The Kruskal-Wallis test can be expressed in an analysis-of-variance-like form. The test can also be expressed in terms of the differences between the observed and expected rank sums. To extend the test to two or more strata, we need the variance-covariance matrices of I − 1 of the rank sums.

There is presently no test available for interaction. The expected values of the rank sums depend on the sample sizes, and these differ from stratum to stratum. The use of weights, to produce a common expected value for each stratum, may provide a test of interaction. This needs further exploration.

REFERENCES

1. ELASHOFF, J.D. and ELASHOFF, M. (1978): Effects of Errors in Statistical Assumptions. In: **International Encyclopedia of Statistics**, Kruskal, W.H. and Tanur, J.M., Eds., Chicago and Stony Brook, NY: Free Press, pp. 229-249.
2. FISZ, M. (1963): **Probability Theory and Mathematical Statistics.** Fisz, Poland (translated by R. Bartaszynski), New York: John Wiley and Sons, ch. 1.6 and 6.2.
3. HEIFETZ, S.B.; MEYERS, R.J.; and KINGMAN, A. (1982): A Comparison of the Anticaries Effectiveness of Daily and Weekly Rinsing with Sodium Caries Prevention – Third-year Results, *Pediatr Dent* 4:300-303.
4. KINGMAN, A. (1979): A Method of Utilizing the Subjects' Initial Caries Experience to Increase Efficiency in Caries Clinical Trials, *Community Dent Oral Epidemiol* 7:87-90.
5. WILCOXON, F. (1945): Individual Comparisons by Ranking Methods, *Biometrics* 1:80-83.
6. MANN, H.B. and WHITNEY, D.R. (1947): On a Test of Whether One of Two Random Variables is Stochastically Larger Than the Other One, *Ann Math Statist* 18:50-60.
7. KRUSKAL, W.H. and WALLIS, W.A. (1952): Use of Ranks in a One-criterion Variance Analysis, *J Amer Statist Assoc* 47:583-621.
8. VARMA, A.O.; FERTIG, J.W.; and CHILTON, N.W. (1979): A Non-parametric Approach to the Comparison of Various Dental Agents in a Stratified Experimental Design, *Pharmacol Therapeut Dent* 4:1-9.

Specific Non-parametric Approaches to Analyzing Caries Clinical Trials: Discussion of Dr. Varma's Presentation

P. B. IMREY

University of Illinois College of Medicine, Urbana, Illinois 61801

J Dent Res 63(Spec Iss):788-790, May, 1984

Dr. Varma merits thanks for an effective discussion, with an interesting example, of the motivation for and application of non-parametric inference to dental clinical trial data. I will supplement his talk with remarks on three areas:

(i) assumptions underlying "standard" non-parametric analyses;
(ii) general approaches to non-parametric analysis of partial association; and
(iii) the meaning of interaction in non-parametric analysis of variance.

Dr. Varma has indicated that violations of the conventional assumptions of normality, equal variances, and, in the analysis of covariance, linearity suggest use of a non-parametric approach. One must take care to avoid that blurring of distinctions between various assumptions which leads to indiscriminate use of non-parametrics as a presumed cure-all for "ill-conditioned" data. Although much work has been done to develop tractable non-parametric procedures for more complex situations, the most desirable properties of the commonly employed methods, such as the Wilcoxon and Kruskal-Wallis tests, depend upon the assumption that all underlying distributions are simple *translations*, or *location shifts*, of a single parent. Equality of variance under the null hypothesis is implied by this assumption. Further, a symmetric parent distribution may be required if non-parametric inferences are desired about a mean. Examples 1-3 illustrate the need for equality of variance, common functional form, and symmetry of parent distribution if the Wilcoxon test is to give valid inferences about a contrast of means. Each of the situations described invalidates the equiprobable permutation model used to generate the usual null distribution of the Wilcoxon statistic.

Example 1. – Consider comparing two Gaussian (normal) distributions with substantially different variances. A Wilcoxon statistic generated by data from such populations will have a null distribution much more peaked about its mean than the usual tabulated referent, because the observations from the more variable group will almost always surround those from the less variable group. The actual level of the test will be well below the nominal level, and, when the true difference in means is small relative to the larger of the within-group standard deviations, the test may have a much lower power than a parametric test.

Example 2. – Consider comparing distributions of different functional forms – for example, a Gaussian with an exponential distribution. In such a situation, the Wilcoxon test is directed at the hypothesis $P(X > Y) = 1/2$, where X and Y are independent random observations from the respective distributions. This hypothesis is not, in general, equivalent to hypotheses equating location parameters of the two populations, *e.g.*, the hypotheses of equal means or of equal medians. When divided by both sample sizes, the Wilcoxon statistic yields an unbiased estimate of $P(X > Y)$. Thus, if Gaussian and exponential forms with identical first and second moments are compared using the Wilcoxon statistic, it may easily be shown that the test will have asymptotically a Type I error rate of 1.0 against the true hypothesis of equal means. Similarly, when Gaussian and exponential forms with identical medians and variances are compared, the Wilcoxon test will have asymptotically a Type I error rate of 1.0 against the hypothesis of equal medians.

Example 3. – Consider the distributions $A = f_{0,\,1/4}$ and $B = f_{1/2,\,3/4}$ from the two-parameter family

$$f_{\delta\epsilon}(x) = \begin{cases} 1-\epsilon \text{ if } x = \delta + 1 \\ \epsilon \text{ if } x = \delta \\ 0 \text{ otherwise} \end{cases}$$

for $-\infty < \delta < \infty$, $0 \leqslant \epsilon \leqslant 1$. A random observation from A has chance 9/16 of exceeding a random observation from B, so the Wilcoxon test will tend to detect differences in medians, even though means, variances, and functional family are common.

These examples emphasize that, although helpful non-parametric procedures are available for a wide variety of situations, the most standard procedures are not as generally useful as one might suppose. Further, general non-parametric inference does not, in principle, provide a solution for the classical Behrens-Fisher problem.

If assumptions underlying the Kruskal-Wallis test are approximately satisfied within strata, it is useful to view Varma's proposed partial association test in a more general context. The test statistic for q strata is based upon expression of the usual Kruskal-Wallis statistic in stratum k as a quadratic form in an (I-1)-vector $\underset{\sim}{D}_k$ of deviations d_i of treatment rank sums from their null hypothesis expectations. The kernel of the quadratic form is the covariance matrix $\underset{\sim}{V}_k$ generated by the null hypothesis-based randomization distribution, conditioning on treatment group sizes. The q-stratum partial association statistic is the quadratic form comparing the deviation vectors summed across strata $(\sum_{k=1}^{q} \underset{\sim}{D}_k)$ with the summed covariance matrices $(\sum_{k=1}^{q} \underset{\sim}{V}_k)$,

$$Q_o = (\Sigma \underset{\sim}{D}_k)' (\Sigma \underset{\sim}{V}_k)^{-1} (\Sigma \underset{\sim}{D}_k).$$

The statistic Q_o tests the average partial association hypothesis $H_o = \underset{\sim}{\mu} = \Sigma \underset{\sim}{\mu}_k = \underset{\sim}{0}$, where $\underset{\sim}{\mu}_k = \underset{\sim}{\mu}_{D_k}$ is the true expectation of the rank sum deviation vector $\underset{\sim}{D}_k$.

Now, H_o is a reasonable hypothesis to test, in the sense that departures from it unexplainable by chance are valid indications of a net treatment effect. However, Kingman (1984) has pointed out at this meeting that, in a stratified situation when interaction may be present, there is no unique test of average or overall effect. Competitors to H_o and Q_o are available corresponding to different sets of weights for the experiences of each treatment in the several strata; the weights may vary from treatment to treatment. Such competitive hypotheses may be specified as

$$H_{o\underset{\sim}{C}} : \underset{\sim}{\mu}_{\underset{\sim}{C}} = \underset{\sim}{C}\underset{\sim}{\mu}_D = \sum_{k=1}^{q} \underset{\sim}{C}_k \underset{\sim}{\mu}_k = \underset{\sim}{0}$$

where $\underset{\sim}{C} = (\underset{\sim}{C}_1, \underset{\sim}{C}_2, \ldots, \underset{\sim}{C}_q)$ is a weight matrix applied to the strung-out expected rank deviation vector

$$\underset{\sim}{\mu}_D = (\underset{\sim}{\mu}_1', \underset{\sim}{\mu}_2', \ldots, \underset{\sim}{\mu}_q')'.$$

The associated test statistic is

$$Q_{\underset{\sim}{C}} = (\underset{\sim}{C}\underset{\sim}{D})' (\underset{\sim}{C}\underset{\sim}{V}_D \underset{\sim}{C}')^{-1} (\underset{\sim}{C}\underset{\sim}{D})$$

where $\underset{\sim}{D} = (\underset{\sim}{D}_1', \underset{\sim}{D}_2', \ldots, \underset{\sim}{D}_k')'$ and $\underset{\sim}{V}_D$ is the block diagonal matrix constructed from the $\underset{\sim}{V}_k$. How does Q_o fare within the general class of test statistics $Q_{\underset{\sim}{C}}$?

It can be argued that Q_o fares poorly whenever treatment-by-stratum cells are not filled proportionately by design. This would include all cases of post-stratification

TABLE
WITHIN-STRATUM AND AVERAGE PARTIAL ASSOCIATION RANDOMIZATION CHI-SQUARED STATISTICS FOR DMFS-INCREMENT AND THREE SCORING SYSTEMS BASED ON RANKS

Treatments	Stratum	DMFS Increment	Marginal Rank Scores (Wilcoxon, Kruskal-Wallis, Benard-vanElteren)	Marginal Ridit Scores (Mack + Skillings, vanElteren)	Combined Ridit Scores
Placebo	1	1.07	0.60	0.60	1.14
vs.	2	0.92	3.25	3.25	3.22
Weekly Rinse	3	3.92	6.02	6.02	5.12
	4	0.84	2.78	2.78	1.17
	Average	5.62	7.54	11.36	10.64
Placebo	1	1.70	0.95	0.95	1.65
vs.	2	2.26	2.20	2.20	2.36
Daily Rinse	3	5.16	9.13	9.13	8.25
	4	0.57	1.53	1.53	0.95
	Average	8.03	6.92	11.24	11.51
Weekly	1	0.28	0.03	0.03	0.05
vs.	2	0.24	0.29	0.29	0.25
Daily Rinse	3	0.05	0.48	0.48	0.46
	4	0.01	0.04	0.04	0.02
	Average	0.21	0.07	0.00	0.01
Placebo	1	2.45	1.06	1.06	2.04
vs.	2	2.33	3.85	3.85	3.93
Weekly Rinse	3	6.49	10.45	10.45	9.05
vs.	4	2.95	4.09	4.09	2.91
Daily Rinse	Average	12.08	9.93	16.09	15.41

adjustment, as well as pre-stratification with unrestricted within-stratum randomization. For, in these circumstances, not only the test statistic, Q_o, but also the hypothesis, H_o, that it tests is based upon a random aspect of the data, the within-stratum distribution of subjects across treatments. The dilemma is the same as that encountered in classical analysis of variance of two-way unbalanced and disproportionate data sets, where main-effects tests address hypotheses depending on the data configuration.

This problem may be simply avoided by constructing test statistics using the approach suggested, but commencing within each stratum with the vector $\underset{\sim}{D}_k$ of mean rank deviations $\bar{d}_i$ for I-1 treatments. If this is done, it is sensible to account for differing total stratum sample sizes by basing an average partial association statistic on the weighted sum of these mean deviation vectors using the stratum sizes as weights. This amounts to choosing

$$\underset{\sim}{C}_k = N_k \, diag\,(n_{k1}^{-1}, n_{k2}^{-1}, \ldots, n_{k,I-1}^{-1}),$$

where n_{ki} is the number of subjects receiving treatment i in stratum k. The result is, for each treatment, an average rank deviation adjusted to the stratum distribution of the entire subject group.

Statistics of the form just described were initially proposed by Cochran (1950) and Mantel and Haenszel (1959) in the special context of 2 x 2 contingency tables, and elaborated more generally by Mantel (1963), Landis *et al.* (1978), Stanish (1978), Landis *et al.* (1979), and Amara (1982). The underlying rationale may, in fact, be applied using arbitrary scoring schemes other than the ranks discussed above, including a variety of scoring schemes derived from rankings and, in appropriate circumstances, from the original data values. The Table gives results of such Cochran-Mantel-Haenszel (CMH) analyses using several scoring schemes. Note that the differences between the rank analyses and those using original data are not substantial. Analyses refer respectively to the actual DMFS increment, the overall within-stratum ranks which underlie the Wilcoxon, Kruskal-Wallis, and Benard-van Elteren (1953) statistics, the overall within-stratum ridits used by Mack and Skillings (1980) and van Elteren (1960), and similar ridit scores derived from the marginal distribution of all strata pooled. The within-stratum ranks produce the Wilcoxon and Kruskal-Wallis statistics used for a single stratum by Varma, and for these data the CMH rank sum statistic is very close to Varma's partial association statistic (since there is virtually equal allocation within each stratum). It should also be recognized that these analyses represent only one of several non-parametric approaches allowing adjustment of treatment comparisons for a concomitant variable, most of which do not involve any linearity assumptions on the covariate relationship. They have been recently reviewed by Koch *et al.* (1982) and extensively studied by Amara (1982), and a SAS macro (GRMM) which implements a general class of randomization analyses for such problems is in the final stages of preparation for release by the University of North Carolina Department of Biostatistics.

What of the issue of interaction? From a certain perspective, it should be realized that interaction is an elusive concept once we move to the use of ranks. Interaction in a continuous data situation is typically defined in the context of a linear model for data on a fixed scale. Such notions of interaction are scale-dependent. A major purpose of using rank procedures is to achieve scale independence, but once we have, concerns about conventional interaction become moot. However, if we define interaction between treatments and strata as variation across strata of some measure of treatment differences, then there is a scale on which interaction may be reasonably expressed using rank data: the "preference probability scale". The quantities involved are $P(T_i > T_{i'})$, the probability that a random individual will respond better to treatment i than to treatment i'. These quantities are estimable by multiples of corresponding Wilcoxon statistics, and the asymptotic covariances of such estimates within strata are easily obtained. The hypothesis of no interaction may be defined as constancy of these probabilities across strata, and appropriate tests derived. This should be pursued, however, only if the preference concept of interaction is of interest in itself. Such circumstances may be infrequent. If the original data scale is a reasonable one for measurement of treatment efficacy, then the extent to which one treatment is superior to another on that scale demands explicit incorporation in any concept of interaction with a claim to practical relevance. In general, presence or absence of interaction on the original response scale, or a transform of it, will have no clear relation to presence or absence of interaction on the preference scale. Rank tests of interaction on a particular location scale pose more complex randomization theory problems than do the tests discussed up to now, and have not seen widespread application. All in all, in extrapolating the distance concepts of conventional parametric modeling to a non-parametric framework, great care is necessary.

REFERENCES

AMARA, I.A. (1982): Strategies for Multivariate Randomization Analyses and Applications to Health Science Data. Ph.D. Dissertation, University of North Carolina Department of Biostatistics.

BENARD, A. and van ELTEREN, P.H. (1953): A Generalization of the Method of m Rankings, *Indacationes Mathematicae* 15:358-369.

COCHRAN, W.G. (1950): Some Methods for Strengthening the Common Chi-square Test, *Biometrics* 10:417-451.

KOCH, G.G.; AMARA, I.A.; DAVIS, G.W.; and GILLINGS, D.B. (1982): A Review of Some Statistical Methods for Covariance Analysis, *Biometrics* 38:563-596.

KINGMAN, A. (1984): Stratification Methods in Caries Clinical Trials, *J Dent Res* 63(Spec Iss):773-777.

LANDIS, J.R.; COOPER, M.M.; KENNEDY, T.; and KOCH, G.G. (1979): A Computer Program for Testing Average Partial Association in Three-way Contingency Tables, *Computer Prog in Biomed* 9:223-246.

LANDIS, J.R.; HEYMAN, E.R.; and KOCH, G.G. (1978): Average Partial Association in Three-way Contingency Tables: A Review and Discussion of Alternative Tests, *Int Statist Rev* 46:237-254.

MACK, G.A. and SKILLINGS, J.H. (1980): A Friedman-type Rank Test for Main Effects in a Two-factor ANOVA, *J Amer Statist Assoc* 75:947-951.

MANTEL, N. (1963): Chi-square Tests with One Degree of Freedom: Extensions of the Mantel-Haenszel Procedure, *J Amer Statist Assoc* 58:690-700.

MANTEL, N. and HAENSZEL, W. (1959): Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease, *J Natl Cancer Inst* 22:719-748.

STANISH, W.M. (1978): Adjustment for Covariables in Categorical Variable Selection and in Multivariate Partial Association Tests. Ph.D. Dissertation, University of North Carolina Department of Biostatistics.

van ELTEREN, P.H. (1960): On the Combination of Independent Two-sample Tests of Wilcoxon, *Bull Int Statist Inst* 37:351-361.