Friedman's test with missing observations

Edyta Mrówka and Przemysław Grzegorzewski
Systems Research Institute, Polish Academy of Sciences
Newelska 6, 01-447 Warsaw, Poland
e-mail: mrowka@ibspan.waw.pl, pgrzeg@ibspan.waw.pl

Abstract

Friedman's test is used traditionally in statistics for testing independence between k orderings ($k > 2$). In the paper we suggest how to generalize Friedman's test for situations with missing information or non-comparable outputs.

Keywords: Friedman's test, IF-sets, missing data, ranks, testing independence.

1 Introduction

Consider a single group of n subjects, each of which is observed under k different conditions. Thus we have a random sample of size n from a k-variate population and a set of data could be presented in the form of a two-way layout of k rows and n columns, with one entry X_{ij} in each cell. Suppose that observation in each row $X_{i1}, X_{i2}, \ldots, X_{in}$ are replaced by their ranking within that row. A typical example is a situation with k judges or experts, called observers, each of whom is presented with the same set of n objects to be ranked. Then Friedman's test will provide a nonparametric test of independence of the k variates.

Classical Friedman's test have been constructed for precise and unambiguous observations, even they are orderings only. But in a real live we often meet vague data and ambiguous answers which abound with missing information and hesitance. Such data exhibit uncertainty which has different source than randomness. Thus traditional statistical tools cannot be directly applied there and new tools which admit missing data and hesitance are strongly required.

The paper is organized as follows: In Sec. 2 we recall the classical Friedman's test. Then we mention basic information on IF-sets (Sec. 3) which are later applied for modelling vague orderings, based on IF-sets (Sec. 4). An finally, in Sec. 5, we suggest how to generalize Friedman's test for incomplete two-way layout.

2 Friedman's test – classical approach

Let $X = \{x_1, \ldots, x_n\}$ denote a finite universe of discourse. Suppose that elements (objects) $x_1, \ldots, x_n$ are ordered according to preferences of k observers $A_1, \ldots, A_k$. Then our data could be presented in the form of a two-way layout (or matrix) M with k rows and n columns. Let R_{ij}, $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, n$, denote the ranked observations so that R_{ij} is the rank given by ith observer to the jth object. Then $R_{i1}, R_{i2}, \ldots, R_{in}$ is a permutation of the first n integers while $R_{1j}, R_{2j}, \ldots, R_{kj}$ is a set of rankings given to object number j by successive observers.

We will represent the data in a tabular form as follows:

$$M = \begin{bmatrix} R_{11} & R_{12} & \ldots & R_{1n} \\ R_{21} & R_{22} & \ldots & R_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ R_{k1} & R_{k2} & \ldots & R_{kn} \end{bmatrix}. \qquad (1)$$

Suppose we are interested in testing the null hypothesis that the k variates are independent or - in other words - that there is no association between rankings given by k observers.

Since the testing problem given above is an extension of the paired-sample problem, one possibility for solving it is to consider $\binom{k}{2}$ tests of the null hypothesis of independence to each pair of rankings. Unfortunately, such method of hypothesis testing is statistically undesirable because test are not independent and the overall probability of the type I error is difficult to determine. Thus we need a single test statistic designed to detect overall dependence between samples with a specified significance level.

One may easily see that each row in (1) is a permutation of numbers $1, 2, \ldots, n$. If, e.g., x_j has the same preference relative to all other objects $x_1, \ldots, x_{j-1}, x_{j+1}, \ldots x_n$ in the opinion of each of the k observers, then all ranks in the jth column will be identical. Therefore, the ranks in each column are indicative for the agreement among observers.

Let $R = (R_1, ..., R_n)$ denote observed column totals, where

$$R_j = \sum_{i=1}^{k} R_{ij}, \quad j = 1, ..., n. \tag{2}$$

and let $\bar{R}$ denote the average column total which equals $\frac{k(n+1)}{2}$ for perfect agreement between rankings. Then the sum of squares of deviations between actually observed column total and average column total for perfect agreement is given by:

$$S(R) = \sum_{j=1}^{n} \left[R_j - \bar{R}\right]^2 = \sum_{j=1}^{n} \left[R_j - \frac{k(n+1)}{2}\right]^2. \tag{3}$$

It can be shown that the value of S for any sets of k rankings ranges between zero and $k^2 n \left(n^2 - 1\right)/12$, with the maximum value attained where there is perfect agreement and the minimum value attained when each observer's rankings are assigned completely at random. Therefore, S maybe used to test the null hypothesis H that the rankings are independent (see, e.g., [3]).

If the null hypothesis holds then the ranges assigned to the n objects are completely random for each of the k observers and the $(n!)^k$ assignments are all equally likely. Thus the probability distribution of S is

$$f_S(s) = \frac{u_s}{(n!)^k}, \tag{4}$$

where u_s is the number of those assignments which yield s as the sum of squares of column total deviations. Some methods of generating values of u_s for given n, k can be applied but the calculations are still considerable. Therefore, an approximation to the distribution of (4) is generally used in practice. A linear function of statistic (3) defined as

$$Q = \frac{12S}{kn(n+1)} \tag{5}$$

can be used to define the rejection region for our hypothesis testing problem. It was proved that statistic (5) approaches the chi-square distribution with $n-1$ degrees of freedom as k increases. Numerical comparisons have shown this to be a good approximation as long as $k > 7$ (see [3]). Therefore, we reject the null hypothesis H if

$$Q \geq \chi^2_{n-1,\alpha}.$$

A test based on S or Q is called *Friedman's test.*

It is easily seen that Friedman's test could be used provided all elements are univocally classified by all observers. However, it may happen that one (or more) observer cannot rank all the elements under study (e.g. he is not familiar with all elements). In other situations someone may have problems with specifying his or her preferences. One way-out is then to remove all objects which are not ordered by all of the observers and not to include them into considerations. However, this approach involves always a loss of information. Moreover, it may happen that if the number of ill-classified objects is large, eliminating them would be preclusive of applying Friedman's test. Below we show how to generalize the classical Friedman's test to make it possible to infer about possible association between rankings with missing information or non-comparable outputs. In our approach we describe vagueness in rankings by the generalization of the standard fuzzy sets, suggested by Atanassov [1].

3 IF-sets – basic definitions

Let X denote a universe of discourse. Then a fuzzy set C in X is defined as a set of ordered pairs

$$C = \{\langle x, \mu_C(x)\rangle : x \in X\}, \quad (6)$$

where $\mu_C : X \to [0,1]$ is the membership function of C and $\mu_C(x)$ is the grade of belongingness of x into C (see [8]). Thus automatically the grade of nonbelongingness of x into C is equal to $1 - \mu_C(x)$. However, in real life the linguistic negation not always identifies with logical negation. This situation is very common in natural language processing, computing with words, etc. Therefore Atanassov [1–2] suggested a generalization of classical fuzzy set, called an intuitionistic fuzzy set. The name suggested by Atanassov is slightly misleading, because his sets have nothing in common with intuitionism known from logic. It seems that other name, e.g. incomplete fuzzy sets (which had the same abbreviation), would be even more adequate for the Atanassov sets. Thus finally, in order to avoid terminology problems, we call the Atanassov sets as IF-sets.

An IF-set C in X is given by a set of ordered triples

$$C = \{\langle x, \mu_C(x), \nu_C(x)\rangle : x \in X\}, \quad (7)$$

where $\mu_C, \nu_C : X \to [0,1]$ are functions such that

$$0 \le \mu_C(x) + \nu_C(x) \le 1 \qquad \forall x \in X. \quad (8)$$

For each x the numbers $\mu_C(x)$ and $\nu_C(x)$ represent the degree of membership and degree of nonmembership of the element $x \in X$ to $C \subset X$, respectively.

It is easily seen that an IF-set $\{\langle x, \mu_C(x), 1 - \mu_C(x)\rangle : x \in X\}$ is equivalent to (6), i.e. each fuzzy set is a particular case of the IF-set. We will denote a family of fuzzy sets in X by $FS(X)$, while $IFS(X)$ stands for the family of all IF-sets in X.

For each element $x \in X$ we can compute, so called, the IF-index of x in C defined as follows

$$\pi_C(x) = 1 - \mu_C(x) - \nu_C(x), \quad (9)$$

which quantifies the amount of indeterminacy associated with x_i in C. It is seen immediately that $\pi_C(x) \in [0,1]$ $\forall x \in X$. If $C \in FS(X)$ then $\pi_C(x) = 0$ $\forall x \in X$.

4 IF-sets in modelling rankings

In this section we will suggest how to apply IF-sets in modelling orderings or rankings. A method proposed below seems to be useful especially if not all elements under consideration could be ranked (see [5–7]). In our approach we will attribute an IF-sets to the ordering corresponding to each observer. For simplicity of notation we will identify orderings expressed by observers with the corresponding IF-sets $A_1, \ldots, A_k$. Thus, for each $i = 1, \ldots, k$ let

$$A_i = \{\langle x_j, \mu_{A_i}(x_j), \nu_{A_i}(x_j)\rangle : x_j \in X\} \quad (10)$$

denote an intuitionistic fuzzy subset of the universe of discourse $X = \{x_1, \ldots, x_n\}$, where membership function $\mu_{A_i}(x_j)$ indicates the degree to which x_j is the most preferred element by ith observer, while nonmembership function $\nu_{A_i}(x_j)$ shows the degree to which x_j is the less preferred element by ith observer.

Here a natural question arises: how to determine these membership and nonmembership functions. Let us recall that the only available information are orderings that admit ties and elements that cannot be ranked (i.e. we deal with orderings which are not necessarily linear orderings because there are elements which are non-comparable). Anyway, for each observer one can always specify two functions $w_{A_i}, b_{A_i} : X \to \{0, 1, \ldots, n-1\}$ defined as follows: for each given $x_j \in X$ let $w_{A_i}(x_j)$ denote the number of elements $x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n$ surely worse than x_j, while $b_{A_i}(x_j)$ let be equal to the number of elements $x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n$ surely better than x_j in the ordering corresponding to the preferences expressed by observer A_i. Using functions $w_{A_i}(x_j)$ and $b_{A_i}(x_j)$ we may determine the requested membership and nonmembership functions as follows

$$\mu_{A_i}(x_j) = \frac{w_{A_i}(x_j)}{n-1}, \quad (11)$$

$$\nu_{A_i}(x_j) = \frac{b_{A_i}(x_j)}{n-1}. \quad (12)$$

It is easily seen that $w_{A_i}(x_j)$, $b_{A_i}(x_j) \in \{0, \ldots, n-1\}$ because we rank n elements and hence for each element $x_j \in X$ there exist no less than zero and no more than $n-1$ elements which are better (worse) than x_j. Moreover, we admit situations when the same rank is assigned to more than one element and elements that are not comparable with the others.

In such a way we get k well defined IF-sets which describe nicely orderings corresponding to k observers. It is seen that $\pi_{A_i}(x_j) = 0$ for each $x_j \in X$ if and only if all elements are ranked by ith observer and there are no ties. Conversely, if there exist such element $x_j \in X$ that $\pi_{A_i}(x_j) > 0$ then it means that there are ties or non-comparable elements in the ordering made by ith observer. Moreover, more ties or elements that are not comparable with the others are present, bigger values of the intuitionistic fuzzy index are observed. One may also notice that $\pi_{A_i}(x_j) = 1$ if and only if element $x_j \in X$ is non-comparable with other element or all elements $x_1, \ldots, x_n$ have obtained the same rank in the ordering made by ith observer.

Hence it is seen that IF-sets seem to be a natural and useful tool for modelling nonlinear orderings.

5 Generalization of Friedman's test

According to (3) the test statistic for testing independence might be expressed in a following way

$$S(R) = d(R, \overline{R^*}), \qquad (13)$$

where $d(R, \overline{R^*})$ denotes a distance between the observed column totals $R = (R_1, \ldots, R_n)$ and the average column totals $\overline{R^*}$ obtained for perfect agreement between rankings. It can be shown that $\overline{R^*} = (\overline{R_1^*}, \ldots, \overline{R_n^*})$ and $\overline{R_j^*} = \frac{k(n+1)}{2}$ for each $j = 1, \ldots, n$.

Now to construct a straightforward generalization of Friedman's test for orderings containing elements that cannot be ranked by all observers, we have to find counterparts of R and $\overline{R^*}$ and a suitable measure of distance between these two objects (see [7]). As we have suggested in the previous section, we would consider appropriate IF-sets $A_1, \ldots, A_k$ for modelling ill-defined rankings. Thus instead of R will also consider an IF-set A, defined as follows

$$A = \{\langle x_j, \mu_A(x_j), \nu_A(x_j)\rangle : x_j \in X\}, \qquad (14)$$

where the membership and nonmembership functions μ_A and ν_A, respectively, are given by

$$\mu_A(x_j) = \frac{1}{k}\sum_{i=1}^{k} \mu_{A_i}(x_j), \qquad (15)$$

$$\nu_A(x_j) = \frac{1}{k}\sum_{i=1}^{k} \nu_{A_i}(x_j). \qquad (16)$$

If there is a perfect agreement within the group of observers and all objects are ranked without ties, then the resulting IF-set is of a form $A^* = \{\langle x_j, \mu_{A^*}(x_j), \nu_{A^*}(x_j)\rangle : x_j \in X\}$ such that the membership function is given by

$$\begin{aligned}
\mu_{A^*}(x_{j_1}) &= \frac{n-1}{n-1} = 1, \qquad (17)\\
\mu_{A^*}(x_{j_2}) &= \frac{n-2}{n-1},\\
\mu_{A^*}(x_{j_3}) &= \frac{n-3}{n-1},\\
&\cdots\\
\mu_{A^*}(x_{j_{n-1}}) &= \frac{1}{n-1},\\
\mu_{A^*}(x_{j_n}) &= 0.
\end{aligned}$$

where $x_{j_1}, \ldots, x_{j_n}$ is a permutation of elements $x_1, \ldots, x_n$ and the nonmembership function is

$$\nu_{A^*}(x_j) = 1 - \mu_{A^*}(x_j) \qquad (18)$$

for each $j = 1, \ldots, n$. Therefore, for perfect agreement between rankings, instead of the average column totals $\overline{R^*}$ we obtain an IF-set $\overline{A^*} = \{\langle x_j, \mu_{\overline{A^*}}(x_j), \nu_{\overline{A^*}}(x_j)\rangle : x_j \in X\}$ such that

$$\mu_{\overline{A^*}}(x_1) = \ldots = \mu_{\overline{A^*}}(x_n) = \frac{1}{2}, \qquad (19)$$

$$\nu_{\overline{A^*}}(x_1) = \ldots = \nu_{\overline{A^*}}(x_n) = \frac{1}{2}. \qquad (20)$$

Now, after substituting R and $\overline{R^*}$ by IF-sets A and $\overline{A^*}$, respectively, we have to choose a suitable distance between these two IF-sets. Several measures of distance between IF-sets were considered in the literature (see, e.g. [4]). In this paper we will apply a distance proposed by Atanassov [2], i.e. such function $d : IFS(X) \times$

$IFS(X) \rightarrow R^+ \cup \{0\}$ which for any two IF-subsets $B = \{\langle x_j, \mu_B(x_j), \nu_B(x_j)\rangle : x_j \in X\}$ and $C = \{\langle x_j, \mu_C(x_j), \nu_C(x_j)\rangle : x_j \in X\}$ of the universe of discourse $X = \{x_1, \ldots, x_n\}$ is defined as

$$d(B,C) = \sum_{j=1}^{n} \Big[(\mu_B(x_j) - \mu_C(x_j))^2 + (\nu_B(x_j) - \nu_C(x_j))^2 \Big]. \quad (21)$$

For actual observed rankings, modelled by IF-sets $A_1, \ldots, A_k$, test statistic is a distance (21) between IF-set A obtained from (14)-(16) and $\overline{A^*}$ obtained from (19)-(20) and is given by

$$S_{IF}(A) = d(A, \overline{A^*}) = \sum_{j=1}^{n} \left[\left(\mu_A(x_j) - \frac{1}{2} \right)^2 + \left(\nu_A(x_j) - \frac{1}{2} \right)^2 \right]. \quad (22)$$

It can be shown that a following lemma holds:

Lemma 5.1 *For any $A \in IFS$ there exist two intuitionistic fuzzy sets A^+ and A^- such that:*

$$S_{IF}(A^-) \leq S_{IF}(A) \leq S_{IF}(A^+) \quad (23)$$

where

$$A^- = \{\langle x_j, \mu_A(x_j), \nu_A(x_j) + \pi_A(x_j)\rangle : x_j \in X\},$$
$$A^+ = \{\langle x_j, \mu_A(x_j) + \pi_A(x_j), \nu_A(x_j)\rangle : x_j \in X\}.$$

Hence our test statistic $S_{IF}(A)$ based on ill-defined data is bounded by two other statistics $S^-_{IF}(A) = S_{IF}(A^-)$ and $S^+_{IF}(A) = S_{IF}(A^+)$ corresponding to situations with perfect rankings. Indeed, for each $x_j \in X$

$$\mu_{A^-}(x_j) = 1 - v_{A^-}(x_j) \Rightarrow \pi_{A^-}(x_j) = 0$$
$$\mu_{A^+}(x_j) = 1 - v_{A^+}(x_j) \Rightarrow \pi_{A^+}(x_j) = 0$$

which means that A^- and A^+ describe situations when all elements are univocally classified. Therefore, there exist two systems of rankings (in a classical sense) R^- and R^+ and one-to-one mapping transforming A^- and A^+ onto R^- and R^+. Next, it could be shown that

$$S^-_{IF}(A) = \frac{2}{k^2 (n-1)^2} S(R^-) \quad (24)$$
$$S^+_{IF}(A) = \frac{2}{k^2 (n-1)^2} S(R^+). \quad (25)$$

Since $S^-_{IF}(A)$ and $S^+_{IF}(A)$ are linear functions of S, we can easily find their distributions necessary for the desired test. Namely, if k is large enough then

$$T_1 = \frac{6k}{n(n-1)(n^2-1)} S^-_{IF} \quad (26)$$
$$T_2 = \frac{6k}{n(n-1)(n^2-1)} S^+_{IF}. \quad (27)$$

are chi-square distributed with $k-1$ degrees of freedom. Traditionally, in hypothesis testing we reject the null hypothesis H if test statistic belongs to critical region or accept H otherwise. In our problem with missing data we get two test statistics T_1 and T_2. Although, by (23) $T_2 \geq T_1$, we have to utilize both statistics for decision making.

Thus finally, the hypothesis H should be rejected on the significance level α if

$$T_1(A) \geq \chi^2_{n-1,\alpha} \quad (28)$$

while there are no reasons for rejecting H (i.e. we accept H) if

$$T_2(A) < \chi^2_{n-1,\alpha}. \quad (29)$$

These two situations are quite obvious. However, it may happen that

$$T_1(A) < \chi^2_{n-1,\alpha} \leq T_2(A). \quad (30)$$

In such a case we are not completely convinced neither to reject nor to accept H. Thus instead of a binary decision we could indicate a degree of conviction that one should accept or reject H. Moreover, going back to theory of necessity we may define a following measure:

$$Ness(\text{reject } H) = \begin{cases} 1 & \text{if } T_1(A) \geq \chi^2_{n-1,\alpha} \\ \frac{T_2(A) - \chi^2_{n-1,\alpha}}{T_2(A) - T_1(A)} & \text{if } T_1(A) < \chi^2_{n-1,\alpha} \leq T_2(A) \\ 0 & \text{if } T_2(A) < \chi^2_{n-1,\alpha} \end{cases}$$

describing the degree of necessity for rejecting H. Simultaneously we get another measure

$$Poss(\text{accept } H) = 1 - Ness(\text{reject } H) \quad (31)$$

describing the degree of possibility for accepting H.

Thus, by (28) $Ness(\text{reject } H) = 1$ means that the null hypothesis should be rejected and hence there is no possibility for accepting H. If condition (29) is fulfilled then $Poss(\text{accept } H) = 1$ and $Ness(\text{reject } H) = 0$. Finally, if (30) holds then $Ness(\text{reject } H) = \xi \in (0, 1)$ shows how strong our data are for or against $(1 - \xi)$ the null hypothesis H.

6 Conclusion

In the paper we have proposed how to generalize the well-known Friedman's test to situations in which not all elements could be ordered. We have discussed Friedman's test as a nonparametric tool for testing independence of k variates. However, this very test could be also applied as nonparametric two-way analysis of variance for the balanced complete block design. In this case kn subjects are grouped into k blocks each containing n subjects and within each block n treatments are assigned randomly to the matched subjects. In order to determine whether the treatment effects are all the same, Friedman's test could be used. It should be stressed that our generalized version of Friedman's test also works for two-way analysis of variance by ranks with missing data.

References

[1] K. Atanassov, "Intuitionistic fuzzy sets", Fuzzy Sets and Systems, vol. 20 (1986), 87–96.

[2] K. Atanassov, "Intuitionistic Fuzzy Sets: Theory and Applications", Physica-Verlag, 1999.

[3] J.D. Gibbons, S. Chakraborti, "Nonparametric Statistical Inference", Marcel Dekker, 2003.

[4] P. Grzegorzewski, "Between intuitionistic fuzzy sets and/or interval-valued fuzzy sets based on the Hausdorff metric", Fuzzy Sets and Systems, vol. 148 (2003), 319-328.

[5] P. Grzegorzewski, "The generalized Spearman's rank correlation coefficient", In: Proceedings of the Tenth International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Perugia, Italy, July 4-9, 2004, pp. 1413-1418.

[6] P. Grzegorzewski, "On measuring association betwwen preference systems", In: Proceedings of the 2004 IEEE International Conference on Fuzzy Systems, Budapest, Hungary, July 25-29, 2004, pp. 133-137.

[7] P. Grzegorzewski, "The Generalized Coefficient of Concordance", In: Current Issues in Data nad Knowledge Engineering, B. De Baets at all. (Eds.) Warszawa 2004, pp. 241-251.

[8] L.A. Zadeh, "Fuzzy sets", Inform. and Control, vol. 8 (1965), 338–353.