Statistica Sinica 2(1992), 285-296

TESTING FOR AGREEMENT AMONG SEVERAL GROUPS OF RATERS: A CONTINGENCY-TABLE APPROACH

S. M. Sadooghi-Alvandi

Shiraz University

Abstract: A contingency-table approach to the problem of testing for agreement among m groups of raters, each ranking k items, is presented. The approach is based on a natural decomposition of the hypothesis of agreement into a hierarchy of subhypotheses. It is argued that unless the samples are unusually large, only a small number of these subhypotheses can actually be tested. A conditional testing procedure is then recommended, but a flexible unconditional procedure is also presented. Both procedures employ the familiar chi-squared statistic and are illustrated with numerical examples.

Key words and phrases: Conditional analysis, ranking, chi-squared test of fit.

1. Introduction

Suppose that a sample is taken from each of m groups of raters (judges) and each rater independently ranks the same set of k items. How do we test agreement among the m groups? For example, suppose the items are ranked by two groups of raters, male and female; how do we test that male and female raters have a common opinion?

There is a large literature on measures of agreement (concordance), for both within and between groups; recent contributions include Gross (1986) and Fegin and Alvo (1986). Nevertheless, no generally accepted nonparametric procedure for testing agreement among several groups has emerged. (For parametric approaches to the problem, see Pettitt (1982) and Tanner and Young (1985).) A statistic for testing agreement was first proposed by Schucany and Frawley (1973). Adopting a different viewpoint, Hollander and Sethuraman (1978) advocated an alternative test statistic. The two tests were discussed by Kraemer (1981), who argued that they are based on different notions of 'agreement', and that "this divergence of what constitutes the 'relevant' hypothesis leads to an irreconcilable conflict of results". In another review, Snell (1983) remarked that "there is no generally accepted statistic estimating agreement *between* populations".

The purpose of this paper is to present a nonparametric contingency-table approach to the problem. This approach is not based on any particular 'measure of agreement', and does not involve restrictive assumptions. The proposed procedures are very simple, both conceptually and computationally, employing the familiar Pearson chi-squared statistic, and allow a more detailed analysis of the data. As will be seen, the approach is also rather flexible.

As noted by Hollander and Sethuraman (1978), the problem of testing for agreement among m groups may be regarded as that of testing the homogeneity of m multinomial distributions, each with $k!$ categories. The difficulty, as they note, is that since $k!$ is usually large, many of the cells of the corresponding two-way table will be empty, and the usual chi-square tests would not be appropriate. (For a recent warning on the use of chi-squared statistics in sparse tables, see Haberman (1988).) The approach presented in this paper, which is based on the work of Plackett (1975), is aimed at resolving this difficulty. In fact the analysis of this situation, where the samples are small compared with $k!$ but large by usual standards, is the main focus of the paper, and it is to be emphasized that it is in this specific context that the approach will be presented. The proposed test procedures are, however, flexible in that they take into account the sizes of the samples, with the tests essentially reducing to the usual chi-squared test if the samples are very large.

The proposed approach focuses on the preferences of the raters (rather than the rankings of the items), and utilizes a natural decomposition, first used by Plackett (1975), of the preference probabilities. For clarity, the basic idea is introduced in Section 2 in a simpler setting: that of testing the interchangeability of the items for a single group. The decomposition is then used in Section 3 to decompose the hypothesis of agreement into a hierarchy of subhypotheses. Each subhypothesis postulates the homogeneity of m multinomials, with few categories, and can, in principle, be tested by a chi-squared test. But, it will be argued that unless the samples are unusually large, not all the subhypotheses can actually be tested; typically only the 'low order' subhypotheses can adequately be tested. A conditional test procedure is then recommended. But we also present, in Section 4, a simple unconditional procedure. Essentially in this procedure attention is restricted to those subhypotheses which are expected, a priori, to be testable. Finally some aspects of the proposed approach are discussed in Section 5.

2. The Basic Approach

The basic approach of the paper will become clear by first considering a single sample. Suppose k items, labelled $1, \dots, k$ are ranked by each of n raters, assumed to be a random sample from some large population. The opinion of a rater is usually described by his ranking vector $\mathbf{r} = (r_1, \ldots, r_k)$ where r_i is his ranking of the ith item. An alternative, and arguably more natural, way of describing his opinion is by his preference vector $\mathbf{t} = (t_1, \ldots, t_k)$, where t_1 is his first preference (the item most preferred), t_2 is his second preference, etc. Note that, as permutations, $\mathbf{t}$ is the inverse of $\mathbf{r}$. In contrast with most previous methods of analysis, this paper focuses on the preferences of the raters $\mathbf{t}$ — rather than the rankings of the items $\mathbf{r}$. As will be seen, it is this shift of focus which makes the proposed approach rather natural and useful.

Throught the paper $\mathbf{t} = (i, j, l, \ldots)$ will be used generically to denote a vector of preferences (similarly with $\mathbf{r}$). The corresponding random vector, i.e. the preference vector of a typical rater will be denoted by $\mathbf{T}$. Let $p_{\mathbf{t}} = \text{pr}\,(\mathbf{T} = \mathbf{t})$, and let $n_{\mathbf{t}}$ be the number of raters with preference vector $\mathbf{t}$. Then $\{n_{\mathbf{t}}\}$ has a multinomial distribution with index n and parameter $\{p_{\mathbf{t}}\}$, written

$$\{n_{\mathbf{t}}\} \sim M(n, \{p_{\mathbf{t}}\}).$$

In a similar notation for rankings, $\{n_{\mathbf{r}}\} \sim M(n, \{p_{\mathbf{r}}\})$.

In order to give some direction to the discussion, consider the problem of testing the hypothesis of (mutual) interchangeability of the items:

$$H_0 : p_{\mathbf{t}} = \frac{1}{k!}, \quad \text{all } \mathbf{t}, \tag{2.1}$$

or, equivalently, $p_{\mathbf{r}} = 1/k!$, all $\mathbf{r}$. This hypothesis, which states that all the items are essentially similar, may be of independent interest (e.g., in market research), but is introduced here because of its similarity with the hypothesis of agreement. Working with preferences, the expected frequencies under H_0 are $n^*_{\mathbf{t}} = n/k!$. But unless n is very large, say $n \geq 5k!$, these will be too small and the usual chi-squared statistic

$$X^2 = \sum_{\mathbf{t}} \{(n_{\mathbf{t}} - n^*_{\mathbf{t}})^2 / n^*_{\mathbf{t}}\} \tag{2.2}$$

would not be appropriate. (Similarly if we work with $\{n_{\mathbf{r}}\}$.) This statistic is inappropriate basically because there are too many parameters relative to the number of observations, so that reliable inferences can not be made about all the parameters. Essentially the same difficulty arises when testing agreement among several groups.

The main advantage of working with preferences – rather than rankings – is that preferences have a natural ordering which allows a natural decomposition of the preference probabilities:

$$p_{ijl\ldots} = p_i p_{ij} p_{ijl} \cdots, \tag{2.3}$$

where $\{p_i\}$ denote the first-preference probabilities:

$$p_i = \mathrm{pr}\,(T_1 = i);$$

$\{p_{ij}\}$ denote the second-preference probabilities conditional on the first preference:

$$p_{ij} = \mathrm{pr}\,(T_2 = j|T_1 = i),$$

etc. This decomposition was first used by Plackett (1975), as a basis for deriving models with relatively few parameters. We shall, however, regard it merely as a useful reparametrization, with $\{p_i\}$, $\{p_{ij}\}$, ... as the new parameters; this will keep the approach nonparametric. It is worth remarking that this type of decomposition has long been used in the analysis of life tables (Kaplan and Meier (1958); see also Cox (1972, 1975)). In fact this work was motivated by the methodology of life tables (cf. Berry (1979)); this will be reflected in the proposed method of analysis. Note that a Plackett decomposition may also be applied to the ranking probabilities — as in Berry (1979). But in the present context (n small compared with $k!$), it is for the preferences that the decomposition is natural and useful. The main point to note is that changing the labelling of the items results in a different decomposition of the ranking probabilities. Thus, unless the items have a natural ordering, the decomposition would be arbitrary. This problem does not arise with preferences, which do have a natural ordering. The question of ordering is of special importance in the present context because, as will be seen, one can make inferences only about the "low order" parameters. For preferences, the order of the parameters $\{p_i\}$, $\{p_{ij}\}$, ... seems to reflect their importance, and information about the first and second preferences, say, would be very useful. To illustrate the difference: it seems more useful to know what the *first and second preferences* of a rater are than to know how he ranks the *first and second items*, especially so if comparison with another rater is intended (as will be the case in the problem of testing for agreement).

We now consider the question of making inferences about the new parameters $\{p_i\}$, $\{p_{ij}\}$, ... In Plackett (1975), conditional inference was suggested as an alternative to unconditional inference. In the present context, however, there are special features which make a conditional approach not only very attractive but almost the only choice: First note that the new parameters $\{p_i\}$, $\{p_{ij}\}$, ... are 'variation-independent' (or 'unrelated'; see Lehmann (1986), p.546). This feature is crucial to the argument for conditioning: Corresponding to the Plackett decomposition (2.3), let n_i be the number of raters with first preference i, n_{ij} be the number of raters with first preference i and second preference j, etc. Then the data may be regarded as having been generated in $k-1$ stages: in the first stage $\{n_i\} \sim M(n, \{p_i\})$ is observed; in the second stage $\{n_{ij}\} \sim M(n, \{p_{ij}\})$,

$i = 1, \ldots, k$, are observed; etc. More formally, we have the factorization

$$n!\prod_{\mathbf{t}}(p_{\mathbf{t}}^{n_{\mathbf{t}}}/n_{\mathbf{t}}!) = \Big\{n!\prod_{i}(p_i^{n_i}/n_i!)\Big\}\Big\{\prod_{i}\Big[n_i!\prod_{j}(p_{ij}^{n_{ij}}/n_{ij}!)\Big]\Big\}\cdots$$

where $\mathbf{t} = (i, j, \ldots)$ is the vector of preferences. This, coupled with the variation-independence of the parameters, strongly suggests a conditional approach: inferences about $\{p_i\}$ should be based on the distribution $\{n_i\} \sim M(n, \{p_i\})$; for each i inferences about $\{p_{ij}\}$ should be based on the conditional distribution $\{n_{ij}\} \sim M(n_i, \{p_{ij}\})$, etc. The need for a conditional approach becomes more apparent by noting that the observations may provide little or no information about higher order parameters. For example, if $n_{12} = 0$, then clearly there is no information about $\{p_{12l}\}$. This possibility may be effectively ruled out only if n is large compared with $k!$. But we are specifically concerned with the case where n is small compared with $k!$ – although large by usual standards. In this case there will be adequate information only about the low order parameters. Note, also, that it is not possible to determine, a priori, the parameters about which there will be adequate information; this will depend on the data. Clearly, a conditional approach is called for.

To illustrate the conditional approach, consider the hypothesis of interchangeability (2.1). Using the Plackett decomposition (2.3), this hypothesis is decomposed into a hierarchy of subhypotheses:

$$p_i = p_i^* = 1/k,\ p_{ij} = p_{ij}^* = 1/(k-1),\ \ldots .$$

Now the hypothesis $p_i = p_i^*$ may be tested by the chi-squared statistic

$$X_1^2 = \sum_i \{(n_i - np_i^*)^2/(np_i^*)\}$$

(on $k-1$ df). For each i, the hypothesis $p_{ij} = p_{ij}^*$ can be adequately tested if n_i is large, in which case the statistic is

$$X_{2i}^2 = \sum_{j \neq i} \{(n_{ij} - n_i p_{ij}^*)^2/(n_i p_{ij}^*)\}$$

(on $k-2$ df). In the same way, each subhypothesis may be tested by a chi-squared test, provided the index of the corresponding multinomial is sufficiently large for the test to be valid (see Section 5). This essentially amounts to a partitioning of the statistic (2.2). However, unless n is unusually large, only the low order subhypotheses can actually be tested.

Since we are dealing with multiple hypotheses, there remains the question of combining the tests. In view of the hierarchical nature of the subhypotheses, a step-down procedure of testing each subhypothesis in turn seems quite natural and provides a fairly detailed analysis. Because of the interrelationships between the conditional distributions on which the individual tests are based, some theoretical difficulties arise when formally combining the tests (see Section 5). It can be shown that the individual chi-squared statistics are asymptotically independent, so the tests may be regarded as approximately independent. An alternative approach to formally combining the tests is to use the Bonferroni inequality (which does not require independence), to divide the overall level of significance among the individual tests. It may also be convenient to pool the statistics for the 'testable' hypotheses into a single overall statistic, with an approximate chi-squared distribution (cf. the unconditional test statistic of Section 4). This statistic will be denoted by X_*^2.

Example 1. We illustrate the proposed procedures in the simplest case, $k = 3$, by applying them to the data of C. Sutton, which previously have been analyzed by Hollander and Sethuraman (1978) and Pettitt (1982). The data concern preferred companions for leisure-time activities of two groups of females, white (group 1) and black (group 2). There are three categories of companions: 'males' (category 1), 'females' (category 2), and 'both sexes' (category 3). Hollander and Sethuraman (1978) gave the data in terms of 'rankings', which we reproduce in terms of 'preferences' in Table 1. For ease of reference, first and second preferences are also given in Table 2 and Table 3 respectively.

For the first group, the chi-squared statistic for testing $p_1 = p_2 = p_3$ is $X_1^2 = 7.43$ on 2 df, which is significant at the 0.05 level but not at the 0.01 level. Next, we consider second preferences. Since $n_1 = 0$, there is no information about $\{p_{12}, p_{13}\}$. But, since $n_2 = 8$ and $n_3 = 6$, there is some information about the other parameters. Although binomial tests would be more appropriate, we use chi-squared tests. For testing $p_{21} = p_{23}$, $X_{22}^2 = 4.5$, and for testing $p_{31} = p_{32}$, $X_{23}^2 = 6$, each on 1 df, which are significant at the 0.05 level. We may now confidently reject the hypothesis of mutual interchangeability. The overall chi-squared value is $X_*^2 = 17.93$, on 4 df, which has a significance level of less than 0.005. For the second group, $X_1^2 = 15.85$ on 2 df, which has a significance level of less than 0.001. Thus for the second group a test of the first preferences is sufficient for the rejection of the hypothesis of mutual interchangeability.

Further analysis shows that for the first group the data are consistent with the hypothesis of interchangeability of categories 2 and 3, and for the second group the data are consistent with the hypothesis of interchangeability of categories 1 and 2. (The procedure for testing these hypotheses is similar to the procedure for testing mutual interchangeability.) We may conclude that the individuals in the first group are indifferent whether their companions are all-male or of both sexes, whereas the individuals in the second group are indifferent between all-male and all-female companions. These conclusions are only tentative and their confirmation requires further data, but at least they indicate that the two groups have different preference patterns (and, therefore, there is no agreement between the two groups).

3. Testing Agreement

The procedure for testing for agreement among m groups should now be clear. Using the superscript g for the parameters and variables associated with the gth group, $g = 1, \ldots, m$, the hypothesis of 'complete agreement'

$$H_0 : p_{\mathbf{t}}^1 = \cdots = p_{\mathbf{t}}^m \quad \text{all } \mathbf{t}$$

is decomposed into a hierarchy of subhypotheses

$$\begin{aligned} &p_i^1 = \cdots = p_i^m \quad \text{(agreement on the first preference)}, \\ &p_{ij}^1 = \cdots = p_{ij}^m \quad \text{(agreement on the second preference)}, \\ &\ldots . \end{aligned} \tag{3.1}$$

These are then tested by considering the corresponding sequence of two-way tables: For testing $p_i^1 = \ldots = p_i^m$, the statistic is

$$X_1^2 = \sum_g \sum_i \{(n_i^g - m_i^g)^2 / m_i^g\} \tag{3.2}$$

where $m_i^g = (n^g n_i^+)/n^+$ and "+" denotes summation over a superscript. For each i, the hypothesis $p_{ij}^1 = \cdots = p_{ij}^m$ is tested by the usual chi-squared statistic, here denoted by X_{2i}^2, provided n_i^g are sufficiently large for a valid test. Similarly for the other subhypotheses. As with the hypothesis of interchangeability, agreement may be tested either by employing a step-down procedure of testing each subhypothesis in turn, or by combining the chi-squared statistics for the testable subhypotheses into an overall statistic, here denoted by X_*^2. (Again, it can be shown that the individual chi-squared statistics are asymptotically independent.)

Example 2. Consider testing agreement between the two groups of Example 1. The table of first preferences, Table 2, yields $X_1^2 = 11.45$ on 2 df, which is significant at the 0.005 level. Thus, the two groups do not even agree on their first preferences. Nevertheless, let us look at the second preferences, given in Table 3. The first two subtables provide no information concerning group differences, but the third subtable is informative and yields $X_{23}^2 = 3.86$ on 1 df. The overall chi-squared value is $X_*^2 = 15.31$ on 3 df, which has a significance level of less than 0.005. Thus, as anticipated by the more detailed analysis of Section 2, the hypothesis of agreement is rejected.

4. An Unconditional Approach

The main feature of the testing problems considered in this paper is that, because of the large number of the parameters, the data may not provide sufficient information to test all the subhypotheses. Our conditional approach was, essentially, to test those subhypotheses which could actually be tested. In this section we present an unconditional approach which, to a large extent, retains the flexibility of the conditional approach. Basically, the idea is to restrict attention to those subhypotheses which are expected, a priori, to be testable; this is determined by the sample size(s). But the approach may also be presented in more conventional terms.

Again, we illustrate the basic idea by first considering the hypothesis of interchangeability, for $k = 4$ items. First suppose that $n = 20$. Then, clearly, the sample will not produce sufficient information about all the second-preference parameters. We may therefore restrict attention to testing $p_i = 1/4$, using the statistic

$$X_1^2 = \sum_i \{(n_i - 5)^2/5\} \tag{4.1}$$

(on 3 df). Now suppose that $n = 60$. Then we may confidently expect the sample to produce adequate information about all the second-preference parameters – but not about the third-preference parameters. Restricting attention to testing

$$p_i = 1/4,\ p_{ij} = 1/3, \tag{4.2}$$

the test statistic is

$$\sum_i \{(n_i - 15)^2/15\} + \sum\sum_{i \neq j} \{(n_{ij} - n_i/3)^2/(n_i/3)\} \tag{4.3}$$

on 11 df. There is, however, a more conventional and more convenient 'version' of this statistic: The hypothesis (4.2) is equivalent to

$$p_{ij++} = 1/12,\ i \neq j,$$

where $p_{ij++} = \mathrm{pr}\,(T_1 = i, T_2 = j)$. Since $\{n_{ij}\} \sim M(60, \{p_{ij++}\})$, this hypothesis may be tested by the usual chi-squared statistic

$$X_2^2 = \sum\sum_{i \neq j} \{(n_{ij} - 5)^2/5\} \tag{4.4}$$

(on 11 df), which is, essentially, a disguised form of the statistic (4.3).

The unconditional approach to testing interchangeability thus amounts to 'collapsing' the original table of frequencies, $\{n_{\mathbf{t}}\}$, over high order preferences. Restricting attention to the first c preferences, the statistic is

$$X_c^2 = \sum_{t_1,\dots,t_c} \{(n_{t_1,\dots,t_c} - m_c)^2/m_c\} \tag{4.5}$$

where $m_c = n\{(k-c)!/k!\}$; cf. (2.2). The value of c is so chosen that m_c is not too small. For example, for $k = 4$, the following rules ensure that $m_c \geq 5$: If $n \geq 120$, no collapsing is needed, $c = 3$. If $60 \leq n < 120$ collapse on the third preference, $c = 2$. If $n < 60$ collapse on the second preference, $c = 1$. These rules are probably too conservative (see Section 5).

Clearly, the same approach may also be used for testing agreement: depending on the sample sizes, attention is restricted to testing agreement on the low-order preferences. Again this amounts to 'collapsing' the original table of frequencies, $\{n_{\mathbf{t}}^g\}$, over high-order preferences.

Example 3. We again consider testing agreement between the two groups of Example 1. With sample sizes 14 and 13, the data are not expected to produce much information about second preferences. The appropriate 'unconditional' test statistic is X_1^2, given by (3.2). Its observed value is 11.45, on 2 df, which has a significance level of less than 0.005, thus leading to the rejection of the hypothesis.

Compared with the conditional approach of Section 3, the present approach does not utilize all the available information. But it is more convenient, and should be quite adequate for most applications, as it tests the most important components of the hypothesis. (It also avoids the theoretical difficulties which arise when formally combining the individual tests in the conditional approach; see Section 5.)

5. Remarks

For simplicity, we have not been specific about how large the 'sample(s)' (the indices of the (conditional) multinomials) should be for an adequate chi-squared test. This question has been investigated in the literature; and guidelines, usually in terms of the size of the minimal expected cell value, have been given. A well known guideline is that it is safe to apply a chi-squared test if all the expected values exceed 5. But more recent studies, e.g. Larntz (1978) and Fienberg (1979), suggest that this rule tends to be somewhat conservative, and the test would be valid even with a minimal cell value of 1. (For a recent review, see

Lewis, Saunders and Westcott (1984), where a more complicated guideline is proposed.)

An alternative to the Pearson chi-squared statistic, $X^2 = \sum\{(O-E)^2/E\}$, is the likelihood ratio statistic, $G^2 = 2\sum\{O\log(O/E)\}$. As pointed out by a referee, G^2 performs poorly in small samples (Larntz (1978)) and X^2 is to be preferred. It is interesting to note, however, that from a theoretical point of view, a treatment based on G^2 would be somewhat neater, since it leads to exact partition. For instance, the sum of the G^2 statistics for testing the subhypotheses of interchangeability, i.e. the G^2 analogues of X_1^2, $X_{21}^2, \ldots,$ is exactly equal to the G^2 statistic for a complete test, i.e. the G^2 analogue of (2.2). Similarly for testing agreement. This exact partitioning should make the conditional approach even more plausible (and further supports the chi-squared approximation to the distribution of the overall statistic X_*^2). Also, the G^2 statistics corresponding to the two versions of the unconditional test statistics of Section 4, e.g. the G^2 analogues of (4.3) and (4.4), are exactly equal.

The analysis presented in this paper clearly shows that unless the samples are unusually large, the hypothesis of agreement can not be fully tested; the proposed tests are only partial tests. It may therefore be useful to indicate, in terms of 'partial agreement', exactly which subhypotheses are being tested: agreement on the first preference, agreement up to the second preference, ..., and finally complete agreement. For example, the statistic (3.2) tests agreement on the first preference. Thus, in the unconditional approach of Section 4, the m sample analogue of (4.5) tests agreement up to the cth preference.

The usual test of the hypothesis of interchangeability of Section 2 is Friedman's (1937) rank sum test, originally developed for comparing k treatments in an analysis of variance context. Clearly the procedures of Section 2 provide an alternative to Friedman's test in the same context, and may therefore be of independent interest. Pairwise comparisons may also be handled by a similar (conditional) approach.

The conditional approach adopted in this paper seems highly plausible. It should be mentioned, however, that there are some difficulties with a formal justification: Although conditional inferences about the individual parameters (e.g. the conditional tests of the individual subhypotheses) may be formally justified by the Conditionality Principle (Cox and Hinkley (1974), Berger and Wolpert (1984), their 'combination' can not be so justified, because of the interrelationships between the conditional distributions. (The difficulties are essentially similar to those discussed by Ford, Titterington and Wu (1985), in a different context). As noted in Section 3, this raises theoretical questions when formally combining the tests. This is an interesting question for which we do not have a completely satisfactory answer (we suspect that an extension of the Conditionality Principle is involved, but this will not be discussed here). In fact it was because of such questions that the unconditional approach of Section 4 was also presented as an alternative.

Acknowledgements

I am grateful to a referee for his constructive comments and criticisms of an earlier version of the paper, which led to substantial improvements. This work was supported by the Research Council of Shiraz University.

Table 1. Preferred companions for leisure time activities of elderly females (data of C. Sutton)

	Preference configuration					
	(1,2,3)	(1,3,2)	(2,1,3)	(2,3,1)	(3,1,2)	(3,2,1)
Observed frequencies for white females	0	0	1	7	0	6
Observed frequencies for black females	1	1	0	0	5	6

Table 2. First-preference observed frequencies for the data of Table 1.

	First preference			
	1	2	3	Total
Group 1	0	8	6	14
Group 2	2	0	11	13

Table 3. Second-preference observed frequencies for the data of Table 1.

First preference		Second preference		Total
		2	3	
1	Group 1	0	0	0
	Group 2	1	1	2
		1	3	
2	Group 1	1	7	8
	Group 2	0	0	0
		1	2	
3	Group 1	0	6	6
	Group 2	5	6	11

References

Berger, J. O. and Wolpert, R. L. (1984). *The likelihood Principle.* Institute of Mathematical Statistics, Hayward, California.

Berry, D. A. (1979). Detecting trends in arrangements of ordered objects: a likelihood approach. *Scand. J. Statist.* **6**, 169–174.

Cox, D. R. (1972). Regression models and life tables. *J. Roy. Statist. Soc. Ser.B* **34**, 187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika* **62**, 269–276.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics.* Chapman and Hall, London.

Fegin, P. D. and Alvo, M. (1986). Intergroup diversity and concordance for ranking data: an approach via metrics for permutations. *Ann. Statist.* **14**, 691–707.

Fienberg, S. E. (1979). The use of chi-squared statistics for categorical data problems. *J. Roy. Statist. Soc. Ser.B* **41**, 54–64.

Ford, I., Titterington, D. M. and Wu, C. F. J. (1985). Inference and sequential design. *Biometrika* **72**, 545–551.

Friedman, M. (1937). The use of ranks to avoid assumption of normality implicit in the analysis of variance. *J. Amer. Statist. Assoc.* **32**, 675–701.

Gross, S. T. (1986). The kappa coefficient of agreement for multiple observers when the number of subjects is small. *Biometrics* **42**, 883–893.

Haberman, S. J. (1988). A warning on the use of chi-squared statistics with frequency tables with small expected cell counts. *J. Amer. Statist. Assoc.* **83**, 555–560.

Hollander, M. and Sethuraman, J. (1978). Testing for agreement between two groups of judges. *Biometrika* **65**, 403-411.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* **53**, 457–481.

Kraemer, H. C. (1981). Intergroup concordance: definition and estimation. *Biometrika* **68**, 641–646.

Larntz, K. (1978). Small sample comparisons of exact levels for chi-squared goodness-of-fit statistics. *J. Amer. Statist. Assoc.* **73**, 253–263.

Lehmann, E. L. (1986). *Testing Statistical Hypotheses*, 2nd edition. John Wiley, New York.

Lewis, T., Saunders, I. W. and Westcott, M. (1984). The moments of the Pearson chi-squared statistic and the minimum expected value in two-way tables. *Biometrika* **71**, 515–522.

Pettitt, A. N. (1982). Parametric tests for agreement amongst groups of judges. *Biometrika* **69**, 365–375.

Plackett, R. L. (1975). The analysis of permutations. *Appl. Statist.* **24**, 193–202.

Schucany, W. R. and Frawley, W. H. (1973). A rank test for two group concordance. *Psychometrika* **38**, 249–258.

Snell, M. (1983). Recent literature on testing for intergroup concordance. *Appl. Statist.* **32**, 134–140.

Tanner, M. A. and Young, M. A. (1985). Modelling agreement among raters. *J. Amer. Statist. Assoc.* **80**, 175–180.

Department of Mathematics and Statistics, Shiraz University, Shiraz 51454, Iran.

(Received July 1989; accepted June 1991)