SCIENCE, STATISTICS, AND PAIRED COMPARISONS\textsuperscript{1,2}

By

Ralph A. Bradley

FSU Technical Report No. M337
ONR Technical Report No. 92

March, 1975
The Florida State University
Department of Statistics
Tallahassee, Florida 32306

\textsuperscript{1} Presented to the Biometric Society, E.N.A.R. as the President's Invited Address, St. Paul, Minnesota, March 25, 1975.

\textsuperscript{2} Research supported at the Florida State University by the Army, Navy and Air Force through ONR Contract N00014-67-A-0235-0006. Reproduction in whole or in part is permitted for any purpose of the United States Government.
SCIENCE, STATISTICS, AND PAIRED COMPARISONS

Ralph A. Bradley
Department of Statistics
Florida State University, Tallahassee, Florida 32306, U.S.A.

SUMMARY

The paper begins with a prologue on science and statistics. It is suggested that more emphasis should be placed on the scientific method in the teaching of science and that statistics is a science and should be taught and practiced as a science. The statistician as a scientist must be concerned with the formulation, modification and verification of stochastic models.

The main sections of the paper deal with a method of paired comparisons proposed originally by Zermelo and rediscovered by Bradley and Terry, Ford, and others. The basic model is reviewed with emphasis on various approaches that lead to the model. Applications and uses of the model are reviewed along with tests of appropriateness of the model. Extensions of the basic model discussed include adjustment for ties, triple comparisons, use of factorial treatment combinations, and the multivariate case for which a numerical example is included. The paper is not a monograph on paired comparisons and methodology is omitted that does not relate to the basic approach of this paper unless introduced for comparison. Nothing is included on details of experimental design or the design of tournaments.

The paper concludes with an epilogue indicating some need for help for those engaged in the development of statistical methodology from those concerned with the mathematics of statistics and probability.

1 Presented to the Biometric Society, E.N.A.R. as the President's Invited Address, St. Paul, Minnesota, March 25, 1975.

2 Research supported at the Florida State University by the Army, Navy and Air Force through ONR Contract N00014-67-A-0235-0006. Reproduction in whole or in part is permitted for any purpose of the United States Government.
1. PROLOGUE

The title of this presentation is presumptuous. It suggests an age-related progression to the stage of pomposity if not senility. It is misleading also in that the subject of this paper is paired comparisons and a narrow consideration of paired comparisons at that. Nevertheless, the occasion presents an opportunity to make some comments on science and statistics that require emphasis and are too often assumed.

The aim of science is to identify and reveal order in nature. For this to be realistic, it must be assumed that order exists in nature, that that same order will exist in the future, and that causal relationships, no matter how imperfectly understood, regulate the occurrence of events.

Reference is made frequently to "the scientific method". Seldom is that method explicitly discussed in science. Steps in the scientific endeavor are thought generally to be recognition of a perplexity, formulation of the problem, development of possible hypotheses or models, selection of a promising solution, and verification through observation. Science develops in an evolutionary way; models are formed and verified within available technology and they are revised and refined as knowledge is accumulated. Thus the scientific method is an iterative one. Models and hypotheses are formulated as approximate relationships among identified variables. Dependence is placed on describing how change in one or more variables is attended by some orderly variation in others. Adequate predictive models over an experimental region of interest are more easily developed than are structural or functional laws. Interpretation of associative relationships as causal ones has led to many of the great controversies of science. Every statistician can cite examples of regression models useful in prediction that imply an effect for one of its independent variables contrary to intuition and scientific fact.
Consulting statisticians are familiar with the consultee who, after describing his proposed experiment in several sentences, has only one question: "How many observations do I need?" The good consultant elicits information on the objectives of the proposed experiment, its nature, and prior information available, often not without difficulty, and assists the consultee to the development of an experimental design in the initial phase of the collaborative research. The consultee, usually rather pleased with the attention received, may remark that "I came with only one simple question and you (the statistician) have asked me many." Much of the effort has been devoted to making explicit the consultee's rather vague and implicit adherence to the scientific method. How much clarity would be introduced into experimental planning in following orderly consideration of planning steps as set forth, for example, by Cochran (1963) in planning a sample survey? Is scientific method taught in science or does the study of science involve assimilation of a large number of facts and techniques from the past?

Mathematics has variously been labelled the Queen of the Sciences, the Handmaiden of the Sciences or the Language of Science. Statistics is a science because it must search for truth in nature through the development of stochastic models representative, descriptive and explanatory of natural phenomena. That stochastic variation is inherent in nature is manifest in many ways - genetic theory is an obvious example. Random variation is not simply a result of failure to develop the proper mathematical model encompassing the proper set of independent variables in the proper way although inadequate models certainly introduce excess variation and bias. Statistics as a science must assist in the formulation of stochastic models that, as they evolve and are refined, represent both the structural relationships and dependencies of the natural process and its inherent stochastic components.
Teachers of statistics must emphasize the scientific aspects of the discipline. The prospective statistician must master more than a collection of algorithms, elegant and useful though they may be. The prospective statistician must master more than the measure theoretic bases of probability, more than the intriguing techniques of statistical distribution theory, more than the currently developed concepts of statistical inference. The professional statistician must master all of these because he must be able to innovate, approximate and refine as he cooperates in the development of stochastic models and engages in the scientific method. The professional statistician should become familiar with one or more areas of investigation concerned with the development of stochastic models. The biometrician must be defined as a statistician with special training and insights in the bio-medical sciences. For these reasons, the Florida State University programs in statistics have sought a balance between the mathematics of probability and statistics and statistical practice, required student experience in supervised statistical consulting and teaching, and utilized the concept of a doctoral "double major" introduced by I. R. Savage which in effect requires that the doctoral candidate study to the research level two areas of specialization, only one of which must be within the statistics program.

R. A. Fisher (1957), in discussing the underworld of probability, considered levels of credibility as had others before him. Perhaps stochastic models merit similar classification. The most primitive model, acceptable for some purposes, must have predictive ability under future similar circumstances and withstand tests of conformity with data. The most sophisticated model must meet the criteria of the primitive model and reflect known structural relationships among the inherent variables.

The method of paired comparisons has been considered by many writers. Interest has been generated through the paired nature of competitions of many kinds,
through the experimental simplicity of sensory comparisons of items observed in pairs, and through combinatorial properties associated with the construction of experimental designs or the arrangement of tournaments. The author's interest arose in reference to a research contract awarded by the United States Department of Agriculture to the Virginia Agricultural Experiment Station for the development of statistical methods for sensory difference tests. A literature review, Bradley (1953), was published. It was apparent that sensory comparisons should be conducted in incomplete block designs with small block sizes because of sensory fatigue and that ranking within such blocks would be effective in view of discriminatory measurement difficulty. Neither the work of Durbin (1951) nor Mosteller (1951a,b,c) was available as research on paired comparisons was initiated.

We are indebted to Davidson and Farquhar (1973) for an extensive bibliography on paired comparisons. Its value would be enhanced through simple coding or annotation and it should then be published because, as Gridgeman (1963) has stated "Any statistical method that engages the interests of experimentalists in various fields and of theoreticians with different specialties is almost bound to generate a raveled literature."

The basic paired comparisons experiment has \( t \geq 2 \) treatments, \( T_1, \ldots, T_t \), considered in pairs with \( \pi_{ij} > 0 \) independent selection decisions on the comparison of \( T_i \) with \( T_j \), \( i < j \), \( i,j = 1, \ldots, t \). We shall write \( T_i \rightarrow T_j \) for "\( T_i \) is selected over \( T_j \)" where the selection may be on the basis of preference, intensity, typicality, \ldots. The probability associated with the selection will be simply \( P(T_i \rightarrow T_j) \). An excellent primitive model postulates the existence of \( \binom{t}{2} \) functionally independent parameters, \( 0 \leq \pi_{ij} \leq 1 \), such that

\[
P(T_i \rightarrow T_j) = \pi_{ij}, \quad \pi_{ij} + \pi_{ji} = 1, \quad i \neq j, \quad i,j = 1, \ldots, t.
\] (1.1)
The whole experiment is regarded as the combination of \( \binom{t}{2} \) sets of unrelated independent Bernoulli trials. A stochastic model has thus been posed. It lacks parsimony in parameterization. It does not in any way account for any available knowledge on treatment formulation nor for the psychophysical behavior of choice or competition. More sophisticated models are available based on heuristic assumption or theory of choice behavior. These models incorporate postulates of psychophysical behavior but largely ignore treatment design and certainly fall short of being structural models incorporating measurable aspects of the treatments or the response mechanism. Much more is needed.

Perhaps this prologue has belabored the obvious but perhaps the obvious should sometimes be emphasized. The remainder of this paper will be specific, reviewing the development of a statistical model for paired comparisons, its psychophysical bases, its extensions, and its association with other methods. A brief epilogue suggests some short-comings of statistical theory and extensions needed.

2. A BASIC MODEL FOR PAIRED COMPARISONS

Bradley and Terry (1952a) and Terry, Bradley, and Davis (1952) proposed a basic model for paired comparisons. Dykstia (1960) extended the method to permit unequal numbers \( n_{ij} \) of selection decisions. The basic approach was heuristic in extension of the notion that \( \pi \) and \( (1 - \pi) \) in some sense measure the relative worths of two items compared in \( n \) independent Bernoulli trials where \( \pi \) is the probability of selection of a designated item of the pair.

Treatment parameters, \( \pi_1, \ldots, \pi_t \), \( \pi_i \geq 0 \), \( i = 1, \ldots, t \), were associated with the treatments, \( T_1, \ldots, T_t \). It was postulated that these parameters represented relative selection probabilities for the treatments so that \( \pi_{ij} \) in (1.1) could be structured,
\[ P(T_i + T_j) = \frac{\pi_{ij}}{\pi_i + \pi_j}, \quad i \neq j, \quad i, j = 1, \ldots, t \quad (2.1) \]

Since the right-hand member of (2.1) is invariant under a change in scale, specificity was achieved by the requirement that

\[ \sum_{i} \pi_i = 1. \quad (2.2) \]

The stochastic model (2.1) imposes structure in that the \(^tC_2\) functionally independent parameters of (1.1) are given now in terms of \((t - 1)\) independent parameters. It is a scientific requirement that the adequacy of this model be checked in applications.

Likelihood estimation was used and likelihood ratio tests were developed. The binomial component of the likelihood function for the \(n_{ij}\) comparisons of \(T_i\) and \(T_j\) is

\[
\left( \frac{\pi_i}{\pi_i + \pi_j} \right)^{a_{ij}} \left( \frac{\pi_j}{\pi_i + \pi_j} \right)^{a_{ji}} = \pi_i^{a_{ij}} \pi_j^{a_{ji}} / (\pi_i + \pi_j)^{n_{ij}}
\]

with \(a_{ij} + a_{ji} = n_{ij}\), \(a_{ij}\) being the number of selections of \(T_i\), ties or non-selections not being permitted in the basic model. The complete likelihood function, from independence, is

\[ L = \prod_{i} \pi_i^{a_i} / \prod_{i<j} (\pi_i + \pi_j)^{n_{ij}} \quad (2.3) \]

where \(a_i = \sum_{j \neq i} a_{ij}\), the total number of selections of \(T_i\). After minor simplifications, maximization of \(\ln L\) subject to (2.2) yields the likelihood equations,
\[
\frac{a_i}{p_i} - \sum_{j \neq i} \frac{n_{ij}}{(p_i + p_j)} = 0, \quad i = 1, \ldots, t
\]

\[
\sum_i p_i = 1,
\]

where \(p_i\) is the estimator of \(\pi_i\).

Solution of equations (2.4) is done iteratively. If \(p_i^{(k)}\) is the \(k\)-th approximation to \(p_i\),

\[
p_i^{(k)} = \frac{p_i^{(k)}}{\sum_i p_i^{(k)}}
\]

where

\[
p_i^{(k)} = \frac{a_i}{\sum_j n_{ij}/(p_i^{(k-1)} + p_j^{(k-1)})}, \quad k = 1, 2, \ldots.
\]

The iteration is started with initial specification of \(p_i^{(0)}\), perhaps \(p_i^{(0)} = 1/t\), \(i = 1, \ldots, t\). Solution of (2.4) was difficult in 1952 but is trivial with modern computers. Bradley and Terry (1952a) and Bradley (1954a) provided tables for \(t = 3, 4, 5\); Dykstia (1956) suggested means of obtaining good initial values \(p_i^{(0)}\).

Ford (1957) proposed the model independently and proved that the iterative procedure converged to a unique maximum for \(L\) if \(\pi_i > 0, \ i = 1, \ldots, t\). El Helbawy (1974) noted that \(a_1, \ldots, a_t\) constitute a set of minimal sufficient statistics for estimation of \(\pi_1, \ldots, \pi_t\).

The major test proposed was that of treatment selection equality,

\[
H_0: \ \pi_1 = \pi_2 = \cdots = \pi_t = 1/t,
\]

against the alternative,

\[
H_a: \ \pi_i \neq \pi_j \text{ for some } i, j, \ i \neq j, \ i, j = 1, \ldots, t.
\]
The likelihood ratio statistic is

\[-2 \ln \lambda_1 = 2N \ln 2 - 2B_1, \quad N = \sum_{i<j} n_{ij}, \]

(2.5)

\[B_1 = \sum_{i<j} n_{ij} \ln(p_i + p_j) - \sum_i a_i \ln p_i.\]

For large \(n_{ij}\), \(-2 \ln \lambda_1\) has the central chi-square distribution with \((t - 1)\) degrees of freedom under \(H_0\). Tables of \(B_1\) were provided with the cited tabled solutions of (2.4). Comparison of use of the large-sample test with small-sample significance values in the tables suggests that the former may be used for modest values of \(n_{ij}\), a situation perhaps comparable to use of the normal approximation to the binomial.

Bradley (1955) investigated large-sample properties of the test with each \(n_{ij} = n\). Davidson and Bradley (1970), considering a multivariate model discussed below, obtained as a special case the more general results. Let \(\mu_{ijN} = n_{ij}/N\) and let \(\lim_{N \to \infty} \mu_{ijN} = \mu_{ij}\). It was shown that the limiting joint distribution function of \(\sqrt{N}(p_1 - \pi_1), \ldots, \sqrt{N}(p_t - \pi_t)\) as \(N \to \infty\) is the singular multivariate normal distribution function of dimensionality \((t - 1)\) in a space of \(t\) dimensions with zero mean vector and dispersion matrix \(\Sigma = [\sigma_{ij}]\) such that

\[\sigma_{ij} = \text{cofactor of } \lambda_{ij} \text{ in } \begin{bmatrix} A & 1' \\ 1 & 0 \end{bmatrix} \begin{bmatrix} A & 1' \\ 1 & 0 \end{bmatrix}^{-1}\]

(2.6)

where \(A = [\lambda_{ij}]\), \(1\) is the \(t\)-dimensional unit row vector, and

\[\lambda_{ii} = \frac{1}{\pi_i} \sum_{j \neq i} \mu_{ij} \frac{\pi_j}{(\pi_i + \pi_j)^2}, \quad i = 1, \ldots, t,\]

\[\lambda_{ij} = -\mu_{ij}/(\pi_i + \pi_j)^2, \quad i \neq j, \quad i,j = 1, \ldots, t.\]
The variance and covariance of (2.6) may be used to calculate approximate variances and covariances of special contrasts among the \( p_i \), or, indeed, among the \( \ln p_i \), the latter having appeal for reasons apparent from Section 3. In the special case with \( \pi_i = 1/t, \ n_{ij} = n \) for all \( i, j \),

\[
\sigma_{ij} = \frac{2(t-1)^2}{t^3} \quad \text{and} \quad \sigma_{ij} = -\frac{2(t-1)}{t^3}, \ i \neq j.
\]  

(2.7)

Use of (2.7) is adequate when the ratios \( \pi_i/(\pi_i + \pi_j) \) are in a central range on the unit interval.

The non-central limiting distribution is available for \(-2 \ln \lambda_1\). Let

\[
\pi_i = \frac{1}{t} + \frac{\delta_iN}{\sqrt{N}}, \ \sum_i \delta_iN = 0, \ \lim_{N \to \infty} \delta_iN = \delta_i.
\]

Then \(-2 \ln \lambda_1\) has the non-central chi-square limiting distribution with \( t-1 \) degrees of freedom and noncentrality parameter,

\[
\lambda^2 = \frac{t^2}{4} \sum_{i<j} \nu_{ij} (\delta_i - \delta_j)^2.
\]

Bradley (1955) obtained the asymptotic relative efficiency for balanced \( n_{ij} = n \) paired comparisons with the analysis of variance for a comparable size completely randomized design given conditions for the latter as \( t/\pi(t-1) \). van Elteren and Noether (1959) make the comparison with the analysis of variance for the matching balanced incomplete block design and obtain \( 2/\pi \) as the asymptotic relative efficiency, the known result for the sign test being a special case.

Bradley and Terry (1952) considered other test situations. Suppose that the paired comparisons experiment may be subdivided into \( g \) subexperiments, each with its own corresponding parameters \( n_{iju}, \pi_{iju} \) and statistics \( p_{iju}, \ B_{lu}, -2 \ln \lambda_{lu}, \ u = 1, \ldots, g \). The subexperiments may have been performed by different judges or groups of homogeneous judges or under conditions otherwise suggesting such subdivision. Tests of hypotheses of interest are
\[ H_0^i: \pi_{iu} = 1/t, \quad i = 1, \ldots, t, \quad u = 1, \ldots, g \] versus \[ H_a^i: \pi_{iu} \neq 1/t, \text{ some } i, u, \] and

\[ H_0^H: \pi_{iu} = \pi_i, \quad i = 1, \ldots, t, \quad u = 1, \ldots, g \] versus \[ H_a^H: \pi_{ij} \neq \pi_i, \text{ some } i, u. \] (2.8)

The test statistics respectively are

\[ \sum_u (-2 \ln \lambda_{1u}) \quad \text{and} \quad \sum_u (-2 \ln \lambda_{1u}) - (-2 \ln \lambda_1) \]

with central chi-square limiting distributions with \( g(t-1) \) and \((g-1)(t-1)\) degrees of freedom under the null hypotheses where \(-2 \ln \lambda_1\) in the second statistic is computed after pooling of the subexperiment data for a single analysis. The test of \( H_0^H \) against \( H_a^H \) is a test of group by treatment interaction. A simple approximate analysis of chi-square table may be formed combining the three tests described. One other test was proposed but it will not be given in detail. It was assumed that the treatments fell into two groups, say \( \pi_1, \ldots, \pi_s = \bar{\pi}, \quad \pi_{s+1}, \ldots, \pi_t \)

\( = (1 - s\bar{\pi})/(t - s) \) and the test is one of the equality of \( \pi \) and \( (1 - s\bar{\pi})/(t - s), \)

again \( \pi_i = 1/t, \quad i = 1, \ldots, t. \)

In the work above, values of \( n_{ij} \) and \( n_{iju} \) may be zero except that there must be linkage of treatment comparisons in each experiment or subexperiment. Badly balanced experiments will yield disparate variances of estimators. Particular experimental designs may be used for special purposes or to achieve certain balance - Kendall (1955), Clatworthy (1955). Bose (1956) developed balanced designs with comparisons so grouped that individual judges may be given \( r \) comparisons,

\[ 1 \leq r \leq t(t-1)/2. \] Dykstia (1960) considered use of the Clatworthy partially balanced designs with the basic paired comparisons model. Wilkinson (1957) used the Bose designs with \( v \) judges in \( g \) groups of judges and generalized the basic tests given above; he was not able to develop a test of judge by treatment interaction but only of group by treatment interaction on the assumption of within-group judge homogeneity. Rai (1971a) uses a designated or control treatment with
replicated comparisons of other treatments with the control; it was not necessary for him to redevelop the analysis for this special case. Various tournament designs exist but they lack sufficient replication.

Beaver (1974) has proposed an alternative test statistic for the test of equality of treatment parameters and included the possibility of ties or non-selection, a topic discussed below. He uses the locally asymptotic most stringent test and claims the advantage that calculation depends only on a matrix inversion and not on iterative solution of likelihood equations. Beaver has shown also that his test has unit asymptotic relative efficiency in comparison with the likelihood ratio test of (2.5). It is not clear that matrix inversion is preferable to iteration with modern computers and either calculation is now easy. The method does not produce estimates of treatment parameters and we suspect that the two tests are asymptotically equivalent as is the basic method described by David (1963, Section 3.5).

3. BASES FOR MODEL FORMULATION

The basic model for paired comparisons has been discovered and rediscovered by various authors. Sometimes it arises as one of the special simple realizations of a generalized model developed from distributional or psychophysical approaches. In this section, we review some of these bases for the model.

3.1 The Heuristic Approach. Reference has been made to the approach of Bradley and Terry and later of Ford in the preceding section. But the first formulation of the model is due to Zermelo (1929) in consideration of chess competition without intervening development. Both Zermelo and Ford concentrate on solution of equations (2.4) and convergence of the iterative solution.
3.2 Linear Models. Thurstone (1927) modeled paired comparisons through the concept of a subjective continuum, an inherent sensation scale on which order but not physical measurement could be discerned. Mosteller (1951a) provides a detailed formulation and analysis of Thurstone's important Case V. With suitable scaling each treatment has a location point on the continuum, say \( \mu_i \) for \( T_i \), \( i = 1, \ldots, t \). An individual is assumed to receive a sensation \( X_i \) in response to \( T_i \) with responses \( X_i \) normally distributed about \( \mu_i \). When an individual compares \( T_i \) and \( T_j \), he in effect reports the order of sensations \( X_i \) and \( X_j \) which may be correlated; \( X_i > X_j \) may be associated with \( T_i > T_j \). Case V takes all such correlations equal and the variances of all \( X_i \) equal. The probability of selection may be written,

\[
P(T_i > T_j) = P(X_i > X_j) = \frac{1}{\sqrt{2\pi}} \int_{-(\mu_i - \mu_j)}^{\infty} e^{-y^2/2} \, dy, \quad i \neq j, \quad i, j = 1, \ldots, t. \quad (3.1)
\]

Mosteller associates observed proportions of selection of \( T_i \) over \( T_j \), \( a_{ij}/n_{ij} \) in our notation, with the probabilities of (3.1) and develops a method of estimation of \( \mu_i \), \( i = 1, \ldots, t \), after arbitrary choice of origin. Bradley (1953) noted that replacement of the normal density function in (3.1) by that for the logistic yields

\[
P(T_i > T_j) = \frac{1}{4} \int_{-(\ln \pi_i - \ln \pi_j)}^{\infty} \text{sech}^2 y/2 \, dy = \frac{\pi_i}{(\pi_i + \pi_j)}. \quad (3.2)
\]

Thus the comparison of models (3.1) and (3.2) is very analogous to a comparison of probit and logit methods in biological assay. Clearly there is a correspondence between \( \mu_i \) and \( \ln \pi_i \); note the earlier remark on contrasts involving \( \ln \pi_i \).
David (1963, Section 1.3) supposes that $T_i$ has "merit" $V_i$ when judged on some characteristic, $i = 1, \ldots, t$, and that these merits may be represented on a merit scale. He notes that it is possible to construct a merit scale such that

$$P(T_i \rightarrow T_j) = H(V_i - V_j) \tag{3.3}$$

where $H$ is a distribution function for a symmetric distribution, $H(-x) = 1 - H(x)$. Models described by (3.3) were called linear models and (3.1) and (3.2) are special cases. Model (3.3) arises again in subsections below. Daniels (1969) considers fair scoring systems in round-robin tournaments and is led to linear models and to our basic model. Moon and Pullman (1970) more formally and generally follow up on the Daniels approach. Bühlmann and Huber (1963) formulate the "correct ranking" problem of paired comparisons in a decision theoretic framework. Let $\Pi = [\pi_{ij}]$, $\pi_{ij} \neq 0$, $\pi_{ij}$ defined in (1.1), and let $W_\Pi(A) = \prod_{i \neq j} \pi_{ij}^{a_{ij}}$ where $A$ denotes the outcome of the experiment in terms of the $a_{ij}$. Let $a = (a_1, \ldots, a_t)$. They prove (Theorem 1) that $W_\Pi(A)$ depends on $A$ only through $a$ if and only if the linear model (3.3) pertains with $H(t) = 1/(1 + e^{-t})$, the logistic function. Thus model (3.2) follows. Note the earlier comment on sufficiency of $a_1, \ldots, a_t$ for the estimation of $\pi_1, \ldots, \pi_t$.

3.3 The Lehmann Model. Lehmann (1953) considered the power of rank tests. He provided distributional models for alternative hypotheses which were themselves distribution-free in a sense and parameter-dependent in a simple way. He was motivated by a desire to provide easy power comparisons rather than a desire to define realistic models. Nevertheless, his models have been used in practice and can be used in paired comparisons.

The Lehmann concepts in simplest form for paired comparisons follows. Let $F(x)$ be any continuous distribution function. If $X_i$ is again a sensation induced by $T_i$, let the distribution function of $X_i$ be
\[ P(X_i \leq x) = F_i^\pi(x), \quad i = 1, \ldots, t, \]  

(3.4)

where \( \pi_i > 0 \). If an individual compares \( T_i \) and \( T_j \), receiving independent responses \( X_i \) and \( X_j \), then

\[ P(T_i \neq T_j) = P(X_i > X_j) = \int_{x_1 \neq x_j} dF_i(x_1) dF_j(x_j) = \frac{\pi_i}{\pi_i + \pi_j}, \]

(3.5)

\[ i \neq j, \quad i, j = 1, \ldots, t. \]

Note that the selection probability in (3.5) does not depend on the unknown distribution function \( F(x) \) and that our basic model again is obtained. Bradley (1965) discussed this application of the Lehmann model and it is implicit in the work of Savage (1956). A more general form permitting relaxation of the independence assumption on \( X_i \) and \( X_j \) was used and triple comparisons were also considered.

Davidson (1969) associates the Lehmann model and the logistic model of (3.2). He took \( F(x) \) in (3.4) to be \( F(x,0) \) where

\[ F(x, \theta) = \exp[-e^{-(x-\theta)}], \quad -\infty < x < \infty, \quad -\infty < \theta < \infty, \]

(3.6)

is an extreme value distribution and noted that \( F(x, \theta) = [F(x,0)]^{\gamma} \) where \( \gamma = e^\theta \). In the Lehmann model, \( \theta \) is associated with \( \ln \pi_i, \quad i = 1, \ldots, t \). In the linear model, it follows that \( H \) must be logistic with \( V_i \) associated with \( \ln \pi_i \).

3.4 A Psychophysical Model. Thompson and Singh (1967) assume that a treatment \( T \) acting as a stimulus activates signals \( u_1, \ldots, u_m \) transmitted to the brain by \( m \) sense receptors. As a first approximation, \( m \) is taken to be fixed and
large and the individual signals are taken to be independent and identically distributed random variables. Two main variables for the response $X$ to treatment $T$ are considered:

(i) $X$ is the sum (or average) of signals $u_1, \ldots, u_m$.

(ii) $X$ is the largest of the signals $u_1, \ldots, u_m$.

Some justifications for each model are given in reference to Stevens (1957).

Limit theory is used to obtain results for both models. For model (i), normality for $X$ results. When $T_i$ and $T_j$ are compared, $X_i$ and $X_j$ are taken to be independent and the Thurstone model (3.1) for paired comparisons results. For model (ii), it is shown that $X$ has one of three distributions depending on whether $G$, the common distribution function of $u_1, \ldots, u_m$ is of the exponential, Cauchy or limited type. The exponential type is most interesting because it includes many common statistical distributions. Then the limiting distribution function of $X$ has the extreme value distribution (3.6). For all three limiting distribution functions for $X$, the basic model (2.1) for paired comparisons results when $X_i$ and $X_j$ are compared; the parameter $\pi_i$ for $T_i$ takes different forms in terms of basic parameters of Thompson and Singh.

Thompson and Singh implicitly anticipate Davidson's association of Lehmann and logistic models because they deal with (3.6) as one of their limiting distributions for model (ii).

It seems not to have been noticed that the Thompson-Singh argument for model (ii) can lead directly to the Lehmann model without resorting to limit theory. Suppose that $T$ evokes signals from exactly $r$ of $m$ sense receptors. Then, under model (ii), the distribution function of $X$ is $G^r(x)$. When $T_i$ and $T_j$ are compared, the responses $X_i$ and $X_j$ have the basic Lehmann model (3.4) with
distribution functions $G_i(x)$ and $G_j(x)$ and $P(T_i \rightarrow T_j) = r_i / (r_i + r_j)$.

Then, for example, it is possible to set $\pi_i = c r_i / m$ where $c$ is chosen to satisfy (2.2). For $r_i$ and $m$ large, $i = 1, \ldots, t$, $\pi_1, \ldots, \pi_t$ so defined will adequately cover their possible ranges.

3.5. Models of Choice and Worth. The basic model of paired comparisons arises from various axioms and postulates of choice behavior.

Luce (1959) poses an axiom of choice. Let $S, T, S \subset T$, be two finite subsets of a universal set $U$ and let $P_S$ and $P_T$ be probability measures of choice defined on $S$ and $T$ respectively. The axiom of choice essentially requires that, for a set $R \subset S$, $P_S(R) = P_T(R | S)$. In essence, the probability measure for choice in $S$ is induced by the conditional probability measure on $S$ based on $P_T$ on $T$. It is a consequence of the axiom that there exists a positive real-valued function $v$ on $T$, unique except for multiplication by a positive constant, such that for every $S \subset T$, for $x \in S$,

$$P_S(x) = v(x) / \sum_{y \in S} v(y).$$

The model for paired comparisons results if $T = \{T_1, \ldots, T_t\}$, $S = \{T_i, T_j\}$, and $v(T_i) = \pi_i$. Then $P(T_i \rightarrow T_j | T_i) = \pi_i / (\pi_i + \pi_j)$, $i \neq j$, $i, j = 1, \ldots, t$.

Audley (1960) sees choice or selection as a dynamic process in time. If selection is to be made from among $t$ items, the stimulus of each item is regarded as a Poisson process with parameter $\lambda_i$ for $T_i$ generating "implicit responses" with probability $\lambda_i \Delta T$ in time interval $(T_i, T_i + \Delta T)$. An implicit response is not well defined but may be regarded to be a fleeting impulse to select $T_i$ at a given small interval in time. A final choice $T_i$ is made when $K$ successive implicit responses indicate choice of $T_i$. The Poisson processes are assumed to
be independent. When \( K = 1 \), and only items \( T_i \) and \( T_j \) are considered,
\[
P(T_i) = \frac{\alpha_i}{\alpha_i + \alpha_j}.
\]
Clearly \( \alpha_i \) is associated with \( c_{ii} \) and the basic model follows. When \( K = 2 \) and only items \( T_i \) and \( T_j \) are considered,
\[
P(T_i) = \frac{\alpha_i}{\alpha_i + \alpha_j} \frac{(\alpha_i + \alpha_j)^2 - \alpha_j^2}{(\alpha_i + \alpha_j)^2 - \alpha_i \alpha_j},
\]
argued to be a stronger (or better considered) choice of \( T_i \) yielding a larger value of \( P(T_i) \) than before when \( \alpha_i > \alpha_j \).

Block and Marschak (1960) develop utility theories associated with choice, selection or ordering. In our notation, the ranking of \( t \) items in a set
\( T = \{ T_1, \ldots, T_t \} \) is considered on the basis of pairwise preferences, \( r_i \leq r_j \) if \( T_i \rhd \) \( T_j \). A real-valued function \( w \) defined on \( T \) is a utility function on \( T \) if \( w(T_i) \geq w(T_j) \) if and only if \( T_i \rhd \) \( T_j \). Weak, strong and strict utilities are defined:

Weak utility: \( w \) is a weak utility on \( T \) if \( w_i \geq w_j \) if and only if
\[
P(T_i + T_j) \geq \frac{1}{2}, \quad w_i = w(T), \quad i = 1, \ldots, t.
\]

Strong utility: \( w \) is a strong utility on \( T \) if there exists a strictly increasing distribution function \( \phi_w \) such that \( \phi_w(w_i - w_j) = P(T_i \rightarrow T_j) \), \( \phi_w(0) = \frac{1}{2} \).

Strict utility: \( w \) is a strict utility on \( T \) if \( P_M(T_i) \), the probability of selection of \( T_i \) from any subset \( M \) of \( T \) containing \( T_i \), is such that
\[
\frac{w_i}{w_j} = P_M(T_i)/P_M(T_j), \quad \text{all } M \subseteq T,
\]
any \( T_i, T_j \in M \).
Note that strong utility may be associated with the linear model concept of Subsection 3.2 and that strict utility is equivalent to the Luce axiom of choice. Location of the origin on the utility scale is important only for strict utility where each \( w_i \geq 0 \). Without giving details, we note that Block and Marschak proceed to state Condition (π) which implies that the probability of any ranking is the product of successive probabilities of first choices. Condition (π) is the basis of one of the models given in Subsection 5.2 on triple comparisons.

Mathematical models for ranking in paired comparisons are considered by Brunk (1960). He considers the score \( S_{ij} \), \( S_{ij} = 1,0 \) as \( T_i \) or \( T_j \) is selected, \( S_{ij} + S_{ji} = 1 \), and bases his modeling on \( e_{ij} = E(S_{ij}) \), \( e_{ij} + e_{ji} = 1 \), \( e_{ii} = \frac{1}{2} \). Brunk assumes that each treatment has an "intrinsic worth" and that these intrinsic worths determine the expected scores \( e_{ij} \). An intrinsic worth model is "weakly regular" if \( e(u,v) \), the expected score of an item of worth \( u \) when compared with an item of worth \( v \), is such that \( e(u,v) \geq \frac{1}{2} \) if \( u > v \); the model is "strongly regular" if \( e(u,v) \) is non-decreasing in \( u \) and non-increasing in \( v \). A sufficient condition for an intrinsic worth model to be strongly regular is given as \( e(u,v) = H(u - v) \) where \( H \) is the distribution function of a symmetric random variable. We are led again to the concept of the linear model of Subsection 3.2.

The treatment of models of choice and worth given here may not be exhaus-
tive. It is intended to provide the flavor of modeling of this type.
4. APPLICATIONS AND UTILIZATIONS

4.1 A Test of the Model. If a statistician is a scientist as claimed, he has an obligation to consider the appropriateness of stochastic models formulated through comparisons with observation. This can be done for the basic method of paired comparisons through a goodness-of-fit test [Bradley (1954b)]. The model (1.1) forms the basis for the procedure since enough parameters are present to fit paired comparisons data perfectly.

Our model for paired comparisons is hypothesized through the null hypothesis,
\[ H_0^*: \pi_{ij} = \pi_i/(\pi_i + \pi_j), \text{ } i \neq j, \text{ } i,j = 1, ..., t, \] (4.1)
and compared with the primitive general model through the alternative hypothesis,
\[ H_a^*: \pi_{ij} \neq \pi_i/(\pi_i + \pi_j), \text{ for some } i,j, \text{ } i \neq j. \] (4.2)

The resulting likelihood ratio test statistic is
\[ -2 \ln \lambda^* = -2\left[ \sum_{i \neq j} a_{ij} \ln(n_{ij}p_i/(p_i + p_j)) - \sum_{i \neq j} a_{ij} \ln a_{ij} \right] \] (4.3)
which has the central chi-square distribution with \( \binom{t}{2} - t + 1 \) degrees of freedom for large \( n_{ij} \) under \( H_0^* \). The statistic can be approximated in the classical form,
\[ -2 \ln \lambda^* \approx \sum_{i \neq j} \left( a_{ij} - \frac{n_{ij}p_i}{p_i + p_j} \right)^2/[n_{ij}p_i/(p_i + p_j)], \]
showing observed frequency \( a_{ij} \) compared with expected frequency \( n_{ij}p_i/(p_i + p_j) \).

The two forms of the statistic are asymptotically equivalent. In the situation where the experiment is subdivided into \( g \) subexperiments, the test of fit may be applied to each subexperiment and a composite or over-all test statistic formed as the sum of \( g \) statistics like that of (4.3) with associated sum of degrees of freedom.
Bradley applied the test of fit to 20 small experiments dealing with the
effects of peanut supplements in hogs' rations on the flavor of fresh roast pork,
the effect of color set on the appearance of apples, and the effect of monosodium
 glutamate on the flavor of apple sauce and apple slices. Of the 20 tests made,
only one was significant at the .05-level of significance. Hopkins (1954)
designed and analyzed a more extensive series of experiments involving basic taste
characteristics (sweetness, bitterness, saltiness, sourness). His summary state-
ment is: "It is therefore perhaps noteworthy that in these trials individuals'
preference rankings of the qualitatively different test solutions were as consis-
tent with the Bradley-Terry conceptual model as were the rankings of single flavor
intensities." Gridgeman (1955) considered preference data and concluded, "The
conclusion that a preference continuum is --- a workable concept is the most
important single outcome of the investigation." The general experience of the
author has been that experiments exhibiting poor fit of the model have one or
more treatments grossly different from the others on some characteristic.

Atkinson (1972), following Cox (1970), considers the log-odds ratio in a
logistic model and expresses

$$\lambda_{ij} = \ln \left( \frac{\pi_{ij}}{1 - \pi_{ij}} \right) = \rho_i - \rho_j , \quad \rho_i = \ln \pi_i , \quad i \neq j , \quad i, j = 1, \ldots, t . \quad (4.4)$$

In this way, (4.4) represents a "linear" model describing our basic model for
paired comparisons. The alternative model to (4.4) is given as

$$\lambda_{ij} = (\rho_i - \rho_j)[1 + \exp(\kappa(\rho_i - \rho_j)^2)] , \quad i \neq j ,$$

and a test of the hypothesis that $\kappa = 0$ is given. The test of the model so
developed is in a sense a test of linearity as described by (4.4) and hence the
alternative to the basic model is very specific. Atkinson uses one of the author's
examples tending to exhibit poor model fit and finds significance for his test and
alternative.
4.2 Applications. The reader may be interested in applications of our method of paired comparisons. Fleckenstein, Freund and Jackson (1958) compare typewriter carbon papers. Jackson and Fleckenstein (1957) use color preference data to compare available methods of analysis. Nakagami (1961) exhibits multivariate data on taste, odor, color and overall preferences for three pharmaceutical products by children and their mothers; univariate analyses are made. Park (1961) compared margarines in consumer studies using both paired and triple comparisons.

4.3 Utilizations. The concept of the basic model has been utilized in some other situations. Davidson and Solomon (1973) consider a Bayesian approach to paired comparison experimentation. Among the results obtained is one that the posterior estimators of \( \pi_1, \ldots, \pi_t \) obtained as the modal point of the posterior distribution result from solution of equations like (2.4) where \( n_{ij} \) and \( a_i \) are replaced by \( n'_{ij} = n_{ij} + n_{ij}^{0} \) and \( a_i' = a_i + a_i^{0} \), \( n_{ij}^{0} \) and \( a_i^{0} \) being information in the prior distribution in equivalences to \( n_{ij} \) and \( a_i \) from the experimentation.

Mallows (1957) considers non-null ranking models. Basically, he seeks the probability of a given ranking of \( n \) objects when information on the known true ranking is summarized in terms of the parameters \( \pi_{ij} \) associated with paired selections. The model (2.1) is used to reduce the number of parameters in the problem; various other simplifying assumptions are considered.

Schönemann and Wang (1972) combine the model (2.1) with the Coombs "unfolding model" to obtain individual differences for the multidimensional analysis of preference data. Distances are formulated in a multidimensional "joint space" (the author would describe this as a factor space) containing \( t \) points representing treatments or stimuli and points representing "ideal points" for each of \( g \) judges or respondents. For judge \( u \),
\[ P(T_i > T_j) = 1/[1 + \exp(-c(d_{ju}^2 - d_{iu}^2))] \]

a representation in logistic form, where \( \pi_{ij} = \exp(-cd_{iu}^2) \) and \( d_{iu} \) is the Euclidean distance between the location of \( T_i \) and the ideal point for judge \( u \) in the joint space. The reader interested in this approach should read also Hayashi (1964).

5. EXTENSIONS OF THE BASIC MODEL

5.1 Adjustment for Ties. The treatment of ties or non-selection responses in the sign test has received considerable attention. Hemelrijk (1952) demonstrated that the most powerful test of significance was obtained by omission of ties and use of a conditional binomial test on the sample results so reduced, but the treatment of ties must depend on experimental objectives [see Gridgeman (1959)] and estimation of potential share of a consumer market surely must require other considerations. Decisions for paired comparisons must be similar to those for the sign test.

Consideration of ties in paired comparisons has taken the form of extension of the basic model to accommodate ties. Rao and Kupper (1967) introduce a parameter \( \Theta \geq 1 \) and adjust probabilities associated with the comparison of \( T_i \) and \( T_j \) to obtain

\[ P(T_i \rightarrow T_j) = \pi_i / (\pi_i + \Theta \pi_j) \]
\[ P(T_j \rightarrow T_i) = \pi_j / (\Theta \pi_i + \pi_j) \]
\[ P(T_i = T_j) = (\Theta^2 - 1) \pi_i \pi_j / (\pi_i + \Theta \pi_j) (\Theta \pi_i + \pi_j) \]  

(5.1)

The derivation of (5.1) is based on the special case (3.2) of the linear model; \( \Theta \) is associated with a threshold parameter \( \eta \), \( \eta = \ln \Theta \) and (3.2) is rewritten,
\[ P(T_i \rightarrow T_j) = \frac{1}{4} \int_{-\infty}^{\infty} \operatorname{sech}^2 \frac{y}{2} dy \quad \] 

The corresponding probability of a tie is
\[ P(T_i = T_j) = \frac{1}{4} \int_{-\ln \pi_i + \ln \pi_j}^{\ln \pi_i - \ln \pi_j} \operatorname{sech}^2 \frac{y}{2} dy \quad . \]

Davidson (1970) wished to retain the odds ratio related to the Luce axiom of choice in the presence of ties. This can be done if the probabilities listed in (5.1) are respectively proportional to \( \pi_i, \pi_j \) and, say, \( c_{ij} \). The choice of \( c_{ij} \) is open but \( c_{ij} = \sqrt{\pi_i \pi_j} \), \( \nu \geq 0 \), is suggested by the logarithmic locations \( \ln \pi_i \) on the linear scale and a desire to introduce only one new parameter. Davidson made this choice and obtained the probabilities,
\[ P(T_i \rightarrow T_j) = \frac{\pi_i}{(\pi_i + \pi_j + \nu \sqrt{\pi_i \pi_j})} \quad , \]
\[ P(T_j \rightarrow T_i) = \frac{\pi_j}{(\pi_i + \pi_j + \nu \sqrt{\pi_i \pi_j})} \quad , \]
\[ P(T_i = T_j) = \frac{\nu \sqrt{\pi_i \pi_j}}{(\pi_i + \pi_j + \nu \sqrt{\pi_i \pi_j})} \quad . \]

Both extended models, (5.1) and (5.2), have the property that the probability of a tie is highest when \( \pi_i / \pi_j = 1 \) and decreases monotonically with departures from unity, a property with some intuitive appeal.

Both Rao and Kupper and Davidson develop appropriate likelihood ratio tests. Davidson compares the tests of treatment equality in the presence of ties for the two models - he notes that they have the same non-central chi-square distribution in the limit under comparable local alternatives. We conjecture that the two tests are in fact asymptotically equivalent. Beaver (1974) emphasizes that his test statistic is the same under either model, a result of its local properties.
and the fact that \( v = 0 \) - 1 when \( \pi_i = 1/t, \ i = 1, \ldots, t \). Beaver and Rao (1972) show that (5.1) may be derived through the Thompson-Singh approach and consider the same problem for triple comparisons.

Adjustment for ties for the Thurstone model (3.1) has been considered by Glenn and David (1960) and others.

5.2 Triple Comparisons. The basic model for paired comparisons can be extended to triple comparisons in several ways. Bradley and Terry (1952b) proposed the model,

\[
P(T_i \rightarrow T_j \rightarrow T_k) = \frac{\pi_i \pi_j}{(\pi_i + \pi_j + \pi_k)(\pi_j + \pi_k)}
\]

(5.3)

for comparison of \( T_i, T_j \) and \( T_k \) in a triplet, \( i \neq j \neq k, \ i,j,k = 1, \ldots, t \). Pendergrass and Bradley (1960) considered the choices for extension from paired comparisons to triple comparisons in some detail and suggested the model,

\[
P(T_i \rightarrow T_j \rightarrow T_k) = \frac{\pi_i^2 \pi_j}{\pi_i^2 (\pi_j + \pi_k) + \pi_j^2 (\pi_i + \pi_k) + \pi_k^2 (\pi_i + \pi_j)}
\]

(5.4)

The basic methodology for the second model was developed in detail along with necessary asymptotic theory. In very limited comparison in application, it seemed that the two models give very similar results.

Both models, (5.3) and (5.4), have desirable characteristics. For (5.3),

\[
P(T_i \rightarrow T_j) = P(T_i \rightarrow T_j | T_k \rightarrow T_i, T_j) = \frac{\pi_i}{(\pi_i + \pi_j)}; \text{ for model (5.4),}
\]

\[
P(T_i \rightarrow T_j | T_i, T_j \rightarrow T_k) = P(T_i \rightarrow T_j | T_k \rightarrow T_i, T_j) = \frac{\pi_i}{(\pi_i + \pi_j)}. \text{ It can be shown that only two of the several desirable probability equalities such as these can be met by any model. The model (5.4) is reversible in that the parameters}
\]

\( \pi_1, \ldots, \pi_t \) representing "worths" may be associated with parameters \( \rho_1, \ldots, \rho_t \),

\( \pi_i/\pi_j = \rho_j/\rho_i \), representing, say, "liabilities"; thus a symmetry is introduced
into the ordering process. Model (5.3) would result if the ordering is a succession of first choices consistent with Condition (π) of Block and Marschak. Model (5.3) results if the Lehmann model is applied to ranking in triple comparisons and from application of the Luce axiom of choice. An important reason for choice of model (5.4) for development is that sums of ranks for treatments constitute a set of sufficient statistics for the estimation of \( \pi_1, \ldots, \pi_t \) whereas this is not true for (5.3).

Park (1961) has applied the Pendergrass-Bradley procedures to experimental data and compared results of analyses with analyses of comparison experiments using paired instead of triple comparisons. He obtained good model fits and estimator agreement but noted that his treatments were similar, tending to differ only in one dimension, and that modeling may be more difficult with more complex treatments. Beaver and Rao (1972) use the Thompson-Singh approach for triple comparisons and obtain model (5.3) with generalization for possible tied ranks. The modified model is given only in reference to details available in Beaver (1970). Rai (1971b) uses model (5.4) and restricts the triplets considered to those including a control treatment and pairs of treatments selected from another set.

Perhaps the verdict is not available on the appropriate extension of the basic model of paired comparisons to triple comparisons. The answer may depend on the appropriate concept of choice in a given ranking situation. An accepted extended model for a complete ranking of \( t \) treatments should solve the problem considered by Mallows (1957).

5.3 Factorials in Paired Comparisons. It was apparent in early applications of paired comparisons to consumer testing that the treatments were often factorial treatment combinations. An effort was made [Abelson and Bradley (1954)] to
consider factorials; a partial model was formulated for a \( 2 \times 2 \) factorial but extension was dropped because of apparent heavy computational requirements.

Recent experience has reinforced the earlier perception of need and the computation is feasible now. El Helbawy (1974) and the author have developed necessary procedures for \( 2^m \) series factorials in paired comparisons. The model is given here and papers on the methodology are in preparation. Extension to other factorial series can be developed; the management of general factorial notation is the main deterrent.

The notation of the basic model for paired comparisons is modified. We set \( t = 2^m \) and redesignate treatments and treatment parameters as \( T_\alpha \) and \( \pi_\alpha \), \( \alpha = (\alpha_1, \ldots, \alpha_m) \), \( \alpha_i = 0, 1 \), \( i = 1, \ldots, m \), so that the vector subscript \( \alpha \) designates the levels of each of the \( m \) factors, \( \sum_\alpha \alpha = 1 \). Reparameterization in terms of "factorial" parameters is effected through a product rule,

\[
\pi_\alpha = \prod_{i=1}^m \pi_i^{\alpha_i} \prod_{i<j} \pi_i^{\alpha_i \alpha_j} \prod_{i_1<\cdots<i_k} \pi_{\alpha_{i_1} \cdots \alpha_{i_k}} \prod_{\alpha_1 \cdots \alpha_m} (5.5)
\]

for all \( \alpha \). Note that the factorial parameters have superscripts. If logarithms of both sides of (5.5) are equated, analogy with the familiar additive model for factorials in the analysis of variance shows that \( \pi_i^{\alpha_i} \) is associated with the main effect of factor \( i \) at level \( \alpha_i \), \( \pi_{\alpha_i \alpha_j} \) is associated with the two-factor interaction of factors \( i \) and \( j \) at levels \( \alpha_i \) and \( \alpha_j \) and so on.

The model (5.5) replaces \( 2^m - 1 \) functionally independent treatment parameters by \( 3^m - 1 \) new factorial parameters. Additional constraints are needed, as in the analysis of variance, to make the transformation one-to-one. Let \( i(k) = (i_1, \ldots, i_k) \) be a vector of \( k \) ordered distinct integers from the first \( m \) integers and let \( \alpha(i(k)) = (\alpha_{i_1}, \ldots, \alpha_{i_k}) \), \( k = 1, \ldots, m \). One way of defining the constraints is
\[ \sum_{\alpha_1} \pi_{\alpha_1} \prod_{i_1} \pi_{\alpha_1 i_2} \prod_{i_2} \pi_{\alpha_1 i_2 i_3} \prod_{i_2} \pi_{\alpha_1 i_2 i_3 i_4} \cdots \prod_{i_2} \pi_{\alpha_1 i_2 i_3 \cdots i_k} = 1 \]

for all \( i(k) \), all \( i_1 \in i(k) \), \( k = 1, \ldots, m \).

In (5.6), \( i_1, \ldots, i_k \) is the vector of integers \( i_1, \ldots, i_k \) in order of magnitude while \( \alpha_1, \ldots, \alpha_k \) is the similarly ordered vector of \( \alpha \)'s. One interpretation of (5.6) can be given. Suppose that \( \pi_{\alpha[i(k)]} \) is the treatment parameter for a \( 2^k \) factorial treatment combination with factor \( i_s \) at level \( \alpha_s \), \( s = 1, \ldots, k \). Then

\[ \pi_{\alpha[i(k)]} = \sum_{\alpha_1} \cdots \sum_{\alpha_k} \pi_{\alpha_1 \cdots \alpha_k}, \text{ all } i(k), \ k = 0, 1, \ldots, m, \tag{5.7} \]

where \( i_1, \ldots, i_{m-k} \) are the \( m-k \) integers from the first \( m \) integers complementary to those integers in \( i(k) \). When \( k = 0 \), we define \( \pi_{\alpha[i(0)]} = 1 \), and (5.7) yields the original constraint, \( \sum_{\alpha} \pi_{\alpha} = 1 \). The identity of factorial parameters is retained whether they be defined for the \( 2^m \) factorial or for a \( 2^k \) factorial with \( k \) of the \( m \) factors.

Statistical methodology has been developed for estimation of factorial parameters in various situations and for tests of significance of main effects and interactions.

Little else has been done on the use of factorials in paired comparisons. Quenouille and John (1971) have considered \( 2^m \) factorials in paired comparisons when an analysis of variance on scores is used. Springall (1973) considers response surface fitting through replacement of \( \ln \pi_i \) in (3.2) with a linear
function, \( \sum_{k=1}^{t} \beta_{ki} x_{ki} \), \( i = 1, \ldots, t \), and uses design points and chooses values of \( n_{ij} \), \( i < j, i,j = 1, \ldots, t \). His example involves a \( 3^2 \) factorial.

5.4 Multivariate Paired Comparisons. Few experiments are conducted that are not multivariate even though univariate analyses may be used. In consumer home-use tests where two variations of a product are compared, it is customary to provide a questionnaire for preference responses on product attributes as well as on overall quality or acceptability. A multivariate paired comparisons experiment has been conducted.

Davidson and Bradley (1969, 1970) have proposed a model for multivariate paired comparisons, one possible extension of the basic univariate model. Suppose that \( T_i \) and \( T_j \) are compared and that a response vector \( s = (s_1, \ldots, s_p) \), \( s_\alpha = i \) or \( j \), is observed on \( p \) attributes where \( s_\alpha = i \) indicates that \( T_i \) is selected (preferred) on attribute \( \alpha \), \( \alpha = 1, \ldots, p \). The model gives probabilities of responses,

\[
p(s|i,j) = p^{(1)}(s|i,j); (s|i,j),
\]

\[
p^{(1)}(s|i,j) = \prod_{\alpha=1}^{p} \frac{\pi_{\alpha i}}{\pi_{\alpha i} + \pi_{\alpha j}}, \tag{5.8}
\]

\[
h(s|i,j) = 1 + \sum_{\alpha<\beta} \delta(s_\alpha,s_\beta) \rho_{\alpha\beta} \left( \frac{\pi_{\alpha i}}{\pi_{\alpha j}} \right)^{-\frac{1}{2}} \left( \frac{\pi_{\beta i}}{\pi_{\beta j}} \right)^{-\frac{1}{2}} \delta(i,s_\beta),
\]

for all \( s, i < j, i,j = 1, \ldots, t \), where \( \pi_{\alpha i} \) is the treatment parameter for \( T_i \) on attribute \( \alpha \), \( \sum_{i} \pi_{\alpha i} = 1, \alpha = 1, \ldots, p \), \( \rho_{\alpha\beta} \) is the "correlation" parameter for attributes \( \alpha \) and \( \beta \), \( \alpha < \beta \), \( \alpha, \beta = 1, \ldots, p \), and \( \delta(\cdot,\cdot) = 1 \) or \(-1\) as the two arguments agree or disagree. Note that \( \rho = 0 \) implies independence.
of responses on attributes where \( \mathbf{p} \) is the vector of element \( \rho_{\alpha \beta} \). The functions in (5.8) are subject to constraints not given here to ensure that \( p(s|i,j) \) is in fact a probability.

The model (5.8) is about as simple a multivariate model as can be devised as an extension of the univariate model. Limited experience suggests that it fits experimental data. The model can be regarded as a second-order model in the classes of discrete multivariate models developed by Bahadur (1961) [see Davidson and Bradley (1969)]. Bivariate generalizations of the logistic model (3.2) and the Lehmann model (3.4) can be found to generate (5.8) but they lose force in justification because they are not themselves unique generalizations to the multivariate case.

The following example is a summary of one given by Davidson and Bradley (1969). A chocolate pudding test was conducted with \( t = 3 \) puddings and preference responses obtained on \( p = 3 \) attributes, 1, taste, 2, color and 3, texture. Responses \( s, (iii),(j,i,i), \ldots, (jjj) \), were observed with frequencies \( f(s|i,j) \) as shown in Table 5.1.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Pair</th>
<th>(iii)</th>
<th>(jii)</th>
<th>(iji)</th>
<th>(jjj)</th>
<th>(iij)</th>
<th>(jj)</th>
<th>(ij)</th>
<th>(jjj)</th>
<th>No. of responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>8</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>9</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(7.93)</td>
<td>(1.09)</td>
<td>(1.15)</td>
<td>(1.69)</td>
<td>(0.76)</td>
<td>(0.97)</td>
<td>(0.37)</td>
<td>(8.03)</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>6</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>9</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(6.25)</td>
<td>(0.60)</td>
<td>(1.24)</td>
<td>(0.92)</td>
<td>(1.12)</td>
<td>(0.62)</td>
<td>(0.64)</td>
<td>(7.61)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(6.92)</td>
<td>(0.37)</td>
<td>(1.26)</td>
<td>(0.60)</td>
<td>(1.70)</td>
<td>(0.75)</td>
<td>(1.10)</td>
<td>(8.31)</td>
<td></td>
</tr>
</tbody>
</table>
Parameter estimates $p_{ai}$ for $\pi_{ai}$ and $\hat{\rho}_{\alpha \beta}$ for $\rho_{\alpha \beta}$ are given in Table 5.2.

<table>
<thead>
<tr>
<th>Attribute $\alpha$</th>
<th>Pudding parameter estimates $i = 1$</th>
<th>$i = 2$</th>
<th>$i = 3$</th>
<th>Correlations</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha = 1$</td>
<td>.312</td>
<td>.360</td>
<td>.328</td>
<td>$\hat{\rho}_{12} = .675$</td>
</tr>
<tr>
<td>2</td>
<td>.307</td>
<td>.321</td>
<td>.372</td>
<td>$\hat{\rho}_{13} = .654$</td>
</tr>
<tr>
<td>3</td>
<td>.338</td>
<td>.288</td>
<td>.374</td>
<td>$\hat{\rho}_{23} = .588$</td>
</tr>
</tbody>
</table>

Frequencies in parentheses in Table 5.1 are calculated from the parameter estimates in Table 5.2 as substituted back into the model (5.8); the model appears to fit the data well in view of the small cell frequencies. Note that estimates of treatment parameters suggest little difference among the puddings on the three attributes. The data primarily reflect the effects of correlations among the attributes. These impressions are confirmed through tests of significance summarized in Table 5.3. Test 1 has $H_0$: $\pi_{ai} = \frac{1}{3}$, $i = 1, 2, 3$, $\alpha = 1, 2, 3$ versus the alternative $H_a$: $\pi_{ai} \neq \frac{1}{3}$ for some $i, \alpha$. Test 2 has $H_0$: $\rho_{\alpha \beta} = 0$, $\alpha < \beta$, $\alpha, \beta = 1, 2, 3$ versus $H_a$: $\rho_{\alpha \beta} \neq 0$ for some $\alpha < \beta$. Test 3 has $H_a$: $p(s|ij)$ given by (5.6) versus separate general multinomial models for each row of Table 5.1. For each test a likelihood ratio statistic $-2 \ln \lambda$ is used with degrees of freedom indicated in the table.

<table>
<thead>
<tr>
<th>Test</th>
<th>$-2 \ln \lambda$</th>
<th>d.f.</th>
<th>Critical level</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Equal preferences</td>
<td>2.362</td>
<td>6</td>
<td>0.88</td>
</tr>
<tr>
<td>2. Zero correlations</td>
<td>62.665</td>
<td>3</td>
<td>&lt;0.0005</td>
</tr>
<tr>
<td>3. Fit of model</td>
<td>9.135</td>
<td>12</td>
<td>0.69</td>
</tr>
</tbody>
</table>
There is little in the literature on formal analysis of multivariate paired comparisons apart from the cited references above. Gary C. Koch and William D. Johnson presented a paper to the American Statistical Association (Montreal, 1972) entitled "Incomplete Contingency Table Approach to Paired Comparison Experiments" and a manuscript is in preparation. Other work in progress by Koch and co-authors on multivariate categorical data impinges on the problem area. Davidson and Bradley (1971) devised a means of relating attribute responses to responses on overall quality in multivariate paired comparisons with a view to assisting the experimenter in determining how changes in attributes might affect acceptability. Sen and David (1968) develop a model for the bivariate case and give test procedures based on asymptotic theory.

6. OTHER METHODS OF PAIRED COMPARISONS

The major effort in this paper has been to review one method of paired comparisons and its extensions. Other models and methods have been introduced only for comparison when relevant. A monograph on paired comparisons has not been attempted.

We note some key papers on other aspects of paired comparisons but give no details. Kendall and Babington Smith (1940) considered circular triads and a coefficient of concordance. Guttman (1946) developed a method of quantification or scaling similar to discriminant analysis. Scheffé (1952) provided an analysis of variance for scored differences from pairs that allows for order of presentation of the items in a pair. Wei (1952) and Kendall (1955) have proposed an iterative scoring system that takes into account not only direct comparisons but also roundabout comparisons involving other items. Bliss et al (1956) used "rankits" in the analysis of paired comparisons. Mehra (1964) and Puri and
Sen (1969) considered the extension of Wilcoxon's (1945) signed ranks to paired comparisons. Details on some of these procedures are given by David (1963) or Kendall (1970). Additional references are provided by Davidson and Farquhar (1973); the suggested annotation of this bibliography would be most helpful in associating papers with basic techniques.

Harris (1957) revised the Thurstone model (3.1) to allow for order of presentation of items in a pair. Similar modification of our model is a suggested topic of investigation.

7. EPILOGUE

Applied statisticians and biometricians through the years have exhibited a great deal of evangelical spirit and enthusiasm for their discipline and its benefits to the scientific endeavor. Step by step outlines on how to apply statistical methods have been prepared for the user to encourage use. While misuse may have resulted occasionally, few would deny that the effort has been beneficial. Mathematical statisticians and probabilists have been less cognizant of the needs of the practitioners of statistics.

The statistical practitioner needs help if he is to be a scientist and not a shoe clerk as discussed by Bross (1974). The applied statistician must develop stochastic models and innovative statistical methodology for new applications. Very often he must rely on asymptotic theory for approximate methods. Theory must be available to him in usable form.

In 1952, in consideration of asymptotic theory for paired comparisons, the author attempted to use the very significant paper of Wald (1943). The paper was impressive, difficult, and involved gaps in development, some noted by Wald himself. Bradley (1954) limited consideration to balanced sets of paired
comparisons because he could not use the Wald theory otherwise. Bradley and Gart (1962) introduced the notion of "associated populations" and extended some of the asymptotic theory. Davidson and Lever (1970) reworked much of the theory with extension to associated populations and with useable, if not most general, regularity conditions. Only recently has an attempt been made [Stroud (1971, 1972)] to study the Wald paper and reexamine his conditions and rigor. The state of theoretical literature from the point of view of the user is not much improved in 1975.

The development of the mathematical theory of statistics and probability is most important to the discipline. But the future of statistics as a discipline depends on the ability of the applied statistician to cope with the needs of science. Good exposition on the modern theory of statistics and probability, including limit theory, is desperately needed. Can we challenge our more mathematical colleagues to meet this need?
REFERENCES


<table>
<thead>
<tr>
<th><strong>REPORT DOCUMENTATION PAGE</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1. REPORT NUMBER</strong></td>
</tr>
<tr>
<td><strong>2. GOVT ACESSION NO.</strong></td>
</tr>
<tr>
<td><strong>3. RECIPIENT’S CATALOG NUMBER</strong></td>
</tr>
<tr>
<td><strong>4. TITLE (and Subtitle)</strong></td>
</tr>
<tr>
<td><strong>5. TYPE OF REPORT &amp; PERIOD COVERED</strong></td>
</tr>
<tr>
<td><strong>6. PERFORMING ORG. REPORT NUMBER</strong></td>
</tr>
<tr>
<td><strong>7. AUTHOR(s)</strong></td>
</tr>
<tr>
<td><strong>8. CONTRACT OR GRANT NUMBER(s)</strong></td>
</tr>
</tbody>
</table>
| **9. PERFORMING ORGANIZATION NAME & ADDRESS** | Department of Statistics  
Florida State University  
Tallahassee, Florida 32306 |
| **10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS** |             |
| **11. CONTROLLING OFFICE NAME & ADDRESS** | Statistics and Probability Program  
Office of Naval Research  
Arlington, Virginia 22217 |
| **12. REPORT DATE**        | March, 1975 |
| **13. NUMBER OF PAGES**   | 39 |
| **14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)** |             |
| **15. SECURITY CLASSIFICATION (of this Report)** | Unclassified |
| **15a. DECLASSIFICATION/DOWNGRADING SCHEDULE** |             |
| **16. DISTRIBUTION STATEMENT (of this Report)** | Approved for public release; distribution unlimited. |
| **17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)** |             |
| **18. SUPPLEMENTARY NOTES** |             |
| **19. KEY WORDS**          | Paired comparisons  
Nonparametric statistics  
Asymptotic methods |
| **20. ABSTRACT**           | The paper begins with a prologue on science and statistics. It is suggested that more emphasis should be placed on the scientific method in the teaching of science and that statistics is a science and should be taught and practiced as a science. The statistician as a scientist must be concerned with the formulation, modification and verification of stochastic models. |
The main sections of the paper deal with a method of paired comparisons proposed originally by Zermelo and rediscovered by Bradley and Terry, Ford, and others. The basic model is reviewed with emphasis on various approaches that lead to the model. Applications and uses of the model are reviewed along with tests of appropriateness of the model. Extensions of the basic model discussed include adjustment for ties, triple comparisons, use of factorial treatment combinations, and the multivariate case for which a numerical example is included. The paper is not a monograph on paired comparisons and methodology is omitted that does not relate to the basic approach of this paper unless introduced for comparison. Nothing is included on details of experimental design or the design of tournaments.

The paper concludes with an epilogue indicating some need for help for those engaged in the development of statistical methodology from those concerned with the mathematics of statistics and probability.