Friedman-Type Statistics and Consistent Multiple Comparisons for Unbalanced Designs With Missing Data

KNUT M. WITTKOWSKI*

A generalization of the Friedman test using the marginal likelihood principle (Kalbfleisch and Prentice 1973) is presented and its asymptotic power given. It allows use of a variety of score functions and handles ties and unbalanced designs. Some well-known statistics [including the tests of Kruskal and Wallis (1952), Prentice (1979), and Rinaman (1983)] are proven to be special cases, whereas others (e.g., Klotz 1980; Groggel and Skillings 1986; Rai 1987; Skillings and Mack 1981) are shown to be less appropriate. Multiple comparisons are considered under both the global hypothesis and alternatives. Evaluating noncentrality parameters under local shift alternatives, the procedures of Klotz (1980) and (for a special case) Skillings and Mack (1981) can be either anticonservative (showing differences that do not exist) or insensitive (ignoring differences that do exist), depending on the distribution of missing data. A new Scheffé-type procedure for arbitrarily missing data is presented and recommended as consistent and more powerful.

KEY WORDS: Incomplete block design; Local shift alternatives; Missing observation; Score function; Ties.

1. INTRODUCTION

Analysis of variance is a common method for analyzing continuous, interval-scaled data in block designs, but it is not appropriate when the observations are ordinally ranked, the relation between observed data and their importance for the observational unit is not linear, or block effects are not additive. For instance, consider situations where different treatments are ranked with respect to quality of life (see Sec. 7 for an example based on Croog et al. 1986), social status is measured in terms of income, or concentrations of dopamine (DA) and epinephrine (E) are compared in the brain areas RMPO, LMPO, AMBH, and PMBH of rats (Sandmann, Wittkowski, and Wuttke 1981). In the latter experiment (Table 1) five rats (nos. 1, 3, 6, 7, and 10) had more DA in the LMPO than the AMBH and four rats (nos. 4, 5, 8, and 9) had more DA in the AMBH than the LMPO. This ratio of 5/4 indicates a *tendency* for DA in the LMPO to exceed DA in the AMBH. The LMPO-to-AMBH ratio, in average DA concentrations of 4.44/6.67, however, suggests a lower *expectation* of DA in the LMPO than the AMBH. Because block effects are apparently not additive, only the order of the data within the same block can be meaningfully interpreted, differences in expectation are less relevant than differences in tendency. A Friedman-type test statistic, where block ranks are taken as a basis for analysis, is appropriate (Wittkowski, in press).

If a complete or balanced randomized block design is planned, missing observations often occur through data loss. Generalizations of the Friedman test (Friedman 1937) to unbalanced designs were proposed by Benard and Van Elteren (1953). Because the test statistic was not given in an easy-to-use form, earlier attention was restricted to balanced designs until algorithms were derived for special unbalanced designs: Bhapkar and Gore (1973) considered the staircase design together with a special score function. Prentice (1979), Klotz (1980), and Skillings and Mack (1981) considered arbitrary patterns of 0 or 1 observations per cell but proposed different procedures for weighting ranks. Groggel and Skillings (1986) and Rai (1987) proposed further noncompatible procedures for designs with at least one observation per cell.

By applying the marginal likelihood principle (Kalbfleisch and Prentice 1973) to the weighting of arbitrary scored ranks [including all score functions of Sen (1968)], a Friedman-type test statistic is given in Sections 3 and 4. This statistic covers well-known cases for complete, proportional, and balanced designs. The procedures of Groggel and Skillings (1986) and Rai (1987) are indicated as less appropriate for unbalanced designs. The new procedure also allows for testing the global hypothesis of no tendency in treatment effects in layouts with arbitrarily tied and missing observations, including the designs considered by Bhapkar and Gore (1973), Klotz (1980), Skillings and Mack (1981), Groggel and Skillings (1986), and Rai (1987). It avoids anticonservative and insensitive results that could occur if ranks are weighted as proposed by these authors. For the Table 1 example, only the differences in DA are significant at the .01 level, whereas the latter weighting procedures [with the exception of Rai (1987)] reject the hypothesis only for E (see Table 4).

If the global null hypothesis is rejected, it is often of interest to analyze treatment pairs by means of multiple comparisons. For special cases, Klotz (1980) and Skillings and Mack (1981) attempted to generalize Scheffé-type multiple-comparison procedures (Scheffé 1959) to unbalanced designs with missing cells. Evaluating the noncentrality parameter, these procedures (as discussed in Secs. 5 and 6) can produce inconsistent results, depending on

* Knut M. Wittkowski is Biostatistician and Assistant Professor of Biomathematics, Department of Medical Biometry, Eberhard-Karls-University, Westbahnhofstrabe 55, D-7400 Tübingen, Federal Republic of Germany. The author gratefully acknowledges the encouragement of K. Dietz. Sincere thanks also go to the associate editor and referees for very useful suggestions.

Table 1. Concentration of Catecholamines [ng/mg protein]

Brain area	Rat number 1	2	3	4	5	6	7	8	9	10
					Dopamine					
RMPO	13	10	14	15	2	17	9	7	10	—
LMPO	5	3	6	3	2	12	5	2	1	4
AMBH	4	—	4	13	6	11	4	6	9	3
PMBH	8	9	5	—	2	—	7	8	—	4
					Epinephrine					
RMPO	3.8	4.2	4.4	1.3	2.5	3.7	3.3	4.1	4.6	—
LMPO	1.7	1.4	.8	1.7	1.3	2.0	1.8	2.2	4.2	1.1
AMBH	.7	—	.8	2.0	1.3	2.1	1.6	1.7	1.3	1.5
PMBH	2.2	1.6	1.2	—	1.3	—	1.9	2.2	—	1.4

NOTE: A dash indicates missing data due to test-tube breakage.

the distribution of missing cells. For the Table 1 example, the latter procedures lead to anticonservative results for RMPO against AMBH (false significance at the .05 level) and insensitive results for RMPO against LMPO (lack of significance at the .05 level). I introduce consistent Scheffé-type multiple comparisons for arbitrary designs. The proposed procedure is conservative for all strongly unimodal densities (e.g., see Hájek and Šidák 1967) and often more powerful than the earlier procedures.

2. ASSUMPTIONS AND NOTATION

Consider an experimental design with observations $\mathbf{X} = (X_{ijk})$, where $i = 1, \ldots, n$ numbers blocks, $j = 1, \ldots, p$ treatments, and $k = 1, \ldots, m_{ij}$ planned observation replications within cell (ij). If $m_{ij} = 1$ for all i and j, the index k is omitted. The actual random number of observations $M_{ij} \leq m_{ij}$ is assumed independent of the effect of treatment j. Let $M_{ij} = 0$ when cell (ij) contains no observations. Note that $\mathbf{M} = (M_{ij})$ is not necessarily balanced. Subscripts $+$ and $\cdot$ indicate sum and average, respectively (i.e., $M_{i+} = \sum_j M_{ij}$ and $M_{i\cdot} = M_{i+}/p$). The indicator function I is defined by $I(x) = 1$ if x is true, and $I(x) = 0$ otherwise. Let $0/0 = 0$ and $\sum_{i=a}^{b} \cdots = 0$ if $b < a$.

Let the null hypothesis H_0 be defined by permutation invariance of the independent column vectors $\mathbf{X}_i = (X_{i11}, \ldots, X_{i1M(i1)}, \ldots, X_{ip1}, \ldots, X_{ipM(ip)})'$ for $i = 1, \ldots, n$. Note that parentheses are used to avoid indexes of indexes. A subscript 0 (E_0, P_0) indicates that H_0 holds, and a superscript $*$ indicates that the expected value under H_0 is subtracted [i.e., $R^* = R - E_0(R)$]. The H_0 is tested against the alternative hypotheses that treatments $j = 1, \ldots, p$ differ in tendency (e.g., see Wittkowski, in press), measured in terms of $\Pr(X_{ijk} > X_{ij''k''})$ for all i, k, and arbitrary j'' and k''.

All limits are taken as $n \to \infty$. The notation $\xrightarrow{\mathcal{L}}$ indicates convergence in law; $N_p(\boldsymbol{\mu}, \mathbf{V})$, a p-variate normal random variable with expectation $\boldsymbol{\mu}$ and covariance matrix $\mathbf{V}$; and $\chi^2_{p-1}(D)$, a chi-squared random variable with $p - 1$ df and noncentrality parameter D.

3. WEIGHTED SCORES

Let R_{ijk} be the rank of X_{ijk} within block i. Where information on the underlying distribution is available (Sen 1968) or H_0 is defined with respect to certain aspects of the data (e.g., Bhapkar and Gore 1973), the ranks $r = 1, \ldots, M_{i+}$ can be transformed by score functions a_i: $r \to a_i(r)$. Because asymptotic results are calculated for $n \to \infty$, the assumptions for the Chernoff–Savage theorems (Govindarajulu, Le Cam, and Raghavachari 1965) can here be replaced by the mild regularity condition of finite scores $a_i(1), \ldots, a_i(m_{i+})$. This class of score functions includes all cases considered by Hájek and Šidák (1967).

The distribution functions of $\mathbf{X}_i$ may be discontinuous. Introducing order statistics, observations within block i can be grouped into G_i ties of W_{ig} (say) equal observations. As an example, rat $i = 10$ of Table 1 yields $G_i = 2$ ties with $W_{i1} = 1$ and $W_{i2} = 2$. Note that $W_{ig} = 1$ is the untied case. Applying the marginal likelihood principle (Kalbfleisch and Prentice 1973), all $W_{ig}!$ rank permutations within tie g and all m_{i+} possible ranks for a missing observation are equally likely under H_0. Let $Q(\mathbf{X}_i)$ be the set of all possible rank vectors $\mathbf{r} = (r_{i11}, \ldots, r_{ipM(ip)})'$, given the nonmissing observations in block i, and let card $Q(\mathbf{X}_i)$ be the cardinal number of this set. Weighted scores $\bar{a}_i(R_{ijk})$ are defined as the average of all card $Q(\mathbf{X}_i)$ possible scores $a_i(r_{ijk})$. Adjusted (centered–weighted) scores $\bar{a}_i^*(R_{ijk})$ are obtained by subtracting the expected score under H_0 (Hájek and Šidák 1967, p. 121).

The following procedure is illustrated in Table 2 for the previous example (rat $i = 10$, Table 1) using the Wilcoxon (1945) score function $a_i(r) = r$:

$$\bar{a}_i(R_{ijk}) = \frac{1}{\text{card } Q(\mathbf{X}_i)} \sum_{\mathbf{r} \in Q(\mathbf{X}_i)} a_i(r_{ijk}),$$

and

$$E_0(\bar{a}_i(R_{ijk})) = \frac{1}{M_{i+}} \sum_{j,k} \bar{a}_i(R_{ijk}) = \frac{1}{M_{i+}} \sum_r a_i(r).$$

The vector of weighted scores $\hat{a}_i(R_{ijk})$ is the average of all card $Q(\mathbf{X}_i) = 8$ rank vectors consistent with vector $\mathbf{X}_i$ of observations.

This procedure generates average scores (Hájek and Šidák 1967, p. 120–121) within ties and weights scores with respect to missing data. The problem for Wilcoxon scores and special designs has been addressed by several authors. For designs with $M_{ij} \leq 1$, Prentice (1979) suggested the weight of $(M_{i+} + 1)^{-1}$ on centered ranks because of computational convenience; Klotz (1980) did not adjust for missing data; and Skillings and Mack (1981, p. 172) proposed the weight of $((M_{i+} + 1)/12)^{-1/2}$, because this yields a simple covariance structure. For designs with $M_{ij} > 0$, Groggel and Skillings (1986, p. 100) proposed unweighted ranks (although the average rank within a cell is named "weighted sum") because of a simple "null mean." Rai (1987) suggested weights of $M_{i+}{}^{-1}$.

Theorem 1. Let $Q(\mathbf{X})$ be the set of possible rank vectors $\mathbf{r} = (r_j)$, given $\mathbf{X} = (X_j)$, $j = 1, \ldots, p$, and R_j the rank of X_j among the $M \leq p$ nonmissing observations. Then

$$\frac{1}{\text{card } Q(\mathbf{X})} \sum_{\mathbf{r} \in Q(\mathbf{X})} r_j = \frac{p + 1}{M + 1} R_j.$$

See the Appendix for the proof.

Table 2. Computation of Adjusted Scores

Brain area	j	X_{ij}	$Q(\mathbf{X}_i)$								$\bar{a}_i(R_{ij})$	$\bar{a}_i^*(R_{ij})$
RMPO	1	—	4	4	3	3	2	2	1	1	20/8 = 2.500	0/8 = .000
LMPO	2	4	2	3	4	2	4	3	4	3	25/8 = 3.125	5/8 = .625
AMBH	3	3	1	1	1	1	1	1	2	2	10/8 = 1.250	−10/8 = −1.250
PMBH	4	4	3	2	2	4	3	4	3	4	25/8 = 3.125	5/8 = .625

By Theorem 1 all but the weights proposed by Prentice (1979) are inappropriate for unbalanced designs, even for the special case of Wilcoxon scores. They give either too much weight (Klotz 1980; Skillings and Mack 1981) or too little weight (Rai 1987) to blocks with missing data (see Sec. 4 for consequences).

4. THE TEST STATISTIC

The test statistic is based on sums of adjusted scores $\mathbf{T}_{ij}$:

$$\mathbf{T}_i = (T_{i1}, \ldots, T_{ip})'$$

$$T_{ij} = \sum_{k=1}^{M_{ij}} \bar{a}_i^*(R_{ijk}). \tag{4.1}$$

Under H_0, it follows that $E_0(\mathbf{T}_i) = \mathbf{0}$. An estimate of the corresponding covariance matrix $\mathbf{V}_i = E_0(\mathbf{T}_i\mathbf{T}_i')$ is given by

$$\mathbf{V}_i = (V_{ijj'}), \qquad j, j' = 1, \ldots, p$$

$$V_{ijj'} = A_{0,i}^2(I(j = j')M_{ij} - M_{ij}M_{ij'}/M_{i+}), \tag{4.2}$$

where $A_{0,i}^2$ denotes the conditional or unconditional block variance estimator under H_0, respectively:

$$A_{0,i}^2 = (M_{i+} - 1)^{-1} \sum_{r=1}^{M_{i+}} \bar{a}_i^*2(r) \qquad \text{unconditional}$$

$$= (M_{i+} - 1)^{-1} \sum_{j=1}^{p} \sum_{k=1}^{M_{ij}} \bar{a}_i^*2(r_{ijk}) \quad \text{conditional.} \tag{4.3}$$

For Wilcoxon scores and balanced designs (where no weighting for missing data is necessary), the conditional form reduces to the well-known block variance estimator with correction for ties:

$$A_{0,i}^2 = M_{i+}(M_{i+} + 1)/12 \times \left(1 - \sum_{g=1}^{G_i} (W_{ig}^3 - W_{ig})/(M_{i+}^3 - M_{i+})\right).$$

In this case, trivial blocks ($A_{0,i}^2 = 0$) may be observed if all observations are tied within one group. These blocks should be omitted. [See the example of Klotz (1980), where all blocks with missing data are trivial.] In this article I use the unconditional estimator, in accordance with Skillings and Mack (1981).

The quadratic-form test statistic W is then computed with a generalized inverse $\mathbf{V}_\cdot^-$ of the average covariance matrix $\mathbf{V}_\cdot$:

$$W = n\mathbf{T}'\mathbf{V}_\cdot^-\mathbf{T}_\cdot = \mathbf{T}_+'\mathbf{V}_+^-\mathbf{T}_+. \tag{4.4}$$

Under H_0, for small designs W may be compared with its exact (conditional) permutation distribution, or for reasonable designs [e.g., if each treatment pair is comparable in at least one block (see Benard and Van Elteren 1953, eq. 2.3.1 and sec. 2.5)] with its limiting $\chi_{p-1}^2(0)$ distribution. Thus an asymptotically distribution-free test is obtained that is invariant with respect to the particular generalized inverse used. For special designs it reduces to the tests of Friedman (1937), Kruskal and Wallis (1952), Skillings and Mack (1981), and Haux, Schumacher, and Weckesser (1984).

Table 3 gives $\mathbf{T}_\cdot$, $\mathbf{V}_\cdot$, and a generalized inverse $\mathbf{V}_\cdot^-$ for the data of Table 1 and differently weighted Wilcoxon scores. Note that for unconditional tests $\mathbf{V}_\cdot$ depends on the distribution of missing cells but not the distribution of ties. For conditional tests, where ties are reflected in $\mathbf{V}_\cdot$, unweighted scores are equivalent to the procedure of Klotz (1980, p. 667, eq. 2.5a) after replacing his $(1 - U_{i'j0})/(I - U_{ij0} - 1)$ with $1/(I - U_{+j0})$. The test statistics obtained from Table 3 are given in Table 4. Note that inappropriate weighting of blocks may lead to both false negative and false positive decisions. With DA unweighted scores and scores proposed by Skillings and Mack (1981) have poor power, whereas the procedure of Rai (1987) exceeds the level of significance. The converse is true for E. For the special case of $M_{ij} > 0$, Groggel and Skillings (1986) and Rai (1987) proposed procedures that differ not only in the weighting of scores but also in the way the quadratic form is computed. Groggel and Skillings (GS, 1986) used average scores within cell (ij) instead of sums of scores. If T_{ij} in (4.1) is divided by the number of replications $M_{ij}(R_i^* = \Sigma_j R_{ij}/n_{ij}$, GS 1986, p. 100), the elements of the covariance matrix $V_i^{(GS)}$ [p. 100, eq. (2.2)] are

$$V_{ijj'}^{(GS)} = I(M_{ij}M_{ij'} > 0)A_{0,i}^2\left[\frac{I(j = j')}{M_{ij}} - \frac{1}{M_{i+}}\right].$$

For $p = 3$, $n = 2$, $M_{1j} = j$, $M_{2j} = 1$, and $A_{0,i}^2 = M_{i+}(M_{i+} + 1)/12$,

$$12\mathbf{V}_+^{(GS)} = \begin{bmatrix} 35 & -7 & -7 \\ -7 & 14 & -7 \\ -7 & -7 & 7 \end{bmatrix} + \begin{bmatrix} 8 & -4 & -4 \\ -4 & 8 & -4 \\ -4 & -4 & 8 \end{bmatrix} = \begin{bmatrix} 43 & -11 & -11 \\ -11 & 22 & -11 \\ -11 & -11 & 15 \end{bmatrix}.$$

Direct calculation shows that $\text{rank}(\mathbf{V}_+^{(GS)}) = p$. With Rao (1973, p. 188) it follows that the test statistic of Groggel and Skillings (1986) is not $\chi_{p-1}^2(0)$ as pretended (p. 100, theorem 2). The procedure of Rai (1987) is invalid for $p > 3$, because covariances have to be taken into account (e.g., see Wittkowski, in press), so the test statistic must not be of the form $T = (p - 1)/p\ \Sigma_j T_{+j}^2/V_0(T_{+j})$ as pretended (p. 295, eq. 5).

Table 3. Intermediate Results for the Global Test Statistic

Brain area	j	$\mathbf{t}_{\cdot}$ DA	$\mathbf{t}_{\cdot}$ E	$\mathbf{V}_{\cdot}$ RMPO	$\mathbf{V}_{\cdot}$ LMPO	$\mathbf{V}_{\cdot}$ AMBH	$\mathbf{V}_{\cdot}$ PMPH	$\mathbf{V}_{\cdot}^{-}$ RMPO	$\mathbf{V}_{\cdot}^{-}$ LMPO	$\mathbf{V}_{\cdot}^{-}$ AMBH	$\mathbf{V}_{\cdot}^{-}$ PMBH
Unweighted scores											
RMPO	1	.85	.95	.89	−.34	−.31	−.24	1.89	1.05	1.06	—
LMPO	2	−.50	−.55	−.34	.96	−.34	−.28	1.05	1.79	1.05	—
AMBH	3	−.55	−.50	−.31	−.34	.89	−.24	1.06	1.05	1.89	—
PMBH	4	.20	.10	−.24	−.28	−.24	.76	—	—	—	—
Scores weighted as in Skillings and Mack (1981)											
RMPO	1	.90	.97	.96	−.38	−.33	−.25	1.80	1.01	1.02	—
LMPO	2	−.53	−.58	−.38	1.04	−.37	−.29	1.01	1.69	1.01	—
AMBH	3	−.58	−.49	−.33	−.37	.96	−.25	1.02	1.01	1.80	—
PMBH	4	.21	.10	−.25	−.29	−.25	.79	—	—	—	—
Marginal likelihood scores											
RMPO	1	.95	1.00	1.04	−.42	−.36	−.26	1.69	.97	.98	—
LMPO	2	−.56	−.62	−.42	1.15	−.42	−.31	.97	1.58	.97	—
AMBH	3	−.60	−.48	−.36	−.42	1.04	−.26	.98	.97	1.69	—
PMBH	4	.21	.10	−.26	−.31	−.26	.83	—	—	—	—
Scores weighted as in Rai (1987)											
RMPO	1	.98	1.02	1.10	−.45	−.38	−.27	1.63	.94	.96	—
LMPO	2	−.58	−.65	−.45	1.22	−.45	−.32	.94	1.51	.94	—
AMPH	3	−.62	−.47	−.38	−.45	1.10	−.27	.96	.94	1.63	—
PMBH	4	.22	.10	−.27	−.32	−.27	.86	—	—	—	—

NOTE: A dash indicates arbitrary values (e.g., 0).

5. ASYMPTOTIC RESULTS FOR ALTERNATIVE HYPOTHESES

As noted by Prentice (1979), in unbalanced designs the Pitman efficiency may depend heavily on the alternative. I thus omit asymptotic relative efficiency results here. Nevertheless, the asymptotic distribution of test statistics under alternatives needs to be investigated to prove consistency of the test statistics; that is, that the power tends to 1 for fixed alternative. Consider a sequence of location parameter alternatives $H_n(\boldsymbol{\theta})$: $F_{ijk}(x) = F_i(x - n^{-1/2}\theta_j)$, where θ_j denotes the effect of treatment j. I consider the case of untied observations and fixed average cell frequencies $M_{\cdot j}$ The probability distribution F_{ijk} of X_{ijk} is assumed to have a strongly unimodal density f_{ijk}. [See Hájek and Šidák (1967) for examples of such densities.] It is further assumed that all pairs of treatments are comparable in at least one block.

Let $\boldsymbol{\mu} = \lim_{n\to\infty} E(n^{1/2}\mathbf{T}_{\cdot} \mid H_n(\boldsymbol{\theta}))$ and $D(W) = \boldsymbol{\mu}_{\cdot}'\mathbf{V}_{\cdot}^{-}\boldsymbol{\mu}_{\cdot}$. Then, straightforward computations (e.g., see Prentice 1979) yield

$$L(n^{1/2}\mathbf{T}_{\cdot}) \xrightarrow[n\to\infty]{\mathcal{L}} N_p(\boldsymbol{\mu}_{\cdot}'\mathbf{V}_{\cdot})$$

and

$$L(W) \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2_{p-1}(D(W)).$$

Table 4. Global Test Statistics

Weighting procedure	Dopamine W	Dopamine p	Epinephrine W	Epinephrine p
Unweighted scores	10.8	.013	12.0	.008*
Skillings and Mack (1981)	11.1	.012	11.6	.009*
Marginal likelihood scores	11.4	.010*	11.2	.011
Rai (1987)	11.5	.009*	10.9	.013

*$p < .01$.

Before we investigate $D(W)$, I define vectors $\boldsymbol{\theta}^{(i)} = (\theta_1^{(i)}, \ldots, \theta_{M(i+)}^{(i)})'$ associated with blocks $\mathbf{X}_i$. Each nonmissing observation X_{ijk} is replaced by the corresponding treatment effect $\theta^{(i)}_{M(i1)+\cdots+M(i,j-1)+k} = \theta_j$.

The expected adjusted scores vectors $\boldsymbol{\mu}_i$ ($i = 1, \ldots, n$) (with factor $n^{1/2}$) are (see, e.g., Sen 1968)

$$\mu_i = (\mu_{i1}, \ldots, \mu_{ip})',$$

$$\mu_{ij} = n^{1/2} I(M_{ij} > 0)(\theta_j - \theta^{(i)}_{\cdot}) A_i^* B_i,$$

$$A_i^* B_i = M_{i+} \sum_{r=1}^{M_{i+}} \bar{a}_i(r) b_i(r),$$

$$b_i(r) = \beta^{(i)}_{r-2}(M_{i+}) - \beta^{(i)}_r(M_{i+}),$$

and

$$\beta^{(i)}_r(M_{i+}) = I(0 \le r \le M_{i+})\binom{M_{i+}-2}{r} \times \int_{-\infty}^{\infty} F_i^r(x)(1 - F_i(x))^{M_{i+}-2-r} f_i^2(x)\, dx.$$

For fixed cell frequencies ($M_{ij} = m$) this reduces to the results of Rinaman [1983, p. 657, eq. (4.3)], where $dF_i(x)$ should be replaced by $f_i^2(x)\, dx$. Note that $\boldsymbol{\theta} = \mathbf{0}$ implies $D(W) = 0$. The converse is true, at least if the score functions are monotone [$a_i(r) < a_i(r+1)$]. Using theorems of Lehmann (1959, p. 305) and Sen (1968), consistency of the test statistic W follows under local shift alternatives. The global hypotheses can thus be consistently tested by straightforward generalizations of well-known results for complete or balanced designs.

Because the average weighted mean effect $n^{-1}\sum_i \theta^{(i)}_{\cdot}$ is not necessarily 0 for unbalanced designs, it must be taken into consideration when computing the noncentrality parameter $D(W)$ under $(H_n(\boldsymbol{\theta}))_n$. In Section 6 I demonstrate that straightforward generalizations of results for complete designs, for example, as proposed by Klotz (1980) and

Table 5. The Counterexample

Treatment	j	θ_j	M_{ij}	T_{ij}	$T_{\cdot j}$
Control	1	0	1 1 1 1 1 0 0 1 1 1	−.833 ··· −1.250	−.792
1	2	1	0 0 0 0 0 1 0 1 1 1	.000 ··· .000	.042
2	3	1	1 1 1 1 0 0 1 1 1 0	.833 ··· .000	.375
3	4	2	0 0 0 0 1 1 1 0 0 1	.000 ··· 1.250	.375

Skillings and Mack (1981), lead to inconsistent multiple-comparison procedures.

6. MULTIPLE COMPARISONS

If the global null hypothesis H_0 is rejected, it is often of interest to explain rejection by comparing treatment subgroups. Let a subhypothesis $H_0^{jj'}$ for comparison between treatments j and j' $(1 \le j < j' \le p)$ be defined by permutation invariance of the random vectors $\mathbf{X}_i$, with X_{ijk} $(k = 1, \ldots, m_{ij})$ and $X_{ij'k'}$ $(k' = 1, \ldots, m_{ij'})$ fixed, and define a corresponding contrast as $k_{jj'} = (\ldots 0, 1_{(j)}, 0, \ldots, 0, -1_{(j')}, 0, \ldots)'$. From the Cauchy–Schwarz inequality we have for all contrasts $\mathbf{k}$

$$\mathbf{T}_{\cdot}'\mathbf{V}_{\cdot}^{-}\mathbf{T}_{\cdot} \le c^2 \Rightarrow (\mathbf{k}'\mathbf{T}_{\cdot})^2/(\mathbf{k}'\mathbf{V}_{\cdot}\mathbf{k}) \le c^2. \quad (6.1)$$

Thus under H_0 (e.g., see Scheffé 1959)

$$P_0(W \le c_{1-\alpha}^2) = 1 - \alpha$$

$$\Rightarrow P_0\{S_{jj'} \le c_{1-\alpha}^2 \text{ for all } 1 \le j < j' \le p\} \ge 1 - \alpha,$$

where $S_{jj'} = n(\mathbf{k}_{jj'}'\mathbf{T}_{\cdot})^2/(\mathbf{k}_{jj'}'\mathbf{V}_{\cdot}\mathbf{k}_{jj'})$. These results have been shown for special cases by Klotz (1980) and others. Thus it is well known that conventional Scheffé-type procedures are consistent under H_0. Because multiple comparisons are of interest only if the global null hypothesis is rejected, this is not sufficient. Consistency has to be proven under subhypotheses $H_0^{jj'}$ as well. Under the same assumptions as in Section 5, it follows that

$$L(n^{1/2}\mathbf{k}_{jj'}'\mathbf{T}_{\cdot}) \xrightarrow[n\to\infty]{\mathcal{L}} N_1(\mathbf{k}_{jj'}'\boldsymbol{\mu}_{\cdot}, \mathbf{k}_{jj'}'\mathbf{V}_{\cdot}\mathbf{k}_{jj'})$$

and

$$L(S_{jj'}) \xrightarrow[n\to\infty]{\mathcal{L}} \chi_{p-1}^2((\mathbf{k}_{jj'}'\boldsymbol{\mu}_{\cdot})^2/\mathbf{k}_{jj'}'\mathbf{V}_{\cdot}\mathbf{k}_{jj'}).$$

Table 6. Test Statistics for Paired Comparisons

Pair			Standard Scheffé		Corrected Scheffé	
Brain area	j	j'	$S_{jj'}$	p	$W_{jj'}$	p
Dopamine						
RMPO vs. LMPO	1	2	7.6	.055	8.5	.036*
RMPO vs. AMBH	1	3	8.5	.036*	6.5	.089
Epinephrine						
RMPO vs. LMPO	1	2	8.7	.033*	7.7	.052
RMPO vs. AMBH	1	3	7.7	.052	8.4	.038*

*$p < .05$.

Evaluation of this chi-squared distribution's noncentrality parameter shows that conventional Scheffé-type test statistics $S_{jj'}$ are invalid for unbalanced designs with missing cells. The numerator is

$$\mathbf{k}_{jj'}'\boldsymbol{\mu}_{\cdot} = n^{-1/2}\sum_{i=1}^{n} A_i^* B_i((\theta_j - \theta_{j'}) - h_{ijj'})$$

$$\begin{aligned} h_{ijj'} &= 0 && \text{if } M_{ij} > 0,\ M_{ij'} > 0 \\ &= \theta_j - \theta^{(i)} && \text{if } M_{ij} > 0,\ M_{ij'} = 0 \\ &= \theta_j - \theta_{j'} && \text{if } M_{ij} = 0,\ M_{ij'} = 0. \end{aligned} \quad (6.2)$$

Under alternative hypotheses the weighted mean effect $\theta^{(i)}$ is not necessarily 0 in the presence of empty cells. Therefore, if a block contains some observation for treatment j but none for treatment j' $(M_{ij} > 0, M_{ij'} = 0)$, the term $h_{ijj'}$ depends on treatment effects $\theta_{j''}$ where $j'' \notin \{j, j'\}$, that is, on effects of treatments not under consideration for this comparison.

Consider a counterexample (Table 5) with three treatments and a control. The control has no effect $(\theta_1 = 0)$, treatments 1 and 2 are equivalent $(\theta_2 = \theta_3 = 1)$, and treatment 3 is best $(\theta_4 = 2)$. Seven blocks contain data for two treatments and the remaining three blocks contain data for three treatments. (A more realistic example is given at the end of this section.) This design is replicated $n/10$ times. Assume that random errors have no effect on

Table 7. Changes in General Well-Being in the Croog Example

	Captopril					Methyldopa					Propranolol				
Occupation	+1	±0	−1	−2	−	+1	±0	−1	−2	−	+1	±0	−1	−2	−
No premedication															
Professional	7	3	4	1	1	4	1	6	3	1	4	2	5	2	1
Administrative	5	2	3	1	1	5	1	6	3	1	5	2	6	2	2
Clerical	3	1	2		1	2	1	3	2	1	4	1	4	2	1
Blue collar	6	2	4	1	1	4	1	6	3	1	5	2	6	2	2
1–2 premedications															
Professional	21	7	12	4	3	10	3	13	7	2	10	4	11	4	3
Administrative	14	5	9	3	2	12	3	15	8	3	11	4	13	5	4
Clerical	8	3	5	1	1	5	1	7	4	2	8	3	9	3	2
Blue collar	17	6	10	3	3	10	2	13	7	3	10	4	12	4	4
3 or more premedications															
Professional	4	1	2	1	1	2		2	1	1	1	1	2	1	
Administrative	3	1	2	1		2	1	2	1	1	2	1	2	1	1
Clerical	2		1								1		1		1
Blue collar	3	1	2	1							2	1	2	1	1

NOTE: Values are numbers of cases. +1, ±0, −1, −2, and − indicate improvement, no change, worsening, withdrawal, and dropout, respectively.

Table 8. Croog Example: Within-Block Results

Treatment	i	t_i	V_i Captopril	V_i Methyldopa	V_i Propranolol	Premedication, occupation
C		1.15	.62	−.32	−.30	(no premedication)
M	1	−.87	−.32	.60	−.28	
P	⋮	−.28	−.30	−.28	.58	(professional)
C		.04	.02	.00	−.02	(3 or more premedications)
M	12	.00	.00	.00	.00	
P		−.04	−.02	.00	.02	(blue collar)

the ranks, so that $\boldsymbol{\mu}_{\cdot} = n^{1/2}\mathbf{T}_{\cdot}$ for all n, and $D(S_{jj'}) \sim n(\mathbf{k}_{jj'}'\mathbf{T}_{\cdot})$.

By (6.2) the power of $S_{jj'}$ is 0 (irrespective of n) if $\theta_j - \theta_{j'} = \sum_i A_i^* B_i h_{ijj'} / \sum_i A_i^* B_i$. [In the counterexample we have $D(S_{3,4}) = n(.375 - .375)^2 = 0$, although $\theta_3 < \theta_4$.] Under $H_0^{jj'}(\theta_j = \theta_{j'})$ the size of $S_{jj'}$ exceeds the level of significance if $\sum_i A_i^* B_i h_{ijj'} \neq 0$. [In the counterexample we have $D(S_{2,3}) \sim n(.042 - .375) \xrightarrow[n\to\infty]{} \infty$, although $\theta_2 = \theta_3$.] Thus conventional Scheffé-type multiple-comparison procedures may be inconsistent, depending on the distribution of missing cells. They may either have very poor power or yield extremely anticonservative results even if the number of blocks is large. For discrete data, $D(S_{jj'})$ can also be computed for small designs (Rao 1973, p. 182). These results are omitted here because they depend on $a_i(r)$, $F_i(x)$, and the distribution of missing cells.

For unbalanced designs with missing cells, unless blocks with missing cells in position j or j' form (accidentally) a balanced design, subhypotheses $H_0^{jj'}$ can only be tested by a modified Scheffé-type test statistic $W_{jj'}$ (say), which must not include blocks with missing cells in position j or j':

$$\begin{aligned} W_{jj'} &= n(\mathbf{k}_{jj'}'\mathbf{T}_{\cdot}^{jj'})^2/(\mathbf{k}_{jj'}'\mathbf{V}_{\cdot}\mathbf{k}_{jj'}) \\ &= n(T_{\cdot j}^{jj'} - T_{\cdot j'}^{jj'})^2/(V_{jj}^{jj'} - 2V_{jj'}^{jj'} + V_{j'j'}^{jj'}) \\ T_{\cdot}^{jj'} &= n^{-1}\sum_{i=1}^{n} I(M_{ij}M_{ij'} > 0)T_i \\ V_{\cdot}^{jj'} &= n^{-1}\sum_{i=1}^{n} I(M_{ij}M_{ij'} > 0)\mathbf{V}_i. \end{aligned} \tag{6.3}$$

By (4.2), the following representation is equivalent to (6.3):

$$W_{jj'} = \frac{\left[\sum_{i=1}^{n} I(M_{ij}M_{ij'} > 0)(T_{ij} - T_{ij'})\right]^2}{\sum_{i=1}^{n} I(M_{ij}M_{ij'} > 0)(M_{ij}^{-1} + M_{ij'}^{-1})A_{0,i}^2}.$$

Note that all subhypotheses can be tested using the same ranks, scores, variance estimators, and critical values used for testing the global hypothesis, and that no matrix inversion is required.

Blocks with missing cells in one of the treatments to be compared must be excluded, although this might seem inefficient. For the example of Table 2, the low DA AMBH score −1.250 hints that the treatment effect is lower in the AMBH than in LMPO and PMBH. The AMBH score is also less than the expected score, given H_0, of .000 assigned to the missing RMPO datum. This, however, does not hint at different effects in AMBH and RMPO. The difference between the observed AMBH score and the expected score, given H_0, depends on LMPO and PMBH data, which are not informative for comparing AMBH and RMPO. Expected scores should be computed not under H_0 but $H_0^{jj'}$. Because the expected score under $H_0^{1,3}$ equals the AMBH score, the block may be omitted without loss of valid information.

Table 6 gives all results significant at the .30 level for conventional and corrected Scheffé-type comparisons in the example of Table 1. Conventional procedures ignore the difference in DA between RMPO and AMBH at the .05 level, but give a false significant result for RMPO versus LMPO. Errors are reversed with E data.

By (6.1), Scheffé-type comparisons are consonant for balanced designs; that is, the size of $S_{jj'}$ cannot exceed the size of W. Although all comparisons are consonant in Table 6, multiple comparisons in unbalanced designs are consonant only for blocks containing information for the smallest subhypothesis. Where some blocks are excluded, the size of $W_{jj'}$ may exceed the size of W.

7. EXAMPLE

I now demonstrate the proposed procedures, using the data of Croog et al. (1986) on well-being following anti-

Table 9. Croog Example: Intermediate Results

Treatment	n	$t_{\cdot}$	$V_{\cdot}$ Captopril	$V_{\cdot}$ Methyldopa	$V_{\cdot}$ Propranolol	$V_{\cdot}^-$ Captopril	$V_{\cdot}^-$ Methyldopa	$V_{\cdot}^-$ Propranolol	W	p
C		2.13	3.08	−1.60	−1.47					
M	10	−1.73	−1.60	3.01	−1.40					
P		−.40	−1.47	−1.40	2.88					
C		1.78	2.57	−1.34	−1.23	.539	.288	—		
M	12	−1.44	−1.34	2.51	−1.17	.288	.552	—	16.5	.001
P		−.33	−1.23	−1.17	2.40	—	—	—		

NOTE: A dash indicates arbitrary values (e.g., 0).

Table 10. Croog Example: Paired Comparisons

Pair							
Treatment	j	j'	$n^{II'}$	$t^{II'}_{\cdot j'} - t^{II'}_{\cdot j'}$	$v^{II'}_{\cdot II} - 2 \cdot v^{II'}_{\cdot II'} + v^{II'}_{\cdot j'j'}$	$W_{II'}$	p
C vs. M	1	2	10	2.13 + 1.73	3.08 + 2 · 1.60 + 3.01	16.0	.001
C vs. P	1	3	12	1.78 + .33	2.57 + 2 · 1.23 + 2.40	7.2	.027
M vs. P	2	3	10	−1.73 + .40	3.01 + 2 · 1.40 + 2.88	2.0	.370

hypertensive therapy. In addition to "improvement," "no change," and "worsening" (p. 1661, table 4), I consider two more classes. First, patients ("withdrawals") who did not complete the study because of adverse drug reactions (p. 1660, table 2) are considered even worse than those who reported worsening. Second, some patients did not complete the study for reasons that can be assumed independent of the treatment (e.g., loss to follow-up). Because the proportion of "dropouts" differs between treatment groups (6.6%, 7.6%, and 10.5%, p. 1658), they must not be ignored.

Quality-of-life ratings are not comparable among people with different social status or medical history. In the Croog et al. (1986) study both occupation and premedications are correlated with the treatment (p. 1659, table 1). Because distribution of confounding variables is not published, it is assumed proportional in Table 7.

The hypothesis of no difference in tendency is here more appropriate than the hypothesis of no difference in distribution. Thus a rank test should be used instead of a chi-squared test. We have four ties (improvement, no change, worsening, and withdrawal), and the proportion of dropouts differs among treatments and strata. By (4.1), (4.2), and Table 7, we obtain the results of Tables 8–10.

If confounding variables are not proportionally distributed, correlation among treatment, occupation, premedication, dropouts, and withdrawals may lead to more or less significant results. At the .05 level, for instance, Captopril and Propranolol might not be different, whereas Propranolol might be better than Methyldopa.

8. SUMMARY AND CONCLUSIONS

Methods for ordinal data are becoming increasingly important: For instance, surveys are often facilitated when respondents are asked to rank alternatives as opposed to answer direct willingness-to-pay questions. A common problem with such surveys is that answers are comparable only within blocks of similar persons, and that some data are missing at random. The proposed method improves on previous approaches of analyzing ordinal data in designs with unbalanced blocks. It leads to consistent results if data (even all data from a cell) are missing at random. The possible improvement and the necessity of publishing more detailed data are demonstrated with an example based on a recent quality-of-life study (Croog et al. 1986). The methodology presented gives a more unified approach based on rank tests that also facilitates knowledge representation in statistical-expert systems (Wittkowski 1986).

APPENDIX: PROOF OF THEOREM 1

The theorem is obviously true for $M = m$. Let the theorem be true for any given $M \leq m$. I assume without loss of generality that $X_1 < \cdots < X_M$ and the $X_{M+1}, \ldots, X_m$ are missing. Now let X_M also be missing. Because all rank vectors in $Q(X)$ must reflect the order within $X_1, \ldots, X_{M-1}$, there are only card $Q(X) = M$ possible rank vectors

$$\frac{m+1}{M+1}\left(1, \ldots, r-1, r+1, \ldots, M-1, r, \frac{M+1}{2}, \ldots, \frac{M+1}{2}\right)',$$

with an average of

$$\frac{m+1}{M+1}\left(1+\frac{1}{M}, \ldots, M-1+\frac{M-1}{M}, \frac{M+1}{2}, \frac{M+1}{2}, \ldots, \frac{M+1}{2}\right)' = \frac{m+1}{M}\left(1, \ldots, M-1, \frac{M}{2}, \ldots, \frac{M}{2}\right)'.$$

Thus if the theorem is true for M it is by induction also true for $M - 1$. In case of ties, the same result is obtained from the average of all rank permutations within ties.

[*Received January 1983. Revised April 1988.*]

REFERENCES

Benard, A., and Van Elteren, P. H. (1953), "A Generalization of the Method of m Rankings," *Indagationes Mathematicae,* 15, 358–369.

Bhapkar, V. P., and Gore, A. P. (1973), "A Distribution-Free Test for Symmetry in Hierarchical Data," *Journal of Multivariate Analysis,* 3, 483–489.

Croog, S. H., Levine, S., Testa, M. A., Brown, B., Bulpitt, C. H., Jenkins, C. D., Klerman, G. L., and Williams, G. H. (1986), "The Effects of Antihypertensive Therapy on the Quality of Life," *New England Journal of Medicine,* 314, 1657–1664.

Friedman, M. (1937), "The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance," *Journal of the American Statistical Association,* 32, 675–701.

Govindarajulu, Z., Le Cam, L., and Raghavachari, M. (1965), "Generalizations of Chernoff–Savage Theorems on Asymptotic Normality of Nonparametric Test Statistics," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1), Berkeley: University of California Press, pp. 608–638.

Groggel, D. J., and Skillings, J. H. (1986), "Distribution-Free Tests for Main Effects in Multifactor Designs," *The American Statistician,* 40, 99–102.

Hájek, J., and Šidák, Z. (1967), *The Theory of Rank Tests,* New York: Academic Press.

Haux, R., Schumacher, M., and Weckesser, G. (1984), "A Rank-Test for Complete Block Designs," *Biometrical Journal,* 26, 567–582.

Kalbfleisch, J. D., and Prentice, R. L. (1973), "Marginal Likelihoods Based on Cox's Regression and Life Model," *Biometrika,* 60, 267–278.

Klotz, J. (1980), "A Modified Cochran–Friedman Test With Missing Observations and Ordered Categorical Data," *Biometrics,* 36, 665–670.

Kruskal, W. H., and Wallis, W. A. (1952), "Use of Ranks in One-Criterion Variance Analysis," *Journal of the American Statistical Association,* 47, 583–621.

Lehmann, E. L. (1959), *Testing Statistical Hypotheses,* New York: John Wiley.

Prentice, M. J. (1979), "On the Problem of *m* Incomplete Rankings," *Biometrika,* 66, 167–170.

Rai, S. C. (1987), "Rank Analysis of Block Designs Having Different Cell Frequencies," *Biometrical Journal,* 29, 293–298.

Rao, C. R. (1973), *Linear Statistical Inference and Its Applications,* New York: John Wiley.

Rinaman, W. C. (1983), "On Distribution-Free Rank Tests for Two-Way Layouts," *Journal of the American Statistical Association,* 78, 655–659.

Sandmann, R., Wittkowski, K., and Wuttke, W. (1981), "Serum LH Levels and Preoptic Catecholamine Turnover Following Estradiol Implantation in the Preoptic Area (MPO)," *Neuroscience Letters Supplement,* 7, 253.

Scheffé, H. (1959), *The Analysis of Variance,* New York: John Wiley.

Sen, P. K. (1968), "Asymptotically Efficient Tests by the Method of *n* Rankings," *Journal of the Royal Statistical Society,* Ser. B, 30, 312–317.

Skillings, J. H., and Mack, G. A. (1981), "On the Use of a Friedman-Type Statistic in Balanced and Unbalanced Block Designs," *Technometrics,* 23, 171–177.

Wilcoxon, F. (1945), "Individual Comparison by Ranking Methods," *Biometrics,* 1, 80–83.

Wittkowski, K. M. (1986), "Generating and Testing Statistical Hypotheses: Strategies for Knowledge Engineering," in *Expert Systems in Statistics,* ed. R. Haux, Stuttgart, NY: Fischer, pp. 139–154.

——— (in press), "Small Sample Properties of Rank Tests for Incomplete Unbalanced Designs," *Biometrical Journal,* 30.