Chapter 6

The One-Way Layout

INTRODUCTION

The procedures of this chapter are designed for statistical analyses in which primary interest is centered on the relative locations (medians) of three or more populations. This development represents a direct generalization of the two-sample location problem (discussed in Chapter 4) to situations in which the data consist of $k\,(\geq 3)$ random samples, one sample from each of k populations. The basic null hypothesis of interest in that of no differences in locations (medians), under which the k samples can be treated as a single (combined) sample from one population. The alternatives considered here correspond to a variety of restricted nonnull relationships between the locations (medians). We encounter two types of data for which such analyses are important. The first of these corresponds to a general setting of k populations (referred to as *treatments* for convenience) with no additional conditions. The second deals with the setting where one of the *treatments* represents a *control* (or placebo) population, and we are interested in detecting which, if any, of the other $(k-1)$ treatments are different from this control.

Section 6.1 presents a distribution-free test directed at general alternatives for the setting of k treatments. A distribution-free test designed for detecting ordered alternatives among k treatments is considered in Section 6.2 and generalized in Section 6.3 to the broader class of umbrella alternatives. In Section 6.4 a distribution-free test procedure is presented for the simultaneous comparison of $(k-1)$ treatments with a control. In Sections 6.5–6.7 we introduce multiple comparison procedures designed to detect which particular populations, if any, differ from one another. Sections 6.5 and 6.6 are devoted to procedures for making the total of $\binom{k}{2}$ pairwise comparisons between all k treatments in the general and ordered alternatives settings, respectively. Section 6.7 presents multiple comparison procedures based on simple random samples for deciding which, if any, of $(k-1)$ treatments are different from a control. Section 6.8 considers estimators of contrasts in the treatment effects, and Section 6.9 deals with simultaneous confidence intervals for simple contrasts. The asymptotic relative efficiencies for translation alternatives of the procedures discussed in this chapter with respect to their normal theory counterparts based on sample averages are discussed in Section 6.10.

Nonparametric Statistical Methods, Third Edition. Myles Hollander, Douglas A. Wolfe, Eric Chicken.
© 2014 John Wiley & Sons, Inc. Published 2014 by John Wiley & Sons, Inc.

Data. The data consist of $N = \sum_{j=1}^{k} n_j$ observations, with n_j observations from the jth treatment, $j = 1, \ldots, k$.

Treatments			
1	2	...	k
X_{11}	X_{12}	...	X_{1k}
X_{21}	X_{22}	...	X_{2k}
⋮	⋮		⋮
$X_{n_1 1}$	$X_{n_2 2}$	...	$X_{n_k k}$

Assumptions

A1. The N random variables $\{X_{1j}, X_{2j}, \ldots, X_{n_j j}\}$, $j = 1, \ldots, k$, are mutually independent.

A2. For each fixed $j \in \{1, \ldots, k\}$, the n_j random variables $\{X_{1j}, X_{2j}, \ldots, X_{n_j j}\}$ are a random sample from a continuous distribution with distribution function F_j.

A3. The distribution functions $F_1, \ldots, F_k$ are connected through the relationship

$$F_j(t) = F(t - \tau_j), \quad -\infty < t < \infty, \tag{6.1}$$

for $j = 1, \ldots, k$, where F is a distribution function for a continuous distribution with unknown median θ and τ_j is the unknown treatment effect for the jth population.

We note that Assumptions A1–A3 correspond directly to the usual one-way layout model commonly associated with normal theory assumptions; that is, Assumptions A1–A3 are equivalent to the representation

$$X_{ij} = \theta + \tau_j + e_{ij}, \quad i = 1, \ldots, n_j, \quad j = 1, \ldots, k,$$

where θ is the overall median, τ_j is the *treatment j effect*, and the N e's form a random sample from a continuous distribution with median 0. (Under the additional assumption of normality, the medians θ and 0 are, of course, also the respective means.)

Hypothesis

The null hypothesis of interest in Sections 6.1–6.4 of this chapter is that of no differences among the treatment effects $\tau_1, \ldots, \tau_k$, namely,

$$H_0 : [\tau_1 = \cdots = \tau_k]. \tag{6.2}$$

This null hypothesis asserts that each of the underlying distributions $F_1, \ldots, F_k$ is the same, corresponding to $F_1 \equiv F_2 \equiv \cdots \equiv F_k \equiv F$ in (6.1).

6.1 A DISTRIBUTION-FREE TEST FOR GENERAL ALTERNATIVES (KRUSKAL–WALLIS)

In this section, we present a procedure for testing H_0 (6.2) against the general alternative that at least two of the treatment effects are not equal, namely,

$$H_1 : [\tau_1, \ldots, \tau_k \text{ not all equal}]. \tag{6.3}$$

Procedure

To compute the Kruskal–Wallis statistic, H, we first combine all N observations from the k samples and order them from least to greatest. Let r_{ij} denote the rank of X_{ij} in this joint ranking and set

$$R_j = \sum_{i=1}^{n_j} r_{ij} \quad \text{and} \quad R_{\cdot j} = \frac{R_j}{n_j}, \quad j = 1, \ldots, k. \tag{6.4}$$

Thus, for example, R_1 is the sum of the joint ranks received by the treatment 1 observations and $R_{\cdot 1}$ is the average rank for these same observations. The Kruskal–Wallis statistic H is then given by

$$\begin{aligned} H &= \frac{12}{N(N+1)} \sum_{j=1}^{k} n_j \left(R_{\cdot j} - \frac{N+1}{2} \right)^2 \\ &= \left(\frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} \right) - 3(N+1), \end{aligned} \tag{6.5}$$

where $(N+1)/2 = \left(\sum_{j=1}^{k} \sum_{i=1}^{n_j} r_{ij} / N \right)$ is the average rank assigned in the joint ranking.

To test

$$H_0 : [\tau_1 = \cdots = \tau_k]$$

versus the general alternative

$$H_1 : [\tau_1, \ldots, \tau_k \text{ not all equal}],$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } H \geq h_\alpha; \text{ otherwise do not reject,} \tag{6.6}$$

where the constant h_α is chosen to make the type I error probability equal to α. The constant h_α is the upper α percentile for the null $(\tau_1 = \cdots = \tau_k)$ distribution of H. Comment 6 explains how to obtain the critical value h_α for k treatments and sample sizes $n_1, \ldots, n_k$ and available levels of α.

Large-Sample Approximation

When H_0 is true, the statistic H has, as $\min(n_1, \ldots, n_k)$ tends to infinity, an asymptotic chi-square (χ^2) distribution with $k-1$ degrees of freedom (see Comment 9 for indications of the proof). The chi-square approximation for procedure (6.6) is

$$\text{Reject } H_0 \text{ if } H \geq \chi^2_{k-1,\alpha}; \quad \text{otherwise do not reject,} \tag{6.7}$$

where $\chi^2_{k-1,\alpha}$ is the upper α percentile point of a chi-square distribution with $k-1$ degrees of freedom. To find $\chi^2_{k-1,\alpha}$, we use the R command `qchisq(`$1-\alpha, k-1$`)`. For example, to find $\chi^2_{4,.025}$, we apply `qchisq (.975, 4)` and obtain $\chi^2_{4,.025} = 11.143$.

Ties

If there are ties among the N X's, assign each of the observations in a tied group the average of the integer runks that are associated with the tied group and compute H with these average ranks. As a consequence of the effect that ties have on the null distribution of H, the following modification is needed to apply either procedure (6.6) or the large-sample approximation in procedure (6.7) when there are tied X's. In either of these procedures, we replace H by

$$H' = \frac{H}{1 - \left(\sum_{j=1}^{g} \left(t_j^3 - t_j\right) / [N^3 - N]\right)}, \tag{6.8}$$

where, in (6.8), H is computed using average ranks, g denotes the number of tied X groups, and t_j is the size of tied group j. We note that an united observation is considered to be a tied group of size 1. In particular, if there are no ties among the X's then $g = N$ and $t_j = 1$ for $j = 1, \ldots, N$. In this case, each term in (6.8) of he form $t_j^3 - t_j$ reduces to zero, the denominator of the right-hand side of expression (6.8)) reduces to 1, and H' (6.8) reduces to H, as given in (6.5).

We note that even the small-sample procedure (6.6) is only approximately, and not exactly, of the significance level α in the presence of tied X observations. To get an exact level α test in this tied setting, see Comment 8.

EXAMPLE 6.1 ***Half-Time of Mucociliary Clearance.***

Thomson and Short (1969) have assessed mucociliary efficiency from the rate of removal of dust in normal subjects, subjects with obstructive airway disease, and subjects with asbestosis. Table 6.1 is based on a subset of the Thomson–Short data. The joint ranks (r_{ij}'s) of the observations are given in Table 6.1 in parentheses after the data values and the treatment rank sums (R_1, R_2, and R_3) are provided at the bottom of the columns.

We are interested in using procedure (6.6) to test if there are any differences in median mucociliary clearance half-times for the three subject populations. For purpose of illustration, we take the significance level to be $\alpha = .0502$. Applying the R command `cKW(`$\alpha, \boldsymbol{n}$`)`, we find `cKW(.0502, c(5, 4, 5), "Exact")` = 5.643; that is, $P_0(H \geq 5.643) = .0502$, and, in the notation of (6.6) with $k = 5, n_1 = 5, n_2 = 4$, and $n_3 = 5$, we have $h_{.0502} = 5.643$ and procedure (6.6) reduces to

$$\text{Reject } H_0 \text{ if } H \geq 5.643.$$

Table 6.1 Half-Time of Mucociliary Clearance (h)

	Subjects with	
Normal subjects	Obstructive airways disease	Asbestosis
2.9 (8)	3.8 (13)	2.8 (7)
3.0 (9)	2.7 (6)	3.4 (11)
2.5 (4)	4.0 (14)	3.7 (12)
2.6 (5)	2.4 (3)	2.2 (2)
3.2 (10)		2.0 (1)
$R_1 = 36$	$R_2 = 36$	$R_3 = 33$

Source: M. L. Thomson and M. D. Short (1969).

Now, we illustrate the computations leading to the sample value of H (6.5). For these data, we have $n_1 = n_3 = 5$, $n_2 = 4$, and $N = 14$. Combining these facts with the treatment rank sums in Table 6.1, we find from (6.5) that

$$H = \frac{12}{14(14+1)}\left(\frac{(36)^2}{5} + \frac{(36)^2}{4} + \frac{(33)^2}{5}\right) - 3(14+1) = .771.$$

As this value of H is less than the critical value 5.643, we do not reject H_0 at the $\alpha = .0502$ level. In fact, from the observed value $H = .771$, we see, using the R command `pKW(mucociliary,"Exact")`, that $P_0(H \geq .771) =$ `pKW(mucociliary,"Exact")` $= .7108$. Thus, the lowest significance level at which we can reject H_0 in favor of H_1 with the observed value of the test statistic $H = .771$ is .7108.

For the large-sample approximation, we compare the value of H (because there are no ties) to the chi-square distribution with $k - 1 = 2$ degrees of freedom. Using the R command `1 - pchisq(.771, 2)`, we find that the observed value of $H = .771$ is approximately the .68 upper percentile for the chi-square distribution with two degrees of freedom. Thus, the approximate P-value for these data and test procedure (6.7) is .68. Both the exact test and the large-sample approximation indicate that there is virtually no sample evidence in support of significant differences in mucociliary clearance half-times for the three subject populations.

Comments

1. *More General Setting.* We could replace Assumptions A1–A3 and H_0 (6.2) with the more general null hypothesis that all $N!/\left(\Pi_{j=1}^{k} n_j!\right)$ assignments of n_1 ranks to the treatment 1 observations, n_2 ranks to the treatment 2 observations, and $\ldots, n_k$ ranks to the treatment k observations are equally likely.
2. *Motivation for the Test.* Under Assumptions A1–A3 and H_0 (6.2), the rank vector $\mathbf{R}^* = (r_{11}, \ldots, r_{n_1 1}, r_{22}, \ldots, r_{n_2 2}, \ldots, r_{1k}, \ldots, r_{n_k k})$ has a uniform distribution over the set of all $N!$ permutations of the vector of integers $(1, 2, \ldots, N)$. It follows that

$$E_0(r_{ij}) = \frac{1}{N!}(N-1)! \sum_{i=1}^{N} i = \frac{N+1}{2},$$

the average rank being assigned in the joint ranking. Thus,

$$E_0(R_{.j}) = E_0\left(\frac{1}{n_j}\sum_{i=1}^{n_j} r_{ij}\right) = \frac{1}{n_j}\sum_{i=1}^{n_j} E_0(r_{ij})$$
$$= \frac{n_j(N+1)}{2n_j} = \frac{N+1}{2}, \text{ for } j = 1, 2, \ldots, k,$$

and we would expect the $R_{.j}$'s to be close to $(N+1)/2$ when H_0 is true. As the test statistic H (6.5) is a constant times a weighted sum of squared differences between the observed treatment average ranks, $R_{.j}$, and their null expected values, $E_0(R_{.j}) = (N+1)/2$, small values of H represent agreement with H_0 (6.2). When the τ's are not all equal, we would expect a portion of the associated treatment average ranks to differ from their common null expectation, $(N+1)/2$, with some tending to be larger and some smaller. The net result (after squaring the observed differences to obtain the $(R_{.j} - (N+1)/2)^2$ terms) would be a large value of H. This suggests rejecting H_0 in favor of H_1 (6.3) for large values of H and motivates procedures (6.6) and (6.7) (see also Comment 3).

3. *Connection to Normal Theory Test.* The Kruskal–Wallis test can also be motivated by considering the usual analysis of variance $\mathcal{F}$ statistic calculated using the ranks, rather than the original observations. The $\mathcal{F}$ statistic can be written as $\mathcal{F} = c(\text{SSB})/(\text{SST} - \text{SSB})$, where c is a constant depending only on the sample sizes, SST is the total sum of squares, and SSB is the between sum of squares. The statistic SSB reduces to $\sum_{j=1}^{k} n_j(R_{.j} - (N+1)/2)^2$ when applied to the ranks rather than the original observations and SST becomes a fixed constant when calculated on the ranks. Using these facts, it can be shown that when $\mathcal{F}$ is calculated for the ranks, $\mathcal{F}$ is an increasing function of H.

4. *Assumptions*. It is important to point out that Assumption A3 stipulates that the k treatment distributions $F_1, \ldots, F_k$ can differ at most in their locations (medians). In particular, Assumption A3 requires that the k underlying distributions belong to the same general family (F) and that they do not differ in scale parameters (variability). (For a discussion of methodology designed for a more general setting where differences in scale parameters are permitted, see Comment 11.)

5. *Special Case of Two Treatments.* For the case of $k = 2$ treatments, the procedures in (6.6) and (6.7) are equivalent to the exact and large-sample approximation forms, respectively, of the two-sided Wilcoxon rank sum test, as discussed in Section 4.1.

6. *Derivation of the Distribution of H under H_0 (No-Ties Case).* The null distribution of H (6.5) can be obtained using the fact that under H_0 (6.2), all $N!/\left(\Pi_{j=1}^{k} n_j!\right)$ assignments of n_1 ranks to the treatment 1 observations, n_2 ranks to the treatment 2 observations, and $\ldots, n_k$ ranks to the treatment k observations are equally likely. We illustrate how the null distribution can be derived in the particular case $k = 3$, $n_1 = n_2 = n_3 = 2$. In this case, we have $H = [\{12/[6(7)]\}\{(R_1^2 + R_2^2 + R_3^2)/2\} - 21] = [(A/7) - 21]$, where $A = R_1^2 + R_2^2 + R_3^2$. We next enumerate 15 of the total possible $\{6!/[(2!)(2!)(2!)]\} = 90$ rank assignments and their corresponding values of A and H.

(a)	I	II	III		(b)	I	II	III	
	1	3	5	$A = 179$		1	3	4	$A = 173$
	2	4	6	$H = 4.57$		2	5	6	$H = 3.71$
(c)	I	II	III		(d)	I	II	III	
	1	3	4	$A = 171$		1	2	5	$A = 173$
	2	6	5	$H = 3.43$		3	4	6	$H = 3.71$
(e)	I	II	III		(f)	I	II	III	
	1	2	4	$A = 165$		1	2	4	$A = 161$
	3	5	6	$H = 2.57$		3	6	5	$H = 2$
(g)	I	II	III		(h)	I	II	III	
	1	2	3	$A = 155$		1	2	5	$A = 171$
	4	5	6	$H = 1.14$		4	3	6	$H = 3.43$
(i)	I	II	III		(j)	I	II	III	
	1	2	3	$A = 153$		1	2	4	$A = 161$
	4	6	5	$H = .86$		5	3	6	$H = 2$
(k)	I	II	III		(l)	I	II	III	
	1	2	3	$A = 153$		1	2	3	$A = 149$
	5	4	6	$H = .86$		5	6	4	$H = .29$
(m)	I	II	III		(n)	I	II	III	
	1	2	4	$A = 155$		1	2	3	$A = 149$
	6	3	5	$H = 1.14$		6	4	5	$H = .29$
(o)	I	II	III						
	1	2	4	$A = 147$					
	6	5	3	$H = 0$					

For each of the foregoing rank configurations, there are five other configurations (corresponding to the six permutations of the names of the samples I, II, and III), which yield the same value of H. This covers the complete total of 90 possible rank assignments. Thus,

$$P_0\{H = 4.57\} = 1/15, \quad P_0\{H = 3.71\} = 2/15, \quad P_0\{H = 3.43\} = 2/15,$$
$$P_0\{H = 2.57\} = 1/15, \quad P_0\{H = 2\} = 2/15, \quad P_0\{H = 1.14\} = 2/15,$$
$$P_0\{H = .86\} = 2/15, \quad P_0\{H = .29\} = 2/15, \quad P_0\{H = 0\} = 1/15.$$

The probability, under H_0, that H is greater than or equal to 3.71, for example, is therefore

$$P_0\{H \geq 3.71\} = P_0\{H = 3.71\} + P_0\{H = 4.57\}$$
$$= \tfrac{1}{15} + \tfrac{2}{15} = .20.$$

Note that we have derived the null distribution of H without specifying the common from (F) of the underlying distribution function for the X's under H_0 beyond the point of requiring that it be continuous. This is why the test procedure (6.6) based on H is called a *distribution-free procedure*. From the null distribution of H, we can determine the critical value h_α and control the probability α of

falsely rejecting H_0 when H_0 is true, and this error probability does not depend on the specific form of the common underlying continuous X distribution.

For a given number of treatments k and sample sizes $n_1, \ldots, n_k$, the R command $\texttt{cKW}(\alpha, \boldsymbol{n})$ can be used to find the available upper-tail critical values h_α for possible values of H. For a given available significance level α, the critical value h_α then corresponds to $P_0(H \geq h_\alpha) = \alpha$ and is given by $\texttt{cKW}(\alpha, \boldsymbol{n})$. Thus, for example, for $k = 5,\ n_1 = 3,\ n_2 = 2,\ n_3 = 3,\ n_4 = 2$, and $n_5 = 3$, we have $P_0(H \geq 8.044) = .0492$, so that $h_{.0492} = 8.044$ for $k = 5,\ n_1 = 3,\ n_2 = 2,\ n_3 = 3,\ n_4 = 2$, and $n_5 = 3$.

7. *Exact Conditional Distribution of H with Ties among the X-Values.* To have a test with the exact significance level even in the presence of tied X's, we need to consider all $N!/\left(\prod_{j=1}^{k} n_j!\right)$ assignments of n_1 ranks to the treatment 1 observations, n_2 ranks to the treatment 2 observations, $\ldots, n_k$ ranks to the treatment k observations, where now these joint ranks are obtained by using average ranks to break the ties. As in Comment 6, it still follows that under H_0 each of these $N!/\left(\prod_{j=1}^{k} n_j!\right)$ assignments is equally likely. For each such assignment, the value of H is computed and the results are tabulated. We illustrate this construction for $k = 3$ and $n_1 = n_2 = 2,\ n_3 = 1$ and the data $X_{11} = 1.3, X_{21} = 1.7, X_{12} = 1.3, X_{22} = 2.0$, and $X_{13} = 2.0$. Using average ranks to break the ties, the observed rank vector is $(r_{11}, r_{21}, r_{12}, r_{22}, r_{13}) = (1.5, 3, 1.5, 4.5, 4.5)$. Thus, $R_1 = 4.5,\ R_2 = 6,\ R_3 = 4.5$, and the attained value of H is

$$H = \left[\frac{12}{5(6)}\left\{\frac{(4.5)^2}{2} + \frac{(6)^2}{2} + \frac{(4.5)^2}{1}\right\} - 3(6)\right] = 1.35.$$

To assess the significance of H, we obtain its conditional null distribution by considering the [5!/(2! 2! 1!)] = 30 equally likely (under H_0) possible assignments of the observed rank vector (1.5, 3, 1.5, 4.5, 4.5) to the three treatments. These 30 assignments and associated values of H are in the following table

I	II	III		I	I	III	
1.5	4.5	1.5		1.5	4.5	1.5	
3	4.5		$H = 3.15$	3	4.5		$H = 3.15$
1.5	3	1.5		1.5	3	1.5	
4.5	4.5		$H = 1.35$	4.5	4.5		$H = 1.35$
1.5	3	1.5		1.5	3	1.5	
4.5	4.5		$H = 1.35$	4.5	4.5		$H = 1.35$
3	1.5	1.5		3	1.5	1.5	
4.5	4.5		$H = 1.35$	4.5	4.5		$H = 1.35$
3	1.5	1.5		3	1.5	1.5	
4.5	4.5		$H = 1.35$	4.5	4.5		$H = 1.35$
4.5	1.5	1.5		4.5	1.5	1.5	
4.5	3		$H = 3.15$	4.5	3		$H = 3.15$
1.5	4.5	3		1.5	1.5	3	
1.5	4.5		$H = 3.60$	4.5	4.5		$H = 0$
1.5	1.5	3		1.5	1.5	3	
4.5	4.5		$H = 0$	4.5	4.5		$H = 0$
1.5	1.5	3		4.5	1.5	3	

I	II	III		I	I	III	
4.5	4.5		$H = 0$	4.5	1.5		$H = 3.60$
1.5	3	4.5		1.5	3	4.5	
1.5	4.5		$H = 3.15$	1.5	4.5		$H = 3.15$
1.5	1.5	4.5		1.5	1.5	4.5	
3	4.5		$H = 1.35$	3	4.5		$H = 1.35$
1.5	1.5	4.5		1.5	1.5	4.5	
3	4.5		$H = 1.35$	3	4.5		$H = 1.35$
1.5	1.5	4.5		1.5	1.5	4.5	
4.5	3		$H = 1.35$	4.5	3		$H = 1.35$
1.5	1.5	4.5		1.5	1.5	4.5	
4.5	3		$H = 1.35$	4.5	3		$H = 1.35$
3	1.5	4.5		3	1.5	4.5	
4.5	1.5		$H = 3.15$	4.5	1.5		$H = 3.15$

As each of these values for H has null probability $\frac{1}{30}$, it follows that

$$P_0(H = 3.60) = \tfrac{2}{30} \qquad P_0(H = 1.35) = \tfrac{16}{30}$$
$$P_0(H = 3.15) = \tfrac{8}{30} \qquad P_0(H = 0) = \tfrac{4}{30}.$$

This distribution is called the *conditional distribution* or the *permutation distribution* of H, given the set of tied ranks {1.5, 1.5, 3, 4.5, and 4.5}. For the particular observed value $H = 1.35$, we have $P_0(H \geq 1.35) = \frac{28}{30}$, so that such a value does not indicate a deviation from H_0.

8. *Large-Sample Approximation*. Define the random variables $T_j = R_{.j} - E_0(R_{.j}) = R_{.j} - (N + 1)/2$, for $j = 1, 2, \ldots, k$. As each $R_{.j} = \sum_{i=1}^{n_j} r_{ij}/n_j$ is an average, it is not surprising (see Kruskal and Wallis (1952), e.g., for justification) that a properly standardized version of the vector $\mathbf{T}^* = (T_1, \ldots, T_{k-1})$ has an asymptotic ($\min(n_1, \ldots, n_k)$ tending to infinity) $(k - 1)$-variate normal distribution with mean vector $\mathbf{0} = (0, \ldots, 0)$ and appropriate covariance matrix $\boldsymbol{\Sigma}$ when the null hypothesis H_0 is true. (Note that $\mathbf{T}^*$ does not include $T_k = R_{.k} - (N + 1)/2$, because T_k can be expressed as a linear combination of $T_1, \ldots, T_{k-1}$. This is the reason that the asymptotic normal distribution is $(k - 1)$-variate and not k-variate.) As the test statistic H (6.5) is a quadratic form in the variables $(T_1, \ldots, T_{k-1})$, it is therefore quite natural that H has an asymptotic ($\min(n_1, \ldots, n_k)$ tending to infinity) chi-square distribution with $k - 1$ degrees of freedom.
9. *Family Monotonicity*. Gabriel (1969) introduced a desirable property of a testing family called *monotonicity* and pointed out that the H statistic does not enjoy the property. We refer the interested user to Gabriel's paper, but we briefly mention here that the problem arises because it is possible that the H statistic computed for a subset can exceed the H statistic computed for a set containing the subset. Gabriel gave the following example. The sample 1 ranks are 8, 9, 10, and 11, the sample 2 ranks are 1, 2, 6, and 7, and the sample 3 ranks are 3, 4, 5, and12. Then H based on samples 1 and 2($k = 2$) is 5.33, whereas H based on samples 1, 2, and 3 ($k = 3$) is 4.77. The same anomaly can arise with the Friedman statistic (Section 7.1).

10. *k-Sample Behrens–Fisher Problem*. Two of the implicit requirements associated with Assumptions A1–A3 are that the underlying distributions belong to the same common family (F) and that they differ within this family at most in their medians. The less restrictive setting, where these assumptions are relaxed to permit the possibility of differences in scale parameters as well as medians (but still requiring the same common family F), is generally referred to as the *k-sample Behrens–Fisher problem*. (Note that this is a direct k-sample extension of the corresponding two-sample Behrens–Fisher problem considered in Section 4.4.) The Kruskal–Wallis procedure (6.6) is no longer distribution-free under these relaxed assumptions permitting unequal scale parameters. Rust and Fligner (1984) proposed a modification of the Kruskal–Wallis statistic H (6.5) to deal with this broader Behrens–Fisher setting. Their procedure is designed as a test for the less restrictive null and alternative hypotheses

$$H_0^* : [\delta_{ij} = \tfrac{1}{2} \text{ for all } i \neq j = 1, \ldots, k] \tag{6.9}$$

and

$$H_1^* : [\delta_{ij} \neq \tfrac{1}{2} \text{ for at least one } i \neq j = 1, \ldots, k], \tag{6.10}$$

respectively, where

$$\delta_{ij} = P(X_{1i} > X_{1j}), \quad \text{for } i \neq j = 1, \ldots, k.$$

The Rust–Fligner modification of the Kruskal–Wallis statistic provides a test procedure that is still exactly distribution-free under the more restrictive null hypothesis H_0 (6.2). However, their modified procedure is also asymptotically ($\min(n_1, \ldots, n_k)$ tending to infinity) distribution-free under the considerably broader null hypothesis H_0^* (6.9) so long as the underlying populations (not necessarily of the same form) are all symmetric. In the special case of $k = 2$ populations, the Rust–Fligner procedure reduces approximately to the Fligner–Policello modifications to the Mann–Whitney–Wilcoxon two-sample test procedure discussed in Section 4.4.

11. *Pairwise Rankings*. The Kruskal–Wallis statistic H (6.5) is based on the treatment rank sums $R_1, \ldots, R_k$ associated with the *joint* ranking of all N sample observations. As an alternative approach, one could just as well choose to compare the k treatments through a combination of all $k(k-1)/2$ *pairwise* rankings. Fligner (1985) proposed such a pairwise ranking analog of the Kruskal–Wallis statistic and demonstrated that the associated pairwise ranking test procedure has some nice efficiency properties. Such pairwise rankings (as opposed to joint rankings) have also proved useful in certain multiple comparison settings (see Sections 6.5 and 6.10 for more in this regard).

12. *Consistency of the H Test*. Replace Assumptions A1–A3 by the less restrictive Assumptions A1′: the X's are mutually independent and A2′ : $X_{1j}, \ldots, X_{n_j j}$ come from the same continuous population $\Pi_j, j = 1, \ldots, k$, but where $\Pi_1, \ldots, \Pi_k$ are not assumed to be identical. Then Kruskal and Wallis (1952) pointed out that (roughly speaking) the test defined by (6.6) is consistent if (and only if) "... there be at least one of the populations for which the limiting probability is not one-half that a random observation from this population is greater than an independent random member of the N sample observations."

Properties

1. *Consistency*. Under Assumptions A1–A3 and equal sample sizes ($n_1 = \cdots = n_k$), the test defined by (6.6) is consistent against the alternative for which $\tau_i \neq \tau_j$ for at least one $i \neq j = 1, \ldots, k$. For arbitrary sample sizes, see Kruskal (1952) and Comment 12.
2. *Asymptotic Chi-Squareness*. See Kruskal and Wallis (1952) and Hettmansperger (1984, pp. 184–185).
3. *Efficiency*. See Andrews (1954), Hodges and Lehmann (1956), and Section 6.10.

Problems

1. Pretherapy training of clients has been shown to have beneficial effects on the process and outcome of counseling and psychotherapy. Sauber (1971) investigated four different approaches to pretherapy training:

1. Control (no treatment).
2. Therapeutic reading (TR) (indirect learning).
3. Vicarious therapy pretraining (VTP) (videotaped, vicarious learning).
4. Group, role induction interview (RII) (direct learning).

Treatment conditions 2–4 were expected to enhance the outcome of counseling and psychotherapy as compared with a control group, those subjects receiving no prior set of structuring procedures. One of the major variables of the study was that of "psychotherapeutic attraction." The basic data in Table 6.2 consist of the raw scores for this measure according to each of the four experimental conditions. Apply procedure (6.7), with the correction for ties given by (6.8).

2. Show that the two expressions for H in (6.5) are indeed equivalent.

3. Show directly, or illustrate by means of an example, that the maximum value of H is $H_{\max} = \{N^3 - \sum_{j=1}^{k} n_j^3\}/\{N(N+1)\}$. For what rank configurations is this maximum achieved?

4. To determine the number of game fish to stock in a given system and to set appropriate catch limits, it is important for fishery managers to be able to assess potential growth and survival of game fish in that system. Such growth and survival rates are closely related to the availability of appropriately sized prey. Young-of-year (YOY) gizzard shad (*Dorosoma cepedianum*) are the primary food source for game fish in many Ohio environments. However, because of their fast growth rate, YOY gizzard shad can quickly become too large for predators to swallow.

Table 6.2 Raw Scores Indicating the Degree of Psychotherapeutic Attraction for Each Experimental Condition

Control	Reading (TR)	Videotape (VTP)	Group (RII)
0	0	0	1
1	6	5	5
3	7	8	12
3	9	9	13
5	11	11	19
10	13	13	22
13	20	16	25
17	20	17	27
26	24	20	29

Source: S. R. Sauber (1971).

Table 6.3 Length of YOY Gizzard Shad from Kokosing Lake, Ohio, Sampled in Summer, 1984 (mm)

Site I	Site II	Site III	Site IV
46	42	38	31
28	60	33	30
46	32	26	27
37	42	25	29
32	45	28	30
41	58	28	25
42	27	26	25
45	51	27	24
38	42	27	27
44	52	27	30

Source: B. Johnson (1984).

Thus to be able to predict predator growth rates in such settings, it is useful to know both the density and the size structure of the resident YOY shad populations. With this in mind, Johnson (1984) sampled the YOY gizzard shad population at four different sites in Kokosing Lake (Ohio) in summer 1984. The data in Table 6.3 are lengths (mm) for a subset of the YOY gizzard shad sampled by Johnson.

Apply procedure (6.7), with the correction for ties given in (6.8), to assess whether there are any differences between the median lengths for the YOY gizzard shad populations in the four Kokosing Lake sites.

5. Suppose $k = 3$ and $n_1 = 2$, $n_2 = n_3 = 6$. Compare the critical region for the exact level $\alpha = .050$ test of H_0 (6.2) based on H with the critical region for the corresponding nominal level $\alpha = .050$ test based on the large-sample approximation. What is the exact significance level of this .050 nominal level test based on the large-sample approximation?

6. Suppose $k = 4$, $n_1 = n_2 = n_3 = 1$, and $n_4 = 2$. Obtain the form of the exact null (H_0) distribution of H for the case of no tied X observations.

7. Suppose $k = 3$, $n_1 = n_2 = n_3 = 2$, and we observe the data $X_{11} = 2.7$, $X_{21} = 3.4$, $X_{12} = 2.7$, $X_{22} = 4.5$, $X_{13} = 4.9$, and $X_{23} = 2.7$. What is the conditional probability distribution of H under H_0 (6.2) when average ranks are used to break the ties among the X's? How extreme is the observed value of H in this conditional null distribution? Compare this fact with that obtained by taking the observed value of H to the (incorrect) unconditional null distribution of H.

8. Leukemia is a disease characterized by proliferation of the white blood cells or leukocytes. One form of chemotherapy used in the treatment of leukemia involves the administration of corticosteroids. Some researchers suggested that forms of leukemia characterized by leukocytes with a large number of glucocorticoid receptor (GR) sites per cell are more effectively controlled by corticosteroids. Other researchers questioned this relationship. In an effort to aid in the resolution of this controversy, Kontula et al. (1980) developed a method for determining more accurately the number of GR sites per cell. In this research and later work by Kontula et al. (1982), this new methodology was used to count the number of GR sites for samples of leukocyte cells from normal subjects, as well as patients with hairy-cell leukemia, chronic lymphatic leukemia, chronic myelocytic leukemia, or acute leukemia. The data in Table 6.4 are a subset of the data considered by the authors in these two publications.

   Use these data to assess whether there are any differences between the median numbers of GR sites per leukocyte cell for the population of normal subjects and the populations of patients with hairy-cell leukemia, chronic lymphatic leukemia, chronic myelocytic leukemia, or acute leukemia.

Table 6.4 Number of Glucocorticoid Receptor (GR) Sites per Leukocyte Cell

Normal subjects	Hairy-cell anemia	Chronic lymphatic leukemia	Chronic myelocytic leukemia	Acute leukemia
3,500	5,710	2,930	6,320	3,230
3,500	6,110	3,330	6,860	3,880
3,500	8,060	3,580	11,400	7,640
4,000	8,080	3,880	14,000	7,890
4,000	11,400	4,280		8,280
4,000		5,120		16,200
4,300				18,250
4,500				29,900
4,500				
4,900				
5,200				
6,000				
6,750				
8,000				

Source: K. Kontula, L. C. Andersson, T. Paavonen, G. Myllyla, L. Teerenhovi, and P. Vuopio (1980) and K. Kontula, T. Paavonen, R. Vuopio, and L. C. Andersson (1982).

9. Generate the conditional permutation distribution of H *using only the last two sample lengths from each of the four sites for the gizzard shad data in Table 6.3.* From this conditional permutation distribution of H, obtain the exact conditional P-value for a test of H_0 (6.2) versus H_1 (6.3) with this subset of data from Table 6.3. Compare this exact conditional P-value with the approximate P-value associated with taking the observed value of H to the unadjusted (for ties) unconditional null distribution of H.
10. Habitat plays an important role in fish behavior, particularly feeding, spawning, and protection/security. One of the modern methods of fisheries management is habitat modification in large, constructed reservoirs. Previous studies have shown that the type of structure introduced is an important factor in such habitat modifications. Of particular relevance in many settings is the size openings or *interstices* in the introduced structure. The data in Table 6.5 represent a subset of that obtained by Kayle (1984) from Alum Creek Lake in Westerville, Ohio, in a study to determine the relative effectiveness of three species of pine trees for habitat modification.

Table 6.5 Mean Interstitial Lengths (mm)

Scotch pine	Blue spruce	White pine
52.2	46.7	75.2
56.4	60.5	63.7
57.1	58.9	73.2
46.9	82.9	66.2
49.1	65.8	67.4
52.5	93.3	69.4
63.0	66.9	70.4
52.0	70.9	72.3
61.1	73.7	63.6
55.3	65.8	61.9
46.2	90.2	74.4
57.2	68.9	70.1

Source: K. A. Kayle (1984).

The measurements in Table 6.5 are averages (mm) of interstitial lengths (distances between midpoints) of 10 pairs of secondary branches for each of 12 scotch pine, 12 blue spruce, and 12 white pine trees. Use an appropriate procedure to test whether there are any differences in median interstitial lengths between secondary branches for the three studied species of pine.

6.2 A DISTRIBUTION-FREE TEST FOR ORDERED ALTERNATIVES (JONCKHEERE–TERPSTRA)

In many practical settings, the treatments are such that the appropriate alternatives to no differences in treatment effects (H_0) are those of increasing (or decreasing) treatment effects according to some natural labeling for the treatments. Examples of such settings include "treatments" corresponding to degrees of knowledge of performance, quality or quantity of materials, severity of disease, amount of practice, drug dosage levels, intensity of a stimulus, and temperature. We note that the Kruskal–Wallis procedure (6.6) does not utilize any such partial prior information regarding a postulated alternative ordering. The statistic H (6.5) takes on the same value for all $k!$ possible labelings of the treatments. In this section, we consider a procedure for testing H_0 (6.2) against the a priori ordered alternatives

$$H_2 : [\tau_1 \leq \tau_2 \leq \cdots \leq \tau_k, \text{ with at least one strict inequality}]. \tag{6.11}$$

The Jonckheere (1954a, 1954b) and Terpstra (1952) test of this section is preferred to the Kruskal–Wallis test in Section 6.1 when the treatments can be labeled a priori in such a way that the experimenter expects any deviation from H_0 (6.2) to be in the particular direction associated with H_2 (6.11). We emphasize, however, that the labeling of the treatments so that the ordered alternatives (6.11) are appropriate *cannot* depend on the observed sample observations. This labeling must correspond completely to a factor(s) implicit in the nature of the *experimental design* and *not* the *observed data*.

Procedure

First, we must label the treatments so that they are in the expected order associated with the alternative H_2 (6.11). (This labeling must be done prior to data collection.) To compute the Jonckheere–Terpstra statistic, J, we calculate the $k(k-1)/2$ Mann–Whitney (see Comment 4.7) counts U_{uv} given by

$$U_{uv} = \sum_{i=1}^{n_u} \sum_{j=1}^{n_v} \phi(X_{iu}, X_{jv}), \quad 1 \leq u < v \leq k, \tag{6.12}$$

where $\phi(a,b) = 1$ if $a < b, 0$ otherwise. (Thus, U_{uv} is the number of sample u before sample v precedences.) The Jonckheere–Terpstra statistic, J, is then the sum of these $k(k-1)/2$ Mann–Whitney counts,

$$J = \sum_{u=1}^{v-1} \sum_{v=2}^{k} U_{uv}. \tag{6.13}$$

To test

$$H_0 : [\tau_1 = \cdots = \tau_k]$$

versus the ordered alternative

$$H_2 : [\tau_1 \le \tau_2 \le \cdots \le \tau_k, \text{ with at least one strict inequality}],$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } J \ge j_\alpha; \text{ otherwise do not reject,} \tag{6.14}$$

where the constant j_α is chosen to make the type I error probability equal to α. The constant j_α is the upper α percentile for the null ($\tau_1 = \cdots = \tau_k$) distribution of J. Comment 17 explains how to obtain the critical value j_α for k treatments and sample sizes $n_1, \ldots, n_k$ and available levels of α.

Large-Sample Approximation

The large-sample approximation is based on the asymptotic ($\min(n_1, n_2, \ldots, n_k)$ tending to infinity) normality of J, suitably standardized. We first need to know the expected value and variance of J when the null hypothesis is true. Under H_0, the expected value and variance of J are:

$$E_0(J) = \frac{N^2 - \sum_{j=1}^k n_j^2}{4} \tag{6.15}$$

and

$$\text{var}_0(J) = \frac{N^2(2N+3) - \sum_{j=1}^k n_j^2(2n_j+3)}{72}, \tag{6.16}$$

respectively. These expressions for $E_0(J)$ and $\text{var}_0(J)$ are verified by direct calculations in Comment 18 for the special case of $k = 3,\ n_1 = n_2 = 1,\ n_3 = 2$. General derivations of both expressions are outlined in Comment 19.

The standardized version of J is

$$J^* = \frac{J - E_0(J)}{\sqrt{\text{var}_0(J)}} = \frac{J - \left[\frac{N^2 - \sum_{j=1}^k n_j^2}{4}\right]}{\left\{\left[N^2(2N+3) - \sum_{j=1}^k n_j^2(2n_j+3)\right] \Big/ 72\right\}^{1/2}}. \tag{6.17}$$

When H_0 is true, J^* has, as $\min(n_1, \ldots, n_k)$ tends to infinity, an asymptotic $N(0,1)$ distribution (see Comment 19 for indications of the proof). The normal theory approximation for procedure (6.14) is

$$\text{Reject } H_0 \text{ if } J^* \ge z_\alpha; \quad \text{otherwise do not reject.} \tag{6.18}$$

Ties

If there are ties among the N X's, replace $\phi(a,b)$ in the calculation of the Mann–Whitney counts U_{uv} by $\phi^*(a,b) = 1, \frac{1}{2}, 0$ if $a <, =,$ or $> b$, respectively, so that for each between-sample comparison where there is a tie, the contribution to the appropriate Mann–Whitney count will be $\frac{1}{2}$. After computing J with these modified Mann–Whitney

counts, use procedure (6.14). Note, however, that this test associated with tied X's is only approximately, and not exactly, of the significance level α.

When applying the large-sample approximation, an additional factor must be taken into account. Although ties in the X's do not affect the null expected value of J, its null variance is reduced to

$$\begin{aligned}\mathrm{var}_0(J) &= \left\{ \frac{1}{72}\left[N(N-1)(2N+5) - \sum_{i=1}^{k} n_i(n_i-1)(2n_i+5) - \sum_{j=1}^{g} t_j(t_j-1)(2t_j+5)\right]\right. \\ &\quad + \frac{1}{36N(N-1)(N-2)}\left[\sum_{i=1}^{k} n_i(n_i-1)(n_i-2)\right]\left[\sum_{j=1}^{g} t_j(t_j-1)(t_j-2)\right] \\ &\quad \left. + \frac{1}{8N(N-1)}\left[\sum_{i=1}^{k} n_i(n_i-1)\right]\left[\sum_{j=1}^{g} t_j(t_j-1)\right]\right\}, \end{aligned} \tag{6.19}$$

where, in (6.19), g denotes the number of tied X groups and t_j is the size of tied group j. We note that an untied observation is considered to be a tied group of size 1. In particular, if there are no ties among the X's, then $g = N$ and $t_j = 1$, for $j = 1, \ldots, N$. In this case, each term in (6.19) that involves the factor $(t_j - 1)$ reduces to zero and (as you are asked to show in Problem 19) the variance expression in (6.19) reduces to the usual null variance of J when there are no ties, as given previously in (6.16).

As a consequence of the effect that ties have on the null variance of J, the following modification is needed to apply the large-sample approximation when there are tied X's. Compute J using the modified Mann–Whitney counts and set

$$J^* = \frac{J - \left[\frac{N^2 - \sum_{j=1}^{k} n_j^2}{4}\right]}{\{\mathrm{var}_0(J)\}^{1/2}}, \tag{6.20}$$

where $\mathrm{var}_0(J)$ is now given by display (6.19). With this modified value of J^*, the approximation (6.18) can be applied.

EXAMPLE 6.2 ***Motivational Effect of Knowledge of Performance.***

Hundal (1969) described a study designed to assess the purely motivational effects of knowledge of performance in a repetitive industrial task. The task was to grind a metallic piece to a specified size and shape. Eighteen male workers were divided randomly into three groups. The subjects in the control group, A, received no information about their output, subjects in group B were given a rough estimate of their output, and subjects in group C were given an accurate information about their output and could check it further by referring to a figure that was placed before them. The basic data in Table 6.6 consist of the numbers of pieces processed by each subject in the experimental period.

We apply the Jonckheere–Terpstra test with the notion that a deviation from H_0 is likely to be in the direction of increased output with increased degree of knowledge of

Table 6.6 Number of Pieces Processed

Control (no information)	Group B (rough information)	Group C (accurate information)
40 (5.5)[a]	38 (2.5)	48 (18)
35 (1)	40 (5.5)	40 (5.5)
38 (2.5)	47 (17)	45 (15)
43 (10.5)	44 (13)	43 (10.5)
44 (13)	40 (5.5)	46 (16)
41 (8)	42 (9)	44 (13)

Source: P. S. Hundal (1969).

[a] Although we do not need to perform the joint ranking to compute the Jonckheere–Terpstra statistic, we give these ranks here for use in Sections 6.4 and 6.7.

performance. Thus, we are interested in using procedure (6.14) with the treatment labels 1 ≡ control (no information), 2 ≡ group B (rough information), and 3 ≡ group C (accurate information). For purpose of illustration, we take the significance level to be $\alpha = .0490$. Applying the R command `cJCK`$(\alpha, \boldsymbol{n})$, we find `cJCK(.0490,c(6,6,6))` $= 75$; that is, $P_0(J \geq 75) = .0490$, and, in the notation of (6.14) with $k = 3$, $n_1 = n_2 = n_3 = 6$, we have $j_{.0490} = 75$, and procedure (6.14) reduces to

$$\text{Reject } H_0 \text{ if } J \geq 75.$$

We now illustrate the computations leading to the sample value of J (6.13). As there are ties in the sample data, we use $\phi^*(a, b) = 1, \frac{1}{2}, 0$ if $a <, =,$ or $> b$, respectively, to compute the $3(2)/2 = 3$ Mann–Whitney counts. We obtain

$$U_{12} = 1.5 + 2.5 + 6 + 5.5 + 2.5 + 4 = 22,$$

$$U_{13} = 6 + 2.5 + 6 + 4.5 + 6 + 5.5 = 30.5,$$

and

$$U_{23} = 6 + 2 + 5 + 4 + 5 + 4.5 = 26.5.$$

From (6.13), it follows that

$$J = 22 + 30.5 + 26.5 = 79.$$

As this value of J is greater than the critical value 75, we reject H_0 at the .0490 level. In fact, from the observed value $J = 79$, we see that the R command `pJCK(motivational.effect)` that $P_0(J \geq 79) =$ `pJCK(motivational.effect)` $= .0231$. Thus, the lowest significance level at which we can reject H_0 in favor of H_2 with the observed value of $J = 79$ is the P-value .0231.

For the large-sample approximation, we need to compute the standardized form of J^* using (6.19) and (6.20), because there are ties in the data. The null expected value for J is $E_0(J) = [(18)^2 - (6^2 + 6^2 + 6^2)]/4 = 54$. For the ties-corrected null variance of J, we note that $g = 11$ and $t_1 = 1, t_2 = 2, t_3 = 4, t_4 = 1, t_5 = 1, t_6 = 2, t_7 = 3, t_8 = 1, t_9 = 1,$

$t_{10} = 1, t_{11} = 1$, for the Hundal data. Hence, using the ties correction in (6.19), we have

$$\begin{aligned}\text{var}_0(J) = \Bigg\{ &\frac{1}{72}[18(17)(41) - 3(6)(5)(17) - 2(2)(1)(9) - 3(2)(11) - 4(3)(13)] \\ &+ \frac{1}{36(18)(17)(16)}[3(6)(5)(4)][3(2)(1) + 4(3)(2)] \\ &+ \frac{1}{8(18)(17)}[3(6)(5)][2(2)(1) + 1(3)(2) + 1(4)(3)]\Bigg\} = 150.29,\end{aligned}$$

from which it follows that the ties-corrected value of J^* (6.20) is

$$J^* = \frac{79 - 54}{\{150.29\}^{1/2}} = 2.04.$$

Thus, using the approximate procedure (6.18) with the ties-corrected value of $J^* = 2.04$ and the R command `pnorm(·)`, we see that the approximate P-value for these data is $P_0(J^* \geq 2.04) \approx 1 - \texttt{pnorm}(2.04) = .0207$. Both the exact test and the large-sample approximation indicate that strong evidence in support of increased output with increase in degree of knowledge of performance for the task considered by Hundal.

Comments

13. *More General Setting*. As with the Kruskal–Wallis procedure in Section 6.1, we could replace Assumptions A1–A3 and H_0 (6.2) with the more general null hypothesis that all $N!/\left(\prod_{j=1}^{k} n_j!\right)$ assignments of n_1 joint ranks to the treatment 1 observations, n_2 joint ranks to the treatment 2 observations, $\ldots, n_k$ joint ranks to the treatment k observations are equally likely.
14. *Motivation for the Test*. Consider J (6.13) and note that the term $\sum_{u=1}^{v-1}\sum_{v=2}^{k} U_{uv}$ takes the postulated ordering into account. Consider, for simplicity, the case $k = 3$. Then $\sum_{u=1}^{v-1}\sum_{v=2}^{3} U_{uv} = U_{12} + U_{13} + U_{23}$ and if $\tau_1 < \tau_2 < \tau_3, U_{12}$ would tend to be larger than $n_1n_2/2$ (its null expectation); U_{13} would tend to be larger than $n_1n_3/2$; U_{23} would tend to be larger than $n_2n_3/2$; and, consequently, $J = U_{12} + U_{13} + U_{23}$ would tend to be larger than its null expectation $(n_1n_2 + n_1n_3 + n_2n_3)/2 = \{[N^2 - (n_1^2 + n_2^2 + n_3^2)]/4\}$. This serves as partial motivation for the J test.
15. *Assumptions*. It is once again (as with the Kruskal–Wallis procedure in Section 6.1) important to point out that Assumption A3 stipulates that the k treatment distributions $F_1, \ldots, F_k$ can differ at most in their locations (medians) (see also Comment 4).
16. *Special Case of Two Treatments*. When there are only two treatments, the procedures in (6.14) and (6.18) are equivalent to the exact and large-sample approximation forms, respectively, of the one-sided upper-tail Wilcoxon rank sum test, as discussed in Section 4.1.
17. *Derivation of the Distribution of J under H_0 (No-Ties Case)*. A little thought will convince the reader that J can be computed from the joint ranking of all $N = \sum_{j=1}^{k} n_j$ observations. That is, although we do not need to perform this

joint ranking in order to compute J, given the ranking, we can, without the knowledge of the actual X_{ij} values, retrieve the value of J. Thus, one way to obtain the null distribution of J is to follow the method of Comment 6; namely, use the fact that under H_0 (6.2) all $N!/\left(\Pi_{j=1}^{k} n_j!\right)$ rank assignments are equally likely, and compute the associated value of J for each possible ranking. Consider how this would work in the small-sample-size case of $k = 3, n_1 = 1, n_2 = 1$, and $n_3 = 2$. The $4!/[1!\ 1!\ 2!] = 12$ possible assignments of the joint ranks 1, 2, 3, and 4 to the three treatments and their associated values of J (6.13) are as follows:

(a)	I	II	III	(b)	I	II	III	(c)	I	II	III	
	1	2	3		2	1	3		1	3	2	
			4				4				4	
	$J = 5$				$J = 4$				$J = 4$			
(d)	I	II	III	(e)	I	II	III	(f)	I	II	III	
	3	1	2		1	4	2		4	1	2	
			4				3				3	
	$J = 3$				$J = 3$				$J = 2$			
(g)	I	II	III	(h)	I	II	III	(i)	I	II	III	
	2	3	1		3	2	1		2	4	1	
			4				4				3	
	$J = 3$				$J = 2$				$J = 2$			
(j)	I	II	III	(k)	I	II	III	(l)	I	II	III	
	4	2	1		3	4	1		4	3	1	
			3				2				2	
	$J = 1$				$J = 1$				$J = 0$			

Thus, the null distribution for J for $n_1 = 1, n_2 = 1, n_3 = 2$, and $k = 3$ is given by

$$P_0\{J = 0\} = \tfrac{1}{12}, \quad P_0\{J = 1\} = \tfrac{2}{12}, \quad P_0\{J = 2\} = \tfrac{3}{12},$$
$$P_0\{J = 3\} = \tfrac{3}{12}, \quad P_0\{J = 4\} = \tfrac{2}{12}, \quad P_0\{J = 5\} = \tfrac{1}{12}.$$

The probability, under H_0, that J is greater than or equal to 4, for example, is therefore

$$P_0\{J \geq 4\} = P_0\{J = 4\} + P_0\{J = 5\}$$
$$= \tfrac{1}{12} + \tfrac{2}{12} = .25.$$

Note that we have derived the null distribution of J without specifying the common form (F) of the underlying distribution function for the X's under H_0 beyond the requirement that it be continuous. This is why the test procedure (6.14) based on J is called a *distribution-free procedure*. From the null distribution of J, we can determine the critical value j_α and control the probability α of

falsely rejecting H_0 when H_0 is true, and this error probability does not depend on the specific form of the common underlying continuous X distribution.

For a given number of treatments k and sample sizes $n_1, \ldots, n_k$, the R command cJCK$(\alpha, \boldsymbol{n})$ can be used to find the available upper-tail critical values j_α for possible values of J. For a given available significance level α, the critical value j_α then corresponds to $P_0(J \geq j_\alpha) = \alpha$ and is given by cJCK$(\alpha, \boldsymbol{n})$. Thus, for example, for $k = 3,\ n_1 = 6,\ n_2 = 5$, and $n_3 = 7$, we have $P_0(J \geq 79) = .0204$, so that $j_{.0204} = 79$ for $k = 3,\ n_1 = 6,\ n_2 = 5$, and $n_3 = 7$.

18. *Calculation of the Mean and Variance of J under the Null Hypothesis H_0.* In displays (6.15) and (6.16), we presented formulas for the mean and variance of J when the null hypothesis is true. In this comment, we illustrate a direct calculation of $E_0(J)$ and $\text{var}_0(J)$ in the particular case of $k = 3$ and $n_1 = n_2 = 1, n_3 = 2$ and no tied observations, using the null distribution of J obtained in Comment 17. (Later, in Comment 19, we present arguments for the general derivations of $E_0(J)$ and $\text{var}_0(J)$.) The null mean, $E_0(J)$, is obtained by multiplying each possible value of J with its probability under H_0. Thus,

$$E_0(J) = 0\left(\tfrac{1}{12}\right) + 1\left(\tfrac{2}{12}\right) + 2\left(\tfrac{3}{12}\right) + 3\left(\tfrac{3}{12}\right) + 4\left(\tfrac{2}{12}\right) + 5\left(\tfrac{2}{12}\right) = 2.5.$$

This is in agreement with what we obtain using (6.15), namely,

$$E_0(J) = \frac{4^2 - \{1^2 + 2^2 + 1^2\}}{4} = 2.5.$$

A check on the expression for $\text{var}_0(J)$ is also easy, using the well-known fact that

$$\text{var}_0(J) = E_0(J^2) - \{E_0(J)\}^2.$$

The value of $E_0(J^2)$, the second moment of the null distribution of J, is again obtained by multiplying possible values (in this case, of J^2) by the corresponding probabilities under H_0. We find

$$E_0(J^2) = 0^2\left(\tfrac{1}{12}\right) + 1^2\left(\tfrac{2}{12}\right) + 2^2\left(\tfrac{3}{12}\right) + 3^2\left(\tfrac{3}{12}\right) + 4^2\left(\tfrac{2}{12}\right) + 5^2\left(\tfrac{1}{12}\right) = \tfrac{49}{6}.$$

Thus,

$$\text{var}_0(J) = \tfrac{49}{6} - (2.5)^2 = \tfrac{23}{12} = 1.92,$$

which agrees with what we obtain using (6.16) directly, namely,

$$\text{var}_0(J) = \frac{\{4^2(2(4) + 3) - [1^2(2(1) + 3) + 1^2(2(1) + 3) + 2^2(2(2) + 3)]\}}{72} = 1.92.$$

19. *Large-Sample Approximation*. From the definition of J (6.13) and U_{uv} (6.12), we see that

$$\begin{aligned} E(J) &= E\left[\sum_{u=1}^{v-1}\sum_{v=2}^{k} U_{uv}\right] = E\left[\sum_{u=1}^{v-1}\sum_{v=2}^{k}\sum_{i=1}^{n_u}\sum_{j=1}^{n_v}\phi(X_{iu}, X_{jv})\right] \\ &= \sum_{u=1}^{v-1}\sum_{v=2}^{k}\sum_{i=1}^{n_u}\sum_{j=1}^{n_v} E[\phi(X_{iu}, X_{jv})] \\ &= \sum_{u=1}^{v-1}\sum_{v=2}^{k}\sum_{i=1}^{n_u}\sum_{j=1}^{n_v} P(X_{iu} < X_{jv}) \\ &= \sum_{u=1}^{v-1}\sum_{v=2}^{k} n_u n_v P(X_{1u} < X_{1v}). \end{aligned} \tag{6.21}$$

Under the null hypothesis H_0 (6.2), $P_0(X_{1u} < X_{1v}) = \frac{1}{2}$ for every $1 \le u < v \le k$. It follows that

$$\begin{aligned} E_0(J) &= \sum_{u=1}^{v-1}\sum_{v=2}^{k}\frac{(n_u n_v)}{2} = \frac{1}{4}\sum_{\substack{u=1\\u\neq v}}^{k}\sum_{v=1}^{k} n_u n_v \\ &= \frac{1}{4}\left[\sum_{u=1}^{k}\sum_{v=1}^{k} n_u n_v - \sum_{t=1}^{k} n_t^2\right] \\ &= \frac{1}{4}\left[N^2 - \sum_{t=1}^{k} n_t^2\right], \end{aligned}$$

which agrees with the general expression stated in (6.15).

It also follows from (6.12) and (6.13) that

$$\begin{aligned} \operatorname{var}(J) &= \operatorname{var}\left(\sum_{u=1}^{v-1}\sum_{v=2}^{k} U_{uv}\right) \\ &= \sum_{u=1}^{v-1}\sum_{v=2}^{k}\operatorname{var}(U_{uv}) + \sum_{\substack{u=1\\(u,v)\neq(s,t)}}^{v-1}\sum_{v=2}^{k}\sum_{s=1}^{t-1}\sum_{t=2}^{k}\operatorname{cov}(U_{uv}, U_{st}). \end{aligned} \tag{6.22}$$

Under H_0 (6.2), it can be shown (we will not here) that

$$\operatorname{var}_0(U_{uv}) = \frac{n_u n_v(n_u + n_v + 1)}{12}, \quad \text{for } 1 \le u < v \le k, \tag{6.23}$$

$$\operatorname{cov}_0(U_{uv}, U_{st}) = 0, \quad \text{for all distinct } u, v, s, t \text{ in } \{1, \ldots, k\}, \tag{6.24}$$

$$\text{cov}_0(U_{uv}, U_{ut}) = \frac{n_u n_v n_t}{12}, \quad \text{for } 1 \le u < v, t \le k,\ v \ne t, \tag{6.25}$$

$$\text{cov}_0(U_{uv}, U_{su}) = \frac{-n_s n_u n_v}{12}, \quad \text{for } 1 \le s < u < v \le k, \tag{6.26}$$

$$\text{cov}_0(U_{uv}, U_{vt}) = \frac{-n_u n_v n_t}{12}, \quad \text{for } 1 \le u < v < t \le k, \tag{6.27}$$

$$\text{cov}_0(U_{uv}, U_{sv}) = \frac{n_u n_v n_s}{12}, \quad \text{for } 1 \le u, s < v \le k,\ u \ne s. \tag{6.28}$$

Combining the results in (6.23)–(6.27), and (6.28) with the expression for $\text{var}(J)$ in (6.22), it follows after significant algebraic manipulation that

$$\text{var}_0(J) = \frac{N^2(2N+3) - \sum_{j=1}^{k} n_j^2(2n_j+3)}{72},$$

which agrees with the general expression stated in (6.16).

The null asymptotic normality of the standardized form

$$J^* = \frac{J - E_0(J)}{\{\text{var}_0(J)\}^{1/2}} = \frac{J - \left[\frac{N^2 - \sum_{j=1}^{k} n_j^2(2n_j+3)}{4}\right]}{\left\{\left[N^2(2N+1) - \sum_{t=1}^{k} n_t^2(2n_t+3)\right]\Big/72\right\}^{1/2}}$$

follows from the fact that J can be expressed as a sum of certain mutually independent combined-samples Mann–Whitney statistics and standard theory for such sums of mutually independent, but not necessarily identically distributed, random variables (see, e.g., Terpstra (1952) or Section 12.1 of Randles and Wolfe (1979)). Asymptotic normality results for J under general alternatives to H_0 are obtainable from standard results in the k-sample U-statistics theory (see, e.g., Lehmann (1975, pp. 401–402)).

20. *Power of the Jonckheere–Terpstra Test.* The Jonckheere–Terpstra procedures (6.14) and (6.18) are quite superior to the Kruskal–Wallis procedures in (6.6) and (6.7) when the conjectured ordering of the treatment effects ($\tau_1 \le \tau_2 \le \cdots \le \tau_k$) is, indeed, appropriate. In addition, small violations in the conjectured ordering for τ_i and τ_j do not seriously affect the power of the Jonckheere–Terpstra tests if i and j correspond to treatment labels near the middle of the conjectured orderings. However, if i and j are both near 1 or k, the effect of such violations can be rather substantial, especially if the magnitude of the difference $|\tau_j - \tau_i|$ is fairly large. Mack and Wolfe (1981) presented the results of a small-sample power study that illustrates this phenomenon about the power of the Jonckheere–Terpstra procedures. In Section 6.3, we will discuss test procedures designed to deal with this possibility of early or late violations of the conjectured orderings $\tau_1 \le \tau_2 \le \cdots \le \tau_k$. The Jonckheere–Terpstra procedures will turn out to be special cases of this class of tests designed for the more general form of alternatives $\tau_1 \le \tau_2 \le \cdots \le \tau_{p-1} \le \tau_p \ge \tau_{p+1} \ge \cdots \ge \tau_k$, known in the literature as *umbrella orderings* for the pictorial shape of the graphed treatment effects.

21. *k-Sample Behrens–Fisher Problem.* Two of the implicit requirements associated with Assumptions A1–A3 are that the underlying distributions belong to the same

common family (F) and that they differ within this family at most in their medians. The less restrictive setting where these assumptions are relaxed to permit the possibility of differences in scale parameters as well as medians within the common family F is referred to as the *k-sample Behrens–Fisher problem*. The Jonckheere–Terpstra procedure (6.14) is no longer distribution-free under this more relaxed Behrens–Fisher setting. Chen and Wolfe (1990a) suggested a modification of the Jonckheere–Terpstra statistic J (6.13) to deal with this less restrictive setting. Their approach is similar to that used by Rust and Fligner (1984) to modify the Kruskal–Wallis statistic H for the same setting (see Comment 10).

22. *Consistency of the J Test*. Replace Assumptions A1–A3 by the less restrictive Assumptions A1′: the X's are mutually independent and A2′ : $X_{1j}, \ldots, X_{n_j j}$ come from the same continuous population $\Pi_j, j = 1, \ldots, k$. The populations $\Pi_1, \ldots, \Pi_k$ need not be identical, but we do assume that

$$\delta_{ij} = P(X_{1j} > X_{1i}) \geq \tfrac{1}{2}, \quad \text{for } 1 \leq i < j \leq k.$$

Then, roughly speaking, the test defined by (6.14) is consistent if and only if there is at least one pair (i, j), with $i < j$, such that $\delta_{ij} > \frac{1}{2}$.

Properties

1. *Consistency*. The condition n_j/N tends to $\lambda_j, 0 < \lambda_j < 1, j = 1, \ldots, k$, is sufficient to ensure that the test defined by (6.14) is consistent against the H_2 (6.11) alternatives. For a more general consistency statement, see Terpstra (1952) and Comment 22.
2. *Asymptotic Normality*. See Randles and Wolfe (1979, pp. 396–397) and Lehmann (1975, pp. 401–402).
3. *Efficiency*. See Puri (1965) and Section 6.10.

Problems

11. Apply the Jonckheere–Terpstra test to the psychotherapeutic attraction data of Table 6.2 using the postulated ordering $\tau_1 \leq \tau_2 \leq \tau_3 \leq \tau_4$. Compare and contrast this result with that obtained for the Kruskal–Wallis test in Problem 1.

12. The statistic J can be computed either from (a) the joint ranking of the $N = \sum_{j=1}^{k} n_j$ observations or from (b) $k(k-1)/2$ "two-sample" rankings. Explain.

13. What are the minimum and maximum values for J? Justify your answer.

14. Suppose $k = 3$ and $n_1 = 4$, $n_2 = 7$, $n_3 = 8$. Compare the critical region for the exact level $\alpha = .0444$ test of H_0 (6.2) based on J with the critical region for the corresponding nominal level $\alpha = .0444$ test based on the large-sample approximation. What is the nominal probability of a type I error assigned by the large-sample approximation to the exact level $\alpha = .0444$ critical region?

15. Suppose $k = 4$, $n_1 = n_2 = n_3 = 1$, and $n_4 = 2$. Obtain the form of the exact null (H_0) distribution of J for the case of no tied observations.

16. Use (6.23)–(6.27), and (6.28) to show that the expression for $\text{var}_0(J)$ in (6.16) follows, under H_0, from the general expression for var (J) in (6.22).

Table 6.7 Average Basal Area Increment (BAI) Values for Oak Stands in Southeastern Ohio

Growing site index interval				
66–68	69–71	72–74	75–77	78–80
1.91	2.44	2.45	2.52	2.78
1.53		2.04	2.36	2.88
2.08		1.60	2.73	2.10
1.71		2.37		1.66

Source: M. Dale (1984).

17. In a project designed to study stand density (i.e., number of trees in a fixed area) and its relationship to other important features of a timber area such as tree growth, wood quality, and total wood production, Dale (1984) collected data on a quantity (related to yearly growth increment in a tree) known as *basal area increment* (*BAI*) for 16 stands of mixed species of oak trees in southeastern Ohio. The 16 stands were grouped according to the value of a second factor called *growing site index*. This index ranges in value from the low 50s to 100s for oak species, and as the value of the site index increases, the growing environment becomes more favorable for a stand of trees. The data in Table 6.7 are a subset of the data obtained by Dale and represent average BAI values for the 16 stands in his study. The BAI data are grouped into five distinct categories according to the associated growing site index values.

Use an appropriate test procedure to evaluate the conjecture that the average basal area increment for a given stand of oak trees is an increasing function of the value of the stand's growing site index.

18. Apply the Kruskal–Wallis test to the knowledge of performance data in Table 6.6. Compare and contrast this result with that obtained by the Jonckheere–Terpstra test in Example 6.2.

19. Show that the expression given in (6.19) for the null variance of J in the case of tied X observations reduces to the usual null variance of J when there are no ties, as given in (6.16).

6.3 DISTRIBUTION-FREE TESTS FOR UMBRELLA ALTERNATIVES (MACK–WOLFE)

In Section 6.2, we introduced the idea of designing test procedures to be especially effective against a restricted class of alternatives. There we considered the special class of monotonically ordered alternatives. In this section, we extend that idea to a broader class of alternatives, which includes the ordered alternatives of Section 6.2 as a special case.

Let $p \in \{1, 2, \ldots, k\}$ be a fixed treatment label. In this section, we consider procedures for testing H_0 (6.2) against the class of umbrella alternatives corresponding to

$$H_3 : [\tau_1 \le \tau_2 \le \cdots \le \tau_{p-1} \le \tau_p \ge \tau_{p+1} \ge \cdots \ge \tau_k, \text{ with at least one strict inequality}]. \tag{6.29}$$

(The label *umbrella* was given to these alternatives by Mack and Wolfe (1981) because of the pictorial configuration of the τ's.) The umbrella in (6.29) is said to have a peak at population p. (Note that the ordered alternatives of Section 6.2 are simply a special case of umbrella alternatives with peak at $p = k$.) These umbrella alternatives are one-way layout analogs to a quadratic regression setting and are appropriate, for example, in

evaluating marginal gain in performance efficiency as a function of time, crop yield as a function of fertilizer applied, reaction to increasing drug dosage levels where a downturn in effect may occur after the optimal dosage is exceeded, effect of age on responses to certain stimuli, etc. (These umbrella alternatives can be effectively used in place of ordered alternatives when one is concerned about possible violations of the monotonic ordering at either the beginning or the end of the sequence of treatment effects. See Comment 20 for further discussions along these lines.)

In Section 6.3A, we present a procedure specifically designed to test H_0 (6.2) against the umbrella alternatives H_3 (6.29), where the peak, p, of the conjectured umbrella is known *prior* to data collection. This procedure is preferred to the general alternatives Kruskal–Wallis test in Section 6.1 when the treatments can be labeled a priori in such a way that the experimenter expects any deviation from H_0 (6.2) to be in the particular direction of H_3 (6.29) with known p. In Section 6.3B, we extend the idea of umbrella alternatives to the more practical setting where it is not necessary to specify the peak, p, of the umbrella configuration prior to data collection. Here, we present a procedure designed to test H_0 (6.2) against the class of umbrella alternatives with peak (p) unspecified, namely,

$$H_4 : [\tau_1 \leq \cdots \leq \tau_{p-1} \leq \tau_p \geq \tau_{p+1} \geq \cdots \geq \tau_k, \text{ with at least one strict inequality, for some } p \in \{1, 2, \ldots, k\}]. \quad (6.30)$$

The Mack–Wolfe procedure in Section 6.3B is preferred to the *peak-known* procedure presented in Section 6.3A for the more common settings when umbrella alternatives are appropriate but where there is some uncertainty about the treatment at which the maximum effect is expected to occur if H_0 (6.2) is not true.

As, with the ordered alternatives in Section 6.2, we emphasize that the labeling of the treatments so that either of the umbrella alternatives H_3 (6.29) or H_4 (6.30) is appropriate *cannot* depend on the observed sample values. This labeling must correspond to a factor (s) associated with the *experimental design* and *not* on the *sample data*. In Section 6.2B, however, the peak of the conjectured umbrella needs not be specified prior to data collection.

6.3A A DISTRIBUTION-FREE TEST FOR UMBRELLA ALTERNATIVES, PEAK KNOWN (MACK–WOLFE)

In this subsection, we present a procedure for testing H_0 (6.2) against the peak-known (at p) umbrella alternatives given by H_3 (6.29).

Procedure

First, we must label the treatments so that they are in the prescribed ordered relationships to the known peak, p, corresponding to the umbrella configuration in H_3 (6.29). To calculate the known-peak umbrella statistic, A_p, we first compute the $p(p-1)/2$ Mann–Whitney counts U_{uv} (6.12) for every pair of treatments with labels less than or equal to the hypothesized peak (i.e., for $1 \leq u < v \leq p$). In addition, we compute the $(k - p + 1)(k - p)/2$ reverse Mann–Whitney counts U_{vu} (6.12) for every pair of treatments with labels greater than or equal to the hypothesized peak (i.e., for $p \leq u < v \leq k$). (Thus, U_{vu} is the number of sample v before sample u precedences. Note that if there

are no ties between the uth sample and vth sample observations, $p \le u < v \le k$, then $U_{vu} = n_u n_v - U_{uv}$.) The Mack–Wolfe peak-known statistic, A_p, is then the sum of the Mann–Whitney counts to the left of the peak and the reverse Mann–Whitney counts to the right of the peak (as appropriate for the umbrella alternatives H_3 (6.29)), namely,

$$A_p = \sum_{u=1}^{v-1} \sum_{v=2}^{p} U_{uv} + \sum_{u=p}^{v-1} \sum_{v=p+1}^{k} U_{vu}. \tag{6.31}$$

To test

$$H_0 : [\tau_1 = \cdots = \tau_k]$$

versus the peak-known (at $p \in \{1, \ldots, k\}$) umbrella alternative

$$H_3 : [\tau_1 \le \tau_2 \le \cdots \le \tau_{p-1} \le \tau_p \ge \tau_{p+1} \ge \cdots \ge \tau_k,$$
$$\text{with at least one strict inequality}],$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } A_p \ge a_{p,\alpha}; \quad \text{otherwise do not reject,} \tag{6.32}$$

where the constant $a_{p,\alpha}$ is chosen to make the type I error probability equal to α. The constant $a_{p,\alpha}$ is the upper α percentile for the null ($\tau_1 = \cdots = \tau_k$) distribution of A_p. Comment 25 explains how to obtain the critical value $a_{p,\alpha}$ for k treatments, known peak p, and sample sizes $n_1, \ldots, n_k$ and available levels of α.

Large-Sample Approximation

The large-sample approximation is based on the asymptotic ($\min(n_1, \ldots, n_k)$ tending to infinity) normality of A_p, suitably standardized. For this purpose, we need to know the expected value and variance of A_p when the null hypothesis is true. Under H_0, the expected value and variance of A_p are

$$E_0(A_p) = \frac{N_1^2 + N_2^2 - \sum_{i=1}^{k} n_i^2 - n_p^2}{4} \tag{6.33}$$

and

$$\text{var}_0(A_p) = \frac{1}{72}\left\{2(N_1^3 + N_2^3) + 3(N_1^2 + N_2^2) - \sum_{i=1}^{k} n_i^2(2n_i + 3)\right.$$
$$\left. - n_p^2(2n_p + 3) + 12n_p N_1 N_2 - 12n_p^2 N\right\}, \tag{6.34}$$

respectively, with $N_1 = \sum_{i=1}^{p} n_i$ and $N_2 = \sum_{i=p}^{k} n_i$. (Note that $N = N_1 + N_2 - n_p$, because the observations in the peak treatment p are counted in both N_1 and N_2.) These expressions for $E_0(A_p)$ and $\text{var}_0(A_p)$ are verified by direct calculations in Comment 27 for the special case of $k = 4$, $p = 3$, $n_1 = n_2 = n_4 = 1$, $n_3 = 2$. General derivations of both expressions are outlined in Comment 28.

The standardized version of A_p is

$$A_p^* = \frac{A_p - E_0(A_p)}{\sqrt{\text{var}_0(A_p)}}$$
$$= \left\{ A_p - \left[\frac{N_1^2 + N_2^2 - \sum_{i=1}^{k} n_i^2 - n_p^2}{4} \right] \right\}$$
$$\div \left\{ \left[2(N_1^3 + N_2^3) + 3(N_1^2 + N_2^2) - \sum_{i=1}^{k} n_i^2(2n_i + 3) \right.\right.$$
$$\left.\left. - n_p^2(2n_p + 3) + 12n_pN_1N_2 - 12n_p^2N \right] \Big/ 72 \right\}^{1/2}. \tag{6.35}$$

When H_0 is true, A_p^* has, as $\min(n_1, \ldots, n_k)$ tends to infinity, an asymptotic $N(0,1)$ distribution (see Comment 28 for indications of the proof). The normal theory approximation to procedure (6.32) is

$$\text{Reject } H_0 \text{ if } A_p^* \geq z_\alpha; \text{ otherwise do not reject.} \tag{6.36}$$

Ties

If there are ties among either the N_1 X's in treatments $1, \ldots, p$ or the N_2 X's in treatments $p, \ldots, k$, replace $\phi(a,b)$ in the calculations of the appropriate Mann–Whitney counts U_{uv} or reverse Mann–Whitney counts U_{vu} by $\phi^*(a,b) = 1, \frac{1}{2}, 0$ if $a <, =, \text{or} > b$, respectively, so that for each between-sample comparison where there is a tie, the contribution to the appropriate Mann–Whitney or reverse Mann–Whitney count will be $\frac{1}{2}$. After computing A_p with these modified counts, use procedure (6.32) with this tie-modified value of A_p. Note, however, that this test associated with tied X's is only approximately, and not exactly, of the significance level α.

When applying the large-sample approximation, an additional factor should be taken into account. Although ties in the X's do not affect the null expected value of A_p, its true null variance is smaller in the case of ties than the numerical value given by the expression in (6.34). However, the appropriate expression for the exact variance of A_p in the case of ties is not available. Therefore, in the case of tied X's and large-sample sizes, we recommend computing A_p using the modified Mann–Whitney counts and then A_p^* via (6.35). With this modified value of A_p^*, the approximation (6.36) can be applied. However, the associated approximate P-value will be larger than what we would obtain if the appropriate expression for the ties-corrected null variance of A_p was available to use in the computation of A_p^*.

EXAMPLE 6.3 ***Fasting Metabolic Rate of White-Tailed Deer.***

Seasonal energy requirements of deer are an important consideration when evaluating wildlife plans for certain habitats. Both nutritional quality of the range and the physiological demands of the deer must be studied in order to prevent starvation during critical seasons and to select optimum harvest strategies. Some aspects of the energy demand were considered by Silver et al. (1969) as they studied the fasting metabolic rate (FMR)

Table 6.8 Fasting Metabolic Rate (FMR) for White-Tailed Deer (kcal/kg/day)

Two-Month Period					
January–February	March–April	May–June	July–August	September–October	November–December
36.0	39.9	44.6	53.8	44.3	31.7
33.6	29.1	54.4	53.9	34.1	22.1
26.9	43.4	48.2	62.5	35.7	30.7
35.8		55.7	46.6	35.6	
30.1		50.0			
31.2					
35.3					

Source: H. Silver, N. F. Colovos, J. B. Holter, and H. H. Hayes (1969).

of white-tailed deer. In particular, one of the questions of interest was whether or not FMR is an increasing function of environmental temperature, for which they collected the data in Table 6.8.

For these data, we expect any deviation from H_0 (6.2) to be in the direction of increasing FMR values from the January–February period up through the warmest 2-month period, July–August, with declining FMR values from July–August through the November–December period. Thus, we are interested in testing H_0 against the peak-known umbrella alternatives (6.29) with treatment labels $1 \equiv$ January–February, $2 \equiv$ March–April, $3 \equiv$ May–June, $4 \equiv$ July–August, $5 \equiv$ September–October, $6 \equiv$ November–December, and known umbrella peak at $p = 4$, corresponding to the warmest (July–August) 2-month period. For the purpose of illustration, we take the significance level to be $\alpha = .0101$. Applying the R command `cUmbrPK`$(\alpha, \boldsymbol{n}, p)$, we find `cUmbrPK(.0101,c(7,3,5,4,4,3),4)` $= 125$; that is, $P_0(A_4 \geq 125) = .0101$, and, in the notation of (6.32) with $k = 6$, $p = 4, n_1 = 7$, $n_2 = 3$, $n_3 = 5$, $n_4 = 4$, $n_5 = 4$, and $n_6 = 3$, we have $a_{4,.0101} = 125$ and procedure (6.32) reduces to

$$\text{reject } H_0 \text{ if } A_4 \geq 125.$$

We now illustrate the computations leading to the sample value of A_4 (6.31). For this purpose, we first need to compute the $4(3)/2 = 6$ Mann–Whitney counts U_{uv}, for $1 \leq u < v \leq 4$, and the $3(2)/2 = 3$ reverse Mann–Whitney counts U_{vu}, for $4 \leq u < v \leq 6$. We obtain

$$U_{12} = 7 + 1 + 7 = 15, \qquad U_{13} = 7 + 7 + 7 + 7 + 7 = 35,$$
$$U_{14} = 7 + 7 + 7 + 7 = 28, \quad U_{23} = 3 + 3 + 3 + 3 + 3 = 15,$$
$$U_{24} = 3 + 3 + 3 + 3 = 12, \quad U_{34} = 3 + 3 + 5 + 1 = 12,$$
$$U_{54} = 4 + 4 + 4 + 4 = 16, \quad U_{64} = 3 + 3 + 3 + 3 = 12,$$
$$U_{65} = 3 + 3 + 3 + 3 = 12.$$

From (6.31), it follows that

$$A_4 = U_{12} + U_{13} + U_{14} + U_{23} + U_{24} + U_{34} + U_{65} + U_{64} + U_{54}$$
$$= 15 + 35 + 28 + 15 + 12 + 12 + 12 + 12 + 16 = 157.$$

As this value of A_4 is greater than the critical value $a_{4,.0101} = 125$, we reject H_0 at the $\alpha = .0101$ level. In fact, from the observed value $A_4 = 157$, we see, using the R command `pUmbrPK(metabolic.rate, 4)` that $P_0(A_4 \geq 157) =$ `pUmbrPK(metabolic.rate, 4)`. Thus, the lowest significance level at which we can reject H_0 in favor of H_3 with the observed value of the test statistic $A_4 = 157$ is $<< .0001$.

For the large-sample approximation, we have $n_1 = 7, n_2 = 3, n_3 = 5, n_4 = 4, n_5 = 4$, and $n_6 = 3$, so that $N_1 = (7 + 3 + 5 + 4) = 19, N_2 = 3 + 4 + 4 = 11$, and $N = (7 + 3 + 5 + 4 + 4 + 3) = 26$. Using these figures in expressions (6.33) and (6.34) for $E_0(A_4)$ and $\text{var}_0(A_4)$, respectively, we see that

$$E_0(A_4) = \frac{(19)^2 + (11)^2 - [(7)^2 + (3)^2 + (5)^2 + (4)^2 + (4)^2 + (3)^2 + (4)^2]}{4}$$
$$= 85.5$$

and

$$\begin{aligned}\text{var}_0(A_4) = \tfrac{1}{72}\{&2[(19)^3 + (11)^3] + 3[(19)^2 + (11)^2] \\ &-[(7)^2(2(7) + 3) + (3)^2(2(3) + 3) \\ &+(5)^2(2(5) + 3) + (4)^2(2(4) + 3) \\ &+(4)^2(2(4) + 3) + (3)^2(2(3) + 3)] \\ &-(4)^2(2(4) + 3) + 12(4)(19)(11) - 12(4)^2(26)\} \\ = \frac{21{,}018}{72} &= 291.92.\end{aligned}$$

Thus, from (6.35), we obtain

$$A_4^* = \frac{A_4 - E_0(A_4)}{\sqrt{\text{var}_0(A_4)}} = \frac{157 - 85.5}{\sqrt{291.92}} = 4.18.$$

Using the R command `pnorm(·)`, the smallest approximate level at which we can reject H_0 in favor of H_3 with the observed value of $A_4^* = 4.18$ (i.e., the approximate P-value) is then given by $P_0(A_4^* \geq 4.18) \approx 1 -$ `pnorm(4.18)` $= 1 - .99999 = .00001$. Both the exact test and the large-sample approximate test provide very strong evidence in support of the claim that FMR for white-tailed deer is an increasing function of environmental temperature. (We note that the Jonckheere–Terpstra procedure from Section 6.2 would not be appropriate for these FMR data even with relabeled treatments, because, e.g., it would be difficult to properly order the temperatures of the March–April and September–October periods.)

Comments

23. *Motivation for the Test.* Notice that the statistic A_p can be viewed as the simple sum of two Jonckheere–Terpstra statistics, one (J_{up}) on treatments 1 through p with the postulated ordering $\tau_1 \leq \cdots \leq \tau_p$ and the second (J_{down}) on treatments

k through p with the postulated reverse ordering $\tau_k \le \tau_{k-1} \le \cdots \le \tau_p$. Thus, the statistic $A_p = J_{\text{up}} + J_{\text{down}}$ will be large if either J_{up} or J_{down} (or both) is large. In view of Comment 14, this serves as partial motivation for the A_p test.

24. *Special Case of Three Treatments*. When there are only $k = 3$ treatments, the umbrella statistic A_p can be viewed in a special way. If $p = 3$, then $A_3 = U_{12} + U_{13} + U_{23}$ is just the usual Jonckheere–Terpstra statistic for the ordered alternatives $\tau_1 \le \tau_2 \le \tau_3$. If $p = 1$, we have $A_1 = U_{31} + U_{32} + U_{21}$, which is the Jonckheere–Terpstra statistic for the reverse ordered alternatives $\tau_3 \le \tau_2 \le \tau_1$. In either of these cases, all the properties of the Jonckheere–Terpstra test procedure (including null distribution and critical values) discussed in Section 6.2 apply directly to tests based on A_1 or A_3, add as appropriate. For the third umbrella setting with $p = 2$, we see that $A_2 = U_{12} + U_{32}$, which is the same as a *single* Mann–Whitney statistic comparing the peak sample (treatment 2) with the combined set of data from treatments 1 and 3. (Thus, A_2 is the number of sample 1 *or* sample 3 before sample 2 precedences.) As a result, if $p = 2$ and $k = 3$, the procedures in (6.32) and (6.36) for sample sizes n_1, n_2, and n_3 are equivalent to the exact and large-sample approximation forms, respectively, of the one-sided upper-tail two-sample Wilcoxon rank sum test (as discussed in Section 4.1) for sample sizes $m = n_1 + n_3$ and $n = n_2$.

25. *Derivation of the Distribution of A_p under H_0 (No Ties).* As with the Jonckheere–Terpstra statistic J (see Comment 17), it is clear that the umbrella peak-known statistic A_p can be computed from the joint ranking of all $N = \sum_{i=1}^{k} n_i$ observations. Thus, one way to obtain the null distribution of A_p is to follow the method of Comments 6 and 17, namely, to compute the value of A_p for each of the $N!/(\prod_{j=1}^{k} = n_j!)$ equally likely (under H_0) rank assignments. We illustrate how this works in the small-sample-size case of $k = 4$, $p = 3$, $n_1 = n_2 = n_4 = 1, n_3 = 2$. The $5!/[1!\ 1!\ 1!\ 2!] = 60$ possible assignments of the joint ranks 1, 2, 3, 4, and 5 to the four treatments and their associated values of A_3 (6.31) are as follows:

1.	I	II	III	IV		2.	I	II	III	IV
	1	2	4	3			2	1	4	3
			5						5	
				$A_3 = 7$						$A_3 = 6$
3.	I	II	III	IV		4.	I	II	III	IV
	1	3	4	2			3	1	4	2
			5						5	
				$A_3 = 7$						$A_3 = 6$
5.	I	II	III	IV		6.	I	II	III	IV
	2	3	4	1			3	2	4	1
			5						5	
				$A_3 = 7$						$A_3 = 6$
7.	I	II	III	IV		8.	I	II	III	IV
	1	2	3	4			2	1	3	4
			5						5	
				$A_3 = 6$						$A_3 = 5$

9.	I	II	III	IV		10.	I	II	III	IV
	1	4	3	2			4	1	3	2
			5						5	
				$A_3 = 6$						$A_3 = 5$
11.	I	II	III	IV		12.	I	II	III	IV
	2	4	3	1			4	2	3	1
			5						5	
				$A_3 = 6$						$A_3 = 5$
13.	I	II	III	IV		14.	I	II	III	IV
	1	3	2	4			3	1	2	4
			5						5	
				$A_3 = 5$						$A_3 = 4$
15.	I	II	III	IV		16.	I	II	III	IV
	1	4	2	3			4	1	2	3
			5						5	
				$A_3 = 5$						$A_3 = 4$
17.	I	II	III	IV		18.	I	II	III	IV
	3	4	2	1			4	3	2	1
			5						5	
				$A_3 = 5$						$A_3 = 4$
19.	I	II	III	IV		20.	I	II	III	IV
	2	3	1	4			3	2	1	4
			5						5	
				$A_3 = 4$						$A_3 = 3$
21.	1	II	III	IV		22.	I	II	III	IV
	2	4	1	3			4	2	1	3
			5						5	
				$A_3 = 4$						$A_3 = 3$
23.	I	II	III	IV		24.	I	II	III	IV
	3	4	1	2			4	3	1	2
			5						5	
				$A_3 = 4$						$A_3 = 3$
25.	I	II	III	IV		26.	I	II	III	IV
	1	2	3	5			2	1	3	5
			4						4	
				$A_3 = 5$						$A_3 = 4$
27.	I	II	III	IV		28.	I	II	III	IV
	1	5	3	2			5	1	3	2
			4						4	
				$A_3 = 5$						$A_3 = 4$
29.	I	II	III	IV		30.	I	II	III	IV
	2	5	3	1			5	2	3	1
			4						4	
				$A_3 = 5$						$A_3 = 4$

31.

I	II	III	IV
1	3	2	5
		4	

$A_3 = 4$

32.

I	II	III	IV
3	1	2	5
		4	

$A_3 = 3$

33.

I	II	III	IV
1	5	2	3
		4	

$A_3 = 4$

34.

I	II	III	IV
5	1	2	3
		4	

$A_3 = 3$

35.

I	II	III	IV
3	5	2	1
		4	

$A_3 = 4$

36.

I	II	III	IV
5	3	2	1
		4	

$A_3 = 3$

37.

I	II	III	IV
2	3	1	5
		4	

$A_3 = 3$

38.

I	II	III	IV
3	2	1	5
		4	

$A_3 = 2$

39.

I	II	III	IV
2	5	1	3
		4	

$A_3 = 3$

40.

I	II	III	IV
5	2	1	3
		4	

$A_3 = 2$

41.

I	II	III	IV
3	5	1	2
		4	

$A_3 = 3$

42.

I	II	III	IV
5	3	1	2
		4	

$A_3 = 2$

43.

I	II	III	IV
1	4	2	5
		3	

$A_3 = 3$

44.

I	II	III	IV
4	1	2	5
		3	

$A_3 = 2$

45.

I	II	III	IV
1	5	2	4
		3	

$A_3 = 3$

46.

I	II	III	IV
5	1	2	4
		3	

$A_3 = 2$

47.

I	II	III	IV
4	5	2	1
		3	

$A_3 = 3$

48.

I	II	III	IV
5	4	2	1
		3	

$A_3 = 2$

49.

I	II	III	IV
2	4	1	5
		3	

$A_3 = 2$

50.

I	II	III	IV
4	2	1	5
		3	

$A_3 = 1$

51.

I	II	III	IV
2	5	1	4
		3	

$A_3 = 2$

52.

I	II	III	IV
5	2	1	4
		3	

$A_3 = 1$

53.	I	II	III	IV	54.	I	II	III	IV
	4	5	1	2		5	4	1	2
			3					3	
				$A_3 = 2$					$A_3 = 1$
55.	I	II	III	IV	56.	I	II	III	IV
	3	4	1	5		4	3	1	5
			2					2	
				$A_3 = 1$					$A_3 = 0$
57.	I	II	III	IV	58.	I	II	III	IV
	3	5	1	4		5	3	1	4
			2					2	
				$A_3 = 1$					$A_3 = 0$
59.	I	II	III	IV	60.	I	II	III	IV
	4	5	1	3		5	4	1	3
			2					2	
				$A_3 = 1$					$A_3 = 0$

Thus, the null distribution for A_3 when $k = 3, n_1 = n_2 = n_4 = 1$, and $n_3 = 2$ is given by

$$P_0\{A_3 = 0\} = \tfrac{3}{60}, \quad P_0\{A_3 = 1\} = \tfrac{6}{60}, \quad P_0\{A_3 = 2\} = \tfrac{9}{60}$$
$$P_0\{A_3 = 3\} = \tfrac{12}{60}, \quad P_0\{A_3 = 4\} = \tfrac{12}{60}, \quad P_0\{A_3 = 5\} = \tfrac{9}{60}$$
$$P_0\{A_3 = 6\} = \tfrac{6}{60}, \quad P_0\{A_3 = 7\} = \tfrac{3}{60}.$$

The probability, under H_0, that A_3 is greater than or equal to 5, for example, is

$$P_0\{A_3 \geq 5\} = P_0\{A_3 = 5\} + P_0\{A_3 = 6\} + P_0\{A_3 = 7\} = \frac{9 + 6 + 3}{60} = .3.$$

Note that we have derived the null distribution of A_3 without specifying the common form (F) of the underlying distribution function for the X's under H_0 beyond the requirement that it be continuous. This is why the test procedure (6.32) based on A_p is called a *distribution-free procedure*. From the null distribution of A_p, we can determine the critical value $a_{p,\alpha}$ and control the probability α of falsely rejecting H_0 when H_0 is true, and this error probability does not depend on the specific form of the common underlying continuous X distribution.

For a given number of treatments k, peak p, and sample sizes $n_1, \ldots, n_k$, the R command `cUmbrPK`$(\alpha, \boldsymbol{n}, p)$ can be used to find the available upper-tail critical values $a_{p,\alpha}$ for possible values of A_p. For a given available significance level α, the critical value $a_{p,\alpha}$ then corresponds to $P_0(A_p \geq a_{p,\alpha}) = \alpha$ and is given by `cUmbrPK`$(\alpha, \boldsymbol{n}, p)$. Thus, for example, for $k = 5,\ p = 3,\ n_1 = n_2 = n_3 = n_4 = n_5 = 4$, we have $P_0(A_3 \geq 68) = .0475$, so that $a_{3,.0475} = 68$ for $k = 5,\ p = 3$, and $n_1 = n_2 = n_3 = n_4 = n_5 = 4$.

26. *Calculation of the Mean and Variance of A_p under the Null Hypothesis H_0.* In displays (6.33) and (6.34), we presented formulas for the mean and variance

of A_p when the null hypothesis is true. In this comment, we provide a direct calculation of $E_0(A_p)$ and $\text{var}_0(A_p)$ in the specific case of $k = 4, p = 3, n_1 = n_2 = n_4 = 1, n_3 = 2$ and no tied observations using the null distribution of A_3 obtained in Comment 25. (Later, in Comment 27, we discuss general derivations of $E_0(A_p)$ and $\text{var}_0(A_p)$.) From the null distribution provided in Comment 25, we see that

$$\begin{aligned} E_0(A_3) &= \left[0\left(\tfrac{3}{60}\right) + 1\left(\tfrac{6}{60}\right) + 2\left(\tfrac{9}{60}\right) + 3\left(\tfrac{12}{60}\right) + 4\left(\tfrac{12}{60}\right)\right. \\ &\qquad \left. +5\left(\tfrac{9}{60}\right) + 6\left(\tfrac{6}{60}\right) + 7\left(\tfrac{3}{60}\right)\right] \\ &= 3.5. \end{aligned}$$

This is in agreement with what we obtain using (6.33), namely,

$$\begin{aligned} E_0(A_3) &= \frac{\{(1+1+2)^2 + (2+1)^2 - [1^2 + 1^2 + 2^2 + 1^2 + 2^2]\}}{4} \\ &= \frac{16 + 9 - 11}{4} = 3.5. \end{aligned}$$

Again using the null distribution in Comment 25, we have

$$\begin{aligned} E_0(A_3^2) &= \left[0^2\left(\tfrac{3}{60}\right) + 1^2\left(\tfrac{6}{60}\right) + 2^2\left(\tfrac{9}{60}\right) + 3^2\left(\tfrac{12}{60}\right) + 4^2\left(\tfrac{12}{60}\right)\right. \\ &\qquad \left. +5^2\left(\tfrac{9}{60}\right) + 6^2\left(\tfrac{6}{60}\right) + 7^2\left(\tfrac{3}{60}\right)\right] \\ &= 15.5. \end{aligned}$$

Using the well-known expression for $\text{var}_0(A_3)$, it follows that

$$\text{var}_0(A_3) = E_0(A_3^2) - \{E_0(A_3)\}^2 = 15.5 - (3.5)^2 = 3.25,$$

which agrees with what we obtain using (6.34) directly, namely,

$$\begin{aligned} \text{var}_0(A_3) &= \{2(4^3 + 3^3) + 3(4^2 + 3^2) + [(3)(1)^2(2(1) + 3) + 2(2)^2(2(2) + 3)] \\ &\qquad +12(2)(4)(3) - 12(2)^2(5)\}/72 \\ &= \frac{182 + 75 - 71 + 288 - 240}{72} = 3.25. \end{aligned}$$

27. *Large-Sample Approximation.* As noted in Comment 23, the umbrella statistic A_p can be expressed as $A_p = J_{\text{up}} + J_{\text{down}}$, where J_{up} is the Jonckheere–Terpstra statistic on treatments 1 through p with the postulated ordering $\tau_1 \le \cdots \le \tau_p$ and J_{down} is the Jonckheere–Terpstra statistic on treatments k through p with the postulated ordering $\tau_k \le \tau_{k-1} \le \cdots \le \tau_p$. Thus, using the previous development

for the Jonckheere–Terpstra statistic in Comment 19, we see that

$$\begin{aligned} E_0(A_p) &= E_0(J_{\text{up}}) + E_0(J_{\text{down}}) \\ &= \frac{1}{4}\left[N_1^2 - \sum_{t=1}^{p} n_t^2\right] + \frac{1}{4}\left[N_2^2 - \sum_{t=p}^{k} n_t^2\right] \\ &= \frac{1}{4}\left[N_1^2 + N_2^2 - \sum_{t=1}^{k} n_t^2 - n_p^2\right], \end{aligned}$$

which agrees with the general expression stated in (6.33).

It also follows from the representation $A_p = J_{\text{up}} + J_{\text{down}}$ that

$$\begin{aligned} \text{var}_0(A_p) &= \text{var}_0(J_{\text{up}} + J_{\text{down}}) \\ &= \text{var}_0(J_{\text{up}}) + \text{var}_0(J_{\text{down}}) + 2\text{cov}_0(J_{\text{up}}, J_{\text{down}}). \end{aligned} \tag{6.37}$$

Now,

$$\begin{aligned} \text{cov}_0(J_{\text{up}}, J_{\text{down}}) &= \text{cov}_0\left(\sum_{u=1}^{v-1}\sum_{v=2}^{p} U_{uv}, \sum_{s=p}^{t-1}\sum_{t=p+1}^{k} U_{ts}\right) \\ &= \text{cov}_0\left(\sum_{u=1}^{v-1}\sum_{v=2}^{p-1} U_{uv} + \sum_{u=1}^{p-1} U_{up}, \sum_{s=p+1}^{t-1}\sum_{t=p+2}^{k} U_{ts} + \sum_{t=p+1}^{k} U_{tp}\right) \\ &= \left[\text{cov}_0\left(\sum_{u=1}^{v-1}\sum_{v=2}^{p-1} U_{uv}, \sum_{s=p+1}^{t-1}\sum_{t=p+2}^{k} U_{ts}\right)\right. \\ &\quad + \text{cov}_0\left(\sum_{u=1}^{v-1}\sum_{v=2}^{p-1} U_{uv}, \sum_{t=p+1}^{k} U_{tp}\right) \\ &\quad + \text{cov}_0\left(\sum_{u=1}^{p-1} U_{up}, \sum_{s=p+1}^{t-1}\sum_{t=p+2}^{k} U_{ts}\right) \\ &\quad \left. + \text{cov}_0\left(\sum_{u=1}^{p-1} U_{up}, \sum_{t=p+1}^{k} U_{tp}\right)\right]. \end{aligned} \tag{6.38}$$

The term $\sum_{u=1}^{v-1}\sum_{v=2}^{p-1} U_{uv}$ involves only X observations from the first $(p-1)$ samples, whereas the terms $\sum_{s=p+1}^{t-1}\sum_{t=p+2}^{k} U_{ts}$ and $\sum_{t=p+1}^{k} U_{tp}$ involve only X observations from samples $p+1, p+2, \ldots, k$ and $p, p+1, \ldots, k$, respectively. As the X observations are mutually independent, it follows that

$$\text{cov}_0\left(\sum_{u=1}^{v-1}\sum_{v=2}^{p-1} U_{uv}, \sum_{s=p+1}^{t-1}\sum_{t=p+2}^{k} U_{ts}\right) = \text{cov}_0\left(\sum_{u=1}^{v-1}\sum_{v=2}^{p-1} U_{uv}, \sum_{t=p+1}^{k} U_{tp}\right) = 0. \tag{6.39}$$

Similarly, the term $\sum_{u=1}^{p-1} U_{up}$ involves only X observations from the first p samples, and the term $\sum_{s=p+1}^{t-1} \sum_{t=p+2}^{k} U_{ts}$ involves only X observations from samples $p+1, p+2, \ldots, k$, leading to

$$\operatorname{cov}_0 \left(\sum_{u=1}^{p-1} U_{up} \sum_{s=p+1}^{t-1} \sum_{t=p+2}^{k} U_{ts} \right) = 0. \tag{6.40}$$

(Note that (6.39) and (6.40) are a consequence of the fact that the sample observations from the peak treatment p are the only data used in both J_{up} and J_{down}.) Combining (6.38), (6.39), and (6.40) with a well-known result about covariances of sums, we obtain

$$\begin{aligned} \operatorname{cov}_0(J_{\text{up}}, J_{\text{down}}) &= \operatorname{cov}_0 \left(\sum_{u=1}^{p-1} U_{up}, \sum_{t=p+1}^{K} U_{tp} \right) \\ &= \sum_{u=1}^{p-1} \sum_{t=p+1}^{k} \operatorname{cov}_0(U_{up}, U_{tp}). \end{aligned} \tag{6.41}$$

From (6.28), it follows that

$$\begin{aligned} \operatorname{cov}_0(J_{\text{up}}, J_{\text{down}}) &= \frac{n_p}{12} \sum_{u=1}^{p-1} \sum_{t=p+1}^{k} n_u n_t = \frac{n_p}{12} \left(\sum_{u=1}^{p-1} n_u \right) \left(\sum_{t=p+1}^{k} n_t \right) \\ &= \frac{n_p (N_1 - n_p)(N_2 - n_p)}{12}. \end{aligned} \tag{6.42}$$

Combining (6.37) and (6.42), we see that

$$\operatorname{var}_0(A_p) = \operatorname{var}_0(J_{\text{up}}) + \operatorname{var}_0(J_{\text{down}}) + \frac{n_p (N_1 - n_p)(N_2 - n_p)}{6}. \tag{6.43}$$

Using the expression in (6.16) for both $\operatorname{var}_0(J_{\text{up}})$ and $\operatorname{var}_0(J_{\text{down}})$, it follows from (6.43) after some algebraic manipulation (see Problem 28) that

$$\begin{aligned} \operatorname{var}_0(A_p) = \frac{1}{72} \Bigg\{ & 2(N_1^3 + N_2^3) + 3(N_1^2 + N_2^2) - \sum_{i=1}^{k} n_i^2 (2n_i + 3) \\ & - n_p^2 (2n_p + 3) + 12 n_p N_1 N_2 - 12 n_p^2 N \Bigg\}, \end{aligned}$$

which agrees with the general expression stated in (6.34).

The null asymptotic normality of the standardized form

$$A_p^* = \frac{A_p - E_0(A_p)}{\{\operatorname{var}_0(A_p)\}_{1/2}}$$

follows from the fact that A_p can be expressed as a sum of certain mutually independent combined-samples Mann–Whitney statistics and standard theory for

such sums of mutually independent, but not necessarily identically distributed, random variables (see, e.g., Mack and Wolfe (1981)). Asymptotic normality results for A_p under general alternatives to H_0 follow directly from work by Archambault, Mack, and Wolfe (1977) on a large class of k-sample statistics.

28. *k-Sample Behrens–Fisher Problem.* Two of the implicit requirements associated with Assumptions A1–A3 are that the underlying distributions belong to the same common family (F) and that they differ within this family at most in their medians. The less restrictive setting where these assumptions are relaxed to permit the possibility of differences in scale parameters as well as medians within the common family F is referred to as the *k-sample Behrens–Fisher problem*. The Mack–Wolfe procedure (6.32) is no longer distribution-free under this more relaxed Behrens–Fisher setting. Chen and Wolfe (1990a) suggested a modification of the Mack–Wolfe statistic A_p (6.31) to deal with this less restrictive setting. Their approach is similar to that used by Rust and Fligner (1984) to modify the Kruskal–Wallis statistic H for the same setting (see Comment 10).

29. *Consistency of the A_p Test.* Replace Assumptions A1–A3 by the less restrictive Assumptions A1′: the X's are mutually independent and A2′ : $X_{1j}, \ldots, X_{n_j j}$ come from the same continuous population $\Pi_j, j = 1, \ldots, k$. The populations $\Pi_1, \ldots, \Pi_k$ need not be identical, but they are restricted to conform with the umbrella alternatives. Letting $\delta_{ij} = P(X_{1j} > X_{1i})$, for $1 \le i < j \le k$, we do assume that

$$\begin{aligned} \delta_{ij} &\ge \tfrac{1}{2}, \quad \text{for } 1 \le i < j \le p \\ \delta_{ij} &\le \tfrac{1}{2}, \quad \text{for } p \le i < j \le k, \end{aligned} \tag{6.44}$$

with no restrictions on δ_{ij} for $i < p$ and $j > p$. Under these conditions on $\Pi_1, \ldots, \Pi_k$, the test defined by (6.32) is, roughly speaking, consistent if and only if at least one of the inequalities in (6.44) is strict.

Properties

1. *Consistency.* The condition n_j/N tends to $\lambda_j, 0 < \lambda_j < 1, j = 1, \ldots, k$, is sufficient to insure that the test defined by (6.32) is consistent against the umbrella alternatives H_3 (6.29). For a more general consistency statement, see Mack and Wolfe (1981) and Comment 29.
2. *Asymptotic Normality.* See Mack and Wolfe (1981).
3. *Efficiency.* See Mack and Wolfe (1981) and Section 6.10.

Problems

20. Survival of stocked tiger muskellunge (*Esox masquinongy*), like other stocked sportfish, is variable to poor in Ohio reservoirs. Previous research with this species suggests three possible reasons for poor survival: (i) predation by largemouth bass, (ii) inability to forage, and (iii) stress-related mortality associated with the stocking process. Among other things, Mather (1984) studied the effect on mortality of three components of the stocking process: netting, confinement, and temperature increase. One portion of her study dealt with the glucose response to the stress of an increase in temperature. A sample of 40 tiger muskellunge were transferred

Table 6.9 Plasma Glucose (mg%)

Hours after 12 °C temperature increase				
0	1	4	24	96
61.08	95.45	205.96	67.74	61.76
86.21	169.19	82.55	79.84	69.12
90.15	216.16	116.60	78.23	77.45
72.91	141.92	107.23	90.23	73.45
83.74	116.16	103.83	64.92	71.08
76.35	172.22	96.60	65.73	52.45
91.63	126.26	112.77	49.60	71.57
56.65	177.78	140.85	77.42	54.90

Source: M. Mather (1984).

from a 15 °C holding tank into a test tank (also held at 15 °C) and allowed 24 h to recover. (This is the period of time that previous experimenters have found to be necessary for the fish's plasma glucose level to return to normal after a dipnet stressor.) Then, a random sample of eight fish were removed from the tank, anesthetized, blood collected, and plasma glucose determined. These data serve as a baseline or control sample. Next, the stressor (a 12 °C temperature increase) was applied to the test tank and blood samples were collected (in the way previously described) for random samples of eight additional fish at each of the time periods 1, 4, 24, and 96 h after the temperature increase. These plasma glucose measurements (mg%) are given in Table 6.9 for the 40 fish in the study.

In anticipation that a 24-h period is also necessary for a tiger muskellunge's plasma glucose level to recover from the 12 °C temperature increase stressor, test the hypothesis of interest using a significance level of .048. What is the P-value for these data?

21. (a) The statistic A_p can be computed from the joint ranking of all N observations. Explain.
(b) The statistic A_p can also be computed from pairwise *two-sample* rankings. Explain.
(c) How many different two-sample rankings are required in (b) to compute A_p?

22. In Example 6.2, we used the Jonckheere–Terpstra procedure to analyze the knowledge of performance data. It is quite reasonable to postulate that "too much information" (e.g., supervisor looking over your shoulder commenting at each step of the process) might actually lead to a downturn in the number of satisfactory pieces produced. Suppose that the following data were collected under such a too much information scenario.

Group D (too much information)
38
41
37
46
39
42

Use both the Jonckheere–Terpstra procedure and the Mack–Wolfe procedure with $p = 3$ to analyze the performance data with these added group D data. Discuss your findings.

23. What are the minimum and maximum values for A_p? Justify your answers.

24. Notice that the statistic A_p (6.31) does not include any Mann–Whitney comparisons between samples from pairs of treatments on opposite sides of the peak treatment p. Discuss the pros and cons of this fact in relation to the Mack–Wolfe test procedure based on A_p.

25. Consider the umbrella statistic A_p for k treatments.

(a) Which value(s) of p requires computation of the maximum number of Mann–Whitney statistics? How many Mann–Whitney statistics are required?

(b) Which value(s) of p requires computation of the fewest number of Mann–Whitney statistics? How many Mann–Whitney statistics are required?

26. Suppose $k = 4$, $n_1 = n_3 = n_4 = 1$, and $n_2 = 2$. Obtain the form of the exact null (H_0) distribution of A_2 for the case of no tied observations. Compare the null distribution of A_2 for $k = 4$, $n_1 = n_3 = n_4 = 1$, $n_2 = 2$ with the null distribution of A_3 for $k = 4$, $n_1 = n_2 = n_4 = 1$, $n_3 = 2$, as obtained in Comment 25. Discuss the differences.

27. Suppose $k = 4$, $n_1 = n_4 = 1$, and $n_2 = n_3 = 2$. Obtain the form of the exact null (H_0) distribution of A_2 for the case of no tied observations.

28. Show that the expression for the null variance (no ties) of A_p given in (6.43) is indeed the same as that stated in (6.34).

29. In many settings, a dose–response relationship needs not be monotonic in the dosage. In *in vitro* mutagenicity assays, for example, experimental organisms may not survive the toxic side effects of high doses of the test agent, thereby actually reducing the number of organisms at risk of mutation and leading to a downturn (i.e., umbrella pattern) in the dose–response curve. The date in Table 6.10 are a subset of the data considered by Simpson and Margolin (1986) in a discussion of the analysis of Ames test results. Plates containing Salmonella bacteria of strain TA98 were exposed to various doses of Acid Red 114. The tabled observations are the numbers of visible revertant colonies on the 18 plates in the study.

Test the null hypothesis H_0 (6.2) against the alternative that the peak of the dose–response curve for Salmonella bacteria of strain TA98 exposure to Acid Red 114 occurs at dosage level 1000 μg/ml.

30. For the Salmonella bacteria strain TA98 data in Table 6.10, test the null hypothesis H_0 (6.2) against the alternative that the peak of the dose–response curve for Salmonella bacteria of strain TA98 exposure to Acid Red 114 occurs at dosage level 333 μg/ml. Compare the result with that from Problem 29.

31. For the Salmonella bacteria strain TA98 data in Table 6.10, test the null hypothesis H_0 (6.2) against the alternative that the number of revertant colonies of the bacteria is a monotone increasing function of the dose level of Acid Red 114 over the range of exposure in Table 6.10. Compare this result with those obtained in Problems 29 and 30.

32. For the Salmonella bacteria strain TA98 data in Table 6.10, use the Kruskal–Wallis procedure to test H_0 (6.2) against the general alternatives H_1 (6.3). Compare this result with those obtained in Problems 29, 30, and 31.

Table 6.10 Number of Revertant Colonies of Salmonella Bacteria of Strain TA98 under Exposure to Various Doses of Acid Red 114, with Hamster Liver Activation

Dose, μg/ml					
0	100	333	1000	3333	10,000
22	60	98	60	22	23
23	59	78	82	44	21
35	54	50	59	33	25

Source: D. G. Simpson and B. H. Margolin (1986).

6.3B A DISTRIBUTION-FREE TEST FOR UMBRELLA ALTERNATIVES, PEAK UNKNOWN (MACK–WOLFE)

In this section, we present a procedure for testing H_0 (6.2) against the general peak-unknown umbrella alternatives H_4 (6.30).

Procedure

We label the treatments so that they are in the proper umbrella relationship to the unknown peak treatment p. To calculate the Mack–Wolfe statistic for this unknown peak setting, we first use the sample data to estimate which of the treatments is most likely to correspond to the peak of the umbrella; that is, we first estimate p from the sample data. To accomplish this, we calculate k combined-samples Mann–Whitney statistics

$$U_{.q} = \sum_{i \neq q} U_{iq}, \quad \text{for } q = 1, \ldots, k, \tag{6.45}$$

where U_{iq} = (number of ith sample observations that precede qth sample observations) is the usual Mann–Whitney statistic for the ith and qth samples. Thus, $U_{.q}$ is itself simply a single Mann–Whitney statistic computed between the qth sample and the remaining $(k-1)$ samples combined (i.e., it equals the number of times an observation from the qth sample exceeds an observation from the other $(k-1)$ combined samples). Next, we standardize each of the $U_{.q}$'s by subtracting off its expected value under the null hypothesis H_0 (6.2) and dividing by its null standard deviation (see Comment 35) to obtain

$$U^*_{.q} = \frac{U_{.q} - E_0(U_{.q})}{\{\text{var}_0(U_{.q})\}^{1/2}} = \frac{U_{.q} - [n_q(N-n_q)/2]}{\left\{\dfrac{n_q(N-n_q)(N+1)}{12}\right\}^{1/2}}, \quad q = 1, \ldots, k. \tag{6.46}$$

Let r equal the number of treatments that are tied for having the maximum $U^*_{.q}$ value and let B be the subset of $\{1, 2, \ldots, k\}$ that corresponds to the r treatments tied for the maximum $U^*_{.q}$ value. (As $U^*_{.1}, \ldots, U^*_{.k}$ are discrete random variables, there are sample size configurations for which the probability is positive that r will be greater than 1. See also Comment 31 and Problem 35). The Mack–Wolfe peak-unknown statistic is then given by

$$A^*_{\hat{p}} = \frac{1}{r} \sum_{j \in B} \left[\frac{A_j - E_0(A_j)}{\{\text{var}_0(A_j)\}^{1/2}} \right], \tag{6.47}$$

where A_j (6.31) is the peak-known statistic with the peak at the jth treatment group and $E_0(A_j)$ and $\text{var}_0(A_j)$ are the null expected value and null variance of A_j given by (6.33) and (6.34), respectively. (Thus, $A^*_{\hat{p}}$ is equal to the average of the r standardized peak-known statistics corresponding to peaks at each of the r samples tied for the maximum $U^*_{.q}$. In most cases, $r = 1$ and $A^*_{\hat{p}}$ is equal to the single standardized peak-known statistic with the peak at the indicated treatment group.)

To test

$$H_0 : [\tau_1 = \cdots = \tau_k]$$

versus the peak-unknown umbrella alternatives

$$H_4 : [\tau_1 \leq \cdots \leq \tau_{p-1} \leq \tau_p \geq \tau_{p+1} \geq \cdots \geq \tau_k,$$
$$\text{with at least one strict inequality, for some } p \in \{1, 2, \ldots, k\}]$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } A_{\hat{p}}^* \geq a_{\hat{p},\alpha}^*; \quad \text{otherwise do not reject,} \tag{6.48}$$

where the constant $a_{\hat{p},\alpha}^*$ is chosen to make the type I error probability equal to α. The constant $a_{\hat{p},\alpha}^*$ is the upper α percentile for the null $(\tau_1 = \cdots = \tau_k)$ distribution of $A_{\hat{p}}^*$. Comment 36 explains how to obtain the critical value $a_{\hat{p},\alpha}^*$ for k treatments, and sample sizes $n_1, \ldots, n_k$ and available levels of α.

Ties

If there are ties among the N X's, replace $\phi(a, b)$ in the calculation of the associated Mann–Whitney counts U_{uv} or reverse Mann–Whitney counts U_{vu} by $\phi^*(a, b) = 1, \frac{1}{2}, 0$ if $a <, =,$ or $> b$, respectively, so that for each between-sample comparison where there is a tie, the contribution to the appropriate Mann–Whitney or reverse Mann–Whitney count will be $\frac{1}{2}$. After computing the $U_{.q}$'s (6.46) and $A_{\hat{p}}^*$ (6.47) with these modified counts, use procedure (6.48) with this tie-modified value of $A_{\hat{p}}^*$. Note, however, that this test associated with tied X's is only approximately, and not exactly, of the significance level α.

EXAMPLE 6.4 ***Learning Comprehension and Age.***

It is generally believed that the ability to comprehend ideas and learn is an increasing function of age up to a certain point, and then it declines with increasing age. The data in Table 6.11 are values in the range typically obtained on the Wechsler Adult Intelligence Scale (WAIS) by males of various ages. (Actually the averages of the five samples agree with the corresponding age group means in Norman (1966).)

With $k = 5$ and $n_1 = \cdots = n_5 = 3$, we wish to test

$$H_0(6.2) \text{ versus } H_4 : [\tau_1 \leq \cdots \leq \tau_p \geq \tau_{p+1} \geq \cdots \geq \tau_5,$$
$$\text{with at least one strict inequality, for some } p \in \{1, 2, \ldots, 5\}],$$

Table 6.11 The Wechsler Adult Intelligence Scale (WAIS) Values

Age group				
16–19	20–34	35–54	55–69	≥ 70
8.62	9.85	9.98	9.12	4.80
9.94	10.43	10.69	9.89	9.18
10.06	11.31	11.40	10.57	9.27

Source: R. D. Norman (1966).

where the five age groups are numbered as treatments in order of increasing age. For the purpose of illustration, we consider the significance level $\alpha = .0495$. Applying the R command `cUmbrPU(`α`,`$\boldsymbol{n}$`)`, we find `cUmbrPU(.0495, c(3, 3, 3, 3, 3)) =` 2.226; that is, $P_0(A_{\hat{p}}^* \geq 2.226) = .0495$, and, in the notation of (6.48) with $k = 5$ and $n_1 = n_2 = n_3 = n_4 = n_5 = 3$, we have $a_{\hat{p},.0495}^* = 2.226$ and procedure (6.48) reduces to

$$\text{reject } H_0 \text{ if } A_{\hat{p}}^* \geq 2.226.$$

We now illustrate the computations leading to the sample value of $A_{\hat{p}}^*$ (6.47). First, we compute all of the $5(4)/2 = 10$ possible Mann–Whitney statistics, obtaining

$$\begin{aligned}
&U_{12} = 1+3+3 = 7,\ U_{13} = 2+3+3 = 8,\ U_{14} = 1+1+3 = 5,\\
&U_{15} = 0+1+1 = 2,\ U_{23} = 1+2+3 = 6,\ U_{24} = 0+1+2 = 3,\\
&U_{25} = 0+0+0 = 0,\ U_{34} = 0+0+1 = 1,\\
&U_{35} = 0+0+0 = 0,\ U_{45} = 0+1+1 = 2.
\end{aligned}$$

In order to estimate the age group at which WAIS values peak, we next need to compute the combined-samples Mann–Whitney statistics $U_{.q}$ (6.45), for $q = 1, \ldots, 5$. Using the fact that $U_{vu} = n_u n_v - U_{uv}$ (because there are no ties in the data), for $u, v = 1, \ldots, 5$, we find that

$$\begin{aligned}
U_{.1} &= U_{21} + U_{31} + U_{41} + U_{51}\\
&= \{[3(3) - U_{12}] + [3(3) - U_{13}] + [3(3) - U_{14}] + [3(3) - U_{15}]\}\\
&= (9-7) + (9-8) + (9-5) + (9-2) = 14.\\
U_{.2} &= U_{12} + U_{32} + U_{42} + U_{52}\\
&= U_{12} + [3(3) - U_{23}] + [3(3) - U_{24}] + [3(3) - U_{25}]\\
&= 7 + (9-6) + (9-3) + (9-0) = 25,\\
U_{.3} &= U_{13} + U_{23} + U_{43} + U_{53}\\
&= U_{13} + U_{23} + [3(3) - U_{34}] + [3(3) - U_{35}]\\
&= 8 + 6 + (9-1) + (9-0) = 31,\\
U_{.4} &= U_{14} + U_{24} + U_{34} + U_{54}\\
&= U_{14} + U_{24} + U_{34} + [3(3) - U_{45}]\\
&= 5 + 3 + 1 + (9-2) = 16,
\end{aligned}$$

and

$$U_{.5} = U_{15} + U_{25} + U_{35} + U_{45} = 2 + 0 + 0 + 2 = 4.$$

For this study, we have equal sample sizes $n_1 = \cdots = n_5 = 3$. This implies that each of the combined-samples Mann–Whitney statistics has the same null mean and null variance; that is, for $q = 1, \ldots, 5$, we have

$$E_0(U_{.q}) = \frac{3(15-3)}{2} = 18, \quad \text{var}_0(U_{.q}) = \frac{3(15-3)(15+1)}{12} = 48.$$

As a result, for this equal-sample-sizes setting, we do not need to compute the standardized forms $U_{.q}^*$ (6.46), as the treatment group with the largest $U_{.q}$ value will also be the one with the largest $U_{.q}^*$ value (see also Comment 31). Therefore, the third age group (35–54) is estimated to be the unique peak group (i.e., $\hat{p} = 3$ and $r = 1$), because

$$U_{.3} = \max\{U_{.1}, U_{.2}, U_{.3}, U_{.4}, U_{.5}\} = 31.$$

The Mack–Wolfe peak-unknown statistic $A_{\hat{p}}^*$ (6.47) with $r = 1$ and $\hat{p} = 3$ becomes

$$A_{\hat{p}}^* = \frac{A_3 - E_0(A_3)}{\{\text{var}_0(A_3)\}^{1/2}}.$$

Using the computational formula (6.35) for the peak-known setting in Section 6.3A, we obtain

$$A_3 = 45, \quad E_0(A_3) = 27, \quad \text{var}_0(A_3) = 58.5,$$

which yields

$$A_{\hat{p}}^* = \frac{45 - 27}{\sqrt{58.5}} = 2.353.$$

As this value is greater than the critical value $a_{\hat{p},.0495}^* = 2.226$, we reject H_0 at the .0495 level and conclude that there is sufficient evidence in support of the claim that the ability to comprehend ideas and learn is an increasing function of age up through the age group 35–54, from which point it declines with further age. In fact, from the observed value $A_{\hat{p}}^* = 2.353$, we see, using the `R` command `pUmbrPU(wechsler)`, that $P_0(A_{\hat{p}}^* \geq 2.353) =$ `pUmbrPU(wechsler)` $=.034$. Thus, the smallest significance level at which we can reject H_0 in favor of H_4 with the observed value of the test statistic $A_{\hat{p}}^* = 2.353$ is .034.

Comments

30. *Motivation for the Test.* The combined-samples Mann–Whitney statistic $U_{.q}$ represents the number of times an observation from the qth sample exceeds an observation from the other $(k - 1)$ combined samples. If the sample sizes are all equal and $\tau_1 < \tau_2 < \cdots < \tau_{p-1} < \tau_p > \tau_{p+1} > \cdots > \tau_k$, then we would expect $U_{.p}$ to be the largest of the k combined-samples Mann–Whitney statistics. Such an outcome would lead to the selection of the pth treatment as the peak group and to $A_{\hat{p}}^* = [A_p - E_0(A_p)]/\{\text{var}_0(A_p)\}^{1/2}$. In view of Comment 23, this provides partial motivation for the $A_{\hat{p}}^*$ test when we have equal sample sizes (see also Comment 31.)
31. *Equal versus Unequal Sample Sizes.* The number of individual comparisons required to produce the value of $U_{.q}$ (6.45) is $n_q(N - n_q)$. If the sample sizes are not all equal, then we will have differing numbers of comparisons leading to the various $U_{.q}$ values. This leads to the undesirable situation where even under the null hypothesis (H_0) those treatments with more sample observations are more likely to be selected as the estimated peak if we use the $U_{.q}$ statistics directly. One way to address this problem is to first standardize the $U_{.q}$'s by subtracting off their null expected values and then dividing by their null standard derivations. The use of these standardized $U_{.q}^*$ statistics to select the peak results

in each treatment having as nearly as possible an equal chance of being selected as the peak under H_0.

If the sample sizes are all equal, say $n_1 = \cdots = n_k = n^*$, then we have

$$E_0(U_{.q}) = \frac{n^*(N - n^*)}{2} \quad \text{and} \quad \text{var}_0(U_{.q}) = \frac{n^*(N - n^*)(N + 1)}{12}$$

for every $q = 1, \ldots, k$. Thus, in order to obtain the standardized $U^*_{.q}$ in such a setting, we would be subtracting the same quantity from each $U_{.q}$ and dividing each of the resulting differences by the same value. The rank order of the resulting $U^*_{.q}$'s would be identical with the rank order of the original $U_{.q}$'s; that is, if we have equal sample sizes and t is such that

$$U_{.t} = \text{maximum}\{U_{.1}, \ldots, U_{.k}\}$$

then it is also true that

$$U^*_{.t} = \text{maximum}\{U^*_{.1}, \ldots, U^*_{.k}\}.$$

As a result, the standardization to obtain the $U^*_{.q}$'s is not necessary in the case of all equal sample sizes as the $U_{.q}$'s themselves can be directly used to select the peak $\hat{p}$.

32. *More General Setting*. As with the other procedures in this chapter, we could replace Assumptions A1–A3 and H_0 (6.2) for the Mack–Wolfe umbrella procedures (both peak-known and peak-unknown) with the more general null hypothesis that all $N!/(\prod_{j=1}^{k} n_j!)$ assignments of n_1 joint ranks to the treatment 1 observations, n_2 joint ranks to the treatment 2 observations, $\ldots, n_k$ joint ranks to the treatment k observations are equally likely.
33. *Assumptions*. As with the other procedures in this chapter, it is important to point out that for the Mack–Wolfe umbrella procedures (both the peak-known and peak-unknown) the k treatment distributions $F_1, \ldots, F_k$ can differ at most in their locations (medians) (see also Comment 4).
34. *Estimation of the Umbrella Peak*. In situations where there is a unique, single treatment label, say t, for which

$$U^*_{.t} = \text{maximum}\{U^*_{.1}, \ldots, U^*_{.k}\},$$

then $r = 1$ in (6.47) and

$$A^*_{\hat{p}} = \frac{A_t - E_0(A_t)}{\{\text{var}_0(A_t)\}^{1/2}}.$$

In this setting, t also provides us with a point estimator for the unknown peak p (i.e., $\hat{p} = t$).

Pan (1996) developed a distribution-free confidence procedure designed to identify those treatments that yield the optimal effects in a one-way layout with umbrella configuration. It utilizes the theory of U-statistics and isotonic regression to provide a random confidence subset of the treatments that contains all the unknown peaks (optimal treatments) within an umbrella ordering with prespecified confidence level.

35. *Null Mean and Variance of Combined-Samples Mann–Whitney Statistics*. The combined-samples statistic $U_{\cdot q}$ (6.45) can be viewed as a single Mann–Whitney statistic between the q-th sample with n_q observations and the remaining $k-1$ samples combined with $N-n_q$ observations. Thus, from the standard formulas for the null mean and null variance of a Mann–Whitney statistic (see the derivation in Comment 19, e.g., particularly (6.23)), we see that

$$E_0(U_{\cdot q}) = \frac{n_q(N-n_q)}{2} \quad \text{and} \quad \text{var}_0(U_{\cdot q}) = \frac{n_q(N-n_q)(N+1)}{12},$$

which agree with the expressions used in (6.46).

36. *Derivation of the Distribution of $A_{\hat{p}}^*$ under H_0 (No Ties)*. As with the peak-known statistic A_p, the peak-unknown statistic $A_{\hat{p}}^*$ can also be computed from the joint ranking of all $N = \sum_{i=1}^{k} n_i$ observations. Thus, one way to obtain the null distribution of $A_{\hat{p}}^*$ is to follow the method of Comments 6, 17, and 25, namely, to compute the value of $A_{\hat{p}}^*$ for each of the $N!/\left(\prod_{j=1}^{k} n_j!\right)$ equally likely (under H_0) rank assignments. We illustrate the development in the specific case of $k=3$, $n_1=1$, $n_2=2$, and $n_3=1$. The $4!/[1!\,2!\,1!] = 12$ possible assignments of the joint ranks 1, 2, 3, and 4 to the three treatments and the associated values of $A_{\hat{p}}^*$ are as follows:

1.	I	II	III	2.	I	II	III
	1	2	4		4	2	1
		3				3	
			$A_{\hat{p}}^* = 1.806$				$A_{\hat{p}}^* = 1.806$
3.	I	II	III	4.	I	II	III
	I	2	3		3	2	1
		4				4	
			$A_{\hat{p}}^* = 0.775$				$A_{\hat{p}}^* = 0.775$
5.	I	II	III	6.	I	II	III
	3	1	4		4	1	3
		2				2	
			$A_{\hat{p}}^* = 0.361$				$A_{\hat{p}}^* = 0.361$
7.	I	II	III	8.	I	II	III
	2	1	4		4	1	2
		3				3	
			$A_{\hat{p}}^* = 1.084$				$A_{\hat{p}}^* = 1.084$
9.	I	II	III	10.	I	II	III
	2	1	3		3	1	2
		4				4	
			$A_{\hat{p}}^* = 0.361$				$A_{\hat{p}}^* = 0.361$
11.	I	II	III	12.	I	II	III
	1	3	2		2	3	1
		4				4	
			$A_{\hat{p}}^* = 1.549$				$A_{\hat{p}}^* = 1.549$

Thus, the null distribution for $A^*_{\hat{p}}$ when $k = 3$, $n_1 = n_3 = 2$, and $n_2 = 1$ is given by

$$P_0\{A^*_{\hat{p}} = 0.361\} = \tfrac{4}{12}, \quad P_0\{A^*_{\hat{p}} = 0.775\} = \tfrac{2}{12}, \quad P_0\{A^*_{\hat{p}} = 1.084\} = \tfrac{2}{12}$$
$$P_0\{A^*_{\hat{p}} = 1.549\} = \tfrac{2}{12}, \quad P_0\{A^*_{\hat{p}} = 1.806\} = \tfrac{2}{12}.$$

The probability, under H_0, that $A^*_{\hat{p}}$ is greater than or equal to 1.549, for example, is therefore

$$P_0\{A^*_{\hat{p}} \geq 1.549\} = P_0\{A^*_{\hat{p}} = 1.549\} + P_0\{A^*_{\hat{p}} = 1.806\} = \tfrac{2+2}{12} = \tfrac{1}{3}.$$

Note that we have derived the null distribution of $A^*_{\hat{p}}$ without specifying the common form (F) of the underlying distribution function for the X's under H_0 beyond the requirement that it be continuous. This is why the test procedure (6.48) based on $A^*_{\hat{p}}$ is called a *distribution-free procedure*. From the null distribution of $A^*_{\hat{p}}$, we can determine the critical value $a^*_{\hat{p},\alpha}$ and control the probability α of falsely rejecting H_0 when H_0 is true, and this error probability does not depend on the specific form of the common underlying continuous X distribution.

For a given number of treatments k and sample sizes $n_1, \ldots, n_k$, the R command `cUmbrPU`$(\alpha, \boldsymbol{n})$ can be used to find the available upper-tail critical values $a^*_{\hat{p},\alpha}$ for possible values of $A^*_{\hat{p}}$. For a given available significance level α, the critical value $a^*_{\hat{p},\alpha}$ then corresponds to $P_0(A^*_{\hat{p}} \geq a^*_{\hat{p},\alpha}) = \alpha$ and is given by `cUmbrPU`$(\alpha, \boldsymbol{n})$. Thus, for example, for $k = 5$, $n_1 = 3$, $n_2 = 2$, $n_3 = 4$, $n_4 = 3$, and $n_5 = 3$, we have $P_0(A^*_{\hat{p}} \geq 2.216) = .0483$, so that $a^*_{\hat{p},.0483} = 2.216$ for $k = 5$, $n_1 = 3$, $n_2 = 2$, $n_3 = 4$, $n_4 = 3$, and $n_5 = 3$.

37. *Powers of the Mack–Wolfe Umbrella Tests.* The Mack–Wolfe unknown-peak umbrella procedure (6.48) based on $A^*_{\hat{p}}$ is generally much superior to the Kruskal–Wallis procedures in (6.6) and (6.7) when the treatment effects do, indeed, follow an umbrella pattern. When the peak is known a priori to be at treatment p, then the peak-known test (6.32) based on A_p has even better power properties. However, if there is serious uncertainty concerning the location of the true peak, the $A^*_{\hat{p}}$ procedure is preferable because the power of the A_p test can be somewhat diminished when p is not the correct peak. Mack and Wolfe (1981) presented the results of a small-sample power study comparing the relative performances of the Kruskal–Wallis, the Jonckheere–Terpstra, and the two Mack–Wolfe procedures for settings where umbrella alternatives pertain.

38. *Inverted Umbrella Alternatives.* The Mack–Wolfe procedures in this section can easily be adapted to provide tests for "inverted umbrella" alternatives of the form $\tau_1 \geq \cdots \geq \tau_{p-1} \geq \tau_p \leq \tau_{p+1} \leq \cdots \leq \tau_k$, with at least one strict inequality, for both p-known and p-unknown situations. To test for such inverted umbrella alternatives, simply redefine the peak-known statistics to be

$$A_p = \sum_{u=1}^{v-1} \sum_{v=2}^{p} U_{vu} + \sum_{u=p}^{v-1} \sum_{v=p+1}^{k} U_{uv}, \quad \text{for } p = 1, \ldots, k,$$

and for the peak-unknown case, redefine the peak selectors to be "valley" selectors of the form

$$U_{q\cdot} = \sum_{i \neq q} U_{qi}, \quad q = 1, 2, \ldots, k.$$

Everything else remains unchanged, including the necessary null distribution tables.

39. *k-Sample Behrens–Fisher Problem*. Two of the implicit requirements associated with Assumptions A1–A3 are that the underlying distributions belong to the same common family (F) and that they differ within this family at most in their medians. The less restrictive setting where these assumptions are relaxed to permit the possibility of differences in scale parameters as well as medians within the common F is referred to as the *k-sample Behrens–Fisher problem*. The Mack–Wolfe peak-unknown procedure (6.48) is no longer distribution-free under this more relaxed Behrens–Fisher setting. Chen and Wolfe (1990a) proposed a modification of the Mack–Wolfe statistic A_p^* (6.47) to deal with this less restrictive setting. Their approach is similar to that used by Rust and Fligner (1984) to modify the Kruskal–Wallis statistic H for the same setting (see Comment 10).

40. *Ordered versus Umbrella Alternatives*. In this section and Section 6.2, we have considered procedures for testing the null hypothesis H_0 (6.2) of no differences in treatment effects against either ordered or, more generally, umbrella alternatives. In some settings, however, what is actually of interest is the ability to distinguish *directly between* a strictly upward trend (ordered alternatives) and an early upward trend with an eventual downturn (umbrella alternatives). This is frequently the case with dose–response data. Simpson and Margolin (1986) proposed a recursive procedure based on the Jonckheere–Terpstra statistic for dealing with such problems.

41. *An Alternative Approach Based on Maximums*. The Mack–Wolfe approach to the setting of umbrella alternatives with unknown peak is to first use the data to estimate the unknown peak and then to base the test of H_0 (6.2) on the peak-known statistic with peak at this estimated value. An alternative approach would be to bypass the first step of estimating the unknown peak and simply assess directly which of the treatments provides the most evidence of an umbrella alternative. To this effect, Chen and Wolfe (1990b) studied competitor test procedures to procedure (6.48) based on the extreme statistic $A_{\max} = \max\{A_1^*, \ldots, A_k^*\}$, with A_p^* given by (6.35). Hettmansperger and Norton (1987) considered similar competitors to (6.48) based on the maximum of certain linear rank statistics. The results of a substantial small-sample power study of these competitors (as well as the Simpson and Margolin (1986) procedure mentioned in Comment 40) are provided in Chen and Wolfe (1990b).

Problems

33. Consider the tiger muskellunge data in Table 6.9. Test the hypothesis of no differences in the plasma glucose values over time against a general umbrella alternative using an approximate significance level of .01. Compare your result with that obtained in Problem 20.

34. Consider the fasting metabolic rate (FMR) data on white-tailed deer in Table 6.8. Test the hypothesis of no difference in FMR over the 2-month periods against a general umbrella alternative. Use an approximate significance level of .01. Compare your result with that obtained in Example 6.3.

35. (a) The statistic $A^*_{\hat{p}}$ can be computed from the joint ranking of all N observations. Explain.
(b) The statistic $A^*_{\hat{p}}$ can also be computed from pairwise two-sample rankings. Explain.

36. Suppose $k = 3, n_1 = n_2 = 1$, and $n_3 = 2$. Obtain the form of the exact null (H_0) distribution of $A^*_{\hat{p}}$ for the case of no tied observations. Compare this null distribution with that of $A^*_{\hat{p}}$ for $k = 3,\ n_1 = n_3 = 1$, and $n_2 = 2$, as obtained in Comment 36.

37. Construct a set of data with no tied observations for which $r > 1$ in the definition of $A^*_{\hat{p}}$ (6.47). Discuss the implications this has for estimation of the umbrella peak.

38. Consider the Acid Red 114 revertant colonies data in Table 6.10. Test the hypothesis of no differences in the number of revertant colonies over the dosage levels against a general umbrella alternative. Use a significance level of .05. Compare this result with those obtained in Problems 29, 30, 31, and 32.

6.4 A DISTRIBUTION-FREE TEST FOR TREATMENTS VERSUS A CONTROL (FLIGNER–WOLFE)

In this section, we discuss a test procedure specifically designed for the setting where one of the treatments corresponds to a control or baseline set of conditions and we are interested in assessing which, if any, of the treatments is better than the control. Without loss of generality, we label the treatments so that the control corresponds to treatment 1. In this setting, the null hypothesis of interest is still H_0 (6.2), but now it corresponds to the statement that none of the treatments $2, \ldots, k$ is different from the control (treatment 1). This is usually expressed as

$$H_0 : [\tau_i = \tau_1,\ i = 2, \ldots, k]. \tag{6.49}$$

(Note that the expression in (6.49) is, indeed, equivalent to the original expression for H_0 (6.2).)

Procedure

To compute the Fligner–Wolfe statistic FW, we first combine all N observations from the k samples and order them from least to greatest. Letting r_{ij} denote the rank of X_{ij} in this joint ranking, the Fligner–Wolfe statistic FW is then the sum of these joint ranks for the noncontrol treatments, namely,

$$\text{FW} = \sum_{j=2}^{k} \sum_{i=1}^{n_j} r_{ij}. \tag{6.50}$$

a. *One-Sided Upper-Tail Test.* To test

$$H_0 : [\tau_i = \tau_1,\ \text{for } i = 2, \ldots, k]$$

versus

$$H_5 : [\tau_i \geq \tau_1,\ \text{for } i = 2, \ldots, k,\ \text{with at least one strict inequality}], \tag{6.51}$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if FW} \geq f_\alpha; \text{ otherwise do not reject,} \tag{6.52}$$

where the constant f_α is chosen to make the type I error probability equal to α. In order to determine the critical value f_α, we note that the statistic FW can be viewed as a two-sample Wilcoxon rank sum statistic (see Section 4.1) computed for the n_1 control treatment observations (playing the role of the X's in the two-sample setting) and the $N^* = \sum_{j=2}^{k} n_j$ combined observations from treatments $2, \ldots, k$ (playing the role of the Y's in the two-sample setting). As a result, the null distribution of FW is the same as that of the Wilcoxon rank sum statistic W with sample sizes $m = n_1$ and $n = N^*$. Thus, the critical value f_α is just the upper αth percentile w_α for the null distribution of the Wilcoxon rank sum statistic with sample sizes $m = n_1$ and $n = N^*$. Values of $f_\alpha = w_\alpha$ in this case can be obtained using the `R` command `pwilcom`, as indicated in Comment 4.3.

b. *One-Sided Lower-Tail Test.* To test

$$H_0 : [\tau_i = \tau_1, \text{ for } i = 2, \ldots, k]$$

versus

$$H_6 : [\tau_i \leq \tau_1, \text{ for } i = 2, \ldots, k, \text{ with at least one strict inequality}], \tag{6.53}$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if FW} \leq N^*(N+1) - f_\alpha; \quad \text{otherwise do not reject.} \tag{6.54}$$

Large-Sample Approximation

As previously noted, when H_0 is true, the statistic FW has the same probability distribution as the null distribution of the two-sample Wilcoxon rank sum statistic W with sample sizes $m = n_1$ and $n = N^*$. Hence, it follows directly from the Large-Sample Approximation discussion of Section 4.1 that the standardized version of FW, namely,

$$\text{FW}^* = \frac{\text{FW} - E_0(\text{FW})}{\{\text{var}_0(\text{FW})\}^{1/2}} = \frac{\text{FW} - \{N^*(N+1)/2\}}{\{n_1 N^*(N+1)/12\}^{1/2}} \tag{6.55}$$

has, as $\min(n_1, N^*)$ tends to infinity, an asymptotic $N(0,1)$ distribution when H_0 is true. The normal theory approximation for procedure (6.52) is

$$\text{Reject } H_0 \text{ if FW}^* \geq z_\alpha; \quad \text{otherwise do not reject,} \tag{6.56}$$

and the normal theory approximation for procedure (6.54) is

$$\text{Reject } H_0 \text{ if FW}^* \leq -z_\alpha; \quad \text{otherwise do not reject.} \tag{6.57}$$

Ties

If there are ties among the X's, assign each of the observations in a tied group the average of the integer ranks that are associated with the tied group. After computing FW with these average ranks, use procedure (6.52) or (6.54) with this tie-averaged value of FW. Note, however, that this test associated with tied X's is only approximately, and not exactly, of the significance level α. (To get an exact level α test even in this tied setting, see Comment 45.)

When applying the large-sample approximation, an additional factor must be taken into account. Although ties in the X's do not affect the null expected value of FW, its null variance is reduced to

$$\text{var}_0(\text{FW}) = \frac{n_1 N^*}{12}\left[N + 1 - \frac{\sum_{j=1}^{g} t_j(t_j-1)(t_j+1)}{N(N-1)}\right], \tag{6.58}$$

where g denotes the number of tied groups and t_j is the size of the tied group j. We note that an united observation is considered to be a tied group of size 1. In particular, if there are no ties among the X's, then $g = N$ and $t_j = 1$ for $j = 1, \ldots, N$. In this case, each term in (6.58) of the form $t_j(t_j - 1)(t_j + 1)$ reduces to zero and the variance expression in (6.58) reduces to the usual null variance of FW when there are no ties, as given previously in (6.55). Note that the term $[n_1 N^*/12N(N-1)]\sum_{j=1}^{g} t_j(t_j - 1)(t_j + 1)$ represents the reduction in the null variance of FW due to the presence of the tied X's.

As a consequence of the effect that ties have on the null variance of FW, the following modification is needed to apply the large-sample approximation when there are tied X's. Compute FW using average ranks and set

$$\text{FW}^* = \frac{\text{FW} - \left\{\frac{N^*(N+1)}{2}\right\}}{\{\text{var}_0(\text{FW})\}^{1/2}}, \tag{6.59}$$

where $\text{var}_0(\text{FW})$ is now given by display (6.58). With this modified value of FW*, approximation (6.56) or (6.57) can be applied.

EXAMPLE 6.5 ***Motivational Effect of Knowledge of Performance—Example 6.2 Continued.***

For Hundal's (1969) study to assess the motivational effects of knowledge of performance, the no information category clearly serves as a control population, and it is very natural to ask if additional performance information of either type (rough or accurate) leads to improved performance as measured by an increase in the number of pieces processed. Thus, we will apply the Fligner–Wolfe procedure (6.51) to the data in Table 6.6 to assess whether there is a deviation from H_0 in the direction of $\tau_2 > \tau_1$ and/or $\tau_3 > \tau_1$, where the treatment numbers are the same as those taken in Example 6.2. For the purpose of illustration, we take the significance level to be $\alpha = .0415$. With $m = N^* = n_2 + n_3 = 6 + 6 = 12$ and $n = n_1 = 6$, we find using the `R` command `pwilcox` (see Comment 4.3) that $f_{.0415} = 133$ and procedure (6.52) reduces to

$$\text{Reject } H_0 \text{ if } \text{FW} \geq 133.$$

Using the joint ranks provided in parentheses beside the data in Table 6.6, we see that

$$\text{FW} = [2.5 + 5.5 + 17 + 13 + 5.5 + 9 + 18 + 5.5 + 15 + 10.5 + 16 + 13] = 130.5.$$

As this value of FW is smaller than the critical value 133, we do not reject H_0 at the .0415 level. (This example illustrates the added power of the Jonckheere–Terpstra procedure relative to that of the Fligner–Wolfe procedure when we are able to utilize the additional piece of information that $\tau_3 \geq \tau_2$. From Example 6.2, the P-value for the Jonckheere–Terpstra procedure applied to these Hundal data is .0231, indicating rejection of H_0 at $\alpha = .0415$.)

For the large-sample approximation, we need to compute the standardized form of FW* using (6.59) because there are ties in the data. The null expected value for FW is $E_0(\text{FW}) = 12(18 + 1)/2 = 114$. For the ties-corrected null variance of FW, we note that $g = 11$ and $t_1 = 1,\ t_2 = 2,\ t_3 = 4,\ t_4 = 1,\ t_5 = 1,\ t_6 = 2,\ t_7 = 3,\ t_8 = 1,\ t_9 = 1,\ t_{10} = 1$, and $t_{11} = 1$ for the Hundal data. Hence, using the ties correction in (6.58), we have that

$$\begin{aligned}\text{var}_0(\text{FW}) &= \frac{6(12)}{12}\left\{18 + 1 - \left[\frac{2(2)(1)(3) + 3(2)(4) + 4(3)(5)}{18(17)}\right]\right\} \\ &= 6\left(19 - \frac{16}{51}\right) = 112.12,\end{aligned}$$

from which it follows that the ties-corrected value of FW* (6.59) is

$$\text{FW}^* = \frac{130.5 - 114}{\{112.12\}^{1/2}} = 1.56.$$

Thus, using the approximate procedure (6.56) with the ties-corrected value of $\text{FW}^* = 1.56$, we see that the approximate P-value for these data is $P_0(\text{FW}^* \geq 1.56) \approx 1 - \texttt{pnorm(1.56)} = 1 - .9406 = .0594$. Thus, we have marginal evidence from the Fligner–Wolfe treatments-versus-control procedure that additional performance knowledge (either rough or accurate) leads to an increase in the number of pieces produced.

Comments

42. *More General Setting*. As with the other procedures of this chapter, we could replace Assumptions A1–A3 and H_0 (6.2) with the more general null hypothesis that all $N!/\left(\prod_{j=1}^{k} n_j!\right)$ assignments of n_1 joint ranks to the control observations, n_2 joint ranks to the treatment 2 observations, $\ldots, n_k$ joint ranks to the treatment k observations are equally likely.
43. *Motivation for the Test*. The statistic FW (6.50) is the sum of the joint ranks assigned to the noncontrol treatments. When some of the τ_i's are strictly greater than the control effect τ_1, we would expect the joint ranks for the observations from those treatments to be larger than the joint ranks for the control observations. The net result would be a larger value of FW. This suggests rejecting H_0 in favor of H_5 (6.51) for large values of FW and motivates procedures (6.52) and (6.56). A similar motivation leads to procedures (6.54) and (6.57). (See also Comment 47.)

44. *Assumptions*. As with the other test procedures of this chapter, Assumption A3 requires that the control and the $(k-1)$ treatment distributions $F_1, \ldots, F_k$ can differ at most in their locations (medians). (See also Comments 4 and 50.)

45. *Exact Conditional Distribution of FW with Ties among the Data*. To get an exact level α test in the presence of ties, we rely on the fact that the null distribution of FW conditional on the observed configuration of joint tied ranks is the same as the corresponding conditional tied ranks null distribution of the Wilcoxon rank sum statistic W with sample sizes $m = n_1$ and $n = N^*$. Therefore, the approach discussed and illustrated in Comment 4.5 can be used to get the exact conditional null distribution of FW and associated exact level α test in the case of ties among the data.

46. *Two-Sided Test*. We note that we have not discussed a test based on the FW statistic that is designed for a two-sided alternative. The "natural" two-sided alternative for this treatment versus control setting corresponds to [either $\tau_i \geq \tau_1$ for all $i = 2, \ldots, k$ or $\tau_i \leq \tau_1$ for all $i = 2, \ldots, k$, with at least one strict inequality]. We feel that it is rather unlikely that we would find ourselves in such a setting where either all the treatments are better than the control or all the treatments are worse than the control, but we have no idea which of the two cases pertains. As a result, a two-sided test based on FW is not presented in this section.

47. *Limitations*. The test procedures in (6.52) and (6.54) deal with very restricted alternatives where *all* the treatments are either at least as good as the control (i.e., $\tau_i \geq \tau_1$ for all $i = 2, \ldots, k$) or *all* the treatments are no better than the control (i.e., $\tau_i \leq \tau_1$ for all $i = 2, \ldots, k$), respectively. They are not appropriate tests when the possibility exists that some of the treatments might be better ($\tau_i > \tau_1$) and some might be worse ($\tau_i < \tau_1$) than the control. For such mixed alternatives, one would need to use the general alternatives Kruskal–Wallis procedure presented in Section 6.1.

48. *Comparisons Between Treatments*. The Fligner–Wolfe procedure (6.52) is a test designed to decide if *any* of treatments $2, \ldots, k$ are better (i.e., $\tau_i > \tau_1$) than the control. It involves no direct comparisons between the various treatments observations themselves. In order to reach conclusions about whether there are any differences among the treatment effects $\tau_2, \ldots, \tau_k$, one would need to apply the Kruskal–Wallis procedure of Section 6.1 (or, if appropriate, the Jonckheere–Terpstra ordered alternatives or Mack–Wolfe umbrella alternatives procedures of Sections 6.2 and 6.3, respectively) to the sample data from treatments $2, \ldots, k$. Under the null hypothesis H_0 (6.2), the Fligner–Wolfe statistic FW is independent of the Kruskal–Wallis statistic H (and also of the Jonckheere–Terpstra statistic J and the Mack–Wolfe statistics A_p and $A_{\hat{p}}^*$). This implies, for example, that if we conduct the Fligner–Wolfe test (6.52) at a significance level α_1 and the Kruskal–Wallis test (6.6) (or the Jonckheere–Terpstra test (6.14), Mack–Wolfe peak-known test (6.32), or Mack–Wolfe peak-unknown test (6.48)) on treatments $2, \ldots, k$ at the significance level α_2, then the probability of incorrectly rejecting H_0 when it is true with at least one of the two tests is exactly $\alpha_1 + \alpha_2 - \alpha_1\alpha_2$. A similar comment applies to procedure (6.54).

49. *Multiple Comparisons*. If test procedure (6.52) leads to rejection of H_0 (6.2), we are led to the conclusion that at least one treatment has a greater effect than the control. However, procedure (6.52) does not address the question of exactly how many treatment effects are greater than that of the control, or does

it provide us information as to which specific treatments are better than the control. For answers to such questions, we turn to treatments-versus-control multiple comparison procedures, as discussed in Sections 6.7 and 6.8. Similar comments apply to the lower-tail test procedure in (6.54).

50. *The Treatments-versus-Control Behrens–Fisher Problem.* Two of the implicit requirements imposed by Assumptions A1–A3 are that the underlying distributions belong to the same common family (F) and that they differ within this family at most in their medians. The less restrictive setting where these assumptions are relaxed to permit the possibility of differences in scale parameters as well as medians within the common family F is referred to as the *k-sample treatments*-versus-control Behrens–Fisher problem. The Fligner–Wolfe procedures (6.52) and (6.54) are no longer distribution-free under this more general Behrens–Fisher setting. If we replace Assumption A3 by the less restrictive Assumption A3*: [The treatments' distribution functions $F_2, \ldots, F_k$ are connected through the relationship

$$F_i(t) = F^*(t - \tau_j), \quad -\infty < t < \infty,$$

for $i = 2, \ldots, k$, where F^* is a distribution function for a continuous distribution that is symmetric about its median θ and, in addition, the control distribution (F_1) is continuous and symmetric about its median $\theta + \tau_1$.], then the Fligner–Policello two-sample robust rank procedure discussed in Section 4.4 can be adapted to provide distribution-free tests of H_0 (6.2) against either H_5 (6.51) or H_6 (6.53) under these more general treatments-versus-control Behrens–Fisher Assumptions A1, A2, and A3*.

51. *Treatments-versus-Control under Umbrella Configurations.* In many settings where we are interested in comparing a number of treatments with a control, we will have additional a priori information regarding the relative magnitude of the treatment effects. One such piece of information might be that the treatment effects are known to follow an umbrella pattern (see Section 6.3) $\tau_1 \leq \cdots \leq \tau_{p-1} \leq \tau_p \geq \tau_{p+1} \geq \cdots \geq \tau_k$ with either known or unknown peak p. (Remember that the ordered pattern of Section 6.2 corresponds to $p = k$ or 1.) In a drug study, for instance, increasing dosage levels may be compared with a zero-dose control. If the treatment effects are not identical to that of the control, then it is often reasonable to assume that the higher the dose of the drug applied, the better (say, higher) will be the resulting effect on a patient, corresponding to monotonically ordered treatment effects. However, it may also be the case that a subject might potentially succumb to toxic effects at high doses, thereby actually decreasing the associated treatment effects. Such a setting would correspond to an ordering in the treatment effects that is monotonically increasing up to a point, followed by a monotonic decrease; that is, an umbrella pattern on the treatment effects. Chen and Wolfe (1993) considered a test procedure designed specifically to compare a number of treatments with a single control under this basic umbrella pattern for the treatment effects. Their test requires an equal number of observations in each of the treatments (i.e., $n_2 = \cdots = n_k = n^*$), but permits a differing number (n_1) of observations from the control setting. The necessary null distribution critical values are provided for a variety of k, n_1, and n^* combinations, and the results of a substantial Monte Carlo simulation power study are presented.

(An example of the type of data for which this Chen–Wolfe procedure would be appropriate is provided by the muskellunge plasma glucose data in Table 6.9.)

52. *Consistency of the FW test*. Replace Assumptions A1–A3 by the less restrictive Assumptions A1′: The X's are mutually independent and A2′ : $X_{1j}, \ldots, X_{n_j j}$ come from the same continuous population Π_j, $j = 1, \ldots, k$. The populations $\Pi_1, \ldots, \Pi_k$ need not be identical, but we do assume that

$$\delta_{ij} = P(X_{1j} > X_{11}) \geq \tfrac{1}{2}, \quad \text{for } j = 2, \ldots, k.$$

Then, roughly speaking, the test defined by (6.52) is consistent if and only if there is at least one $j \in \{2, 3 \ldots, k\}$ for which $\delta_{1j} > \frac{1}{2}$.

Properties

1. *Consistency*. The condition n_j/N tends to λ_j, $0 < \lambda_j < 1$, $j = 1, \ldots, k$, is sufficient to ensure that the tests defined by (6.52) and (6.54) are consistent against the H_5 (6.51) and H_6 (6.53) alternatives, respectively. For a more general consistency statement, see Comment 52.
2. *Asymptotic Normality*. See Fligner and Wolfe (1982).
3. *Efficiency*. See Fligner and Wolfe (1982) and Section 6.10.

Problems

39. Apply the appropriate Fligner–Wolfe test to the psychotherapeutic attraction data of Table 6.2. Compare and contrast this result with that obtained for the Kruskal–Wallis test in Problem 1.

40. Apply the appropriate Fligner–Wolfe procedure to the glucocorticoid receptor data for the leukemia patients in Table 6.4, using the normal subjects as the control. Compare and contrast with the result obtained from the Kruskal–Wallis test in Problem 8.

41. Apply the appropriate Fligner–Wolfe test to the muskellunge plasma glucose data in Table 6.9. Compare and contrast with the result obtained from the Mack–Wolfe test in Problem 20. (See also Comment 51.)

RATIONALE FOR MULTIPLE COMPARISON PROCEDURES

In Sections 6.1–6.4 of this chapter, we have discussed procedures designed to test the null hypothesis H_0 (6.2) against a variety of alternative hypotheses. Upon rejection of H_0 with one of these test procedures for a given set of data, our conclusions range from the general statement that there are some unspecified differences among the treatment effects (associated with the Kruskal–Wallis procedure discussed in Section 6.1) to the more informative relationships between the treatment effects associated with test procedures designed for the ordered or umbrella alternatives or the treatments-versus-control setting. However, in none of these test procedures are our conclusions pair-specific; that is, the tests in Sections 6.1–6.4 are not designed to enable us to reach conclusions about specific pairs of treatment effects. The relative sizes of the specific treatment effects τ_1 and τ_2, for example, cannot be inferred from the conclusions reached by any of the test procedures in Sections 6.1–6.4. To elicit such pairwise specific information, we turn to the class of multiple comparison procedures. In Section 6.5 we present such two-sided

all-treatments multiple comparison procedures for the omnibus setting corresponding to the general alternatives H_1 (6.3). In Section 6.6 we deal with one-sided all-treatments multiple comparison procedures associated with the restricted ordered alternatives H_2 (6.11). Finally, in Section 6.7 we discuss an approach for making treatments-versus-control multiple comparison decisions.

6.5 DISTRIBUTION-FREE TWO-SIDED ALL-TREATMENTS MULTIPLE COMPARISONS BASED ON PAIRWISE RANKINGS—GENERAL CONFIGURATION (DWASS, STEEL, AND CRITCHLOW–FLIGNER)

In this section we present a multiple comparison procedure based on pairwise two-sample rankings that is designed to make decisions about individual differences between pairs of treatment effects (τ_i, τ_j), for $i < j$, in a setting where general alternatives H_1 (6.3) are of interest. Thus, the multiple comparison procedure of this section would generally be applied to one-way layout data *after* rejection of H_0 (6.2) with the Kruskal–Wallis procedure from Section 6.1. In this setting, it is important to reach conclusions about all $\binom{k}{2} = k(k-1)/2$ pairs of treatment effects, and these conclusions are naturally two-sided in nature.

Procedure

For each pair of treatments (i, j), let

$$W_{ij} = \sum_{b=1}^{n_j} R_{ib}, \quad \text{for } 1 \le i < j \le k, \tag{6.60}$$

where $R_{i1}, \ldots, R_{in_j}$ are the ranks of $X_{1j}, \ldots, X_{n_j j}$, respectively, among the combined ith and jth samples; that is, W_{ij} is the Wilcoxon rank sum of the jth sample ranks in the joint two-sample ranking of the ith and jth sample observations. Compute

$$W_{ij}^* = \sqrt{2}\left[\frac{W_{ij} - E_0(W_{ij})}{\{\text{var}_0(W_{ij})\}^{1/2}}\right] = \frac{W_{ij} - \frac{n_j(n_i+n_j+1)}{2}}{\{n_i n_j (n_i + n_j + 1)/24\}^{1/2}}, \quad \text{for } 1 \le i < j \le k. \tag{6.61}$$

(Thus, W_{ij}^* is the standardized (under H_0) version of W_{ij} multiplied by $\sqrt{2}$.)

At an experimentwise error rate of α, the Steel–Dwass–Critchlow–Fligner two-sided all-treatments multiple comparison procedure reaches its $k(k-1)/2$ pairwise decisions, corresponding to each (τ_u, τ_v) pair $1 \le u < v \le k$, by the criterion

$$\text{Decide } \tau_u \ne \tau_v \text{ if} |W_{uv}^*| \ge w_\alpha^*; \quad \text{otherwise decide } \tau_u = \tau_v, \tag{6.62}$$

where the constant w_α^* is chosen to make the experimentwise error rate equal to α; that is, w_α^* satisfies the restriction

$$P_0(|W_{uv}^*| < w_\alpha^*, u = 1, \ldots, k-1; v = u+1, \ldots, k) = 1 - \alpha, \tag{6.63}$$

where the probability $P_0(\cdot)$ is computed under H_0 (6.2). Equation (6.63) stipulates that the $k(k-1)/2$ inequalities $|W_{uv}^*| < w_\alpha^*$, corresponding to all pairs (u, v) of treatments

with $u < v$, hold simultaneously with probability $1 - \alpha$ when H_0 (6.2) is true. Comment 55 explains how to obtain the critical value w_α^* for k treatments, sample sizes $n_1, \ldots, n_k$, and available experimentwise error rates α.

Large-Sample Approximation

When H_0 is true, the $[k(k-1)/2]$-component vector $(W_{12}^*, W_{13}^*, \ldots, W_{k-1,k}^*)$ has, as $\min(n_1, \ldots, n_k)$ tends to infinity, an asymptotic multivariate normal distribution with mean vector $\mathbf{0}$. It then follows (see Comment 58 for indications of the proof) that w_α^* can be approximated for large-sample sizes by q_α, where q_α is the upper αth percentile point for the distribution of the range of k independent $N(0,1)$ variables. Thus, the large-sample approximation for procedure (6.62) is

$$\text{Decide } \tau_u \neq \tau_v \text{ if } |W_{uv}^*| \geq q_\alpha; \quad \text{otherwise decide } \tau_u = \tau_v. \tag{6.64}$$

To find q_α for k treatments, we use the R command `cRangekNorm`(α, k). For example, to find $q_{.05}$ for $k = 6$ treatments, we apply `cRangekNorm(.05, 6)` and obtain $q_{.05} = 4.031$.

Ties

If there are ties among the X observations, use average ranks in computing the individual Wilcoxon rank sum statistics W_{ij} (6.60). In addition, replace the term $\text{var}_0(W_{ij})/2 = n_i n_j (n_i + n_j + 1)/24$ in the denominator of W_{ij}^* (6.61) by

$$\frac{\text{var}_0(W_{ij})}{2} = \frac{n_i n_j}{24}\left[n_i + n_j + 1 - \frac{\sum_{b=1}^{g_{ij}} (t_b - 1)t_b(t_b + 1)}{(n_i + n_j)(n_i + n_j - 1)}\right], \tag{6.65}$$

where, for $1 \leq i < j \leq k$, g_{ij} denotes the number of tied groups in the joint ranking of the ith and jth sample observations and t_b is the size of tied group b in this joint ranking. Furthermore, an untied observation is considered to be a tied group of size 1. In particular, if there are no tied observations in the ith and jth combined samples, then $g_{ij} = n_i + n_j$ and $t_b = 1$ for $b = 1, \ldots, n_i + n_j$, in which case each term of the form $(t_b - 1)t_b(t_b + 1)$ reduces to 0 and $\text{var}_0(W_{ij})/2$ reduces to $n_i n_j (n_i + n_j + 1)/24$, the appropriate expression when there are no ties in the ith and jth combined samples.

EXAMPLE 6.6 ***Length of YOY Gizzard Shad.***

Consider the length of YOY gizzard shad data discussed in Problem 4. Applying the Kruskal–Wallis procedure to the length data from the four sites in Kokosing Lake yields highly significant differences between the median YOY lengths at the four sites. To examine which particular sites differ in median YOY lengths, we apply the approximate procedure (6.64) with the appropriate corrections for ties given in (6.65). For this study, we have $k = 4$, $n_1 = n_2 = n_3 = n_4 = 10$, and we must compute $k(k-1)/2 = 4(3)/2 = 6$ standardized W_{ij}^* statistics. For the sake of illustration, we take our experimentwise error

rate to be $\alpha = .01$. With $k = 4$, we find $q_{.01} =$ `cRangekNorm(.01, 4)` = 4.404 and procedure (6.64) reduces to

$$\text{Decide } \tau_u \neq \tau_v \text{ if } |W_{uv}^*| \geq 4.404.$$

Next, we compute the six W_{ij}^* statistics. For the sample observations from sites I and II (populations 1 and 2, respectively), the combined-samples ranking yields the sum of ranks for the site II data to be

$$W_{12} = 9.5 + 20 + 3.5 + 9.5 + 13.5 + 19 + 1 + 17 + 9.5 + 18 = 120.5.$$

For this pair of samples, there are tied observations, and we have $g_{12} = 14$ and $t_1 = t_2 = 1,\ t_3 = 2,\ t_4 = t_5 = t_6 = 1,\ t_7 = 4,\ t_8 = 1,\ t_9 = t_{10} = 2,$ and $t_{11} = t_{12} = t_{13} = t_{14} = 1$. From (6.65), we find

$$\begin{aligned}\frac{\text{var}_0(W_{12})}{2} &= \frac{10(10)}{24}\left[10 + 10 + 1 - \frac{3(1)(2)(3) + (3)(4)(5)}{(10+10)(10+10-1)}\right] \\ &= \frac{25}{6}\left[\frac{7,980 - 78}{380}\right] = 86.64.\end{aligned}$$

Using this result in (6.61), we obtain

$$W_{12}^* = \frac{[120.5 - 10(21)/2]}{\sqrt{86.64}} = 1.67.$$

For the other five population pairs, similar calculations yield the following:
Site I and Site III

$$\begin{aligned}W_{13} &= 13.5 + 11 + 2.5 + 1 + 8 + 8 + 2.5 + 5 + 5 + 5 = 61.5, \\ g_{13} &= 13, \quad t_1 = 1, \quad t_2 = 2, \quad t_3 = t_4 = 3, \quad t_5 = t_6 = t_7 = 1, \quad t_8 = 2, \\ t_9 &= t_{10} = t_{11} = t_{12} = 1, \quad t_{13} = 2, \\ \text{var}_0(W_{13}) &= \frac{10(10)}{24}\left[10 + 10 + 1 - \frac{3(1)(2)(3) + 2(2)(3)(4)}{(10+10)(10+10-1)}\right] = 86.78, \\ W_{13}^* &= \frac{[61.5 - 10(21)/2]}{\sqrt{86.78}} = -4.67.\end{aligned}$$

Site I and Site IV

$$\begin{aligned}W_{14} &= 11 + 9 + 4.5 + 7 + 9 + 2.5 + 2.5 + 1 + 4.5 + 9 = 60, \\ g_{14} &= 15, \quad t_1 = 1, \quad t_2 = t_3 = 2, \quad t_4 = t_5 = 1, \quad t_6 = 3, \\ t_7 &= t_8 = t_9 = t_{10} = t_{11} = t_{12} = t_{13} = t_{14} = 1, \quad t_{15} = 2, \\ \text{var}_0(W_{14}) &= \frac{10(10)}{24}\left[10 + 10 + 1 - \frac{3(1)(2)(3) + 2(3)(4)}{(10+10)(10+10-1)}\right] = 87.04, \\ W_{14}^* &= \frac{[60 - 10(21)/2]}{\sqrt{87.04}} = -4.82.\end{aligned}$$

Site II and Site III

$$W_{23} = 12 + 11 + 2.5 + 1 + 8.5 + 8.5 + 2.5 + 5.5 + 5.5 + 5.5 = 62.5,$$
$$g_{23} = 13, \quad t_1 = 1, \quad t_2 = 2, \quad t_3 = 4, \quad t_4 = 2, \quad t_5 = t_6 = t_7 = 1, \quad t_8 = 3,$$
$$t_9 = t_{10} = t_{11} = t_{12} = t_{13} = 1,$$
$$\text{var}_0(W_{23}) = \frac{10(10)}{24}\left[10 + 10 + 1 - \frac{2(1)(2)(3) + 2(3)(4) + 3(4)(5)}{(10+10)(10+10-1)}\right] = 86.45,$$
$$W_{23}^* = \frac{[62.5 - 10(21)/2]}{\sqrt{86.45}} = -4.57.$$

Site II and Site IV

$$W_{24} = 11 + 9 + 5 + 7 + 9 + 2.5 + 2.5 + 1 + 5 + 9 = 61,$$
$$g_{24} = 13, \quad t_1 = 1, \quad t_2 = 2, \quad t_3 = 3, \quad t_4 = 1, \quad t_5 = 3, \quad t_6 = t_7 = 1, \quad t_8 = 3,$$
$$t_9 = t_{10} = t_{11} = t_{12} = t_{13} = 1,$$
$$\text{var}_0(W_{24}) = \frac{10(10)}{24}\left[10 + 10 + 1 - \frac{1(2)(3) + 3(2)(3)(4)}{(10+10)(10+10-1)}\right] = 86.64,$$
$$W_{24}^* = \frac{[61 - 10(21)/2]}{\sqrt{86.64}} = -4.73.$$

Site III and Site IV

$$W_{24} = 18 + 16 + 9 + 14 + 16 + 3 + 3 + 1 + 9 + 16 = 105,$$
$$g_{34} = 10, \quad t_1 = 1, \quad t_2 = 3, \quad t_3 = 2, \quad t_4 = 5, \quad t_5 = 2, \quad t_6 = 1,$$
$$t_7 = 3, \quad t_8 = t_9 = t_{10} = 1,$$
$$\text{var}_0(W_{34}) = \frac{10(10)}{24}\left[10 + 10 + 1 - \frac{2(1)(2)(3) + 2(2)(3)(4) + 4(5)(6)}{(10+10)(10+10-1)}\right] = 85.53,$$
$$W_{34}^* = \frac{[105 - 10(21)/2]}{\sqrt{85.53}} = 0.$$

Taking absolute values and referring them to the critical value $q_{.01} = 4.403$, we see that

$$|W_{12}^*| = 1.67 < 4.404 \implies \text{decide } \tau_1 = \tau_2,$$
$$|W_{13}^*| = 4.67 > 4.404 \implies \text{decide } \tau_1 \neq \tau_3,$$
$$|W_{14}^*| = 4.82 > 4.404 \implies \text{decide } \tau_1 \neq \tau_4,$$
$$|W_{23}^*| = 4.57 > 4.404 \implies \text{decide } \tau_2 \neq \tau_3,$$
$$|W_{24}^*| = 4.73 > 4.404 \implies \text{decide } \tau_2 \neq \tau_4,$$
$$|W_{34}^*| = 0 < 4.404 \implies \text{decide } \tau_3 = \tau_4.$$

Thus, at an experimentwise error rate of .01, the six multiple comparison decisions can be summarized by the statement $(\tau_1 = \tau_2) \neq (\tau_3 = \tau_4)$. This multiple comparison procedure provides more detailed information about the lengths of the YOY gizzard shad population in Kokosing Lake. We now know that sites I and II may be viewed as providing similar living environments for gizzard shad. The same conclusion holds for sites III and IV. However, we also know that the common living environment at sites I and II is significantly different from the common living environment at sites III and IV.

Comments

53. *Rationale for Multiple Comparison Procedures*. We think of the methods of this section as multiple comparison procedures. The aim of applying such procedures goes beyond the point of deciding whether the treatments are equivalent to the (often more important) problem of selecting which, if any, treatments differ from one another. Thus, the user makes $k(k-1)/2$ decisions, one for each pair of treatments. Equation (6.63) states that the probability of making all correct decisions when H_0 is true is controlled to be $1-\alpha$. That is, when using procedure (6.62), the probability of at least one incorrect decision, when H_0 is true, is controlled to be α. This error rate is derived under the assumption that H_0 is true, but it does not depend on the particular underlying distribution F. This is why we call (6.62) a distribution-free multiple comparison procedure.

    Multiple comparison procedures can be interpreted as hypothesis tests. If we consider the test that rejects H_0 if the inequality of (6.62) holds for at least one (u, v) pair and accepts H_0 if, for every (u, v) pair, the inequality of (6.62) is not satisfied, this is a distribution-free test of size α for H_0 (6.2).

54. *Experimentwise Error Rate*. The use of an experimentwise error rate represents a very conservative approach to multiple comparisons. We are insisting that the probability of making only correct decisions be $1-\alpha$ when the hypothesis H_0 (6.2)) of treatment equivalence is true. Thus, we have a high degree of protection when H_0 is true, but we often apply such techniques when we have evidence (perhaps based on a priori information or perhaps obtained by applying the Kruskal–Wallis test, as in Example 6.6) that H_0 is not true.

    This protection under H_0 also makes it harder for the procedure to judge treatments as differing significantly when in fact H_0 is false, and this difficulty becomes more severe as k increases. We justify our use of an experimentwise error rate in much the same way as Kurtz et al. (1965). The rate provides a precise measure of a level of uncertainty, and statements at higher or lower levels are readily obtained.

    Anscombe (1965), although not advocating the use of such rates, mentioned an interesting hypothetical situation (which he attributed to Richard Olshen) in defense of such a conservative approach. Anscombe was commenting on simultaneous confidence intervals proposed by Kurtz et al., but his statements would also apply to multiple comparison procedures of the type discussed here. We quote from his comments. "A panacea manufacturer advertises on television that trials have shown his product to be more effective than any other leading brand. Such trials (if they are not a downright fabrication) certainly seem to present

a situation of the third type.* Their objective is not to help the manufacturer reach a decision, but hopefully to permit him to make a multiple comparison statement that will impress the public and boost sales. He could appropriately use the simultaneous confidence intervals of this paper; indeed, the Food and Drug Administration could appropriately require him to do so. The more equally ineffective other leading brands there were, the harder would it be for him to obtain the evidence he needed, and the more trials he would have to conduct and suppress before achieving a favorable one. Thus would Statistics and Economics go hand in hand to protect the public."

55. *Critical Values* w_α^*. The w_α^* critical values can be obtained by using the fact that under H_0 (6.2), all $N!/(\prod_{j=1}^k n_j!)$ *joint* (of all N sample observations) rank assignments of n_1 ranks to the treatment 1 observations, n_2 ranks to the treatment 2 observations, $\ldots, n_k$ ranks to the treatment k observations are equally likely. (Although the standardized pairwise Wilcoxon statistics W_{ij}^* (6.61) are formally defined in terms of pairwise two-sample ranks, it is clear that all $k(k-1)/2$ W_{ij}^* values can also be computed from the joint ranks of all N observations.) Thus, to obtain the probability, under H_0, that $|W_{uv}^*| < c$, for all $u < v$, we simply count the number of configurations for which the event $A = \{|W_{uv}^*| < c, \text{ for all } u < v\}$ occurs, and divide this count by $N!/\left[\prod_{j=1}^k n_j!\right]$. For an illustration, we return to Comment 6 and use the 15 joint rank configurations displayed there for the case $k = 3$ and $n_1 = n_2 = n_3 = 2$. (Again, we can reduce the number of configurations that need to be considered from 90 to 15 by the same reasoning as in Comment 6.) For each of these 15 configurations, we now display the values of $|W_{12}^*|$, $|W_{13}^*|$, and $|W_{23}^*|$.

(a)	$\|W_{12}^*\| = 2.1909$ $\|W_{13}^*\| = 2.1909$ $\|W_{23}^*\| = 2.1909$	(b)	$\|W_{12}^*\| = 2.1909$ $\|W_{13}^*\| = 2.1909$ $\|W_{23}^*\| = 1.0954$	(c)	$\|W_{12}^*\| = 2.1909$ $\|W_{13}^*\| = 2.1909$ $\|W_{23}^*\| = 0$
(d)	$\|W_{12}^*\| = 1.0954$ $\|W_{13}^*\| = 2.1909$ $\|W_{23}^*\| = 2.1909$	(e)	$\|W_{12}^*\| = 1.0954$ $\|W_{13}^*\| = 2.1909$ $\|W_{23}^*\| = 1.0954$	(f)	$\|W_{12}^*\| = 1.0954$ $\|W_{13}^*\| = 2.1909$ $\|W_{23}^*\| = 0$
(g)	$\|W_{12}^*\| = 1.0954$ $\|W_{13}^*\| = 1.0954$ $\|W_{23}^*\| = 1.0954$	(h)	$\|W_{12}^*\| = 0$ $\|W_{13}^*\| = 2.1909$ $\|W_{23}^*\| = 2.1909$	(i)	$\|W_{12}^*\| = 1.0954$ $\|W_{13}^*\| = 1.0954$ $\|W_{23}^*\| = 0$
(j)	$\|W_{12}^*\| = 0$ $\|W_{13}^*\| = 1.0954$ $\|W_{23}^*\| = 2.1909$	(k)	$\|W_{12}^*\| = 0$ $\|W_{13}^*\| = 1.0954$ $\|W_{23}^*\| = 1.0954$	(l)	$\|W_{12}^*\| = 1.0954$ $\|W_{13}^*\| = 0$ $\|W_{23}^*\| = 0$
(m)	$\|W_{12}^*\| = 0$ $\|W_{13}^*\| = 0$ $\|W_{23}^*\| = 2.1909$	(n)	$\|W_{12}^*\| = 0$ $\|W_{13}^*\| = 0$ $\|W_{23}^*\| = 1.0954$	(o)	$\|W_{12}^*\| = 0$ $\|W_{13}^*\| = 0$ $\|W_{23}^*\| = 0$

* The term *third type* is used by Anscombe to refer to experiments intended to give fundamental knowledge or insight into some phenomenon but not to aid in a particular job of decision making.

Thus, for example,

$$\begin{aligned} P_0\{|w^*_{uv}| < 2.1909, (u,v) = (1,2),(1,3),(2,3)\} \\ &= P_0\{|W^*_{12}| < 2.1909; |W^*_{13}| < 2.1909; |W^*_{23}| < 2.1909\} \\ &= \tfrac{6}{15} = 1 - .6, \end{aligned}$$

because for 6 of the 15 configurations—[(g), (i), (k), (1), (n), and (o)]—the event $\{|W^*_{12}| < 2.1909; |W^*_{13}| < 2.1909; |W^*_{23}| < 2.1909\}$ occurs. Similarly, $P_0\{|W^*_{uv}| < 1.0954, (u,v) = (1,2),(1,3),(2,3)\} = \frac{1}{15} = 1 - .9333$, as the event $\{|W^*_{12}| < 1.0954; |W^*_{13}| < 1.0954; |W^*_{23}| < 1.0954\}$ occurs only for the single configuration (o). Hence, for $k = 3$ and $n_1 = n_2 = n_3 = 2$, we have $w^*_{.6000} = 2.1909$ and $w^*_{.9333} = 1.0954$, and the values .6000 and .9333 are the only available experimentwise error rates for the Dwass–Steel–Critchlow–Fligner procedure (6.62) in this setting.

For a given number of treatments k and sample sizes $n_1, \ldots, n_k$, the R command `cSDCFlig`$(\alpha, \boldsymbol{n})$ can be used to find the available critical values w^*_α. For a given available experimentwise error rate α, the critical value w^*_α is given by `cSDCFlig`$(\alpha, \boldsymbol{n})$. Thus, for example, for $k = 3$ and $n_1 = 3$, $n_2 = 5$, and $n_3 = 7$, we have $w^*_{.0331} =$ `cSDCFlig(.0331,c(3,5,7))` $= 3.330$.

56. *Historical Development*. The multiple comparison procedures (6.62) and (6.64) based on the Wilcoxon rank sum statistics were first proposed independently by Steel (1960, 1961) and Dwass (1960) for the setting of equal sample sizes $n_1 = \cdots = n_k$. Critchlow and Fligner (1991) presented a natural generalization of these Steel–Dwass procedures when the n_i are not all equal and provided the exact critical values w^*_α for $k = 3$ and $2 \le n_1 \le n_2 \le n_3 \le 7$.

57. *Maximum Type I Error Rate*. The multiple comparison procedure (6.62) is designed so that the experimentwise error rate (see Comment 54) is controlled to be equal to α; that is, the probability of falsely declaring any pair of treatment effects to be different, when in fact *all* of the treatment effects are the same, is equal to α. However, it also satisfies the more stringent *maximum type I error rate* requirement that the probability of falsely declaring any pair of treatment effects to be different, regardless of the values of the other $k - 2$ treatment effects, is no larger than the stated α. This requires controlling the probability of making false declarations about treatment effect differences even in situations when *not all* of the treatment effects are the same. For example, if $\tau_1 < \tau_2 = \tau_3$, the probability of incorrectly deciding that $\tau_2 \ne \tau_3$ is still controlled to be α by multiple comparison procedure (6.62). Similar comments apply to the approximate procedure in (6.64).

58. *Large-Sample Approximation.* Let $\mathbf{W}^* = \left(W^*_{12}, W^*_{13}, \ldots, W^*_{k-1,k}\right)$, where W^*_{ij} is given by (6.61) for $1 \le i < j \le k$. Then it can be shown that $\mathbf{W}^*$ has, as $\min(n_1, \ldots, n_k)$ tends to infinity, an asymptotic multivariate normal distribution with mean vector $\mathbf{0}$ and appropriate covariance matrix $\boldsymbol{\Sigma}$ (see Miller (1981a) for further details). It follows directly from this result (again, see Miller (1981a)) that the procedure in (6.64) has an asymptotic experimentwise error rate equal to α when $n_1 = n_2 = \cdots = n_k$. Critchlow and Fligner (1991) used a result by Hayter (1984) to establish the fact that the asymptotic experimentwise error

rate for procedure (6.64) is also bounded above by α when we have unequal sample sizes.

When H_0 is true and $n_1 = n_2 = \cdots = n_k$, the asymptotic correlation matrix $\mathbf{\Sigma}_1$ (say) of the $\binom{k}{2}W_{ij}$'s is the same as the correlation matrix $\mathbf{\Sigma}_2$ (say) of the $\binom{k}{2}$ differences $Z_i - Z_j$, $1 \le i < j \le k$, where $Z_1, \ldots, Z_k$ are independent $N(0, 1)$ random variables (cf. Miller (1966), pp. 155–156). It follows that the asymptotic distribution of

$$\sqrt{2} \max_{1 \le i < j \le k} \left\{ \frac{|W_{ij} - E_0(W_{ij})|}{[\text{var}_0(W_{ij})]^{1/2}} \right\} = \max_{1 \le i < j \le k} |W_{ij}^*|$$

can be approximated by the distribution of

$$\max_{1 \le i < j \le k} |Z_i - Z_j| = \text{range } (Z_1, \ldots, Z_k).$$

The $\sqrt{2}$ occurs because the variance of $Z_i - Z_j$ equals 2. This justifies the use of approximation (6.64) in the equal-sample-size case. When the sample sizes are unequal, the asymptotic correlation matrix of the $\binom{k}{2}W_{ij}$'s will not in general agree with $\mathbf{\Sigma}_2$, but (6.64)) can be justified via a Tukey–Kramer approximation (see, e.g., Tukey (1953), Kramer (1956, 1957) and pages 91–93 of Hochberg and Tamhane (1987)).

59. *Joint Ranking Approach.* The multiple comparison procedure discussed in this section is based on $k(k-1)/2$ separate two-sample rankings. However, it is also reasonable to consider all-treatments multiple comparisons based on a single joint ranking of all N observations. Let $R_{.j}$ (6.4), $j = 1, \ldots, k$, be the average rank for the jth treatment sample in the joint ranking of all N observations. The joint ranking analog to procedure (6.62) is then given by

$$\text{Decide } \tau_u \ne \tau_v \text{ if } N^*|R_{.u} - R_{.v}| \ge y_\alpha; \quad \text{otherwise decide } \tau_u = \tau_v, \qquad (6.66)$$

where N^* is the least common multiple of $n_1, \ldots, n_k$ and the constant y_α is chosen to make the experimentwise error rate equal to α; that is, y_α satisfies the restriction

$$P_0(N^*|R_{.u} - R_{.v}| < y_\alpha, u = 1, \ldots, k-1; \; v = u+1, \ldots, k) = 1 - \alpha, \quad (6.67)$$

where the probability $P_0(\cdot)$ is computed under H_0 (6.2). As with the multiple comparison procedures based on pairwise rankings, (6.67) stipulates that the $k(k-1)/2$ inequalities $N^*|R_{.u} - R_{.v}| < y_\alpha$, corresponding to all pairs (u, v) of treatments with $u < v$, hold simultaneously with probability $1 - \alpha$ when H_0 (6.2) is true.

Nemenyi (1963) first proposed procedure (6.67) for the special case of equal sample sizes, in which case $N^*|R_{.u} - R_{.v}| = |R_u - R_v|$, where R_j (6.4) is the sum of the joint ranks for the treatment j observations. The general form of (6.67) for arbitrary sample sizes was considered by Damico and Wolfe (1987).

The y_α critical values can be obtained in exactly the same way as the w_α^* values for procedure (6.63). Proceeding as in Comment 55, we simply count the number of joint rank configurations for which the event

$B = \{N^*|R_{.u} - R_{.v}| < c, \text{ for all } u < v\}$ occurs and divide this count by $N!/\left[\Pi_{j=1}^{k} n_j!\right]$ to obtain the probability, under H_0, that $N^*|R_{.u} - R_{.v}| < c$ for all $u < v$. For an illustration, we again return to Comment 6 and use the 15 joint rank configurations displayed there for the case $k = 3$ and $n_1 = n_2 = n_3 = 2$. (For this setting, $N^*|R_{.u} - R_{.v}| = |R_u - R_v|$, and we can once again reduce the number of configurations that need to be considered from 90 to 15 by the same reasoning as in Comment 6.) For each of these 15 configurations, we now display the values of $|R_1 - R_2|$, $|R_1 - R_3|$, and $|R_2 - R_3|$.

(a)	$\|R_1 - R_2\| = 4$ $\|R_1 - R_3\| = 8$ $\|R_2 - R_3\| = 4$	(b)	$\|R_1 - R_2\| = 5$ $\|R_1 - R_3\| = 7$ $\|R_2 - R_3\| = 2$	(c)	$\|R_1 - R_2\| = 6$ $\|R_1 - R_3\| = 6$ $\|R_2 - R_3\| = 0$
(d)	$\|R_1 - R_2\| = 2$ $\|R_1 - R_3\| = 7$ $\|R_2 - R_3\| = 5$	(e)	$\|R_1 - R_2\| = 3$ $\|R_1 - R_3\| = 6$ $\|R_2 - R_3\| = 3$	(f)	$\|R_1 - R_2\| = 4$ $\|R_1 - R_3\| = 5$ $\|R_2 - R_3\| = 1$
(g)	$\|R_1 - R_2\| = 2$ $\|R_1 - R_3\| = 4$ $\|R_2 - R_3\| = 2$	(h)	$\|R_1 - R_2\| = 0$ $\|R_1 - R_3\| = 6$ $\|R_2 - R_3\| = 6$	(i)	$\|R_1 - R_2\| = 3$ $\|R_1 - R_3\| = 3$ $\|R_2 - R_3\| = 0$
(j)	$\|R_1 - R_2\| = 1$ $\|R_1 - R_3\| = 4$ $\|R_2 - R_3\| = 5$	(k)	$\|R_1 - R_2\| = 0$ $\|R_1 - R_3\| = 3$ $\|R_2 - R_3\| = 3$	(l)	$\|R_1 - R_2\| = 2$ $\|R_1 - R_3\| = 1$ $\|R_2 - R_3\| = 1$
(m)	$\|R_1 - R_2\| = 2$ $\|R_1 - R_3\| = 2$ $\|R_2 - R_3\| = 4$	(n)	$\|R_1 - R_2\| = 1$ $\|R_1 - R_3\| = 1$ $\|R_2 - R_3\| = 2$	(o)	$\|R_1 - R_2\| = 0$ $\|R_1 - R_3\| = 0$ $\|R_2 - R_3\| = 0$

Thus, for example,

$$
\begin{aligned}
&P_0\{|R_u - R_v| < 8, (u, v) = (1,2), (1,3), (2,3)\} \\
&= P_0\{|R_1 - R_2| < 8; |R_1 - R_3| < 8; |R_2 - R_3| < 8\} \\
&= \tfrac{14}{15} = 1 - .067,
\end{aligned}
$$

because for 14 of the configurations—all but configuration (a)—the event $\{|R_1 - R_2| < 8; |R_1 - R_3| < 8; |R_2 - R_3| < 8\}$ occurs. Similarly, $P_0\{|R_u - R_v| < 7; (u, v) = (1,2), (1,3), (2,3)\} = \frac{12}{15} = .80$, because the event $\{|R_1 - R_2| < 7; |R_1 - R_3| < 7; |R_2 - R_3| < 7\}$ occurs for 12 of the configurations—all but (a), (b), and (d). Hence, for $k = 3$ and $n_1 = n_2 = n_3 = 2$, we have $y_{.067} = 8$ and $y_{.200} = 7$. Values of y_α are available in Damico and Wolfe (1987) for available experimentwise error rates (α) closest to but not exceeding .001, .005, .01 (.005) .05 (.01) .15 and most useful combinations of either $k = 3, 1 \le n_1 \le n_2 \le n_3 \le 6$ or $k = 4, 1 \le n_1 \le n_2 \le n_3 \le n_4 \le 6$. For the special cases of equal sample sizes, these tabled values agree with those previously given by Nemenyi (1963) and McDonald and Thompson (1967). An approximation to y_α for large common sample size is discussed in Miller (1966). A related approximate procedure based on joint ranks and appropriate for large unequal sample size is suggested by Dunn (1964).

The joint ranking multiple comparison procedure given by (6.66) is a good deal simpler computationally than the corresponding pairwise ranking multiple comparison procedure in (6.62). Both procedures maintain the designated experimentwise error rate α. However, the joint ranking procedure does not provide the additional maximum type I error rate protection level α guarantee associated with the pairwise ranking procedure (see Comment 57). A second drawback for the joint ranking procedure is the fact that the absolute differences $|R_{\cdot u} - R_{\cdot v}|$ depend on the values of the observations from the other $k - 2$ treatments, in addition to the observations from treatments u and v. Thus, in the case of $k = 3$, the decision concerning treatments 1 and 2, for example, depends on the treatment 3 observations. This difficulty is discussed in Miller (1966) and Gabriel (1969).

Properties

1. *Asymptotic Multivariate Normality.* See Hayter (1984) and Critchlow and Fligner (1991).
2. *Efficiency.* See Sherman (1965) and Section 6.10.

Problems

42. Apply procedure (6.62) to the mean interstitial length data of Table 6.5.

43. Procedure (6.62) is defined specifically in terms of the $k(k-1)/2$ pairwise two-sample rankings. However, it can be applied to settings where only the joint ranks of all N observations are available. Explain.

44. Apply procedure (6.62) to the half-time of mucociliary clearance data of Table 6.1.

45. Apply the approximate procedure (6.64) to the glucocorticoid receptor data of Table 6.4.

46. For the case $k = 3, \alpha = .05$, and $n_1 = n_2 = n_3 = 6$, compare procedures (6.62) and (6.64).

47. Apply the approximate procedure (6.64) to the psychotherapeutic attraction data of Table 6.2.

48. Find the totality of all available experimentwise error rates α and the associated critical values w_α^* for procedure (6.62) when $k = 4$, $n_1 = 1$, and $n_2 = n_3 = n_4 = 2$.

49. Consider the joint ranking procedure (6.66) discussed in Comment 59. Find the totality of all available experimentwise error rates α and the associated critical values y_α for this procedure when $k = 4$, $n_1 = 1$ and $n_2 = n_3 = n_4 = 2$.

50. Consider the YOY gizzard shad data discussed in Example 6.6. Find the smallest (available) approximate experimentwise error rate at which the most significant difference in treatment effects (i.e., that between site I and site IV) would be detected.

51. Consider the mean interstitial length data in Table 6.5. Find the smallest (available) approximate experimentwise error rate at which we would declare that the typical mean interstitial length for white pines is different from that for Scotch pines.

6.6 DISTRIBUTION-FREE ONE-SIDED ALL-TREATMENTS MULTIPLE COMPARISONS BASED ON PAIRWISE RANKINGS-ORDERED TREATMENT EFFECTS (HAYTER–STONE)

In this section, we discuss a multiple comparison procedure based on pairwise two-sample rankings that is designed to make decisions about individual differences between pairs of

treatment effects (τ_i, τ_j), for $i < j$, in a setting where ordered alternatives H_2 (6.11) are of interest. Thus, the multiple comparison procedure of this section would be appropriate for one-way layout data *after* rejection of H_0 (6.2) with the Jonckheere–Terpstra procedure from Section 6.2. As with the procedure for general alternatives discussed in Section 6.5, we will once again reach conclusions about all $\binom{k}{2} = k(k-1)/2$ pairs of treatment effects. However, here these conclusions are naturally one-sided, in accordance with the ordered alternatives setting.

Procedure

For each pair of treatments $(i,j), 1 \le i < j \le k$, let W_{ij} be defined by expression (6.60); that is, W_{ij} is the Wilcoxon rank sum of the jth sample ranks in the two-sample ranking of the ith and jth sample observations. Compute the standardized form W_{ij}^* given in (6.61) for each treatment pair combination (i,j) with $i < j$.

At an experimentwise error rate of α, the Hayter–Stone one-sided all-treatments multiple comparison procedure reaches its $k(k-1)/2$ pairwise decisions, corresponding to each (τ_u, τ_v) pair, $1 \le u < v \le k$, by the criterion

$$\text{Decide } \tau_v > \tau_u \text{ if } W_{uv}^* \ge c_\alpha^*; \quad \text{otherwise decide } \tau_u = \tau_v, \tag{6.68}$$

where the constant c_α^* is chosen to make the experimentwise error rate equal to α; that is, c_α^* satisfies the restriction

$$P_0(W_{uv}^* < c_\alpha^*, u = 1, \ldots, k-1; v = u+1, \ldots, k) = 1 - \alpha, \tag{6.69}$$

where the probability $P_0(\cdot)$ is computed under H_0 (6.2). Equation (6.69) requires that the $k(k-1)/2$ inequalities $W_{uv}^* < c_\alpha^*$, corresponding to all pairs (u, v) of treatments with $u < v$, hold simultaneously with probability $1 - \alpha$ when H_0 (6.2) is true. Comment 62 explains how to obtain the critical value c_α^* for k treatments, sample sizes $n_1, \ldots, n_k$, and available experimentwise error rates α.

Large-Sample Approximation

When H_0 is true, the $k(k-1)/2$ component vector $(W_{12}^*, W_{13}^*, \ldots, W_{k-1,k}^*)$ has, as $\min(n_1, \ldots, n_k)$ tends to infinity, an asymptotic multivariate normal distribution with mean vector $\mathbf{0}$. It then follows (see Hayter and Stone (1991), e.g., for an indication of the proof) that c_α^* can be approximated for large sample sizes by d_α, where d_α is the upper αth percentile point for the distribution of

$$D = \underset{1 \le i < j \le k}{\text{maximum}} \left[\frac{Z_j - Z_i}{\left\{ \frac{n_i + n_j}{2 n_i n_j} \right\}^{1/2}} \right],$$

where $Z_1, \ldots, Z_k$ are mutually independent and Z_i has an $N(0, 1/n_i)$ distribution, for $i = 1, \ldots, k$. Thus, the large-sample approximation for procedure (6.68) is

$$\text{Decide } \tau_v > \tau_u \text{ if } W_{uv}^* \le d_\alpha; \quad \text{otherwise decide } \tau_u = \tau_v. \tag{6.70}$$

To find d_α for k treatments, we use the R command `cHayStonLSA`(α, k). For example, to find $d_{.05}$ for $k = 6$ treatments, we apply `cHayStonLSA(.05, 6)` and obtain $d_{.05} = 3.719$ (see also Comment 64).

Ties

If there are ties among the X observations, use average ranks in computing the individual Wilcoxon rank sum statistics W_{ij} (6.60). In addition, replace the term $\text{Var}_0(W_{ij})/2 = n_i n_j (n_i + n_j + 1)/24$ in the denominator of W_{ij}^* (6.61) by the expression in (6.65).

EXAMPLE 6.7 ***Motivational Effect of Knowledge of Performance–Example 6.2 Continued.***

For Hundal's (1969) study to assess the motivational effects of knowledge of performance, we found in Example 6.2 (using the Jonckheere–Terpstra test procedure) that there was sufficient evidence in the sample data to conclude that $\tau_1 \le \tau_2 \le \tau_3$ with at least one strict inequality. To examine which of the types of information (none, rough, or accurate) lead to differences in median numbers of pieces processed, we apply procedure (6.68) with the appropriate corrections for ties, as given in (6.65). For this study, we have $k = 3, n_1 = n_2 = n_3 = 6$, and we must compute $k(k-1)/2 = 3(2)/2 = 3$ standardized W_{ij}^* statistics. For the sake of illustration, we take our experimentwise error rate to be $\alpha = .0553$. With $k = 3$ and $n_1 = n_2 = n_3 = 6$, we find $c_{.0553}^* =$ `cHayston(.0553, c(6, 6, 6))` = 2.9439 and procedure (6.68) reduces to

$$\text{Decide } \tau_v > \tau_u \text{ if } W_{uv}^* \ge 2.9439.$$

Next, we compute the three W_{ij}^* statistics. For the control (no information) and group B (partial information) sample observations, the combined-samples ranking yields the sum of ranks for the group B data to be

$$W_{12} = 2.5 + 5 + 12 + 10.5 + 5 + 8 + 43.$$

For this pair of samples, there are tied observations and we have $g_{12} = 8$ and $t_1 = 1, t_2 = 2, t_3 = 3, t_4 = t_5 = t_6 = 1, t_7 = 2$, and $t_8 = 1$. From (6.65), we obtain

$$\frac{\text{var}_0(W_{12})}{2} = \frac{6(6)}{24}\left[6 + 6 + 1 - \frac{2(1)(2)(3) + 2(3)(4)}{(6+6)(6+6-1)}\right]$$
$$= \frac{3}{2}\left[\frac{1716 - 36}{132}\right] = 19.09.$$

Using this result in (6.61), we find

$$W_{12}^* = \frac{[43 - 6(13)/2]}{\sqrt{19.09}} = .92.$$

For the other two population pairs, similar calculations lead to the following.

Control (No Information) and Group C (Accurate Information)

$$W_{13} = 12 + 3.5 + 10 + 6.5 + 11 + 8.5 = 51.5,$$
$$g_{13} = 9, \quad t_1 = t_2 = 1, \quad t_3 = 2, \quad t_4 = 1, \quad t_5 = t_6 = 2, \quad t_7 = t_8 = t_9 = 1,$$
$$\text{var}_0(W_{13}) = \frac{6(6)}{24}\left[6 + 6 + 1 - \frac{3(1)(2)(3)}{(6+6)(6+6-1)}\right] = 19.30,$$
$$W_{13}^* = \frac{[51.5 - 6(13)/2]}{\sqrt{19.30}} = 2.85.$$

Group B (Partial Information) and Group C (Accurate Information)

$$W_{23} = 12 + 3 + 9 + 6 + 10 + 7.5 = 47.5,$$
$$g_{23} = 9, \quad t_1 = 1, \quad t_2 = 3, \quad t_3 = t_4 = 1, \quad t_5 = 2, \quad t_6 = t_7 = t_8 = t_9 = 1,$$
$$\text{var}_0(W_{23}) = \frac{6(6)}{24}\left[6 + 6 + 1 - \frac{1(2)(3) + 2(3)(4)}{(6+6)(6+6-1)}\right] = 19.16,$$
$$W_{23}^* = \frac{[47.5 - 6(13)/2]}{\sqrt{19.16}} = 1.94.$$

Referring these W_{ij}^* values to the critical point $c_{.0553}^* = 2.9439$, we see that

$$W_{12}^* = .92 < 2.9439 \Rightarrow \text{decide } \tau_1 = \tau_2,$$
$$W_{13}^* = 2.85 < 2.9439 \Rightarrow \text{decide } \tau_1 = \tau_3,$$
$$W_{23}^* = 1.94 < 2.9439 \Rightarrow \text{decide } \tau_2 = \tau_3.$$

Thus, at an experimentwise error rate of .0553, we have reached the conclusion that $\tau_1 = \tau_2 = \tau_3$ (i.e., there are no differences in median numbers of pieces processed between the different levels of information), in contradiction with the conclusion from the Jonckheere–Terpstra test that $\tau_1 \le \tau_2 \le \tau_3$ with at least one strict inequality. Even though the P-value for the Jonckheere–Terpstra test procedure for these data is .0231, we are not able to detect any individual differences between treatment effects with the multiple comparison procedure (6.68), even with an experimentwise error rate as high as .0553. Such occurrences are, unfortunately, rather common in practice because of the conservative nature of the multiple comparison procedures (see Comment 54). For this reason, we often conduct our multiple comparison procedure at an experimentwise error rate that is higher than a typical significance level (such as .01 or .05) for a hypothesis test. If we have previously conducted a hypothesis test (such as the Jonckheere–Terpstra test in the example) and rejected H_0, we would at least like to know the *most* significant difference between pairs of treatment effects. For this reason, it is always informative in such cases to find the smallest experimentwise error rate at which the first pairwise difference in treatment effects would become significant. For the Hundal data, that corresponds to treatments 1 (no information) and 3 (accurate information) with an observed value $W_{13}^* = 2.85$. Using the R command `pHaySton(motivational.effect)`, we find that the smallest experimentwise error rate (among the limited number available) at which we would decide $\tau_3 > \tau_1$ (and thus conclude that accurate information is more effective than no information) is `pHaySton(motivational.effect)$p.val [2]` = .0850.

Comments

60. *Rationale for Multiple Comparison Procedures*. The general rationale for the multiple comparison procedures of this section is the same as that given in Comment 53 for the two-sided-all-treatments multiple comparison procedures of Section 6.5. The only additional factor here is that the procedures of this section yield decisions that are one-sided by nature in line with their association with the ordered restriction ($\tau_1 \leq \cdots \leq \tau_k$) on the treatment effects.

61. *Experimentwise Error Rate*. The use of an experimentwise error rate represents a very conservative approach to multiple comparisons. We are insisting that the probability of making only correct decisions be $1 - \alpha$ when the hypothesis H_0 (6.2) of treatment equivalence is true. Thus, we have a high degree of protection when H_0 is true, but we often apply the techniques of this section when we have evidence (perhaps based on a priori information or perhaps obtained by applying the Jonckheere–Terpstra test, as in Example 6.7) that H_0 is not true. (For additional general remarks about experimentwise error rates, see Comment 54.)

62. *Critical Values* c_α^*. The c_α^* critical values can be obtained by using the fact that under H_0 (6.2), all $N!/\left(\prod_{j=1}^k n_j!\right)$ *joint* (of all N sample observations) rank assignments of n_1 ranks to the treatment 1 observations, n_2 ranks to the treatment 2 observations, $\ldots, n_k$ ranks to the treatment k observations are equally likely. (Although the standardized pairwise Wilcoxon statistics W_{ij}^* (6.61) are formally defined in terms of pairwise two-sample ranks, it is clear that all $k(k-1)/2$ W_{ij}^* statistics can also be computed from the joint ranks of all N observations.) Thus, to obtain the probability, under H_0, that $W_{uv}^* < t$, for all $u < v$, we simply count the number of configurations for which the event $A = \{W_{uv}^* < t, \text{ for all } u < u\}$ occurs and divide this count by $N!/\left[\prod_{j=1}^k n_j!\right]$. For an illustration, we return to Comment 17 and use the 12 joint rank configurations displayed there for the case $k = 3, n_1 = 1, n_2 = 1$, and $n_3 = 2$. For each of these 12 configurations, we now display the values of W_{12}^*, W_{13}^*, and W_{23}^*.

(a)	$W_{12}^* = 1.4142$ $W_{13}^* = 1.7321$ $W_{23}^* = 1.7321$	(b)	$W_{12}^* = -1.4142$ $W_{13}^* = 1.7321$ $W_{23}^* = 1.7321$	(c)	$W_{12}^* = 1.4142$ $W_{13}^* = 1.7321$ $W_{23}^* = 0$
(d)	$W_{12}^* = -1.4142$ $W_{13}^* = 0$ $W_{23}^* = 1.7321$	(e)	$W_{12}^* = 1.4142$ $W_{13}^* = 1.7321$ $W_{23}^* = -1.7321$	(f)	$W_{12}^* = -1.4142$ $W_{13}^* = -1.7321$ $W_{23}^* = 1.7321$
(g)	$W_{12}^* = 1.4142$ $W_{13}^* = 0$ $W_{23}^* = 0$	(h)	$W_{12}^* = -1.4142$ $W_{13}^* = 0$ $W_{23}^* = 0$	(i)	$W_{12}^* = 1.4142$ $W_{13}^* = 0$ $W_{23}^* = -1.7321$
(j)	$W_{12}^* = -1.4142$ $W_{13}^* = -1.7321$ $W_{23}^* = 0$	(k)	$W_{12}^* = 1.4142$ $W_{13}^* = -1.7321$ $W_{23}^* = -1.7321$	(l)	$W_{12}^* = -1.4142$ $W_{13}^* = -1.7321$ $W_{23}^* = -1.7321$

Thus, for example,

$$\begin{aligned} &P_0\{W_{uv}^* < 1.7321, (u,v) = (1,2), (1,3), (2,3)\} \\ &= P_0\{W_{12}^* < 1.7321; W_{13}^* < 1.7321; W_{23}^* < 1.7321\} \\ &= \tfrac{6}{12} = 1 - .5, \end{aligned}$$

because for 6 of the 12 configurations—(g), (h), (i), (j), (k), and (l)—the event $\{W_{12}^* < 1.7321; W_{13}^* < 1.7321; W_{23}^* < 1.7321\}$ occurs. Similarly, $P_0\{W_{uv}^* < 1.4142, (u,v) = (1,2), (1,3), (2,3)\} = \frac{3}{12} = 1 - .75$, because the event $\{W_{12}^* < 1.4142; W_{13}^* < 1.4142; W_{23}^* < 1.4142\}$ occurs only for the three configurations (h), (j), and (l). Finally, $P_0\{W_{uv}^* < 0, (u,v) = (1,2), (1,3), (2,3)\} = \frac{1}{12} = 1 - .9167$, corresponding to the single configuration (l). Hence, for $k = 3$, $n_1 = 1$, $n_2 = 1$, and $n_3 = 2$, we have $c_{.5000}^* = 1.7321$, $c_{.7500}^* = 1.4142$, and $c_{.9167}^* = 0$, and the values .5000, .7500, and .9167 are the only available experimentwise error rates for the Hayter–Stone procedure (6.68) in this setting.

For a given number of treatments k and sample sizes $n_1, \ldots, n_k$, the R command `cHayston(`α`,`$\boldsymbol{n}$`)` can be used to find the available critical values c_α^*. For a given available experimentwise error rate α, the critical value c_α^* is given by `cHayston(`α`,`$\boldsymbol{n}$`)`. Thus, for example, for $k = 3$ and $n_1 = 3$, $n_2 = 4$, and $n_3 = 6$, we have $c_{.0295}^* =$ `cHayston(.0295,c(3,4,6))` $= 3.015$.

63. *Maximum Type I Error Rate*. The multiple comparison procedure (6.68) is designed so that the experimentwise error rate (see Comment 61) is controlled to be equal to α; that is, the probability of falsely declaring any pair of treatment effects to be different, when in fact *all* the treatment effects are the same, is equal to α. However, it also satisfies the more stringent *maximum type I error rate* requirement that the probability of falsely declaring any pair of treatment effects to be different, regardless of the values of the other $k - 2$ treatment effects, is no larger than the stated α. This requires controlling the probability of making false declarations about treatment effect differences even in situations when *not all* the treatment effects are the same. For example, if $\tau_1 < \tau_2 = \tau_3$ the probability of incorrectly deciding that $\tau_2 \neq \tau_3$ is still controlled to be α by multiple comparison procedure (6.68). Similar comments apply to the approximate procedure in (6.70).

64. *Large and Unequal Sample Sizes*. In order to obtain the large-sample approximate critical values d_α for use in procedure (6.70) when we have an unbalanced setting (i.e., where the sample sizes are not all equal), we must evaluate a $(k-1)$-dimensional integral expression. In view of this difficulty (even with the availability of high-speed computers) and the fact that there is a large number of possible unequal-sample-size combinations for each fixed k and N, the evaluation of these approximate critical values is practically feasible only for a small percentage of the necessary cases. To make matters even more complicated, the use of an appropriate equal-sample-size asymptotic critical value when we actually have unequal sample sizes does not result in a conservative procedure, as it does for the Dwass–Steel–Critchlow–Fligner two-sided all-treatments multiple comparison procedure in Section 6.5 (see, e.g., Comment 58). In the case of the Hayter–Stone one-sided procedure (6.70), use of a particular equal-sample-size asymptotic critical value d_α may result in either a conservative or liberal (i.e.,

experimentwise error rate $\leq \alpha$ or $> \alpha$, respectively) procedure, depending on the particular unequal sample size configurations involved. Thus, for k and unequal $(n_1, \ldots, n_k)$ configurations beyond those for which the exact critical values c_α^* can reasonably be obtained from the R program `cHaySton`$(\alpha, \boldsymbol{n})$, Hayter and Stone (1991) recommended that computer simulation techniques be used to obtain appropriate asymptotic critical values.

Properties

1. *Asymptotic Multivariate Normality*. See Hayter (1984) and Hayter and Stone (1991).

Problems

52. Apply procedure (6.70) to the psychotherapeutic attraction data of Table 6.2 using the postulated order $\tau_1 \leq \tau_2 \leq \tau_3 \leq \tau_4$.

53. Procedure (6.68) is defined specifically in terms of the $k(k-1)/2$ pairwise two-sample rankings. However, it can be applied to settings where only the joint ranks of all N observations are available. Explain.

54. Apply procedure (6.68) to the average basal area increment data in Table 6.7. Use only the growing site index intervals 72–74, 75–77, and 78–80 with the postulated ordering $\tau_{72-74} \leq \tau_{75-77} \leq \tau_{78-80}$.

55. For the case $k = 3$, $\alpha = .05$, and $n_1 = n_2 = n_3 = 6$, compare procedures (6.68) and (6.70).

56. Find the totality of all available experimentwise error rates α and the associated critical values c_α^* for procedure (6.68) when $k = 4$, $n_1 = 1$, and $n_2 = n_3 = n_4 = 2$.

57. Consider the psychotherapeutic attraction data of Table 6.2 with the postulated ordering $\tau_1 \leq \tau_2 \leq \tau_3 \leq \tau_4$. Find the smallest (available) approximate experimentwise error rate at which the most significant difference in treatment effects would be detected.

58. Consider the average basal area increment data in Table 6.7. Using only the growing site index intervals 72–74, 75–77, and 78–80 with the postulated ordering $\tau_{72-74} \leq \tau_{75-77} \leq \tau_{78-80}$, find the smallest available experimentwise error rate at which we would declare $\tau_{78-80} > \tau_{72-74}$.

6.7 DISTRIBUTION-FREE ONE-SIDED TREATMENTS-VERSUS-CONTROL MULTIPLE COMPARISONS BASED ON JOINT RANKINGS (NEMENYI, DAMICO–WOLFE)

In this section our attention turns to a multiple comparison procedure designed to make decisions about individual differences between the median effect for a single, baseline control population and the median effects for each of the remaining $k-1$ treatments. This treatment versus control multiple comparison procedure is based on the joint ranking of all N sample observations and can be applied to one-way layout data containing a single control sample *after* rejection of H_0 (6.2) with any of the test procedures in Sections 6.1–6.4. Its application leads to conclusions about the differences between each of the $k-1$ treatment effects and the control effect, and these conclusions are naturally one-sided in nature.

Procedure

For simplicity of notation, we let treatment 1 assume the role of the single baseline control. In addition, let N^* be the least common multiple of the sample sizes $n_1, \ldots, n_k$. Jointly rank all N of the sample observations and let $R_{\cdot 1}, \ldots, R_{\cdot k}$ be the averages of these joint ranks associated with treatments $1, \ldots, k$, respectively. (Thus, $R_{\cdot 1}, \ldots, R_{\cdot k}$ are as originally defined in (6.4) in conjunction with the Kruskal–Wallis statistic.) For each of the $k-1$ noncontrol treatments, calculate the difference $R_{\cdot u} - R_{\cdot 1}$, $u = 2, \ldots, k$.

At an experimentwise error rate of α, the Nemenyi–Damico–Wolfe one-sided treatments-versus-control multiple comparison procedure (see Comment 65) reaches its $k-1$ pairwise decisions, corresponding to each (τ_1, τ_u) pair, $u = 2, \ldots, k$, by the criterion

$$\text{Decide } \tau_u > \tau_1 \text{ if } N^*(R_{\cdot u} - R_{\cdot 1}) \geq y_\alpha^*; \quad \text{otherwise decide } \tau_u = \tau_1, \tag{6.71}$$

where the constant y_α^* is chosen to make the experimentwise error rate equal to α; that is, y_α^* satisfies the restriction

$$P_0\{N^*(R_{\cdot u} - R_{\cdot 1}) < y_\alpha^*,\ u = 2, \ldots, k\} = 1 - \alpha, \tag{6.72}$$

where the probability $P_0(\cdot)$ is computed under H_0 (6.2). Equation (6.72) stipulates that the $k-1$ inequalities $N^*(R_{\cdot u} - R_{\cdot 1}) < y_\alpha^*$, corresponding to all pairs $(1, u)$ of noncontrol treatments ($u = 2, \ldots, k$) with the control treatment 1, hold simultaneously with probability $1 - \alpha$ when H_0 (6.2) is true. Comment 68 explains how to obtain the critical value y_α^* for $(k-1)$ noncontrol treatments, sample sizes $n_1, \ldots, n_k$, and available experimentwise error rates α.

Large-Sample Approximations

When H_0 is true, the $(k-1)$ component vector $(R_{\cdot 2} - R_{\cdot 1}, R_{\cdot 3} - R_{\cdot 1}, \ldots, R_{\cdot k} - R_{\cdot 1})$ has, as $\min(n_1, \ldots, n_k)$ tends to infinity, an asymptotic $(k-1)$-variate normal distribution with mean vector $\mathbf{0}$. (For an indication of the proof, see Miller (1966).) For the special case of $n_1 = b$ and $n_2 = \cdots = n_k = n$, with both n and b large, the critical value y_α^* can be approximated by $[N(N+1)/12]^{1/2}[(1/b) + (1/n)]^{1/2} N^* m_\alpha^*$, where m_α^* is the upper αth percentile point for the distribution of the maximum of $(k-1)N(0,1)$ variables with common correlation $\rho = n/(b+n)$. Thus, the large-sample approximation for procedure (6.71) when we have equal treatment sample sizes $n_2 = \cdots = n_k = n$ (possibly different from $b = n_1$) is

$$\text{Decide } \tau_u > \tau_1 \text{ if } (R_{\cdot u} - R_{\cdot 1}) \geq m_\alpha^* \left[\frac{N(N+1)}{12}\right]^{1/2} \left(\frac{1}{b} + \frac{1}{n}\right)^{1/2};$$
$$\text{otherwise decide } \tau_u = \tau_1, u = 2, \ldots, k. \tag{6.73}$$

To find m_α^* for $k-1$ noncontrol treatments, number of control observations b and an equal number, n, of observations from each of the noncontrol treatments, we use the R command `cMaxNorm(`α`, k - 1, n/(b+n))`. For example, to find $m_{.0310}^*$ for $k - 1 = 4$ noncontrol treatments, $b = 9$, and $n = 3$, we have $\rho = 3/(9+3) = .25$ and apply `cMaxNorm(.0310, 4, 0.25)` to obtain $m_{.0310}^* = 2.40$.

For the general setting of arbitrary (not necessarily equal) treatments sample sizes, Dunn (1964) used Bonferroni's Inequality to provide the large-sample approximation to procedure (6.71) given by

$$\text{Decide } \tau_u > \tau_1 \text{ if } (R_{.u} - R_{.1}) \geq z_{\alpha^*} \left[\frac{N(N+1)}{12}\right]^{1/2} \left(\frac{1}{n_1} + \frac{1}{n_u}\right)^{1/2};$$
$$\text{otherwise decide } \tau_u = \tau_1, u = 2, \ldots, k, \tag{6.74}$$

where $\alpha^* = \alpha/(k-1)$. (We note that this general approximate procedure can often be quite conservative in practice, as a direct result of the conservative nature of the Bonferroni Inequality.)

Ties

If there are ties among the X observations, use average ranks in computing the individual treatment sums of ranks $R_1, \ldots, R_k$.

EXAMPLE 6.8 ***Motivational Effect of Knowledge of Performance—Example 6.2 Continued.***

Once again we consider Hundal's (1969) study to assess the motivational effects of knowledge of performance. We previously found in Example 6.2 (using the Jonckheere–Terpstra test procedure) that there was sufficient evidence in the sample data to conclude that $\tau_1 \leq \tau_2 \leq \tau_3$ with at least one strict inequality. To further investigate which (if either) of the two types of additional information (rough or accurate) lead to differences in median numbers of pieces processed relative to the no information control (treatment 1), we apply procedure (6.71). Here, we have $k = 3$ and $n_1 = n_2 = n_3 = N^* = 6$. For the sake of illustration, we take our experimentwise error rate to be $\alpha = .0554$. With $k = 3$ and $n_1 = n_2 = n_3 = 6$, we find $y^*_{.0554} =$ `cNDWol(.0554, c(6, 6, 6))` $= 35$ and procedure (6.71) reduces to

$$\text{Decide } \tau_u > \tau_i \text{ if } 6(R_{.u} - R_{.1}) = (R_u - R_1) \geq 35.$$

Using the joint ranks (with average ranks to break ties among the observations) provided in parentheses beside the data in Table 6.6, we see that

$$R_1 = 5.5 + 1 + 2.5 + 10.5 + 13 + 8 = 40.5,$$

$$R_2 = 2.5 + 5.5 + 17 + 13 + 5.5 + 9 = 52.5,$$

and

$$R_3 = 18 + 5.5 + 15 + 10.5 + 16 + 13 = 78.$$

Thus, $(R_2 - R_1) = (52.5 - 40.5) = 12$ and $(R_3 - R_1) = (78 - 40.5) = 37.5$. Referring these rank sum differences to the critical point $y^*_{.0554} = 35$, we see that

$$(R_2 - R_1) = 12 < 35 \quad \Rightarrow \text{decide } \tau_2 = \tau_1,$$

$$(R_3 - R_1) = 37.5 \geq 35 \Rightarrow \text{decide } \tau_3 > \tau_1.$$

Thus at an experimentwise error rate of .0554, we have reached the conclusion that accurate information leads to significantly more pieces processed than the no information control. (We note that the smallest experimentwise error rate at which we would reach this conclusion is .0426, as $y^*_{.0426} =$ `cNDWol(.0426, c(6, 6, 6))` $= 37$ and $y^*_{.0371} =$ `cNDWol(.0371, c(6, 6, 6))` $= 38$.)

For the sake of illustration for the associated large-sample approximation (with equal sample sizes) procedure in (6.73), we note that $\rho = n/(b+n) = 6/(6+6) = \frac{1}{2}$ (which is always the case with equal sample sizes in the control and the noncontrol treatments). Using an approximate experimentwise error rate of $\alpha = .05183$ with $(k-1) = 2$, we see that `cMaxNorm(.05183, 2, 0.5)` $= m^*_{.05183} = 1.90$. Thus, we have that

$$\left[\frac{N(N+1)}{12}\right]^{1/2}\left(\frac{1}{b}+\frac{1}{n}\right)^{1/2} m^*_{.05183} = \left[\frac{18(19}{12}\right]^{1/2}\left(\frac{1}{6}+\frac{1}{6}\right)^{1/2}(1.90) = 5.856$$

and procedure (6.73) becomes

$$\text{Decide } \tau_u > \tau_1 \text{ if } (R_{.u} - R_{.1}) \geq 5.856$$

or, equivalently,

$$\text{Decide } \tau_u > \tau_1 \text{ if } (R_u - R_1) = 6(R_{.u} - R_{.1}) \geq 6(5.856) = 35.14.$$

Thus, for $k = 3$ and $n_1 = n_2 = n_3 = 6$, the exact procedure (6.71) and the large-sample approximation (6.73) are virtually identical and lead to the same conclusions that $\tau_2 = \tau_1$ and $\tau_3 > \tau_1$.

We note that the treatment-versus-control procedure (6.71) yields the conclusion that $\tau_3 > \tau_1$ at a considerably smaller experimentwise error rate (as low as .0426) than is the case with the Hayter–Stone one-sided all-treatments multiple comparison procedure (as detailed in Example 6.7), where the smallest experimentwise error rate leading to this conclusion is .0850. This situation is due primarily to the fact that the Hayter–Stone procedure is required to make an additional decision about the relative magnitude of τ_2 and τ_3, which, for these data, do not appear to be significantly different.

Comments

65. *Rationale for Treatments-versus-Control Multiple Comparison Procedures.* The general rationale for the multiple comparison procedures of this section is the same as that given in Comment 53 for the two-sided all-treatments multiple comparison procedures of Section 6.5. The only additional factor here is that the treatment-versus-control procedures of this section do not compare all treatments, but only each noncontrol treatment with the control on a directional bias. This situation arises, for example, in drug screening in the examination of many new treatments in hopes of improving on a standard, and there is no initial reason to perform between treatment comparisons. Of course, similar comparisons would be carried out later between treatments that were selected as being better than the control.

66. *Experimentwise Error Rate.* The use of an experimentwise error rate represents a very conservative approach to multiple comparisons. We are insisting that the probability of making only correct decisions be $1-\alpha$ when the hypothesis H_0 (6.2) of treatment equivalence is true. Thus, we have a high degree of protection when H_0 is true, but we often apply the techniques of this section when we have evidence (perhaps based on a priori information or perhaps obtained by applying a previous test procedure, as in Example 6.8) that H_0 is not true. (For additional general remarks about experimentwise error rates, see Comment 54.)

67. *Opposite Direction Decisions.* Procedures (6.71), (6.73), and (6.74) are designed for the one-sided case where the decisions are $\tau_u > \tau_1$ versus $\tau_u = \tau_1,\ u = 2,\ldots,k$. To handle the analogous one-sided situation where the decisions involve $\tau_u < \tau_1$ versus $\tau_u = \tau_1,\ u = 2,\ldots,k$, we use (6.71), (6.73), and (6.74) with $(R_{.u} - R_{.1})$ replaced by $(R_{.1} - R_{.u})$ for $u = 2,\ldots,k$.

68. *Critical Values* y_α^*. The y_α^* critical values can be obtained by using the fact that under H_0 (6.2), all $N!/\left[\prod_{j=1}^{k} n_j!\right]$ rank assignments are equally likely. However, in this one-sided treatments-versus-control setting, we must work a little harder than in the two-sided all-treatments case (see Comments 55 and 59 in Section 6.5) as the values $R_{.u} - R_{.1},\ u = 2,\ldots,k$, are, in general, changed when we relabel the control treatment. (In either of the previous two-sided all-treatments cases, the relevant statistic is unaffected by treatment relabelings.) As a result, we will have to take the complete enumeration approach employed in Comment 62 for the one-sided all-treatments setting, where the relevant statistic is also not invariant with respect to treatment relabelings.

    For an illustration, we return to Comment 17 and use the 12 rank configurations displayed there for the case $k = 3$, $n_1 = 1$, $n_2 = 1$, and $n_3 = 2$. (Here, $N^* = 2$.) For each of these 12 configurations, we now display the values of $2(R_{.2} - R_{.1})$ and $2(R_{.3} - R_{.1})$.

(a)	$2(R_{.2} - R_{.1}) = 2$ $2(R_{.3} - R_{.1}) = 5$	(b)	$2(R_{.2} - R_{.1}) = -2$ $2(R_{.3} - R_{.1}) = 3$	(c)	$2(R_{.2} - R_{.1}) = 4$ $2(R_{.3} - R_{.1}) = 4$
(d)	$2(R_{.2} - R_{.1}) = -4$ $2(R_{.3} - R_{.1}) = 0$	(e)	$2(R_{.2} - R_{.1}) = 6$ $2(R_{.3} - R_{.1}) = 3$	(f)	$2(R_{.2} - R_{.1}) = -6$ $2(R_{.3} - R_{.1}) = -3$
(g)	$2(R_{.2} - R_{.1}) = 2$ $2(R_{.3} - R_{.1}) = 1$	(h)	$2(R_{.2} - R_{.1}) = -2$ $2(R_{.3} - R_{.1}) = -1$	(i)	$2(R_{.2} - R_{.1}) = 4$ $2(R_{.3} - R_{.1}) = 0$
(j)	$2(R_{.2} - R_{.1}) = -4$ $2(R_{.3} - R_{.1}) = -4$	(k)	$2(R_{.2} - R_{.1}) = 2$ $2(R_{.3} - R_{.1}) = -3$	(l)	$2(R_{.2} - R_{.1}) = -2$ $2(R_{.3} - R_{.1}) = -5$

    Thus, for example,

$$
\begin{aligned}
&P_0\{2(R_{.u} - R_{.1}) < 6, u = 2, 3\} \\
&= P_0\{2(R_{.2} - R_{.1}) < 6 \text{ and } 2(R_{.3} - R_{.1}) < 6\} \\
&= \tfrac{11}{12} = 1 - .0833,
\end{aligned}
$$

because the event $\{2(R_{.2} - R_{.1}) < 6$ and $2(R_{.3} - R_{.1}) < 6\}$ occurs for all but configuration (e). Similarly, $P_0\{2(R_{.u} - R_{.1}) < 2, u = 2, 3\} = \frac{5}{12} = 1 - .5833$, because the event $\{2(R_{.2} - R_{.1}) < 2$ and $2(R_{.3} - R_{.1}) < 2\}$ occurs only for the five configurations (d), (f), (h), (j), and (l). Hence, for $k = 3$, $n_1 = 1$, $n_2 = 1$, and $n_3 = 2$, we have $y^*_{.0833} = 6$ and $y^*_{.5833} = 2$. The other possible experimentwise error rates (there are 10, including 1, all together) and the associated critical values for this setting are obtained through the same type of enumeration.

For a given number of noncontrol treatments $k - 1$ and sample sizes $n_1, \ldots, n_k$, the R command `cNDWol`$(\alpha, \boldsymbol{n})$ can be used to find the available critical values y^*_α. For a given available experimentwise error rate α, the critical value y^*_α is given by `cNDWol`$(\alpha, \boldsymbol{n})$. Thus, for example, for $k = 3$ and $n_1 = n_2 = n_3 = 6$, we have $y^*_{.0795} =$ `cNDWol(.0795, c(6,6,6))` $= 32$.

69. *Interpretation as Hypothesis Tests*. Procedures (6.71), (6.73), and (6.74) can also be interpreted as hypothesis tests of H_0 (6.2). (For example, the procedure that rejects H_0 if at least one of the $k - 1$ inequalities of (6.71) holds is a distribution-free test of level α for H_0.) However, they are generally more effective at detecting differences between individual treatment effects when applied to data for which the null hypothesis H_0 has previously been rejected by one of the test procedures is Sections 6.1–6.4.

70. *Dependence on Observations from Other Noninvolved Treatments*. The differences $(R_{.u} - R_{.1})$ depend on the values of the observations from the other $k - 2$ treatments, in addition to the observations from the control and treatment u. Thus, the multiple comparison procedures in (6.71), (6.73), and (6.74) all have the disadvantage that the decision concerning treatment u and the control can be affected by changes only in the observations from one or more of the other $k - 2$ noninvolved treatments. This difficulty has been emphasized by Miller (1966) and Gabriel (1969).

71. *Two-Sided Treatments-versus-Control Multiple Comparison Procedures*. All the multiple comparison procedures of this section are one-sided by nature, resulting in decisions between $\tau_u = \tau_1$ and $\tau_u > \tau_1$ for every $u = 2, \ldots, k$ (or between $\tau_u = \tau_1$ and $\tau_u < \tau_1$ for every $u = 2, \ldots, k$, as noted in Comment 67). We view such one-sided comparisons to be the most natural approach for treatments-versus-control settings. In such situations, we are generally interested in seeing which, if any, of the proposed new treatments are better than a standard control or placebo. In most practical applications, *better* is synonymous with one-sided comparisons (all in one direction or all in the other)—thus our emphasis on such procedures in this section. However, a two-sided treatments-versus-control analog to procedure (6.71) has been developed in the literature and corresponds to the criterion

$$\text{Decide } \tau_u \neq \tau_1 \text{ if } N^*|R_{.u} - R_{.1}| \geq y^{**}_\alpha; \quad \text{otherwise decide } \tau_u = \tau_1, \quad (6.75)$$

where the constant y^{**}_α is chosen to make the experimentwise error rate equal to α; that is,

$$P_0\{N^*|R_{.u} - R_{.1}| < y^{**}_\alpha,\ u = 2, \ldots, k\} = 1 - \alpha,$$

where the probability $P_0(\cdot)$ is computed under H_0 (6.2). However, the required critical values y^{**}_α are available only in a very limited fashion. Leach (1972) has

provided such critical values y_α^{**} for the very special case of $k = 3$ and equal sample sizes $n_1 = n_2 = n_3 = 2(1)6$. Associated large-sample approximations to (6.75) for equal and unequal noncontrol treatment sample sizes have been considered by Miller (1966) and Dunn (1964), respectively. For further discussion of these two-sided treatments-versus-control multiple comparison procedures, see Miller (1966).

72. *Pairwise Ranking Approach*. The treatments-versus-control multiple comparison procedures discussed in this section are based on the joint ranking of all N of the sample observations. They suffer from the same drawbacks as do other one-way layout multiple comparison procedures based on joint rankings. For example, they do not provide the maximum type I error rate protection level α guarantee and decisions between treatment u and the control depend on the values of the observations from the other $k - 2$ treatments (for more details, see, e.g., Fligner (1984)).

    Steel (1959) developed a competitor of these Nemenyi–Damico–Wolfe procedures that takes the pairwise ranking approach discussed in Sections 6.5 and 6.6. His procedure is based on $k - 1$ separate two-sample rankings between the control sample and each of the $k - 1$ noncontrol samples and has the form

$$\text{Decide } \tau_u > \tau_1 \text{ if } W_{1u}^* \geq b_\alpha^*; \quad \text{otherwise decide } \tau_u = \tau_1, u = 2, \ldots, k, \quad (6.76)$$

    where $W_{12}^*, \ldots, W_{1k}^*$ are defined by (6.61) and b_α^* is chosen to make the experimentwise error rate equal to α. This pairwise ranking treatments-versus-control procedure has many of the nice properties of the analogous pairwise rankings all-treatments multiple comparison procedures discussed in Sections 6.5 and 6.6, including proper control of the maximum type I error rate (see Comments 57 and 63).

Properties

1. *Asymptotic Multivariate Normality*. See Miller (1966).
2. *Efficiency*. See Sherman (1965) and Section 6.10.

Problems

59. Apply the approximate procedure (6.73) to the psychotherapeutic attraction data in Table 6.2.

60. For the case $k = 3$, $\alpha = .01$, $n_1 = n_2 = n_3 = 6$, compare procedures (6.71) and (6.73).

61. For the psychotherapeutic attraction data in Table 6.2, find the smallest approximate experimentwise error rate at which we would decide $\tau_4 > \tau_1$ using procedure (6.73).

62. Consider the mucociliary clearance data in Table 6.1. Use procedure (6.71) to decide whether or not either obstructive airways disease or asbestosis (or both) lead to a deterioration (slowdown) in median mucociliary clearance half-times.

63. Apply the approximate procedure (6.74) to the glucocorticoid receptor site data in Table 6.4.

64. For the glucocorticoid receptor site data in Table 6.4, find the smallest approximate experimentwise error rate at which we would decide $\tau_5 > \tau_1$ using procedure (6.74).

65. Apply the approximate procedure (6.73) to the plasma glucose data in Table 6.9.

66. For the plasma glucose data in Table 6.9, find the smallest approximate experimentwise error rate at which we would decide $\tau_3 > \tau_1$ using procedure (6.73).

67. For the plasma glucose data in Table 6.9, find the smallest approximate experimentwise error rate at which we would decide $\tau_5 < \tau_1$ with an appropriate treatments-versus-control multiple comparison procedure (see Comment 67).

68. Apply the approximate procedure (6.73) to the revertant colonies data in Table 6.10.

69. For the revertant colonies data in Table 6.10, find the smallest approximate experimentwise error rate at which we would decide $\tau_4 > \tau_1$ using procedure (6.73).

70. Consider the revertant colonies data in Table 6.10. Find the smallest approximate experimentwise error rate at which the most significant difference in treatment (dosage) effects would be detected.

71. Find the totality of all available experimentwise error rates α and the associated critical values y_α^* for procedure (6.71) when $k = 4$, $n_1 = 1$, and $n_2 = n_3 = n_4 = 2$.

6.8 CONTRAST ESTIMATION BASED ON HODGES–LEHMANN TWO-SAMPLE ESTIMATORS (SPJøTVOLL)

In this section we discuss a method for the point estimation of certain linear combinations of treatment effects known in the literature as *contrasts*. We define such a contrast in the treatment effects $\tau_1, \ldots, \tau_k$ to be any linear combination of the form

$$\theta = \sum_{i=1}^{k} a_i \tau_i, \tag{6.77}$$

where $a_1, \ldots, a_k$ are any specified set of constants such that $\sum_{i=1}^{k} a_i = 0$. Equivalently, we can write θ in terms of the individual differences in treatment effects (known in the literature as *simple contrasts*)

$$\Delta_{hj} = \tau_h - \tau_j, \quad h = 1, \ldots, k; \; j = 1, \ldots, k, \tag{6.78}$$

by noting that

$$\theta = \sum_{h=1}^{k} \sum_{j=1}^{k} d_{hj} \Delta_{hj}, \tag{6.79}$$

where

$$d_{hj} = \frac{a_h}{k}, \; h = 1, \ldots, k; \quad j = 1, \ldots, k. \tag{6.80}$$

For a given setting, decisions about which contrasts to estimate can be related either to a priori interest in particular linear combinations of the τ's or the results of one of the multiple comparison procedures discussed in Sections 6.5–6.7.

Procedure

For each pair of treatments $(h, j), h \neq j = 1, \ldots, k$, define the pairwise estimators

$$Z_{hj} = \text{median } \{X_{\alpha h} - X_{\beta j}, \alpha = 1, \ldots, n_h; \beta = 1, \ldots, n_j\}. \tag{6.81}$$

As $Z_{hj} = -Z_{jh}$, we need to calculate only the $k(k-1)/2$ estimators Z_{hj} corresponding to $h < j$. We refer to Z_{hj} as the raw or unadjusted estimator of the simple contrast $\Delta_{hj} = \tau_h - \tau_j$. (Note that Z_{hj} is exactly the Hodges–Lehmann two-sample estimator defined in Section 4.2, as applied to the hth sample (playing the role of the Y's) and the jth sample (playing the role of the X's). For example, Z_{13} is simply the median of the $n_1 n_3$ differences $X_{\alpha 1} - X_{\beta 3}$ obtained from the treatments 1 and 3 observations.) Next, we obtain the set $\overline{\Delta}_1, \ldots, \overline{\Delta}_k$ of individual weighted average of these unadjusted estimators Z_{hj} corresponding to

$$\overline{\Delta}_h = \sum_{j=1}^{k} \left(\frac{n_j}{N}\right) Z_{hj}, \ h = 1, \ldots, k, \tag{6.82}$$

where we note that $Z_{hh} = 0$ for $h = 1, \ldots, k$.

The weighted-adjusted estimator of the contrast θ (6.77) is given by

$$\widehat{\theta} = \sum_{i=1}^{k} a_i \overline{\Delta}_i, \tag{6.83}$$

or, equivalently,

$$\widehat{\theta} = \sum_{h=1}^{k} \sum_{j=1}^{k} d_{hj} (\overline{\Delta}_h - \overline{\Delta}_j) = \sum_{h=1}^{k} \sum_{j=1}^{k} d_{hj} \ W_{hj}, \tag{6.84}$$

where

$$W_{hj} = \overline{\Delta}_h - \overline{\Delta}_j = \widehat{\Delta}_{hj} \tag{6.85}$$

is the weighted-adjusted estimator of the simple contrast $\Delta_{hj} = \tau_h - \tau_j$. We note that in the special case $n_1 = n_2 = \cdots = n_k$, $\overline{\Delta}_h$ (6.82) reduces to

$$Z_{h\cdot} = \frac{\sum_{j=1}^{k} Z_{hj}}{k}, \quad h = 1, \ldots, k, \tag{6.86}$$

and $W_{hj} = \overline{\Delta}_h - \overline{\Delta}_j$ (6.85) can be simplified to

$$W_{hj} = Z_{h\cdot} - Z_{j\cdot}, \quad h \neq j = 1, \ldots, k. \tag{6.87}$$

EXAMPLE 6.9 ***Motivational Effect of Knowledge of Performance—Examples 6.2 and 6.8 Continued.***

Consider the Hundal knowledge of performance data originally presented in Example 6.2. In the application of the Nemenyi–Damico–Wolfe one-sided treatments-versus-control multiple comparison procedure (Example 6.8) to these data, we concluded that the group

receiving accurate information about their output produced significantly more (experimentwise error rate .0554) pieces than the group that received no information. Thus, it is of interest to use the knowledge of performance data in Table 6.6 to estimate the simple contrast $\theta = \tau_{\text{accurate information}} - \tau_{\text{no information}} = \tau_3 - \tau_1$, thereby providing an idea of the increased output that might be expected for this task by providing accurate information to the workers.

From Table 6.6 and (6.81), the three pairwise estimators are

$$\begin{aligned} Z_{12} &= \text{median}\{2, 0, -7, -4, 0, -2, -3, -5, -12, -9, -5, -7, 0, -2, -9, \\ &\qquad -6, -2, -4, 5, 3, -4, -1, 3, 1, 6, 4, -3, 0, 4, 2, 3, 1, -6, -3, 1, -1\} \\ &= -1.5, \\ Z_{13} &= \text{median}\{-8, 0, -5, -3, -6, -4, -13, -5, -10, -8, -11, -9, -10, -2, -7, \\ &\qquad -5, -8, -6, -5, 3, -2, 0, -3, -1, -4, 4, -1, 1, -2, 0, -7, 1, -4, -2, -5, -3\} \\ &= -4, \end{aligned}$$

and

$$\begin{aligned} Z_{23} &= \text{median}\{-10, -2, -7, -5, -8, -6, -8, 0, -5, -3, -6, -4, -1, 7, 2, \\ &\qquad 4, 1, 3, -4, 4, -1, 1, -2, 0, -8, 0, -5, -3, -6, -4, -6, 2, -3, -1, -4, -2\} \\ &= -3. \end{aligned}$$

From expression (6.82), or equivalently (as $n_1 = n_2 = n_3 = 6$) from (6.86), we have

$$\overline{\Delta}_1 = \frac{Z_{11} + Z_{12} + Z_{13}}{3} = \frac{0 - 1.5 - 4}{3} = -\frac{11}{6},$$

$$\overline{\Delta}_2 = \frac{Z_{21} + Z_{22} + Z_{23}}{3} = \frac{1.5 + 0 - 3}{3} = -.5,$$

and

$$\overline{\Delta}_3 = \frac{Z_{31} + Z_{32} + Z_{33}}{3} = \frac{4 + 3 + 0}{3} = \frac{7}{3}.$$

(Note that in calculating $\overline{\Delta}_2$ and $\overline{\Delta}_3$, we have used the fact that $Z_{21} = -Z_{12}, Z_{31} = -Z_{13}$, and $Z_{32} = -Z_{23}$.) The weighted-adjusted estimator of $\theta = \tau_3 - \tau_1$ is now obtained from (6.83) with $a_1 = -1, a_2 = 0$, and $a_3 = 1$. We find

$$\widehat{\theta} = W_{31} = \overline{\Delta}_3 - \overline{\Delta}_1 = \tfrac{7}{3} - \left(-\tfrac{11}{6}\right) = \tfrac{25}{6} = 4.17 \text{ pieces}.$$

(We note that the values of the raw estimator Z_{31} and the classical estimator $\overline{X}_3 - \overline{X}_1$ are 4.00 and 4.17, respectively, for these data.)

Comments

73. *Ambiguities with the Unadjusted Estimators*. The unadjusted estimators Z_{hj} (6.81) lead to ambiguities in contrast estimation because they do not satisfy the linear relations that are satisfied by the contrasts they estimate. For example, $\Delta_{13} = \tau_1 - \tau_3 = (\tau_1 - \tau_2) + (\tau_2 - \tau_3) = \Delta_{12} + \Delta_{23}$, but in general $Z_{13} \neq Z_{12} + Z_{23}$. Thus, the two "reasonable" estimators Z_{13} and $Z_{12} + Z_{23}$ of $\Delta_{13} = \tau_1 - \tau_3$ can give different estimates. This was pointed out by Lehmann (1963a), who called the unadjusted estimators incompatible.
74. *Compatible, but Inconsistent Estimators*. Lehmann (1963a) removed the incompatibility difficulty discussed in Comment 73 by using the estimators $W_{hj} = Z_{h\cdot} - Z_{j\cdot}$ (6.87). These estimators are obtained by minimizing the sum of squares $\sum\sum_{h\neq j}[Z_{hj} - (\tau_h - \tau_j)]^2$. Although these estimators are compatible, Lehmann also pointed out that two additional difficulties now arise. First, the estimator $Z_{h\cdot} - Z_{j\cdot}$ of $\Delta_{hj} = \tau_h - \tau_j$ depends, in addition to the observations from samples h and j, on the observations from the other $k - 2$ samples. Furthermore, in the case of $k = 3$, for example, the estimator $Z_{1\cdot} - Z_{2\cdot}$ (of $\tau_1 - \tau_2$) is not consistent when n_1 and n_2 tend to infinity unless n_3 also tends to infinity.
75. *Consistency*. Spjøtvoll (1968) removed the nonconsistency difficulty by obtaining the weighted-adjusted estimators $W_{hj} = \overline{\Delta}_h - \overline{\Delta}_j$ (6.85). These estimators minimize the sum of squares $N^{-2}\sum\sum_{h\neq j} n_h n_j [Z_{hj} - (\tau_h - \tau_j)]^2$. Spjøtvoll's estimators do, however, retain the disadvantage that the estimator of $\tau_h - \tau_j$ depends on unrelated observations from the other samples.
76. *Competitor Contrast Estimator*. Spjøtvoll (1968) also proposed weighted-adjusted estimators that minimize

$$\sum_{h\neq j}\sum \left(\frac{N}{n_h} + \frac{N}{n_j}\right)^{-1} [Z_{hj} - (\tau_h - \tau_j)]^2, \tag{6.88}$$

    using the asymptotic variances of the Z_{hj}'s as weights in the sum of squares. These estimators are more difficult to compute than the estimators W_{hj} (6.85). Furthermore, Spjøtvoll showed that the weighted-adjusted estimators W_{hj} (6.85) and those obtained by minimizing (6.88) have the same asymptotic properties when n_j tends to infinity in such a way that (n_j/N) tends to λ_j with $0 < \lambda_j < 1$, for $j = 1, \ldots, k$.
77. *Equivalence with Equal Sample Sizes*. Spjøtvoll pointed out that the estimator of Δ_{hj} obtained by minimizing (6.88) and the estimator W_{hj} (6.85) both reduce to Lehmann's estimator $Z_{h\cdot} - Z_{j\cdot}$ (6.88) when $n_1 = n_2 = \cdots = n_k$.

Properties

1. *Standard Deviation of $\widehat{\theta}$ (6.83)*. For the asymptotic standard deviation of $\widehat{\theta}$ (6.83), see Spjøtvoll (1968).
2. *Asymptotic Normality*. See Spjøtvoll (1968) and Lehmann (1963a).
3. *Efficiency*. See Spjøtvoll (1968), Lehmann (1963a), and Section 6.10.

Problems

72. Estimate the simple contrast $\theta = \tau_4 - \tau_1$ for the psychotherapeutic attraction data in Table 6.2.

73. Estimate the simple contrasts $\theta_1 = \tau_2 - \tau_1, \theta_2 = \tau_4 - \tau_1$, and $\theta_3 = \tau_5 - \tau_1$ for the glucocorticoid receptor sites data in Table 6.4.

74. Estimate all possible simple contrasts for the mean interstitial lengths data in Table 6.5.

75. Estimate the simple contrasts $\tau_2 - \tau_1, \tau_4 - \tau_1$, and $\tau_5 - \tau_1$ for the BAI data in Table 6.7.

76. Take $k = 4$ and construct a data example where $Z_{13} + Z_{24} \neq Z_{14} + Z_{23}$. (Note that $\Delta_{13} + \Delta_{24} = \Delta_{14} + \Delta_{23}$. See also Comment 73.)

77. As suggested by the application of the Dwass–Steel–Critchlow–Fligner multiple comparison procedure (in Example 6.6), estimate the contrast $\theta = [\frac{1}{2}(\tau_1 + \tau_2) - \frac{1}{2}(\tau_3 + \tau_4)]$ for the gizzard shad data in Table 6.3.

78. Estimate the contrast $\theta = [\frac{1}{3}(\tau_2 + \tau_3 + \tau_4) - \frac{1}{3}(\tau_1 + \tau_5 + \tau_6)]$ for the revertant colonies data in Table 6.10.

79. Estimate the simple contrasts $\tau_2 - \tau_1$ and $\tau_3 - \tau_1$ for the plasma glucose data in Table 6.9.

80. Estimate all contrasts found to be of interest in Problem 59 for the psychotherapeutic attraction data in Table 6.2.

6.9 SIMULTANEOUS CONFIDENCE INTERVALS FOR ALL SIMPLE CONTRASTS (CRITCHLOW–FLIGNER)

A contrast θ (6.77) is said to be a simple contrast if it involves only two treatment effects (i.e., all but two of the a_i coefficients are zero). In this section, we present a method for obtaining simultaneous confidence intervals for the entire collection, C, of all $\binom{k}{2}$ simple contrasts given by

$$C = \{\Delta_{uv} : \Delta_{uv} = \tau_v - \tau_u, 1 \leq u < v \leq k\}. \tag{6.89}$$

Procedure

For each pair of treatments $(u, v), u \neq v = 1, \ldots, k$, define the sample differences

$$D_{ij}^{uv} = X_{jv} - X_{iu}, \quad i = 1, \ldots, n_u^*; \quad j = 1, \ldots, n_v. \tag{6.90}$$

Let $D_{(1)}^{uv} \leq D_{(2)}^{uv} \leq \cdots \leq D_{(n_u n_v)}^{uv}$ denote the ordered values of the $n_u n_v$ D_{ij}^{uv} differences, for $u \neq v = 1, \ldots, k$. Let w_α^* be the upper αth percentile for the distribution of maximum $\{|W_{uv}^*|, u \neq v = 1, \ldots, k\}$ under H_0 (6.2), where W_{uv}^* is the standardized two-sample rank sum statistic (multiplied by $\sqrt{2}$) for the uth and vth samples, as defined previously in (6.61) for the two-sided all-treatments multiple comparisons setting. Comment 55 explains how to obtain the critical values w_α^* for k treatments, sample sizes $n_1, \ldots, n_k$, and available experimentwise error rates α.

For $1 \leq u < v \leq k$, set

$$a_{uv} = \frac{n_u n_v}{2} - w_\alpha^* \left[\frac{n_u n_v (n_u + n_v + 1)}{24}\right]^{1/2} + 1 \tag{6.91}$$

and

$$b_{uv} = a_{uv} - 1. \tag{6.92}$$

The simultaneous $100(1-\alpha)\%$ confidence intervals for the collection C (6.89) of all simple contrasts are then

$$\{[D^{uv}_{(\langle a_{uv}\rangle)}, D^{uv}_{(n_u n_v - \langle b_{uv}\rangle)}), 1 \le u < v \le k\}, \tag{6.93}$$

where $\langle t \rangle$ denotes the greatest integer less than or equal to t. This set of intervals satisfies the condition

$$\begin{aligned} P_{\tau_1,\ldots,\tau_k}(D^{uv}_{(\langle a_{uv}\rangle)} \le \tau_v - \tau_u < D^{uv}_{(n_u n_v - \langle b_{uv}\rangle)}, \text{ for } 1 \le u < v \le k) \\ = 1 - \alpha, \quad \text{for all } -\infty < \tau_i < \infty, i = 1, \ldots, k. \end{aligned} \tag{6.94}$$

(For simultaneous lower confidence bounds for the collection C that are appropriate under the ordered alternatives setting of Section 6.2, see Comment 79.)

Large-Sample Approximation

When H_0 is true, the $[k(k-1)/2]$-component vector $(W^*_{12}, W^*_{13}, \ldots, W^*_{k-1,k})$ has, as min $(n_1, \ldots, n_k)$ tends to infinity, an asymptotic multivariate normal distribution with mean vector $\mathbf{0}$. It then follows (see Comment 58) that w^*_α can be approximated for large sample sizes by q_α, where q_α is the upper αth percentile point for the distribution of the range of k independent $N(0,1)$ variables. Thus, the large-sample approximate simultaneous $100(1-\alpha)\%$ confidence intervals for C (6.89) are given by (6.93) with w^*_α replaced by q_α in the expressions for a_{uv} (6.91) and b_{uv} (6.92). To find q_α for k treatments, we use the R command `cRangekNorm(`α`, k)`. For example, to find $q_{.05}$ for $k = 6$ treatments, we apply `cRangekNorm(.05, 6)` and obtain $q_{.05} = 4.30$.

EXAMPLE 6.10 ***Motivational Effect of Knowledge of Performance—Examples 6.2, 6.8, and 6.9 Continued.***

Consider the Hundal knowledge of performance data originally presented in Example 6.2. In this example, we wish to find simultaneous $100(1-\alpha)\%$ confidence intervals for the $3(2)/2 = 3$ simple contrasts

$$C = \{\tau_2 - \tau_1, \tau_3 - \tau_1, \tau_3 - \tau_2\}.$$

For the sake of illustration, we take $\alpha = .1041$. Using the R program `cSDCFlig(`α`, n)` with $k = 3$ and $n_1 = n_2 = n_3 = 6$, we have $w^*_{.1041} =$ `cSDCFlig(.1041,c(6, 6, 6))` $=$ 2.9439. It follows from expressions (6.91) and (6.92) that

$$a_{12} = a_{13} = a_{23} = \frac{6(6)}{2} - 2.944\left[\frac{6(6)(6+6+1)}{24}\right]^{1/2} + 1 \approx 6.00$$

and

$$b_{12} = b_{13} = b_{23} = 6.00 - 1 = 5.00.$$

Thus, the simultaneous 89.51% confidence intervals for the simple contrasts $\Delta_{12} = \tau_2 - \tau_1, \Delta_{13} = \tau_3 - \tau_1$, and $\Delta_{23} = \tau_3 - \tau_2$ correspond to $[D^{12}_{(\langle 6\rangle)}, D^{12}_{(36-\langle 5\rangle)}) = [D^{12}_{(6)}, D^{12}_{(31)}), [D^{13}_{(6)}, D^{13}_{(31)})$, and $[D^{23}_{(6)}, D^{23}_{(31)})$, respectively. Using the individual differences already computed in Example 6.9 to obtain a point estimate of the contrast $\tau_3 - \tau_1 = \tau_{\text{accurate information}} - \tau_{\text{no information}}$, we see that the three sets of $n_u n_v = 36$ ordered $D^{uv}_{(t)}$'s are given by

$$D^{12}_{(t)} : \{-6, -5, -4, -4, -3, -3, -3, -2, -2, -1, -1, -1, 0, 0, 0, 0, 1, 1,$$

$$2, 2, 2, 3, 3, 3, 4, 4, 4, 5, 5, 6, 6, 7, 7, 9, 9, 12\},$$

$$D^{13}_{(t)} : \{-4, -3, -1, -1, 0, 0, 0, 1, 1, 2, 2, 2, 2, 3, 3, 3, 4, 4,$$

$$4, 5, 5, 5, 5, 5, 6, 6, 7, 7, 8, 8, 8, 9, 10, 10, 11, 13\},$$

and

$$D^{23}_{(t)} : \{-7, -4, -4, -3, -2, -2, -1, -1, 0, 0, 0, 1, 1, 1, 2, 2, 2, 3,$$

$$3, 3, 4, 4, 4, 4, 5, 5, 5, 6, 6, 6, 7, 8, 8, 8, 10\}.$$

Hence, the simultaneous 89.51% confidence intervals for the simple contrasts $\Delta_{12} = \tau_2 - \tau_1, \Delta_{13} = \tau_3 - \tau_1$, and $\Delta_{23} = \tau_3 - \tau_2$ for the Hundal data are

$$[D^{12}_{(6)}, D^{12}_{(31)}) = [-3, 6),$$

$$[D^{13}_{(6)}, D^{13}_{(31)}) = [0, 8),$$

and

$$[D^{23}_{(6)}, D^{23}_{(31)}) = [-2, 6),$$

respectively.

For the sake of illustration for the associated large-sample approximation, we take an approximate α value of .10. With $k = 3$, we find $q_{.10} = $ `cRangeKNorm(.10, 3)` $= 2.902$. The associated approximate values for the a_{uv}'s and b_{uv}'s are

$$a_{12} = a_{13} = a_{23} \approx \frac{6(6)}{2} - 2.902\left[\frac{6(6)(6+6+1)}{24}\right]^{1/2} + 1 = 6.19$$

and

$$b_{12} = b_{13} = b_{23} \approx 6.19 - 1 = 5.19.$$

As $\langle 6.19\rangle = 6$ and $\langle 5.19\rangle = 5$, we see that the approximate 90% simultaneous confidence intervals for the simple contrasts $\tau_2 - \tau_1, \tau_3 - \tau_1$, and $\tau_3 - \tau_2$ are identical with the exact 89.51% simultaneous confidence intervals for these Hundal data. This provides some indication that the common sample size of six observations is already large enough to

enable the large-sample approximation to be effective for these simultaneous confidence intervals.

Comments

78. *Relationship of Simultaneous Confidence Intervals to Two-Sided All-Treatments Multiple Comparisons*. The simultaneous $100(1-\alpha)\%$ confidence intervals (6.93) for the collection C (6.89) of all simple contrasts are directly related to the Dwass–Steel–Critchlow–Fligner two-sided all-treatments multiple comparison procedure (6.62) discussed in Section 6.5. In fact, for *every* (u, v) pair, $1 \le u < v \le k$, the two-sided multiple comparison procedure (6.62) yields the decision $\tau_u \ne \tau_v$ at an experimentwise error rate α if and only if 0 does not belong to the corresponding simultaneous $100(1-\alpha)\%$ confidence interval $[D^{uv}_{(\langle a_{uv}\rangle)}, D^{uv}_{(n_u n_v - \langle b_{uv}\rangle)})$ for $\tau_v - \tau_u$. Thus, each of the $2^{\binom{k}{2}}$ sets of possible multiple comparison decisions associated with procedure (6.62) at an experimentwise error rate α corresponds to a collection of simultaneous $100(1-\alpha)\%$ confidence intervals (6.93) for C (6.89) for which the particular (u, v) intervals not containing the value 0 match exactly with those treatment pairs for which procedure (6.62) leads to the decision $\tau_u \ne \tau_v$.

79. *Simultaneous* $100(1-\alpha)\%$ *Lower Confidence Bounds*. In situations where an order restriction $\tau_1 \le \tau_2 \le \cdots \le \tau_k$ on the treatment effects is appropriate (see Sections 6.2 and 6.6 for further details), it is more natural to seek out simultaneous $100(1-\alpha)\%$ lower confidence bounds (rather than two-sided intervals) for the collection C (6.89) of simple contrasts. In such a setting, let c^*_α be the critical value for the Hayter–Stone one-sided all-treatments multiple comparison procedure (6.68) and set

$$h_{uv} = \frac{n_u n_v}{2} - c^*_\alpha \left[\frac{n_u n_v (n_u + n_v + 1)}{24}\right]^{1/2} + 1. \tag{6.95}$$

The simultaneous $100(1-\alpha)\%$ lower confidence bounds for the collection C (6.89) suggested by Hayter and Stone (1991) are then given by

$$\left\{[D^{uv}_{(\langle h_{uv}\rangle)}, \infty), 1 \le u < v \le k\right\}, \tag{6.96}$$

where, once again, $\langle t\rangle$ denotes the greatest integer less than or equal to t and the ordered $D^{uv}_{(t)}$'s are as defined in the Procedure of this section. When either the number of treatments exceeds 3 or $k = 3$ and one or more of the sample sizes is larger than 7, Hayter and Stone (1991) suggest approximating c^*_α in expression (6.95) by d_α, the upper αth percentile point for the distribution of

$$D = \underset{1 \le i < j \le k}{\text{maximum}} \left[\frac{Z_j - Z_i}{\left\{\frac{n_i + n_j}{2 n_i n_j}\right\}^{1/2}}\right],$$

where $Z_1, \ldots, Z_k$ are mutually independent and Z_i has an $N(0, 1/n_i)$ distribution, for $i = 1, \ldots, k$. To find d_α for k treatments, we use the R command

`cHayStonLSA(`α, k`)`. For example, to find $d_{.05}$ for $k = 6$ treatments, we apply `cHayStonLSA(.05, 6)` and obtain $d_{.05} = 3.725$.

The relationship between these simultaneous $100(1-\alpha)\%$ lower confidence bounds (6.96) for C (6.89) and the Hayter–Stone one-sided all-treatments multiple comparison procedure (6.68) at experimentwise error rate α is identical to that described in Comment 78 for the simultaneous $100(1-\alpha)\%$ confidence intervals (6.93) for C (6.89) and the two-sided all-treatments multiple comparison procedure (6.62) at experimentwise error rate α.

80. *Pairwise versus Joint Rankings*. In the latter portion of Comment 59, we discussed some of the pros and cons of pairwise rankings versus joint rankings in the one-way layout setting. The simultaneous $100(1-\alpha)\%$ confidence intervals (6.93) for the collection of all simple contrasts (and the analogous simultaneous lower confidence bounds (6.96) discussed in Comment 79) are clearly associated with pairwise rankings. This provides an additional advantage to the use of pairwise rankings, as the joint ranking approach discussed in Comment 59 does not lead directly to such simultaneous confidence intervals or bounds for C (6.89).

Properties

1. *Distribution-Freeness*. For populations satisfying Assumptions A1–A3, (6.94) holds. Hence, we can control the simultaneous coverage probability to be $1-\alpha$ without having more specific knowledge about the form of the underlying F. As a result, the intervals in (6.93) are distribution-free simultaneous confidence intervals for the collection C (6.89) of all simple contrasts over a very large class of populations.
2. *Asymptotic Multivariate Normality*. See Hayter (1984) and Critchlow and Fligner (1991).

Problems

81. Consider the length of YOY gizzard shad data discussed in Problem 4. Find a set of approximate simultaneous 95% confidence intervals for the set of all simple contrasts.

82. Consider the Hundal knowledge of performance data originally discussed in Example 6.2. Find a set of simultaneous 88.11% lower confidence bounds for the three simple contrasts $\tau_2 - \tau_1, \tau_3 - \tau_1$, and $\tau_3 - \tau_2$ (see Comment 79). Compare with the set of 89.59% simultaneous confidence intervals obtained in Example 6.10 for these same simple contrasts.

83. Consider the Acid Red 114 revertant colonies data in Table 6.10. Find a set of approximate simultaneous 90% confidence intervals for the set of all simple contrasts.

84. Consider the tiger muskellunge plasma glucose data in Table 6.9. Find a set of approximate simultaneous 95% confidence intervals for the set of all simple contrasts.

85. Consider the white-tailed deer fasting metabolic rate data in Table 6.8. Find a set of approximate simultaneous 80% confidence intervals for the set of all simple contrasts.

86. Consider the half-time of mucociliary clearance data in Table 6.1. Find a set of approximate simultaneous 91.81% confidence intervals for the set of all simple contrasts. Without further calculations, what decisions would be reached for these data by the multiple comparison procedure (6.62) at experimentwise error rate $\alpha = .0819$? (See Comment 78.)

87. Consider the average basal area increment data in Table 6.7. Find a set of approximate simultaneous 90% confidence intervals for the set of all simple contrasts. Do you have any concerns about the application of this procedure to these data?

88. Consider the Wechsler Adult Intelligence Scale data in Table 6.11. For the age groups 16–19, 20–34, 35–54, and 55–69 years only, find a set of approximate simultaneous 90% lower confidence bounds for the set of all simple contrasts for these four age groups. Without further calculations, what decisions would be reached for these data by the multiple comparison procedure (6.70) at approximate experimentwise error rate $\alpha = .10$? (See Comments 78 and 79.)

6.10 EFFICIENCIES OF ONE-WAY LAYOUT PROCEDURES

The Pitman asymptotic relative efficiencies for translation alternatives of most of the nonparametric procedures discussed in this chapter with respect to the corresponding normal theory procedures are given by the expression

$$e_F = 12\sigma_F^2 \left\{ \int_{-\infty}^{\infty} f^2(u)du \right\}^2, \tag{6.97}$$

where σ_F^2 is the variance of the common underlying (continuous) distribution F (6.1) and $f(\cdot)$ is the probability density function corresponding to F. The parameter $\int_{-\infty}^{\infty} f^2(u)\, du$ is the area under the curve associated with $f^2(\cdot)$, the square of the common probability density function. We note that this same expression (6.97) also yields the corresponding Pitman efficiencies in the one-sample and two-sample location settings (see Sections 3.11 and 4.5).

In particular, the Pitman asymptotic relative efficiency of the Kruskal–Wallis test based on H (6.5) with respect to the normal theory one-way layout $\mathcal{F}$-test was found to be e_F (6.97) by Andrews (1954). The asymptotic relative efficiency of the Jonckheere–Terpstra test for ordered alternatives based on the statistic J (6.13) with respect to a suitable normal theory competitor was found by Puri (1965) to be e_F (6.97) as well. Mack and Wolfe (1981) found the same expression to hold for the asymptotic relative efficiency of their peak-known umbrella test procedure based on A_p (6.31) relative to an analogous normal theory procedure based on sample averages. Fligner and Wolfe (1982) found the same to be case for the treatments-versus-control test based on FW (6.50).

Sherman (1965) obtained e_F (6.97) as the asymptotic relative efficiency of the two-sided all-treatments and the one-sided treatments-versus-control multiple comparison procedures discussed in Sections 6.5 and 6.7 with respect to the corresponding classical normal theory procedures based on sample means. Spjøtvoll (1968) showed that, when n_j/N tends to ρ_j, with $0 < \rho_j < 1$, the estimators W_{hj} (6.85) have the same asymptotic properties as the estimators $[Z_{h\cdot} - Z_{j\cdot}]$ (see Comment 74). It then follows from Lehmann's (1963a) results that e_F (6.97) is the asymptotic relative efficiency of the estimator $\widehat{\theta}$ (6.83) with respect to the least squares estimator $\overline{\theta} = \sum_{h=1}^{k} \sum_{j=1}^{k} d_{hj}(\overline{X}_{\cdot h} - \overline{X}_{\cdot j})$, where

$$\overline{X}_{\cdot t} = \sum_{i=1}^{n_t} \frac{X_{it}}{n_t}, \quad \text{for } t = 1, \ldots, k.$$

As noted in both Sections 3.11 and 4.5, the asymptotic relative efficiency e_F (6.97) is always greater than or equal to .864 and can be infinite. See expression (3.116) for the value of e_F (6.97) for a variety of underlying F populations.

We do not know of any results for the asymptotic relative efficiencies of the Mack–Wolfe peak-unknown umbrella test (Section 6.3B), the Hayter–Stone one-sided all-treatments multiple comparison procedure (Section 6.6), or the Critchlow–Fligner procedure for simultaneous confidence intervals for all simple contrasts (Section 6.9).