2

Ordinal Data

2.1 Introduction

A categorical variable has a measurement scale consisting of a set of categories. Categorical variables that have ordered categories are called *ordinal* (Agresti, 2002). They appear, for example, whenever the condition of a patient cannot be measured by a metric variable and has to be classified or rated as "critical", "serious", "fair", or "good". The measurements on ordered categorical scales can be ordered by size, but the scales lack any algebraic structure; that is, the distances between categories are unknown. Although a patient categorized as "fair" is more healthy than a patient categorized as "serious", no numerical value describes how much more healthy that patient is.

Let X and Y denote two categorical variables, X with r categories and Y with c categories. The classification of n measurements on both variables has rc possible combinations, which can be represented in an $r \times c$ contingency table (see Table 2.1), where $\{m_{i,j}, i = 1, \ldots, r, j = 1, \ldots, c\}$ represents cell frequencies, with row and column margins $n_i = \sum_{j=1}^{c} m_{i,j}$ and $t_j = \sum_{i=1}^{r} m_{i,j}$, respectively.

Table 2.1. $r \times c$ contingency table.

	1	$\cdots$	j	$\cdots$	c	
1	$m_{1,1}$	$\cdots$	$m_{1,j}$	$\cdots$	$m_{1,c}$	n_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	$m_{i,1}$	$\cdots$	$m_{i,j}$	$\cdots$	$m_{i,c}$	n_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
r	$m_{r,1}$	$\cdots$	$m_{r,j}$	$\cdots$	$m_{r,c}$	n_r
	t_1	$\cdots$	t_j	$\cdots$	t_c	n

D. Basso et al., *Permutation Tests for Stochastic Ordering and ANOVA*, Lecture Notes in Statistics, 194, DOI 10.1007/978-0-387-85956-9_2,

There is a large body of literature concerning the analysis of categorical data for which the row and column variables are ordinal measurements. In recent years, statisticians increasingly have recognized that many benefits can result from using methods that take into account orderings among categories in contingency tables. One way to utilize ordered categories is to assume inequality constraints on parameters for those categories that describe dependence structure.

In many applications, one would typically expect larger values of Y to be associated with larger values of X. One can describe the positive dependence of the discrete bivariate distribution of (X, Y) using various types of odds ratios, referred to as generalized odds ratios. Three of them, which are the most commonly used in application problems, are defined below. Let $i = 1, \ldots, r-1$, $j = 1, \ldots, c-1$. Then:

1. *Local odds ratios*:

$$\theta_{i,j}^{L} = \frac{\Pr(Y = j|X = i)\Pr(Y = j+1|X = i+1)}{\Pr(Y = j|X = i+1)\Pr(Y = j+1|X = i)}.$$

2. *Cumulative odds ratios*:

$$\theta_{i,j}^{C} = \frac{\Pr(Y \le j|X = i)\Pr(Y > j|X = i+1)}{\Pr(Y \le j|X = i+1)\Pr(Y > j|X = i)}.$$

3. *Global odds ratios*:

$$\theta_{i,j}^{G} = \frac{\Pr(Y \le j|X \le i)\Pr(Y > j|X > i)}{\Pr(Y \le j|X > i)\Pr(Y > j|X \le i)}.$$

These definitions show that generalized odds ratios are odds ratios for 2×2 tables obtained from the $r \times c$ table by collapsing adjacent categories if necessary, as displayed in Figure 2.1.

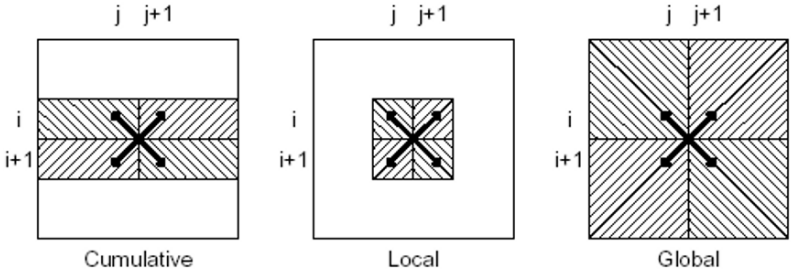


Fig. 2.1. Generalized odds ratios.

If $\boldsymbol{\theta}$ denotes the column vector of any one of the generalized odds ratios, then $\boldsymbol{\theta} = \mathbf{1}_{(r-1)(c-1)}$ if and only if X and Y are independent, where $\mathbf{1}_k$ is

used to denote a k-dimensional column vector of ones. Relationships among the positive dependencies considered are

$$\boldsymbol{\theta}_1^L \gneq \mathbf{1}_{(r-1)(c-1)} \Rightarrow \boldsymbol{\theta}_1^C \gneq \mathbf{1}_{(r-1)(c-1)} \Rightarrow \boldsymbol{\theta}_1^G \gneq \mathbf{1}_{(r-1)(c-1)},$$

where $\boldsymbol{\theta} \gneq \mathbf{1}_{(r-1)(c-1)}$ means that $\theta_{i,j} \geq 1$, $i = 1, \ldots, r-1$, $j = 1, \ldots, c-1$, with at least one strict inequality.

To introduce the notation to be adopted throughout, suppose that the $r \times c$ table arises from the comparison of r increasing levels of a treatment for which the *response* variable Y is ordinal with c categories. The c levels can be thought of as c ordinal categories of an *explanatory* variable X. A primary aim of many studies is to compare conditional distributions of Y at various levels of the explanatory variable X. For $i = 1, \ldots, r$ and $j = 1, \ldots, c$, let

$$\pi_i(j) = \Pr(Y = j | X = i), \qquad F_i(j) = \sum_{l=1}^{j} \pi_i(l),$$

be the probability mass function (p.m.f.) and cumulative distribution function (c.d.f.) of Y_i, respectively, where Y_i denotes a random variable whose distribution is the conditional distribution of Y given $X = i$.

One of the earliest definitions of stochastic ordering was given by Lehmann (1955).

Definition 2.1. *The random variable Y_{i+1} is said to dominate Y_i according to the simple stochastic ordering, or Y_{i+1} is stochastically larger than Y_i, written $Y_i \overset{st}{\leq} Y_{i+1}$, if*

$$F_i(j) \geq F_{i+1}(j), \qquad j = 1, \ldots, c-1.$$

In some cases, a pair of distributions may satisfy a stronger condition.

Definition 2.2. *The random variable Y_{i+1} is said to dominate Y_i according to the likelihood ratio ordering, written $Y_i \overset{lr}{\leq} Y_{i+1}$, if*

$$\pi_i(j)\pi_{i+1}(j+1) \geq \pi_i(j+1)\pi_{i+1}(j), \quad j = 1, \ldots, c-1.$$

Clearly, if $\boldsymbol{\theta}^L \geq \mathbf{1}_{(r-1)(c-1)}$ and $\boldsymbol{\theta}^C \geq \mathbf{1}_{(r-1)(c-1)}$, then the rows satisfy the likelihood ratio ordering $Y_1 \overset{lr}{\leq} \ldots \overset{lr}{\leq} Y_r$ and the simple stochastic ordering $Y_1 \overset{st}{\leq} \ldots \overset{st}{\leq} Y_r$, respectively (and vice versa). Each of these constraints defines a form of monotone order of the rows involving only two rows at a time. By contrast, the constraint $\boldsymbol{\theta}^G \geq \mathbf{1}_{(r-1)(c-1)}$ define a monotone relationship that involves more than two rows at a time; this constraint is equivalent to the notion of positive quadrant dependence (PQD, Lehmann, 1955).

Definition 2.3. *We shall say that the pair (X, Y) is positive quadrant dependent if*

$$\Pr(Y \leq j | X \leq i) \geq \Pr(Y \leq j | X > i) \quad \forall\, i, j.$$

The reason why PQD is a positive dependence concept is that X and Y are more likely to be large together or to be small together compared with X' and Y', where $X \stackrel{d}{=} X'$, $Y \stackrel{d}{=} Y'$, and X' and Y' are independent.

In many applications in which it is believed that certain constraints on the distributions exist, it is reasonable to assume a *stochastic ordering*. The statistical information arising from these constraints, if properly incorporated, makes the statistical inference more efficient than its counterparts, wherein such constraints are ignored.

In this chapter, several data examples are presented to motivate and to provide an overview of the topics.

2.2 Testing Whether Treatment is "Better" than Control: $2 \times c$ Contingency Tables

Patefield (1982) reported the results of a double-blind study concerning the use of Oxprenolol in the treatment of examination stress. Thirty-two students were entered in the study: fifteen were treated with Oxprenolol (treatment) and seventeen were given Diazepam (control). The examination grades were compared with their tutor's prediction; the results are given in Table 2.2.

Table 2.2. Examination results compared with tutor's predictions

		Worse	Same	Better	
		1	2	3	Total
Control	1	6	11	0	17
Treatment	2	2	8	5	15
		8	19	5	32

For this example, one wishes to test the null hypothesis that the treatment and control effects are the same against the one-sided alternative that treatment is in some sense"better" than the placebo. One obstacle to the development of suitable tests is that it is often difficult to be specific as to the notion of "better" (Cohen and Sackrowitz, 2000). In fact, a precise definition of "better" and hence a precise definition of an alternative hypothesis is often not even mentioned in instances where such a testing problem is encountered. In contrast, Cohen et al. (2000) offered various formal definitions of "better" in order to improve understanding of the alternative hypothesis.

Among the formally defined notions for a $2 \times c$ table, the less stringent is the simple stochastic order. Let the testing problem be

$$H_0 : Y_1 \stackrel{d}{=} Y_2 \Leftrightarrow \boldsymbol{\theta}^C \in \Theta_0 = \{\boldsymbol{\theta}^C : \boldsymbol{\theta}^C = \mathbf{1}_{c-1}\}, \tag{2.1}$$

where "$\stackrel{d}{=}$" means "equal in distribution", against the one-sided alternative

$$H_1 : Y_1 \stackrel{st}{\lesssim} Y_2 \Leftrightarrow \boldsymbol{\theta}^C \in \Theta_1 = \{\boldsymbol{\theta}^C : \boldsymbol{\theta}^C \gtrsim \mathbf{1}_{c-1}\}. \tag{2.2}$$

Suppose that a test rejects H_0. Then it does follow that there is sufficient statistical evidence to support the claim that there is a difference between the treatment and the control, but it does not follow that there is statistical evidence to accept that the treatment is better than the control (H_1 is true). However, if we make the prior assumption that

the treatment is at least as good as the control

(that is, either H_0 or H_1 is true), then the rejection of "no difference between the treatment and the control" together with the prior assumption that "the treatment is at least as good as the control" (that is, $Y_1 \stackrel{st}{\lesssim} Y_2 \Leftrightarrow \boldsymbol{\theta}^C \in \Theta_0 \cup \Theta_1$) would lead to the conclusion that the treatment is better than the control (Silvapulle and Sen, 2005). In other words, we consider a model specifying the distribution of Y to be stochastically ordered with respect to the value of the explanatory variable X; that is, $\{Y_x, Y_{x'} : Y_x \stackrel{st}{\lesssim} Y_{x'} \text{ if } x < x'\}$.

2.2.1 Conditional Distribution

Let $\boldsymbol{M}_i = (M_{i,1}, \ldots, M_{i,c})$, $i = 1, 2$, be independent random vectors having multinomial distributions with cell probabilities $\boldsymbol{\pi}_i = (\pi_i(1), \ldots, \pi_i(c))$. Under the product multinomial model (that is, given the row totals $\boldsymbol{n} = (n_1, n_2)$), when the null hypothesis H_0 is true, the column total $\boldsymbol{t} = (t_1, \ldots, t_c)$ is a completely sufficient statistic (row and column total in the full multinomial model). Testing is carried out by conditioning on row and column totals. This allows the two models, product multinomial and full multinomial, to be treated simultaneously since the conditional distributions are the same (Cohen and Sackrowitz, 2000).

Here, for convenience of notation, we drop the row index i from $\theta_{i,j}$. The conditional distribution of $(M_{1,1}, \ldots, M_{1,c-1})$ given the row and column totals $(\boldsymbol{n}, \boldsymbol{t})$ is the *multivariate noncentral hypergeometric distribution* with p.m.f.

$$\begin{aligned}
&\Pr_{\boldsymbol{\theta}^L}(M_{1,1} = m_{1,1}, \ldots, M_{1,c-1} = m_{1,c-1} | \boldsymbol{n}, \boldsymbol{t}) \\
&= \frac{\binom{t_1}{m_{1,1}} \cdots \binom{t_j}{n_1 - \sum_{j=1}^{c-1} m_{1,j}} \prod_{j=1}^{c-1} \left(\prod_{l=j}^{c-1} \theta_l^L \right)^{m_{1,j}}}{\sum_{(m_{1,1}, \ldots, m_{1,c-1}) \in \mathcal{M}} \binom{t_1}{m_{1,1}} \cdots \binom{t_j}{n_1 - \sum_{j=1}^{c-1} m_{1,j}} \prod_{j=1}^{c-1} \left(\prod_{l=j}^{c-1} \theta_l^L \right)^{m_{1,j}}},
\end{aligned}$$

where $m_{1,j} \geq 0$, $j = 1, \ldots, c-1$, and

$$\mathcal{M} = \left\{ (m_{1,1}, \ldots, m_{1,c-1}) : n_1 - \sum_{j=1}^{c-1} m_{1,j} \geq 0; t_j - m_{1,j} \geq 0; \right. \tag{2.3}$$
$$\left. n + \sum_{j=1}^{c-1} m_{1,j} - \sum_{j=1}^{c-1} t_j - n_1 \right\}.$$

Note that the conditional distribution has a simple exponential family form that depends only on the natural parameters $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_{(c-1)})$, where

$$\nu_j = \log\left(\frac{\pi_1(j)\pi_2(c)}{\pi_2(j)\pi_1(c)}\right) = \log\left(\prod_{l=j}^{c-1} \theta_l^L\right).$$

As the conditional distribution depends on $(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2)$ only through $\boldsymbol{\nu}$, the conditional hypotheses $H_0^{|(\boldsymbol{n},\boldsymbol{t})}$ and $H_1^{|(\boldsymbol{n},\boldsymbol{t})}$ must be formulated in terms of $\boldsymbol{\nu}$. We have $H_0 : \boldsymbol{\theta}^C = \mathbf{1}_{c-1} \Leftrightarrow H_0^{|(\boldsymbol{n},\boldsymbol{t})} : \boldsymbol{\nu} = \mathbf{0}_{c-1}$, but $H_1 : \boldsymbol{\theta}^C \gtrsim \mathbf{1}_{c-1} \Rightarrow H_1^{|(\boldsymbol{n},\boldsymbol{t})} : \nu_1 > 0$ (or the first nonzero element of $\boldsymbol{\nu}$ when $c > 3$). In contrast, for the likelihood order alternative, $H_1 : \boldsymbol{\theta}^L \gtrsim \mathbf{1}_{c-1} \Leftrightarrow H_1^{|\boldsymbol{n},\boldsymbol{t}} : \boldsymbol{\nu} \gtrsim \mathbf{0}_{c-1}$.

In Patefield's example, there are 54 table configurations with the same marginal totals as Table 2.2. The conditional probabilities are given by

$$\Pr_{\theta_1^L, \theta_2^L}(M_{1,1} = m_{1,1}, M_{1,2} = m_{1,2} | (17, 15), (8, 19, 5))$$
$$= \frac{\binom{8}{m_{1,1}}\binom{19}{m_{1,2}}\binom{5}{17-m_{1,1}-m_{2,2}}(\theta_1^L\theta_2^L)^{m_{1,1}}(\theta_2^L)^{m_{1,2}}}{\sum_{m_{1,1}=0}^{8}\sum_{m_{1,2}=12}^{17}\binom{8}{m_{1,1}}\binom{19}{m_{1,2}}\binom{5}{17-m_{1,1}-m_{2,2}}(\theta_1^L\theta_2^L)^{m_{1,1}}(\theta_2^L)^{m_{1,2}}},$$

where $(m_{1,1}, m_{1,2}) \in \mathcal{M} = \{(m_{1,1}, m_{1,2}) : m_{1,1} = 0, \ldots, 8; m_{1,1} + m_{1,2} = 12, \ldots, 17\}$.

2.2.2 Linear Test Statistics: Choice of Scores

A popular class of linear test statistics is based on the explicit or implicit assignment of scores to the c categories (Graubard and Korn, 1987).

Gail (1974) called a "value system" any real function on the sample space of a multinomial random variable. Consider a nondecreasing real function of Y_i, $w(\cdot) : \{1, \ldots, c\} \to \mathbb{R} : -\infty < w(1) \leq \ldots \leq w(c) < \infty$. Denote by $w(j) := w_j$ the score attached to the jth category and by $\boldsymbol{w} = (w_1, \ldots, w_c)$ the vector of scores, and observe that $\mathrm{E}[w(Y_i)] = \sum_{j=1}^{c} w_j \pi_i(j)$. Thus, the hypotheses are

$$H_0^{\boldsymbol{w}} : \mathrm{E}[w(Y_1)] = \mathrm{E}[w(Y_2)] \Leftrightarrow \sum_{j=1}^{c-1}(w_{j+1} - w_j)(F_1(j) - F_2(j)) = 0$$

against

$$H_1^{\boldsymbol{w}} : \mathrm{E}\left[w(Y_1)\right] < \mathrm{E}\left[w(Y_2)\right] \Leftrightarrow \sum_{j=1}^{c-1} (w_{j+1} - w_j)\left(F_1(j) - F_2(j)\right) > 0,$$

by which simplicity is achieved by reducing the multiparameter inference problem to one that involves only a scalar parameter. While this simplifies the testing problem, it could also be expected to have low power at points away from the chosen (through $\boldsymbol{w}$) direction in the alternative space. Note that because of our prior assumption $\{F_1(j) \geq F_2(j), j = 1, \ldots, c-1\}$, H_0 in (3.1) implies $H_0^{\boldsymbol{w}}$ and $H_1^{\boldsymbol{w}}$ implies H_1 in (3.2), and when $w_1 < \ldots < w_c$, also $H_0^{\boldsymbol{w}} \Rightarrow H_0$ and $H_1 \Rightarrow H_1^{\boldsymbol{w}}$.

The class of linear test statistics based on $\boldsymbol{w}$ is given by

$$T_{\boldsymbol{w}} = \frac{\left(\frac{(n-2)n_1n_2}{n}\right)^{\frac{1}{2}} \left(\sum_{j=1}^{c} \frac{m_{2,j}w_j}{n_2} - \sum_{j=1}^{c} \frac{m_{1,j}w_j}{n_1}\right)}{\left(\sum_{j=1}^{c}(w_j)^2 t_j - \frac{1}{n_1}(\sum_{j=1}^{c} w_j m_{1,j})^2 - \frac{1}{n_2}(\sum_{j=1}^{c} w_j m_{2,j})^2\right)^{\frac{1}{2}}};$$

that is, the usual two-sample t statistic based on assigning a set of scores to the c categories. It is straightforward to show that, for any linear transformation of scores that preserves the monotonicity, $T_{a\mathbf{1}+b\boldsymbol{w}} = T_{\boldsymbol{w}}$ with $a \in \mathcal{R}, b \in \mathcal{R}^+$. Hence, we may consider standardized scores $\boldsymbol{w}$ obtained by transforming the original scores via $a = -w_1/(w_c - w_1)$ and $b = 1/(w_c - w_1)$ to the $[0, 1]$ interval. Permutationally equivalent formulations of $T_{\boldsymbol{w}}$ are

- Graubard and Korn (1987): $T_{\boldsymbol{w}} = \sum_{j=1}^{c} m_{2,j}w_j$,
- goodness of fit statistics: $T_{\boldsymbol{w}} = \sum_{k=1}^{c-1}(w_{j+1} - w_j)\left(\hat{F}_1(j) - \hat{F}_2(j)\right)$, and
- Mantel (1963): $T_{\boldsymbol{w}} = (n-1)^{\frac{1}{2}}\hat{\rho}$,

where $\hat{F}_i(j) = (\sum_{l=1}^{j} m_{i,j})/n_i$ denotes the empirical c.d.f. of the ith group and $\hat{\rho}$ the Pearson correlation coefficient based on the scores $\boldsymbol{w}$ and values 0 and 1 assigned to the control and the treatment, respectively.

Widely used scoring systems in data analysis include equal-spacing scores $\boldsymbol{w} = (1, \ldots, c)$, midrank scores $\boldsymbol{w} = (\bar{r}_1, \ldots, \bar{r}_c)$, where $\bar{r}_1 = \frac{t_1+1}{2}$ and $\bar{r}_j = \sum_{l=1}^{j-1} t_l + \frac{t_j+1}{2}$, $j = 2, \ldots, c$, and Anderson-Darling scores $(w_{j+1} - w_j) = 1/(\hat{F}(j)(1 - \hat{F}(j)))^{1/2}$, $j = 1, \ldots, c$, where $\hat{F}(j) = [n_1\hat{F}_1(j) + n_2\hat{F}_2(j)]/n$.

The use of midrank scores seems appealing since it yields to the Wilcoxon-Mann-Whitney (WMW) test statistic. This test statistic can also be viewed as Spearman's correlation coefficient between X and Y. However, midrank scores do not necessarily provide distances between categories that correspond to a "reasonable" metric. In particular, for highly unbalanced response frequencies, adjacent categories having relatively few observations necessarily have similar midrank scores. For example, suppose few subjects fall in the first categories on the scale "bad", "fair", "good", "very good", "excellent"; mid-ranks then have similar scores for categories "bad" and "good".

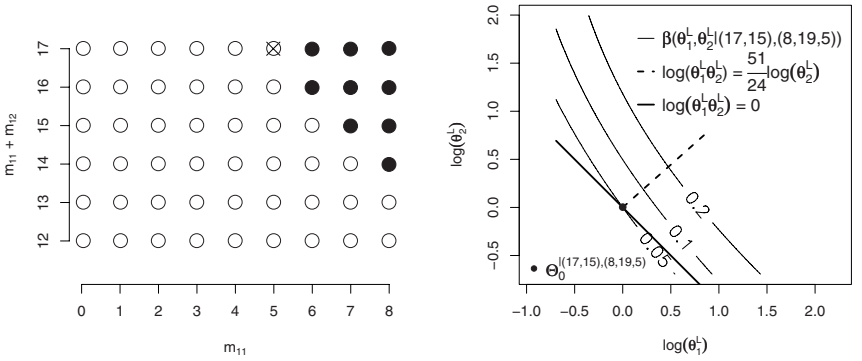
For Patefield's data, the critical function of the randomized WMW test based on the statistic $T_{(0,27/51,1)}$ is

$$\phi_{T_{(0,27/51,1)}} = \begin{cases} 1 & \bullet \text{ if } 27m_{1,1} + 24(m_{1,1} + m_{2,2}) > 543 \\ 0.53 & \otimes \text{ if } 27m_{1,1} + 24(m_{1,1} + m_{2,2}) = 543 \\ 0 & \circ \text{ if } 27m_{1,1} + 24(m_{1,1} + m_{2,2}) < 543 \end{cases},$$

and the rejection region at a significance level of $\alpha = 0.05$ is given in Figure 2.2 (a). The power of the randomized WMW test ϕ_T given $\boldsymbol{n} = (17, 15), \boldsymbol{t} = (8, 19, 5)$, the *conditional power*, is given by

$$\begin{aligned} \beta(\theta_1^L, \theta_2^L | (17,15),(8,19,5)) &= \mathrm{E}_{\theta_1^L,\theta_2^L}\left[\phi_{T_{(0,27/51,1)}} | (17,15),(8,19,5)\right] \\ &= \Pr_{\theta_1^L,\theta_2^L}(27M_{1,1} + 24(M_{1,1} + M_{1,2}) > 543 | (17,15),(8,19,5)) \\ &\quad + 0.53 \Pr_{\theta_1^L,\theta_2^L}(27M_{1,1} + 24(M_{1,1} + M_{1,2}) = 543 | (17,15),(8,19,5)) \end{aligned}$$

for $(\theta_1^L, \theta_2^L) : \boldsymbol{\nu} = (\log(\theta_1^L\theta_2^L), \log(\theta_2^L))^t \in \Theta_0^{|(\boldsymbol{n},\boldsymbol{t})} \cup \Theta_1^{|(\boldsymbol{n},\boldsymbol{t})}$, where the conditional null and alternative parameter spaces are $\Theta_0^{|(\boldsymbol{n},\boldsymbol{t})} : \{\boldsymbol{\nu} : \log(\theta_1^L\theta_2^L) = 0, \log(\theta_2^L) = 0\}$ and $\Theta_1^{|(\boldsymbol{n},\boldsymbol{t})} : \{\boldsymbol{\nu} : \log(\theta_1^L\theta_2^L) > 0\}$, respectively. The conditional power is depicted in Figure 2.2 (b) as a function of $(\log(\theta_1^L), \log(\theta_2^L))$.



(a) 0.047 level rejection region (nonrandomized test)

(b) Contour plot of the conditional power function for the WMW test

Fig. 2.2. WMW test.

Observe that:

- The WMW test is conditionally biased; that is, $\beta(\theta_1^L, \theta_2^L | (17,15),(8,19,5)) < \alpha = 0.05$ for some $(\theta_1^L, \theta_2^L) : \boldsymbol{\nu} \in \Theta_1^{|(\boldsymbol{n},\boldsymbol{t})}$ (see also Berger and Ivanova, 2002b). Note that for the likelihood ratio conditional alternative $H_1^{|\boldsymbol{n},\boldsymbol{t}} : \boldsymbol{\nu} = (\log(\theta_1^L\theta_2^L), \log(\theta_2^L))^t \gneq (0,0)^t$, we have $\Theta_0^{|(\boldsymbol{n},\boldsymbol{t})} \cup \Theta_1^{|(\boldsymbol{n},\boldsymbol{t})} = \{\boldsymbol{\nu} :$

$\log(\theta_1) \geq 0, \log(\theta_2) \geq 0\}$. Cohen and Sackrowitz (1991) showed that the WMW test (among others) is conditionally (and hence unconditionally) unbiased.

- The WMW test is conditionally Bayes with respect to a prior putting all its mass on the set $\{\boldsymbol{\nu} \in \Theta_1^{|(\boldsymbol{n},\boldsymbol{t})} : \log(\theta_1^L \theta_2^L) = \frac{51}{24} \log(\theta_2^L)\}$; that is, the WMW test is very powerful for alternatives near this direction (Cohen and Sackrowitz, 1998, see).

Often it is unclear how to assign scores because the power of the test depends on them. Indeed, the permutation tests listed in Chapter 7 of the StatXact-8 User Manual can be with general scores or with MERT scores (see Section 7.13). Podgor et al. (1996) consider a robust test from several test statistics $T_{\boldsymbol{w}}$ based on different sets of scores. The maximin efficient robust test (MERT) has the property of maximin efficiency in that its lowest asymptotic efficiency relative to each of the possible tests is higher than the lowest such efficiency for any other statistic based on any set of scores. The MERT considers a linear combination of the pair of test statistics with minimum correlation. However, Podgor's MERT is itself a linear rank test, which is ironic since it was proposed to correct the weaknesses of the class of linear rank tests (Berger and Ivanova, 2002a).

To handle the ambiguities arising from the choice of scoring, Kimeldorf et al. (1992) obtained the minimum and the maximum of the $T_{\boldsymbol{w}}$ test statistic over all possible assignments of nondegenerate nondecreasing scores $\boldsymbol{w}$. If the range of min and max values does not include the critical value of the test statistic (they term this case "nonstraddling"), then it can be immediately concluded that the results of the analysis remain the same no matter the choice of increasing scores used. However, if the range includes the critical value (termed the "straddling" case), the choice of scores used in the analysis must be carefully justified.

The scores $\boldsymbol{w}^{\max}$ that maximize $T_{\boldsymbol{w}}$ can be found by considering two cases:

- If $\hat{F}_2 \geq \hat{F}_1$, $\boldsymbol{w}^{\max}$ is one of the $c-1$ *monotone extreme points*

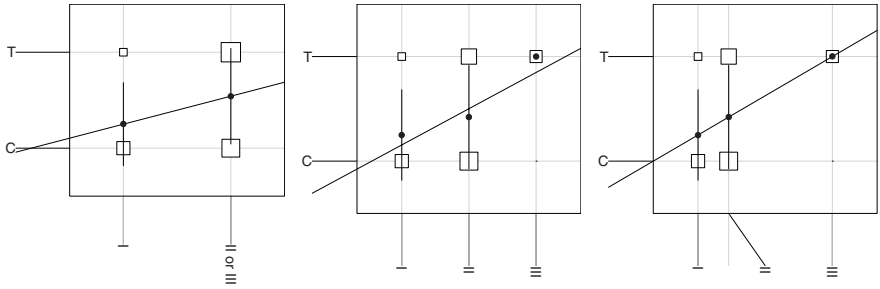
$$\begin{cases} w_l^{\max} = 0, 1 \leq l \leq j \\ w_l^{\max} = 1, j+1 \leq l \leq c \end{cases} \quad j = 1, \ldots, c-1.$$

- Otherwise, $w_j^{\max}$, $j = 1, \ldots, c-1$, are given by the *isotonic regression* of $m_{2,j}/t_j$ with weights t_j, denoted by $P_{\boldsymbol{t}}\left(\frac{\boldsymbol{m}_2}{\boldsymbol{t}} | \mathcal{I}\right)$, the solution that minimizes the weighted sum of squares

$$\min_{\boldsymbol{w} \in \mathcal{I}} \sum_{j=1}^{c} \left(\frac{m_{2,j}}{t_j} - w_j \right)^2 t_j;$$

that is, the weighted least squares projection of $\boldsymbol{m}_2/\boldsymbol{t}$ onto the closed convex cone $\mathcal{I} = \{\boldsymbol{w} \in I\!R^c : w_1 \leq \ldots \leq w_c\}$ with weights $\boldsymbol{t}$. The simple and elegant pool adjacent violators algorithm (PAVA) can be used (see Robertson et al., 1988).

For Patefield's data, the empirical distribution of the treatment can be shown to be stochastically larger than the empirical distribution of the control; that is, $\hat{F}_2(j) \leq \hat{F}_1(j)$, $j = 1, 2, 3$. The scores that minimize and maximize $T^{\boldsymbol{w}}$ are $\boldsymbol{w}^{\min} = (0, 1, 1)$ and $\boldsymbol{w}^{\max} = (0, 0.228, 1)$, respectively, with corresponding p-values of 0.1534 and 0.0052. We find that there are, in this straddling case, some scores that produce significance and some others that do not. A graphical representation in terms of $T_{\boldsymbol{w}} = \hat{\rho}$ is given in Figure 2.3.



(a) $w_2^{\min} = 1.0$, $\hat{\rho} = 0.253$ (b) $w_2 = 0.5$, $\hat{\rho} = 0.437$ (c) $w_2^{\max} = 0.23$, $\hat{\rho} = 0.480$

Fig. 2.3. Correlation coefficients as a function of $\boldsymbol{w}$.

However, we cannot consider $T_{\boldsymbol{w}}$ with fixed scores $\boldsymbol{w}^{\max} = (0, 0.23, 1)$ because data-snooping bias arises. Gross (1981) suggested that an "analysis based on [...] data-dependent scores may yield procedures that compare favorably to fixed-scores procedures". An adaptive test (Hogg, 1974; Berger and Ivanova, 2002a) based on the test statistic

$$T_{\max} = T_{\boldsymbol{w}^{\max}} = \max(\hat{\rho} : \boldsymbol{w} \in \mathcal{I}, w_1 < w_c) \tag{2.4}$$

can be constructed by computing the data-dependent scores $\boldsymbol{w}^{\max}$ at each permutation of the data. For instance, $\boldsymbol{w}^{\max} = (0, 0.41, 1)$ from the contingency table $\{(6, 10, 1); (2, 9, 4)\}$ obtained by exchanging one control value from "same" to "better" and one treatment value from "better" to "same" (that is, an arbitrary bth permutation of the data), obtaining $T^*_{\max}(b) = 0.288$.

For Patefield's data, the permutation distributions of $T_{\boldsymbol{w}} = \hat{\rho}$ by using different scoring systems are displayed in Figure 2.4, and results are given in Table 2.3. We can see that the permutation distributions of $T_{(0,.5,1)}$ and $T_{\max}$ are rather discrete (there were only 12 different realized values out of 54), making the tests automatically more conservative. Possible solutions are to make use of the mid-p-value (Lancaster, 1961) or a backup statistic (see Cohen et al., 2003).

The critical function of the randomized adaptive test is

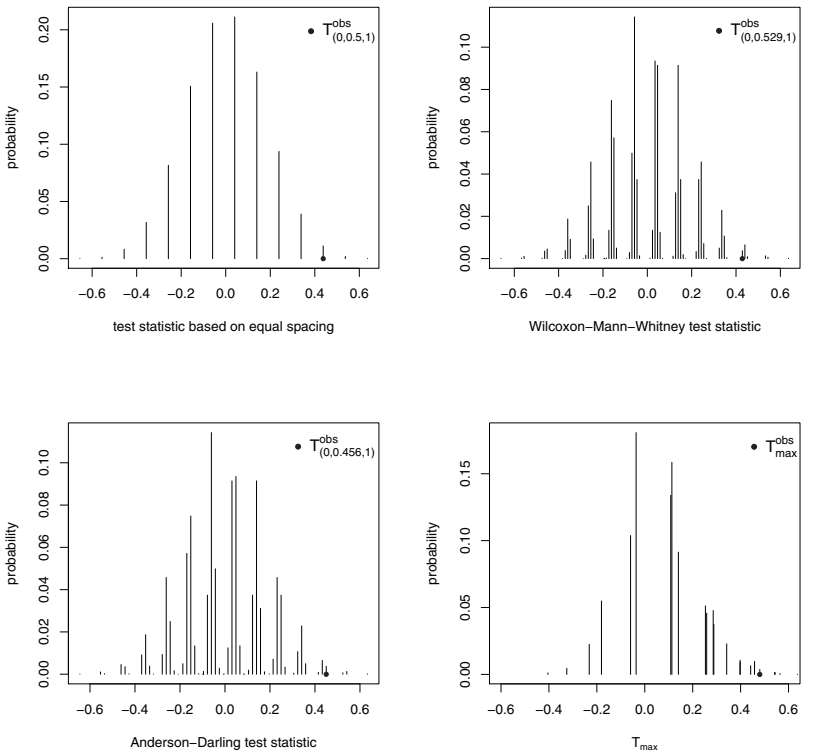


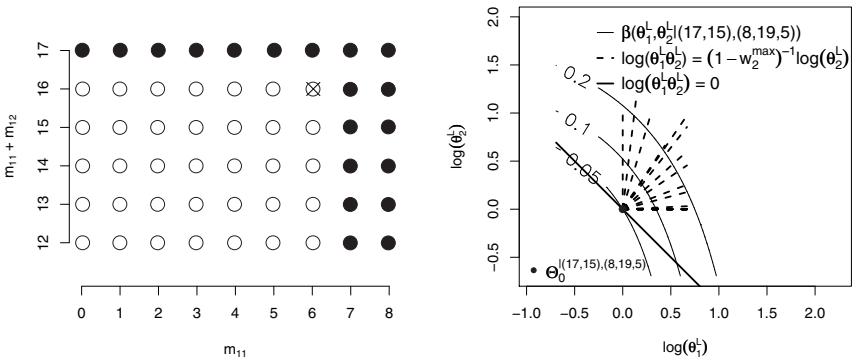
Fig. 2.4. Permutation distributions of $T_{\boldsymbol{w}} = \hat{\rho}$.

Table 2.3. Results for Patefield's data.

Value System	$\boldsymbol{w}$	$T_{\boldsymbol{w}}^{obs}$	p^{obs}
equal spacing	(0,0.500,1)	0.438	0.0133
midranks	(0,0.529,1)	0.428	0.0133
Anderson-Darling	(0,0.456,1)	0.450	0.0059
adaptive	$\boldsymbol{w}^{\max}$	0.480	0.0073

$$\phi_{T_{\max}} = \begin{cases} 1 & \text{if } T_{\max} > .342 \\ 0.28 & \text{if } T_{\max} = .342 \\ 0 & \text{if } T_{\max} < .342 \end{cases}.$$

The rejection region and the conditional power function of the adaptive test are given in Figures 2.5 (a) and (b), respectively. The adaptive test can be viewed as conditionally Bayes with respect to a prior putting its mass on the



(a) 0.043 level rejection region (nonrandomized test)

(b) Contour plot of the conditional power function for the adaptive test

Fig. 2.5. Adaptive test.

set $\{\boldsymbol{\nu} \in \Theta_1^{|(\boldsymbol{n},\boldsymbol{t})} : \log(\theta_1^L\theta_2^L) = \frac{1}{(1-w_2^{\max})}\log(\theta_2^L)\}$, where $w_2^{\max}$ are given in Table 2.4.

Table 2.4. Values of $w_2^{\max}$.

$w_2^{\max}$	0	0.108	0.228	0.398	0.407	0.438	0.526	0.594	0.723	0.789	0.955	1
prob.	0.343	0.037	0.003	0.001	0.022	0.091	0.001	0.006	0.001	0.045	0.0106	0.435

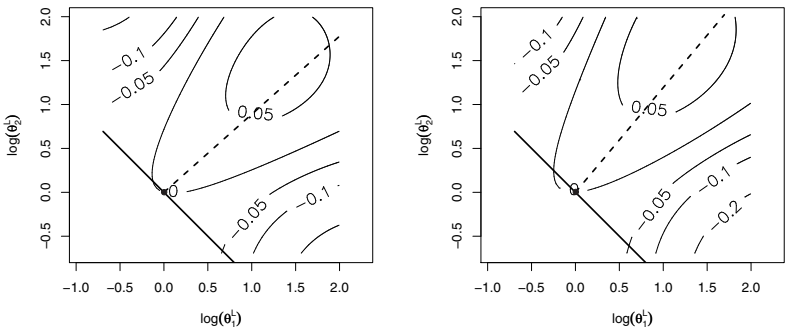
We can see from the difference between conditional powers in Figure 2.6 that the adaptive test distributes the power more uniformly over the entire alternative space.

The power of $T_{\boldsymbol{w}}$ given $\boldsymbol{n} = (17, 15)$ (i.e., the *unconditional power* for fixed sample sizes) is given by

$$\beta(\boldsymbol{\theta}^C|\boldsymbol{n} = (17,15)) = \mathrm{E}_{\boldsymbol{\theta}^C}\left[\phi_{T_{\boldsymbol{w}}}|\boldsymbol{n} = (17,15)\right], \quad \forall\, \boldsymbol{\theta}^C = (\theta_1^C, \theta_2^C)^t \in \Theta_0 \cup \Theta_1.$$

We replicate the unconditional power study in Cohen and Sackrowitz (2000) by using the algorithm of Patefield (1981) for generating all the possible tables. We compare the empirical power of Cramér-von Mises (equal spaced scores), WMW, Anderson-Darling, and adaptive tests with the most powerful test (Berger, 1998), which is based on $\lambda = (\nu_1 - \nu_2)/\nu_1 \in \mathcal{R}$, and it can be expressed as

$$T_{\boldsymbol{w}} : \begin{cases} \boldsymbol{w} = (-\lambda/(1-\lambda), 0, 1) & \text{if } \lambda < 0 \\ \boldsymbol{w} = (0, \lambda, 1) & \text{if } \lambda \in [0,1] \\ \boldsymbol{w} = (0, 1, 1/\lambda) & \text{if } \lambda > 1 \end{cases}.$$



(a) WMW minus $T_{\max}$ (b) Anderson-Darling minus $T_{\max}$

Fig. 2.6. Difference between conditional powers.

Simulations are based on 2000 Monte Carlo replicates with a nominal level of $\alpha = 0.05$, where the row probabilities and the stochastic order relationship (i.e., likelihood ratio "lr", hazard ratio "hr", and simple stochastic order "st") are given in Table 2.5.

2.2.3 Applications with R functions

Here we provide assistance in doing the statistical tests illustrated in Subsection 2.2.2 using the `R` language. Create Patefield's data with equal-spaced scores $(w_1, w_2, w_3) = (1, 2, 3)$, where `X` represents class labels and `Y` the vector of data,

```
> X <- c(rep(1,17),rep(2,15))
> Y <- c(rep(1,6),rep(2,11),rep(1,2),rep(2,8),rep(3,5))
```

and obtain w_2 for the mid-rank and Anderson-Darling scores

```
> mr <- rank(Y)
> w2.mr <- (sort(unique(mr))[2]-min(mr))/(max(mr)-min(mr))
> w2.mr
[1] 0.5294118
> F <- ecdf(Y)
> varF <-sort(unique(F(Y)))*(1-sort(unique(F(Y))))
> w2.ad<-(1/sqrt(varF[1]))/(1/sqrt(varF[1])+1/sqrt(varF[2]))
> w2.ad
[1] 0.4560859
```

Perform the `ptest2s` function for comparing two independent samples based on Student's t statistic $T_{\boldsymbol{w}}$ with equal-spaced scores

Table 2.5. Unconditional power comparisons.

$(\boldsymbol{\pi}_1;\boldsymbol{\pi}_2)$	$\overset{\circ}{\preceq}$	C-vM	WMW	A-D	$T^{\max}$	MP	λ
$(0.3,0.3,0.4)$ $(0.3,0.3,0.4)$	H_0	0.033	0.042	0.040	0.040	-	-
$(0.2,0.5,0.3)$ $(0.1,0.1,0.8)$	hr	0.692	0.814	0.674	0.823	0.915	-0.547
$(0.2,0.5,0.3)$ $(0.1,0.3,0.6)$	lr	0.391	0.465	0.409	0.456	0.485	0.131
$(0.2,0.7,0.1)$ $(0.2,0.3,0.5)$	hr	0.430	0.540	0.448	0.695	0.840	-0.526
$(0.3,0.1,0.6)$ $(0.1,0.1,0.8)$	lr	0.320	0.323	0.345	0.340	0.350	0.792
$(0.3,0.2,0.5)$ $(0.1,0.3,0.6)$	st	0.198	0.202	0.262	0.279	0.377	1.174
$(0.4,0.4,0.2)$ $(0.1,0.7,0.2)$	st	0.248	0.358	0.274	0.447	0.627	1.403
$(0.4,0.5,0.1)$ $(0.3,0.1,0.6)$	hr	0.584	0.634	0.664	0.853	0.960	-0.635
$(0.6,0.2,0.2)$ $(0.4,0.2,0.4)$	lr	0.265	0.292	0.294	0.295	0.297	0.369
$(0.6,0.3,0.1)$ $(0.1,0.4,0.5)$	lr	0.940	0.956	0.949	0.948	0.962	0.611
$(0.6,0.3,0.1)$ $(0.2,0.1,0.7)$	lr	0.947	0.953	0.961	0.974	0.975	0
$(0.4,0.5,0.1)$ $(0.4,0.4,0.2)$	hr	0.067	0.080	0.103	0.117	0.175	-0.322

```
> source("ptest2s.R")
> set.seed(0)
> B <- 5000
> T <- ptest2s(Y,X,B,"Student")
> p.obs <- sum(T[-1]>=T[1])/(B-1)
> p.obs
[1] 0.01140228
```

and with midrank or Anderson-Darling scores

```
> Y.mr <- Y
> Y.mr[Y.mr==1] <- 0
> Y.mr[Y.mr==2] <- w2.mr
> Y.mr[Y.mr==3] <- 1
```

```
> set.seed(0)
> T <- ptest2s(Y.mr,X,B,"Student")
> p.obs <- sum(T[-1]>=T[1])/(B-1)
> p.obs
[1] 0.01140228
> Y.ad <- Y.mr
> Y.ad[Y.ad==w2.mr] <- w2.ad
> set.seed(0)
> T <- ptest2s(Y.ad,X,B,"Student")
> p.obs <- sum(T[-1]>=T[1])/(B-1)
> p.obs
[1] 0.00620124
```

The adaptive test $T_{\boldsymbol{w}^{\max}}$ can be performed by computing the data-dependent scores $\boldsymbol{w}^{\max}$ at each permutation of the data using the function `Tmax`:

```
> source("Tmax.R")
> set.seed(0)
> T <- ptest2s(Y,X,B,"Tmax")
> p.obs <- sum(T[-1]>=T[1])/(B-1)
> p.obs
[1] 0.00680136
```

2.2.4 Concordance Monotonicity

Likelihood inference is perhaps the default approach for many statistical models. Recently, there have been debates about the suitability of different test procedures: Perlman and Chaudhuri (2004b) argue in favor of likelihood ratio tests, whereas Cohen and Sackrowitz (2004) argue in favor of the so-called class of directed tests. The likelihood ratio test (LRT) statistic is given by

$$T_{LR} = 2\sum_{i=1}^{2}\sum_{j=1}^{c} m_{i,j}\left\{\log[\hat{\pi}_i^{|H_1}(j)] - \log[\hat{\pi}_i^{|H_0}(j)]\right\},$$

where $\hat{\pi}_i^{|H_1}(j)$ and $\hat{\pi}_i^{|H_0}(j)$ are the maximum likelihoood (ML) estimates of $\pi_i(j)$ under H_1 and H_0, respectively. When $m_{i,j} > 0$, $i = 1, 2$, $j = 1, \ldots, c$, Dykstra et al. (1996) showed that ML estimates can be expressed in terms of a weighted least squares projection,

$$\hat{\pi}_1^{|H_1}(j) = \frac{\boldsymbol{m}_1}{n_1}\left\{\frac{n_1}{n} + \frac{n_2}{n}P_{\frac{\boldsymbol{m}_1}{n_1}}\left(\frac{\boldsymbol{m}_1}{\boldsymbol{m}_2}|\mathcal{I}\right)\right\},$$
$$\hat{\pi}_2^{|H_1}(j) = \frac{\boldsymbol{m}_2}{n_2}\left\{\frac{n_2}{n} + \frac{n_1}{n}P_{\frac{\boldsymbol{m}_2}{n_2}}\left(\frac{\boldsymbol{m}_2}{\boldsymbol{m}_1}|\mathcal{D}\right)\right\},$$

where $\mathcal{I} = \{\boldsymbol{w} \in \mathcal{R}^c : w_1 \leq \ldots \leq w_c\}$ and $\mathcal{D} = \{\boldsymbol{w} \in \mathcal{R}^c : w_1 \geq \ldots \geq w_c\}$.

Then, a least favorable null value for the asymptotic distribution of the LRT assigns probability $\frac{1}{2}$ for the first and the last ordinal categories (see Silvapulle and Sen, 2005, Proposition 6.5.1), and

$$\sup_{H_0} \lim_{n\to\infty} \Pr(T_{LR} \geq t|H_0) = \frac{1}{2}(\Pr(\chi^2_{c-2} \geq t) + \Pr(\chi^2_{c-1} \geq t)).$$

To bypass possibly poor asymptotic approximations, mostly for small or unbalanced sample sizes, Agresti and Coull (2002) suggest performing the permutation test based on the LRT statistic. The cell entries in Tables I and II of Table 2.6 represent two permutations of Patefield's data.

Table 2.6. Two permutations of Patefield's data.

Table I	W	S	B	Total
C	5	11	1	17
T	3	8	4	15
	8	19	5	32

Table II	W	S	B	Total
C	0	16	1	17
T	8	3	4	15
	8	19	5	32

Note that Table II is created from Table I by exchanging five control values from "worse" to "same" and five treatment values from "same" to "worse". The LRT statistic for Tables I and II is 2.777 and 22.652, respectively. This seems to contradict intuition because the control performance is improved while simultaneously the treatment is made to perform worse. Then we would expect the p-value to increase. The LRT does not have this property (i.e., is not *concordant monotone*, Cohen and Sackrowitz, 1998), meaning that the p-value decreases if any entry in the first row, say $m_{1,j}$, increases while $m_{1,l}$ decreases for $j < l$, holding all row and column totals fixed.

As an alternative to LRT, Cohen et al. (2003) developed the directed chi-square, which is concordant monotone and is defined as

$$T_{\overrightarrow{\chi}^2} = \inf_{\boldsymbol{u}\in\mathcal{A}} \sum_{i=1}^{2}\sum_{j=1}^{c} \frac{\left(u_{i,j} - \frac{n_i t_j}{n}\right)^2}{\frac{n_i t_j}{n}},$$

where $\mathcal{A} = \{u_{1,1} + \ldots + u_{1,j} \geq m_{1,1} + \ldots + m_{1,j}, \sum_{j=1}^{c} u_{i,j} = n_i, u_{1,j} + u_{2,j} = t_j, i = 1, 2, j = 1, \ldots, c\}$. Therefore, the directed chi-squared test rejects H_0 if the minimum of the Pearson chi-square for tables in $\mathcal{A}$ is large. Cohen et al. (2003) showed that a permutationally equivalent formulation is given by

$$T_{\overrightarrow{\chi}^2} = \sum_{j=1}^{c} (w_j)^2 t_j,$$

where $\boldsymbol{w} = P_{\boldsymbol{t}}\left(\frac{\boldsymbol{m_1}}{\boldsymbol{t}}|\mathcal{D}\right)$ and $\mathcal{D} = \{\boldsymbol{w} \in I\!R^c : w_1 \geq \ldots \geq w_c\}$.

2.2.5 Applications with R functions

Here we provide assistance in doing the statistical tests illustrated in Subsection 2.2.4. Create Table I and Table II representing two permutations of Patefield's data by using equal-spaced scores $(w_1, w_2, w_3) = (1, 2, 3)$, where `X` represents class labels and `Y` the vector of data

```
> X <- c(rep(1,17),rep(2,15))
> YI <- c(rep(1,5),rep(2,11),3,rep(1,3),rep(2,8),rep(3,4))
> YII <- c(rep(2,16),rep(3,1),rep(1,8),rep(2,3),rep(3,4))
```

and perform the likelihood ratio test T_{LR} by `LRT`

```
> source("LRT.R")
>  LRT(YI[X==1],YI[X==2])
[1] 2.783381
>  LRT(YII[X==1],YII[X==2])
[1] 19.4776
```

and the directed chi-squared test $T_{\overrightarrow{\chi^2}}$ by `DChisq`

```
> source("DChisq.R")
>  DChisq(YI,X)
[1] 0.07446918
>  DChisq(YII,X)
[1] 0.07037817
```

Tests based on linear test statistics are also concordant monotone. For example,

```
> source("studT.R")
> studT(YI[X==1],YI[X==2])
[1] 1.348210
> studT(YI[X==1],YI[X==2])
[1] -1.460447
> source("Tmax.R")
> Tmax(YI[X==1],YI[X==2])
[1] 0.2882637
> Tmax(YII[X==1],YII[X==2])
[1] 0.2856531
```

2.2.6 Multiple Testing

In this multiparameter problem, following Roy's (1953) union-intersection principle , it might be possible to look upon the null hypothesis as the intersection of several component hypotheses and the alternative hypothesis as the union of the same number of component alternatives, in symbols

$$H_0 : \boldsymbol{\theta}_1^C = \mathbf{1}_{c-1} \Leftrightarrow \bigcap_{j=1}^{c-1} \{H_{0,j}\} : \bigcap_{j=1}^{c-1} \left\{\theta_{1j}^C = 1\right\},$$

stating that H_0 is true if all $H_{0,j}$ are true, and

$$H_1 : \boldsymbol{\theta}_1^C \gtrsim \mathbf{1}_{c-1} \Leftrightarrow \bigcup_{j=1}^{c-1} \{H_{1,j}\} : \bigcup_{j=1}^{c-1} \left\{\theta_{1j}^C > 1\right\},$$

stating that H_1 is true if at least one $H_{1,j}$ is true.

To provide an interpretation of this, let us consider the $c-1$ possible 2×2 subtables that can be formed by dichotomizing the column variable: the first column vs. all the rest, the first two columns pooled vs. the others, and so on. Thus $H_{0,j}$ and $H_{1,j}$ define the hypotheses of interest for the jth subtable, $j = 1, \ldots, c-1$ (Table 2.7).

Table 2.7. jth subtable.

	$\leq j$	$> j$	
1	$\sum_{l=1}^{j} m_{1,l}$	$\sum_{l=j+1}^{c} m_{1,l}$	n_1
2	$\sum_{l=1}^{j} m_{2,l}$	$\sum_{l=j+1}^{c} m_{2,l}$	n_2
	$\sum_{l=1}^{j} t_l$	$\sum_{l=j+1}^{c} t_l$	n

For testing $H_{0,j}$ against $H_{1,j}$, we may consider Fisher's test statistic $T_j = \sum_{l=1}^{j} m_{1,l}$ or its standardized formulation

$$T_j = \left(\frac{n_1 n_2}{n^2/(n-1)}\right) \frac{\hat{F}_1(j) - \hat{F}_2(j)}{(\hat{F}(j)[1-\hat{F}(j)])^{\frac{1}{2}}}.$$

For any $K \subseteq \{1, \ldots c-1\}$, let $H_{0,K} : \bigcap_{j \in K} \{H_{0,j}\}$ denote the hypothesis that all $H_{0,j}$ with $j \in K$ are true. The closure method of Marcus et al. (1976) allows strong control of FWE if we know how to test each intersection hypothesis $H_{0,K}$. Let T_K be a test statistic for $H_{0,K}$ that can be a function of test statistics T_j or p-values p_j. For instance, by using the standardized Fisher statistic, the "sum-T" combined test $T_K = \sum_{j \in K} T_j$ yields the Anderson-Darling statistic, whereas by using $T_j = \hat{F}_1(j) - \hat{F}_2(j)$, the "max-T" combined test $T_K = \max_{j \in K} T_j$ yields the Smirnov statistic.

For the analysis of Patefield's data, by applying Fisher's exact tests, we obtain $p_1^{obs} = 0.1536$ and $p_2^{obs} = 0.0149$. Depending on the combined test used to test H_0, obtaining p-value p^{obs}, from the closed testing principle we have $p_1^{adj} = \max(0.1536, p^{obs})$ and $p_2^{adj} = \max(0.0149, p^{obs})$. When $p^{obs} \leq \alpha$, as

happens with the tests considered in Subsection 2.2.2, the "individual" hypothesis $H_{0,2} : \theta_2^C = 1$ can be rejected while controlling the FWE, supporting the alternative $H_{1,2} : \theta_2^C > 1$.

2.3 Independent Binomial Samples: $r \times 2$ Contingency Tables

In a typical dose-response study, several increasing doses of a treatment are randomly assigned to the subjects, with each subject receiving only one dose throughout the study. We discuss the case of a binary response variable Y with a single regressor X having r ordered levels. Let $Y_1, \ldots, Y_r$ be r independent binomial variables with $Y_i \sim \text{Binomial}(n_i, \pi_i)$, where the probability of "success" is $\pi_i := \pi_i(2) = \Pr(Y = 2|X = i)$. We are often interested in detecting inequalities between the parameters π_i, $i = 1, \ldots, r$.

Graubard and Korn (1987) consider a prospective study of maternal drinking and congenital malformations. After the first three months of pregnancy, the women in the sample completed a questionnaire about alcohol consumption (average number of drinks per day). Following childbirth, observations were recorded on the presence or absence of congenital sex organ malformations. The data are displayed in Table 2.8.

Table 2.8. Maternal drinking and congenital malformations data.

X	Alcohol Consumption	Malformation Absent	Malformation Present	
1	0	17,066	48	17,114
2	< 1	14,464	38	14,502
3	1–2	788	5	793
4	3–5	126	1	127
5	≥ 6	37	1	38
		32,481	93	32,574

The goal is to test for a dose-response relationship. For example, when investigating a dose-response relationship of the form $\text{logit}(\pi_i) = \gamma + \beta d_i$, one would typically have in mind a biologically or clinically meaningful slope, say β , above which one could claim the existence of a trend in the data. Specifically, it is of interest to test against a *simple order restriction*,

$$H_0 : \pi_1 = \ldots = \pi_r \quad \text{against} \quad H_1 : \pi_1 \leq \ldots \leq \pi_r, \tag{2.5}$$

with at least one strict inequality. An efficient test of the null hypothesis is the Cochran-Armitage test of trend (Cochran, 1954; Armitage, 1955), in which

the test statistic is

$$T_{\boldsymbol{d}} = \sum_{i=1}^{r-1} m_{i,2} d_i,$$

where the d_i's are prespecified scores that may correspond to doses in a dose-response setting. It is known (Agresti, 2002, pp. 181−182) that the Cochran-Armitage statistic is equivalent to the score statistic for testing $H_0 : \beta = 0$ in the linear logit model. Cochran (1954) noted that "any set of scores gives a valid test, provided that they are constructed without consulting the results of the experiment. If the set of scores is poor, in that it badly distorts a numerical scale that really does underlie the ordered classification, the test will not be sensitive. The scores should therefore embody the best insight available about the way in which the classification was constructed and used." Ideally, the scale is chosen by a consensus of experts, and subsequent interpretations use that same scale. When uncertain about this choice, the adaptive scores used in Subsection 2.2.2 may be considered.

Alcohol consumption, measured as the average number of drinks per day, is an ordinal explanatory variable. This groups a naturally continuous variable, and we first use the scores $\boldsymbol{d} = (0, .5, 1.5, 4, 7)$, the last score being somewhat arbitrary. For this choice, the p-value is 0.014. By contrast, for the equally spaced row scores, $\boldsymbol{d} = (1, 2, 3, 4, 5)$, giving a much weaker conclusion ($p = 0.104$). Midrank scores yield an even weaker conclusion ($p = 0.319$). Why does this happen? Adjacent categories having relatively few observations necessarily have similar midranks. This scoring scheme treats the alcohol consumption level 1−2 drinks as much closer to consumption level ≥ 6 drinks than to consumption level 0 drinks. This seems inappropriate since it is usually better to select scores that reflect distances between doses. However, by using the adaptive scores $\boldsymbol{d} = \boldsymbol{w}^{\max}$, the p-value is 0.022, supporting the adaptive test when the choice of scores is uncertain.

Peddada et al. (2001) consider a study investigating the effects of several treatments on the reproductive condition of the redbacked salamander (*Plethodon cinereus*). Female salamanders were randomly assigned to either the control or one of three treatment groups. The treatments consisted of injections of either follicle-stimulating hormone ($i = 2$), luteinizing hormone ($i = 3$), or, for animals in the control group ($i = 1$), saline solution. The remaining treatment group ($i = 4$) was fed exactly twice the amount of food as salamanders in all other groups. The reproductive condition of each animal was later evaluated by measuring the size of the ova through the abdominal wall of the animal. If the ova were larger than 2 mm, then the animal was declared to be in a reproductive condition. Data are displayed in Table 2.9.

The hypothesis of interest is whether the salamanders in the treatment groups had a greater probability of being in reproductive condition than those in the control group. No ordering was hypothesized between treatment groups, and hence we wish to test against a *simple tree order restriction*

Table 2.9. Reproductive condition of the redbacked salamander.

	Nonreproductive	Reproductive	Total
1	4	9	13
2	8	4	12
3	7	6	13
4	1	13	14
	20	22	42

$$H_0 : \pi_1 = \ldots = \pi_r \quad \text{against} \quad H_1 : \bigcup_{i=1}^{r-1} \{\pi_1 < \pi_{i+1}\}. \tag{2.6}$$

In the case of rejection of the global null hypothesis that none of the treatments is an improvement over the control, answering the question "Is there any evidence of the treatment effect?" one usually wants to know which of the treatments show a significant difference, answering the more specific question "For which treatments is the response larger than the response in the control group?"; that is, to test simultaneously the hypotheses

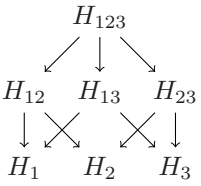
$$H_{0,i} : \pi_1 = \pi_{i+1} \quad \text{against} \quad H_{1,i} : \pi_1 < \pi_{i+1}, \qquad i = 1, \ldots, r-1. \tag{2.7}$$

A multiple comparison procedure can be used for this purpose. No type I error should be made in any of these comparisons because otherwise a treatment that is actually inferior to the control may be recommended. Thus, in this case, strong control of the FWE is required.

For any $K \subseteq \{1, \ldots r-1\}$, let $H_{0,K} : \bigcap_{i \in K} H_{0,i}$ denote the hypothesis that all $H_{0,i}$ with $i \in K$ are true. Note that for testing the global null hypothesis H_{0,K_0} with $K_0 = \{1, \ldots, r-1\}$, all permutations of the observations among the r groups are equally likely. However, for testing the intersection hypothesis $H_{0,K}$, we consider only the permutations that, under that hypothesis, become equally likely. In particular, for each hypothesis $H_{0,i}$, one has to permute only within the control and the $(i+1)$th treatment. Thus we should consider a closed testing procedure that uses the valid permutations depending on the intersection hypothesis under testing. For testing $H_{0,i}$, we consider as test statistics

$$T_i = \frac{\hat{\pi}_{i+1} - \hat{\pi}_1}{\left[\left(\frac{n_1\hat{\pi}_1 + n_{i+1}\hat{\pi}_{i+1}}{n_1 + n_{i+1}}\right)\left(1 - \frac{n_1\hat{\pi}_1 + n_{i+1}\hat{\pi}_{i+1}}{n_1 + n_{i+1}}\right)\right]^{\frac{1}{2}}}, \qquad i = 1, \ldots, r-1,$$

where $\hat{\pi}_i = m_{i2}/n_i$, and for testing H_K the "max-T" combined test statistic $T_K = \max(T_i : i \in K)$. Results are given in the following diagram to better illustrate the closed testing method. The result indicates that no individual hypothesis can be rejected at a nominal level $\alpha = 5\%$.



$p_{123} = 0.1636$

$p_{12} = 0.9643$ $p_{13} = 0.1268$ $p_{23} = 0.1020$

$p_1 = 0.9851$ $p_2 = 0.9439$ $p_3 = .01448$

Note that in all examples there is prior belief in the shape of the exposure-outcome curve. The prior belief relates to a restricted alternative to the "no effect" hypothesis. For instance, in the salamander example, *a priori expectations* were that all three treatment groups would result in increased ova development compared with a control.

2.3.1 Applications with `R` functions

This paragraph provides assistance in using the statistical methods illustrated in Section 2.3. Create the malformations data with equal-spaced scores, where `X` represents class labels and `Y` the vector of data,

```
> X <- c(rep(1,17114),rep(2,14502),rep(3,793),rep(4,127),
+ rep(5,38))
> Y<-c(rep(0,17066),rep(1,48),rep(0,14464),rep(1,38),
+ rep(0,788),rep(1,5),rep(0,126),rep(1,1),rep(0,37),rep(1,1))
```

and perform the `ptest2s` function based on $T_{\boldsymbol{d}}$ by switching the input arguments `X` and `Y`; with equally spaced, scores, we obtain

```
> source("ptest2s.R")
> set.seed(0)
> B <- 1000
> T <- ptest2s(X,Y,B,"Student")
> p.obs <- sum(T[-1]>=T[1])/(B-1)
> p.obs
[1] 0.1041041
```

whereas with $\boldsymbol{d} = (0, .5, 1.5, 4, 7)$ and midrank scores, we obtain

```
> set.seed(0)
> X.d <- c(rep(0,17114),rep(.5,14502),rep(1.5,793),
+ rep(4,127),rep(7,38))
> T <- ptest2s(X.d,Y,B,"Student")
> p.obs <- sum(T[-1]>=T[1])/(B-1)
> p.obs
[1] 0.01401401
> set.seed(0)
> X.mr <- rank(X)
> T <- ptest2s(X.mr,Y,B,"Student")
```

```
> p.obs <- sum(T[-1]>=T[1])/(B-1)
> p.obs
[1] 0.3193193
```

Finally, by using adaptive scores $\boldsymbol{d} = \boldsymbol{w}^{\max}$, we obtain

```
> source("Tmax.R")
> set.seed(0)
> T <- ptest2s(X,Y,B,"Tmax")
> p.obs <- sum(T[-1]>=T[1])/(B-1)
> p.obs
[1] 0.02202202
```

To set up the redbacked salamander data, type

```
> X <- c(rep(1,13),rep(2,12),rep(3,13),rep(4,14))
> Y <- c(rep(0,4),rep(1,9),rep(0,8),rep(1,4),
+ rep(0,7),rep(1,6),rep(0,1),rep(1,13))
```

To test the global null hypothesis $H_{0,\{1,2,3\}} : \bigcap_{i=1}^{3}\{\pi_1 = \pi_{i+1}\}$, perform the `ptestRs` function based on the combined statistic $T_{\{1,2,3\}} = \max(T_1, T_2, T_3)$

```
> source("ptestRs.R")
> set.seed(0)
> B <- 5000
> T123 <- ptestRs(Y,X,B,combi="max")
> p.obs <- sum(T123[-1]>=T123[1])/(B-1)
> p.obs
[1] 0.1672334
```

and for the intersection hypotheses $H_{0,\{1,2\}}$, $H_{0,\{1,3\}}$, and $H_{0,\{2,3\}}$

```
> T12 <- ptestRs(Y[X!=4],X[X!=4],B,combi="max")
> sum(T12[-1]>=T12[1])/(B-1)
[1] 0.9643929
> T13 <- ptestRs(Y[X!=3],X[X!=3],B,combi="max")
> sum(T13[-1]>=T13[1])/(B-1)
[1] 0.1254251
> T23 <- ptestRs(Y[X!=2],X[X!=2],B,combi="max")
> sum(T23[-1]>=T23[1])/(B-1)
[1] 0.1028206
```

Finally, to test $H_{0,1}$, $H_{0,2}$ and $H_{0,3}$, perform the two-sample comparisons by using the `ptest2s` function

```
> T1 <- ptest2s(Y[X!=3&X!=4],X[X!=3&X!=4],B,"Student")
> sum(T1[-1]>=T1[1])/(B-1)
[1] 0.9867974
> T2 <- ptest2s(Y[X!=2&X!=4],X[X!=2&X!=4],B,"Student")
> sum(T2[-1]>=T2[1])/(B-1)
```

```
[1] 0.9461892
> T3 <- ptest2s(Y[X!=2&X!=3],X[X!=2&X!=3],B,"Student")
> sum(T3[-1]>=T3[1])/(B-1)
[1] 0.1386277
```

2.4 Comparison of Several Treatments when the Response is Ordinal: $r \times c$ Contingency Tables

Table 2.10 displays data appearing in Chuang-Stein and Agresti (1997). Five ordered categories ranging from "death" to "good recovery" describe the clinical outcome of patients who experienced trauma. In the literature on critical care, these five categories are often called the Glasgow Outcome Scale (GOS). We have four treatment groups: three intravenous doses for the medication (low, medium, and high) and a vehicle infusion serving as the control.

Table 2.10. Glasgow Outcome Scale.

Treatment group	X	Death	Vegetative state	Major disability	Minor disability	Good recovery	Total
Placebo	1	59	25	46	48	32	210
Low dose	2	48	21	44	47	30	190
Medium dose	3	44	14	54	64	31	207
High dose	4	43	4	49	58	41	195
		194	64	193	217	134	802

Investigation of a dose-response relationship is of primary interest in many drug-development studies. Here the outcome of interest is measured at several (increasing) dose levels, among which there is a control group. One study objective was to determine whether a more favorable GOS outcome tends to occur as the dose increases; that is, testing

$$H_0 : Y_1 \stackrel{d}{=} \ldots \stackrel{d}{=} Y_r \quad \text{against} \quad H_1 : Y_1 \stackrel{st}{\leq} \ldots \stackrel{st}{\leq} Y_r$$

with at least one "$\stackrel{st}{\lesssim}$". Note that the dose-response curve is assumed to be monotone; i.e., the GOS increases as the dose level increases.

Other questions usually asked in dose-response studies are: "For which doses is the response higher from the response in the control group?"; that is, testing the many-to-one comparisons

$$H_{0,i} : Y_1 \stackrel{d}{=} Y_{i+1} \quad \text{against} \quad H_{1,i} : Y_1 \stackrel{st}{\lesssim} Y_{i+1}, \qquad i = 2, \ldots, r,$$

or "What are the strict inequalities in the stochastic ordering relationship?", that is, testing all pairwise comparisons

$$H_{0,(i,i')} : Y_i \overset{d}{=} Y_{i'} \quad \text{against} \quad H_{1,(i,i')} : Y_i \overset{st}{\lesssim} Y_{i'}, \qquad i < i'.$$

Gatekeeping procedures (see Dmitrienko and Tamhane, 2007, for an overview) have become popular in recent years as they provide a convenient way to handle logical relationships between multiple objectives that clinical trials are often required to address. In a gatekeeping strategy, the k null hypotheses are divided into h ordered families $\mathcal{F}_l$, $l = 1, \ldots, h$. Generally, familywise error rate control at a designated level α is desired for the family of all k hypotheses.

Westfall and Krishen (2001) proposed procedures for the *serial* gatekeeping problem in which the hypotheses in $\mathcal{F}_{l+1}$ are tested if and only if all hypotheses in $\mathcal{F}_l$ are rejected. Dmitrienko et al. (2003) proposed procedures for the *parallel* gatekeeping problem in which the hypotheses in $\mathcal{F}_{l+1}$ are tested if at least one hypothesis in $\mathcal{F}_l$ is rejected.

For the many-to-one comparisons, the serial gatekeeping procedure may exploit the hierarchy of the stochastic ordering relationship $Y_1 \overset{st}{\leq} \ldots \overset{st}{\leq} Y_r$ by starting from the comparison between the highest dose and the control to the comparison between the lowest dose and the control. Here, the ordered families are simply $\mathcal{F}_1 = \{H_{0,r}\}, \ldots, \mathcal{F}_{r-1} = \{H_{0,2}\}$. The procedure stops at the dose level where the null hypothesis is not rejected at the nominal level $\alpha = 5\%$.

For the pairwise comparisons, a parallel gatekeeping procedure may exploit the distance $(i, i') = i' - i$ in the stochastic ordering relationship by testing first the hypothesis $H_{(1,r)}$ comparing the highest dose with the control and, if rejected, both $H_{(1,r-1)}$ and $H_{(2,r)}$, and if at least one is rejected, the three hypotheses $H_{0,(1,r-2)}$, $H_{0,(2,r-1)}$, and $H_{0,(3,r-2)}$, and so on. Here the ordered families are $\mathcal{F}_1 = \{H_{0,(1,r)}\}, \mathcal{F}_2 = \{H_{0,(1,r-1)}, H_{0,(2,r)}\}, \ldots, \mathcal{F}_l = \{H_{0,(i,i')} : i - i' = r - l\}, \ldots, \mathcal{F}_{r-1} = \{H_{0,(i,i+1)}, i < i'\}$. In this procedure, the first $r - 2$ families are tested using the Bonferroni single-step adjustment that tests $\mathcal{F}_l$ at level $\alpha\gamma_l$. The family $\mathcal{F}_{r-1}$ is tested at level $\alpha\gamma_l$ using Holm's stepdown adjustment. Here γ_l is the so-called rejection gain factor for $\mathcal{F}_l$, given by $\gamma_1 = 1$, $\gamma_l = \prod_{j=1}^{l-1} \left(\frac{\text{rejected}(\mathcal{F}_j)}{\text{cardinality}(\mathcal{F}_j)} \right)$, where "rejected($\mathcal{F}_j$)" is the number of rejected hypotheses in $\mathcal{F}_j$; thus γ_l is the product of the proportions of rejected hypotheses in $\mathcal{F}_1$ through $\mathcal{F}_{l-1}$. If no hypotheses are rejected in some family $\mathcal{F}_l$, then $\gamma_j = 0$ for all $j > l$, and all hypotheses in $\mathcal{F}_j$ for $j > l$ are automatically accepted. On the other hand, if all hypotheses are rejected in $\mathcal{F}_1$ through $\mathcal{F}_{l-1}$, then $\gamma_l = 1$ and thus a full α level is used to test $\mathcal{F}_l$, no part of α being used up by the rejected hypotheses ("use it or lose it" principle).

To illustrate the implementation of gatekeeping procedures, consider the GOS example. Raw p-values for the six hypotheses computed from a two-sample $T_{\boldsymbol{w}}$ test (with equally spaced or adaptive scores) are displayed in Table 2.11.

Table 2.11. Raw p-values for GOS data.

	$H_{0,(1,4)}$	$H_{0,(1,3)}$	$H_{0,(2,4)}$	$H_{0,(1,2)}$	$H_{0,(2,3)}$	$H_{0,(3,4)}$
Equally spaced	0.0026	0.0282	0.0194	0.2819	0.1206	0.1666
Adaptive	0.0018	0.0220	0.0162	0.4563	0.1258	0.1466

For many-to-one comparisons, by the serial gatekeeping procedure we reject at $\alpha = 5\%$ both the hypotheses $H_{0,(1,4)}$ and $H_{0,(1,3)}$ but not the comparison between the lowest dose and control; that is, $H_{0,(1,2)}$. Note that by performing Holm's procedure, with adaptive scores we reject both $H_{0,(1,4)}$ and $H_{0,(1,3)}$ ($p_{(1,4)} = 0.0018 < \alpha/3$ and $p_{(1,3)} = 0.0220 < \alpha/2$) but with equally spaced scores we can reject $H_{0,(1,4)}$ only ($p_{(1,3)} = 0.0282 > \alpha/2$).

For the six pairwise comparisons, with Holm's procedure we reject only the hypothesis $H_{0,(1,4)}$. However, because there are logical implications among the hypotheses and alternatives, Holm's procedure can be improved to obtain a further increase in power (Shaffer, 1986). By considering all possible configurations of true and false hypotheses, all six hypotheses may be true at the first step, but because the hypothesis $H_{0(1,4)}$ is rejected ($p_{(1,4)} \leq \alpha/6$), at least three must be false since if any two distributions differ, at least one of them must differ from the remaining ones. By exploiting logical implications and using p-values from tests based on adaptive scores, we can also reject $H_{0,(2,4)}$ ($p_{(2,4)} = 0.0162 < \alpha/3$). By performing the parallel gatekeeping procedure, with adaptive scores we reject at $\alpha = 5\%$ all the hypotheses in the families $\mathcal{F}_1 = \{H_{0,(1,4)}\}$ and $\mathcal{F}_2 = \{H_{0,(1,3)}, H_{0,(2,4)}\}$ but none of the hypotheses in the family $\mathcal{F}_3 = \{H_{0,(1,2)}, H_{0,(2,3)}, H_{0,(3,4)}\}$, whereas with equally spaced scores we reject only the hypotheses $H_{0,(1,4)}$ and $H_{0,(2,4)}$ because $p_{0,(2,3)} = 0.0282 > \alpha/2$.

http://www.springer.com/978-0-387-85955-2

Permutation Tests for Stochastic Ordering and ANOVA
Theory and Applications with R
Basso, D.; Pesarin, F.; Salmaso, L.; Solari, A.
2009, XIV, 218 p., Softcover
ISBN: 978-0-387-85955-2