{t}{2} .$$

We wish to test the null hypothesis

$$H_o : \pi_{i.} = \tfrac{1}{2} \qquad (i = 1,\ldots,t),$$

against the general alternative hypothesis

$$H_a : \pi_{i.} \neq \tfrac{1}{2} \text{ for some } i,$$

while assuming that the true preference probabilities π_{ij} do not change between repetitions. (This is equivalent to the Bradley-Terry (1952) pooled analysis situation.) To test H_o against H_a, it is natural to base a test-criterion on the sum of squares of the deviations of the treatment scores from their expected values under H_o.

It was noted in the previous section that when a linear model (e.g., the Thurstone-Mosteller model) is appropriate, H_o is equivalent to

$$H_o' : \pi_{ij} = \tfrac{1}{2} \qquad (i,j = 1,\ldots,t;\ i \neq j).$$

The distribution of the test-criterion will be obtained under H_o'. The resulting test is, of course, then suitable for situations in which a linear model is appropriate. For the non-linear model case, in which H_o and H_o' are not equivalent, it will be demonstrated that the test is still applicable, but in this situation it is a conservative test. All subsequent statements in this section are made under the assumption that H_o' is true.

Since each treatment has a probability of $\frac{1}{2}$ of being preferred in each of its $n(t-1)$ comparisons with the other $(t-1)$ treatments, we have

$$(2.2.2) \qquad E(a_i) = \tfrac{1}{2}n(t-1) .$$

Also, due to the linear restriction (2.3.1) on the scores,

$$(2.2.3) \qquad \bar{a} = \sum_{i=1}^{t} a_i/t = n\binom{t}{2}/t = \tfrac{1}{2}n(t-1),$$

and, therefore

$$(2.2.4) \qquad \bar{a} = E(a_i) \qquad (i = 1,\ldots,t) .$$

We multiply the sum of squares,

$$(2.2.5) \qquad \sum_{i=1}^{t} (a_i - \bar{a})^2$$

by 4/(nt) to obtain the proposed test-statistic,

$$(2.2.6) \qquad D = 4\sum_{i=1}^{t} (a_i - \bar{a})^2/(nt) = 4\left[\sum_{i=1}^{t} a_i^2 - \tfrac{1}{4}n^2t(t-1)^2\right]/(nt).$$

It will be shown later that the factor 4/(nt) is needed to give the test-statistic a desirable asymptotic distribution.

It should be mentioned that a general statistic proposed by Durbin (1951) for use in the rank analysis of balanced incomplete block designs with block size k ($2 \le k \le t - 1$) is for our special case, k = 2, the same as D. Also, for n = 1, D may be written

$$(2.2.7) \qquad D = 8\left[(1/12)t(t-1)(2t-1) - c - (1/8)t(t-1)^2\right]/t$$

where c is the number of circular triads occurring in the experiment [cf. Kendall and Babington Smith (1940)].

Durbin (1951) has suggested that when the value of t is moderately large and m, the number of times each treatment is ranked [$m = n(t-1)$ in the paired-comparisons case], is large, one can consider his general statistic to be distributed

as a χ^2-variate with (t-1) degrees of freedom. This suggestion is based on the fact that the statistic, like our special case, D, is a function of a corrected sum of squares with one restriction and its mean and variance are (t-1) and $2(t-1)\left[1-\frac{k(t-1)}{mt(k-1)}\right]$, respectively. Benard and Van Elteren (1953), in a paper considering the general m-comparisons type of experiment which has the balanced incomplete block design discussed by Durbin (1951) as a special case, give a method for proving that the limiting distribution of Durbin's statistic is the suggested χ^2-distribution.

The approximate test of H_o suggested by Bradley and Terry (1952) is based on $T = -2\ln\lambda$, where λ is a likelihood ratio statistic. Without actually using the expression we denote by D, Bradley (1955) shows that T converges to D as $n\to\infty$. Since T is known to have an asymptotic χ^2-distribution with (t-1) degrees of freedom, Bradley has, in effect, proved that as $n\to\infty$ the limiting distribution of D is the suggested χ^2-distribution.

For completeness, a short proof of the asymptotic property of the distribution of D will be presented in the next paragraph.

The score, a_i (i = 1,...,t), is a binomial variate with the parameters $\left[\frac{1}{2}, n(t-1)\right]$. It follows that

$$d_i = 2\left[a_i - \tfrac{1}{2}n(t-1)\right] / \sqrt{(nt)} \qquad (2.2.8)$$

has mean zero and variance

$$\sigma^2 = \frac{t-1}{t} . \qquad (2.2.9)$$

Knowing that $\Sigma d_i = 0$, and that the d_i are, by symmetry, equicorrelated with correlation coefficient ρ (say), we have

$$\text{(2.2.10)} \qquad \mathrm{Var}(\textstyle\sum d_i) = t\sigma^2 + 2\binom{t}{2}\rho\sigma^2 = 0 \ .$$

Solving for the covariance $\rho\sigma^2$ of any two d_i, we have

$$\text{(2.2.11)} \qquad \rho\sigma^2 = -\frac{1}{t} \ .$$

Hence, the dispersion matrix of the variable $\underline{d} = (d_1, d_2, \ldots, d_t)$ is the t x t matrix

$$\text{(2.2.12)} \qquad \sum_{\underline{d}} = \begin{bmatrix} \frac{t-1}{t} & -\frac{1}{t} & \cdot & \cdot & \cdot & -\frac{1}{t} \\ -\frac{1}{t} & \frac{t-1}{t} & & & & \cdot \\ \cdot & & \cdot & & & \cdot \\ \cdot & & & \cdot & & \cdot \\ \cdot & & & & \cdot & -\frac{1}{t} \\ -\frac{1}{t} & \cdot & \cdot & \cdot & -\frac{1}{t} & \frac{t-1}{t} \end{bmatrix} \ .$$

From the generalized central limit theorem (see Appendix Section A), we have that the limiting distribution of $\underline{d}$ as $n \to \infty$ is $N(\underline{0}, \sum_{\underline{d}})$. The matrix of the Helmert transformation of t variables is

$$\text{(2.2.13)} \qquad H = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} & 0 & \cdot & \cdot & \cdot & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{-2}{\sqrt{6}} & 0 & \cdot & \cdot & 0 \\ \cdot & & & \cdot & & & \\ \cdot & & & & \cdot & & \\ \cdot & & & & & \cdot & \\ \frac{1}{\sqrt{t(t-1)}} & \frac{1}{\sqrt{t(t-1)}} & \cdot & \cdot & \cdot & \frac{1}{\sqrt{t(t-1)}} & \frac{-(t-1)}{\sqrt{t(t-1)}} \\ \frac{1}{\sqrt{t}} & \frac{1}{\sqrt{t}} & \cdot & \cdot & \cdot & & \frac{1}{\sqrt{t}} \end{bmatrix} \ .$$

We now find

$$(2.2.14) \qquad H\Sigma_{\underline{d}}H' = \begin{bmatrix} 1 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & 1 & & & & \cdot \\ \cdot & & \cdot & & & \cdot \\ \cdot & & & \cdot & & \cdot \\ \cdot & & & & 1 & 0 \\ 0 & \cdot & \cdot & \cdot & 0 & 0 \end{bmatrix} .$$

Since H is an orthogonal matrix, (2.2.14) demonstrates that $\Sigma_{\underline{d}}$ has (t-1) characteristic roots equal to one and that the remaining root is zero. As shown in Cramer (1946, Section 24.5), the above results are sufficient to prove $D = \Sigma d_i^2$ has a limiting χ^2-distribution with (t-1) degrees of freedom as $n \to \infty$.

Since a_i is a binomial variate with parameters $\left[\frac{1}{2}, n(t-1)\right]$, $d_i = 2\left[\bar{a}_i - \frac{1}{2}n(t-1)\right]/\sqrt{(nt)}$ is distributed approximately as a $N(0,(t-1)/t)$-variate when t is moderately large. The Helmert transformation $\underline{y} = H\underline{d}$ yields t uncorrelated linear functions y_i of the d_i with covariance matrix $H\Sigma_{\underline{d}}H'$. Therefore, we have that the uncorrelated variates $y_1, \ldots, y_{t-1}$, are approximately N(0,1)-variates and that y_t has a point distribution at the origin. This leads one to expect that the statistic $D = \sum_{i=1}^{t} d_i^2 = \sum_{i=1}^{t-1} y_i^2$ is distributed approximately as a χ^2-variate with (t-1) degrees of freedom for moderately large values of t.

Now, knowing the approximate distribution of D under the null hypothesis for moderately large values of n and t, we have the basis for an approximate test of H_o using the D-statistic. The test proceeds in the following manner:

1. Specify the significance level, α.

2. From tables of the χ^2-distribution, find the 100 $(1-\alpha)$ percentile, $\chi^2_{(t-1),\alpha}$ say, of the χ^2-distribution with $(t-1)$ degrees of freedom.
3. Compute $D = 4\left[\sum a_i^2 - \frac{1}{4}n^2t(t-1)^2\right] / (nt)$.
4. If $D \geq \chi^2_{(t-1),\alpha}$, we accept the alternative hypothesis H_a; otherwise, we fail to reject H_o.

In the non-linear case, where H_o is not equivalent to H_o', the test of H_o based on D is conservative. This is seen from the fact that in the non-linear case we have

$$(2.2.15)\quad E(D) = 4\sum E(a_i - \bar{a})^2/(nt)$$

$$= 4\sum_{i=1}^{t}\left[\frac{1}{4}n(t-1) - n\sum_{j\neq i}(\pi_{ij} - \tfrac{1}{2})^2\right]/(nt)$$

$$= (t-1) - 4\sum_{i=1}^{t}\sum_{j\neq i}^{t}(\pi_{ij} - \tfrac{1}{2})^2/t$$

$$\leq (t-1) = E(\chi^2_{t-1}).$$

Hence, we see that the expected value of D is now less than it was under the linear model. Further, consider the extreme case with three treatments with preference probabilities

$$(2.2.16)\qquad \pi_{12} = 1, \quad \pi_{23} = 1, \quad \pi_{31} = 1 .$$

In this case, H_o is true but D has a point distribution, i.e., $Pr(D = 0) = 1$. This indicates that both the mean and the variance of D decrease as we depart from the linear model.

For experiments falling within the range of the tables of Bradley and Terry (1952), Bradley (1954a) and David (1959) an exact test can be made when a linear model is appropriate.

The value of $Pr(D \geq D_o | H_o')$, where D_o is the observed value of D, can be obtained from one of the above tables. If $Pr(D \geq D_o | H_o')$ is less than the significance level set for the test, one would reject H_o. Again this test is conservative for the non-linear model situation.

2.3 Accuracy of the Approximate Test Based on D

As implied in Section 2.2, a general test proposed by Durbin (1951) for balanced incomplete block designs reduces to the test based on the D-statistic when the block size is two. Durbin stated that the distribution of his general statistic could be approximated by that of a χ^2-variate with (t-1) degrees of freedom when the number of treatments, t, is moderately large and the number of blocks containing any particular treatment $[\bar{n}(t-1)$ in the paired-comparisons case$]$ is large. However, Durbin gave no indication of the accuracy of the test or how large the experiment must be before the use of the approximate test can be justified. In this section, the answers to these questions will be indicated for our special case, paired comparisons (i.e., block size equal two).

When the combination (t,n) lies within the range of the tables prepared by Bradley and Terry (1952), Bradley (1954a), and David (1959), one can use these tables to determine $Pr(D \geq D_o | H_o')$, where D_o is the observed value of the statistic D. By comparing $Pr(D \geq D_o | H_o')$, from the above tables, with $Pr(\chi^2_{t-1} \geq D_o)$, one can determine how well the distribution of D is fitted by the χ^2-distribution for small values of n and t. Such a comparison is made for ten combinations (t,n) in Table 2.1. These ten combinations include the largest experiments for which the distributions of scores are tabulated. The accuracy of the approximate test of H_o

based on D is of more interest than the overall fit; so, in many cases, only the approximations in the critical part of the distribution (i.e., $0.5\% \leq \Pr[\overline{D} \geq D_o / H_o'] \leq 15\%$) are considered.

Since the possible values of the discrete variable D cannot differ by less than $8/(nt)$, one might consider the use of a continuity correction that would lead one to compare $\Pr[\overline{\chi}^2_{t-1} \geq D_o - 4/(nt)]$ with $\Pr[\overline{D} \geq D_o / H_o']$. Table 2.1 was constructed without the use of the continuity correction, since, as the table indicates, $\Pr(\chi^2_{t-1} \geq D_o)$ already gives a conservative approximation (i.e., larger than $\Pr[\overline{D} \geq D_o | H_o']$) for values of D_o such that $\Pr(D \geq D_o | H_o')$ is in the interval $(0, 0.05)$ and the correction would only tend to increase this error in this critical part of the distribution.

Table 2.1 shows that the χ^2-approximation is rather poor for these small values of n and t when D_o is such that $\Pr(D \geq D_o | H_o') \geq 15\%$. Nevertheless, the table also demonstrates that the approximation in the critical part of the distribution used in testing H_o; namely, $0 \leq \Pr[\overline{D} \geq D_o | H_o'] \leq 10\%$, becomes reasonably accurate for the larger values of n associated with $t = 3$, $t = 4$, and $t = 5$. In fact, for the three combinations (3,10), (4,8), and (5,5), $\Pr(\chi^2_{t-1} \geq D_o)$ differs from $\Pr(D \geq D_o | H_o')$ by less than 0.0050 at the critical values of D for a 5% significance level test and by less than 0.0030 at the critical values of D for a 1% significance level test. Due to the asymptotic property of the distribution of D, it is evident for $t = 3$, 4, and 5 that the accuracy of the χ^2-approximation will, in general, be adequate for testing purposes when n takes on values outside the range for which the distributions are tabulated.

Little can be said about the value of n that is necessary to assure a reasonably accurate significance level approximation

when t is larger than five. However, the asymptotic property of the distribution of D may be taken as a basis for suggesting that the approximations should be at least as accurate for the combinations ($t \geq 5$, $n \geq 5$) as they were for the combination (5,5). Also, we find in the case of (8,1) that our approximations are conservative and not in error by more than 0.015 in the critical part of the distribution. For many experiments, this degree of accuracy would be sufficient.

TABLE 2.1

Accuracy of the χ^2-Approximation to the Distribution of D

$t = 3, n = 5$				$t = 3, n = 8$			
$\frac{1}{4}ntD_o$	$Pr(D \geq D_o \mid H_o')$	$Pr(\chi_2^2 \geq D_o)$	Difference	$\frac{1}{4}ntD_o$	$Pr(D \geq D_o \mid H_o')$	$Pr(\chi_2^2 \geq D_o)$	Difference
50	0.0002	0.0013	-0.0011	128	0.0000	0.0000	0.0000
42	.0020	.0037	- .0017	114	.0000	.0001	- .0001
38	.0057	.0063	- .0006	104	.0000	.0002	- .0002
32	.0112	.0140	- .0028	98	.0001	.0003	- .0002
26	.0386	.0312	+ .0074	96	.0001	.0003	- .0002
24	.0569	.0408	.0161	88	.0001	.0007	- .0006
18	.1000	.0907	.0093	86	.0003	.0008	- .0005
14	.2464	.1546	.0918	78	.0008	.0015	- .0007
8	.4039	.3442	.0597	74	.0015	.0021	- .0006
6	.6053	.4493	.1560	72	.0020	.0025	- .0005
2	.9313	.7659	.1654	62	.0046	.0057	- .0011
0	1.0000	1.0000	0.0000	56	.0091	.0094	- .0003
				54	.0118	.0111	.0007
				50	.0158	.0155	.0003
				42	.0324	.0302	.0022
				38	.0560	.0421	.0139
				32	.0758	.0695	.0063
				26	.1419	.1146	.0273
				24	.1808	.1353	.0455
				18	.2447	.2231	.0216
				14	.4211	.3114	.1097
				8	.5631	.5134	.0497
				6	.7293	.6065	.1228
				2	.9559	.8465	.1096
				0	1.0000	1.0000	0.0000

$$\left[\frac{1}{4}ntD_o = \sum (a_i - \bar{a})^2\right]$$

TABLE 2.1 (continued)

t = 3, n = 10				t = 4, n = 6			
$\frac{1}{4}ntD_o$	$Pr(D \geq D_o \mid H_o')$	$Pr(\chi_2^2 \geq D_o)$	Difference	$\frac{1}{4}ntD_o$	$Pr(D \geq D_o \mid H_o')$	$Pr(\chi_3^2 \geq D_o)$	Difference
128	0.0000	0.0002	-0.0002	80	0.0023	0.0040	-0.0017
126	.0001	.0002	- .0001	78	.0028	.0046	- .0018
122	.0001	.0003	- .0002	76	.0031	.0054	- .0023
114	.0002	.0005	- .0003	74	.0048	.0063	- .0015
104	.0005	.0010	- .0005	72	.0055	.0074	- .0019
98	.0012	.0015	- .0003	70	.0064	.0086	- .0022
96	.0014	.0017	- .0003	68	.0077	.0101	- .0024
86	.0025	.0032	- .0007	66	.0106	.0117	- .0011
78	.0046	.0055	- .0009	64	.0108	.0137	- .0029
74	.0074	.0072	.0002	62	.0152	.0159	- .0007
72	.0090	.0082	.0008	58	.0168	.0216	- .0048
62	.0157	.0160	- .0003	56	.0204	.0252	- .0048
56	.0261	.0239	.0022	54	.0294	.0293	.0001
54	.0320	.0273	.0047	52	.0321	.0341	- .0020
50	.0399	.0357	.0042	50	.0432	.0396	.0036
42	.0674	.0608	.0066	48	.0444	.0460	- .0016
38	.1035	.0794	.0241	46	.0536	.0534	.0002
32	.1306	.1184	.0122	44	.0590	.0620	- .0030
26	.2112	.1767	.0345	42	.0718	.0719	- .0001
24	.2571	.2019	.0552	40	.0796	.0833	- .0037
18	.3250	.3012	.0238	38	.1069	.0964	.0105
14	.5009	.3932	.1077	36	.1202	.1117	.0085
8	.6299	.5866	.0433	34	.1455	.1291	.0164
6	.7762	.6703	.1059	32	.1530	.1490	.0040
2	.9644	.8752	.0892	30	.1888	.1718	.0170

$$\left\lfloor \tfrac{1}{4}ntD_o = \Sigma (a_i - \bar{a})^2 \right\rfloor$$

TABLE 2.1 (continued)

t = 4, n = 8				t = 5, n = 2			
$\frac{1}{4}ntD_o$	$Pr(D \geq D_o \mid H'_o)$	$Pr(\chi^2_3 \geq D_o)$	Difference	$\frac{1}{4}ntD_o$	$Pr(D \geq D_o \mid H'_o)$	$Pr(\chi^2_4 \geq D_o)$	Difference
98	0.0054	0.0066	-0.0012	40	0.0001	0.0030	-0.0029
96	.0056	.0074	- .0018	38	.0006	.0043	- .0037
94	.0070	.0081	- .0011	36	.0009	.0061	- .0052
90	.0091	.0104	- .0013	34	.0030	.0087	- .0057
88	.0096	.0117	- .0021	32	.0048	.0123	- .0075
86	.0125	.0132	- .0007	30	.0093	.0173	- .0080
84	.0138	.0148	- .0010	28	.0126	.0244	- .0118
82	.0154	.0166	- .0012	26	.0289	.0342	- .0053
80	.0162	.0186	- .0024	24	.0424	.0477	- .0053
78	.0182	.0208	- .0026	22	.0691	.0663	.0028
76	.0193	.0233	- .0040	20	.0926	.0916	.0010
74	.0257	.0261	- .0004	18	.1319	.1257	.0062
72	.0279	.0293	- .0014	16	.1948	.1712	.0236
70	.0312	.0328	- .0016	14	.3053	.2311	.0742
68	.0350	.0367	- .0017	12	.3403	.3084	.0319
66	.0437	.0411	.0026	10	.4989	.4060	.0929
64	.0444	.0460	- .0016	8	.6122	.5249	.0873
62	.0556	.0515	.0041	6	.7727	.6627	.1100
58	.0593	.0643	- .0050	4	.8866	.8088	.0778
56	.0676	.0719	- .0043	2	.9926	.9385	.0541
54	.0865	.0803	.0062				
52	.0918	.0897	.0021				
50	.1100	.1000	.0100				
48	.1123	.1116	.0007				
46	.1280	.1244	.0036				
44	.1370	.1386	- .0016				
42	.1571	.1544	.0027				
40	.1686	.1718	- .0032				

$$\left[\frac{1}{4}ntD_o = \sum (a_i - \bar{a})^2\right]$$

TABLE 2.1 (continued)

t = 5, n = 3				t = 5, n = 5			
$\frac{1}{4}ntD_o$	$Pr(D \geq D_o \mid H'_o)$	$Pr(\chi^2_4 \geq D_o)$	Difference	$\frac{1}{4}ntD_o$	$Pr(D \geq D_o \mid H'_o)$	$Pr(\chi^2_4 \geq D_o)$	Difference
62	0.0007	0.0024	-0.0017	90	0.0042	0.0061	-0.0019
60	.0009	.0030	- .0021	88	.0050	.0070	- .0020
58	.0015	.0038	- .0023	86	.0064	.0081	- .0017
56	.0022	.0048	- .0026	84	.0070	.0093	- .0023
54	.0032	.0061	- .0029	82	.0085	.0107	- .0022
52	.0038	.0077	- .0039	80	.0098	.0123	- .0025
50	.0060	.0098	- .0038	78	.0111	.0141	- .0030
48	.0069	.0123	- .0054	76	.0128	.0162	- .0034
46	.0117	.0155	- .0038	74	.0167	.0186	- .0019
44	.0153	.0194	- .0041	72	.0185	.0213	- .0028
42	.0186	.0244	- .0058	70	.0225	.0244	- .0019
40	.0250	.0306	- .0056	68	.0244	.0279	- .0035
38	.0348	.0382	- .0034	66	.0294	.0320	- .0026
36	.0415	.0477	- .0062	64	.0348	.0366	- .0018
34	.0624	.0595	+ .0029	62	.0411	.0418	- .0007
32	.0755	.0739	.0016	60	.0441	.0477	- .0036
30	.0958	.0916	.0042	58	.0526	.0545	- .0019
28	.1081	.1132	- .0051	56	.0616	.0622	- .0006
26	.1566	.1395	.0171	54	.0733	.0708	.0025
24	.1885	.1712	.0173	52	.0789	.0805	- .0016
22	.2436	.2093	.0343	50	.0957	.0916	.0041
20	.2825	.2548	.0277	48	.1021	.1040	- .0019
				46	.1272	.1180	.0092
				44	.1433	.1338	.0095
				42	.1560	.1514	.0046
				40	.1781	.1712	.0069
				38	.2071	.1932	.0139
				36	.2248	.2178	.0070

$$\left[\frac{1}{4}ntD_o = \sum(a_i - \bar{a})^2\right]$$

TABLE 2.1 (continued)

t = 7, n = 1				t = 8, n = 1			
$\frac{1}{4}ntD_o$	$Pr(D \geq D_o \mid H'_o)$	$(\chi^2_6 \geq D_o)$	Difference	$\frac{1}{4}ntD_o$	$Pr(D \geq D_o \mid H'_o)$	$(\chi^2_7 \geq D_o)$	Difference
28	0.0024	0.0138	-0.0114	42	0.0002	0.0038	-0.0036
26	.0064	.0214	- .0150	40	.0005	.0056	- .0051
24	.0168	.0330	- .0162	38	.0014	.0082	- .0068
22	.0328	.0504	- .0176	36	.0028	.0120	- .0092
20	.0689	.0760	- .0071	34	.0064	.0174	- .0110
18	.1120	.1130	- .0010	32	.0112	.0251	- .0139
16	.1977	.1657	+ .0320	30	.0226	.0360	- .0134
14	.2874	.2381	+ .0493	28	.0370	.0512	- .0142
				26	.0626	.0721	- .0095
				24	.0938	.1006	- .0068
				22	.1528	.1386	+ .0142
				20	.2077	.1886	+ .0191
				18	.2989	.2527	+ .0462

$$\left[\frac{1}{4}ntD_o = \sum (a_i - \bar{a})^2\right]$$

2.4 Comparison with Other Approximate Tests

Three methods, other than the one based on the D-statistic, are available to the experimenter as approximate tests of

$$H_o : \pi_{i.} = \tfrac{1}{2} \qquad (i = 1,\ldots,t),$$

against

$$H_a : \pi_{i.} \neq \tfrac{1}{2} \qquad \text{for some } i,$$

under the assumption of no change in the preference probabilities between repetitions. One of these methods is applicable only in experiments involving one repetition.

All of the approximate tests are reasonably accurate if the experiment is sufficiently large. However, an indication of their accuracy for moderate - size experiments will be obtained through comparison of approximate significance levels with exact significance levels for the largest experiments with tabled distributions. Also, each of the three methods will be compared with the D-method on the basis of ease of computation.

(i) Durbin's Test Based on the F-Distribution

Durbin (1951) suggested another approximate test for balanced incomplete block designs based on a statistic that reduces to

$$(2.4.1) \qquad F = [n(t-1) - 3]\cdot 4\cdot \sum_{i=1}^{t} (a_i-\bar{a})^2 / [n^2 t(t^2-1) - 12 \sum_{i=1}^{t} (a_i-\bar{a})^2]$$

for the paired-comparison case. Under the null hypothesis,

the first two moments of F are the same as the first two moments of the F-distribution with

$$(2.4.2) \qquad v_1 = \frac{nt\left[\bar{n}(t^2-1) - 3t - 3\right] + 12}{n(t+1)(nt-2)} \quad ,$$

and

$$(2.4.3) \qquad v_2 = \left[\frac{\bar{n}(t+1)}{3} - 1\right] v_1 \quad ,$$

degrees of freedom. Also, the third and fourth moments of the statistic are approximately the same as those of the F-statistic for large values of $n(t-1)^2$. Hence, the test of H_o consists of comparing the observed value of F with $F_{v_1,v_2;\alpha}$, the $100(1-\alpha)$ percentile of the F-distribution with v_1 and v_2 degrees of freedom, and rejecting H_o when the observed F is not less than $F_{v_1,v_2;\alpha}$.

It should be noted that although the calculation of F, v_1, and v_2 is not difficult, it is not as simple as finding D. However, due to the fact that v_1 and v_2 are usually fractions, the value of $F_{v_1,v_2;\alpha}$ will generally have to be found by two-way non-linear interpolation in a table of the α-level critical values of the F-distribution. Similarly, if one wants to find the significance level of the observed F by comparing it with the F-distribution, one must perform a three-way non-linear interpolation in Pearson's (1934) Tables of the Incomplete Beta Function.

As was observed in Section 2.2, the D-statistic has only its first moment equal and its second moment asymptotically equal to those of a χ^2-variate with (t-1) degrees of freedom. Since more moments are equated both exactly and approximately in the case of the F-statistic, Durbin states that the F-distribution will give a more accurate approximation to the distribution of F than the χ^2-distribution

will give to the distribution of D. However, we are only interested in the fit in the critical part of the distribution, and a better overall fit does not necessarily mean a better approximation in the critical region. In Table 2.2, comparisons are made between the approximate significance levels obtained by the methods associated with the D and F (with continuity correction*) statistics and the exact value, for experiment sizes (3,10), (4,8), and (8,1), and values of $\sum_{i=1}^{t}(a_i-\bar{a})^2$ that are near the 1% and 5% significance levels.

The comparisons indicate that the D-statistic gives a more conservative approximation to the true significance level than does the Durbin F. The corrected F-statistic gave better approximations for the two comparisons in the experiment size (8,1). However, D gave a slightly better approximation for three of the other four comparisons. It, therefore, appears that there is very little difference in the accuracy of the two methods for moderate experiment sizes with the possible exception of the experiments in which n is near one.

---

* It was found that the error in the F-approximation could be reduced in each of the six cases considered by the use of a continuity correction suggested by Kendall (1955; p.99, Example 6.4). The corrected statistic is

$$F_c = \frac{4\sqrt{n}(t-1) - 3\sqrt{\sum(a_i-a)^2} - 4}{n^2t(t^2-1) - 12\sum(a_i-a)^2 + 36} \quad . \tag{2.4.4}$$

Because of the improvement in accuracy, all approximations in the "Durbin F" column of Table 2.2 are values of $\Pr(F_{v_1,v_2} \geq F_c)$.

(ii) Bradley-Terry T-Test

Another approximate test of H_o can be made by employing the previously mentioned Bradley-Terry (1952) T-statistic. As stated earlier, $T = -2\ln \lambda$, where λ is the likelihood ratio statistic corresponding to the null and alternative hypotheses mentioned at the beginning of this section and derived under the assumption that the experiment satisfies the Bradley-Terry model. The test procedure is the same as that for the D-test; namely, compare the observed value of T with $\chi^2_{t-1,\alpha}$, and reject H_o if T is not less than $\chi^2_{t-1,\alpha}$.

There are three reasons for preferring the test based on the D-statistic to the one based on T. First, a more general model may be assumed when the D-test is employed. The next reason is that while there is little practical difference in accuracy, the D-test is generally conservative while the T-test shows the outcome to be slightly more significant than is true. The third reason is that the value of T is usually more difficult to calculate than the value of D since iteration is usually required to find T. (A paper by Dykstra (1956) gives a method for reducing the iteration required in the T-test.)

(iii) The Kendall and Babington Smith Test Based on Circular Triads

The approximate test of H_o proposed by Kendall and Babington Smith (1940) for experiments in which $n = 1$ is based on the number of circular triads occurring in the experiment. The test-statistic is

$$X^2 = \frac{8}{t-4}\left\{\tfrac{1}{4}\binom{t}{3} - c + \tfrac{1}{2}\right\} + \nu \,, \tag{2.4.5}$$

where

(2.4.6) $\qquad \nu = t(t-1)(t-2)/(t-4)^2$

and c is the number of circular triads occurring in the experiment. Solving relation (2.2.7) for c and substituting into (2.4.5), we obtain

(2.4.7) $\qquad X^2 = \frac{1}{t-4}\{4 + t(D-t+1)\} + \nu .$

The X^2-statistic has approximately a χ^2-distribution with ν degrees of freedom for moderately large t. Kendall and Babington Smith used this test-statistic to test the consistency of the judge of the repetition. However, the judge cannot be expected to be consistent unless there are differences in the treatment stimuli.

The test procedure again consists of comparing the observed X^2 with $\chi^2_{\nu,\alpha}$, and rejecting H_o when X^2 is not less than $\chi^2_{\nu,\alpha}$. Because of the fractional values that ν assumes, interpolation is usually necessary to obtain the critical value $\chi^2_{\nu,\alpha}$. A similar difficulty is encountered in calculating the significance level of the observed X^2.

Table 2.2 gives a comparison of the X^2-approximation with the D and F_c-approximations and the exact significance level for two values of D in an experiment of size (8,1). In comparing the accuracy of the D and X^2-approximations, we find the D-approximations are conservative [i.e., larger than $Pr(D \geq D_o | H_o')$], while the X^2-approximations are not large enough. In terms of absolute distance from the true value, it appears that the X^2-approximation is closer when the true significance level is 1% or less, while the D-approximation is closer when the true significance level is near 5%.

To summarize the results of the comparisons of the D-test with the other three approximate tests, one can say that the D-method is always simpler to perform, and is generally at least as accurate for moderate-size experiments, and, of course, sufficiently accurate for large experiments. The possible exception to the above statement is when n is one, or is close to one, in which case, for moderate t, it appears that the more complicated $\tilde{F}_c$-test would give a better approximation.

TABLE 2.2

Comparison of the Bradley-Terry T, the Durbin F, and the Kendall-Babington Smith X^2 Approximations with Our D Approximation and the Exact Significance Level

$t = 3,\ n = 10$

$\frac{1}{4}ntD_o$	$2B_1$*	T**	B-T Approx. $Pr(\chi^2_2 \geq T)$	Durbin $Pr(F_{v_1 v_2} \geq F_c)$	Our D $Pr(\chi^2_2 \geq D)$	Exact $Pr(D \geq D_o \mid H_o')$
128	9.098	20.640	0.0000	-	0.0002	0.0000
126	9.108	20.617	.0000	-	.0002	.0001
122	9.430	19.875	.0000	-	.0003	.0001
114	10.282	17.913	.0001	-	.0005	.0002
104	11.068	16.104	.0003	-	.0010	.0005
98	11.576	14.934	.0006	-	.0015	.0012
96	11.668	14.772	.0006	-	.0017	.0014
86	12.476	12.862	.0016	-	.0032	.0025
78	13.050	11.540	.0031	-	.0055	.0046
74	13.330	10.895	.0043	-	.0072	.0074
72	13.490	10.527	.0052	-	.0082	.0090
62	14.180	8.938	.0115-	0.0136	.0160	.0157
56	14.582	8.012	.0182	-	.0239	.0261
54	14.714	7.708	.0212	-	.0273	.0320
50	14.982	7.087	.0288	-	.0357	.0399
42	15.504	5.889	.0526	.0599	.0608	.0674
38	15.758	5.304	.0705+	-	.0794	.1035
⋮	⋮	⋮	⋮			⋮
8	17.594	1.077	.5842	-	.5866	.6299
⋮	⋮	⋮	⋮			⋮
2	17.964	0.266	.8755	-	.8752	.9644

$t = 4,\ n = 8$

$\frac{1}{4}ntD_o$	$2B_1$*	T**	B-T Approx. $Pr(\chi^2_3 \geq T)$	Durbin $Pr(F_{v_1 v_2} \geq F_c)$	Our D $Pr(\chi^2_3 \geq D)$	Exact $Pr(D \geq D_o \mid H_o')$
98	23.118	13.311	0.0040	-	0.0066	0.0054
96	23.188	13.149	.0043	-	.0074	.0056
94	23.380	12.707	.0053	-	.0081	.0070
90	23.650	12.086	.0071	-	.0104	.0091
88	23.678	12.021	.0073	0.0091	.0117	.0096

TABLE 2.2 (continued)

t = 4, n = 8 (continued)

$\frac{1}{4}ntD_o$	$2B_1$*	T**	B-T Approx. $Pr(\chi^2_3 \geq T)$	Durbin $\tilde{F}$ $Pr(F_{v_1 v_2} \geq \tilde{f}_c)$	Our D $Pr(\chi^2_3 \geq D_o)$	Exact $Pr(D \geq D_o \mid H_o')$
86	23.910	11.487	0.0094	-	0.0132	0.0125
84	23.992	11.298	.0102	-	.0148	.0138
82	24.170	10.888	.0123	-	.0166	.0154
80	24.256	10.690	.0135	-	.0186	.0162
78	24.354	10.465	.0150	-	.0208	.0182
76	24.472	10.193	.0170	-	.0233	.0193
⋮	⋮	⋮	⋮			⋮
62	25.354	8.162	.0427	0.0504	.0515	.0556
⋮	⋮	⋮	⋮			⋮
50	26.086	6.477	.0906	-	.1000	.1100

t = 5, n = 5

$\frac{1}{4}ntD_o$	$2B_1$*	T**	B-T Approx. $Pr(\chi^2_4 \geq T)$	Our D $Pr(\chi^2_4 \geq D_o)$	Exact $Pr(D \geq D_o \mid H_o')$
80	24.014	14.020	0.0072	0.0123	0.0098
78	24.188	13.619	.0086	.0141	.0111
76	24.376	13.187	.0104	.0162	.0128
74	24.548	12.790	.0123	.0186	.0167
⋮	⋮	⋮	⋮	⋮	⋮
60	25.662	10.225	.0368	.0477	.0441
58	25.830	9.839	.0432	.0545	.0526
56	25.986	9.479	.0501	.0622	.0616

t = 8, n = 1

$\frac{1}{4}ntD_o$	Durbin $\tilde{F}$ $Pr(F_{v_1 v_2} \geq \tilde{f}_c)$	Kendall & Smith X^2 $Pr(\chi^2_\nu \geq X^2)$	Our D $Pr(\chi^2_7 \geq D_o)$	Exact $Pr(D \geq D_o \mid H_o')$
⋮				
34	0.0079	0.0042	0.0174	0.0064
⋮	⋮	⋮	⋮	⋮
28	.0438	.0219	.0512	.0370

* The B_1 value used in calculating T is the largest B_1 corresponding to an outcome for which $\sum(a_i-\bar{a})^2 = D_o nt/4$.

** $T = nt(t-1)\ln 2 - 2B_1 \ln 10$.

2.5 Asymptotic Power of D-Test

As mentioned in Section 2.2, Bradley (1955) has shown that the Bradley-Terry T-statistic converges to D as $n \to \infty$. This demonstration of convergence was presented in an argument leading to the asymptotic power of the T-test described in Subsection 2.4(ii). Because of the convergence of T to D, the asymptotic power of the test based on T is the same as that of the test based on D when the Bradley-Terry model presented in Subsection 1.2(ii) is valid.

Bradley (1955) considered the alternative hypothesis

$$H_a : \pi_i = t^{-1} + n^{-\frac{1}{2}}\delta_{in} ,$$

which in our notation is

$$H_a : \pi_{ij} = (n^{\frac{1}{2}} + t\,\delta_{in}) \,/\, [2n^{\frac{1}{2}} + t(\delta_{in} + \delta_{jn})] ,$$

where δ_{in} represents a sequence of constants converging to δ_i as $n \to \infty$ and $\sum_{i=1}^{t} \delta_i = 0$. He found that the limiting distribution of T (and D) under H_a as $n \to \infty$ is a non-central χ^2-distribution with (t-1) degrees of freedom and parameter of non-centrality.

$$\eta = \tfrac{1}{4}t^3 \sum_{i=1}^{t} \delta_i^2 . \tag{2.5.1}$$

Hence, the asymptotic power of the test based on T (or D) is given by

$$\beta(\eta \mid \alpha, t-1, \infty) = \int_{\chi^2_{t-1,\alpha}}^{\infty} f(T)dT$$

where $\chi^2_{t-1,\alpha}$ is defined in Section 2.2 and f(T) is the

density function of the non-central χ^2-variate with (t-1) degrees of freedom and parameter of non-centrality η. It must be remembered that this is the asymptotic power of the D-test only when the Bradley-Terry model is valid.

The asymptotic power of the D-test depends on the distribution of treatment responses. Since the distribution is not specified by our model in Section 2.1, it is not possible to give a general expression for the asymptotic power of the D-test for all possible cases covered by the model.

2.6 Tests of Grouped Repetitions*

The paired-comparison experiment can be partitioned into τ groups of repetitions ($1 \leq \tau \leq n$); each group representing its particular judge, or the time at which it was tested, or any other common factor that changes between groups and might cause a change in the true preference probabilities π_{ij} when the treatments do not produce equal stimuli. If the experimenter wishes to test

$$H_o : \pi_{i.} = \tfrac{1}{2} \qquad (i = 1,\ldots,t),$$

against

$$H_a : \pi_{i.} \neq \tfrac{1}{2} \qquad \text{for some } i,$$

he must decide before the experiment whether or not there is a possibility of a change in preference probabilities between groups.

---

* Throughout the remainder of this chapter, Greek subscripts will refer to groups; Roman subscripts, to treatments.

(i) Pooled Analysis

If the experimenter decides before the experiment that there is no reason to expect changes in the π_{ij} between groups of repetitions, he should perform the test described in Section 2.2 provided the experiment is sufficiently large for that type of test. Of course, an exact test is available if the experiment lies within the range of the tables of distributions.

(ii) Combined Analysis

If the experimenter decides before the experiment that the true preference probabilities might change between groups of repetitions when the null hypothesis is not true, he should not use a pooled analysis. If the combinations (t, n_γ), where n_γ is the number of repetitions in group $\gamma (\gamma = 1, \ldots, T)$, lie within the range of the Bradley-Terry (1952) and Bradley (1954a) tables, and if the experiment can be assumed to satisfy the Bradley-Terry model, the exact combined analysis test described by Bradley and Terry (1952) should be used. If, however, the combinations do not lie within the range of the above tables and are of such size that individual D tests can be performed on each group, a combined analysis should be employed that makes use of our knowledge of the D statistic.

Calculate D for each group γ, call it D_γ, and let

$$(2.6.1) \qquad D_c = \sum_{\gamma=1}^{T} D_\gamma = \sum_{\gamma=1} \Big[4 \Big\{ \sum_{i=1}^{t} a_i^2 - \tfrac{1}{4} n_\gamma^2 \, t(t-1)^2 \Big\} / (n_\gamma \, t) \Big],$$

where $a_{i\gamma}$ is the score of treatment i in group γ. Since the D_γ's are independent and approximately χ^2-variates with (t-1) degrees of freedom, D_c is distributed approximately

as a χ^2-variate with $r(t-1)$ degrees of freedom. Therefore, the combined analysis test, for the case where the individual D_γ 's are approximately distributed as χ^2-variates, consists of comparing D_c with the $100(1-\alpha)$ percentile of the χ^2-distribution with $r(t-1)$ degrees of freedom, and rejecting the null hypothesis if D_c is not less than $\chi^2_{(t-1),\alpha}$. As before, the test becomes conservative in the non-linear model situation.

The necessity for a combined analysis is illustrated by the following extreme hypothetical example.

Example 2.1: Consider the experiment with two judges, representing two groups; each making 20 repetitions of paired comparisons on three varieties of apples according to their preference of the flavor. The null hypothesis is that there is no difference in flavor between the three varieties.

Now if the judges can detect that variety A is sweet, variety B is mild, and variety C is sour, and if judge 1 prefers sweet apples while judge 2 prefers sour apples, one might expect the following scores:

$a_{A_1} = 40$ $\qquad$ $a_{A_2} = 0$

$a_{B_1} = 20$ $\qquad$ $a_{B_2} = 20$

$a_{C_1} = 0$ $\qquad$ $a_{C_2} = 40$

If the groups were pooled, the resulting value of the D-statistic would be zero and this would lead to the false conclusion that one must fail to reject H_o with the evidence at hand.

If the combined analysis were used on the data, one would find

$$D_c = \sum_{\gamma=1}^{2} D = \sum_{\gamma=1}^{2} [4\{\sum_{i=A}^{C} a_i^2 - \tfrac{1}{4}\cdot 20^2\cdot 3\cdot 2^2\}/20\cdot 3]$$

$$= 106\tfrac{2}{3}\ .$$

This is, of course, a highly significant value of the χ^2-variate with four degrees of freedom, and therefore the null hypothesis would be rejected.

If there is no reason to expect changes in the true preference probabilities between groups, one would lose power and accuracy using the combined analysis instead of the pooled analysis. Accuracy is lost in summing approximations. Power is lost in comparing the observed value of the test statistic with a χ^2-variate with Γ times as many degrees of freedom. For instance, if the judges in Example 2.1 had been instructed to prefer the apple that is sweeter, redder, firmer, etc., the pooled analysis would have been the proper one to use and would have given a more significant result than the combined analysis.

Bradley and Terry (1952) have also suggested a test statistic based on their model of the type $(-2\ln \lambda)$ for the combined analysis situation. As in the case of their T-statistic, the value of the statistic is generally more difficult to calculate than D_c. A comparison of the accuracy of these two methods was not feasible, but, on the basis of simplicity, the approximate combined analysis test based on the D-statistic is to be preferred.

2.7 Extension of the Previous Test Methods to Other Designs

The paired-comparison experiment is a ranking method in balanced incomplete block designs with blocks of size two. The two test methods, pooled analysis and combined analysis, can be extended to ranking methods in balanced incomplete

block designs with block size $k(2 \leq k \leq t)$. (Of course, we are dealing with a randomized block design when $k = t$.)

The number of times each treatment is ranked is denoted by m. The treatments in each block are ranked in order of preference. Let y_i $(i=1,\ldots,t)$ denote the sum of the ranks obtained by treatment i in the experiment. [Note that for $k = 2$, $\sum(a_i-\bar{a})^2 = \sum(y_i-\bar{y})^2$.] Friedman (1937) has shown for the randomized block design, and Durbin (1951), for the incomplete block design, that the statistic

$$(2.7.1) \qquad G = \frac{12(t-1)\sum_{i=1}^{t}(y_i-\bar{y})^2}{mt(k^2-1)} \quad *,$$

is under the null hypothesis of equivalent treatment stimuli distributed approximately as a χ^2-variate with (t-1) degrees of freedom for moderate t and large m.

Hence, what has been referred to as a pooled analysis in testing

H_o : the treatment stimuli are all equal,

against

H_a : the treatment stimuli are not all equal,

is in this general case a comparison of the observed G with $\chi^2_{t-1,\alpha}$. If G is greater than the critical value, we accept H_a. This test by Friedman and Durbin is, of course, a large-experiment procedure.

---

* G is the statistic referred to in Sections 2.2 and 2.3 as Durbin's general statistic. When k=2, we have m=n(t-1) and, hence, $G = \frac{12\cdot(t-1)\sum(y_i-\bar{y})^2}{n(t-1)\cdot t\cdot 3} = \frac{4}{nt}\sum(a_i-\bar{a})^2 = D$.

If the experimenter suspects that a change in stimuli relationships may occur between groups of replications, he should make each group sufficiently large that the G for each group is distributed approximately as a χ^2-variate with (t-1) degrees of freedom. Let G_γ ($\gamma = 1,\ldots,\Gamma$) be the value of G calculated from the results in group γ. Then

$$(2.7.2) \qquad G_c = \sum_{\gamma=1}^{\Gamma} G$$

is distributed approximately as a χ^2-variate with $\Gamma(t-1)$ degrees of freedom. Hence, the combined analysis in this case consists of accepting H_a when G_c is not less than $\chi^2_{\Gamma(t-1),\alpha}$, and otherwise accepting H_o.

2.8 Test of Agreement Between Groups

To test the agreement of preference probabilities between groups, that is, to test

$$H_o : \pi_{ij\gamma} = \pi_{ij} \qquad (\gamma = 1,\ldots,\Gamma;\ i,j = 1,\ldots,t),$$

against

$$H_a : \pi_{ij\gamma},$$

where $\pi_{ij\gamma}$ represents the probability that treatment i will be preferred to treatment j in group γ, one can make use of the theory of χ^2 homogeneity tests. Consider the 2 x Γ tables,

(2.8.1)

		GROUPS				
		1	2	...	Γ	Totals
TREATMENTS	i	x_{ij1}	x_{ij2}	...	$x_{ij\Gamma}$	x_{ij}
	j	x_{ji1}	x_{ji2}	...	$x_{ji\Gamma}$	x_{ji}
Totals		n_1	n_2	...	n_Γ	n

where $i < j$ and $x_{ij\gamma}$ is the number of times treatment i is preferred to treatment j in group γ. Under the null hypothesis, the probability that treatment i is preferred to treatment j will be the same for each group. Hence, by the theory of the χ^2 homogeneity test, we have that

$$(2.8.2) \quad C_{ij} = \sum_{\gamma=1}^{} \frac{(x_{ij\gamma} - x_{ij}n_\gamma/n)^2}{x_{ij}n_\gamma/n} - \frac{(x_{ji\gamma} - x_{ji}n_\gamma/n)^2}{x_{ji}n_\gamma/n}$$

$$= n^2 \sum_{\gamma=1}^{} \frac{(x_{ij\gamma} - x_{ij}n_\gamma/n)^2}{n_\gamma x_{ij}(n-x_{ij})}$$

tends to be distributed as a χ^2-variate with $(r-1)$ degrees of freedom as the expected cell frequencies become large. Since the C_{ij}'s $(i < j)$ are independent variates, we have

$$(2.8.3) \quad C_T = \sum_{i=1}^{t-1} \sum_{j=i+1}^{t} C_{ij} \approx \chi^2_{\frac{1}{2}t(t-1)(r-1)},$$

where $\approx$ is read "is asymptotically distributed as".

The approximate test of H_o is now seen to be the comparison of C_T with the α level significance point of the χ^2-variate with $\binom{t}{2}(r-1)$ degrees of freedom; if C_T is not less than $\chi^2_{\binom{t}{2}(r-1);\alpha}$, the null hypothesis is rejected.

A possible difficulty that may arise in the use of this test is that the expected cell frequencies may be small. A paper by Cochran (1954) gives some recommendations on how small the minimum expected cell frequencies may be in various possible cases, all pertaining to the χ^2-test for a single contingency table. The amount that these restrictions might be relaxed when summing the test statistics of several tables is unknown. One should be dubious of significant results obtained by the above χ^2-test when Cochran's recommendations are not met by some of the tables.

Cochran (1954) suggests that when the number of degrees of freedom is greater than 30, and the expected cell frequencies are too small for an accurate χ^2-test, one should consider the test-statistic to be normally distributed. Haldane (1939) has derived expressions for the exact mean and variance of the test-statistic. In our case, these expressions become

$$(2.8.4) \qquad E(C_{ij}) = (r-1)n/(n-1) ,$$

and

$$(2.8.5) \qquad \mathrm{Var}(C_{ij}) = \frac{n^2}{(n-1)^2(n-2)(n-3)} \Big[2(r-1)n^2 + 2r(2r+1)n - 6r^2 - 6n(n-1)\sum_{\gamma=1}^{r} n_\gamma^{-1} + \frac{n(n-1)}{x_{ij}x_{ji}} \Big\{ n(n+1)\sum_{\gamma=1}^{r} n_\gamma^{-1} - (r^2 + 2r - 2)n + r(r-2) \Big\} \Big] .$$

When $n_\gamma = n/r$ ($\gamma = 1,\ldots,r$), the expression for the variance reduces to

$$(2.8.6) \qquad \mathrm{Var}(C_{ij}) = \frac{2(r-1)n^3(n-r)}{(n-1)^2(n-2)(n-3)} \Big[1 - (n-1)/(x_{ij}x_{ji}) \Big] .$$

The statistic

$$(2.8.7) \qquad Z = \frac{C_T - \binom{t}{2}(r-1)n/(n-1)}{\Big[\sum_{i<j} \mathrm{Var}\,(C_{ij}) \Big]}$$

is asymptotically distributed as a N(0,1)-variate as $\binom{t}{2}(r-1) \to \infty$. When the number of degrees of freedom,

$\binom{t}{2}(r-1)$, exceeds 30 and the χ^2-test is not applicable, it appears best to follow Cochran's suggestion and compare Z with the $100(1-\alpha)$ percentile of the N(0,1)-distribution. If Z is not less than the critical value, the alternative hypothesis that the comparison probabilities change between groups of repetitions is accepted.

Another facet of the low expected cell frequency difficulty is the extreme case in which $x_{ij} = 0$ or n. When this occurs, C_{ij} is undefined, and the tests must either be abandoned or modified. If only a few of the C_{ij}'s are undefined when $\binom{t}{2}$ is large, one might exclude the undefined C_{ij}'s and redefine the test-statistic to be

$$C_T' = \sum_{i<j}{}' C_{ij} \qquad (2.8.8)$$

where $\sum'$ denotes summation over defined C_{ij}'s only. If the remaining contingency tables have sufficiently high expected cell frequencies, one might compare C_T' with $\chi^2_{[\binom{t}{2} - K](r-1);\alpha}$ (where K is the number of undefined C_{ij}'s) to test the null hypothesis. If the remaining C_{ij}'s are such that the Z-test should be used, we would redefine the test statistic to be

$$Z' = \frac{C_T' - [\binom{t}{2} - K][(r-1)n/(n-1)]}{\sqrt{\sum' \text{Var } C_{ij}}} \cdot \qquad (2.8.9)$$

The reduced tests are liable to show significance too often since we are taking out contingency tables that indicate no change between groups. Hence, one should be suspicious of significant results when the reduced procedures are employed.

Bradley and Terry (1952) proposed an approximate test of H_o based on a statistic of the form $(-2 \ln \lambda)$, where λ is

again the likelihood ratio statistic for the null and alternative hypotheses under consideration. Their test is performed by comparing the observed value of $(-2 \ln \lambda)$ with the critical value of the χ^2-distribution with $(t-1)(r-1)$ degrees of freedom. As with all tests of this form, the experiment must be large before reasonable accuracy can be expected. Due to lack of knowledge of the exact distribution, a comparison of the accuracy of the test based on $(-2 \ln \lambda)$ with those based on C_T is not feasible. However, due to the iteration generally needed to find $(-2 \ln \lambda)$, the tests based on C_T are preferable with regard to ease of computation.

III. TESTS OF INDIVIDUAL TREATMENTS AND GROUPS OF TREATMENTS

3.1 Introduction

Answers to the following five questions will be proposed in this chapter.

1. If, prior to the experiment, one of the t treatments in the paired-comparison experiment is of particular interest to the experimenter, how can the experimenter use the results to test whether the treatment is better (worse, different) than the average of the t treatments?

2. If, before the experiment, there is a special interest in whether or not two specified treatments produce different stimuli, how does one use the results of the t-treatment paired-comparison experiment to test for the difference?

3. How does one test whether the treatment receiving the highest (lowest) score in the paired-comparison experiment is significantly better (worse) than the average?

4. How does one separate the t treatments in a paired-comparison experiment into significantly different groups?

5. After rejecting the null hypotheses, $H_o : \pi_{i.} = \frac{1}{2}(i=1,\ldots,t)$, how does one determine whether a contrast of treatment scores is significant?

Scheffé (1952) has used Tukey's test based on allowances to separate the treatment parameters into significantly different groups, and thereby answer question (4), in his analysis of variance approach to the paired-comparisons problem. The answer to question (4) given in this chapter

will be for the more general case in which the degree of preference is not necessarily scaled and the distribution of treatment responses is not necessarily multivariate normal.

The methods that are introduced to answer the above questions rely on the model introduced in Section 2.1. As in Chapter II, the tests are developed under the hypothesis

$$H_o' : \pi_{ij} = \tfrac{1}{2} \qquad (i,j=1,\ldots,t;\ i\neq j),$$

and become conservative when the actual null hypothesis differs from H_o.

3.2 Test of a Pre-Assigned Treatment

Because of cost, or some other characteristic of treatment r ($1 \leq r \leq t$), the experimenter may be particularly interested in knowing whether that treatment is better than average, that is, if

$$(3.2.1) \qquad \pi_{r.} = \sum_{\substack{j=1 \\ j\neq r}}^{t} \pi_{rj} \,/\, (t-1) > \tfrac{1}{2} \ .$$

For this purpose, a method for testing

$$H_o : \pi_{r.} = \tfrac{1}{2} \qquad (i=1,\ldots,t),$$

against

$$H_a : \pi_{r.} > \tfrac{1}{2} \ ,$$

is developed.

Under the hypothesis H_o' mentioned in the previous section, the score of treatment r is a binomial variate with parameters $\left[\frac{1}{2},\ n(t-1)\right]$. If a_r^o is the observed score of treatment r, its significance level under H_o' is

$$(3.2.2) \qquad \Pr(a_r \geq a_r^o | H_o') = 2^{-n(t-1)} \sum_{k=a_r^o}^{n(t-1)} \binom{n(t-1)}{k} .$$

[For all but small experiments, a normal approximation can be used in (3.2.2).]

If $\pi_{r.} = \frac{1}{2}$, but not all the π_{ri}'s are equal, a_r is a generalized binomial variate (see Cramér (1946) Section 16.6) with the same expected value as under H_o', but with a variance that is less than it was under H_o' by the amount

$$n \sum_{i \neq r}^{t} (\pi_{ri} - \tfrac{1}{2})^2 .$$

Therefore, for any positive integer u greater than the mean $\frac{1}{2}n(t-1)$, we have

$$(3.2.3) \qquad \Pr(a_r \geq u | H_o') \geq \Pr(a_r \geq u | H_o) .$$

Hence, a conservative test procedure for testing H_o against H_a is obtained by applying the usual binomial test to the score $a_{r.}$, i.e., accept H_a if $\Pr(a_r \geq a_r^o | H_o) \leq \alpha$.

It should perhaps be pointed out that the fact that a treatment is better than average is not always very meaningful. The reason for this is that one extremely poor treatment can lower the average so far that the second poorest treatment might be better than average.

Conservative tests of H_o against the alternative hypotheses

H_a : treatment r is worse than average, i.e., $\pi_{r.} < \frac{1}{2}$,

and

H_a : treatment r is different from the average, i.e., $\pi_{r.} \neq \frac{1}{2}$,

may also be made by treating a_r as a binomial variate with parameters $[\frac{1}{2}, n(t-1)]$ and making the appropriate binomial test for the considered alternative hypothesis.

3.3 Test of Two Pre-Assigned Treatments

Consider the case in which interest is expressed in testing the difference between treatments r and s before the experiment. David (1959) has developed a method for testing

$$H_o : \pi_{r.} = \pi_{s.} ,$$

against the alternative hypothesis

$$H_a : \pi_{r.} \neq \pi_{s.} ,$$

in paired-comparison experiments with one repetition. We shall generalize his method to the case of n repetitions. The test procedure is again developed under $H_o' : \pi_{ij} = \frac{1}{2}$ for all (i,j).

For any positive integer d, let

(3.3.1) $$Q_{rsd} = \Pr(a_r - a_s = d \mid H_o')$$

$$= \sum_{p=o}^{n} [\Pr \text{ (In all comparisons between treatments r and s, r is preferred k=2p-n more times than is s) x Pr (treatment r is preferred (d-k) more times than is s in comparisons with the other (t-2) treatments)}].$$

$$= \sum_{p=o}^{n} \binom{n}{p} \cdot 2^{-n} \sum_{q=d-2p+n}^{n(t-2)} \binom{n(t-2)}{q} \binom{n(t-2)}{q-d+2p-n} 2^{-2n(t-2)}$$

$$= 2^{3n-2nt} \sum_{p=o}^{n} \binom{n}{p} \sum_{q=d-2p+n}^{n(t-2)} \binom{n(t-2)}{q} \binom{n(t-2)}{q-d+2p-n}.$$

Through application of the convention

$$\binom{a}{b} = 0 \quad \text{when } b < 0,$$

we may write

$$(3.3.2)\quad \sum_{q=d-2p+n}^{n(t-2)} \binom{n(t-2)}{q}\binom{n(t-2)}{q-d+2p-n} = \sum_{q=o}^{n(t-2)} \binom{n(t-2)}{q}\binom{n(t-2)}{q-d+2p-n}.$$

It should be observed that the R.H.S. of (3.3.2) is the coefficient of $x^{-d+2p-n}$ in the expansion of the L.H.S. of the identity

$$(3.3.3)\quad (1+x)^{n(t-2)}\cdot(1+x^{-1})^{n(t-2)} \equiv \frac{(1+x)^{2n(t-2)}}{x^{n(t-2)}} .$$

An expansion of the R.H.S. of the identity shows that the coefficient of $x^{-d+2p-n}$ can also be expressed as

$$(3.3.4)\quad \binom{2n(t-2)}{n(t-2)-d+2p-n}$$

Therefore,

$$(3.3.5)\quad \sum_{q=d-2p+n}^{n(t-1)} \binom{n(t-2)}{q}\binom{n(t-2)}{q-d+2p-n} = \binom{2n(t-2)}{n(t-3)-d+2p} ,$$

and, substituting into (3.3.1), we have

$$(3.3.6)\quad Q_{rsd} = 2^{3n-2nt} \sum_{p=o}^{n} \binom{n}{p}\binom{2n(t-2)}{n(t-3)-d+2p} .$$

Now, for any positive integer m, let

$$(3.3.7)\quad P_{rsm} = \tfrac{1}{2}Pr(|a_r-a_s| \geq m/H_o) = Pr(a_r-a_s \geq m/H_o)$$
$$= \sum_{d=m}^{n(t-1)} Q_{rsd} .$$

If $\pi_{r\cdot} = \pi_{s\cdot}$, but not all the π_{ri} and π_{sj} are $\frac{1}{2}$, the variance of $(a_r - a_s)$ is less than under H_o'*, while its mean is the same as under H_o', namely, zero. Therefore,

$$(3.3.8) \quad \Pr(|a_r - a_s| \geq m \mid H_o) \leq \Pr(|a_r - a_s| \geq m \mid H_o') \ .$$

A conservative test of H_o against H_a can now be described by the following four-step procedure.

1. Specify the significance level α.
2. Find m_c, the smallest value of m for which $2P_{rsm}$ does not exceed α.
3. Calculate $|d| = |a_r - a_s|$.
4. Accept H_a if $|d|$ is not less than m_c.

To test H_o against the new alternative hypothesis

H_a : the stimulus produced by treatment r is preferable to that produced by treatment s, i.e., $\pi_{r\cdot} > \pi_{s\cdot}$,

a one-sided test is necessary. The four-step procedure for the one-sided test is as follows:

1. Specify the significance level α.
2. Find m_c', the smallest value of m for which P_{rsm} does not exceed α.
3. Calculate $d = a_r - a_s$.
4. Accept H_a if d is not less than m_c'.

Again this test is conservative.

Since the results from different repetitions are stochastically independent, the characteristic function of

$$(3.3.9) \qquad d = a_r - a_s$$

is n times the characteristic function of the difference d_1(say) of the scores of treatments r and s in a single

---

* See Appendix Section B.

repetition. The characteristic function of d_1 under H_o' was found by David (1959) to be

$$(3.3.10)\quad \phi_{d_1}(u) = (\cos\tfrac{1}{2}u)^{2t-4}\cos u \,.$$

Hence,

$$(3.3.11)\quad \phi_{d}(u) = (\cos\tfrac{1}{2}u)^{2n(t-2)}\cos^n u \,.$$

From (3.3.11), one finds $E(d) = 0$ and $E(d^2) = \frac{1}{2}nt$. Now consider the characteristic function of $d/\sqrt{(\frac{1}{2}nt)}$,

$$(3.3.12)\quad \phi_{d/\sqrt{(\frac{1}{2}nt)}}(u) = \phi_d\left[u/\sqrt{(\tfrac{1}{2}nt)}\right]$$

$$= (\cos\tfrac{1}{2}u\sqrt{2/nt})^{2n(t-2)}(\cos u\sqrt{2/nt})^n$$

$$= \left[1 - \frac{u^2}{2\cdot 4}\cdot\frac{2}{nt} + O\left(\frac{1}{t^2n^2}\right)\right]^{2n(t-2)}$$

$$\cdot\left[1 - \tfrac{1}{2}u^2\cdot 2/nt + O\left(\frac{1}{t^2n^2}\right)\right]^n .$$

Taking the limit as $n\to\infty$,

$$(3.3.13)\quad \lim_{n\to\infty}\phi_{d/\sqrt{\frac{1}{2}nt}}(u) = e^{\frac{-t+2}{t}\frac{u^2}{2}}\cdot e^{\frac{-u^2}{t}} = e^{-\frac{1}{2}u^2}.$$

Also,

$$(3.3.14)\quad \lim_{t\to\infty}\phi_{d/\sqrt{\frac{1}{2}nt}}(u) = e^{-\frac{1}{2}u^2}\cdot 1 \,.$$

Hence, the distribution of d tends to the normal distribution with parameters $(0,\frac{1}{2}nt)$. As David (1959) pointed out, this trend toward the normal distribution is quite rapid.

Table 3.1 gives the critical values of d for small values of n and t when the desired significance levels are 0.01 and 0.05. For these two significance levels one may use the normal approximation with continuity correction to obtain the critical values of d when the experiment is outside the range of the table. Jordan's (1932) table of the values of $\binom{m}{n}$ for n = 1(1)10, m = 1(1)110 is often useful in evaluating P_{rsm} to obtain the significance level of an observed m under H_o'.

TABLE 3.1

Critical Values of m.

Experiment Size		$\alpha = 0.01$		$\alpha = 0.05$	
		one-sided test	two-sided test	one-sided test	two-sided test
n	t	m'_c	m_c	m'_c	m_c
1	≤ 4	no significant values		no significant values	
1	5	4	none possible	4	4
1	6	5	5	4	4
1	7	5	5	4	5
1	8	5	6	4	5
1	9	6	6	4	5
1	10	6	7	5	5
1	11	6	7	5	6
1	12	7	7	5	6
1	13	7	7	5	6
1	14	7	8	5	6
1	15	7	8	5	6
1	16	7	8	6	6
2	3	no significant values		4	4
2	4	5	6	4	5
2	5	6	6	5	5
3	3	6	6	4	5
3	4	6*	7	5	6
4	3	6**	7	5	6
4	4	7	8	6	6
All larger values of n or t		$\left[m'_c = \text{smallest integer} \geq 2.33\sqrt{\frac{1}{2}nt} + 0.5\right]$	$\left[m_c = \text{smallest integer} \geq 2.56\sqrt{\frac{1}{2}nt} + 0.5\right]$	$\left[m'_c = \text{smallest integer} \geq 1.64\sqrt{\frac{1}{2}nt} + 0.5\right]$	$\left[m_c = \text{smallest integer} \geq 1.96\sqrt{\frac{1}{2}nt} + 0.5\right]$

* $P_{rs6} = 0.0103$

** $P_{rs6} = 0.01001$

3.4 A Test of the Highest Score

After running a paired-comparison experiment, the experimenter may wish to know whether the treatment with the highest score is significantly better than average. If the treatment with the highest score is denoted as treatment (1), one tests the null hypothesis

$$H_o : \pi_{(1)\cdot} = \tfrac{1}{2} ,$$

against the alternative hypothesis

$$H_a : \pi_{(1)\cdot} > \tfrac{1}{2} .$$

As in previous sections, the test will be developed under $H_o' : \pi_{ij} = \frac{1}{2}$ for all (i,j), and then applied to the above case. To perform a test of H_o' against H_a one needs information concerning at least the critical part of the distribution of the largest score, $a_{(1)}$, under H_o.

If we let A_i be the event $a_i \geq m$ $[0 \leq m \leq n(t-1)]$, then the elementary law of probability concerning the sum of events, and the fact that the events are symmetric under H_o', yields

$$(3.4.1) \quad \Pr(a_{(1)} \geq m) = \Pr\Big(\sum_{i=1}^{t} A_i\Big) = \sum_{j=1}^{t} (-1)^{j-1} \binom{t}{j} \Pr(A_1 \cdot A_2 \cdots A_j).$$

There are two cases in which it is easy to evaluate $\Pr(a_{(1)} \geq m)$. They are

1. The trivial case, $m = 0$, for which $\Pr(a_{(1)} \geq m) = 1$, and
2. The case, $m > n(t-1) - \frac{1}{2}n$, for which

$$(3.4.2) \quad \Pr(a_{(1)} \geq m) = \binom{t}{1} \Pr(A_i) = t \cdot 2^{-n(t-1)} \sum_{k=m}^{n(t-1)} \binom{n(t-1)}{k},$$

due to the fact that no two treatment scores can simultaneously exceed $\lfloor n(t-1) - \frac{1}{2}n \rfloor$.
When

$$(3.4.3) \quad 0 < m \leq n(t-1) - \tfrac{1}{2}n ,$$

it is difficult to determine the joint probabilities on the R.H.S. of (3.4.1) since the scores are correlated binomial variates. We shall, therefore, use an approximation.

For values of m such that $Pr(A_i) < \frac{1}{t}$, one can apply the Bonferroni inequality to obtain

$$(3.4.4) \quad tPr(A_i) - \binom{t}{2}Pr(A_i \cdot A_j) \leq Pr(a_{(1)} \geq m) \leq tPr(A_i) .$$

Since the sum of the treatment scores is a constant, we have

$$(3.4.5) \quad Pr(A_i | A_j) \leq Pr(A_i) ,$$

and, therefore,

$$(3.4.6) \quad Pr(A_i \cdot A_j) \leq \lfloor Pr(A_i) \rfloor^2 .$$

Substituting $\lfloor Pr(A_i) \rfloor^2$ for $Pr(A_i \cdot A_j)$ in (3.4.4), we have the relation [cf. David (1956)]

$$(3.4.7) \quad tPr(A_i) - \binom{t}{2}\lfloor Pr(A_i) \rfloor^2 \leq Pr(a_{(1)} \geq m) \leq tPr(A_i) .$$

Now, since

$$(3.4.8) \quad Pr(A_i) = 2^{-n(t-1)} \sum_{k=m}^{n(t-1)} \binom{n(t-1)}{k}$$

can be calculated directly or found in a table of the cumulative binomial probability distributions [see Harvard

Univ.(1955)_7, we have limits on $Pr(a_{(1)} \geq m | H_o')$ that are easy to obtain.

To make an approximate α-level significance test of H_o' against H_a, one should choose as the critical value that positive integer value of m, say m_β, for which

$$(3.4.9) \quad tPr(a_i \geq m_\beta \,|H_o') = \beta \leq \alpha < tPr(a_i \geq m_\beta - 1 | H_o') .$$

If the highest score, $a_{(1)}$, is not less than m_β, one should reject H_o' and conclude that the treatment with score $a_{(1)}$ is better than average.

Since $Pr(A_i)$ is multiplied by the factor t in relations (3.4.7) and (3.4.9), the error in these relations, due to the use of the normal approximation to the binomial to find $Pr(A_i)$, would be t times the error in the normal approximation. For this reason, the normal approximation should not be used in (3.4.7) to find the limits on $Pr(a_{(1)} \geq m)$ nor in (3.4.9) to obtain m_β.

When $\beta = 0.05$ and t is large, we find from relations (3.4.7) and (3.4.9) that

$$0.04875 \leq Pr(a_{(1)} \geq m_{(0.05)} | H_o') \leq 0.05 .$$

The range between the limits decreases as t or β or both decrease. Hence, the true significance level of the above approximate test is known to be between quite narrow bounds.

If more accuracy is necessary in testing H_o' against H_a, one might apply a generalization of a test suggested by David (1959) for the case n = 1. The generalized form of his joint density function of $s (s \leq t)$ scores is

$$(3.4.10) \quad f(a_i, a_j, \ldots, a_k) = 2^{-\frac{1}{2}ns(2t-s-1)} \sum_{|p|}^{i,j,\ldots,k} \binom{t-s}{a_p - a_p'} ,$$

where a'_p is the score of treatment p in a sub-experiment between the s treatments, and the sum is over all outcomes of sub-experiments compatible with the final score. By use of (3.4.10) to find the joint probabilities on the R.H.S. of (3.4.1), one obtains the exact value of $Pr(a_{(1)} \geq m | H'_o)$. With this information, the exact test of H'_o follows in the same way as the approximate test described above. However, it should be observed that this method will become increasingly difficult as n and t increase.

Because the scores a_i change from binomial to generalized binomial variates when we switch from H'_o to

$$H_o : \pi_{(1).} = \tfrac{1}{2} ,$$

we have

$$(3.4.11) \quad Pr(a_{(1)} \geq m_\beta \mid H_o) \leq Pr(a_{(1)} \geq m_\beta \mid H'_o) .$$

From (3.4.11), we see that if the observed $a_{(1)}$ is equal m_β, its significance level under H_o is not greater than β. Hence, the use of m_β as the critical value of $a_{(1)}$ gives a conservative test of H_o against H_a.

A test to determine whether the lowest score is worse than the average will proceed in a manner analogous to the above test of the highest score.

3.5 A Multiple-Range Test of Treatment Scores

To separate significantly unequal treatment stimuli, a multiple-range test-procedure is introduced for paired-comparison experiments that is analogous to Tukey's test based on allowance [see Federer (1955) p.297] for separating significantly different independent normally distributed sample means. Tukey's test uses the analysis of variance table based on k independent samples of size p from normal

populations with equal variances to obtain an estimate of the standard error of a mean, $s_m = \sqrt{\frac{\text{error sum of squares}}{rp}}$, with r degrees of freedom. After obtaining this estimate, the test consists of performing the following three operations.

1. Choose a significance level, α.
2. Compute the value of $M = \sqrt{2}\, s_m\, q_{m,\alpha}$, where $q_{m,\alpha}$ is the 100α % point of the studentized range with r degrees of freedom for a sample of size k.
3. If the difference between any two sample means exceeds M, it will be concluded that the two samples are from populations with different means.

The probability of declaring at least one difference to be significant when all the samples are in fact from the same normal population is α.

The corresponding multiple range method for a paired-comparison experiment is as follows:

1. Choose a significance level, α.
2. Find a positive integer $R_{\beta(\alpha)}$ (say) such that

(3.5.1)
$$\Pr\left[\bar{a}_{(1)} - a_{(t)} \geq R_{\beta(\alpha)}\right] = \beta \leq \alpha < \Pr\left[\bar{a}_{(1)} - a_{(t)} \geq (R_{\beta(\alpha)} - 1)\right],$$

where $a_{(i)}$ denotes the i th largest score, and the probabilities are calculated under the hypothesis

$$H_0' : \pi_{ij} = \tfrac{1}{2} \qquad (i \neq j;\ i,j = 1,\ldots,t).$$

3. Any pairwise difference in the scores not less than $R_{\beta(\alpha)}$ is considered significant.

To declare two scores significantly different is the same as stating that the corresponding treatment stimuli are not equal. Hence, a significant difference is a rejection of the null hypothesis, H_o, of treatment equality. Since the probabilities in step 2 are calculated under H_o' and the variance of (a_i-a_j) is a maximum under H_o', the above test procedure is conservative when H_o is not equivalent to H_o', that is, when the model is non-linear.

Step 2 of the above test procedure requires a knowledge of the distribution of the range, $a_{(1)}-a_{(t)}$, of the t scores under H_o'. There are t(t-1) differences of the form (a_i-a_j); we are interested in the distribution of the maximum of the t(t-1) differences. Equation (3.4.1) gives the probability that the maximum of t scores will not be less than a given quantity as the sum of k-fold (k=1,...,t) joint probabilities. If we let R_{ij} be the event $a_i-a_j \geq R$, where R is a positive integer, then the substitution of t(t-1) and the R_{ij}'s in the R.H.S. of (3.4.1) for t and the A_i's respectively, gives an expression for $Pr(a_{(1)}-a_{(t)} \geq R)$. When

$$(3.5.2) \qquad R > n(t-1) - \tfrac{1}{2}n \ ,$$

the R_{ij}'s are mutually exclusive, and, since every R_{ij} is equally likely under H_o',

$$(3.5.3) \qquad Pr(a_{(1)}-a_{(t)} \geq R) = t(t-1)Pr(R_{ij}) \ .$$

Or, by (3.3.6) and (3.3.7), when R satisfies condition (3.5.2)

$$(3.5.4) \qquad Pr(a_{(1)}-a_{(t)} \geq R)$$

$$= t(t-1)\cdot 2^{3n-2nt} \sum_{d=R}^{\frac{n(t-1)}{2}} \sum_{p=o}^{n} \binom{n}{p} \binom{2n(t-2)}{n(t-3)-R+2p} \ .$$

However, to find $Pr(a_{(1)}-a_{(t)} \geq R)$ when

$$0 < R \leq n(t-1) - \tfrac{1}{2}n ,$$

it is necessary to evaluate joint probability terms of the form $Pr(R_{ij} \cdots R_{gh})$. Rather than attempt this difficult task, a relatively simple approximation will be developed.

Some observations should be made before considering the approximation. One is that $Pr(a_{(1)}-a_{(t)} \geq R)$, by virtue of the Bonferoni inequality, will never exceed the value of the expression on the R.H.S. of (3.5.4) for any value of R. Values of the R.H.S. of (3.5.4) are presented in Table 3.2 in the "U(R)" column. Also, values of R for which $Pr(a_{(1)}-a_{(t)} \geq R)$ is in the 0.005 to 0.075 interval, were obtained from the tables of Bradley and Terry (1952), Bradley (1954a), and David (1959), and are listed in Table 3.2 along with their exact significance levels. These significance levels, obtained from the tabled distributions of scores, are listed in the "$Pr(a_{(1)}-a_{(t)} \geq R | H_o')$" column of Table 3.2. Since, we are able to obtain the significance level of R for these small experiments, we need only require that an approximation to $Pr(a_{(1)}-a_{(t)} \geq R)$ be reasonably accurate in the critical part of the distribution of the range for experiments of moderate and large sizes.

The d_i (i=1,...,t) variates discussed in Section 2.2 have equal variances, $\sigma^2 = (t-1)/t$, equal correlations, $\rho = -1/(t-1)$, and limiting normal distributions as $n(t-1) \to \infty$. We may expect the distribution of the range [cf. Hartley (1950)]

$$(3.5.5) \qquad d_{(1)}-d_{(t)} = \sqrt{\tfrac{4}{nt}}(a_{(1)}-a_{(t)})$$

to be asymptotically the same as that of t independent

observations from a normal population with variance

$$(3.5.6) \qquad \sigma^2(1-\rho) = (t-1)/t \left[1 + 1/(t-1)\right] = 1 .$$

The c.d.f. of the range W_t (say) of t independent observations (t = 2(1)20) from a normal population with variance one is given as Table 23 in the Biometrika Tables for Statisticians, Vol. I., by Pearson and Hartley (1954). Hence, $\Pr(W_t \geq \sqrt{\frac{4}{nt}}\, R)$ is an easily obtained approximation to $\Pr(a_{(1)}-a_{(t)} \geq R)$ and will increase in accuracy as n(t-1) becomes larger.

The range W_t is a continuous variate; whereas, the range $(d_{(1)}-d_{(t)})$ is a discrete variate with distinct values differing by not less than $\sqrt{4/nt}$. This suggests that a continuity correction might improve the approximation. The usual correction would subtract $\frac{1}{2}\sqrt{\frac{4}{nt}}$ from the observed range of the d_i's, $\sqrt{\frac{4}{nt}}\, R$. However, it has been found empirically that more accurate results are obtained in the interval of significance levels (0.005, 0.075) through the subtraction of $\frac{1}{4}\sqrt{\frac{4}{nt}}$ from $\sqrt{\frac{4}{nt}}\, R$.

Table 3.2 gives comparisons between $\Pr\left[W_t \geq \sqrt{\frac{4}{nt}}\,(R-\frac{1}{4})\right]$ and $\Pr(a_{(1)}-a_{(t)} \geq R)$ for some of the larger experiments with tabled distributions. The table indicates that the approximation is conservative for values of R with significance levels in the neighborhood of 1% or less. For such values of R, it is wise to compare the W_t approximation with the upper limit U(R) of $\Pr(a_{(1)}-a_{(t)} \geq R)$, mentioned earlier, and use the smaller value as the significance level of R.

Based on Table 3.2 and the asymptotic property of the distribution of $(d_{(1)}-d_{(t)})$, it is safe to predict that the W_t approximation will give at least two place accuracy when the true significance level is in the interval

(0.005, 0.075) and $n(t-1) \geq 20$. Even for an experiment as small as $t = 8$, $n = 1$, we find that for $R = 6$ the W_t approximation differs from the true significance level of 0.0738 by only 0.0040. For $R = 7$, in the experiment of size (8,1), one would not use the W_t approximation since $R > n(t-1) - \frac{1}{2}n$, and, therefore, the exact value of $Pr(a_{(1)}-a_{(t)} \geq R)$ can be easily obtained by calculating the value of $U(R)$.

The method of finding $R_{\beta(\alpha)}$ and β, or an approximation to β, can be summarized as follows:

a. If the experiment is sufficiently small, $R_{\beta(\alpha)}$ and β may be obtained from the "R" and "$Pr(a_{(1)}-a_{(t)} \geq R|H_o')$" columns of Table 3.2, or calculated directly from the tables of Bradley and Terry (1952), Bradley (1954a), and David (1959).

b. If the experiment is too large for method (1), find the $100\alpha\%$ point, $W_{t(\alpha)}$ say, of the W_t-distribution (Biometrika Tables, Table 22) and solve

$$W_{t(\alpha)} = \sqrt{\frac{4}{nt}}\,(R^* - \tfrac{1}{4}) \tag{3.5.7}$$

for R^*. If R^+, the smallest integer not less than R^*, is larger than $\lceil n(t-1) - \frac{1}{2}n \rceil$, calculate $U(R^+)$, $U(R^+-1)$,..., $U(R^+-p)$, where p is such that

$$U(R^+ - p + 1) \leq \alpha < U(R^+ - p). \tag{3.5.8}$$

Then $R_{\beta(\alpha)} = R^+ - p + 1$ and β will either equal $U(R^+ - p + 1)$ or be approximately equal to it depending on whether or not $(R - p + 1)$

is greater than $\lfloor \overline{n}(t-1) - \frac{1}{2}n \rfloor$. The value of p will seldom be greater than one, and, therefore, we usually find $R_{\beta(\alpha)} = R^+$ and $\beta = U(R^+)$.

c. If the experiment is too large for method (1) and W_t is such that R^+ is not greater than $\lfloor \overline{n}(t-1) - \frac{1}{2}n \rfloor$, $R_{\beta(\alpha)} = R^+$ and β is approximately equal to

$$\Pr\lfloor W_t \geq \sqrt{(4/nt)(R^+ - \tfrac{1}{4})} \rfloor .$$

However, if R^+ and $\lfloor \overline{n}(t-1) - \frac{1}{2}n \rfloor$ differ by less than ten, one should also calculate $U(R^+)$ and use

$$\min \left\{ \Pr\lfloor W_t \geq \sqrt{(4/nt)(R^+ - \tfrac{1}{4})} \rfloor, \; U(R^+) \right\}$$

as the estimate of β.

As an example consider the paired-comparison experiment with five treatments and ten repetitions for which the scores are

$$a_1 = 15, \quad a_2 = 10, \quad a_3 = 30, \quad a_4 = 20, \quad a_5 = 25.$$

We will consider a 5% significance level test. Method (a) cannot be used to find $R_{\beta(\alpha)}$ because the experiment is too large. From the Biometrika Tables (1954, Table 22), we obtain $W_{5(0.05)} = 3.86$. Hence,

$$R^* = \sqrt{(10 \times 5/4)} \times 3.86 + \tfrac{1}{4} = 13.897,$$

and

$$R^+ = 14 < n(t-1) - \tfrac{1}{2}n = 35 .$$

Since R^+ is 21 less than $\lfloor \overline{n}(t-1) - \frac{1}{2}n \rfloor$, we use method (c) which gives $R_{\beta(0.05)} = 14$ and $\beta \doteqdot \Pr\lfloor \overline{W}_t \geq \sqrt{(4/50)}x(14 - \frac{1}{4}) \rfloor$. From the Biometrika Tables (Table 23), we obtain $\beta \doteqdot 0.0464$. Now, applying step 3 of the test procedure, we have

10	15	20	25	30
a_2	a_1	a_4	a_5	a_3

where any two scores not underlined by the same line are declared significantly different.

TABLE 3.2

Approximations to $Pr(a_{(1)} - a_{(t)} \geq R)$

t	n	R	$Pr(a_{(1)}-a_{(t)} \geq R \mid H_o')$**	U(R)	$Pr(W_t \geq \sqrt{\frac{4}{nt}(R-\frac{1}{4})})$
3	1-2	(No possible range R exists for which significance level is less than 0.09.)			
	3	6*	0.0117	0.0117	
	4	8*	.0015	.0015	
	4	7*	.0132	.0132	
	4	6	.0513	.0615	
	5	9*	.0020	.0020	
	5	8*	.0112	.0112	
	5	7	.0386	.0423	
	6	9	.0082	.0086	
	6	8	.0282	.0294	
	7	10	.0062 ± .00003	.0064	
	7	9	.0185	.0203	
	7	8	.0504	.0545	
	8	11	.0046	.0046	
	8	10	.0131 ± .00003	.0139	
	8	9	.0324	.0364	
	9	11	.0093	.0096	
	9	10	.0226 ± .00003	.0245	
	9	9	.0490	.0564	
	10	12	.0062 ± .00003	.0066	0.0078
	10	11	.0157	.0165	.0152
	10	10	.0340 ± .00003	.0378	.0317
4	1	(No possible range R exists for which the significance level is less than 0.1.)			
	2	6*	0.0116 ± .00003	0.0117	
	3	8*	.0048 ± .00004	.0048	
	3	7	.0279 ± .00004	.0300	
	4	9	.0085 ± .00006	.0088	
	4	8	.0314 ± .00004	.0351	

TABLE 3.2 (continued)

t	n	R	$Pr(a_{(1)}-a_{(t)} \geq R \mid H_o')$**	U(R)	$Pr(W_t \geq \sqrt{\frac{4}{nt}}(\bar{R}-\frac{1}{4}))$
4	5	10	0.0097 ± .00006	0.0103	
	5	9	.0295 ± .00005	.0335	
	5	8	.0756 ± .00004	.0931	
	6	11	.0094	.0101	0.0104
	6	10	.0258	.0292	.0252
	6	9	.0616	.0746	.0560
	7	12	.0085 ± .00008	.0091	.0092
	7	11	.0217 ± .00006	.0243	.0212
	7	10	.0493 ± .00005	.0587	.0454
	8	13	.0075 ± .00009	.0079	.0078
	8	12	.0178 ± .00008	.0198	.0174
	8	11	.0389 ± .00006	.0456	.0363
5	1	(No possible range with significance level less than 0.1.)			
	2	8*	.0012	.0012	
	2	7	.0154	.0159	
	2	6	.0811	.0989	
	3	9	.00935 ± .00009	.0100	
	3	8	.0368 ± .00007	.0436	
	4	11	.0045 ± .00013	.0048	
	4	10	.0164 ± .00011	.0182	
	4	9	.0474 ± .00009	.0575	
	5	12	.0071 ± .00015	.0075	.0079
	5	11	.0199 ± .00013	.0226	.0200
	5	10	.0494 ± .00011	.0607	.0460
6	1	5*	.0586	.0586	
	2	9		.0019	
	2	8		.0159	
	3	11		.0031	
	3	10		.0140	
	3	9		.0514	
	4	12		.0087	.0091
	4	11		.0273	.0235
	4	10		.0763	.0552

TABLE 3.2 (continued)

t	n	R	$Pr(a_{(1)}-a_{(t)} \geq R \mid H_o')$**	U(R)	$Pr(W_t \geq \sqrt{\frac{4}{nt}}(R-\frac{1}{4}))$
7	1	6*	0.0205	0.0205	0.0345
	2	9		.0139	
	2	8		.0663	
	3	12		.0042	
	3	11		.0163	
	3	10		.0541	
	4	13		.0123	.0116
	4	12		.0349	.0282
	4	11			.0619
8	1	7*	.0068	.0068	.0169
	1	6	.0738	.0889	.0778
	2	10		.0113	
	2	9		.0502	
	3	13		.0049	.0057
	3	12		.0172	.0159
	3	11		.0529	.0403
9	1	8*		.0022	
	1	7		.0330	
	2	11		.0087	
	2	10		.0367	
	3	14		.0052	.0057
	3	13		.0169	.0153
	3	12		.0493	.0373
10	1	8		.0117	.0190
	1	7		.0947	
	2	12		.0065	
	2	11		.0262	
	2	10		.0910	.0635
	3	15		.0051	.0054
	3	14		.0156	.0141
	3	13		.0442	.0336

TABLE 3.2 (continued)

* These values of R are such that

$$R > n(t-1) - \tfrac{1}{2}n,$$

and therefore $U(R) = \Pr(a_{(1)} - a_{(t)} \geq R | H_o')$

** The values of $\Pr(a_{(1)}-a_{(t)} \geq R | H_o')$ were calculated from the tables of Bradley and Terry (1952), Bradley (1954a), and David (1959). Because of the construction of the Bradley-Terry and Bradley tables, it was sometimes necessary to subtract probabilities of smaller ranges from the cumulative probability entered in these tables. Each subtraction could cause two possible errors of as much as ± 0.00005. Hence, many of the entries in the "$\Pr(a_{(1)}-a_{(t)} \geq R | H_o')$" column are of the form $(\beta \pm \sigma)$, where σ is the standard deviation of the sum of 2k (k being the number of subtractions) independent variates with rectangular distributions over the range (-0.00005, 0.00005). Hence

$$\sigma = 0.00005\sqrt{\frac{k}{3}} .$$

3.6 A Method for Judging Contrasts of Treatment Scores

If in the analysis of variance, the F-test rejects the null hypothesis of equal means, then it also rejects the null hypothesis that the values of all contrasts are zero. A method for making further inferences concerning the contrasts when H_o is rejected by the F-test has been developed by Scheffé (1953). The Scheffé method of determining the significance of a contrast applies to those contrasts chosen after studying the data, as well as to those contrasts of interest prior to the experiment. The method proceeds as follows after the null hypothesis has been rejected at the α-level by the F-test:

1. Calculate the contrast $C = \sum_{i=1}^{m} L_i m_i$, where $\sum_{i=1}^{m} L_i = 0$, and m_i is the sample mean of the i th sample.
2. Calculate the variance of C,

$$s_c^2 = \sum_{i,j=1}^{t} a_{ij} L_i L_j s_m^2 , \tag{3.6.1}$$

where s_m^2 is the analysis of variance estimate of the variance of the mean with r degrees of freedom, and $\mathrm{cov}(m_i, m_j) = a_{ij}\,\sigma_m^2$.

3. Calculate

$$T = (m-1)\, F_{m-1,r;\alpha} , \tag{3.6.2}$$

where $F_{m-1,r;\alpha}$ is the $100(1-\alpha)$ percentile of the F-distribution with (m-1) and r degrees of freedom.

4. If $|C| \geq \sqrt{(Ts_c^2)}$, the contrast is declared significantly different from zero.

An argument similar to that employed by Scheffé (1953) will be used to develop an analogous method for judging contrasts of the treatment scores after finding the observed value of D in the paired-comparison experiment to be significant.

In Section 2.2, we found that under the null hypothesis, $H_o' : \pi_{ij} = \frac{1}{2}$ $(i,j = 1,\ldots,t)$, the d_i have variances

$$(3.6.3) \qquad \sigma_i^2 = \frac{t-1}{t}$$

and correlations

$$(3.6.4) \qquad \rho_{ij} = \frac{-1}{t-1} .$$

Consider the set of orthogonal contrasts

$$(3.6.5) \qquad Q_k = \sum_{i=1}^{t} L_{i(k)} d_i \qquad (k = 1,\ldots,t-1) ,$$

where

$$(3.6.6) \qquad \sum_{i=1}^{t} L_{i(k)} = 0 .$$

The variance of Q_k is

$$(3.6.7) \qquad S_k = \sum_{i=1}^{t} \sum_{j=1}^{t} \rho_{ij} L_{i(k)} L_{j(k)} \sigma_i^2 = \sum_{i=1}^{t} L_{i(k)}^2 .$$

From the theory of orthogonal contrasts, we have

$$(3.6.8) \qquad D = \sum_{i=1}^{t} d_i^2 = \sum_{k=1}^{t-1} Q_k^2 / S_k ,$$

for every set (Q_k) of (t-1) orthogonal contrasts. Since the terms in the sum on the R.H.S. of (3.6.8) are all non-negative,

we have

$$(3.6.9) \qquad D \geq Q_k^2/S_k \qquad (k = 1,\ldots,t-1) ,$$

with the equality occurring when $L_{i(k)} = d_i \quad (i = 1,\ldots,t)$. Hence, if D_α is the upper $100\alpha\%$ point of the D distribution under H_o',

$$\Pr(D < D_\alpha) = \Pr \text{ (For every possible contrast, } Q_k^2/S_k < D_\alpha) = 1 - \alpha .$$

Hence, after performing the D-test and rejecting H_o at the α significance level, the experimenter may judge all contrasts of treatment scores in which he is interested with the following procedure:

1. Calculate

$$Q^2 = (\sum_{i=1}^{t} L_i d_i)^2 = (\sum_{i=1}^{t} L_i \sqrt{\tfrac{4}{nt}}\, a_i)^2 ,$$

where $\sum_{i=1}^{t} L_i = 0$.

2. Calculate

$$S = \sum_{i=1}^{t} L_i^2 .$$

3. Calculate SD_α, where D_α is the critical value used in the preceding D-test.
4. If $Q^2 \geq S \cdot D_\alpha$, declare the contrast significantly different from zero.

For small experiments, α and D_α are determined exactly from the tabled distributions, and, for larger experiments, the approximation $D_\alpha = \chi^2_{(t-1);\alpha}$ is used.

IV. FACTORIALS WITH PAIRED COMPARISONS

4.1 Introduction

If the treatments in a paired-comparison experiment represent factorial combinations, it is desirable to test the various factorial effects separately rather than to test $H_o' : \pi_{ij} = \frac{1}{2}$, for all (i,j), against its general alternative.

The idea of paired-comparison experiments involving factorials is first discussed by Abelson and Bradley (1954). They employ the Bradley-Terry model and a maximum likelihood approach to obtain estimators of the "true ratings" of the levels of each factor. Also, through the use of statistics of the form $-2\ln\lambda$, where λ is a likelihood ratio function, they are able to analyze a 2 x 2 factorial. Their method allows one to test the main effect of one factor without regard to the other effects, the main effect of the second factor assuming the main effect of the first factor to be zero, and the interaction effect of the two factors. They found that their method for testing and estimation becomes quite difficult for factorials larger than the 2 x 2.

Bliss, Greenwood, and White (1956) have suggested an analysis of variance technique using rankits to test the factorial effects in a 2 x 2 factorial. Their analysis of variance gives tests of the main effects, the interaction effect, and the order and replicate effects. The possibility of correlations between test statistics is not discussed. It appears that their method may be extended to larger factorials. The rankit analysis may be based on either the Thurstone-Mosteller model or the Scheffé model.

Dykstra (1958) has applied Scheffé's (1952) analysis of variance method to experiments involving factorials.

Through the use of contrasts over the estimators $\hat{\alpha}_i$ of the treatment location parameters, he is able to test all the factorial effects. For 2^p factorials, he proposes fractional replication and a blocking method to reduce the required number of paired comparisons. The method, of course, depends upon the applicability of the Scheffé model.

The method that is developed in the following section is based on the model presented in Section 2.1.

4.2 Tests of Factorial Effects

The u x v x...x z factorials that will be considered are those in which only quantitative factors appear at more than two levels, and the levels form an arithemetic progression in the scales of measurement of their respective factors. Approximations similar to those in Chapter II will be employed. For this reason, the size of the experiment should be at least moderately large, and, on the basis of Table 2.1, we suggest

$$n(t-1) \geq 20 \quad . \tag{4.2.1}$$

To facilitate the discussion of paired-comparison experiments in which the treatments are factorial combinations, the following new symbols will be employed.

(1) $a_{ij\ldots m}$ is the score of the treatment representing the i th level of factor U, the j th level of factor V, etc.

(2) $d_{ij\ldots m} = \sqrt{(4/nt)}(a_{ij\ldots m} - \bar{a})$, where, as usual, $\bar{a} = \sum a_{ij\ldots m}/t = \frac{1}{2}n(t-1)$.

(3) $Q_{U,q} = \sum_{i,j,\ldots,m}^{u,\ldots,z} L_{i,q}\, d_{ij\ldots m}$, where $\sum_{i=1}^{u} L_{i,q} = 0$.

$$(4)\quad Q_{V,r} = \sum_{i,j,\ldots,m}^{u,v,\ldots,z} M_{j,r} d_{ij\ldots m} \text{ , where } \sum_{j=1}^{v} M_{j,r} = 0 \text{ .}$$

$$(5)\quad Q_{UV\ldots Y,s} = \sum_{i,j,\ldots,m}^{u,v,\ldots,z} N_{ij\ldots k,s} d_{ij\ldots m} \text{ , where}$$

$$k \leq m, \text{ and } \sum_{i=1}^{u} N_{ij\ldots k} = \sum_{j=1}^{v} N_{ij\ldots k} = \ldots = \sum_{k=1}^{y} N_{ij\ldots k}$$

$$= 0 \text{ .}$$

$$(6)\quad S_{UV\ldots Y,s} = \sum_{i,j,\ldots m}^{u,v,\ldots,z} N^2_{ij\ldots k,s} \text{ .}$$

The null hypothesis

$$H_o' : \pi_{ij} = \tfrac{1}{2} \qquad (i,j = 1,\ldots,t; \quad i \neq j)$$

is now equivalent to

H_o : All factorial effects, main and interaction, are zero.

Hence, we may apply some of the results of Section 3.6 to the contrasts over the $d_{ij\ldots m}$'s. When H_o is true, any such contrast,

$$(4.2.2)\qquad Q = \sum_{i,j,\ldots,m}^{u,v,\ldots,z} K_{ij\ldots m} d_{ij\ldots m} \qquad \text{(where}$$

$$\sum_{i,j,\ldots,m} K_{ij\ldots m} = 0) \text{ ,}$$

has mean zero and variance

$$(4.2.3)\qquad S = \sum_{i,j\ldots,m} K^2_{ij\ldots m} \text{ .}$$

Due to the asymptotic normality of the $d_{ij\ldots m}$'s, the statistic Q^2/S is distributed approximately as a χ^2-variate with one degree of freedom.

To avoid excess notation, we will consider a two-factor $\underline{u}$ x $\underline{v}$ factorial in discussing further properties of the contrasts. If $Q_j = \sum_{i,j}^{u,v} J_{ij} d_{ij}$ and $Q_k = \sum_{i,j}^{u,v} K_{ij} d_{ij}$ are

orthogonal contrasts (i.e., $\sum_{i,j} J_{ij} = \sum_{i,j} K_{ij} = \sum_{i,j} J_{ij} K_{ij} = 0$), then under H_o we have

$$(4.2.4) \quad \mathrm{Cov}(Q_J, Q_K) = \sum_{i,j}^{u,v} J_{ij} K_{ij} \, \mathrm{Var}(d_{ij}) + \sum_{\substack{i,j \\ (i,j)\neq(g,h)}}^{u,v} \sum_{g,h}^{u,v} J_{ij} K_{gh} \, \mathrm{Cov}(d_{ij}, d_{gh})$$

$$= \left(\frac{t-1}{t}\right) \sum_{i,j}^{u,v} J_{ij} K_{ij} - \frac{1}{t} \left[\sum_{i,j}^{u,v} J_{ij} K_{ij} \right] = 0 .$$

From this result, we see that under the null hypothesis any two orthogonal contrasts are uncorrelated. Since the orthogonal contrasts are asymptotically normal, they are also asymptotically independent.

For the two-factor $\underline{u}$ x $\underline{v}$ factorial, we have, of course,

$$(4.2.5) \qquad t = uv,$$

and

$$(4.2.6) \qquad (u-1) + (v-1) + (u-1)(v-1) = (t-1) .$$

From the theory of orthogonal contrasts, we know that it is possible to find (u-1) contrasts $Q_{U,q}$ (definition 3), (v-1) contrasts $Q_{V,v}$ (definition 4), and (u-1)(v-1) contrasts

$Q_{UV,s}$ (definition 5) such that all (t-1) of the contrasts are mutually orthogonal. Let us consider the properties of these contrasts under the alternative hypothesis

H_{aU} : the main effect of the factor U is not zero, while all other effects are zero.

The statistic a_{ij} is now a generalized binomial variate representing the outcome of n(t-1) comparisons with probability of success p_{ik}, where p_{ik} is the probability that any treatment with factor U at the $\underline{i}$ th level will be preferred to any treatment with U at the $\underline{k}$ th level. In comparing two distinct treatments both with factor U at the $\underline{i}$ th level, we have $p_{ii} = \frac{1}{2}$. Hence,

$$(4.2.7) \qquad \delta_i = E(d_{ij}) = \sqrt{\frac{4}{nt}} \left[\frac{t}{2}(v-1) + v\sum_{k\neq i}^{u} p_{ik} - \tfrac{1}{2}(t-1)\right],$$

$$(4.2.8) \qquad \sigma_i^2 = Var(d_{ij}) = (v-1)/t + 4v/t \sum_{k\neq i}^{u} p_{ik}\, q_{ik}$$

$$(q_{ik} = 1-p_{ik}),$$

and

$$(4.2.9) \qquad \sigma_{ig} = Cov(d_{ij},d_{gk}) = \begin{cases} -1/t & \text{if } i=g,\ j\neq h \\ -4p_{ig}q_{ig}/t & \text{if } i\neq g \end{cases}.$$

The fact that the R.H.S.'s of the above equations are independent of the level of the factor V will be used repeatedly in the following discussion of the variances and covariances of the contrasts.

We have found that under H_o the orthogonal contrasts Q_k are uncorrelated and have means zero and variances $\sum_{ij} K_{ij}^2$. We shall now consider the properties of the

orthogonal contrasts corresponding to effects other than the main effect of U under H_{aU}.

$$(4.2.10)\quad E(Q_{V,r}) = \sum_{i=1}^{u}\sum_{j=1}^{v} M_{j,r}\,\delta_i = \sum_{i=1}^{u}\delta_i \sum_{j=1}^{v} M_{j,r} = 0 .$$

$$(4.2.11)\quad E(Q_{UV,s}) = \sum_{i=1}^{u}\delta_i \sum_{j=1}^{v} N_{ij,s} = 0 .$$

$$(4.2.12)\quad \mathrm{Var}(Q_{V,r}) = \sum_{i=1}^{u}\sum_{j=1}^{v} M_{j,r}^2\,\sigma_i^2 + \sum_{i=1}^{u}\sum_{j=1}^{v}\sum_{h\neq j} M_{j,r}M_{h,r}\,\sigma_{ii} + \sum_{i=1}^{u}\sum_{j=1}^{v}\sum_{g\neq i}\sum_{h=1}^{v} M_{j,r}\,M_{h,r}\,\sigma_{ij}$$

$$= \sum_{i=1}^{u}(1/t + \sigma_i^2)\sum_{j=1}^{v} M_{j,r}^2 .$$

Similarly,

$$(4.2.13)\quad \mathrm{Var}(Q_{UV,s}) = \sum_{i=1}^{u}(1/t + \sigma_i^2)\sum_{j=1}^{v} N_{ij,s}^2 .$$

Equation (4.2.8) shows that $\sigma_i^2 < (t-1)/t$ under H_{aU}, and, therefore, the variances of $Q_{V,r}$ and $Q_{UV,s}$ are less than they were under the null hypothesis. Hence, $Q_{V,r}^2/S_{V,r}$ is distributed approximately as $(\mathrm{Var}\ Q_{V,r})/S_{V,r}$ times a χ^2 variate with one degree of freedom. An analogous statement is true for $Q_{UV,s}^2/S_{UV,s}$.

For testing purposes, we will be interested in whether the various contrasts are correlated with each other under H_{aU}.

$$(4.2.14)\quad \mathrm{Cov}(Q_{U,q},Q_{V,r}) = \sum_{i=1}^{u}\sum_{j=1}^{v} L_{i,q}M_{j,r}\,\sigma_i^2$$

$$+ \sum_{i=1}^{u}\sum_{j=1}^{v}\sum_{g=1}^{u}\sum_{h=1}^{v} L_{i,q}M_{h,r}\,\sigma_{ig}$$

$$= \sum_{i=1}^{u} L_{i,q}\,\sigma_i^2 \sum_{j=1}^{v} M_{j,r} + 1/t\sum_{i}^{u}\sum_{j}^{v} L_{i,q}M_{j,r}$$

$$+ \sum_{i=1}^{u}\sum_{j=1}^{v}\sum_{g\neq i}^{u} L_{i,q}\,\sigma_{ig}\sum_{h=1}^{v} M_{h,r} \;=\; 0\,.$$

$$(4.2.15)\quad \mathrm{Cov}(Q_{U,q},Q_{UV,s}) = \sum_{i=1}^{u}\sum_{j=1}^{v} L_{i,q}N_{ij,s}\,\sigma_i^2$$

$$+ \sum_{\substack{i=1\\(i,j)\neq(g,h)}}^{u}\sum_{j=1}^{v}\sum_{g=1}^{u}\sum_{h=1}^{v} L_{i,q}N_{gh,s}\,\sigma_{ig}$$

$$= \sum_{i=1}^{u} L_{i,q}\,\sigma_i^2 \sum_{j=1}^{v} N_{ij,s} + 1/t\sum_{i=1}^{u}\sum_{j=1}^{v} L_{i,q}N_{ij,s}$$

$$+ \sum_{i=1}^{u}\sum_{j=1}^{v}\sum_{g\neq i}^{u} L_{i,q}\,\sigma_{ig}\sum_{h=1}^{v} N_{gh,s} \;=\; 0\,.$$

Similarly,

$$(4.2.16)\quad \mathrm{Cov}(Q_{V,r},Q_{UV,s}) = \sum_{i=1}^{u}\sum_{j=1}^{v} M_{j,r}N_{ij,s}\,\sigma_i^2$$

$$+ \sum_{\substack{i=1\\(i,j)\neq(g,h)}}^{u}\sum_{j=1}^{v}\sum_{g=1}^{u}\sum_{h=1}^{v} M_{j,r}N_{gh,s}\,\sigma_{ig} \;=\; 0\,.$$

$$(4.2.17)\quad \mathrm{Cov}(Q_{V,r}Q_{V,s}) = \sum_{i=1}^{u}\sum_{j=1}^{v} M_{j,r}M_{j,s}\,\sigma_i^2$$

$$+ \sum_{\substack{i=1\\(i,j)\neq(g,h)}}^{u}\sum_{j=1}^{v}\sum_{g=1}^{u}\sum_{h=1}^{v} M_{j,r}M_{h,s}\,\sigma_{ig}$$

$$= \sum_{i=1}^{u}\sigma_i^2\sum_{j=1}^{v} M_{j,r}M_{j,s} + 1/t\sum_{i=1}^{u}\sum_{j=1}^{v} M_{j,r}M_{j,s}$$

$$+ \sum_{i=1}^{u}\sum_{g\neq 1}^{u}\sigma_{ig}\sum_{j=1}^{v} M_{j,r}\sum_{h=1}^{v} M_{h,s} \;=\; 0\,.$$

$$(4.2.18)\quad \mathrm{Cov}(Q_{UV,r},Q_{UV,s}) = \sum_{i=1}^{u}\sum_{j=1}^{v} N_{ij,r}N_{ij,s}\,\sigma_i^2$$

$$+ \sum_{i=1}^{u}\sum_{g\neq i}^{u}\sigma_{ig}\sum_{j=1}^{v} N_{ij,r}\sum_{h=1}^{v} N_{gh,s}$$

$$+ 1/t\sum_{i=1}^{u}\sum_{j=1}^{v} N_{ij,r}N_{ij,s}$$

$$= \sum_{i=1}^{u}\sigma_i^2\sum_{j=1}^{v} N_{ij,r}N_{ij,s}\,.$$

Now

$$(4.2.19)\quad N_{ij,r} = L_{i,q}\cdot M_{j,t}\,,$$

and

$$(4.2.20)\quad N_{ij,s} = L_{i,p}\cdot M_{j,m} \qquad [(q,t)\neq(p,m)]\,.$$

If in equations (4.2.19) and (4.2.20), m = t, then

$$(4.2.21)\quad \sum_{j=1}^{v} N_{ij,r}N_{ij,s} \neq 0 \qquad \text{for some } i\,.$$

If in equations (4.2.19) and (4.2.20), $m \neq t$, then

$$(4.2.22) \quad \sum_{j=1}^{v} N_{ij,r}N_{ij,s} = 0 \qquad \text{for every } i \text{ .}$$

Equations (4.2.14) and (4.2.15) demonstrate that the contrasts corresponding to the non-zero main effect of U are uncorrelated with those corresponding to the V and UV effects. Equations (4.2.16) and (4.2.17) show that any contrast for the main effect of V is uncorrelated with the contrasts for the interaction UV and with other orthogonal contrasts for the main effect of V. We find, however, that some of the orthogonal contrasts pertaining to the interaction effect UV are correlated if one of the factors in the interaction has a non-zero main effect and occurs at more than two levels. When this correlation exists, it is a function of the unknown p_{ik}'s. Because of this elusive property of the correlation, the distribution of

$$\sum\nolimits_{UV}^{2} = \sum_{s=1}^{(u-1)(v-1)} Q_{UV,s}^{2}/S_{UV,s}$$

has not yet been determined under H_{aU} when $u > 2$.

It is suspected that a comparison of $\sum_{UV}^{2}$ with the critical value of the χ^2 variate with $(u-1)(v-1)$ degrees of freedom may be a conservative test when correlation occurs since $E(Q_{UV,s}^{2}/S_{UV,s}) < 1$. However, lack of more exact information concerning the distribution of $\sum_{UV}^{2}$ forces us to restrict our attention to factorial experiments in which the only factors that may appear at more than two levels are those that are known not to interact with the other factors.

If the interaction (UV) effect is non-zero, equations (4.2.7), (4.2.8), and (4.2.9) no longer apply and $\mathrm{Var}(d_{ij})$ depends on both i and j, while $\mathrm{Cov}(d_{ij},d_{gh})$ is contingent on all four subscripts. Therefore, the correlation of any

two orthogonal contrasts will be a function, not identically zero, of the contrast coefficients and the variances and covariances of the d_{ij}'s. For larger factorials, if one observes a significant k-order interaction, all orthogonal contrasts for lower order interactions and main effects involving only factors in the k-order interaction will be correlated. Hence, in testing, we should start with the highest order interaction and work down to the main effects. If an interaction is found significant, the lower order interactions and main effects for the particular factors in the interaction would not be tested. This is not a serious drawback since the significant interaction requires a separate study of the effect of each factor for each combination of the levels of the other factors.

The preceding discussion of the variances and correlations of the orthogonal contrasts under various conditions has indicated the type of factorial that we may test and the order of testing. The complete testing procedure will now be presented for a $\underline{u}$ x $\underline{v}$ x...x $\underline{z}$ factorial in which no factor is at more than two levels unless the experimenter is reasonably sure that it will not interact with the other factors.

(1) Compare $Q^2_{UV\ldots Y}/S_{UV\ldots Y}$, where $Q_{UV\ldots Y}$ is the contrast corresponding to the highest order interaction considered possible by the experimenter, with the $100\alpha\%$ point $\chi^2_{1,\alpha}$ of a χ^2 variate with one degree of freedom. If the null hypothesis of no effects is accepted (i.e., $Q^2_{U\ldots Y}/S_{U\ldots Y} < \chi^2_{1,\alpha}$) proceed to step (2). If the alternative hypothesis that there is a non-zero (UV...Y) interaction effect is accepted (i.e., $Q^2_{UV\ldots Y}/S_{UV\ldots Y} \geq \chi^2_{1,\alpha}$), proceed to step (3).

(2) If the contrast corresponding to the highest order interaction (U...Y) has been found non-significant, use the same test procedure to test the contrasts of the next lower order interactions. If a contrast is found non-significant continue testing through the lower order interactions. If a contrast is significant, make no further tests on the lower order interactions and main effects of the factors involved.

(3) For each factor W, not involved in a non-zero interaction effect, calculate the sum of squares

$$\sum_W^2 = \sum_{r=1}^{(w-1)} Q_{W,r}^2 / S_{W,r} \text{ and compare with } \chi^2_{w-1,\alpha}.$$

(4) Perform the D-test of Section 2.3.

The contrasts considered in the above procedure are, of course, mutually orthogonal. Also, as pointed out earlier, the tests are conservative when a non-zero effect is present. A significant D, when all other tests give non-significance, indicates that the assumption of zero interaction effects for interactions involving factors appearing at more than two levels is false. The above test procedure is illustrated in Table 4.1 for a 2^3 x 4 factorial in which the factors U, V, and W are at two levels each and may interact with each other, while the factor X, at four levels, is assumed not to interact with the other factors.

TABLE 4.1

Test Procedure for a 2^3 x 4 Factorial*

Test	Hypotheses	Statistic	d.f.	Critical Value	When to Test
U x V x W interaction	H_o : No non-zero factorial effects. H_{a1}^o : A non-zero U x V x W interaction effect is present.	Q^2_{UVW}/S_{UVW}	1	$\chi^2_{1,\alpha}$	Always.
U x V interaction	H_o : No non-zero factorial effects. H_{a2}^o : A non-zero U x V interaction effect is present.	Q^2_{UV}/S_{UV}	1	$\chi^2_{1,\alpha}$	If H_{a1} is not accepted.
U x W interaction	H_o : No non-zero factorial effects. H_{a3}^o : A non-zero U x W interaction effect is present.	Q^2_{UW}/S_{UW}	1	$\chi^2_{1,\alpha}$	If H_{a1} is not accepted.
V x W interaction	H_o : No non-zero factorial effects. H_{a4}^o : A non-zero V x W interaction effect is present.	Q^2_{WV}/S_{VW}	1	$\chi^2_{1,\alpha}$	If H_{a1} is not accepted.
U factor	H_o : No non-zero factorial effects. H_{a5}^o : A non-zero U main effect is present.	Q^2_U/S_U	1	$\chi^2_{1,\alpha}$	If H_{a1}, H_{a2}, and H_{a3} are not accepted.
V factor	H_o : No non-zero factorial effects. H_{a6}^o : A non-zero V main effect is present.	Q^2_V/S_V	1	$\chi^2_{1,\alpha}$	If H_{a1}, H_{a2}, and H_{a4} are not accepted.
W factor	H_o : No non-zero factorial effects. H_{a7}^o : A non-zero W main effect is present.	Q^2_W/S_W	1	$\chi^2_{1,\alpha}$	If H_{a1}, H_{a3}, and H_{a4} are not accepted.

TABLE 4.1 (continued)

Test	Hypotheses	Statistic	d.f.	Critical Value	When to Test
X factor	H_o : No non-zero factorial effects. H_{a8}: A non-zero X main effect is present.	$\sum_{r=1}^{3} Q^2_{X,r}/S_{X,r}$	3	$\chi^2_{3,\alpha}$	Always
Treatment effects	H_o : $\pi_{ij} = \frac{1}{2}$ $(i,j=1,\ldots,24;\ i\neq j)$ H_a : $\pi_{ij} \neq \frac{1}{2}$ for some (i,j)	D	31	$\chi^2_{31,\alpha}$	Always

(* On prior knowledge, the experimenter is able to assume that factor X, the factor at 4 levels, will not interact with U, V, or W.)

For the case of a 2^p factorial, the binomial test may be substituted for the χ^2-test. This is accomplished by summing the treatments at one of the two levels of the effect to be tested and then treating the sum as a binomial variate with parameters $\left[\frac{r}{2}, n\binom{t}{2}\right]$. The same sequence of testing is followed as is stated in steps (1) - (3).

An example using the testing methods presented in this section is given in Section 5.12.

4.3 Comparison with Other Methods

In comparison with the methods mentioned in Section 4.1, the procedure of Section 4.2 may be considered as a "quick and dirty" analysis. The computation is simpler and the model is more general for our method, but the results are not, in general, as conclusive or extensive as obtained by the other methods when the data satisfy their models.

V. NUMERICAL EXAMPLES

5.1 Introduction

A recent paper by Fleckenstein, Freund, and Jackson (1958) describes a paired-comparison experiment performed to determine the relative quality of various brands of carbon paper. We shall analyze the data from their experiment using the methods developed in Chapters II and III.

A hypothetical example of a paired-comparison experiment involving factorials is analyzed by the method presented in Chapter IV.

5.2 Carbon Paper Problem

Five brands of carbon paper are used by various departments of a company. For the purpose of reducing inventory costs, the company wishes to standardize on the one or two brands of highest quality. A side issue is that two brands (1 and 2) are considered expensive, one (3) moderately priced, and the remaining two (4 and 5) relatively inexpensive.

5.3 Procedure Used by Fleckenstein, Freund, and Jackson

Five secretaries with varied clerical backgrounds were selected from each of six departments of the company. Fifteen arrangements of the ten possible pairs of brands were assigned to fifteen of these secretaries. The order of comparison in each pair was fixed. The remaining fifteen secretaries received the same fifteen random arrangements, but the pairs were in reverse order to those in the first group.

The secretaries were requested to use a seven point scoring system in listing their preference. This scoring

system is described under the Scheffé method in Section 1.2. The results are listed in Table 5.1.

With the information in Table 5.1, Fleckenstein, Freund, and Jackson applied Scheffé's analysis of variance. They found the sample brand means on the response scale to be as follows:

$$\hat{\alpha}_1 = 0.133$$
$$\hat{\alpha}_2 = -0.180$$
$$\hat{\alpha}_3 = 0.873$$
$$\hat{\alpha}_4 = -1.080$$
$$\hat{\alpha}_5 = 0.253$$

Using Tukey's test based on allowances at the 5% significance level and with an h.s.d. of 0.562, they obtained the ordering

$$\underline{\hat{\alpha}_4 \quad \hat{\alpha}_2 \quad \hat{\alpha}_1 \quad \hat{\alpha}_5} \qquad \underline{\hat{\alpha}_3},$$

where any two means not underlined by the same line are significantly different.

The use of Scheffé's method required the assumption of normality of treatment responses and, also, the ability to scale the preferences. When we apply the methods of Chapters II and III, the assumptions may be weakened to those of Section 2.1.

5.4 Test of Uniformity of Brand Quality

In order to apply the tests described in Chapters II and III, we must disregard the scaling system and simply count the number of times each brand is preferred to each of the other brands. Also, judgments falling in the zero

TABLE 5.1

Scores in the Pairwise Comparison of 5 Carbon Papers

Pair i,j	Groups*	Frequency of Preference						
		-3	-2	-1	0	1	2	3
1,2	I	3	2	0	4	3	2	1
	II	1	2	0	1	2	6	3
1,3	I	4	7	0	3	0	0	1
	II	1	5	4	3	0	2	0
1,4	I	0	0	0	2	3	7	3
	II	0	2	2	0	2	8	1
1,5	I	3	1	3	3	1	2	2
	II	1	3	2	1	1	6	1
2,3	I	2	5	1	3	1	2	1
	II	4	4	2	1	0	2	2
2,4	I	0	2	0	4	2	5	2
	II	2	2	1	2	2	3	3
2,5	I	1	3	3	4	0	3	1
	II	4	2	3	2	1	1	2
3,4	I	0	0	0	2	1	5	7
	II	0	1	0	1	0	8	5
3,5	I	1	2	3	3	3	1	2
	II	0	4	0	1	0	10	0
4,5	I	6	3	3	3	0	0	0
	II	0	9	2	0	1	3	0
Totals		33	59	29	43	23	76	37

(* Group I: compared in order i,j.
Group II: compared in order j,i.)

category for each pair must be assigned in some way to the elements of the pair. Miss Fleckenstein made random assignments of the zero judgments in order to apply the Bradley-Terry and the Thurstone-Mosteller techniques (unpublished). The scores she obtained are set out in Table 5.2 with the results for each of the six departments shown separately.

The objective of the experiment determines the type of test that should be performed on the resulting data. The objective of the above paired-comparison experiment is to find the one or two brands of highest quality. Of the test procedures proposed in this thesis, the appropriate method is the one given in Section 3.5 and, in Section 5.10, that method is used on the data. However, to illustrate the other test methods of Chapters II and III, new objectives are suggested and the corresponding tests are performed. In this section, it is assumed that the goal of the experiment is to determine whether or not the five brands of carbon paper are of equal quality.

Since the brands of carbon paper being tested were ones used by the six departments, it would seem possible that secretaries might recognize the characteristics of the brand they normally use and be prejudiced for or against it. Hence, one might suspect before performing the experiment that the true preference probabilities may differ between departments if there is an actual difference in the brands. For this reason, one should use a combined analysis to test the null hypothesis

$$H_0 : \pi_{i\cdot} = \tfrac{1}{2} \qquad (i = 1,\ldots 5)$$

against its general alternative. Because of the way in which the comparisons were performed, there is little reason to

TABLE 5.2

Data of Table 5.1 Set Out as Preference Scores,
With Random Allotments of No-Preference Judgments

	Departments (γ)						
	I	II	III	IV	V	VI	Total
Comparisons i : j	Scores						$x_{ij}:x_{ji}$
1 : 2	5:0	3:2	4:1	2:3	4:1	2:3	20:10
1 : 3	1:4	2:3	0:5	0:5	2:3	1:4	6:24
1 : 4	5:0	3:2	3:2	4:1	5:0	5:0	25:5
1 : 5	3:2	2:3	3:2	2:3	1:4	4:1	15:15
2 : 3	2:3	2:3	0:5	2:3	2:3	2:3	10:20
2 : 4	3:2	4:1	2:3	4:1	3:2	4:1	20:10
2 : 5	3:2	1:4	1:4	3:2	1:4	2:3	11:19
3 : 4	5:0	3:2	5:0	4:1	5:0	5:0	27:3
3 : 5	4:1	3:2	3:2	2:3	3:2	3:2	18:12
4 : 5	1:4	3:2	0:5	1:4	1:4	0:5	6:24
$a_{1\gamma}$	14	10	10	8	12	12	a_1= 66
$a_{2\gamma}$	8	9	4	12	7	11	a_2= 51
$a_{3\gamma}$	16	12	18	14	14	15	a_3= 89
$a_{4\gamma}$	3	8	5	4	3	1	a_4= 24
$a_{5\gamma}$	9	11	13	12	14	11	a_5= 70
Total	50	50	50	50	50	50	a_i=300

expect an ordering effect. (This reasoning appears valid as the Scheffé analysis by Fleckenstein, Freund, and Jackson (1958) did not show a significant ordering effect.)

The combined analysis method for testing H_o (cf. Section 2.6) proceeds as follows:

1. Significance level for test: 5%.
2. Critical value of $\chi^2_{\pi(t-1)}$: $\chi^2_{24,0.05} = 36.42$.
3. $$D_c = \sum_{\gamma=1}^{6} D_\gamma = \sum_{\gamma=1}^{6} 4\left[\sum_{i=1}^{5} a_i^2 - \tfrac{1}{4}\cdot 5^2 \cdot 5 \cdot 4^2\right]/(5.5)$$
   $$= 83.2.$$
4. $D_c > \chi^2_{24,0.05}$.
5. Conclusion: The five brands of carbon paper are not all of equal quality.

(The exact Bradley-Terry combined analysis can be performed on these data. It also gives a significant result.)

5.5 Test of Interaction between Groups (Departments) and Brands

In Section 5.4, we mentioned that there is reason to suspect changes in true preference probabilities between departments. A method for testing this suspicion is given in Section 2.8. The test of

$$H_o : \pi_{ij\gamma} = \pi_{ij} \qquad (i \neq j;\ i,j=1,\ldots,5;\ \gamma=1,\ldots,6),$$

against

$$H_a : \pi_{ij\gamma},$$

is as follows:

1. Significance level of test : 5%.

2. Critical values: $\chi^2_{50;0.05} = 79.49$; $Z_{0.05} = 1.64$.

3. $C_{12} = 30^2(25/9 + 1/9 + 4/9 + 16/9 + 4/9 + 16/9)$
$/(5 \times 200) = 6.60$

$C_{13} = 900(2)/(5 \times 144) = 2.50$

$C_{14} = 900(174/36)/(5 \times 125) = 6.96$

$C_{15} = 900(22/4)/(5 \times 225) = 4.40$

$C_{23} = 900(30/9)/(5 \times 200) = 3.00$

$C_{24} = 900(30/9)/(5 \times 200) = 3.00$

$C_{25} = 900(174/36)/(5 \times 209) = 4.16$

$C_{34} = 900(14/4)/(5 \times 81) = 7.78$

$C_{35} = 900(2)/(5 \times 216) = 1.67$

$C_{45} = 900(6)/(5 \times 144) = 5.00$

$$C_T = \sum_{i=1}^{4} \sum_{j=i+1}^{5} C_{ij} = 45.57$$

Since there are only five repetitions per group, the expected cell frequencies, $x_{ij}/6$, are too small to satisfy Cochran's (1954) recommendations. Hence, the Z test of Section 2.8 will be used instead of the comparison of C_T with $\chi^2_{50;0.05}$.

From equations (2.8.4), (2.8.6), and (2.8.7), we obtain

$$E(C_T) = \sum_{i=1}^{4} \sum_{j=i+1}^{5} E(C_{ij}) = 10 \cdot \frac{5 \cdot 30}{29} = 51.72$$

$$\mathrm{Var}(C_T) = \sum_{i=1}^{4} \sum_{j=i+1}^{5} \mathrm{Var}(C_{ij}) = 83.27$$

$$Z = \frac{C_T - E(C_T)}{\sqrt{\mathrm{Var}\ C_T}} = \frac{45.57 - 51.72}{\sqrt{83.27}} < Z_{0.05} = 1.64$$

4. Our conclusion is that we may accept the null hypothesis of no preference probability changes between repetitions (i.e., no interaction between departments and brands).

It should be pointed out that a test based on C_T could also be used to test the ordering effect.

5.6 Test of Pre-Assigned Treatment

Let us suppose that before the experiment brand 5 received considerable recommendation because of its low cost and reputed quality. In such a case, the experimenter is particularly interested in testing whether brand 5 is better than the average of the 5 brands considered. The test of

$$H_o : \pi_{5.} = \tfrac{1}{2}$$

against

$$H_a : \text{Treatment 5 is better than the average (i.e., } \pi_{5.} > \tfrac{1}{2}),$$

proceeds as follows (cf. Section 3.2):

1. Significance level : 5%.
2. Since $n(t-1) = 120$ is large, we may use the normal approximation to the binomial to find the critical value for the score of brand 5.

$$a_c = 1.64\sqrt{n(t-1)/4} + \tfrac{1}{2}n(t-1) + \tfrac{1}{2} = 1.64\sqrt{30} + 60 + \tfrac{1}{2} = 69.48.$$

3. $a_5 = 70$. (The actual significance level of this score is 0.0412 under H_o'.)
4. $a_5 > a_c$
5. Conclusion: Brand 5 is superior to the average of the 5 brands.

5.7 Test of Equality of Two Pre-Assigned Treatments

Because brand 4 is less expensive than brand 2, we might suppose that there may have been some interest expressed prior to the experiment on whether or not brand 2 is actually superior in quality to brand 4. If this were the case, we would test

$$H_o : \pi_{4.} = \pi_{2.},$$

against

$$H_a : \pi_{2.} \quad \pi_{4.}$$

This is an example in which we would use the one-sided test method described in Section 3.3.

1. Significance level : 5%.
2. $1.64\sqrt{nt/2} + 0.5 = 1.64\sqrt{75} + 0.5 = 14.7$; $m_c = 15$.
3. $a_2 - a_4 = 51 - 24 = 27$.
4. $a_2 - a_4 > m_c$.
5. Conclusion: Brand 2 is better than brand 4.

As mentioned in Section 3.3, with the present hypothesis, the above method gives a conservative test.

5.8 Test of Brand Receiving Highest Score

In the carbon paper experiment under discussion, brand 3 obtained the highest score. It is natural to wonder whether brand 3 is significantly better than the average in quality. To test

H_o : brand 3 is of average quality, i.e., $\pi_{(1).} = \frac{1}{2}$,

against

H_a : brand 3 is better than average, i.e., $\pi_{(1).} > \frac{1}{2}$,

we use the method presented in Section 3.4.

1. Significance level : 5% .
2. From binomial tables [Harvard Univ. (1955)] we find for n = 120 and p = $\frac{1}{2}$, that

   $P(a_i \geq 74) = 0.033/5$,

   and $P(a_i \geq 73) > 0.05/5$.

   Therefore, $m_\beta = 74$ and $\beta = 0.033$.
3. $a_3 = a_{(1)} = 89$
4. $a_{(1)} > m_\beta$
5. Conclusion: We reject the null hypothesis and declare brand 3 significantly better than average.

As discussed in Section 3.4, the present null hypothesis makes the above procedure conservative.

5.9 Test of Brand Receiving Minimum Score

If the experimenters, after seeing the low score of brand 4, wish to find out whether it is significantly worse

than average, they may proceed in a manner similar to that of Section 5.8.

1. Significance level : 5% .
2. From the cumulative binomial probability tables [Harvard Univ. (1955)] we find that $m_\beta = 46$, and $\beta = 0.033$.
3. $a_4 = a_{(5)} = 24$.
4. $a_{(5)} < m_\beta$.
5. Conclusion: Brand 4 is significantly worse than the average.

5.10 Separation of Brands on the Basis of Quality

After obtaining the data in Table 5.2, one would want to separate the brands into significantly different groups, if a difference exists. (This is, of course, the question the company actually wanted answered.) To do this, one may use the method presented in Section 3.5.

1. Significance level : 5% .
2. $W_{5;0.05} = 3.86$ (obtained from Biometrika Tables)

   $R = W_{5;0.05} \cdot \sqrt{nt/4} + \frac{1}{4} = 3.86 \sqrt{150/4} + \frac{1}{4} = 23.887$.

   $R^+ = 24 < n(t-1) - \frac{1}{2}n = 105$.

   $\Pr[\overline{W}_t \geq \sqrt{4/nt}\ (R^+ - \frac{1}{4})] = \Pr[\overline{W}_5 \geq 3.878] = 0.048$.

   Since R^+ is so much less than $[\overline{n}(t-1) - \frac{1}{2}n]$, one need not calculate $U(R^+)$.

   Hence, $R_{\beta(\alpha)} = 24$ and $\beta = 0.048$.
3. a_4 a_2 a_1 a_5 a_3

   24 51 66 70 89

4. Any two brands not underlined by the same line in (3) are significantly different.

The corresponding result obtained by Fleckenstein, Freund, and Jackson (1958) was

$$\underline{a_4 \quad a_2 \quad a_1 \quad a_5} \qquad \underline{a_3}\ .$$

Hence, the Scheffé procedure is slightly more discriminating than our method. This additional discrimination is justified if the data satisfy the Scheffé model.

5.11 Contrasts of Treatment Scores

Since

$$D = 4/150\ [20{,}354 - 18{,}000] = 62.77$$

is greater than $\chi^2_{4,0.05} = 9.488$, it is significant, and we may go on to judge contrasts in the manner described in Section 3.6. There is little reason to be interested in contrasts in this experiment but, for illustrative purposes, we shall consider the following contrasts:

(a) All contrasts of the form $(d_i - d_j)$,
(b) $(d_1 + d_2 - d_4 - d_5)$.

The second contrast (b) gives indication of whether carbon paper quality changes with cost.

For (a) we have (using $\alpha = 0.05$)

(1) $S_k = 1^2 + 1^2 = 2$

(2) $D_\alpha = \chi^2_{4;0.05} = 9.488$, $S_k D_\alpha = 18.976$.

(3) $(d_3-d_2)^2 = 4(89-51)^2/150 = 38.5 > S_k D_\alpha$

$(d_3-d_1)^2 = 4(89-66)^2/150 = 14.1 < S_k D_\alpha$

$(d_5-d_2)^2 = 4(70-51)^2/150 = 9.6 < S_k D_\alpha$

$(d_2-d_4)^2 = 4(51-24)^2/150 = 19.4 > S_k D_\alpha$

Hence,

$(4/150)\cdot(24-60)$ $(4/150)\cdot(51-60)$ $(4/150)\cdot(66-60)$ $(4/150)\cdot(70-60)$ $(4/150)\cdot(89-60)$

$\underline{d_4}$ $\underline{d_2}$ $\underline{d_1}$ $\underline{d_5}$ d_3

(4) Contrasts over any two d_i's not underlined by the same line are declared significant.

In general, the pairwise discrimination between treatments with the contrast method will not be as good as with the method used in Section 5.10, but in this case the results are the same.

For contrast (b),

(1) $Q^2 = (\sum L_i \cdot \sqrt{4/nt}\ a_i)^2 = 4(66 + 51 - 24 - 70)^2/150$
$= 14.11$

(2) $S = 4.$

(3) $SD_\alpha = 4 \times 9.488 = 37.95.$

(4) The contrast is non-significant. We fail to reject the null hypothesis that cost does not affect the brand quality.

5.12 A Paired-Comparison Experiment Involving Factorial Combinations

The methods of Chapter IV are employed in the following hypothetical paired-comparison experiment.

A paired-comparison experiment is used to determine the effects of storage, curing, and concentration of insecticide on the sweetness of two varieties, A and B say, of sweet

potatoes. The factor corresponding to varieties is denoted as U. The curing factor (V) is taken at two levels, 0 and 12 days. The storage factor (W) is also considered at two levels, 0 and 14 weeks. On the basis of prior experience, the horticulturist assumes that the strength-of-insecticide factor (X) will not interact with the other factors and he is, therefore, free to use the three levels of concentration: 10, 20, and 30 percent.

The $2^3 \times 3$ factorial has, of course, 24 factorial combinations (i.e., $t = 24$). One repetition of the experiment is performed by each of six judges. The judges are instructed to prefer the sweeter member in each paired comparison.

The list of the orthogonal contrasts to be used and the hypothetical results to be tested are given in Table 5.3. An analysis of the results is given in Table 5.4.

The conclusions that may be drawn from the analysis (Table 5.4) are that storage and strength of insecticide affect sweetness, and that the two varieties differ in their reaction to the curing process. A linear decrease in sweetness with increase in insecticide concentration, observed in the large negative value of $Q_{x,1}$, is the cause for significance in testing the X effect. No tests were made of the main effects of curing and varieties because their non-zero interaction effect causes the contrasts Q_{UV}, Q_U, and Q_V to be correlated. The non-significant value of D tends to confirm the horticulturist's assumption that X will not interact with the other factors.

The reader is again reminded that the data and the assumption alluded to the horticulturist are strictly hypothetical and are not intended to represent the true situation.

TABLE 5.3

Construction of Contrasts from the 24 Factorial Combinations in a 2^3 x 3 Factorial Sweet Potato Experiment

Factors	Treatment Combinations																							
U (varieties)	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B
V (days of curing)	0	0	12	12	0	0	12	12	0	0	12	12	0	0	12	12	0	0	12	12	0	0	12	12
W (weeks of storage)	0	0	0	0	14	14	14	14	0	0	0	0	14	14	14	14	0	0	0	0	14	14	14	14
X (% concentration of insecticide)	10	10	10	10	10	10	10	10	20	20	20	20	20	20	20	20	30	30	30	30	30	30	30	30
Code Numbers	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
CONTRASTS										COEFFICIENTS														
Q_U	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1
Q_V	-1	-1	1	1	-1	-1	1	1	-1	-1	1	1	-1	-1	1	1	-1	-1	1	1	-1	-1	1	1
Q_W	-1	-1	-1	-1	1	1	1	1	-1	-1	-1	-1	1	1	1	1	-1	-1	-1	-1	1	1	1	1
$Q_{X,1}$	-1	-1	-1	-1	-1	-1	-1	-1	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
$Q_{X,2}$	-1	-1	-1	-1	-1	-1	-1	-1	2	2	2	2	2	2	2	2	-1	-1	-1	-1	-1	-1	-1	-1
Q_{UV}	-1	1	1	-1	-1	1	1	-1	-1	1	1	-1	-1	1	1	-1	-1	1	1	-1	-1	1	1	-1
Q_{UW}	-1	1	-1	1	1	-1	1	-1	-1	1	-1	1	1	-1	1	-1	-1	1	-1	1	1	-1	1	-1
Q_{VW}	1	1	-1	-1	-1	-1	1	1	1	1	-1	-1	-1	-1	1	1	1	1	-1	-1	-1	-1	1	1
Q_{UVW}	1	-1	-1	1	-1	1	1	-1	1	-1	-1	1	-1	1	1	-1	1	-1	-1	1	-1	1	1	-1
SCORES	62	74	72	68	70	84	80	75	65	70	68	66	70	75	71	64	64	66	63	60	64	65	70	70

TABLE 5.3 (continued)

	U	V	W	X,1	X,2	UV	UW	VW	UVW
S	24	24	24	16	48	24	24	24	24
Q	-3.00	-0.33	+10.00	-10.50	-1.50	+10.00	+0.33	+1.00	+0.67
Q^2/S	.38	.00	4.17	6.89	.05	4.17	.00	.04	.02

TABLE 5.4

Analysis of Treatment Scores

Source of Variation	d.f.	Sum of Squares	Critical Value of $\chi^2_{(0.05)}$	Decision
Interactions				
UVW	1	$Q^2_{UVW}/S_{UVW} = .02$	3.84	Non-significant
UV	1	$Q^2_{UV}/S_{UV} = 4.17$	3.84	Significant
UW	1	$Q^2_{UW}/S_{UW} = .00$	3.84	Non-significant
VW	1	$Q^2_{VW}/S_{VW} = .04$	3.84	Non-significant
Main Effects				
U	1	$Q^2_{U}/S_{U} = .38$	3.84	No Test
V	1	$Q^2_{V}/S_{V} = .00$	3.84	No Test
W	1	$Q^2_{W}/S_{W} = 4.17$	3.84	Significant
X	2	$\sum_{i=1}^{2} Q^2_{X,i}/S_{X,i} = 6.94$	5.99	Significant
D	23	20.9	35.17	Non-significant

VI. SUMMARY

A paired comparison is the process of stating a preference for one member of a pair of treatments on the basis of some common characteristic. This dissertation is confined to balanced experiments in which all possible pairs of treatments are compared one or more times. Methods are developed that do not rely on a specific treatment response distribution, but instead are based on a very general model. The proposed test procedures are, in general, quite easy to apply.

Two tests of the null hypothesis that all treatments in the experiment produce equal stimuli against the general alternative are discussed. One test is for the case in which the property of no interaction between repetitions and preference probabilities (the probabilities of the pairwise preferences of each treatment over each other treatment) may be assumed prior to the experiment. The other test is for the case in which the "no interaction" assumption cannot be made. The number of times a treatment is preferred is called its score. For the "no interaction" case, the test is based on a statistic that is a simple function D of the corrected sum of squares of the treatment scores. In the other case, the value of D is calculated for each homogeneous group of repetitions and then the values are summed to give the new test statistic. The applicability of a χ^2-approximation to the critical values of the two test-statistics for experiments outside the range of the tabled distributions is established. The chief advantage of these two tests over other approximate tests of treatment equality is the ease with which they may be performed. It is shown that the

approximate test based on D is, in general, at least as accurate with respect to errors of the first kind as are the other applicable approximate tests. Bradley (1955) has shown that the approximate test due to Bradley and Terry (1952) for the "no interaction" case has the same asymptotic power as the test based on D under the Bradley-Terry model. The power of the test based on D has not been obtained for other models.

To test the null hypothesis of no interaction between preference probabilities and repetitions against the general alternative, a test based on the theory of χ^2 homogeneity tests is introduced. Again, this is a relatively simple test to perform.

Methods are developed for testing whether (1) a particular treatment stimulus is better than average; (2) two particular treatments have stimuli that are not equal; and (3) the treatment receiving the highest score produces a stimulus that is better than the average of the stimuli in the experiment. The critical values of the test statistics are found under the hypothesis that all preference probabilities are equal. Because of this approach the tests are conservative. These tests are useful to the experimenter interested in the performance of one or two treatments in relation to several others.

A multiple range test analogous to Tukey's test based on allowances is developed to test the null hypothesis of equal treatment stimuli and to separate significantly different treatment scores when the null hypothesis is rejected. This test is of particular importance when the prime purpose of the paired-comparison experiment is to separate out the better (or poorer) treatments from a group of treatments.

A procedure for judging linear contrasts of treatment scores is proposed that is similar to Scheffé's (1953) method for judging contrasts in the analysis of variance. The change from the Scheffé method to the paired-comparisons method is obtained essentially by substituting multiples of treatment scores for the sample means and by using the D-test in place of the analysis of variance F-test. This analysis may be used for such things as finding whether treatment preference varies with the cost of the treatment.

The use of paired-comparison experiments to test factorial effects is discussed and a test method based on orthogonal contrasts of the treatment scores is suggested. Because of correlations that arise, it is necessary to restrict this method to cases in which the only factors that are allowed to appear at more than two levels are those that will not interact with the other factors in the experiment. This method may be useful in preliminary investigations of subjective factorial effects.

The data from a paired-comparison experiment involving the quality of carbon papers are employed to illustrate all the above methods other than the final one, which is used to test factorial effects in a hypothetical experiment.

VII. BIBLIOGRAPHY

Abelson, R. M. and Bradley, R. A. (1954). A 2 x 2 factorial with paired comparisons. Biometrics, 10,487.

Anderson, T. W. (1958). An Introduction to Multivariate Statistical Analysis. New York: John Wiley & Sons, Inc.

Benard, A. and Van Elteren Ph. (1953). A generalization of the method of m rankings. Kon. Ned. Ak. Wet., A, No. 4, and Indag. Math., 15, No. 4, 358.

Bliss, C. I., Greenwood, M. L., and White, E. S. (1956). A rankit analysis of paired comparisons for measuring the effect of sprays on flavor. Biometrics, 12,381.

Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs. I. The method of paired comparisons. Biometrika, 39,324.

Bradley, R. A. (1953). Some statistical methods in taste testing and quality evaluation. Biometrics, 9,22.

Bradley, R. A. (1954a). Rank analysis of incomplete block designs, II. Additional tables for the method of paired comparisons. Biometrika, 41,502.

Bradley, R. A. (1954b). Incomplete block rank analysis. On the appropriateness of the model for a method of paired comparisons. Biometrics, 10,375.

Bradley, R. A. (1955). Rank analysis of incomplete block designs. III. Some large-sample results on estimation and power for a method of paired comparisons. Biometrika, 42,450.

Cochran, W. G. (1954). Some methods for strengthening the common χ^2 tests. Biometrics, 10,417.

Cramer, H. (1946). Mathematical Methods of Statistics. Princeton: Princeton University Press.

David, H. A. (1956). On the application to statistics of an elementary theorem in probability. Biometrika, 43,85.

David, H. A. (1959). Tournaments and paired comparisons. (To be published in Biometrika, 46).

Durbin, J. (1951). Incomplete blocks in ranking experiments. Brit. J. Psychol. (Statist. Sect.), 4,85.

Dykstra, O. (1956). A note on the Rank Analysis of Incomplete Block Designs - Applications Beyond the scope of existing tables. Biometrics, 12,301.

Dykstra, O. (1958). Factorial experimentation in Scheffé's analysis of variance for paired comparisons. J. Amer. Statist. Ass., 53,529.

Federer, W. T. (1955). Experimental Design. New York: The Macmillan Company.

Fleckenstein, M., Freund, R. A. and Jackson, J. E. (1958). A paired comparison test of typewriter carbon papers. Tappi, 41,128.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. J. Amer. Statist. Ass., 32,675.

Gulliksen, H. and Tukey, J. W. (1958). Reliability for the law of comparative judgment. Psychometrika, 23,95.

Haldane, J. B. S. (1939). The mean and variance of χ^2 when used as a test of homogeneity, when expectations are small. Biometrika, 31,346.

Hartley, H. O. (1950). The use of range in analysis of variance. Biometrika, 37,271.

Harvard University Computation Laboratory (1955). Tables of the Cumulative Binomial Probability Distribution. Cambridge, Mass.: Harvard University Press.

Jackson, J. E. and Fleckenstein, M. (1957). An evaluation of some standard techniques used in the analysis of paired comparison data. Biometrics, 13,51.

Jordan, Charles (1932). Approximation and graduation according to the principle of least squares by orthogonal polynomials. Ann. Math. Statist., 3,257.

Kendall, M. G. and Babington Smith, B. (1940). On the method of paired comparisons. Biometrika, 31,324.

Kendall, M. G. (1955). Rank Correlation Methods. New York: Hafner Publishing Company.

Mosteller, F. (1951a). Remarks on the method of paired comparisons: I. The least squares solution assuming equal standard deviations and equal correlations. Psychometrika, 16,3.

Mosteller, F. (1951b). Remarks on the method of paired comparisons: III. A test of significance for paired comparisons when equal standard deviations and equal correlations are assumed. Psychometrika, 16,203.

Pearson, E. S. and Hartley, H. O. (1954). Biometrika Tables for Statisticians. Cambridge: University Press.

Pearson, Karl (1934). Tables of the Incomplete Beta-Function. Cambridge: University Press.

Scheffé, H. (1952). An analysis of variance for paired comparisons. J. Amer. Statist. Ass., 47,381.

Scheffé, H. (1953). A method for judging all contrasts in the analysis of variance. Biometrika, 40,87.

Thurstone, L. L. (1927). Psychophysical analysis. Amer. J. Psychol., 38,368.

VIII. APPENDIX

A. Application of the Multivariate Central Limit Theorem to the $\underline{d}$-Statistic

In Section 2.2, working under H_0' : $\pi_{ij} = \frac{1}{2}$ $(i,j=1,\ldots,t;\ i\neq j)$, it is shown that the random variable $\underline{d} = (d_1,\ldots,d_t)$ has mean $\underline{0}$ and covariance matrix

$$\text{(A.1)} \qquad \sum = \begin{bmatrix} \frac{t-1}{t} & -\frac{1}{t} & \cdot & \cdot & \cdot & -\frac{1}{t} \\ -\frac{1}{t} & \frac{t-1}{t} & & & & \\ \cdot & & \cdot & & & \\ \cdot & & & \cdot & & \\ \cdot & & & & & -\frac{1}{t} \\ -\frac{1}{t} & & & & -\frac{1}{t} & \frac{t-1}{t} \end{bmatrix}$$

In the same section, it is stated that the multivariate central limit theorem may be applied to show that $\underline{d}$ has a multivariate normal limiting distribution as $n\to\infty$. The purpose of this appendix is to prove that statement under H_0'.

The multivariate central limit theorem: Let the t-component vectors $\underline{Y}_1,\underline{Y}_2,\ldots$ be independently and identically distributed with means $E\,\underline{Y}_\gamma = \underline{\mu}$ and covariance matrices $E(\underline{Y}_\gamma - \underline{\mu})(\underline{Y}_\gamma - \underline{\mu})' = T$. Then the limiting distribution of $(1/\sqrt{n})\sum_{\gamma=1}^{n}(\underline{Y}_\gamma - \underline{\mu})$ as $n\to\infty$ is $N(\underline{0},T)$. [A proof is given by Anderson (1958).]

If we let

$$(A.2) \qquad d_{i\gamma} = (2/\sqrt{t})\left[a_{i\gamma} - \tfrac{1}{2}(t-1)\right] \qquad \begin{matrix}(i=1,\ldots,t;\\ \gamma=1,\ldots;n)\end{matrix}$$

where $a_{i\gamma}$ is the score of treatment i in the γ th repetition, then

$$(A.3) \qquad \underline{d}_{\gamma} = \begin{bmatrix} d_{1\gamma} \\ d_{2\gamma} \\ \vdots \\ d_{t\gamma} \end{bmatrix}$$

may be considered as a $\underline{d}$ statistic for a paired-comparison experiment with only one repetition. Therefore, using the results of Section 2.2,

$$(A.4) \qquad E\,\underline{d}_{\gamma} = \underline{0}$$

and

$$(A.5) \qquad E\,\underline{d}_{\gamma}\,\underline{d}'_{\gamma} = \sum .$$

Since the n repetitions of a paired-comparison experiment may be considered under H'_0 as n independent t-variate observations from the same parent population, the $\underline{d}_{\gamma}$'s are independent and identically distributed. Hence, we have from the above multivariate central limit theorem that the limiting distribution of

$$(A.6) \qquad (1/\sqrt{n})\sum_{\gamma=1}^{n} \underline{d}_{\gamma}$$

as $n \longrightarrow \infty$ is $N(\underline{0}, \sum)$. But

$$(A.7) \qquad \underline{d} = (1/\sqrt{n})\sum_{\gamma=1}^{n} \underline{d}_{\gamma} \ ,$$

which proves our statement that $\underline{d}$ has a limiting multivariate normal distribution.

B. Conditions for Maximum Variance of $(a_i - a_j)$

In Section 3.3, we state that under $H_o : \pi_{i.} = \pi_{j.}$ $(i \neq j)$ the variance of $(a_i - a_j)$ is a maximum when

$$H_o' : \pi_{m,n} = \tfrac{1}{2} \qquad (m,n=1,\ldots,t; \quad m \neq n)$$

is also true. A proof of that statement is given here.

As previously defined, $x_{m,n}$ is the number of comparisons in which treatment m is preferred to treatment n.

$$(B.1) \qquad (a_i - a_j) = x_{ij} - x_{ji} + \sum_{k \neq i,j}^{t} (x_{ik} - x_{jk})$$

$$(B.2) \qquad \mathrm{Var}(a_i - a_j) = \mathrm{Var}(x_{ij} - x_{ji}) + \sum_{k \neq i,j}^{t} \mathrm{Var}(x_{ik} - x_{jk})$$

$$= n\Big\{4\,\pi_{ij}\,\pi_{ji} + \sum_{k \neq i,j}^{t} \big[\pi_{ik}(1 - \pi_{ik}) + \pi_{jk}(1 - \pi_{jk})\big]\Big\} .$$

Hence, to maximize the variance of $(a_i - a_j)$ subject to the condition $\pi_{i.} = \pi_{j.}$, we must find the values of $\pi_{m,n}$ that maximize

$$(B.3) \qquad Q = 4\,\pi_{ij}\,\pi_{ji} + \sum_{k \neq i,j} \big[\pi_{ik}\,\pi_{ki} + \pi_{jk}\,\pi_{kj}\big]$$

$$+ \lambda(\pi_{i.} - \pi_{j.}) .$$

$$(B.4) \qquad \frac{\partial Q}{\partial \pi_{ij}} = 4(1 - 2\pi_{ij}) + \lambda(1 + 1) .$$

$$(B.5) \qquad \frac{\partial Q}{\partial \pi_{ik}} = 1 - 2\pi_{ik} + \lambda .$$

$$(B.6) \qquad \frac{\partial Q}{\partial \pi_{jk}} = 1 - 2\pi_{jk} - \lambda .$$

$$(B.7) \qquad \frac{\partial Q}{\partial \lambda} = \pi_{i.} - \pi_{j.} .$$

Setting the partial derivatives equal zero, we have

$$(B.8) \qquad 2 - 4\pi_{ij} + \lambda = 0$$

$$(B.9) \qquad 1 - 2\pi_{ik} + \lambda = 0$$

$$(B.10) \qquad 1 - 2\pi_{jk} - \lambda = 0$$

$$(B.11) \qquad \pi_{i.} = \pi_{j.}$$

Subtracting the sum of (B.9) over $k \neq i$ from the sum of (B.10) over $k \neq j$, one obtains

$$2(t-1)\pi_{i.} - 2(t-1)\pi_{j.} - 2(t-1)\lambda = 0 .$$

But, since $\pi_{i.} = \pi_{j.}$,

$$\lambda = 0 .$$

Hence, the solution to equations (B.8) through (B.11) is obtained by setting all the preference probabilities equal $\frac{1}{2}$. The resulting variance under this condition, H_0', is $\frac{1}{2}nt$.

The fact that $\frac{1}{2}nt$ is a maximum value rather than a minimum is seen from the three-treatment experiment with

$$\pi_{12} = 1, \quad \pi_{23} = 1, \quad \pi_{31} = 1 ,$$

for which the variance of $(a_i - a_j)$ is zero.

IX. ACKNOWLEDGMENTS

The author wishes to express his appreciation to Dr. Herbert A. David for suggesting the thesis topic and for his assistance and advice during the preparation of the dissertation.

The encouragement and counsel offered by Dr. Boyd Harshbarger is appreciated. The author is indebted to Dr. Ralph A. Bradley for his helpful suggestions and criticisms. The author is also grateful for the experimental data supplied by Mr. J. E. Jackson and Miss Mary Fleckenstein.

Thanks are due to Mrs. H. Barns Copenhaver for the care with which she typed the final manuscript.

This study was supported, in part, by the Office of Ordnance Research, U. S. Army, Contract No. D.A.-36-034-ORD-1527 RD.

The vita has been removed from the scanned document

ABSTRACT

TESTS OF SIGNIFICANCE FOR EXPERIMENTS
INVOLVING PAIRED COMPARISONS

by

Thomas Harold Starks B.A., M.S.

Thesis submitted to the graduate faculty of the
Virginia Polytechnic Institute
in candidacy for the degree of

DOCTOR OF PHILOSOPHY
in
STATISTICS

December, 1958

ABSTRACT

New methods for testing hypotheses in paired-comparison experiments are presented in this dissertation. The methods are developed on the basis of a very general mathematical model and they are, in general, quite easy to employ.

Two tests of the null hypothesis that all treatments have equal stimuli, against its general alternative, are proposed. One test is for the case in which it is assumed prior to the experiment that no interaction will take place between repetitions and preference probabilities (the probabilities of the possible comparison preferences). The other test is for the case in which the above assumption cannot be made. The number of times a treatment is preferred is called its score. For the "no interaction" case, the test procedure is based on a test statistic that is a function D of the corrected sum of squares of the treatment scores. In the other case, the value of D is calculated for each group of homogeneous repetitions and then the values are summed to give the new test statistic. It is established that a χ^2-approximation may be used to determine the critical value of the test statistic for experiments outside the range of the tabled distributions. This test procedure is shown to be simpler than other approximate tests and, in general, at least as accurate with respect to errors of the first kind.

It is shown that the two test methods discussed above may be extended to ranking experiments in balanced incomplete block designs with more than two treatments per block.

To test the null hypothesis of no interaction between preference probabilities and repetitions, against its general alternative, a test method based on the theory of χ^2 homogeneity tests is introduced.

Means are presented for testing whether (1) a particular treatment is better than the average of the treatment stimuli; (2) two particular treatment stimuli are not equal; and (3) the treatment receiving the highest score is better than the average. The three test procedures are based essentially on the binomial distribution of the treatment scores under the null hypothesis. In each case, the test procedure is conservative.

A procedure analogous to Tukey's test based on allowances is developed to test the null hypothesis of equal treatment stimuli and to separate the significantly different treatment scores when it rejects the null hypothesis.

A method for judging contrasts of treatment scores similar to Scheffé's (1953) method for judging contrasts in the analysis of variance is proposed. The test method based on D, mentioned earlier, is used in place of the F-test employed in the Scheffé method.

The use of paired-comparison experiments to test factorial effects is discussed and a test method based on orthogonal contrasts of the treatment scores is suggested. Because of correlations that arise, it is necessary to restrict this method to cases in which the only factors that are allowed to appear at more than two levels are those that will not interact with the other factors in the experiment.

The test methods are illustrated through application on the data from two paired-comparison experiments.