COMMUN. STATIST.-THEOR. METH., 15(5), 1417-1433 (1986)

COMPARING MEAN RANKS FOR REPEATED MEASURES DATA

Alan Agresti and Jane Pendergast

Department of Statistics
University of Florida
Gainesville, FL 32611

Key Words and Phrases: Friedman test, Hotelling test, ordered categorical data, randomized blocks, rank transform, repeated measures, ridits.

ABSTRACT

Rank tests are considered that compare t treatments in repeated measures designs. A statistic is given that contains as special cases several that have been proposed for this problem, including one that corresponds to the randomized block ANOVA statistic applied to the rank transformed data. Another statistic is proposed, having a null distribution holding under more general conditions, that is the rank transform of the Hotelling statistic for repeated measures. A statistic of this type is also given for data that are ordered categorical rather than fully ranked. Unlike the Friedman statistic, the statistics discussed in this article utilize a single ranking of the entire sample. Power calculations for an underlying normal distribution indicate that the rank transformed ANOVA test can be substantially more powerful than the Friedman test.

0361-0926/86/1505-1417$3.50/0

1. INTRODUCTION

Let $\underset{\sim}{X}_i = (X_{i1},\ldots,X_{it})'$, i=1,...,n, be n independent observations taken from a t-dimensional continuous distribution. For each subject i, we shall regard $\underset{\sim}{X}_i$ as repeated measures, one observation for each of t treatments. This paper considers ways of testing for treatment effects, using rankings of the data.

Denote the distribution function of $\underset{\sim}{X}_i$ by G. In Section 2 we consider tests that are appropriate when the hypothesis of no treatment effects is expressed as the exchangeability condition $G(x_1,\ldots,x_t) = G(x_{i_1},\ldots,x_{i_t})$ for all $\underset{\sim}{x}$ and for any permutation $i_1,\ldots,i_t$) of (1,...,t). The most commonly used rank test for this situation is the Friedman test (see, e.g., Lehmann 1975, p. 263), which compares treatment mean ranks when ranks 1 through t are assigned to the treatments separately for each subject. Alternative tests formulated for this situation by Sen (1967), Koch (1969), Raviv (1978), Lam and Longnecker (1983), and for randomized complete blocks by Iman, Hora, and Conover (1984) compare treatment mean ranks when the entire set of N = tn observations is ranked. For t=2 these tests correspond to making pairwise comparisons of all (X_{i1}, X_{j2}) (as would be done by a Mann-Whitney test for independent samples) rather than sign test-type comparisons of only the natural pairs (X_{i1}, X_{i2}). We shall show that some of these statistics are similar, in the sense that they are special cases of a statistic we derive for this problem.

In Section 3, we present a test statistic that is appropriate when the hypothesis of no treatment effects is more broadly expressed as the marginal homogeneity condition $G_1 \equiv G_2 \equiv \ldots \equiv G_t$, where $G_1,\ldots,G_t$ denote the one-dimensional marginal distributions of G. Like the ones in Section 2, this statistic compares treatment mean ranks when there is a single ranking scheme. However, to obtain the asymptotic null distribution of the statistic, it is only necessary to assume marginal homogeneity rather than to make strong assumptions about the joint distribution (such as a common correlation between pairs of

treatments). The relationship between this rank test and the ones presented in Section 2 is analogous to the relationship for the $\{X_{ij}\}$ between the Hotelling test applied for the comparison of means in a repeated measures context and the more structured (compound-symmetry based) repeated measures ANOVA. See, for instance, Morrison (1976, pp. 141-151). In fact, this statistic is simply the Hotelling statistic applied to the ranks, whereas the statistic proposed by Iman, Hora, and Conover (1984) is the ANOVA statistic applied to the ranks.

In Section 4 we give a statistic for testing marginal homogeneity for an ordered categorical response, for which the data take the form of counts in a t-dimensional cross-classification table. In that case, the test also compares mean ranks based on a single ranking scheme for the combined sample, and it can be expressed in terms of mean ridits for the one-dimensional marginal distributions.

For several distributional forms, Iman, Hora, and Conover (1984) made power comparisons between their statistic and others, including the Friedman statistic and the parametric ANOVA statistic. In Section 5, we make some additional comparisons of size and power for these statistics and for the Hotelling rank transform statistic. These indicate for a multivariate normal model that the Iman, Hora, and Conover rank transformed statistic (i) can be substantially more powerful than the Friedman statistic, and (ii) is fairly robust under an autoregressive structure for the treatment correlations.

2. ANOVA RANK TRANSFORM TEST

Let R_{ia} denote the rank of X_{ia} when it is ranked among the entire set of $\{X_{uv}, 1\le u\le n, 1\le v\le t\}$. Midranks are assigned when ties occur. Suppose that the null hypothesis of "no treatment effects" is interpreted in the strict sense to mean that G satisfies $G(x_1,\ldots,x_t) = G(x_{i_1},\ldots,x_{i_t})$ for all $\underset{\sim}{x}$ and all permutations $(i_1,\ldots,i_t)$ of $(1,\ldots,t)$. When H_0 is true,

$\text{Corr}(R_{ia}, R_{ib})$ is identical for all $a \neq b$ and $\text{Corr}(R_{ia}, R_{jb})$ is identical for all a and b with $i \neq j$. Denote these values by $\rho = \text{Corr}(R_{ia}, R_{ib})$ and $\lambda = \text{Corr}(R_{ia}, R_{jb})$.

Under H_0, R_{ia} is equally likely to be any of the ranks $1,\ldots,tn = N$, so that $E(R_{ia}) = (N+1)/2$ and $\sigma^2 = \text{Var}(R_{ia}) = (N^2-1)/12$. Now $\text{Var}(\sum_i R_{ia}) = [n+n(n-1)\lambda]\sigma^2$, and for $a \neq b$, $\text{Cov}(\sum_i R_{ia}, \sum_j R_{jb}) = [n\rho + n(n-1)\lambda]\sigma^2$. Also, $\text{Var}(\sum\sum R_{ia}) = 0$ implies that $\lambda = (1+(t-1)\rho)/(n-1)t$. Let $\bar{R}_{.a} = \sum_i R_{ia}/n$ for $1 \leq a \leq t$. These treatment means satisfy $\sum \bar{R}_{.a}/t = (N+1)/2$. The covariance matrix of the mean ranks is

$$[\sigma^2(t-1)(1-\rho)/N] \begin{bmatrix} 1 & -1/(t-1) & \cdots & -1/(t-1) \\ -1/(t-1) & 1 & \cdots & -1/(t-1) \\ \vdots & \vdots & & \vdots \end{bmatrix} \qquad (2.1)$$

When the asymptotic distribution of $\bar{R}' = (\bar{R}_{.1},\ldots,\bar{R}_{.t-1})$ is multivariate normal, it follows that

$$T = n \sum_a \{\bar{R}_{.a} - (N+1)/2\}^2/\sigma^2(1-\rho) \qquad (2.2)$$

has an asymptotic chi-squared distribution with df = t-1.

Let $r_{ia} = R_{ia}/(N+1)$ and let $\bar{r}_{.a} = \sum_i r_{ia}/n$. When ρ is replaced by an estimate and σ^2 is replaced by an estimate or by its null value, T becomes a statistic that can be used to detect whether any of $\{E\bar{r}_{.a}, a=1,\ldots,t\}$ differ from 1/2. Conditional on the ranks assigned to each subject, the $(t!)^n$ permutations of ranks are equally likely under H_0. For small samples, an exact test can be based on the permutation distribution of the numerator of T.

Lam and Longnecker (1983) suggested a statistic for the paired data (t=2) case, based on this single ranking scheme. When the sample Spearman correlation of the $\{(X_{i1}, X_{i2}), i=1,\ldots,n\}$ is substituted for ρ and the null value is used for σ^2, T is related to their statistic W_p by $W_p^2 = T(N-1)/N$. Lam and Longnecker showed that their statistic is also asymptotically equivalent to one proposed by Raviv (1978). The Lam and Longnecker statistic is actually the special case of a statistic proposed by Sen (1967) applied with "Wilcoxon scores."

Koch (1969) presented statistics for repeated measures that use aligned ranks. His statistic for this setting (formula (38) for v=1), when applied instead with the regular ranks, equals

$$\bar{W}^* = n^2(t-1) \sum [\bar{R}_{.a} - (N+1)/2]^2 \,/\, \{\sum\sum (R_{ia} - \bar{R}_{i.})^2\} \qquad (2.3)$$

where $\bar{R}_{i.} = \sum_a R_{ia}/t$. A little algebra shows that T reduces to $\hat{W}^*$ if we replace σ^2 by its null value and ρ by $\sum_{a \neq b}\sum r_{ab}/t(t-1)$, with

$$r_{ab} = \{\sum_i R_{ia}R_{ib}/n - ((N+1)/2)^2\}/\sigma^2.$$

Yet another statistic related to T was proposed by Iman, Hora, and Conover (1984) in a rank analysis for randomized complete blocks. Their statistic

$$F_R = \frac{n \sum_a \{\bar{R}_{.a} - (N+1)/2\}^2/(t-1)}{\sum_i \sum_a \{R_{ia} - \bar{R}_{i.} - \bar{R}_{.a} + (N+1)/2\}^2/(t-1)(n-1)} \qquad (2.4)$$

has ranks substituted for the $\{X_{ia}\}$ in the ANOVA statistic for randomized complete blocks. This statistic is also the rank analog of the ANOVA F statistic for a repeated measures design. They gave conditions under which $\bar{R}$ is asymptotically multivariate normal and the null sampling distribution of $(t-1)F_R$ is asymptotically chi-squared with df = t-1 as $n \to \infty$. Their simulations showed that the behavior of F_R is closely approximated by the F distribution with $df_1 = t-1$ and $df_2 = (t-1)(n-1)$. If in (2.2) we replace ρ by the ratio of the average sample covariance

ınd average sample variance

$$\frac{(1/t(t-1))\{\sum\sum_{a\neq b}(R_{ia}-\bar{R}_{.a})(R_{ib}-\bar{R}_{.b})/(n-1)\}}{(1/t)\{\Sigma\Sigma(R_{ia}-\bar{R}_{.a})^2/(n-1)\}} \qquad (2.5)$$

nd replace σ^2 by the average sample variance, then T simplifies o $(t-1)F_R$. In fact, the denominator of the F_R statistic has xact expectation $\sigma^2(1-\rho)$ under H_0.

3. RANK TRANSFORM HOTELLING TEST

In Section 2 the condition of no treatment effects was xpressed as exchangeability in the joint distribution. The symptotic distribution theory required at least a simple orrelation structure for the $\{R_{ia}\}$, so that (2.1) holds. ever, most studies are primarily concerned with whether the arginal distributions have similar locations, even if more tringent assumptions about the joint distribution do not hold. or instance, suppose that the subjects must receive the reatments in a certain time order. Then even if there is no reatment effect in some average sense (such as identical $\{E\bar{r}_{.a}\}$), oservations closer together in time may be more strongly orrelated. Hence, if we mainly require the ability to detect ifferences in marginal location, we might prefer a statistic ose null distribution applies more generally than under compound mmetry-type conditions. In this section we construct a test in ich lack of treatment effects is expressed as the broader pothesis H_0: $G_1 \equiv \ldots \equiv G_t$.

For the $\{X_{ij}\}$, there are alternatives to repeated measures OVA for which one can make comparisons of treatment means thout using such strong null conditions. For instance, the telling test for a single multivariate mean can be used to mpare a vector such as $(\bar{X}_{.1}-\bar{X}_{.2}, \bar{X}_{.2}-\bar{X}_{.3}, \ldots, \bar{X}_{.t-1}-\bar{X}_{.t})$ to the ll value of $\underset{\sim}{0}$. Issues pertaining to the choice between the OVA and multivariate approaches are discussed in detail in Koch et al. (1980) and in Barcikowski and Robey (1984). In particular, suppose that there is no randomization of treatments, such as in longitudinal studies. Then it is difficult to make assumptions about covariance structure, so the multivariate approach may be preferred.

In analogy to the Iman, Hora, and Conover (1984) proposal of a rank transformed version of randomized complete blocks ANOVA, here we consider a version for mean ranks of Hotelling's test. When $G_1 \equiv \ldots \equiv G_t$, it is not necessarily true that $\{\mathrm{Corr}(R_{ia}, R_{ib})\}$ or $\{\mathrm{Corr}(R_{ia}, R_{jb})\}$ are the same for all a and b. Therefore the simple covariance matrix (2.1) for the mean ranks is no longer applicable, and we instead use an estimated covariance matrix S/n, where S has entries

$$s_{ab} = \sum_{i=1}^{n}(R_{ia}-\bar{R}_{.a})(R_{ib}-\bar{R}_{.b})/(n-t+1).$$

Suppose that conditions hold under which $\bar{\underset{\sim}{R}}$ is asymptotically normal as $n \to \infty$. Let $\underset{\sim}{\nu} = E\bar{\underset{\sim}{R}}$ and

$$C = \begin{bmatrix} 1 & -1 & 0 & \ldots. & 0 & 0 \\ 0 & 1 & -1 & \ldots. & 0 & 0 \\ . & . & . & & . & . \\ 0 & 0 & 0 & & 1 & -1 \end{bmatrix} .$$

The null hypothesis H_0: $G_1 \equiv \ldots \equiv G_t$ implies that $C\underset{\sim}{\nu} = \underset{\sim}{0}$ and

$$n(C\bar{\underset{\sim}{R}})'(CSC')^{-1}C\bar{\underset{\sim}{R}} \xrightarrow{d} \chi^2_{t-1}. \qquad (3.1)$$

We use the adaptation of this statistic that corresponds to the rank version of the Hotelling statistic for the $\{X_{ia}\}$, namely

$$F_H = (n/(t-1))(C\bar{\underset{\sim}{R}})'(CSC')^{-1}C\bar{\underset{\sim}{R}} . \qquad (3.2)$$

Following the Iman, Hora and Conover suggestion of treating F_R as

an approximate F statistic with $df_1 = t - 1$ and $df_2 = (t-1)(n-1)$, we recommend treating F_H as an approximate F statistic with $df_1 = t-1$ and $df_2 = n-t+1$. Our simulations involving tail percentage points have indicated that this is a reasonable approximation.

Let $\bar{G} = \sum_a G_a/t$, and let $\underset{\sim}{\mu}' = (\mu_1,\ldots,\mu_t)$ with $\mu_a = E\bar{G}(X_{ia})$. The components of $\underset{\sim}{\mu}$ provide descriptions of the degree to which the $\{G_a\}$ have different locations. The value μ_a gives the probability that a randomly selected observation on treatment a is larger than an independent observation selected at random on one of the t treatments. Under either the marginal homogeneity or exchangeability null hypotheses, $\underset{\sim}{\mu}' = .\underset{\sim}{5}$, and both the rank transformed Hotelling and ANOVA statistics are designed to detect whether at least one $\mu_a \neq .5$.

The rank transformed Hotelling statistic and the rank transformed ANOVA statistic satisfy $F_H = F_R$ when t=2. Under H_0: $G_1 \equiv \ldots \equiv G_t$ with t>2, though, $(t-1)F_R$ may no longer have a chi-square asymptotic distribution, but $(t-1)F_H$ would. Thus, the F_R statistic could be inadequate if we are mainly concerned with departures from this broader null hypothesis. Similarly, under marginal homogeneity the t! possible Friedman rankings for each subject need not be equally likely, so the Friedman test is inappropriate. For instance, when t=2, H_0: $G_1 \equiv G_2$ does not imply that $P(X_{i1} > X_{i2}) = P(X_{i1} < X_{i2})$, so for many joint distributions the Friedman test has probability of rejection converging to one as $n \to \infty$ even though there is marginal homogeneity.

There are many theoretical questions that are suggested by the rank transformed Hotelling statistic. We conjecture that the following results hold quite generally:

1. Under the exchangeability null hypothesis, $F_H - F_R \xrightarrow{p} 0$.

2. The vector $\underset{\sim}{\bar{r}}$ converges in probability to $\underset{\sim}{\mu}$.

3. Asymptotic 100(1-α)% Scheffe-type simultaneous confidence intervals for contrasts of the form $\underset{\sim}{b}'\underset{\sim}{\mu}$ are given by

$$\underset{\sim}{b}'\underset{\sim}{\bar{r}} \pm \sqrt{\underset{\sim}{b}'S\underset{\sim}{b}/n}\; T_{\alpha,t-1,n-t+1}\;/(N+1)$$

where $T^2_{\alpha,t-1,n-t+1} = (n-1)(t-1)F_{\alpha,t-1,n-t+1}/(n-t+1)$. (See Morrison 1976, p.147, for an analogous result for the $\{X_{ia}\}$)

4. The correlation between R_{ia} and R_{ib} converges asymptotically to $\mathrm{Corr}(\bar{G}(X_{ia}), \bar{G}(X_{ib}))$.

5. Under H_0: $G_1 \equiv \ldots \equiv G_t$, the covariance matrix of $\sqrt{n}\,(\underset{\sim}{\bar{r}} - .\underset{\sim}{5})$ converges asymptotically to the matrix with a$\neq$bth element $\rho_s(X_{ia}, X_{ib})/12$, where $\rho_s(.,.)$ denotes the Spearman correlation.

Further research is needed to find conditions under which these conjectures hold and under which $\underset{\sim}{\bar{R}}$ is asymptotically normal and the null distribution given in (3.1) holds.

4. MARGINAL HOMOGENEITY FOR ORDERED CATEGORICAL DATA

Midranks are used in the formulas presented above whenever it is not possible to fully rank the observations. In the extreme (fully discrete) case, the observations are made on a scale consisting of c ordered categories. Then the data on the t repeated measurements are summarized by counts in a c^t cross-classification table, and different formulas are appropriate for testing marginal homogeneity. For $\underset{\sim}{i}' = (i_1,\ldots,i_t)$, let $\{\pi_{\underset{\sim}{i}}\}$ and $\{p_{\underset{\sim}{i}}\}$ denote the population and sample cell proportions. The null hypothesis of marginal homogeneity is expressed as

$$H_0: \pi_{i+\ldots+} = \cdots = \pi_{++\ldots+i},\ i=1,\ldots,c;$$

where $\pi_{1+\ldots+}$ is the proportion in the ith category of the first marginal distribution, $\pi_{+i+\ldots+}$ is the proportion in the ith category of the second marginal distribution, and so forth.

Koch et al. (1977) proposed several statistics for testing this and other hypotheses for categorical repeated measures. They argued that this hypothesis of marginal homogeneity is often of

greater practical relevance than the more structured one of symmetry ($\pi_{\underset{\sim}{i}} = \pi_{\underset{\sim}{j}}$ whenever $\underset{\sim}{j}$ is a permutation of $\underset{\sim}{i}$) in cell proportions. For ordinal classifications, they gave a statistic that detects differences in marginal means, when a fixed set of scores is assigned to the ordered categories. Here, we also formulate a test sensitive to inequalities in means of marginal distributions. The ridit scores for our statistic are data-based rather than pre-assigned, however, and are analogous to the rank scores for the statistics in the previous sections. The ridits are average cumulative probabilities for an average marginal distribution; that is, they are related to midranks calculated on an ordered categorical scale. See Bross (1958) for an introduction to analyses using ridits.

Let $\bar{G}$ denote the distribution function corresponding to the probabilities $\{(\pi_{1+\ldots+} + \ldots + \pi_{+\ldots+1})/t, \ldots, (\pi_{c+\ldots+} + \ldots + \pi_{\ldots+c})/t\}$, and denote its corresponding values by $(\bar{G}(1), \ldots, \bar{G}(c)=1)$. Let $s_i = [\bar{G}(i-1) + \bar{G}(i)]/2$, $i=1,\ldots,c$, where $\bar{G}(0) = 0$. The $\{s_i\}$ are ridit scores for the average marginal distribution. The mean ridit score for margin 1 is $\mu_1 = \sum_{i=1}^{c} s_i \pi_{i+\ldots+}$, and similarly mean ridits $\{\mu_1, \ldots, \mu_t\}$ can be defined for each margin. Analogous mean ridits $\{\bar{r}_1 = \sum_{i=1}^{c} s_i p_{i+\ldots+}, \ldots\}$ apply to the sample. The following properties hold for the mean ridits:

- Marginal homogeneity implies $\mu_1 = \ldots = \mu_t = .5$.
- $\mu_a = \frac{1}{t} \sum_{b=1}^{t} \mu_{a(b)}$, where $\mu_{a(b)}$ is the mean ridit for margin a when the ridits are calculated using the distribution in margin b.
- $\sum_{a=1}^{t} \mu_a/t = .5$.
- Let X_a, $a=1,\ldots,t$ denote independent observations from the t marginal distributions, let $\tau_{ab} = P(X_a > X_b) - P(X_b > X_a)$, and let $\tau_a = \sum_{b=1}^{t} \tau_{ab}/t$. Then $\mu_{a(b)} = (1+\tau_{ab})/2$ and $\mu_a = (1+\tau_a)/2$. In particular, H_0 implies $\tau_1 = \ldots = \tau_t = 0$. Also μ_a takes on its minimum value $\mu_a = 1/2t$ when $\tau_{ab} = -1$ for all $b \neq a$ and it takes on its maximum value $\mu_a = 1 - 1/2t$ when $\tau_{ab} = 1$ for all $b \neq a$. Note that $\mu_a - .5$ has the same sign as τ_a, and that $\tau_a = P(X_a > Z) - P(Z > X_a)$, where Z is an independent observation from $\bar{G}$.
- μ_a can be regarded as an approximation for the probability that a randomly selected observation from the continuous distribution underlying margin a exceeds an independent observation from the continuous distribution underlying $\bar{G}$.

For a random sample of subjects, the asymptotic normality of the vector of sample cell proportions induces an asymptotic normal distribution for the sample version $\bar{\underset{\sim}{r}}$ of $\underset{\sim}{\mu}$, which can be used to test H_0. Let $\underset{\sim}{\pi}$ be the $\{\pi_{ij}\}$ written in column vector form, let $\underset{\sim}{\Sigma} = \mathrm{Diag}(\underset{\sim}{\pi}) - \underset{\sim}{\pi}\underset{\sim}{\pi}'$ where $\mathrm{Diag}(\underset{\sim}{\pi})$ is the diagonal matrix having $\underset{\sim}{\pi}$ on the main diagonal. Let D be the $t \times c^t$ matrix whose ith row has the partial derivatives of μ_i taken with respect to $\underset{\sim}{\pi}'$. Specifically,

$$\partial \mu_i / \partial \pi_{j_1 \ldots j_t} = 1 + s_{j_i} - \sum_{a=1}^{t} s_{j_a}^{(i)}/t,$$

where $s_{j_a}^{(i)}$ denotes the ridit score for category j_a when the i^{th} marginal distribution alone is used for forming the ridits.

By the delta method, the asymptotic covariance matrix of $\sqrt{n}(\bar{\underset{\sim}{r}} - \underset{\sim}{\mu})$ is $D\underset{\sim}{\Sigma}D'$. For the $(t-1)\times t$ matrix of contrasts C, $C(\bar{\underset{\sim}{r}} - \underset{\sim}{\mu})$ is asymptotically normal with covariance $CD\underset{\sim}{\Sigma}D'C'$. The hypothesis of marginal homogeneity can be tested by

$$n(C\bar{\underset{\sim}{r}})'(C\hat{D}\hat{\underset{\sim}{\Sigma}}\hat{D}'C')^{-1}C\bar{\underset{\sim}{r}}, \qquad (4.1)$$

where $\hat{D}$ and $\hat{\underset{\sim}{\Sigma}}$ are D and $\underset{\sim}{\Sigma}$ calculated for the sample proportions. This statistic has an asymptotic null χ^2_{t-1} distribution. It also has the same form as (3.1) for continuous data, and it is a generalization of a statistic presented by

Agresti (1983) for the case t=2.

5. POWER COMPARISONS

The Monte Carlo study conducted by Iman, Hora, and Conover (1984) indicated that their ANOVA rank transform statistic has superior power to the Friedman statistic for many distributional forms. Their simulations did not consider the effect of the correlation between treatments on this power. We wondered whether a sufficiently high correlation might provide more of an advantage to the Friedman approach of using only within-block rankings. We were also curious about how poorly the ANOVA rank transform statistic and the Friedman statistic would perform if there were marginal homogeneity but not exchangeability. In particular, we wanted to see whether the Hotelling-type rank transform statistic would be much superior to these statistics in terms of matching the nominal α-level, since its asymptotic distribution is appropriate even for this broader condition.

To help answer these questions, we conducted some simulations for a model in which $\underset{\sim}{X}_i$ has a multivariate normal distribution with $EX_{ia} = m_a$, correlation ρ_{ab} between X_{ia} and X_{ib}, and unit standard deviations. We conducted 10,000 simulations for $\alpha = .05$ level tests at all combinations of the following:

$t=2$ and $t=5$

$\rho_{ab} = \tau$ and $\rho_{ab} = \tau^{|b-a|}$ for $\tau = .2$ and $\tau = .8$

$n = 10, 30, 50$

$m_{i+1} - m_i = \Delta$, $i=1,\ldots,t-1$ with $\Delta=0$ and $\Delta=.316$ for $t=2$, $\Delta=.086$ for $t=5$

The Δ values were chosen so that the powers would range from about .05 to .95 over the conditions indicated. These simulations were done using the GGNSM random number generator in the IMSL on IBM 3081D and IBM 3033N computers. Approximate powers were recorded for the regular ANOVA F statistic (F), the Friedman statistic (F_F), the ANOVA rank transform (F_R), and the Hotelling rank transform (F_H).

Table I contains the estimated powers for the condition $\{\rho_{ab} = \tau\}$. For all these cases, the true size of the test is very close to (within .01 of) the nominal value of .05 for F, F_R, and F_H. However, for t=2, the Friedman statistic was overly conservative for n=10 and overly liberal at n=50. For this latter case it is clearly much less powerful than the other statistics. The poorness of the null F approximation for the Friedman statistic was also observed by Iman, Hora, and Conover. The Friedman test did not have higher power than the ANOVA rank transform even in what one would expect to be its most favorable case -- t=5 and $\tau=.8$. In fact, the ANOVA rank transform test performed nearly as well as its parametric analog in all these cases. The Hotelling rank transform statistic satisfies $F_H = F_R$ for t=2. Not surprisingly, for t=5 it fares somewhat more poorly than F_R, since it loses degrees of freedom from not exploiting the simple correlation structure. However, even then, it is comparable in power to the Friedman approach, the discrepancy for the case $\tau=.8$ and n=10 perhaps explained by the larger actual α-level for F_F.

Table II contains results for the autoregressive structure $\rho_{ab} = \tau^{|b-a|}$ for t=5. Since the exchangeability condition does not hold in this case, the null distributions of the F, F_R, and F_F statistics may be poorly approximated by the F distribution. For $\tau=.2$ the violation is very weak and the pattern of results is similar to that obtained when all $\rho_{ab} = .2$. When $\tau=.8$ the actual α-levels depart more from the nominal level and it is less appropriate to compare non-null powers. However, each statistic seems to be relatively robust, particularly F_F. Although asymptotically only F_H is guaranteed to have .05 α-level, in this simulation only minor improvement is achieved by using it.

Table I

Approximate* Powers when the Underlying Distribution is t-variate Normal with Means $m_{i+1} = m_i + \Delta$, Unit Variances, and Treatment Correlations $\rho_{ab} = \tau$.

			$\tau = .2$		$\tau = .8$	
t	n	Test	$\Delta = 0$	$\Delta = .3155$	$\Delta = 0$	$\Delta = .3155$
2	10	F	.0489	.1031	.0479	.2905
		$F_R=F_H$	.0481	.1042	.0467	.2553
		F_F	.0182	.0456	.0208	.1349
	30	F	.0510	.2596	.0494	.7519
		$F_R=F_H$	.0520	.2447	.0491	.6954
		F_F	.0408	.1676	.0403	.5533
	50	F	.0485	.4008	.0461	.9369
		$F_R=F_H$	.0483	.3855	.0467	.8998
		F_F	.0664	.3234	.0643	.8296
t	n	Test	$\Delta = 0$	$\Delta = .0855$	$\Delta = 0$	$\Delta = .0855$
5	10	F	.0520	.0997	.0541	.2642
		F_R	.0535	.0985	.0520	.2343
		F_H	.0518	.0805	.0431	.1334
		F_F	.0534	.0919	.0539	.2210
	30	F	.0498	.2162	.0478	.7402
		F_R	.0492	.2038	.0482	.6795
		F_H	.0451	.1893	.0431	.6062
		F_F	.0486	.1825	.0475	.6343
	50	F	.0487	.3583	.0462	.9454
		F_R	.0467	.3389	.0477	.9115
		F_H	.0502	.3163	.0475	.8869
		F_F	.0485	.2891	.0469	.8783

* based on 10,000 replications

NOTE: F = ANOVA F statistic, F_R = Iman et al. rank statistic, F_H = Hotelling-type rank statistic, F_F = Friedman statistic.

Table II

Approximate* Powers when the Underlying Distribution is t-variate Normal with Means $m_{i=1} = m_i + \Delta$, Unit Variances, and Treatment Correlations $\rho_{ab} = \tau^{|b-a|}$.

			$\tau = .2$		$\tau = .8$	
t	n	Test	$\Delta = 0$	$\Delta = .0855$	$\Delta = 0$	$\Delta = .0855$
5	10	F	.0550	.0995	.0758	.1915
		F_R	.0571	.0979	.0745	.1762
		F_H	.0536	.0737	.0448	.0690
		F_F	.0526	.0886	.0640	.1455
	30	F	.0531	.2022	.0736	.4296
		F_R	.0521	.1927	.0702	.4017
		F_H	.0473	.1445	.0473	.1942
		F_F	.0487	.1619	.0588	.3100
	50	F	.0504	.3195	.0716	.6254
		F_R	.0502	.3035	.0705	.5884
		F_H	.0496	.2331	.0465	.3294
		F_F	.0497	.2483	.0565	.4707

* based on 10,000 replications

As in the parametric case, F_H would be expected to be relatively more advantageous than F_R as the correlations become more disparate. For instance, we conducted simulations for t=5 in which $\rho_{12} = \rho_{13} = \rho_{23} = .8$ and all other $\rho_{ab} = 0$ when a≠b. For n=50 the estimated probabilities of rejection for the null case were .087 for F, .084 for F_R, .048 for F_H, and .052 for F_F. For the nonnull case Δ=.086, the powers were .462 for F, .443 for F_R, .626 for F_H, and .544 for F_F. Hence, for this structure, the Hotelling test seems to perform best.

In summary, in our simulations the F_R and F_H statistics

behaved much like their parametric analogs. When t=2, F_R came much closer than the Friedman statistic to matching the nominal α-level, and it appeared to have better power. It maintained a slight power advantage even when the number of treatments or the correlation increased. When the treatment correlations are highly irregular, the F_H statistic has better asymptotic justification than F_R, and for this case it may be desirable to develop adjustments for F_R analogous to those sometimes used in the parametric case. Also, primary interest in marginal effects dictates that F_H is more appropriate than the Friedman statistic, which has asymptotic probability of rejection equal to 1 for some joint distributions that exhibit marginal homogeneity. The results in this paper give further support to the arguments presented by Iman, Hora, and Conover (1984) against unquestioned use of the Friedman test for rank analysis of repeated measures.

ACKNOWLEDGEMENTS

The authors would like to thank Ming Yang for computing assistance, and Dr. William Parr and Dr. Ronald Randles for helpful discussions. Dr. Agresti's research was partially supported by a grant from the National Institutes of Health.

BIBLIOGRAPHY

Agresti, A. (1983) "Testing Marginal Homogeneity for Ordinal Categorical Variables," Biometrics, 39, 505-510.

Barcikowski, R. S. and Robey, R. R. (1984) "Decisions in Single Group Repeated Measures Analysis: Statistical Tests and Three Computer Packages," The American Statistician, 38, 148-150.

Cross, I. D. J. (1958) "How to Use Ridit Analysis," Biometrics, 14, 18-38.

Iman, R. L., Hora, S. C., and Conover, W. J. (1984) "Comparison of Asymptotically Distribution Free Procedures for the Analysis of Complete Blocks," J. Amer. Statist. Assoc., 79, 674-685.

Koch, G. G. (1969) "Some Aspects of the Statistical Analysis of Split Plot Experiments in Completely Randomized Layouts," J. Amer. Statist. Assoc., 64, 485-505.

Koch, G. G., Amara, I. A., Stokes, M. E., and Gillings, D. B. (1980) "Some Views on Parametric and Non-Parametric Analysis for Repeated Measurements and Selected Bibliography," International Statistical Review, 48, 249-265.

Koch, G. G., Landis, J. R., Freeman, J. L., Freeman, D. H., and Lehnen, R. G. (1977) "A General Methodology for the Analysis of Experiments with Repeated Measurement of Categorical Data," Biometrics, 33, 133-158.

Lam, F. C., and Longnecker, M. T. (1983) "A Modified Wilcoxon Rank Sum Test for Paired Data," Biometrika, 70, 510-513.

Lehmann, E. (1975) Nonparametrics: Statistical Methods Based on Ranks, San Francisco: Holden-Day.

Morrison, D. F. (1976) Multivariate Statistical Methods, 2nd ed., New York: McGraw-Hill.

Raviv, A. (1978) "A Non-parametric Test for Comparing Two Non-independent Distributions," J. R. Statist. Soc.,B 40, 253-261.

Sen, P.K. (1967) "Nonparametric Tests for Multivariate Interchangeability. Part I: Problems of Location and Scale in Bivariate Distributions," Sankhya, A29, 351-372.

Received by Editorial Board member September, 1984, Revised February, 1986.

Recommended by Ronald L. Iman, Sandia National Laboratories, Albuquerque, NM.

Refereed by Stephen C. Hora, University of Hawaii at Hilo, Hilo, HI.