Chapter 11
Nonparametric tests for the randomized complete block design with ordered categorical variables

Livio Corain and Luigi Salmaso

11.1 Introduction

In many scientific disciplines and industrial fields, when dealing with comparisons between two or more treatments, researchers and practitioners are often faced with theoretical and practical problems within the framework of Randomized Complete Block (RCB) design with ordered categorical response variables. This situations can arise very often in the field of the evaluation of educational services or quality of products, for example in connection with the sensorial testing studies, where several useful experimental performance indicators, especially in the food and body care industry, are provided by individual sensorial evaluations by trained people (panelists) during a so-called sensory test (Meilgaard et al., 2006). Within this framework the experimental design typically handles panelists as blocks.

In general, the requirement to take into consideration a RCB design occurs when the experimental units are heterogeneous, hence the notion of blocking is used to control the extraneous sources of variability. The major criteria of blocking are characteristics associated with the experimental material and the experimental setting. The purpose of blocking is to sort experimental units into blocks, so that the variation within a block is minimized while the variation among blocks is maximized. An effective blocking not only yields more precise results than an experimental design of comparable size without blocking, but also increases the range of validity of the experimental results.

In this contribution we propose a general solution within the Nonparametric Combination (NPC) of Dependent Permutation Tests (Pesarin, 2001) which is

Livio Corain
Department of Management and Engineering, University of Padua, Str. S. Nicola 4, 36100 Vicenza, Italy, e-mail: livio.corain@unipd.it

Luigi Salmaso
Department of Management and Engineering, University of Padua, Str. S. Nicola 4, 36100 Vicenza, Italy, e-mail: salmaso@gest.unipd.it

M. Bini et al. (eds.), *Statistical Methods for the Evaluation of Educational Services and Quality of Products*, Contributions to Statistics,
DOI 10.1007/978-3-7908-2385-1_11,

particularly suitable for the RCB design, especially in case of ordered categorical response variables such that used for sensorial studies. In the next section, we present an update review of the procedures proposed in the literature for the hypothesis testing on the RCD design. In Sect. 11.3 we present the proposed permutation solution for the RCB Design. In Sects. 11.4 and 11.5 a comparative simulation study and a real case study are presented. Finally, we conclude, in Section 6, with some directions of current and future research.

11.2 Overview on procedures proposed in the literature for the RCB design

Let us consider the experimental design where there are n blocks and, within each block, experimental units are randomly assigned to the C treatments ($C > 2$) and exactly one unit is assigned to each of the C treatments. The statistical model (with fixed effects) for the randomized complete block (RCB) design can be represented as follows:

$$Y_{ij} = \mu + \beta_i + \tau_j + \varepsilon_{ij}, \varepsilon_{ij} \sim IID(0, \sigma^2), i = 1, ..., n, j = 1, ..., C, \tag{11.1}$$

where β_i, τ_j and Y_{ij}, are respectively the effect of the i-th block, the effect of the j-th treatment and the response variable for the i-th block and the j-th treatment. The random term ε_{ij} represents the experimental error with zero mean, variance σ^2 and unknown continuous distribution P. The usual side-conditions for effects are given by the constrains $\sum_i \beta_i = \sum_j \tau_j = 0$.

Model (11.1) is called "effect model" (Montgomery, 2005). If we define $\mu_j = \mu + \tau_j$, $j = 1, ..., C$, an alternative representation of model (11.1) is the so called "mean model", i.e.

$$Y_{ij} = \mu_j + \beta_i + \varepsilon_{ij}. \tag{11.2}$$

The resulting inferential problem of interest is concerned with the following hypotheses: $H_0 : \{\tau_j = 0, \forall\, j\}$, against $H_1 : \{\exists j : \tau_j \neq 0\}$. Note that this hypothesis is referred to a global test; if H_0 is rejected, it is of interest to perform inference on each pairwise comparison between couples of treatments, i.e. $H_{0(jh)} : \tau_j = \tau_h,\ j, h = 1, ..., C$, $j \neq h$, against $H_{1(jh)} : \tau_j \neq \tau_h$; with reference to model (11.2), an equivalent representation of $H_{0(jh)}$ is the following: $H_{0(jh)} : \mu_j - \mu_h = 0,\ j, h = 1, ..., C$, $j \neq h$, against $H_{1(jh)} : \mu_j - \mu_h \neq 0$.

We recall that in the framework of RCB designs there is usually no interest in testing the block effect which is handled as a nuisance factor. Note that, since no interaction effect between treatments and blocks is here supposed to exist, expressions (11.1) and (11.2) do not consider any interaction effect.

In the framework of traditional parametric methods, when assuming random normal components, it is appropriate to test the equality of all treatment means by using the traditional F statistic:

$$F = \frac{SS_{\text{Treatments}}/(C-1)}{SS_{\text{E}}/(n-1)(C-1)}, \tag{11.3}$$

where $SS_{\text{Treatments}} = n \sum_{j=1}^{C} (\overline{Y}_{\cdot j} - \overline{Y}_{\cdot\cdot})^2$, $SS_{\text{E}} = \sum_{i=1}^{n} \sum_{j=1}^{C} (Y_{ij} - \overline{Y}_{\cdot j} - \overline{Y}_{i\cdot} + \overline{Y}_{\cdot\cdot})^2$ and $\overline{Y}_{\cdot j}$ is the mean of the n experimental units in the j-th treatment, $\overline{Y}_{i\cdot}$ is the block mean for the i-th block, and $\overline{Y}_{\cdot\cdot}$ is the overall mean. The F statistic is distributed as $F_{C-1,(C-1)(n-1)}$ if the null hypothesis H_0 is true, hence we would reject H_0, at the significance level α, if $F_0 > F_{\alpha;(C-1),(C-1)(n-1)}$. If the analysis indicates a significant difference in treatment means, we are usually interested in multiple comparisons to find out which treatment means differ. That is, when the global null hypothesis H_0 would be rejected we would consider the post-hoc set of $C(C-1)/2$ individual $H_{0(jh)}$ null hypotheses. Under normality, Bonferroni adjusted t-tests or Tukey's tests are the most recommended procedures. We recall that when carrying out multiple testing, there should be a formal guarantee against incorrect decisions. The so called multiplicity problem is particularly relevant in multiple comparison problems, since omitting to consider the multiplicity issue can often cause biased statistical analyses (Westfall et al., 1999).

Since the normality assumption is often questionable, if we do not assume the normality of random errors we can take into consideration a nonparametric approach. In the framework of nonparametric rank-based testing procedures, one of the earlier tests has been proposed by Friedman (1937). A general form of the Friedman's statistic T, which incorporates a correction for ties (Lehmann and D'Abrera, 2006), is given by:

$$T = \frac{(C-1)\sum_{j=1}^{C} [R_{+j} - n(C+1)/2]^2}{\sum_{i=1}^{n}\sum_{j=1}^{C} (R_{ij})^2 - nC(C+1)^2/4}, \tag{11.4}$$

where R_{ij} is the rank of Y_{ij} among the experimental units in block i and $R_{+j} = \sum_j R_{ij}$ is the sum of the ranks for the j-th treatment over the n blocks. Under the null hypothesis, the R_{+j}'s should be close to $n(C+1)/2$ which is the average of the R_{+j}. Since T has an asymptotic Chi-square distribution with $C-1$ degree of freedom, we would reject the null hypothesis H_0 if $T_0 > \chi^2_{\alpha,C-1}$. After rejection of H_0, the comparisons between pairs of treatments can be performed via absolute differences of the sums of within-blocks ranks. This set of values have to be compared with an appropriate value r_α which is function of C and n. For small values of C and n, r_α has been tabulated whereas, as n tends to infinity, it can be approximated by the distribution of the range of independent standard normal variables. This procedure, called Wilcoxon-Nemenyi-McDonald-Thompson procedure (Hollander and Wolfe, 1999), has been designed in order to maintain an appropriate Maximum Experimentwise

Error Rate (MEER) α, where EER is defined as the probability to reject at least one true hypotheses in the set of $C(C-1)/2$ individual $H_{0(jh)}$ null hypotheses.

Following Lehmann and D'Abrera (2006), the formula (11.4) can be replaced by:

$$T = nd'\Sigma_0^{-1}d, \tag{11.5}$$

where $\Sigma_0 = (\sigma_{jj'})$ is the covariance matrix under the null hypothesis of $R_i = (R_{i1}, \cdots, R_{i,C-1})$, that is the rank order of the first $C-1$ treatments, and

$$d' = \left[R_{+1} - (C+1)/2, R_{+2} - (C+1)/2, \cdots, R_{+(C-1)} - (C+1)/2\right], \tag{11.6}$$

where $R_{+j} = \sum_j R_{ij}$. Sepansky (2007) suggests a modification of (11.6), by the following test statistic:

$$T_P = nd'\widehat{\Sigma}^{-1}d, \tag{11.7}$$

where $\widehat{\Sigma}^{-1} = (s_{jj'})$ is the sample covariance matrix of the R_i. Note that T_P is an Hotelling-type T^2 statistic and its limiting distribution is the χ^2 distribution with $C-1$ degrees of freedom (see Hollander and Wolfe, 2003). Sepansky (2007) examines also the covariance matrix in the test statistic (11.7) when the number of blocks or sample size is small and he claims that the null hypothesis of no treatment difference should be rejected when the sample covariance matrix is singular. It is worth noting that while the Friedman test statistic is well defined when n is less than C, T_P is not since the sample covariance matrix is singular for all possible data matrices in this case. The idea of Sepanky of rejecting the null hypothesis when the sample covariance matrix is questionable and he does not support this statement with any kind of formal proof and the motivation he provided is quite debatable. Moreover, the simulation results presented by author clearly show that, especially for small values of n, his test statistic does not maintain the nominal levels under the null hypothesis. Hence, this proposal might be unreliable to properly perform inference for RCB designs.

Another approach, refereed as aligned rank test (Lehmann and D'Abrera, 2006), is to make all blocks comparable so that comparisons between treatments in different blocks are meaningful. This can be done by subtracting the median or mean value of the experimental units in the block from all experimental units in that block. After this alignment is completed, the aligned experimental units are ranked over all blocks and treatments. It can be shown that, under the null hypothesis, the following statistic is a χ^2_{C-1} for large samples:

$$S = \frac{(C-1)n^2 \sum_{j=1}^{C} \left(\overline{R}_{\cdot j} - \overline{R}^2_{\cdot\cdot}\right)^2}{\sum_{i=1}^{n}\sum_{j=1}^{C} \left(R_{ij} - \overline{R}_{i\cdot}\right)^2}, \tag{11.8}$$

where now R_{ij} denotes the aligned rank for Y_{ij}, $\overline{R}_{i\cdot}$ is the average rank for the i-th block, $\overline{R}_{\cdot j}$ is the average rank for the j-th treatment and $\overline{R}_{\cdot\cdot}$ is the overall average rank.

In the literature there are a few other test statistics proposed for the RCB design. Among others, Quade (1979) proposed a test based on within-block rankings that gives greater weights to blocks that have greater variability. However, since several simulations studies (Fawcett and Salter, 1984); Groggel (1987) have shown that the Quade procedure is not well performing in some situations, hence as suggested by O'Gorman (2001), it will be not included in the simulations we will present afterwards in this work. O'Gorman (2001) reviews and evaluates several tests for RCB design, including the F-test, Friedman's test, and a few aligned rank tests. His simulations show that Friedman's test has low power compared with the aligned rank tests if the number of treatments does not exceed six and a novel aligned rank-based F-test proposed by the author shows relatively high power for several skewed distributions if there is a large number of experimental units.

11.3 Permutation tests for multivariate RCB design

When dealing with complex designs conditional nonparametric methods can represent a reasonable approach. We recall that traditional unconditional parametric testing methods (such as t test or F test) may be available, appropriate and effective only when a set of restrictive conditions are satisfied. Accordingly, just as there are circumstances in which unconditional parametric testing procedures may be appropriate, there are others where they may be unsuitable or even impossible to be properly applied. In conditional testing procedures, provided that exchangeability of data with respect to groups is satisfied in the null hypothesis, permutation methods play a central role. This is because they allow for quite efficient solutions, are useful when dealing with many difficult problems, provide clear interpretations of inferential results, and allow for weak extensions of conditional to unconditional inferences. For a detailed discussion on the topic of the comparison between permutation conditional inferences with traditional unconditional inferences we refer to Pesarin (2002).

In this chapter we propose a novel solution for the whole set of hypotheses of interest within the nonparametric framework of NonParametric Combination (NPC) of dependent permutation tests (Pesarin, 2001; Corain and Salmaso, 2004).

In order to better explain the proposed approach let us denote an $(n \times C)$ data set $\mathbf{Y}$ as:

$$\mathbf{Y} = \left[\mathbf{Y}_1, ..., \mathbf{Y}_j, ..., \mathbf{Y}_C\right] = \begin{bmatrix} Y_{11} & ... & Y_{1j} & ... & Y_{1C} \\ ... & ... & ... & ... & ... \\ Y_{i1} & ... & Y_{ij} & ... & Y_{iC} \\ ... & ... & ... & ... & ... \\ Y_{n1} & ... & Y_{nj} & ... & Y_{nC} \end{bmatrix},$$

where Y_{ij} represents the ijth observed response for ith block and jth treatment, $i = 1,...,n$, $j = 1,...,C$, $(C \geq 2)$.

In the framework of NonParametric Combination (NPC) of dependent permutation tests we suppose that, if the global null hypothesis H_0 is true, the hypothesis of exchangeability of random errors within the same block holds. Hence, the following set of mild conditions should be jointly satisfied:

(i) Suppose that for $\mathbf{Y} = [\mathbf{Y}_1,...,\mathbf{Y}_C]$ an appropriate distribution P_j exists, $P_j \in \mathscr{F}$, $j = 1,...,C$, belonging to a (possibly non-specified) family $\mathscr{F}$ of non-degenerate probability distributions;

(ii) The null hypothesis H_0 states the equality in distribution of the response variable in all C groups:

$$H_0 = [P_1,...,P_C] = [\mathbf{Y_1} \overset{d}{=} ... \overset{d}{=} \mathbf{Y_C}].$$

Null hypothesis H_0 implies the exchangeability, within each block, of the individual data with respect to the C groups. Moreover H_0 is supposed to be properly decomposed into $C \times (C-1)/2$ sub-hypotheses $H_{0(jh)}$, $j,h = 1,...,C$, $j \neq h$, each one related to the jhth pairwise comparison between couples of treatments:

$$H_0 = [\bigcap_{\substack{j,h=1 \\ j \neq h}}^{C} \mathbf{Y}_j \overset{d}{=} \mathbf{Y}_h] = [\bigcap_{\substack{j,h=1 \\ j \neq h}}^{C} H_{0(jh)}].$$

H_0 is called the *global* or *overall null hypothesis*, and $H_{0(jh)}$, $j,h = 1,...,C$, $j \neq h$, are the *partial null hypotheses*.

(iii) The alternative hypothesis H_1 is represented by the union of partial $H_{1(jh)}$ sub-alternatives:

$$H_1 = [\bigcup_{\substack{j,h=1 \\ j \neq h}}^{C} H_{1(jh)}] = [\bigcup_{\substack{j,h=1 \\ j \neq h}}^{C} H_{1(jh)}],$$

so that H_1 is true if at least one of sub-alternatives is true.
In this context, H_1 is called the *global* or *overall alternative*, and $H_{1(jh)}$, $j,h = 1,...,C$, $j \neq h$, are called the partial alternatives.

(iv) Let $\mathbf{T} = \mathbf{T}(\mathbf{Y})$ represent a vector of test statistics, whose components $T_{(jh)}$, $j,h = 1,...,C$, $j \neq h$, represent the partial univariate and non-degenerate *partial test* appropriate for testing the sub-hypothesis $H_{0(jh)}$ against $H_{1(jh)}$. Without loss of generality, all partial tests are assumed to be marginally unbiased, consistent and significant for large values (for more details see Pesarin, 2001).

At this point, in order to test the global null hypothesis H_0 and the $C \times (C-1)/2$ hypotheses $H_{0(jh)}$, we perform the partial (univariate) tests and then we combine them, with an appropriate combining function, in order to test the global null hypothesis H_0.

However, we should observe that in most real problems when the number of blocks is large enough, there might be computational difficulties in calculating the conditional permutation distribution. This means that it is not possible to calculate

the exact p-value of observed statistic $T_{(jh)0}$. This drawback is overcome by using the Conditional Monte Carlo (CMC) Procedure. The CMC on the pooled data set $\mathbf{Y}$ is a random simulation of all possible permutations of the same data under H_0 (for more details refer to Pesarin, 2001). Hence, in order to obtain an estimate of the permutation distribution under H_0 of all test statistics, a CMC can be used. It should be emphasized that CMC only considers permutations of individual data vectors within each individual block, so that all underlying dependence relations which are present in the component variables are preserved. From this point of view, the CMC is essentially a multivariate procedure.

A suitable algorithm for calculating the proposed permutation test is composed of the following steps:

(a) For each pairwise comparison between couples of treatments calculate the vector of the observed values of test statistics ${}^{o}\mathbf{T}(\mathbf{Y})$, whose components ${}^{o}T_{jh} = T(\mathbf{Y}_j, \mathbf{Y}_h)$, $j, h = 1, \ldots, C$, $j \neq h$, are appropriate for testing the sub-hypothesis $H_{0(jh)}$ against $H_{1(jh)}$.

(b) Consider $\mathbf{Y}^*$ as a permutation of the data set $\mathbf{Y}$, carried out within each ith block in order to preserve the dependence structure of data, then calculate the permutation value of the test statistics:

$$T^*_{jh} = T\left(\mathbf{Y}^*_j, \mathbf{Y}^*_h\right),\ j, h = 1, \ldots, C,\ j \neq h.$$

(c) Carry out B independent repetitions (i.e. Conditional Monte Carlo, CMC, iterations) of step (b). The set of CMC results $\left\{{}_{b}T^*_{jh},\ b = 1, \ldots, B\right\}$ is thus a random sampling from the permutation distribution of the test statistics.

(d) Obtain the p-value from each partial sub-hypothesis $H_{0(jh)}$:

$$\lambda_{jh} = \#\left(T^*_{jh} \geq^{o} T_{jh}\right)/B, \quad b = 1, \ldots, B,\ j, h = 1, \ldots, C,\ j \neq h.$$

(e) The combined observed value of the global or overall null hypothesis H_0 is:

$${}^{o}T'' = \psi(\lambda_{11}, \ldots, \lambda_{(C-1)C}).$$

(f) The combined value is then computed by:

$$T''^{*} = \psi\left(\lambda^*_{11}, \ldots, \lambda^*_{(C-1)C}\right).$$

where $\lambda^*_{jh} = \#\left(T''^{*}_{jh} \geq {}_{b}T''^{*}_{jh}\right)/B, \quad b = 1, \ldots, B.$

(g) The global p-value is computed as:

$$\lambda'' = \#(T''^{*} \geq {}^{o}T''), \quad b = 1, \ldots, B.$$

Matlab routines implementing permutation test for RCB design are available upon request by authors.

It can be seen that under the general null hypothesis the CMC procedure provides a consistent estimation of the permutation distributions, both marginal and

combined, of the S partial tests. In the nonparametric combination procedure, Fisher's combination function is usually considered, principally for its good properties which are both finite and asymptotic (Pesarin, 2001). Of course, if it were considered appropriate, it would be possible to take into consideration any other combining function. The combined test is unbiased and consistent.

A general characterization of the class of combining functions is given by the following three main features for the combining function ψ:

(a) it must be non-increasing in each argument:

$$\psi(\ldots,\lambda_s,\ldots) \geq \psi(\ldots,\lambda_s',\ldots) \text{ if } \lambda_s < \lambda_s', s \in \{1,\ldots,S\};$$

(b) it must attain its supreme value, possibly not finite, even when only one argument reaches zero:

$$\psi(\ldots,\lambda_s,\ldots) \to \overline{\psi} \text{ if } \lambda_s \to 0, s \in \{1,\ldots,S\};$$

(c) $\forall \alpha > 0$, the critical value of every ψ is assumed to be finite and strictly smaller than the supreme value:

$$T_\alpha'' < \overline{\psi}.$$

The above properties define the class C of combining functions. Some of the functions most often used to combine independent tests (Fisher, Lancaster, Liptak, Tippett, Mahalanobis, etc.) are included in this class. For a detailed description on how to build partial and global permutation tests refer to Pesarin (2001) and Corain and Salmaso (2004).

11.4 Simulation study

In order to validate the proposed method and to evaluate its performance in comparison with either the traditional parametric (F and t test) and the nonparametric approach (Friedman and aligned rank tests), in this section we perform a comparative simulation study. The goal is focused either on the global test H_0 and on the related treatment pairwise comparisons (hypotheses $H_{0(jh)}$).

The real context we are referring to is a typical sensorial study where the number of blocks (panel lists) usually ranges around 10–15 people and the sensorial evaluation is provided with a Likert 1–5 rating ordinal scale, where we suppose that the 0.5 scores are admitted as well. Note that we are actually considering a 9 point ordered categorical response variable.

Let us consider the following setting:

- 1,000 independent simulations;
- number of blocks: $n = 6, 10, 20$; number of treatment: $C = 3,5,7$;
- block effect β_i, $i = 1,...,n$, generated from a discrete uniform distribution with values $(-1, -0.5, 0, 0.5, 1)$;

- with reference to model (11.2), the treatment effects μ_j, $j = 1,..,C$, are set in Fig. 11.1.
- three type of random errors: normal, exponential (as an example of an asymmetric distribution) and Student's t with 2 degree of freedom (as an example of an heavy tailed distribution). The variability of random errors has been calibrated to the value of $\sigma = 2$, with the aim of properly reveal and compare the power among the considered procedures. Finally, in order to better represent a genuine ordinal scale, before being added to the true effects the random errors were rounded to the nearest integer.

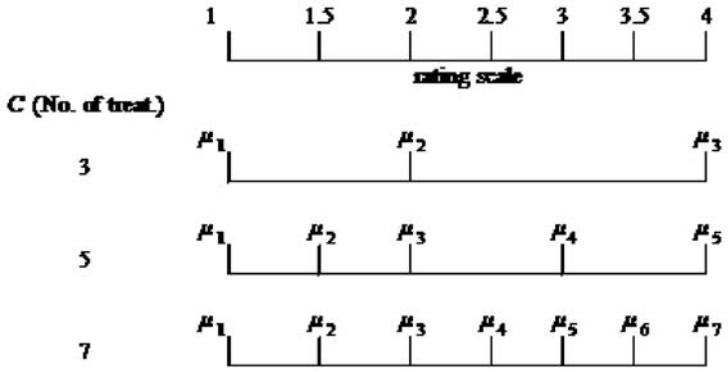

Fig. 11.1 Scheme of treatment effects for the simulation study.

For each simulation we performed the permutation tests (with 1,000 CMC), using the Fisher combining function, and we considered as counterparts the traditional F-test, the Friedman test and finally the Mean Aligned Rank (MAR) test proposed by O'Gorman (2001). The considered significance level was $\alpha = 0.05$. In case of rejection of the global null hypothesis H_{0k}, in order to perform the treatment pairwise comparisons, we considered permutation tests for two paired samples. Least Significant Difference (LSD) for the difference of mean ranks and t-tests as post-hoc procedures respectively for Friedman test and F-test and MAR have been considered as well. We recall that all post-hoc pairwise procedures should take into account for the problem of multiplicity (Westfall et al., 1999) hence they have to be well defined in order to maintain at the desired α-level the type I error probability of the main global hypothesis H_0. For this goal, for permutation tests we adopted a multiplicity correction strategy by using the closed testing approach (Marcus et al., 1976) via Tippett combining function (i.e. the so called minP procedure, Westfall et al., 1999) which is particularly suitable to be implemented within the framework of permutation tests (Finos and Salmaso, 2007), while for all other pairwise procedures we adopted the Bonferroni correction. Table 11.1 summarizes the obtained rejection rates ($\alpha = 0.05$). Note that, in order to be able to properly compare the

performances of the compared procedures with different values of C (i.e. no. of treatments), rejection rates of pairwise comparisons are presented in terms of *delta* (δ), that is of the true differences (in term of σ) between treatment effects, where delta is defined as

$$\delta_{jh} = \tau_j - \tau_h,\ j,h = 1,...,C,\ j \neq h.$$

For example we get $\delta = 1\sigma$ for $C = 3$ from the difference between μ_2 and μ_1, whereas we get $\delta = 1\sigma$ for $C = 5$ from the differences $\mu_3 - \mu_1$, $\mu_4 - \mu_3$, $\mu_5 - \mu_4$.

As first remark for the simulation study, we can observe that under null hypothesis all procedures appear to properly behave according to the nominal level. From a general point of view, as expected, the power for the global hypothesis increases when increasing the number of blocks and the number of active treatments. On the contrary, power for pairwise comparisons decreases from 3 to 5 treatments and slightly increases from 5 to 7. This is probably due to a drawback of the multiplicity correction strategy which is too much conservative.

Obviously, F-test shows a better behaviour under normality, but in case of exponential errors and particularly of Student's t errors, all nonparametric procedures show a greater power. Among nonparametric tests, the worst one is the Friedman test whereas a good behaviour is provided by the Mean Aligned Rank test. It should be noted that Friedman test is actually not satisfactory when data have ties as in case of ordered categorical variables we considered in this chapter. In fact, the continuity correction proposed by several authors is valid only asymptotically and for finite samples it does not provides a conservative test. Permutation test has an intermediate performance which is denoted by some strength and weakness aspects: it is particularly powerful when the number of treatments is not too high and the number of blocks is around ten. An advantage of the permutation method is that it can be easily extended to the multivariate case, i.e. when the response variable in multidimensional, by means of the nonparametric combination methodology (Pesarin, 2001).

11.5 Case study

In this section we face a real case study proposed in the literature. Suppose, as in Lamond (1970), p. 28, that we wish to compare the flavour of meat from three breeds of geese X, Y, and Z on a five point scale with categories ranging from "excellent" to "very poor" and that the data from eight consumers shown in Table 11.2 are obtained, where we have labelled the ordered categories as 1–5 scores.

When applying the considered RCB procedures to meat flavour data we can obtain results reported in Table 11.3, where we performed pairwise comparisons only if the global test had been rejected ($\alpha = 0.05$). Note that, in addition to the Fisher combining function, we considered here for the global test Tippett and Liptak combining functions (Pesarin, 2001).

Table 11.1 Rejection rates ($\alpha = 0.05$) and nominal levels (only for global test)

Test	n	H_1 (rejection rates)												H_0 (nominal level)		
		$C=3$				$C=5$				$C=7$				Glob. test		
			δ				δ				δ					
		Glob	1	2	3	Glob	1	2	3	Glob	1	2	3	$C=3$	$C=5$	$C=7$
							Normal errors									
	6	0.485	0.034	0.175	0.393	0.532	0.020	0.107	0.309	0.557	0.024	0.057	0.245	0.050	0.043	0.045
F	10	0.813	0.092	0.377	0.761	0.823	0.036	0.236	0.632	0.832	0.048	0.127	0.533	0.043	0.054	0.050
	20	0.982	0.191	0.730	0.977	0.993	0.100	0.614	0.968	0.996	0.140	0.357	0.941	0.047	0.056	0.060
	6	0.382	0.017	0.094	0.292	0.406	0.005	0.043	0.164	0.440	0.005	0.020	0.127	0.044	0.041	0.030
Friedman	10	0.701	0.032	0.220	0.609	0.721	0.009	0.116	0.454	0.736	0.018	0.058	0.363	0.049	0.048	0.048
	20	0.959	0.066	0.530	0.935	0.975	0.043	0.417	0.878	0.987	0.078	0.223	0.878	0.045	0.047	0.054
	6	0.415	0.058	0.206	0.365	0.442	0.025	0.107	0.277	0.475	0.022	0.055	0.206	0.051	0.052	0.038
Mean AR	10	0.714	0.104	0.370	0.657	0.744	0.033	0.202	0.536	0.751	0.040	0.105	0.455	0.050	0.052	0.058
	20	0.964	0.174	0.651	0.943	0.975	0.086	0.532	0.912	0.987	0.127	0.315	0.899	0.048	0.050	0.057
	6	0.311	0.018	0.083	0.185	0.347	0.007	0.031	0.060	0.363	0.008	0.016	0.048	0.030	0.033	0.032
Permutation	10	0.738	0.091	0.350	0.643	0.721	0.027	0.135	0.357	0.729	0.036	0.073	0.301	0.044	0.039	0.049
	20	0.985	0.259	0.761	0.973	0.988	0.113	0.566	0.935	0.991	0.158	0.329	0.909	0.048	0.055	0.047
							Exponential errors									
	6	0.567	0.057	0.252	0.489	0.571	0.024	0.123	0.375	0.584	0.030	0.069	0.310	0.040	0.045	0.046
F	10	0.815	0.104	0.390	0.756	0.812	0.049	0.253	0.620	0.846	0.058	0.141	0.563	0.046	0.051	0.046
	20	0.980	0.211	0.728	0.972	0.991	0.103	0.600	0.955	0.991	0.144	0.369	0.931	0.050	0.044	0.043
	6	0.554	0.018	0.092	0.460	0.597	0.005	0.054	0.326	0.637	0.009	0.027	0.261	0.035	0.027	0.029
Friedman	10	0.850	0.041	0.241	0.785	0.904	0.014	0.188	0.680	0.936	0.036	0.111	0.618	0.054	0.056	0.039
	20	0.997	0.139	0.657	0.993	1.000	0.050	0.603	0.987	1.000	0.134	0.400	0.979	0.049	0.036	0.048
	6	0.596	0.137	0.296	0.545	0.643	0.035	0.177	0.462	0.684	0.046	0.095	0.404	0.042	0.039	0.040
Mean AR	10	0.861	0.207	0.470	0.820	0.917	0.059	0.360	0.769	0.943	0.087	0.221	0.734	0.059	0.067	0.045
	20	0.997	0.325	0.840	0.995	1.000	0.140	0.767	0.993	1.000	0.222	0.538	0.990	0.053	0.043	0.052
	6	0.421	0.028	0.155	0.260	0.449	0.012	0.033	0.085	0.478	0.016	0.018	0.074	0.026	0.031	0.027
Permutation	10	0.900	0.156	0.486	0.792	0.914	0.058	0.253	0.545	0.940	0.086	0.155	0.520	0.056	0.056	0.043
	20	0.988	0.314	0.768	0.966	0.994	0.150	0.602	0.917	0.995	0.238	0.422	0.914	0.053	0.042	0.050
							Student's t errors									
	6	0.189	0.019	0.050	0.134	0.165	0.006	0.026	0.067	0.139	0.006	0.008	0.029	0.036	0.033	0.040
F	10	0.261	0.037	0.079	0.203	0.263	0.011	0.042	0.133	0.222	0.006	0.015	0.072	0.030	0.026	0.039
	20	0.478	0.052	0.195	0.395	0.430	0.017	0.080	0.237	0.379	0.012	0.033	0.144	0.030	0.025	0.053
	6	0.216	0.013	0.044	0.154	0.197	0.004	0.017	0.071	0.234	0.003	0.010	0.052	0.046	0.034	0.046
Friedman	10	0.352	0.031	0.101	0.254	0.397	0.008	0.049	0.183	0.388	0.007	0.022	0.115	0.039	0.034	0.051
	20	0.693	0.058	0.265	0.624	0.746	0.024	0.155	0.501	0.790	0.030	0.086	0.443	0.050	0.042	0.056
	6	0.241	0.049	0.091	0.192	0.232	0.011	0.045	0.116	0.273	0.013	0.022	0.086	0.051	0.043	0.054
Mean AR	10	0.372	0.068	0.155	0.294	0.438	0.018	0.082	0.231	0.422	0.014	0.034	0.162	0.047	0.042	0.061
	20	0.703	0.095	0.316	0.641	0.756	0.040	0.209	0.543	0.802	0.047	0.111	0.495	0.057	0.046	0.061
	6	0.155	0.018	0.029	0.089	0.146	0.006	0.014	0.029	0.172	0.007	0.007	0.022	0.029	0.028	0.039
Permutation	10	0.403	0.066	0.155	0.312	0.428	0.014	0.057	0.171	0.412	0.011	0.024	0.120	0.039	0.033	0.040
	20	0.550	0.096	0.266	0.458	0.559	0.032	0.134	0.329	0.593	0.037	0.071	0.300	0.039	0.045	0.051

It is interesting to observe that not all procedures agree to reject the global null hypothesis ($\alpha = 0.05$). Moreover, the use of different combining functions for permutation tests seems to provide decision rules which are potentially more or less powerful.

It can be proved that the combined permutation test obtained using Fisher, Liptak or Tippet combining functions are so called 'admissible' combination, i.e. it does not exist any other type of combination which is uniformly more powerful. Note that if several combining functions are admissible they are equivalent as well.

Table 11.2 Category ratings for meat flavour for three breeds of geese

Consumer	*X*	*Y*	*Z*
1	3	2	3
2	4	5	4
3	3	2	3
4	1	4	2
5	2	4	2
6	1	3	3
7	2	5	4
8	2	5	2

Table 11.3 Category ratings for meat flavour for three breeds of geese

Test		Pairwise comparisons		
	Global	*X* vs. *Y*	*X* vs. *Z*	*Y* vs. *Z*
F	0.028	0.026	0.675	0.292
Friedman	0.152	–	–	–
Mean AR	0.158	–	–	–
Permutation				
Fisher	0.049	0.048	0.235	0.113
Tippet	0.107	–	–	–
Liptak	0.019	0.026	0.256	0.103

11.6 Conclusions

In this chapter we have presented a combination-based permutation solution for hypothesis testing within the framework of randomized complete block design. The proposed solution may suggest to practitioners in the field of evaluation for educational services and quality of products an effective approach, especially when using ordered categorical variables, such as in the case of sensorial evaluations. As confirmed by the presented simulation study, the nonparametric tests are certainly good alternatives, in particular respect to the traditional parametric F and t test. In fact, even in case of normality, the power of permutation tests is nearly the same as that of the parametric tests, while in case of asymmetric or heavy tailed error distributions permutation tests can provide higher power. Hence, in each practical situation where the normality assumption is hard to justify, the proposed nonparametric procedure can be considered a valid solution.

Finally, as suggested by the real case study, a possible way to improve power of permutation tests is to better investigate the role of the combining functions. Note that our proposed permutation test applies a combining function two times: at first in order to combine the partial pairwise permutation tests to obtain a global test, then we apply a combining function in order to perform a suitable multiplicity correction strategy for pairwise permutation p-values.