

# Missing Data in Cross-Over Trials

Edyta Winciorek<sup>1</sup>, Quanticate, Warsaw, Poland

## 1 Abstract

A cross-over trial is a design, in which each patient is given a sequence of at least two treatments. The purpose of such trials is to investigate the differences between the individual treatments, which make up the sequences. Such experiments are most useful in investigating long-lasting, stable diseases such as asthma, rheumatism or epilepsy. It should be stressed that in real experiments patients often do not receive an equal number of treatments. This, for example, may be a consequence of patients' fundamental right to withdraw from a trial at any time. Therefore, the issue of imperfect and missing data should constantly be borne in mind during designing and handling cross-over projects. In general, classical statistical procedures assume that data are complete, and the topic of how to handle missing data is not often discussed outside statistical journals. Therefore the most common and easiest approach is to omit the data with missing values. Using this approach when only a very few observations are missing does little harm. However, in case of a large number of observations with missing data, omitting patients without full data, results in a large proportion of the data being discarded, and eventually leads to biased results of statistical analyses. In this presentation we discuss alternative ways of dealing with missing data in cross-over trials.

## 2 Cross-Over Trials

### 2.1 Introduction to AB/BA Design

The simplest type of cross-over design is *AB/BA* **cross-over** also known as the **two treatment, two periods cross-over**. In such studies the number of  $n_1$  patients were given drug *A* for a fixed period followed by the period without any treatment and after that given the drug *B*. The drug *B* is administered after the **wash-out period**, i.e. the period in a trial which the effect of a treatment given previously is assumed to disappear. On the other hand the number of  $n_2$  patients were given the drug *B* followed by period without any treatment, and then the treatment *A* after the wash-out period.

In case of more than two treatments a very common sort of design assumes that the patients were randomized to all possible combination of sequences. According this design in a cross-over trial with three treatments *A, B* and *C* for example, the patients will be randomized to all six possible sequences of one treatment each, as was illustrated [3] :

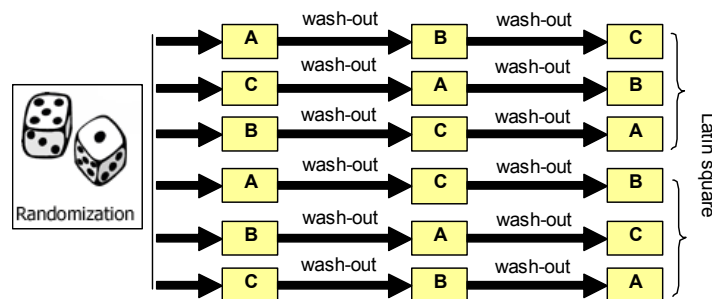


Figure 1. Subject randomization model for cross-over design with three treatments.

<sup>1</sup>PhD candidate in Polish Academy of Science.

In some cases, especially for quite big number of treatments, to examine so-called **balanced incomplete blocks** design comes in useful. In such design incompleteness results from fact that each patient receives only  $k_1$  of  $k$  treatments, and not all possible sequences are considered. However, the sequences are chosen so that every treatment appeared equally often in each of the  $k_1$  periods, the patient received  $k_1$  treatments, and every possible pair of treatments appeared in an equal number of sequences. Therefore such design is balanced because each pair of treatments is represented equally often.

The table (1) is the most common structure used in the medical literature [4] for the representation of cross-over design. All the data are split into columns which represent the appropriate periods and grouped in respect of the received treatment:

| Group 1   |           |     |           |
|-----------|-----------|-----|-----------|
|           | Period    |     |           |
| Subject   | $p_1$     | ... | $p_n$     |
| $S_1$     | $Y_{111}$ | ... | $Y_{1n1}$ |
| ...       | ...       | ... | ...       |
| $S_i$     | $Y_{11i}$ | ... | $Y_{1ni}$ |
| ...       | ...       | ... | ...       |
| Group $i$ |           |     |           |
| $S_j$     | $Y_{i1j}$ | ... | $Y_{inj}$ |
| ...       | ...       | ... | ...       |
| $S_k$     | $Y_{i1k}$ | ... | $Y_{ink}$ |

Table 1: Period orientated structure.

The value  $Y_{ikn}$  represents of subject  $S_n$  in period  $p_k$  in group (sequence)  $i$ . In cross-over models we assume that:

$$Y_{ikn} = E[Y_{ikn}] + \varepsilon_{ikn} \quad (1)$$

where  $\varepsilon_{ikn}$  is a **random error** resulting from variety of phenomena while  $E[Y_{ikn}]$  is **expected response of patient  $n$  in period  $k$** . The random error by definition is supposed to have  $E[\varepsilon_{ikn}] = 0$ . Moreover, we assume that  $\varepsilon_{ip}$  are independently normally distributed with constant variance  $\sigma^2$ . However, in order to model  $E[Y_{ikn}]$  we have to take into consideration the following factors:

- $\mu$  – an **effect due to patient**,
- $\tau$  – an **expected effect due to treatment**,
- $\pi$  – an **expected effect due to period**,
- $\lambda$  – a **carry over effect due to treatment**.

## 2.2 Analysis of Treatment Effect

As we can see in the analysis of treatment difference we need to eliminate the period and carry over effect. Generally speaking the carry-over effect is the most serious potential drawback of cross-over trials. Some estimators efficiently eliminate the the effects of periods but are susceptible to the problem of carry-over effect. Therefore, the standard analysis of the cross-over consists first of all of performing a statistical test on data to examine the possibility of

carry-over having occurred. In the purpose of this article we will assume the carry over effect did not affect the output results.

For two treatment, two period designs the most popular analysis is based on period difference. Namely, for each subject the  $p1 - p2$  difference is calculated. If there is any treatment difference, then the size of the  $p1 - p2$  in the first group should be the same as those in the second group. Therefore, comparing the size of the period difference between two groups will lead us to test the null hypothesis:

$$H : \tau_1 - \tau_2 = 0. \tag{2}$$

Assuming normal distribution one may accomplish such analysis using **Two sample  $t$  test**. However, if data are not normal distributed, then a nonparametric analysis should be applied. The most common approach refers to the **Wilcoxon rank sum test** [2].

There is no general approach to the analysis of cross-over data in more than two treatments. The most convenient approach is to reduce the analysis to one that is based on the analysis of the two treatments. However, because of the potential for period effect we cannot compare the treatments directly. Therefore, we first have to remove the period effect and after that investigate the treatment effect. If  $\hat{\pi}_n$  is an estimator of period  $k$ , then

$$Y'_{ik}(n) = Y_{ink} - \hat{\pi}_n$$

is a response of subject  $S_k$  from group  $i$  on a treatment given in period  $n$ . It is easy to notice, that  $Y'_{ik}(n)$  is free of the period effect. In such way the data of the trial can be arranged in the following table:

| Group 1   |              |     |              |
|-----------|--------------|-----|--------------|
|           | Treatment    |     |              |
| Subject   | $t_1$        | ... | $t_k$        |
| $S_1$     | $Y'_{11}(1)$ | ... | $Y'_{11}(n)$ |
| ...       | ...          | ... | ...          |
| $S_i$     | $Y'_{1i}(1)$ | ... | $Y'_{1i}(n)$ |
| ...       | ...          | ... | ...          |
| Group $i$ |              |     |              |
| $S_j$     | $Y'_{ij}(1)$ | ... | $Y'_{ij}(n)$ |
| ...       | ...          | ... | ...          |
| $S_k$     | $Y'_{ik}(1)$ | ... | $Y'_{ik}(n)$ |

Table 2: Treatment orientated structure

Since the design 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 the 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. A Friedman's test seems to be a suitable approach in this situation.

Suppose that response on treatments  $t_1, \dots, t_n$  are ordered according to  $k$  subjects  $S_1, \dots, S_k$ . Then our data could be presented in the form of a two-way layout (or matrix)  $M$  with  $k$  rows and  $n$  columns. Let  $R_{ij}$ ,  $i = 1, 2, \dots, k$ ,  $j = 1, 2, \dots, n$ , denote the ranked observations so that  $R_{ij}$  is the rank obtain for subject  $i$  to the  $j$ th treatment. Then  $R_{i1}, R_{i2}, \dots, R_{in}$  is a permutation of the first  $n$  integers while  $R_{1j}, R_{2j}, \dots, R_{kj}$  is a set of rankings assigned to treatment  $j$ .

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

$$M = \begin{bmatrix} R_{11} & R_{12} & \dots & R_{1n} \\ \dots & \dots & \dots & \dots \\ R_{k1} & R_{k2} & \dots & R_{kn} \end{bmatrix}. \quad (3)$$

Suppose we are interested in testing the null hypothesis that there is no difference between ranking of treatments. If  $R = (R_1, \dots, R_n)$  denotes observed column totals, where

$$R_j = \sum_{i=1}^k R_{ij}, \quad j = 1, \dots, n. \quad (4)$$

and let  $\bar{R}$  denotes 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:

$$S(R) = \sum_{j=1}^n [R_j - \bar{R}]^2 = \sum_{j=1}^n \left[ R_j - \frac{k(n+1)}{2} \right]^2. \quad (5)$$

It can be shown that the value of  $S$  for any sets of  $k$  rankings ranges  $< 0, k^2n(n^2 - 1)/12 >$ , with minimum value attained when each subject's rankings are assigned completely at random. Therefore,  $S$  maybe used to test our null hypothesis  $H$ . If the null hypothesis holds then following statistic

$$Q = \frac{12S}{kn(n+1)} \quad (6)$$

can be used to define the rejection region for our hypothesis testing problem. It was proved that statistic (6) approaches the chi-square distribution with  $n - 1$  degrees of freedom as  $k$  increases.

It is easily seen that Friedman's test could be used provided all elements are univocally classified for all subjects. However, it may happen that for some subjects we cannot rank all the treatments responses under study because some are missing. One way out is then to remove all objects which are not ordered for all of the subjects 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 data is large, eliminating them we may have some difficulties to apply the Friedman's test . Suppose for example we have 5 subjects and 5 treatments. The following matrix shows how the

subjects' response on treatments:

$$M = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ (?) & 1 & 2 & 3 & 4 \\ 1 & (?) & 2 & 3 & 4 \\ 1 & 2 & (?) & 3 & 4 \\ 1 & 2 & 3 & (?) & 4 \end{bmatrix}, \quad (7)$$

where symbol (?) means that considered object has not been ranked. In this example following the above mentioned rule lead us to the situation where all of columns in matrix (7) should be eliminated. On the other hand, it is obvious that the treatment effects differ form each other.

Taking above described situations into consideration 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 Incomplete Data Analysis Methods

#### 4 IF-sets in Modeling rankings

A method proposed below seems to be useful especially if not all elements under consideration could be ranked. In suggested approach we will attribute an IF-sets to the ordering corresponding to each subject. For simplicity of notation we will identify orderings assigned to subjects with the corresponding IF-sets  $S_1, \dots, S_k$ . Thus, for each subject let

$$S_i = \{ \langle t_j, \mu_{S_i}(t_j), \nu_{S_i}(t_j) \rangle : t_j \in T \} \quad (8)$$

denote an IF-set, where function  $\mu_{S_i}(t_j)$  (called membership function) indicates the degree to which treatment  $t_j$  give the most preferred response for  $i$ th subject, while function  $\nu_{S_i}(t_j)$  (called nonmembership function) shows the degree to which  $t_j$  is the less preferred response for  $i$ th subject.

Here a natural question arises: how to determine these membership and nonmembership functions. Namely, for each subject one can always specify two functions  $w_{S_i}, b_{S_i} : T \rightarrow \{0, 1, \dots, n-1\}$  defined as follows: for each given  $t_j$  let  $w_{S_i}(t_j)$  denote the number of treatments surely worse than  $t_j$ , while  $b_{S_i}(t_j)$  let be equal to the number of treatments surely better than  $t_j$  in the ordering corresponding to subject  $S_i$ . Then, we have

$$\mu_{S_i}(t_j) = \frac{w_{S_i}(t_j)}{n-1}, \nu_{S_i}(t_j) = \frac{b_{S_i}(t_j)}{n-1} \quad (9)$$

In such a way we get  $k$  well defined IF-sets which describe nicely the treatment orderings corresponding to  $k$  subject. It should be stress that  $0 \leq \mu_{S_i}(t_j) + \nu_{S_i}(t_j) \leq 1$ . Therefore,  $\pi_{S_i}(t_j) = 1 - \mu_{S_i}(t_j) - \nu_{S_i}(t_j)$ , called the IF-index, quantifies the amount of indeterminacy associated with  $S_i$  and  $t_j$ . Therefore  $\pi_{S_i}(t_j) = 0$  for each treatment  $t_j$  if and only if all treatments are ranked for  $i$ th subject and there are no ties. Conversely, if there exist such treatment that  $\pi_{S_i}(t_j) > 0$  then it means that there are ties or non-comparable elements in the ordering. One may also notice that  $\pi_{S_i}(t_j) = 1$  if and only if treatment  $t_j$  is non-comparable with other element or all treatments have obtained the same rank.

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

#### 5 Generalization of Friedman's test

According to (5) the test statistic for testing independence might be expressed in a following way

$$S(R) = d(R, \overline{R^*}), \quad (10)$$

where  $d(R, \overline{R^*})$  denotes a distance between the observed column totals  $R = (R_1, \dots, R_n)$  and the average column totals  $\overline{R^*}$  obtained for perfect agreement between rankings. It can be shown that  $\overline{R^*} = (\overline{R_1^*}, \dots, \overline{R_n^*})$  and  $\overline{R_j^*} = \frac{k(n+1)}{2}$  for each  $j = 1, \dots, n$ . Now to construct a straightforward generalization of Friedman's test we have to find counterparts of  $R$  and  $\overline{R^*}$  and a suitable measure of distance between these two objects. Thus instead of  $R$  we will consider an IF-set  $S$ , defined as follows

$$S = \{ \langle t_j, \mu_s(t_j), \nu_s(t_j) \rangle : t_j \in T \}, \quad (11)$$

where the membership and nonmembership functions  $\mu_s$  and  $\nu_s$ , respectively, are given by

$$\mu_s(t_j) = \frac{1}{k} \sum_{i=1}^k \mu_{S_i}(t_j) \quad \text{and} \quad \nu_s(t_j) = \frac{1}{k} \sum_{i=1}^k \nu_{S_i}(t_j). \quad (12)$$

It can be proved that for perfect agreement between rankings, instead of the average column totals  $\overline{R^*}$  we obtain an IF-set  $\overline{S^*} = \{ \langle t_j, \mu_{\overline{S^*}}(t_j), \nu_{\overline{S^*}}(t_j) \rangle : t_j \in X \}$  such that

$$\mu_{\overline{S^*}}(t_1) = \dots = \mu_{\overline{S^*}}(t_n) = \frac{1}{2} \quad \text{and} \quad \nu_{\overline{S^*}}(t_1) = \dots = \nu_{\overline{S^*}}(t_n) = \frac{1}{2}. \quad (13)$$

Now, after substituting  $R$  and  $\overline{R^*}$  by IF-sets  $S$  and  $\overline{S^*}$ , 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. In this paper a distance proposed by Atanassov is recommended. Therefore, for actual observed rankings, modeled by IF-sets  $S_1, \dots, S_k$ , test statistic is a distance (??) between IF-set  $S$  obtained from (11)-(12) and  $\overline{S^*}$  obtained from (??)-(13) and is given by

$$\tilde{D}(S) = d(S, \overline{S^*}) = \sum_{j=1}^n \left( \mu_S(t_j) - \frac{1}{2} \right)^2 + \left( \nu_S(t_j) - \frac{1}{2} \right)^2. \quad (14)$$

It can be shown that for any IF-set  $S$  there exist two IF-sets  $S^\sharp$  and  $S^\flat$  such that:

$$\min \left\{ \tilde{D}(S^\flat), \tilde{D}(S^\sharp) \right\} \leq \tilde{D}(S) \leq \max \left\{ \tilde{D}(S^\flat), \tilde{D}(S^\sharp) \right\} \quad (15)$$

where

$$S^\flat = \{ \langle t_j, \mu_S(t_j), \nu_S(t_j) + \pi_S(t_j) \rangle : t_j \in T \}, \quad (16)$$

$$S^\sharp = \{ \langle t_j, \mu_S(t_j) + \pi_S(t_j), \nu_S(t_j) \rangle : t_j \in T \}. \quad (17)$$

Hence our test statistic  $\tilde{D}(S)$  based on ill-defined data is bounded by two other statistics  $\tilde{D}(S^\flat)$  and  $\tilde{D}(S^\sharp)$  corresponding to situations with perfect rankings. Indeed, for each  $x_j \in X$

$$\mu_{S^\flat}(t_j) = 1 - \nu_{S^\flat}(t_j) \Rightarrow \pi_{S^\flat}(t_j) = 0$$

$$\mu_{S^\sharp}(t_j) = 1 - \nu_{S^\sharp}(t_j) \Rightarrow \pi_{S^\sharp}(t_j) = 0$$

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

$$\tilde{D}(S^\flat) = \frac{2}{k^2(n-1)^2} D(R^\flat) \quad \text{and} \quad \tilde{D}(S^\sharp) = \frac{2}{k^2(n-1)^2} D(R^\sharp). \quad (18)$$

Since  $\tilde{D}(S^\flat)$  and  $\tilde{D}(S^\sharp)$  are linear functions of  $D$ , we may proved that

$$T_1(S) = \frac{6k(n-1)^2}{n(n+1)} \tilde{D}(S^\flat) \quad \text{and} \quad T_2(S) = \frac{6k(n-1)^2}{n(n+1)} \tilde{D}(S^\sharp), \quad (19)$$

are chi-square distributed with  $n-1$  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 statistics  $T_1$  and  $T_2$  and we have to utilize them for decision making.

According to (15) we get the following inequality

$$T_{\min} \leq T(S) \leq T_{\max} \quad (20)$$

where  $T_{\min} = \min \{T_1(S), T_2(S)\}$ ,  $T_{\max} = \max \{T_1(S), T_2(S)\}$  and  $T(S) = \frac{6k(n-1)^2}{n(n+1)} \tilde{D}(S)$ . Let us denote by  $q_\alpha^{\min}$  and  $q_\alpha^{\max}$  the critical values of  $T_{\min}$  and  $T_{\max}$ , respectively. It means that with  $q_\alpha^{\min}$  and  $q_\alpha^{\max}$  we will find as the solutions of the following equations:

$$P(T_{\min} > q_\alpha^{\min}) = \alpha \quad \text{and} \quad P(T_{\max} > q_\alpha^{\max}) = \alpha. \quad (21)$$

In order to find the critical values  $q_\alpha^{\min}$  and  $q_\alpha^{\max}$  we have to determine the random sampling distribution of  $T_{\min}$  and  $T_{\max}$  under the assumption of independence. For this purpose the Monte Carlo simulation has been run. Numerical comparisons have shown that  $T_{\min}$  and  $T_{\max}$  approaches the gamma distribution as a dimension of  $M$  increases.

Going back to our hypothesis testing problem, we should reject  $H$  on the significance level  $\alpha$  if

$$T(S) \geq q_{\alpha}^{\max} \quad (22)$$

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

$$T(S) < q_{\alpha}^{\min} \quad (23)$$

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

$$q_{\alpha}^{\min} \leq T(S) < q_{\alpha}^{\max}.$$

In such a case we are not completely convinced to neither reject nor accept  $H$  (see Figure 2).

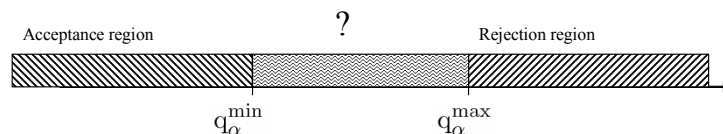


Figure 2. Rejection region for hypothesis testing problem with missing data.

Thus instead of a binary decision we could indicate a degree of conviction that one should accept or reject  $H$ . The measure describing degree of necessity for rejecting  $H$  is given by following formula:

$$Ness(\text{reject } H) = \begin{cases} 1 & \text{if } q_{min} \geq T(S) \\ \frac{q_{max} - T(S)}{q_{max} - q_{min}} & \text{if } q_{min} < T(S) \leq q_{max} \\ 0 & \text{if } q_{max} < T(S) \end{cases} \quad (24)$$

Simultaneously we get another measure

$$Poss(\text{accept } H) = 1 - Ness(\text{reject } H) \quad (25)$$

describing the degree of possibility for accepting  $H$ .

## 6 Conclusion

In the paper we have discussed well-known Friedman's test as a nonparametric tool for testing the treatment effect in cross-over studies with more than two treatments. The Friedman's test give us a single test statistic designed to detect overall dependence between treatments with a specified significance level. Moreover, we have shown how to generalize this test to situations in which not all elements could be ordered. Presented approach refers to so called fuzzy statistic, a quite new discipline in statistical data analysis.

## References

- [1] K. Atanassov (1986): *Intuitionistic fuzzy sets*, Fuzzy Sets and Systems **20**, 87–96.
- [2] G.G. Koch (1972): *The use of non-parametric methods in the statistical analysis of the two-period change-over design*, Biometrics **28**, 577-84.
- [3] Max M. B. and Lynn J. (2004): *The Symptom Research interactive clinical research textbook contains over 20 chapters on different aspects of clinical pain and symptom research*, Interactive Textbook on Clinical Symptom Research, <http://symptomresearch.nih.gov>
- [4] S.J. Senn, (1993): *Cross-over Trials in Clinical Research*, Chichester and New York, Wiley.
- [5] L.A. Zadeh (1965): *Fuzzy sets*, Inform. and Control **8** , 338–353.