CHAPTER 8

Comparing Multiple Dependent Groups

This chapter covers basic methods for comparing dependent groups, including both a between-by-within and a within-by-within design. Three-way designs are covered as well where one or more factors involve dependent groups.

As noted in Chapter 5, when comparing dependent groups based on some measure of location, there are three general approaches that might be used. The first is to compare measures of location associated with the marginal distributions. The second is to make inferences based on a measure of location associated with the difference scores. And the third focuses on measures of location associated with the distribution of the difference between two dependent random variables. When comparing means, it makes no difference which view is adopted, but when using robust measures of location, this is no longer the case. Methods relevant to all three approaches are described and comments on their relative merits are provided.

Note that when comparing measures of location associated with the marginal distributions, there are two types of estimators that might be used. The first estimates a measure of location for each marginal distribution, ignoring the other variables under study. That is, for p-variate data X_{ij} ($i = 1, \ldots, n$; $j = 1, \ldots, p$), compute the trimmed mean or some other measure of location using the n values associated with each j. This is in contrast to using a location estimator that takes into account the overall structure of the data when dealing with outliers, such as the OP-estimator in Section 6.5. The bulk of the methods in this chapter are based on the former type of estimator. A multiple comparison procedure that deals with the latter type of estimator is described at the end of Section 8.2.7.

8.1 Comparing Trimmed Means

This section focuses on nonbootstrap methods for testing hypotheses about trimmed means. Methods that use other estimators based on the marginal distributions, such as robust M-estimators, are described in Section 8.2.

Introduction to Robust Estimation and Hypothesis Testing. DOI: 10.1016/B978-0-12-386983-8.00008-1

8.1.1 Omnibus Test Based on the Trimmed Means of the Marginal Distributions

For J dependent groups, let μ_{tj} be the population trimmed mean associated with the jth group. That is, μ_{tj} is the trimmed mean associated with the jth marginal distribution. The goal in this section is to test

$$H_0 : \mu_{t1} = \cdots = \mu_{tJ},$$

the hypothesis that the trimmed means of J dependent groups are equal. The method used here is based on a generalization of the Huynh–Feldt method for means which is designed to handle violations of the sphericity assumption associated with the standard F-test. (See Kirk, 1995, for details about sphericity. For simulation results on how the test for trimmed means performs, see Wilcox, 1993c.) The method begins by Winsorizing the values in essentially the same manner described in Section 5.9.3. That is, fix j, let $X_{(1)j} \leq X_{(2)j} \leq \cdots \leq X_{(n)j}$ be the n values in the jth group written in ascending order, and let

$$Y_{ij} = \begin{cases} X_{(g+1)j} & \text{if } X_{ij} \leq X_{(g+1)j} \\ X_{ij} & \text{if } X_{(g+1)j} < X_{ij} < X_{(n-g)j} \\ X_{(n-g)j} & \text{if } X_{ij} \geq X_{(n-g)j}, \end{cases}$$

where g is the number of observations trimmed or Winsorized from each end of the distribution corresponding to the jth group. The test statistic, F, is computed as described in Table 8.1, and Table 8.2 describes how to compute the degrees of freedom.

8.1.2 R Function rmanova

The R function

rmanova(x,tr=.2,grp=c(1:length(x))

tests the hypothesis of equal population trimmed means among J dependent groups using the calculations in Tables 8.1 and 8.2. The data are stored in any variable x, which can be either an n-by-J matrix, the jth column containing the data for the jth group, or an R variable having list mode. In the latter case, x[[1]] contains the data for group 1, x[[2]] contains the data for group 2, and so on. As usual, tr indicates the amount of trimming which defaults to 0.2, and grp can be used to compare a subset of the groups. If the argument grp is not specified, the trimmed means of all J groups are compared. If, for example, there are five groups, but the goal is to test $H_0 : \mu_{t2} = \mu_{t4} = \mu_{t5}$, the command rmanova(x,grp=c(2,4,5)) accomplishes this goal using 20% trimming.

Example

Section 8.6.2 reports measures of hangover symptoms for participants belonging to one of two groups, with each participant consuming alcohol on three different occasions.

Table 8.1: Test Statistic for Comparing the Trimmed Means of Dependent Groups.

Winsorize the observations in the jth group, as described in this section, yielding Y_{ij}. Let $h = n - 2g$ be the effective sample size, where $g = [\gamma n]$, and γ is the amount of trimming. Compute

$$\bar{X}_t = \frac{1}{J}\sum \bar{X}_{tj}$$

$$Q_c = (n-2g)\sum_{j=1}^{J}(\bar{X}_{tj} - \bar{X}_t)^2$$

$$Q_e = \sum_{j=1}^{J}\sum_{i=1}^{n}(Y_{ij} - \bar{Y}_{.j} - \bar{Y}_{i.} + \bar{Y}_{..})^2,$$

where

$$\bar{Y}_{.j} = \frac{1}{n}\sum_{i=1}^{n} Y_{ij}$$

$$\bar{Y}_{i.} = \frac{1}{J}\sum_{j=1}^{J} Y_{ij}$$

$$\bar{Y}_{..} = \frac{1}{nJ}\sum_{j=1}^{J}\sum_{i=1}^{n} Y_{ij}.$$

The test statistic is

$$F = \frac{R_c}{R_e},$$

where

$$R_c = \frac{Q_c}{J-1}$$

$$R_e = \frac{Q_e}{(h-1)(J-1)}.$$

For present purposes, focus on group 1 (the control group) with the goal of comparing the responses on the three different occasions. The function rmanova reports a p-value of .09.

■

8.1.3 Pairwise Comparisons and Linear Contrasts Based on Trimmed Means

Suppose that for J dependent groups, it is desired to compute a $1-\alpha$ confidence interval for

$$\mu_{tj} - \mu_{tk},$$

Table 8.2: How to Compute Degrees of Freedom when Comparing Trimmed Means?

Let

$$v_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_{ij} - \bar{Y}_{.j})(Y_{ik} - \bar{Y}_{.k})$$

for $j = 1, \ldots, J$ and $k = 1, \ldots, J$, where Y_{ij} is the Winsorized observation corresponding X_{ij}. When $j = k$, $v_{jk} = s_{wj}^2$, the Winsorized sample variance for the jth group, and when $j \neq k$, v_{jk} is a Winsorized analog of the sample covariance.

Let

$$\bar{v}_{..} = \frac{1}{J^2}\sum_{j=1}^{J}\sum_{k=1}^{J} v_{jk}$$

$$\bar{v}_{d} = \frac{1}{J}\sum_{j=1}^{J} v_{jj}$$

$$\bar{v}_{j.} = \frac{1}{J}\sum_{k=1}^{J} v_{jk}$$

$$A = \frac{J^2(\bar{v}_{d} - \bar{v}_{..})^2}{J-1}$$

$$B = \sum_{j=1}^{J}\sum_{k=1}^{J} v_{jk}^2 - 2J\sum_{j=1}^{J}\bar{v}_{j.}^2 + J^2\bar{v}_{..}^2$$

$$\hat{\epsilon} = \frac{A}{B}$$

$$\tilde{\epsilon} = \frac{n(J-1)\hat{\epsilon} - 2}{(J-1)[n - 1 - (J-1)\hat{\epsilon}]}.$$

The degrees of freedom are

$$\nu_1 = (J-1)\tilde{\epsilon}$$

$$\nu_2 = (J-1)(h-1)\tilde{\epsilon},$$

where h is the effective sample size for each group.

for all $j < k$. That is, the goal is to compare all pairs of trimmed means. One possibility is to compare the jth trimmed mean to the kth trimmed mean using the R function yuend in Chapter 5, and control the familywise error (FWE) rate, (the probability of at least one type I error) with the Bonferroni inequality. That is, if C tests are to be performed, perform each test at the α/C level. A practical concern with this approach is that the actual probability of at least one type I error can be considerably less than the nominal level. For example, if $J = 4$, $\alpha = 0.05$, and sampling is from independent normal distributions, the actual probability of at least one type I error is approximately .019 when comparing 20% trimmed means with

$n = 15$. If each pair of random variables has correlation 0.1, the probability of at least one type I error drops to .014, and it drops even more as the correlations are increased. Part of the problem is that the individual tests for equal trimmed means tends to have type I error probabilities less than the nominal level, so performing each test at the α/C level makes matters worse. In fact, even when sampling from heavy-tailed distributions, power can be low compared to using means, even though the sample mean has a much larger standard error (Wilcox, 1997a). One way of improving on this approach is to use results in Rom (1990) to control FWE.

Momentarily consider a single linear contrast

$$\Psi = \sum_{j=1}^{J} c_j \mu_j,$$

where $\sum c_j = 0$ and the goal is to test

$$H_0 : \Psi = 0.$$

Let Y_{ij} $(i = 1, \ldots, n;\ j = 1, \ldots, J)$ be the Winsorized values which are computed as described in Section 8.1.1. Let

$$A = \sum_{j=1}^{J} \sum_{k=1}^{J} c_j c_k d_{jk},$$

where

$$d_{jk} = \frac{1}{h(h-1)} \sum_{i=1}^{n} (Y_{ij} - \bar{Y}_j)(Y_{ik} - \bar{Y}_k),$$

and $h = n - 2g$ is the number of observations left in each group after trimming. Let

$$\hat{\Psi} = \sum_{j=1}^{J} c_j \bar{X}_{tj}.$$

The test statistic is

$$T = \frac{\hat{\Psi}}{\sqrt{A}}$$

and the null hypothesis is rejected if $|T| \geq t$, where t is the $1 - \alpha/2$ quantile of a Student's t-distribution with $\nu = h - 1$ degrees of freedom.

When testing C hypotheses, the following method, motivated by results in Rom (1990), appears to be relatively effective at controlling FWE. Let p_k be the p-value associated with the kth hypothesis and put these C p-values in descending order yielding $p_{[1]} \geq \cdots \geq p_{[C]}$. Then

Table 8.3: Critical Values, d_k, for Rom's Method.

k	$\alpha = 0.05$	$\alpha = 0.01$
1	0.05000	0.01000
2	0.02500	0.00500
3	0.01690	0.00334
4	0.01270	0.00251
5	0.01020	0.00201
6	0.00851	0.00167
7	0.00730	0.00143
8	0.00639	0.00126
9	0.00568	0.00112
10	0.00511	0.00101

1. Set k=1.
2. If $p_{[k]} \leq d_k$, where d_k is read from Table 8.3, stop and reject all C hypotheses; otherwise, go to step 3. (When $k > 10$, then $d_k = \alpha/k$.)
3. Increment k by 1. If $p_{[k]} \leq d_k$, stop and reject all hypotheses having p-values less than or equal to d_k
4. If $P_{[k]} > d_k$, repeat step 3.
5. Continue until you reject or all C hypotheses have been tested.

Note that Table 8.3 is limited to $k \leq 10$. If $k > 10$, here FWE is controlled with Hochberg's (1988) method. That is, proceed as just indicated, but rather than use d_k read from Table 8.3, use $d_k = \alpha/k$.

8.1.4 Linear Contrasts Based on the Marginal Random Variables

The method just described is readily extended to a situation that contains comparisons based on difference scores as a special case. Let

$$D_{ik} = \sum_{j=1}^{J} c_{jk} X_{ij},$$

where for any k ($k = 1, \ldots, C$), $\sum c_{jk} = 0$, and let μ_{tk} be the population trimmed mean of the distribution from which the random sample $D_{1k}, \ldots, D_{nk}$ was obtained. For example, if $c_{11} = 1$, $c_{21} = -1$, and $c_{31} = \cdots = c_{J1} = 0$, then

$$D_{i1} = X_{i1} - X_{i2},$$

the difference scores for groups 1 and 2, and μ_{t1} is the (population) trimmed mean associated with this difference. Similarly, if $c_{22} = 1$, $c_{32} = -1$, and $c_{12} = c_{41} = \cdots = c_{J1} = 0$, then

$$D_{i2} = X_{i2} - X_{i3}$$

and μ_{t2} is the corresponding (population) trimmed mean. The goal is to test

$$H_0 : \mu_{tk} = 0$$

for each $k = 1, \ldots, C$ such that FWE is approximately α. Each hypothesis can be tested using results in Chapter 4, but there is the added goal of controlling FWE. Here, Rom's method, described in Section 8.1.3, is used to accomplish this goal.

It should be noted that the multiple comparison procedures in this chapter are designed to control the probability of one or more type I errors. As was the case in Chapter 7, the expectation is that the actual probability of one more type I error will be reduced if the multiple comparison procedures in this chapter are used contingent on a global test rejecting at the α level. That is, power might be adversely affected (cf. Bernhardson, 1975).

Section 5.3.4 described ξ, a robust, heteroscedastic measure of effect size based on the notion of explanatory power. One way of characterizing the difference between two dependent groups is to again use this measure of effect size, which can be done for all pairs of groups via the R function esmcp in Section 7.1.2.

8.1.5 R Function rmmcp and rmmismcp

The R function

$$\text{rmmcp(x,con} = 0, \text{ tr} = 0.2, \text{ alpha} = 0.05, \text{ dif=T)}$$

performs multiple comparisons among dependent groups using trimmed means and Rom's method for controlling FWE. By default, difference scores are used. Setting dif=F results in comparing the marginal trimmed means. When α differs from both 0.05 and 0.01, FWE is controlled with Hochberg's (1988) method. That is, proceed as indicated in Section 8.1.3 but rather than use d_k from Table 8.3, use $d_k = \alpha/k$.

When there are values missing at random, method M2 in Section 5.9.13 can be used to perform multiple comparisons via the R function

$$\text{rmmismcp(x,y} = \text{NA, alpha} = 0.05, \text{ con} = 0, \text{ est} = \text{tmean, plotit} = \text{T, grp} = \text{NA, nboot} = 500, \text{ SEED} = \text{T, xlab} = \text{“Group 1”, ylab} = \text{“Group 2”, pr} = \text{F, } \ldots),$$

which was introduced in Section 5.9.14 and controls the probability of one or more type I errors using Hochberg's method. By default, 20% trimmed means are used, but other robust estimators can be used via the argument est.

8.1.6 Judging the Sample Size

Let $D_{ijk} = X_{ij} - X_{ik}$ and let μ_{tjk} be a trimmed mean corresponding to D_{ijk}. If when testing $H_0 : \mu_{tjk} = 0$ for any $j < k$, a non-significant result is obtained, this might be because the null hypothesis is true, or of course, a type II error might have been committed due to a sample size that is too small. To help determine whether the latter explanation is reasonable, an extension of Stein's (1945) two-stage method for means might be used. Suppose it is desired to have all-pairs power greater than or equal to $1 - \beta$ when for any $j < k$, $\mu_{tjk} = \delta$. That is, the probability of rejecting H_0 for all $j < k$ for which $\mu_{tjk} = \delta$ is to be at least $1 - \beta$. The goal here is to determine whether the sample size used, namely n, is large enough to accomplish this goal, and if not, the goal is to determine how many more observations are needed. The following method performs well in simulations (Wilcox, 2004b).

Let $C = (J^2 - J)/2$ and

$$d = \left(\frac{\delta}{t_\beta - t_{1-\alpha/(2C)}} \right)^2,$$

where t_β is the β quantile of Student's t distribution with $\nu = n - 2g - 1$ degrees of freedom, and g is the number of observations trimmed from each tail. (So $n - 2g$ is the number of observations not trimmed.) Let

$$N_{jk} = \max(n, \left[\frac{s^2_{wjk}}{(1-2\gamma)^2 d} \right] + 1)$$

where s^2_{wjk} is the Winsorized variance of the D_{ijk} values. Then the required sample size in the second stage is

$$N = \max N_{jk},$$

the maximum being taken over all $j < k$. So if $N = n$, the sample size used is judged to be adequate for the specified power requirement.

In the event the additional $N - n$ vectors of observations can be obtained, familiarity with Stein's (1945) original method suggests how H_0 should be tested, but in simulations, a slight modification performs a bit better in terms of power. Let S_{wjk} be the Winsorized variance based on all N of the observations, where the amount of Winsorizing is equal to the amount of trimming. Let $\hat{\mu}_{tjk}$ be the trimmed mean based on all N D_{ijk} differences and let

$$T_{jk} = \frac{\sqrt{N}(1-2\gamma)\hat{\mu}_{tjk}}{S_{wjk}}.$$

Then reject $H_0 : \mu_{tjk} = 0$ if $|T_{jk}| \geq t_{1-\alpha/(2C)}$. So as would be expected based on Stein's method, the degrees of freedom depend on the initial sample size, n, not the ultimate sample

size, N. But contrary to what is expected based on Stein's method, the Winsorized variance when computing T_{jk} is based on all N observations. (All indications are that no adjustment for β is needed when computing d when multiple tests are performed and the goal is to have all-pairs power greater than or equal to $1 - \beta$. Also, a variation of the method aimed at comparing the marginal trimmed means has not been investigated.)

8.1.7 R Functions stein1.tr and stein2.tr

Using the method just described, the R function

```
stein1.tr(x,del,alpha=0.05,pow=0.8,tr=0.2)
```

determines the required sample size needed to achieve all-pairs power equal to the value indicated by the argument pow for a difference specified by the argument del which corresponds to δ. In the event additional data are needed to achieve the desired amount of power, and if these additional observations can be acquired,

```
stein2.tr(x,y,alpha=0.05,tr=0.2)
```

tests all pairwise differences. Here the first-stage data are stored in x (which is a vector or a matrix with J columns) and y contains the second-stage data.

8.2 Bootstrap Methods Based on Marginal Distributions

This section focuses on bootstrap methods aimed at making inferences about measures of location associated with the marginal distributions. (Section 8.3 takes up measures of location associated with difference scores.) As in previous chapters, two general types of bootstrap methods appear to deserve serious consideration in applied work. (As usual, this is not intended to suggest that all other variations of the bootstrap have no practical value for the problems considered here, only that based on extant studies, the methods covered here seem to perform relatively well.) The first type uses estimated standard errors and reflects extensions of the bootstrap-t methods in Chapter 5; they are useful when comparing trimmed means. The other is an extension of the percentile bootstrap method where estimated standard errors do not play a direct role. When comparing robust M-measures of location, this latter approach is the only known way of controlling the probability of a type I error for a fairly wide range of distributions.

8.2.1 Comparing Trimmed Means

Let μ_{tj} be the population trimmed mean associated with the jth marginal distribution and consider the goal of testing

$$H_0 : \mu_{t1} = \cdots = \mu_{tJ},$$

An extension of the bootstrap-t method to this problem is straightforward. Set

$$C_{ij} = X_{ij} - \bar{X}_{tj}$$

with the goal of estimating an appropriate critical value, based on the test statistic F in Table 8.1, when the null hypothesis is true. The remaining steps are as follows:

1. Generate a bootstrap sample by randomly sampling, with replacement, n rows of data from the matrix

$$\begin{pmatrix} C_{11}, \ldots, C_{1J} \\ \vdots \\ C_{n1}, \ldots, C_{nJ} \end{pmatrix}$$

   yielding

$$\begin{pmatrix} C^*_{11}, \ldots, C^*_{1J} \\ \vdots \\ C^*_{n1}, \ldots, C^*_{nJ} \end{pmatrix}.$$

2. Compute the test statistic F in Table 8.1 based on the C^*_{ij} values generated in step 1, and label the result F^*.
3. Repeat steps 1 and 2 B times and label the results $F^*_1, \ldots, F^*_B$.
4. Put these B values in ascending order and label the results $F^*_{(1)} \leq \cdots \leq F^*_{(B)}$.

The critical value is estimated to be $F^*_{(u)}$, where $u = (1-\alpha)B$ rounded to the nearest integer. That is, reject the hypothesis of equal trimmed means if

$$F \geq F^*_{(u)},$$

where F is the statistic given in Table 8.1 based on the X_{ij} values.

8.2.2 R Function rmanovab

The R function

$$\text{rmanovab(x, tr} = 0.2, \text{alpha} = 0.05, \text{grp} = 0, \text{nboot} = 599)$$

performs the bootstrap-t method just described.

8.2.3 Multiple Comparisons Based on Trimmed Means

This section describes bootstrap methods for performing multiple comparisons based on trimmed means. First consider the goal of performing all pairwise comparisons. That is, the goal is to test

$$H_0 : \mu_{tj} = \mu_{tk}$$

for all $j < k$. A bootstrap-t method is applied as follows. Generate bootstrap samples as was done in Section 8.2.1 yielding

$$\begin{pmatrix} C_{11}^*, \ldots, C_{1J}^* \\ \vdots \\ C_{n1}^*, \ldots, C_{nJ}^* \end{pmatrix}.$$

For every $j < k$, compute the test statistic T_y, given by Eq. (5.24), using the values in the jth and kth columns of the matrix just computed. That is, perform the test for trimmed means corresponding to two dependent groups using the data $C_{1j}^*, \ldots, C_{nj}^*$ and $C_{1k}^*, \ldots, C_{nk}^*$. Label the resulting test statistic T_{yjk}^*. Repeat this process B times yielding $T_{yjk1}^*, \ldots, T_{yjkB}^*$. Because these test statistics are based on data generated from a distribution for which the trimmed means are equal, they can be used to estimate an appropriate critical value. In particular, for each b, set

$$T_b^* = \max|T_{yjkb}^*|,$$

the maximum being taken over all $j < k$. Let $T_{(1)}^* \leq \cdots \leq T_{(B)}^*$ be the T_b^* values written in ascending order and let $u = (1-\alpha)B$, rounded to the nearest integer. Then $H_0 : \mu_{tj} = \mu_{tk}$ is rejected if $T_{yjk} > T_{(u)}^*$. That is, for the jth and kth groups, test the hypothesis of equal trimmed means using the method in Section 5.9.5, only the critical value is $T_{(u)}^*$, which was determined so that the probability of a least one type I error is approximately equal to α. Alternatively, the confidence interval for $\mu_{tj} - \mu_{tk}$ is

$$(\bar{X}_{tj} - \bar{X}_{tk}) \pm T_{(u)}^* \sqrt{d_j + d_k - 2d_{jk}},$$

where $\sqrt{d_j + d_k - 2d_{jk}}$ is the estimate of the standard error of $\bar{X}_{tj} - \bar{X}_{tk}$, which is computed as described in Section 5.9.5. The simultaneous probability coverage is approximately $1-\alpha$. Probability coverage appears to be reasonably good with n as small as 15 when using 20% trimming with $J = 4$, $\alpha = 0.05$, $B = 599$ (Wilcox, 1997a). When there is no trimming, probability coverage can be poor, and no method can be recommended. Also, the power of the bootstrap method, with 20% trimmed means, compares well to an approach based on means and the Bonferroni inequality.

The method is easily extended to situations where the goal is to test C linear contrasts, $\Psi_1, \ldots, \Psi_C$, where

$$\Psi_k = \sum c_{jk}\mu_{tj},$$

and c_{jk} ($j = 1, \ldots, J$ and $k = 1, \ldots, C$) are constants chosen to reflect some hypothesis of interest. As before, Ψ_k is estimated with $\hat{\Psi}_k = \sum c_{jk}\bar{X}_{tj}$, but now the squared standard error is estimated with

$$A_k = \sum_{j=1}^{J}\sum_{\ell=1}^{J} c_{jk}c_{\ell k}d_{j\ell},$$

where

$$d_{jk} = \frac{1}{h(h-1)} \sum (Y_{ij} - \bar{Y}_j)(Y_{ik} - \bar{Y}_k),$$

and Y_{ij} are the Winsorized observations for the jth group. (When $j = k$, $d_{jk} = d_j^2$.)

To compute a $1 - \alpha$ confidence interval for Ψ_k, generate a bootstrap sample yielding C_{ij}^* and let

$$T_{yk}^* = \frac{\hat{\Psi}_k^*}{\sqrt{A_k^*}},$$

where $\hat{\Psi}_k^*$ and A_k^* are computed with the bootstrap observations. Repeat this bootstrap process B times yielding T_{ykb}^*, $b = 1, \ldots B$. For each b, let $T_b^* = \max|T_{ykb}^*|$, the maximum being taken over $k = 1, \ldots, C$. Put the T_b^* values in order yielding $T_{(1)}^* \leq \cdots \leq T_{(B)}^*$, in which case an appropriate critical value is estimated to be $T_{(u)}^*$, where $u = (1 - \alpha)B$, rounded to the nearest integer. Then an approximate $1 - \alpha$ confidence interval for Ψ_k is

$$\hat{\Psi}_k \pm T_{(u)}^* \sqrt{A_k}.$$

8.2.4 R Functions pairdepb and bptd

The R function

pairdepb(x,tr=0.2,alpha=0.05,grp=0,nboot=599)

performs all pairwise comparisons among J dependent groups using the bootstrap method just described. The argument x can be an n-by-J matrix of data, or it can be an R variable having list mode. In the latter case, x[[1]] contains the data for group 1, x[[2]] contains the data for group 2, and so on. The argument tr indicates the amount of trimming, which, if unspecified, defaults to 0.2. The value for α defaults to alpha=0.05, and B defaults to nboot=599. The argument grp can be used to test the hypothesis of equal trimmed means using a subset of the groups. If missing values are detected, they are eliminated via the function elimna described in Section 1.9.1.

Example

For the alcohol data reported in Section 8.6.2, suppose it is desired to perform all pairwise comparisons using the time 1, time 2, and time 3 data for the control group. The R function pairdepb returns

```
$test:
     Group  Group       test        se
[1,]     1      2  -2.115985  1.693459
[2,]     1      3  -2.021208  1.484261
[3,]     2      3   0.327121  1.783234

$psihat:
     Group  Group      psihat   ci.lower    ci.upper
[1,]     1      2  -3.5833333  -7.194598  0.02793158
[2,]     1      3  -3.0000000  -6.165155  0.16515457
[3,]     2      3   0.5833333  -3.219376  4.38604218

$crit:
[1] 2.132479
```

Thus, none of the pairwise differences is significantly different at the 0.05 level.

■

Assuming the data are stored in the R variable dat, the command pairdepb(dat,grp=c(1,3)) would compare groups 1 and 3, ignoring group 2. It is left as an exercise to show that if the data are stored in list mode, the command ydbt(dat[[1]],dat[[3]]) returns the same confidence interval.

The function

bptd(x,tr=0,alpha=0.05,con=0,nboot=599)

computes confidence intervals for each of C linear contrasts, Ψ_k, $k = 1, \ldots, C$, such that the simultaneous probability coverage is approximately $1 - \alpha$. The only difference between bptd and pairedpb is that bptd can handle a set of specified linear contrasts via the argument con. The argument con is a J-by-C matrix containing the contrast coefficients. The kth column of con contains the contrast coefficients corresponding to Ψ_k. If con is not specified, all pairwise comparisons are performed. So for this special case, pairdepb and bptd always produce the same results.

■ Example

If there are three dependent groups, and con is a 3-by-1 matrix with the values 1, −1, and 0, and if the data are stored in the R variable xv, the command bptd(xv,con=con) will compute a confidence interval for $\Psi = \mu_{t1} - \mu_{t2}$, the difference between the 20% trimmed means corresponding to the first two groups. If xv has list mode, the command ydbt(xv[[1]],xv[[2]]) returns the same confidence interval. (The function ydbt was described in Section 5.9.8.)

■

8.2.5 Percentile Bootstrap Methods

This section describes two types of percentile bootstrap methods that can be used to compare J dependent groups based on any measures of location, θ, associated with the marginal distributions. Included as special cases are M-measures of location and trimmed means. The goal is to test

$$H_0 : \theta_1 = \cdots = \theta_J. \tag{8.1}$$

Method RMPB3

The first method uses the test statistic

$$Q = \sum (\hat{\theta}_j - \bar{\theta})^2,$$

where $\bar{\theta} = \sum \hat{\theta}_j / J$. An appropriate critical value is estimated using an approach similar to the bootstrap-t technique. First, set $C_{ij} = X_{ij} - \hat{\theta}_j$. That is, shift the empirical distributions so that the null hypothesis is true. Next a bootstrap sample is obtained by resampling, with replacement, as described in step 1 of Section 8.2.1. As usual, label the results

$$\begin{pmatrix} C_{11}^*, \ldots, C_{1J}^* \\ \vdots \\ C_{n1}^*, \ldots, C_{nJ}^* \end{pmatrix}.$$

For the jth column of the bootstrap data just generated, compute the measure of location that is of interest and label it $\hat{\theta}_j^*$. Compute

$$Q^* = \sum (\hat{\theta}_j^* - \bar{\theta}^*)^2,$$

where $\bar{\theta}^* = \sum \hat{\theta}_j^* / J$, and repeat this process B times yielding $Q_1^*, \ldots, Q_B^*$. Put these B values in ascending order yielding $Q_{(1)}^* \leq \cdots \leq Q_{(B)}^*$. Then reject the hypothesis of equal measures of location if $Q > Q_{(u)}^*$, where again $u = (1 - \alpha)B$ rounded to the nearest integer.

Method RMPB4

If the null hypothesis is true, then all J groups have a common measure of location, θ. The next method estimates this common measure of location and then checks to see how deeply it is nested within the bootstrap values obtained when resampling from the original values. That is, in contrast to method RMPB3, the data are not centered, and bootstrap samples are obtained by resampling rows of data from

$$\begin{pmatrix} X_{11}, \ldots, X_{1J} \\ \vdots \\ X_{n1}, \ldots, X_{nJ} \end{pmatrix}$$

yielding

$$\begin{pmatrix} X_{11}^*, \ldots, X_{1J}^* \\ \vdots \\ X_{n1}^*, \ldots, X_{nJ}^* \end{pmatrix}.$$

For the jth group (or column of bootstrap values) compute $\hat{\theta}_j^*$. Repeating this process B times yields $\hat{\theta}_{jb}^*$, $(j = 1, \ldots, J; b = 1, \ldots, B)$. The remaining calculations are performed as outlined in Table 8.4.

Table 8.4: Repeated Measures ANOVA Based on the Depth of the Grand Mean.

Goal: Test the hypothesis

$$H_0 : \theta_1 = \cdots = \theta_J.$$

1. Compute

$$S_{jk} = \frac{1}{B-1} \sum_{b=1}^{B} (\hat{\theta}_{jb}^* - \bar{\theta}_j^*)(\hat{\theta}_{kb}^* - \bar{\theta}_k^*),$$

where

$$\bar{\theta}_j^* = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}_{jb}^*.$$

(The quantity S_{jk} is the sample covariance of the bootstrap values corresponding to the jth and kth groups.)

2. Let

$$\hat{\theta}_b^* = (\hat{\theta}_{1b}^*, \ldots, \hat{\theta}_{Jb}^*)$$

and compute

$$d_b = (\hat{\theta}_b^* - \hat{\theta}) \mathbf{S}^{-1} (\hat{\theta}_b^* - \hat{\theta})',$$

where $\mathbf{S}$ is the matrix corresponding to S_{jk}, $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_J)$, $\hat{\theta}_j$ is the estimate of θ based on the original data for the jth group (the X_{ij} values, $i = 1, \ldots, n$), and $\hat{\theta}_b = (\hat{\theta}_{1b}, \ldots, \hat{\theta}_{Jb})$. The value of d_b measures how far away the bth bootstrap vector of location estimators is from $\hat{\theta}$, which is roughly the center of all B bootstrap values.

3. Put the d_b values in ascending order: $d_{(1)} \leq \cdots \leq d_{(B)}$.
4. Let $\hat{\theta}_G = (\bar{\theta}, \ldots, \bar{\theta})$, where $\bar{\theta} = \sum \hat{\theta}_j / J$, and compute

$$D = (\hat{\theta}_G - \hat{\theta}) \mathbf{S}^{-1} (\hat{\theta}_G - \hat{\theta})'.$$

D measures how far away the estimated common value is from the observed measures of location (based on the original data).

5. Reject if $D \geq d_{(u)}$, where $u = (1 - \alpha)B$, rounded to the nearest integer.

For completeness, yet another approach to comparing dependent groups is to use a *mixed linear model* in conjunction with the regression MM-estimator introduced in Chapter 10. Heritier, Cantoni, Copt, and Victoria-Feser (2009, Section 4.5) summarize the relevant details and computations. The mixed linear model has the form

$$Y = \mathbf{X}\alpha + \sum Z_j \beta_j + \epsilon,$$

where Y is a vector of N measurements, $\mathbf{X}$ is an $n \times q$ design matrix for the fixed effects, $\mathbf{Z}_j$ is an $N \times q_j$ design matrix for the random effects β_j, and ϵ is an N-vector of independent residual errors. Evidently it is unknown what advantages this approach might have, in terms of type I errors and power, over the other methods covered in this chapter. (Some concerns about the regression MM-estimator are described in Chapter 10.) Copt and Heritier (2007) derived a (nonbootstrap) method for testing hypotheses that is based in part on an appropriate estimate of the standard errors. However, a general pattern regarding M-estimators seems to be that nonbootstrap methods that use a test statistic based on an estimate of the standard error can perform poorly in terms of type I errors and probability coverage when dealing with skewed distributions. Perhaps the MM-estimator, in the context of the mixed linear model, is an exception, but this has not been investigated.

8.2.6 R Functions bd1way and ddep

The R functions

$$\text{bd1way(x, est} = \text{onestep, nboot} = 599, \text{alpha} = 0.05)$$

and

$$\text{ddep(x, alpha} = 0.05, \text{est} = \text{onestep, grp} = \text{NA, nboot} = 500)$$

perform the percentile bootstrap methods just described. The first function performs method RMPB3; it uses by default the one-step M-estimator of location (based on Huber's Ψ), but any other estimator can be used via the argument est. As usual, x is any R variable that is a matrix or has list mode, nboot is B, the number of bootstrap samples to be used, and grp can be used to analyze a subset of the groups, with the other groups ignored. (That is, grp is used as illustrated in Section 8.1.2.) The function ddep performs method RMPB4 described in Table 8.4.

When there are values missing at random, method M2 in Section 5.9.13 can be used to perform multiple comparisons via the R function

$$\text{rmmismcp(x,y} = \text{NA, alpha} = 0.05, \text{con} = 0, \text{est} = \text{tmean, plotit} = \text{T, grp} = \text{NA, nboot} = 500, \text{SEED} = \text{T, xlab} = \text{“Group 1”, ylab} = \text{“Group 2”, pr} = \text{F, } \ldots).$$

By default, 20% trimmed means are used, but other robust estimators can be used via the argument est.

Example

Table 6.5 shows the weight of cork borings taken from north, east, south, and west sides of the 28 trees. Assuming the data are stored in the R matrix cork, the command bd1way(cork) returns:

```
$test:
17.08

$crit:
34.09
```

So comparing one-step M-estimators, we fail to reject the hypothesis that the typical weight of a cork boring is the same for all four sides of a tree. If we compare groups using MOM in conjunction with method RMPB4, the p-value is .385. (Compare this result to the Example in Section 8.2.8.)

■

Example

Again consider the hangover data used to illustrate rmanova in Section 8.1.2. (The data are listed in Section 8.6.2.) Comparing M-measures of location results in an error because there are too many tied values resulting in MAD=0 within the bootstrap. Assuming the data are stored in x, the command bd1way(x,est=hd) compares medians based on the Harrell–Davis estimator. The function reports that $Q = 9.96$ with a 0.05 critical value of 6.3, so the null hypothesis is rejected at the 0.05 level.

■

8.2.7 Multiple Comparisons Using M-estimators or Skipped Estimators

Next consider C linear contrasts involving M-measures of location where the kth linear contrast is

$$\Psi_k = \sum_{j=1}^{J} c_{jk}\mu_{mj},$$

and, as usual, the c_{jk} values are constants that reflect linear combinations of the M-measures of location that are of interest and for fixed k, $\sum c_{jk} = 0$. The goal is to compute a confidence interval for Ψ_k, $k = 1, \ldots, C$, such that the simultaneous probability coverage is

approximately $1-\alpha$. Alternatively, test $H_0 : \Psi_k = 0$ with the goal that the probability of at least one type I error is α.

First, set $C_{ij} = X_{ij} - \hat{\mu}_{mj}$. Next, obtain a bootstrap sample by sampling, with replacement, n rows of data from the matrix C_{ij}. Label the bootstrap values C^*_{ij}. Use the n values in the jth column of C^*_{ij} to compute $\hat{\mu}^*_{mj}$, $j = 1, \ldots, J$. Repeat this process B times yielding $\hat{\mu}^*_{mjb}$, $b = 1, \ldots, B$. Next, compute the J-by-J covariance matrix associated with the $\hat{\mu}^*_{mjb}$ values. That is, compute

$$\hat{\tau}_{jk} = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\mu}^*_{mjb} - \bar{\mu}^*_j)(\hat{\mu}^*_{mkb} - \bar{\mu}^*_k),$$

where $\bar{\mu}^*_j = \sum \hat{\mu}^*_{mjb}/B$. Let

$$\hat{\Psi}_k = \sum_{j=1}^{J} c_{jk}\hat{\mu}_{mj},$$

$$\hat{\Psi}^*_{kb} = \sum_{j=1}^{J} c_{jk}\hat{\mu}^*_{mjb},$$

$$S^2_k = \sum_j \sum_\ell c_{jk}c_{\ell k}\hat{\tau}_{j\ell},$$

$$T^*_{kb} = \frac{\hat{\Psi}^*_{kb}}{S_k},$$

and

$$T^*_b = \max|T^*_{kb}|,$$

the maximum being taken over $k = 1, \ldots, C$. Then a confidence interval for Ψ_k is

$$\hat{\Psi}_k \pm T^*_{(u)}S_k,$$

where as usual $u = (1-\alpha)B$, rounded to the nearest integer, and $T^*_{(1)} \le \cdots \le T^*_{(B)}$ are the T^*_b values written in ascending order. The simultaneous probability coverage is approximately $1-\alpha$, but for $n \le 21$ and $B = 399$, the actual probability coverage might be unsatisfactory. Under normality, for example, there are situations where the probability of at least one type I error exceeds .08 with $J = 4$, $\alpha = 0.05$, and $n = 21$, and where all pairwise comparisons are performed. Increasing $B = 599$ does not correct this problem. It seems that $n > 30$ is required if the probability of at least one type I error is not to exceed .075 when testing at the 0.05 level (Wilcox, 1997a).

An alternative approach, which appears to have some practical advantages over the method just described, is to use a simple extension of the percentile bootstrap method in

Section 5.9.11. Let $\hat{p}_k^*$ be the proportion of times $\hat{\Psi}_{kb}^* > 0$ among the B bootstrap samples. Then a (generalized) p-value for H_0: $\Psi_k = 0$ is $2\min(\hat{p}_k^*, 1 - \hat{p}_k^*)$. When using M-estimators or MOM, however, a bias adjusted estimate of the p-value appears to be beneficial; see Section 5.9.11. (With trimmed means, this bias adjustment appears to be unnecessary; see Wilcox & Keselman, 2002). FWE can be controlled with method SR outlined in Section 7.6.2. Again, for large sample sizes, say greater than 80, Hochberg's method (mentioned in Section 8.1.3) appears to be preferable.

Note that all of the methods described so far are based on measures of location that do not take into account the overall structure of the data when dealing with outliers. The skipped estimators in Section 6.5 do take the overall structure of the data into account and situations might be encountered where this makes a practical difference. A basic percentile bootstrap method can be used to test H_0: $\Psi_k = 0$ and appears to control the probability of a type I error reasonably well when using the OP estimator in Section 6.5.

8.2.8 R Functions lindm and mcpOV

Using the first method described in the Section 8.2.7, the R function lindm computes confidence intervals for C linear contrasts involving M-measures of location corresponding to J dependent groups. (The second method is performed by the R function in Section 8.3.3.) The function has the form

lindm(x,con=0,est=onestep,grp=0,alpha=0.05,nboot=399, . . .).

The argument x contains the data and can be any n-by-J matrix, or it can have list mode. In the latter case, x[[1]] contains the data for group 1, x[[2]] the data for group 2, and so on. The optional argument con is a J-by-C matrix containing the contrast coefficients. If not specified, all pairwise comparisons are performed. The argument est is any statistic of interest. If unspecified, a one-step M-estimator is used. The argument grp can be used to select a subset of the groups for analysis. As usual, alpha is α and defaults to 0.05, and nboot is B which defaults to 399. The argument . . . is any additional arguments that are relevant to the function est.

Example

Again consider the hangover data (reported in Section 8.6.2) where two groups of participants are measured at three different times. Suppose the first row of data is stored in DAT[[1]], the second row in DAT[[2]], the third in DAT[[3]], and so forth, but it is desired to perform all pairwise comparisons using only the group 1 data at times 1, 2, and 3. Then the command lindm(DAT,grp=c(1:3)) attempts to perform the comparisons, but eventually the function terminates with the error message "missing values in x not allowed." This error arises because there are so many tied values,

bootstrap samples yield $\text{MAD} = 0$ which in turn makes it impossible to compute $\hat{\mu}^*$. This error can also arise when there are no tied values but one or more of the sample sizes are smaller than 20.

■

The command lindm(DAT,est=hd,gpr=c(1:3)) compares medians instead, using the Harrell–Davis estimator, and returns

```
     con.num       psihat   ci.lower    ci.upper        se
[1,]       1  -3.90507026  -7.880827  0.07068639  2.001571
[2,]       2  -3.82383677  -9.730700  2.08302610  2.973775
[3,]       3   0.08123349  -6.239784  6.40225079  3.182279

$crit:
[1] 1.986318

$con:
     [,1]  [,2]  [,3]
[1,]    1     1     0
[2,]   -1     0     1
[3,]    0    -1    -1
```

Because the argument con was not specified, the function creates its own set of linear contrasts assuming all pairwise comparisons are to be performed. The resulting contrast coefficients are returned in the R variable $con. Thus, the first column, containing the values 1, −1, and 0, indicates that the first contrast corresponds to the difference between the medians for times 1 and 2. The results in the first row of $con.num indicate that the estimated difference between these medians is −3.91, and the confidence interval is (−7.9, 0.07). In a similar fashion, the estimated difference between the medians at times 1 and 3 is −3.82, and for time 2 versus time 3 the estimate is 0.08. The command lindm(DAT,est=hd,gpr=c(1:3),q=.4) would compare 0.4 quantiles.

The R function

mcpOV(x,alpha=0.05,nboot=NA,grp=NA,est=smean,con=0,bhop=F,SEED=T, ...).

is like the R function the function lindm, only it is designed to handle skipped estimators that take into account the overall structure of the data when checking for outliers. By default it uses the OP-estimator, which is based on the projection method for detecting outliers.

8.3 Bootstrap Methods Based on Difference Scores

The following method, based on difference scores, has been found to have practical value, particularly in terms of controlling type I error probabilities when sample sizes are very small.

First consider the goal of testing the hypothesis that a measure of location associated with the difference scores $D_{ij} = X_{ij} - X_{i,j+1}$ has the value zero. That is, use the difference between the ith observation in group j and the ith observation in group $j+1$, $j = 1, \ldots, J-1$. Let θ_j be any measure of location associated with the D_{ij} values. So, for example, θ_1 might be an M-measure of location corresponding to the difference scores between groups 1 and 2, and θ_2 might be the M-measure of location associated with difference scores between groups 2 and 3. A simple alternative to Eq. (8.1) is to test

$$H_0 : \theta_1 = \cdots = \theta_{J-1} = 0, \tag{8.2}$$

the hypothesis that the typical difference scores do not differ and are all equal to zero. However, a criticism of this approach is that the outcome can depend on how we order the groups. That is, rather than take differences between groups 1 and 2, we could just as easily take differences between groups 1 and 3, which might alter our conclusions about whether to reject. We can avoid this problem by instead taking differences among all pairs of groups. There are a total of

$$L = \frac{J^2 - J}{2}$$

such differences which are labeled $D_{i\ell}$, $i = 1, \ldots, n$; $\ell = 1, \ldots, L$.

Example

For four groups ($J = 4$), there are $L = 6$ differences given by

$$D_{i1} = X_{i1} - X_{i2},$$

$$D_{i2} = X_{i1} - X_{i3},$$

$$D_{i3} = X_{i1} - X_{i4},$$

$$D_{i4} = X_{i2} - X_{i3},$$

$$D_{i5} = X_{i2} - X_{i4},$$

$$D_{i6} = X_{i3} - X_{i4}.$$

■

The goal is to test

$$H_0 : \theta_1 = \cdots = \theta_L = 0, \tag{8.3}$$

where θ_ℓ is the population measure of location associated with the ℓth set of difference scores, $D_{i\ell}$ ($i = 1, \ldots, n$). To test H_0 given by Eq. (8.3), resample vectors of D values, but unlike the

bootstrap-t, observations are not centered. That is, a bootstrap sample now consists of resampling with replacement n rows from the matrix

$$\begin{pmatrix} D_{11}, \ldots, D_{1L} \\ \vdots \\ D_{n1}, \ldots, D_{nL} \end{pmatrix}.$$

yielding

$$\begin{pmatrix} D^*_{11}, \ldots, D^*_{1L} \\ \vdots \\ D^*_{n1}, \ldots, D^*_{nL} \end{pmatrix}.$$

For each of the L columns of the D^* matrix, compute whatever measure of location is of interest, and for the ℓth column label the result $\hat{\theta}^*_\ell$ ($\ell = 1, \ldots, L$). Next, repeat this B times yielding $\hat{\theta}^*_{\ell b}$, $b = 1, \ldots, B$ and then determine how deeply the vector $\mathbf{0} = (0, \ldots, 0)$, having length L, is nested within the bootstrap values $\hat{\theta}^*_{\ell b}$. For two groups, this is tantamount to determining how many bootstrap values are greater than zero, which leads to the (generalized) p-value described in Section 5.4. The computational details when dealing with more than two groups are relegated to Table 8.5.

8.3.1 R Function rmdzero

The R function

$$\text{rmdzero(x,est} = \text{mom, grp} = \text{NA, nboot} = \text{NA, } \ldots)$$

performs the test on difference scores outlined in Table 8.5.

Example

For the cork data in Table 6.5, rmdzero returns a p-value of .044, so in particular reject with $\alpha = 0.05$. That is, conclude that the typical difference score is not equal to zero for all pairs of groups. This result is in sharp contrast to comparing marginal measures of location based on a robust M-estimator or MOM and the method in Table 8.4; see the Example in Section 8.2.6.

■

8.3.2 Multiple Comparisons

Multiple comparisons based on a percentile bootstrap method and difference scores can be addressed as follows. First generate a bootstrap sample as described at the beginning of this

Table 8.5: Repeated Measures ANOVA Based on Difference Scores

Goal: Test the hypothesis given by Eq. (8.3).

1. Let $\hat{\theta}_\ell$ be the estimate of θ_ℓ. Compute bootstrap estimates as described in Section 8.3 and label them $\hat{\theta}^*_{\ell b}$, $\ell = 1, \ldots, L$; $b = 1, \ldots, B$.
2. Compute the L-by-L matrix

$$S_{\ell\ell'} = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}^*_{\ell b} - \hat{\theta}_\ell)(\hat{\theta}^*_{\ell' b} - \hat{\theta}_{\ell'}).$$

   Readers familiar with multivariate statistical methods might notice that $S_{\ell\ell'}$ uses $\hat{\theta}_\ell$ (the estimate of θ_ℓ based on the original difference values) rather than the seemingly more natural $\bar{\theta}^*_\ell$, where

$$\bar{\theta}^*_\ell = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^*_{\ell b}.$$

   If $\bar{\theta}^*_\ell$ is used, unsatisfactory control over the probability of a type I error can result.
3. Let $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_L)$, $\hat{\theta}^*_b = (\hat{\theta}^*_{1b}, \ldots, \hat{\theta}^*_{Lb})$ and compute

$$d_b = (\hat{\theta}^*_b - \hat{\theta})\mathbf{S}^{-1}(\hat{\theta}^*_b - \hat{\theta})',$$

   where $\mathbf{S}$ is the matrix corresponding to $S_{\ell\ell'}$.
4. Put the d_b values in ascending order: $d_{(1)} \leq \cdots \leq d_{(B)}$.
5. Let

$$\mathbf{0} = (0, \ldots, 0)$$

   having length L.
6. Compute

$$D = (\mathbf{0} - \hat{\theta})\mathbf{S}^{-1}(\mathbf{0} - \hat{\theta})'.$$

   D measures how far away the null hypothesis is from the observed measures of location (based on the original data). In effect, D measures how deeply $\mathbf{0}$ is nested within the cloud of bootstrap values.
7. Reject if $D \geq d_{(u)}$, where $u = (1-\alpha)B$, rounded to the nearest integer.

section yielding $D^*_{i\ell}$, $\ell = 1, \ldots, L$. When all pairwise differences are to be tested, $L = (J^2 - J)/2$, $\ell = 1$ corresponds to comparing group 1 to group 2, $\ell = 2$ is comparing group 1 to group 3, and so on. Let $\hat{p}^*_\ell$ be the proportion of times among B bootstrap resamples that $D^*_{i\ell} > 0$. As usual, let

$$\hat{p}^*_{m\ell} = \min(\hat{p}^*_\ell,\ 1 - \hat{p}^*_\ell),$$

in which case $2\hat{p}^*_{m\ell}$ is the estimated (generalized) p-value for the ℓth comparison.

One approach to controlling FWE is to put the p-values in descending order and to make decisions about which hypotheses are to be rejected using method SR outlined in

Section 7.6.2. That is, once the $\hat{p}^*_{mc}$ are computed, reject the hypothesis corresponding to $\hat{p}^*_{mc}$ if $\hat{p}^*_{mc} \leq \alpha_c$, where α_c is read from Table 7.13.

As for linear contrasts, consider any specific linear contrast with contrast coefficients $c_1, \ldots, c_J$, set

$$D_i = \sum c_j X_{ij},$$

and let θ_d be some (population) measure of location associated with this sum. Then $H_0 : \theta_d = 0$ can be tested by generating a bootstrap sample from the D_i values, repeating this B times, computing $\hat{p}^*$, the proportion of bootstrap estimates that are greater than zero, in which case $2\min(\hat{p}^*, 1 - \hat{p}^*)$ is the estimated significance level. Then FWE can by controlled in the manner just outlined.

When comparing groups using MOM or M-estimators, at the moment it seems that the method based on difference scores often provides the best power versus testing hypotheses based on measures of location associated with the marginal distributions. Both approaches do an excellent job of avoiding type I error probabilities greater than the nominal α level. But when testing hypotheses about measures of location associated with the marginal distributions, the actual type I error probability can drop well below the nominal level in situations where the method based on difference scores avoids this problem. This suggests that the method based on difference scores will have more power, and indeed, there are situations where this is the case even when the two methods have comparable type I error probabilities. It is stressed, however, that a comparison of these methods, in terms of power, needs further study. Also the bias adjusted critical value mentioned in Section 5.9.7 appears to help increase power.

While comparing marginal measures of location based on MOM or an M-estimator seems to result in relatively low power, there is weak evidence that comparing marginal measures of location based on the OP-estimator, via a percentile bootstrap method as mentioned at the end of Section 8.2.7, performs relatively well. But the extent this is true needs additional study.

When dealing with trimmed means, again the percentile bootstrap method just described can be used and appears to be relatively good choice provided the amount of trimming is not too small. With a small amount of trimming, use a bootstrap-t method instead. Controlling the probability of at least one type I error can be done with Hochberg's method.

8.3.3 R Functions rmmcppb, wmcppb, dmedpb, and lindepbt

The R function

rmmcppb(x,y = NA, alpha = 0.05, con = 0, est = mom, plotit = T, dif = T, grp = NA, nboot = NA, BA=F,hoch=F, ...)

performs multiple comparisons among dependent groups using the percentile bootstrap methods just described. The argument dif defaults to T (for true) indicating that difference scores will be used, in which case Hochberg's method is used to control FWE. If dif=F, measures of location associated with the marginal distributions are used instead. If dif=F and BA=T, the bias adjusted estimate of the generalized p-value (described in Section 5.9.7) is applied; using BA=T (when dif=F) is recommended when comparing groups with M-estimators and MOM, but it is not necessary when comparing 20% trimmed means (Wilcox & Keselman, 2002). If hoch=F, then FWE is controlled using method SR in Section 7.6.2 if the sample size is less than 80, otherwise Hochberg's method is used as described in Section 8.1.3. If hoch=T, Hochberg's method is used regardless of the sample size. If no value for con is specified, then all pairwise differences will be tested. As usual, if the goal is to test hypotheses other than all pairwise comparisons, con can be used to specify the linear contrast coefficients.

When comparing trimmed means, it appears that Hochberg's method is preferable to method SR in terms of controlling the probability of at least one type I error. For convenience, the R function

wmcppb(x, alpha = 0.05, con = 0, est = tmean, plotit = T, dif = T, grp = NA, nboot = NA, BA=F,hoch=T, ...)

is supplied. It is the same as the R function rmmcppb, only it defaults to comparing 20% trimmed means, and by default it uses Hochberg's method rather than method SR. (The R function dtrimpb is the same as the function wmmcppb.)

The R function

dmedpb(x,y=NA,alpha=0.05,con=0,est=median,plotit=T,dif=F,grp=NA, hoch=T,nboot=NA,xlab="Group 1",ylab="Group 2",pr=T,SEED=T,BA=F, ...)

is similar to the R function rmmcppb, only it defaults to comparing medians and it is designed to handle tied values. Hochberg's method is used to control FWE. With a small sample size, say less than 30, setting the argument BA=T seems advisable, meaning that the p-value is adjusted as described in 5.9.11 (Wilcox, 2006b).

The R function

lindepbt(x, con = NULL, tr = 0.2, alpha = 0.05,nboot=599,dif=T,SEED=T)

performs multiple comparisons based on trimmed means using a bootstrap-t method. When the amount of trimming is small, a bootstrap-t method is preferable to a percentile bootstrap method, but it is unclear at what point this will be case. The function reports critical p-values based on Rom's method for controlling the probability of one or more type I errors. The function returns confidence intervals, but they are not adjusted so that the simultaneously

probability coverage is $1-\alpha$. Rather, each confidence interval is designed to have probability coverage $1-\alpha$.

Example

For the cork boring data in Table 6.5, the R function wmcppb (with the argument dif=F as well as dif=T) finds no significant results when the probability of at least one type I error is taken to be .05. But the R function mcpOV (in Section 8.2.8), which compares marginal measures of location via the OP-estimator, finds three significant results, the only point being that the choice of method can make a practical difference. Again it is stressed that little is known about the extent the OP-estimator might have higher power compared to the many other methods that might be used to compare dependent groups.

■

8.4 Comments on which Method to Use

No single method in this chapter dominates based on various criteria used to compare hypothesis testing techniques. However, a few comments can be made about their relative merits that might be useful. First, the expectation is that in many situations where groups differ, all methods based on means perform poorly, in terms of power, relative to approaches based on some robust measure of location such as MOM or a 20% trimmed mean. Currently, with a sample size as small as 21, the bootstrap-t method in Section 8.2.2, which is performed by the R function rmanovab, appears to provide excellent control over the probability of a type I error when used in conjunction with 20% trimmed means. Its power compares reasonably well to most other methods that could be used, but as noted in previous chapters, different methods are sensitive to different features of the data and arguments for some other measure of location, such as an M-estimator, have been made.

The percentile bootstrap methods in Section 8.2.5 also do an excellent job of avoiding type I errors greater than the nominal level, but there are indications that when using method RMPB3, and the sample size is small, the actual probability of a type I error can be substantially less than α suggesting that some other method might provide better power. Nevertheless, if there is specific interest in comparing M-estimators associated with the marginal distributions, it is suggested that method RMPB3 be used when the sample size is greater than 20. Also, it can be used to compare groups based on MOM, but with very small sample sizes power might be inadequate relative to other techniques that could be used. Given the goal of testing some omnibus hypothesis, currently, among the techniques covered in this chapter, it seems that the two best methods for controlling type I error probabilities and

simultaneously providing relatively high power are the bootstrap-t method based on 20% trimmed means and the percentile bootstrap method in Table 8.5, which is based in part on difference scores; the computations are performed by the R function rmdzero. (But also consider the multiple comparison procedure in Section 8.1.4.) With near certainty, situations arise where some other technique is more optimal, but typically the improvement is small. But again, comparing groups with MOM is not the same as comparing means, trimmed means or M-estimators and certainly there will be situations where some other estimator has higher power than any method based on MOM or a 20% trimmed mean. If the goal is to maximize power, several methods are contenders for routine use. With sufficiently large sample sizes, trimmed means can be compared without resorting to the bootstrap-t method, but it remains unclear just how large the sample size must be. Roughly, as the amount of trimming increases from 0% to 20%, the smaller the sample size must be to control the probability of a type I error without resorting to a bootstrap technique. But if too much trimming is done, power might be relatively low. When comparing medians, a percentile bootstrap method is recommended when dealing with tied values.

As for the issue of whether to use difference scores rather than robust measures of location based on the marginal distributions, each approach provides a different perspective on how groups differ and they can give different results regarding whether groups are significantly different. There is some evidence that difference scores typically provide more power and better control over the probability of a type I error, but situations are encountered where the reverse is true. A more detailed study is needed to resolve this issue.

As previously mentioned, method RMPB4 performed by the R function ddep in Section 8.2.6 is very conservative in terms of type I errors, meaning that when testing at the 0.05 level, say, often the actual probability of a type I error will be less than or equal to α and typically smaller than any other method described in this chapter. So a concern is that power might be low relative to the many other methods that might be used.

Regarding methods designed for performing multiple comparisons in a manner that controls the probability of at least one type I error, currently it seems that using the R function wmcppb, which compares groups based on trimmed means, is a good choice, particularly in terms of maximizing power. But again, there are exceptions. Perhaps the method in Section 8.2.7, which is based on the OP-estimator and performed by the R function mcpOV, generally competes well with the R function wmcppb, but little is known about the extent to which this is true. In general, no single method is always best. As in previous chapters, in terms of controlling the probability of at least one type I error, the method used by wmcppb and mcpOV does not assume or require that one first test and reject the global hypothesis that all groups have identical population trimmed means. It is not necessary, for example, that the function rmanovab (described in Section 8.2.2) returns a significant result before using the function wmcppb.

8.5 Some Rank-Based Methods

This section describes two rank-based methods for testing

$$H_0 : F_1(x) = \cdots = F_J(x),$$

the hypothesis that J dependent groups have identical marginal distributions. Friedman's test is the best-known test of this hypothesis, but no details are given here. The first of the two methods was derived by Agresti and Pendergast (1986) and has higher power than Friedman's test when sampling from normal distributions, and their test can be expected to have good power when sampling from heavy-tailed distributions, so it is included here. The other method stems from Brunner, Domhof, and Langer (2002, Section 7.2.2), which appears to have an advantage over the Agresti–Pendergast method in terms of power (Tian & Wilcox, 2007).

Method AP

The calculations for the Agresti–Pendergast method are shown in Table 8.6.

Method BPRM

Let R_{ij} be defined as in Table 8.6 and let $\mathbf{R}_i = (R_{i1}, \ldots, R_{iJ})'$ be the vector of ranks for the ith participant, where $(R_{i1}, \ldots, R_{iJ})'$ is the transpose of $(R_{i1}, \ldots, R_{iJ})$. Let

$$\bar{\mathbf{R}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{R}_i$$

be the vector of ranked means, let

$$\bar{R}_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n} R_{ij}$$

denote the mean of the ranks for group j and let

$$\mathbf{V} = \frac{1}{N^2(n-1)}\sum_{i=1}^{n}(\mathbf{R}_i - \bar{\mathbf{R}})(\mathbf{R}_i - \bar{\mathbf{R}})'.$$

The test statistic is

$$F = \frac{n}{N^2 \mathrm{tr}(\mathbf{PV})}\sum_{j=1}^{J}\left(\bar{R}_{\cdot j} - \frac{N+1}{2}\right)^2, \tag{8.4}$$

where

$$\mathbf{P} = \mathbf{I} - \frac{1}{J}\mathbf{J},$$

$\mathbf{J}$ is a $J \times J$ matrix of all ones, and $\mathbf{I}$ is the identity matrix.

Table 8.6: Computing the Agresti–Pendergast Test Statistic.

Pool all the observations and assign ranks. Let R_{ij} be the resulting rank of the ith observation in the jth group. Compute

$$\bar{R}_j = \frac{1}{n}\sum_{i=1}^{n} R_{ij}$$

$$s_{jk} = \frac{1}{n-J+1}\sum_{i=1}^{n}(R_{ij}-\bar{R}_j)(R_{ik}-\bar{R}_k).$$

Let the vector $\boldsymbol{R}'$ be defined by

$$\mathbf{R}' = (\bar{R}_1, \ldots, \bar{R}_J),$$

and let $\mathbf{C}$ be the $(J-1)$-by-J matrix given by

$$\begin{pmatrix} 1 & -1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & -1 & \ldots & 0 & 0 \\ . & . & . & . & . & . \\ 0 & 0 & 0 & \ldots & 1 & -1 \end{pmatrix}.$$

The test statistic is

$$F = \frac{n}{J-1}(\mathbf{CR})'(\mathbf{CSC}')^{-1}\mathbf{CR},$$

where

$$\mathbf{S} = (s_{jk}).$$

The degrees of freedom are $\nu_1 = J-1$ and $\nu_2 = (J-1)(n-1)$, and you reject if $F > f_{1-\alpha}$, the $1-\alpha$ quantile of an F distribution with ν_1 and ν_2 degrees of freedom.

Decision Rule

Reject the hypothesis of identical distributions if

$$F \geq f,$$

where f is the $1-\alpha$ quantile of an F distribution with degrees of freedom

$$\nu_1 = \frac{[\mathrm{tr}(\mathbf{PV})]^2}{\mathrm{tr}(\mathbf{PVPV})}$$

and $\nu_2 = \infty$. Note that based on the test statistic, a crude description of method BPRM is that it is designed to be sensitive to differences among the average ranks.

8.5.1 R Functions apanova and bprm

The R function

apanova(x,grp=0)

performs the Agresti–Pendergast test of equal marginal distributions using the calculations in Table 8.6. As usual, x can have list mode, or x can be an n-by-J matrix, and the argument grp can be used to specify some subset of the groups. If grp is unspecified, all J groups are used. The function returns the value of the test statistic, the degrees of freedom, and the p-value. For example, the command apanova(dat,grp=c(1,2,4)) would compare groups 1, 2, and 4 using the data in the R variable dat.

The R function

bprm(x)

performs method BPRM; it returns a p-value.

8.6 Between-by-Within and Within-by-Within Designs

This section describes some methods for testing hypotheses in a between-by-within (or split-plot) design. That is, a J-by-K ANOVA design is being considered where the J levels of the first factor correspond to independent groups (between subjects), and the K levels of the second factor are dependent (within subjects). Within-by-within designs are covered as well.

8.6.1 Analyzing a Between-by-Within Design Based on Trimmed Means

We begin with a between-by-within design. For the jth level of factor A, let Σ_j be the K-by-K population Winsorized covariance matrix for the K dependent random variables associated with the second factor. The better-known methods for analyzing a split-plot design are based on the assumption that $\Sigma_1 = \cdots = \Sigma_J$, but violating this assumption can result in problems controlling the probability of a type I error. Keselman, Keselman, and Lix (1995) found that a method derived by Johansen (1980), that does not assume there is a common covariance matrix, gives better results, so a generalization of Johansen's method, to trimmed means, is described here. (For related results, see Keselman, Algina, Kowalchuk, & Wolfinger, 1999; Keselman, Carriere, & Lix, 1993; Livavcic-Rojas, Vallejo, & Fernández, 2010.)

As in the case of a two-way design for independent groups, it is easier to describe the method in terms of matrices. Main effects and interactions are examined by testing

$$H_0 : \mathbf{C}\mu_t = \mathbf{0},$$

where $\mathbf{C}$ is a k-by-JK contrast matrix having rank k that reflects the null hypothesis of interest. Let $\mathbf{C}_m$ and $\mathbf{j}'$ be defined as in Section 7.3. Then for factor A, $\mathbf{C} = \mathbf{C}_J \otimes \mathbf{j}'_K$, and $k = J - 1$. For factor B, $\mathbf{C} = \mathbf{j}'_J \otimes \mathbf{C}_K$ and $k = K - 1$, and the test for no interactions uses $\mathbf{C} = \mathbf{C}_J \otimes \mathbf{C}_K$.

For every level of factor A, there are K dependent random variables, and each pair of these dependent random variables has a Winsorized covariance that must be estimated. In symbols, let X_{ijk} be the ith observation randomly sampled from the jth level of factor A and the kth level of factor B. For fixed j, the Winsorized covariance between the mth and ℓth levels of factor B is estimated with

$$s_{jm\ell} = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (Y_{ijm} - \bar{Y}_{.jm})(Y_{ij\ell} - \bar{Y}_{.j\ell}),$$

where

$$Y_{ijk} = \begin{cases} X_{(g+1),jk} & \text{if } X_{ijk} \leq X_{(g+1),jk} \\ X_{ijk} & \text{if } X_{(g+1),jk} < X_{ij} < X_{(n-g),jk} \\ X_{(n-g),jk} & \text{if } X_{ijk} \geq X_{(n-g),jk}, \end{cases}$$

and

$$\bar{Y}_{.jm} = \frac{1}{n} \sum_{i=1}^{n} Y_{ijm}.$$

For fixed j, let $\mathbf{S}_j = (s_{jm\ell})$. That is, $\mathbf{S}_j$ estimates Σ_j, the K-by-K Winsorized covariance matrix for the jth level of factor A. Let

$$\mathbf{V}_j = \frac{(n_j - 1)\mathbf{S}_j}{h_j(h_j - 1)}, \quad j = 1, \ldots, J,$$

and let $\mathbf{V} = \text{diag}(\mathbf{V}_1, \ldots, \mathbf{V}_J)$ be a block diagonal matrix. The test statistic is

$$Q = \bar{\mathbf{X}}'\mathbf{C}'(\mathbf{C}\mathbf{V}\mathbf{C}')^{-1}\mathbf{C}\bar{\mathbf{X}}, \tag{8.5}$$

where $\bar{\mathbf{X}}' = (\bar{X}_{t11}, \ldots, \bar{X}_{tJK})$. Let $\mathbf{I}_{K\times K}$ be a K-by-K identity matrix, let $\mathbf{Q}_j$ be a JK by JK block diagonal matrix (consisting of J blocks, each block being a K-by-K matrix), where the tth block ($t = 1, \ldots, J$) along the diagonal of $\mathbf{Q}_j$ is $\mathbf{I}_{K\times K}$ if $t = j$, and all other elements are zero. (For example, if $J = 3$ and $K = 4$, then $\mathbf{Q}_1$ is a 12-by-12 matrix block diagonal matrix where the first block is a 4-by-4 identity matrix, and all other elements are zero. As for $\mathbf{Q}_2$, the second block is an identity matrix, and all other elements are zero.) Compute

$$A = \frac{1}{2} \sum_{j}^{J} \{\text{tr}[(\mathbf{V}\mathbf{C}'(\mathbf{C}\mathbf{V}\mathbf{C}')^{-1}\mathbf{C}\mathbf{Q}_j)^2] + [\text{tr}(\mathbf{V}\mathbf{C}'(\mathbf{C}\mathbf{V}\mathbf{C}')^{-1}\mathbf{C}\mathbf{Q}_j)]^2\}/(h_j - 1).$$

where tr indicates trace, and let

$$c = k + 2A - \frac{6A}{k+2}.$$

When the null hypothesis is true, Q/c has, approximately, an F distribution with $\nu_1 = k$ and $\nu_2 = k(k+2)/(3A)$ degrees of freedom, so reject if $Q/c > f_{1-\alpha}$, the $1-\alpha$ quantile. Recent simulation results reported by Livavcic-Rojas et al. (2010) indicate that this method performs relatively well, in terms of controlling the probability of a type I error, when comparing means. However, their results do not consider the effect of having different amounts of skewness.

Example

Consider a 2-by-3 design where for the first level of factor A, observations are generated from a multivariate normal distribution with all correlations equal to zero. For the second level of factor A, the marginal distributions are lognormal that have been shifted to have mean zero. Further suppose that the covariance matrix for the second level is three times larger that the covariance matrix for the first level. If the sample sizes are $n_1 = n_2 = 30$ and the hypothesis of no main effects for factor A, based on means, is tested at the 0.05 level, the actual level is approximately 0.088. If the sample sizes are $n_1 = 40$ and $n_2 = 70$ and for the first level of factor A the marginal distributions are g-and-h distributions with $g = h = 0.5$, the probability of a type I error, again testing at the 0.05 level, is approximately .188. Comparing 20% trimmed means instead, the actual type I error probability is approximately .035.

■

8.6.2 R Functions bwtrim and tsplit

The R function

bwtrim(J,K,x,tr=.2,grp=c(1:p),p=J*K)

tests the hypotheses of no main effects and no interactions in a between-by-within (split-plot) design, where J is the number of independent groups, K is the number of dependent groups, and the argument x contains the data stored in list mode, or a matrix, or a data frame. The optional argument, tr, indicates the amount of trimming, which defaults to 0.2 if unspecified.

The groups are assumed to be ordered as described in Section 7.2.1. If the data are not stored in the proper order, grp can be used to indicate how they are stored. For example, if a 2-by-2 design is being used, the R command

bwtrim(2,2,x,grp=c(3,1,2,4))

indicates that the data for the first level of both factors are stored in x[[3]], the data for level 1 of Factor A and level 2 of Factor B are in x[[1]], and so forth. If x is a matrix or a data frame, the first K columns correspond to the first level of Factor A and the K levels of Factor B, the next K columns correspond to the second level of Factor A, and so on.

The R function tsplit, which is described in earlier editions of this book, performs the same analysis. The function bwtrim was added to match naming conventions used by other functions to be described.

Example

In a study on the effect of consuming alcohol, hangover symptoms were measured for two independent groups, with each subject consuming alcohol and being measured on three different occasions. One group (group 2) consisted of sons of alcoholics and the other was a control group. Here, $J = 2$ and $K = 3$. The results were as follows.

Group 1, Time 1	0 32 9 0 2 0 41 0 0 0 6 18 3 3 0 11 11 2 0 11
Group 1, Time 2	4 15 26 4 2 0 17 0 12 4 20 1 3 7 1 11 43 13 4 11
Group 1, Time 3	0 25 10 11 2 0 17 0 3 6 16 9 1 4 0 14 7 5 11 14
Group 2, Time 1	0 0 0 0 0 0 0 0 1 8 0 3 0 0 32 12 2 0 0 0
Group 2, Time 2	2 0 7 0 4 2 9 0 1 14 0 0 0 0 15 14 0 0 7 2
Group 2, Time 3	1 0 3 0 3 0 15 0 6 10 1 1 0 2 24 42 0 0 0 2

Suppose the first row of data is stored in DAT[[1]], the second row in DAT[[2]], the third in DAT[[3]], the fourth in DAT[[4]], the fifth in DAT[[5]], and the sixth in DAT[[6]]. That is, the data are stored as assumed by the function bwtrim. Then the command bwtrim(2,3,DAT,tr=0) will compare the means and returns

```
$Qa:
[1] 3.277001

$Qa.siglevel:
[1] 0.149074

$Qb:
[1] 0.7692129

$Qb.siglevel:
[1] 0.5228961

$Qab:
[1] 0.917579

$Qab.siglevel:
[1] 0.4714423
```

The test statistic for factor A is $Qa = 3.28$, and the corresponding p-value is .159. For factor B the p-value is .52, and for the interaction it is .47.

■

This section described one way of comparing independent groups in a between-by-within subjects design. Another approach is simply to compare the J independent groups for each level of Factor B. That is, do not sum or average the data over the levels of Factor B as was done here. So the goal is to test

$$H_0 : \mu_{t1k} = \cdots \mu_{tJk}$$

for each $k = 1, \ldots, K$. The next example illustrates that in applied work, the choice between these two methods can make a practical difference. (Yet another strategy for comparing the levels of Factor A is to apply the robust MANOVA method in Section 7.10.)

■ Example

Section 7.8.4 reported data from a study comparing schizophrenics to a control group based on a measure taken at two different times. Analyzing the data with the function tsplit, no main effects for Factor A are found, the p-value being .245. So no difference between the schizophrenics and the control group was detected at the 0.05 level. But if the groups are compared using the first measurement only, using the function yuen (described in Chapter 5), the p-value is .012. For the second measurement, ignoring the first, the p-value is .89. (Recall that in Chapter 6, using some multivariate methods, again a difference between the schizophrenics and the control group was found.)

■

8.6.3 Data Management: R Function bw2list

Imagine a situation where data are stored in a matrix or a data frame, say x, with one column indicating the levels of the between factor, but the K levels of the within group factor are stored in K columns of the matrix x. In order to use the R function bwtrim, it is necessary to store the data in the format that is allowed. The R function

bw2list(x, grp.col, lev.col)

is provided to help accomplish this goal. The argument grp.col indicates the column containing information about the levels of the independent groups. The values in this column can be numeric or character data. And the argument lev.col indicates the K columns where the within group data are stored. The function returns the data stored in list mode, which can then be used by bwtrim as well as other functions aimed at dealing with a between-by-within design. The function will store the data sorted in ascending (or alphabetical) order based on

the values found in the column of x indicated by the argument grp.col. The next example illustrates this feature.

Example

Imagine that three medications are being investigated regarding their effectiveness to lower cholesterol and that column 3 of the matrix m indicates which medication a participant received. Moreover, columns 5 and 8 contain the participants' cholesterol level at times 1 and 2, respectively. The R command

```
z=bw2list(m,3,c(5,8))
```

will store the data in z in list mode. If column 3 contains the character values "P", "CH", and "BN", then z[[1]] and z[[2]] will contain the data for times 1 and 2, respectively, corresponding to level "BN" of Factor A, z[[3]] and z[[4]] will contain the data for level "CH", and z[[5]] and z[[6]] will contain the data for level "P". The R command

```
bwtrim(3,2,z)
```

will compare the groups based on 20% trimmed means.

■

8.6.4 Bootstrap-t Method for a Between-by-Within Design

To apply a bootstrap-t method, when working with trimmed means and dealing with a between-by-within design, first center the data in the usual way. In the present context, this means you compute

$$C_{ijk} = X_{ijk} - \bar{X}_{tjk},$$

$i = 1, \ldots, n_j$; $j = 1, \ldots, J$; and $k = 1, \ldots, K$. That is, for the group corresponding to the jth level of factor A and the kth level of factor B, subtract the corresponding trimmed mean from each of the observations. Next, for each j, generate a bootstrap sample based on the C_{ijk} values by resampling with replacement n_j vectors of observations from the n_j rows of data corresponding to level j of factor A. That is, for each level of factor A, you have an n_j-by-K matrix of data, and you generate a bootstrap sample from this matrix of data as described in Section 8.2.5 where for fixed j, resampling is based on the C_{ijk} values. Label the resulting bootstrap samples C^*_{ijk}. Compute the test statistic Q, based on the C^*_{ijk} values as described in Section 8.6.1 and label the result Q^*. Repeat this B times yielding $Q^*_1, \ldots, Q^*_B$ and then put these B values in ascending order yielding $Q^*_{(1)} \leq \cdots \leq Q^*_{(B)}$. Next, compute Q using the original data (the X_{ijk} values) and reject if $Q \geq Q^*_{(c)}$, where $c = (1 - \alpha)$ rounded to the nearest integer.

A crude rule that seems to apply to a wide variety of situations is: the more the distributions (associated with groups) differ, the more beneficial it is to use some type of bootstrap method,

at least when sample sizes are small. Keselman et al. (2000) compared the bootstrap method just described to the nonbootstrap method for a split-plot design covered in Section 8.6.1. For the situations they examined, this rule did not apply; it was found that the bootstrap offered little or no advantage. Their study included situations where the correlations (or covariances) among the dependent groups differ across the independent groups being compared. However, the more complicated the design, the more difficult it becomes to consider all the factors that might influence the operating characteristics of a particular method. One limitation of their study was that the differences among the covariances were taken to be relatively small. Another issue that has not been addressed is how the bootstrap method performs when distributions differ in skewness. Having differences in skewness is known to be important when dealing with the simple problem of comparing two groups only. There is no reason to assume that this problem diminishes as the number of groups increases, and indeed there are reasons to suspect that it becomes a more serious problem. So currently, it seems that if groups do not differ in any manner, or the distributions differ slightly, it makes little difference whether you use a bootstrap versus a nonbootstrap method for comparing trimmed means. However, if distributions differ in shape, there is indirect evidence that a bootstrap method might offer an advantage when using a split-plot design, but the extent to which this is true is not well understood.

8.6.5 R Functions bwtrimbt and tsplitbt

The R function

```
tsplitbt(J,K,x,tr=0.2,alpha=0.05,JK=J*K,grp=c(1:JK),nboot=599)
```

performs a bootstrap-t method for a split-plot design as just described. The data are assumed to be arranged as indicated in conjunction with the R function tsplit (as described in Section 8.6.2), and the arguments J, K, tr, and alpha have the same meaning as before. The argument JK can be ignored, and grp can be used to rearrange the data if they are not stored as expected by the function. (For an R function that might help when dealing with organizing the data in a manner that is accepted by tsplitbt, see Section 8.6.3.)

The R function

```
bwtrimbt(J,K,x,tr=0.2,JK=J*K,grp=c(1:JK),nboot=599)
```

is the same as tsplitbt, only bwtrimbt reports p-values rather than α level critical values.

8.6.6 Percentile Bootstrap Methods for a Between-by-Within Design

Comparing groups based on MOMs, medians, and M-estimators in a between-by-within design is possible using extensions of percentile bootstrap methods already described. And they provide yet another way of comparing trimmed means.

Again consider a two-way design where factor A consists of J independent groups and Factor B corresponds to K dependent groups. First consider the dependent groups. One approach to comparing these K groups, ignoring Factor A, is to simply form difference scores and then apply the method in Section 8.3. More precisely, imagine you observe X_{ijk} ($i = 1, \ldots, n_j$; $j = 1, \ldots, J$; $k = 1, \ldots, K$). That is, X_{ijk} is the ith observation in level j of Factor A and level k of Factor B. Note that if we ignore the levels of Factor A, we can write the data as Y_{ik}, $i = 1, \ldots, N$; $k = 1, \ldots, K$, where $N = \sum n_j$. Now consider levels k and k' of Factor B ($k < k'$) and set

$$D_{ikk'} = Y_{ik} - Y_{ik'},$$

and let $\theta_{kk'}$ be some measure of location associated with $D_{ikk'}$. Then the levels of Factor B can be compared, ignoring Factor A, by testing

$$\theta_{12} = \cdots = \theta_{k-1,k} = 0 \tag{8.6}$$

using the method in Section 8.3. In words, the null hypothesis is that the typical difference score between any two levels of Factor B, ignoring Factor A, is zero.

As for Factor A, ignoring Factor B, one approach is as follows. Momentarily focus on the first level of Factor B and note that the levels of Factor A can be described as in Chapter 7. That is, the null hypothesis of no differences among the levels of Factor A is

$$H_0 : \theta_{11} = \theta_{21} = \cdots = \theta_{J1},$$

where of course these J groups are independent, and a percentile bootstrap method can be used. More generally, for any level of Factor B, say the kth, the hypothesis of no main effects is

$$H_0 : \theta_{1k} = \theta_{2k} = \cdots = \theta_{Jk},$$

($k = 1, \ldots, K$), and the goal is to determine whether these K hypotheses are simultaneously true. Here we take this to mean that we want to test

$$H_0 : \theta_{11} - \theta_{21} = \cdots \theta_{J-1,1} - \theta_{J1} = \cdots = \theta_{J-1,K} - \theta_{JK} = 0. \tag{8.7}$$

In this last equation, there are $C = K(J^2 - J)/2$ differences, all of which are hypothesized to be equal to zero. Proceeding along the lines in Chapter 7, for each level of Factor A, generate bootstrap samples as is appropriate for K dependent groups and then test Eq. (8.7). Label the C differences based on the observed data as $\delta_1, \ldots, \delta_C$ and then denote bootstrap estimates by $\hat{\delta}_c^*$ ($c = 1, \ldots, C$). For example, $\hat{\delta}_1^* = \theta_{11}^* - \theta_{21}^*$. Then we test Eq. (8.5) by determining how deeply the vector $(0, \ldots, 0)$, having length C, is nested within the B bootstrap values, which is done as described in Table 8.5. However, a criticism of this method is that control over the

probability of a type I error can be unsatisfactory it can (exceed .075 when testing at the 0.05 level) when the sample size is small.

For Factor A, an alternative approach, which seems more satisfactory in terms of type I errors, is to base the analysis on the average measures of location across the K levels of Factor B. In symbols, let

$$\bar{\theta}_{j.} = \frac{1}{K}\sum_{k=1}^{K}\theta_{jk},$$

in which case the goal is to test

$$H_0 : \bar{\theta}_{1.} = \cdots = \bar{\theta}_{J.}.$$

Again for each level of Factor A, generate B samples for the K dependent groups as described in Section 8.2.5 in conjunction with method RMPB4. Let $\bar{\theta}^*_{j.}$ be the bootstrap estimate for the jth level of Factor A. For levels j and j' of Factor A, $j < j'$, set $\delta^*_{jj'} = \bar{\theta}^*_{j.} - \bar{\theta}^*_{j'.}$, Then you determine how deeply $\mathbf{0}$, having length $(J^2 - J)/2$, is nested within the B bootstrap values for $\delta^*_{jj'}$ using the method described in Table 8.5. When dealing with Factor A, this approach seems to be more satisfactory than the strategy described in the previous paragraph.

As for interactions, again there are several approaches one might adopt. Here an approach based on difference scores among the dependent groups is used. To explain, first consider a 2-by-2 design, and for the first level of Factor A let $D_{i1} = X_{i11} - X_{i12}$, $i = 1, \ldots, n_1$. Similarly, for level 2 of Factor A let $D_{i2} = X_{i21} - X_{i22}$, $i = 1, \ldots, n_2$, and let θ_{d1} and θ_{d2} be the population measure of location corresponding to the D_{i1} and D_{i1} values, respectively. Then the hypothesis of no interaction is taken to be

$$H_0 : \theta_{d1} = \theta_{d2},$$

which of course is the same as

$$H_0 : \theta_{d1} - \theta_{d2} = 0. \tag{8.8}$$

Again the basic strategy for testing hypotheses is generating bootstrap estimates and determining how deeply 0 is embedded in the B values that result. For the more general case of a J-by-K design, there are a total of

$$C = \frac{J^2 - J}{2} \times \frac{K^2 - K}{2}$$

equalities, one for each pairwise difference among the levels of Factor B and any two levels of Factor A.

8.6.7 R Functions sppba, sppbb, and sppbi

The R function

sppba(J,K,x,est=onestep,grp = c(1:JK),avg=T,nboot=500,MC=F,MDIS=T, ...)

argument avg to T (for true) indicates that the averages of the measures of location (the $\bar{\theta}_{j.}$ values) will be used. That is, $H_0 : \bar{\theta}_{1.} = \cdots = \bar{\theta}_{J.}$ is tested. Otherwise, the hypothesis given by Eq. (8.6) is tested. By default, the argument MDIS=T, meaning that the depths of the points in the bootstrap cloud are based on Mahalanobis distance. Otherwise a projection distance is used, which was described in Section 6.2.5. If MDIS=F and MC=T, a multi-core processor will be used if one is available. The remaining arguments have their usual meaning.

The R function

sppbb(J,K,x,est=onestep,grp = c(1:JK),nboot=500, ...)

tests the hypothesis of no main effects for Factor B (as described in the previous section) and

sppbi(J,K,x,est=onestep,grp = c(1:JK),nboot=500, ...)

tests the hypothesis of no interactions.

Example

We examine once more the EEG measures for murderers versus a control group reported in Table 6.1, only now we use the data for all four sites in the brain where measures were taken. If we label the typical measures for the control group as $\theta_{11}, \ldots, \theta_{14}$, and the typical measures for the murderers as $\theta_{21}, \ldots, \theta_{24}$, we have a 2-by-4, between-by-within design and a possible approach to comparing the groups is testing

$$H_0 : \theta_{11} - \theta_{21} = \theta_{12} - \theta_{22} = \theta_{13} - \theta_{23} = \theta_{14} - \theta_{24} = 0.$$

This can be done with the R function sppba with the argument avg set to F. If the data are stored in a matrix called eeg having eight columns, with the first four corresponding to the control group, then the command sppba(2,4,eeg,est=mom) performs the calculations based on the MOM measure of location and returns a significance level of 0.098. An alternative approach is to average the value of MOM over the four brain sites for each group, and then compare these averages. That is, test $H_0 : \bar{\theta}_{1.} = \bar{\theta}_{2.}$, where $\bar{\theta}_{j.} = \sum \theta_{jk}/4$. This can be done with the command sppba(2,4,eeg,avg=T). Now the *p*-value is .5 illustrating that the *p*-value can vary tremendously depending on how groups are compared.

■

8.6.8 Multiple Comparisons

When dealing with multiple comparisons associated with a between-by-within design, there are several approaches that might be taken that answer different questions. This section outlines some of the possibilities.

Method BWMCP

Focusing on trimmed means, multiple comparisons, when dealing with a between-by-within design, can be tested using linear contrasts, which are created in the same manner as outlined in Section 7.4.3. Consider any linear contrast Ψ, with the understanding that multiple linear contrasts are generally of interest. As usual, the goal is to test

$$H_0 : \Psi = 0,$$

and among the C linear contrasts of interest, often it is desired to have the probability of one or more type I errors equal to some specified value, α. Two methods for accomplishing this goal seem to be relatively effective: a bootstrap-t method and a percentile bootstrap method.

For convenience, we write the $L = JK$ trimmed means as $\bar{X}_{t1}, \ldots, \bar{X}_{tL}$. The estimate of

$$\Psi = \sum c_\ell \mu_{t\ell}$$

is

$$\hat{\Psi} = \sum c_\ell \bar{X}_{t\ell},$$

where $c_1, \ldots c_L$ are the linear contrast coefficients.

To test hypotheses using a bootstrap-t method, first note that the variances and covariances among the sample trimmed means can be estimated using results in Section 5.9.5. (Of course, when two sample trimmed means are independent, their covariance is taken to be zero.) Let $\mathbf{S}$ denote this L-by-L covariance matrix. (So the diagonal elements are the estimated squared standard errors.) Let $\mathbf{C}$ be a column matrix having length L that contains the contrast coefficients. Then the squared standard error of $\hat{\Psi}$ is estimated with

$$s^2_{\hat{\Psi}} = \mathbf{C}'\mathbf{S}\mathbf{C}$$

and an appropriate test statistic is

$$W = \frac{|\hat{\Psi}|}{s_{\hat{\Psi}}}.$$

A bootstrap-t method is used to estimate the null distribution of W. First, compute

$$Y_{ijk} = X_{ijk} - \bar{X}_{tjk},$$

where $\bar{X}_{tjk}$ is the trimmed mean corresponding to level j of Factor A and level k of Factor B. Next, take bootstrap samples based on the Y_{ijk} values. So for level j of Factor A, n_j rows of data are sampled with replacement. Based on this bootstrap sample, compute the test statistic W, which is labeled W^*. Repeat this process B times yielding $W_1^*, \ldots, W_B^*$. Let $c = (1-\alpha)B$, rounded to the nearest integer. Then a $1-\alpha$ confidence interval for Ψ is

$$\hat{\Psi} \pm W_{(c)}^* \frac{s_{\hat{\Psi}}}{\sqrt{n}}.$$

Alternatively, reject the null hypothesis if $W \geq W_{(c)}^*$.

Section 4.4.3 made a distinction between a symmetric bootstrap-t confidence interval and an equal-tailed confidence interval. Here, a symmetric confidence interval is used. For the situation at hand, there are no results on whether an equal-tailed confidence interval ever offers a practical advantage.

A percentile bootstrap method can be applied as well. As usual, no standard errors are used. For each hypothesis to be tested, corresponding to some linear contrast, a p-value can be computed as indicated at the end of Section 8.2.7.

Method BWAMCP: Comparing Levels of Factor A for Each Level of Factor B

To provide more detail about how groups differ, another strategy is to focus on a particular level of Factor B and perform all pairwise comparisons among the levels of Factor A. Of course, this can be done for each level of Factor B.

Example

Consider again a 3-by-2 design where the means are arranged as follows:

		Factor B	
		1	2
Factor A	1	μ_1	μ_2
	2	μ_3	μ_4
	3	μ_5	μ_6

For level 1 of Factor B, method BWAMCP would test H_0: $\mu_1 = \mu_3$, H_0: $\mu_1 = \mu_5$ and H_0: $\mu_3 = \mu_5$. For level 2 of Factor B, the goal is to test H_0: $\mu_2 = \mu_4$, H_0: $\mu_2 = \mu_6$ and H_0: $\mu_4 = \mu_6$. These hypotheses can be tested by creating the appropriate linear contrasts and using the R function lincon, which can be done with the R function bwamcp described in the next section.

■

Method BWBMCP: Dealing with Factor B

When dealing with Factor B, there are four variations of method BWMCP that might be used, which are described here under the appellation method BWBMCP. The first two variations ignore the levels of Factor A and test hypotheses based on the trimmed means. The first variation uses difference scores and the second uses the marginal trimmed means. Both of these variations begin by pooling the data over the levels of Factor A. In essence, Factor A is ignored. The other two variations do not pool the data over the levels of Factor A, but rather perform an analysis based on difference scores or the marginal trimmed means for each level of Factor A. In more formal terms, consider the jth level of Factor A. Then there are $(K^2 - K)/2$ pairs of groups that can be compared. If for each of the J levels of Factor A, all pairwise comparisons are performed, the total number of comparisons is $J(K^2 - K)/2$.

Example

Consider a 2-by-2 design where the first level of Factor A has 10 pairs of observations and the second has 15. So we have a total of 25 pairs of observations with the first 10 corresponding to level 1 of Factor A. When analyzing Factor B, pooling the data means the goal is to compare either the difference scores corresponding to all 25 pairs of observations, or to compare the marginal trimmed means, again based on all 25 observations. Not pooling means that for level 1 of Factor A either test hypotheses based on difference scores or compare the marginal trimmed means. And the same could be done for level 2 of Factor A.

■

Method BWIMCP: Interactions

As for interactions, we focus on a 2-by-2 design with the understanding that the same analysis can be done for any two levels of Factor A and any two levels of Factor B. Rather than define interactions as done when using Method BWMCP, difference scores might be used instead. To elaborate, consider the first level of Factor A. There are n_1 pairs of observations corresponding to the two levels of Factor B. Form the difference scores, which for level j of Factor A are denoted by

$$D_{ij},$$

$(i = 1, \ldots, n_j)$ and let μ_{tj} be the population trimmed means associated with these difference scores. Then one way of stating the hypothesis of no interaction is

$$H_0 : \mu_{t1} = \mu_{t2}.$$

In words, the hypothesis of no interaction corresponds to the trimmed means of the difference scores associated with level 1 of Factor A being equal to the trimmed means of the difference

scores associated with level 2 of Factor A. When either factor has more than two levels, a possible goal is to test all similar hypotheses (associated with any two levels of Factor A and Factor B) in a manner that controls FWE, which might be done using Rom's method or Hochberg's method.

Methods SPMCPA, SPMCPB, and SPMCPI

If it is desired to compare groups based on using a percentile bootstrap method, which appears to be the best method when comparing groups based on an M-estimator or MOM, analogs of methods BWAMCP, BWBMCP, and BWIMCP can be used, which are called methods SPMCPA, SPMCPB, and SPMCPI, respectively.

8.6.9 R Functions bwmcp, bwamcp, bwbmcp, bwimcp, spmcpa, spmcpb, and spmcpi

The R function

bwmcp(J, K, x, tr = 0.2, alpha = 0.05,con=0, nboot=599)

performs method BWMCP described in the previous section. By default, it creates all relevant linear contrasts for main effects and interactions by calling the R function con2way.

The R function

bwamcp(J, K, x, tr = 0.2, alpha = 0.05)

performs multiple comparisons associated with Factor A using the (bootstrap-t) method BWAMCP, described in the previous section. The function creates the appropriate set of linear contrasts and calls the R function lincon. The function returns three sets of results corresponding to Factor A, Factor B, and all interactions. The critical value reported for each of the three set of tests in designed to control the probability of at least one type I error.

The R function

spmcpa(J,K,x,est=tmean,JK=J*K,grp=c(1:JK),con=0,avg=F,alpha=.05, nboot=NA,pr=T, ...)

is like the R function bwamcp, only a percentile bootstrap method is used.

The R function

bwbmcp(J, K, x, tr = 0.2, con = 0, alpha = 0.05, dif = T,pool=F)

uses method BWBMCP to compare the levels of Factor B. If the argument pool=T, the function pools the data for you and then calls the function rmmcp. If the argument dif=F, the marginal trimmed means are compared instead. By default, pool=F meaning that

$$H_0 : \mu_{tjk} = \mu_{tjk'}$$

is tested for all $k < k'$ and $j = 1, \ldots, J$. For each level of Factor A, the function simply selects data associated with the levels of Factor B and tests hypotheses via the R function rmmcp. "Critical p-values" are reported in the column headed by p.crit. That is, p.crit indicates how small the p-value must be in order to reject, given the goal that FWE be equal to some specified α value. The R function

spmcpb(J,K,x,est=tmean,JK=J*K,grp=c(1:JK),dif=T,alpha=0.05, nboot=NA,pr=T, ...)

is like bwbmcp, only a percentile bootstrap method is used.

As for interactions, the R function

bwimcp(J, K, x, tr = 0.2, alpha = 0.05)

compares trimmed means using a nonbootstrap method. The R function

spmcpi(J,K,x,est=tmean,JK=J*K,grp=c(1:JK),alpha=0.05,nboot=NA, SEED=T,pr=T, ...)

uses a percentile bootstrap technique instead.

The R function

bwmcppb(J, K, x, tr = 0.2, alpha = 0.05, nboot = 500, bhop = F)

simultaneously performs all multiple comparisons related to all main effects and interactions using a percentile bootstrap method. Unlike spmcpa, spmcpb, and spmcpi, the function bwmcppb is designed for trimmed means only and has an option for using the Benjamini–Hochberg method via the argument bhop.

The R function

bwmcppb(J, K, x, tr = 0.2, alpha = 0.05, nboot = 500, bhop = F)

tests the same hypotheses as done by the R function bwmcp, only a percentile bootstrap method used. Unlike spmcpa, spmcpb, and spmcpi, the function bwmcppb is designed for trimmed means only and has an option for using the Benjamini–Hochberg method via the argument bhop.

8.6.10 Within-by-Within Designs

The methods for dealing with a between-by-within design are readily extended to a within-by-within design. That is, all JK groups being compared are dependent. For example, the method in Section 8.6.1 can be modified to handle this situation by taking $\mathbf{V}$ in Eq. (8.5) to be the Winsorized variance–covariance of all JK variables under study. (That is, for a between-by-within design, $\mathbf{V}$ was a block diagonal matrix, but for the situation at hand, generally this is no longer the case.) A similar extension can be used when dealing with linear

contrasts. Note that when dealing with linear contrasts, again there are two basic goals that might be of interest. The first is to test hypotheses about linear contrasts stated in terms of the measures of location associated with the marginal distributions. Section 8.1.3 provides explicit details when dealing with trimmed means that can be used to analyze a within-by-within design. The second strategy is to use an extension of methods based on difference scores. That is, now the hypotheses of interest take the form described in Section 8.1.4. R functions specifically designed for within-by-within design are described in the next section.

8.6.11 R Functions wwtrim, wwtrimbt, wwmcppb, and wwmcpbt

The R function

wwtrim(J, K, x, grp = c(1:p), p = J * K, tr = 0.2)

tests for main effects and interactions in a within-by-within design using a modification of the method for trimmed means described in Section 8.6.1. (The modification simply takes into account the possibility that all JK variables might be dependent.) The R function

wwtrimbt(J, K, x, tr = 0.2, JKL = J * K, grp = c(1:JK), nboot = 599, SEED = T, ...)

is the same as the R function wwtrim, only a bootstrap-t method is used. The R function

wwmcp(J,K,x,tr=0.2,alpha=0.05,dif=T)

performs multiple comparisons relevant to both main effects and interactions. (The function creates the appropriate linear contrasts and then uses the R function rmmcp.) By default, linear contrasts are created along the lines described in Section 8.1.4. To use linear contrasts based on the marginal trimmed means, set the argument dif=F. The R function

wwmcppb(J,K,x, alpha = 0.05, con = 0,est=tmean, plotit = F, dif = T, grp = NA, nboot = NA, BA = T, hoch = T, xlab = "Group 1", ylab = "Group 2", pr = T, SEED = T, ...)

is like the R function wwmcp, only a percentile bootstrap method is used. It defaults to using a 20% trimmed mean, but other measures of location can be used via the argument est. (When using an M-estimator, setting the argument hoch=F is suggested.) This function (using default settings) appears to be a relatively good choice, particularly when dealing with a small sample size. When the amount of trimming is small, use the R function

wwmcpbt(J,K,x, tr=0.2, alpha = 0.05, nboot = 599),

which uses a bootstrap-t method.

8.6.12 A Rank-Based Approach

This section describes a rank-based approach to a split-plot (or between-by-within subjects) design taken from Brunner et al. (2002, Chapter 8). There are other rank-based approaches

(e.g., Beasley, 2000; Beasley & Zumbo, 2003), but it seems that the practical merits of these competing methods, versus the method described here, have not been explored.

Main effects for Factor A are expressed in terms of

$$\bar{F}_{j.}(x) = \frac{1}{K}\sum_{k=1}^{K} F_{jk}(x),$$

the average of the distributions among the K levels of Factor B corresponding to the jth level of Factor A. The hypothesis of no main effects for Factor A is

$$H_0 : \bar{F}_{1.}(x) = \bar{F}_{2.}(x) = \cdots = \bar{F}_{J.}(x).$$

for any x. Letting

$$\bar{F}_{.k}(x) = \frac{1}{J}\sum_{j=1}^{J} F_{jk}(x)$$

be the average of the distributions for the kth level of Factor B, the hypothesis of no main effects for Factor B is

$$H_0 : \bar{F}_{.1}(x) = \bar{F}_{.2}(x) = \cdots = \bar{F}_{.K}(x).$$

As for interactions, first consider a 2-by-2 design. Then no interaction is taken to mean that for any x,

$$F_{11}(x) - F_{12}(x) = F_{21}(x) - F_{22}(x).$$

More generally, the hypothesis of no interactions among all JK groups is

$$H_0 : F_{jk}(x) - \bar{F}_{j.}(x) - \bar{F}_{.k}(x) + \bar{F}_{..}(x) = 0,$$

for any x, all j ($j = 1, \ldots, J$) and all k ($k = 1, \ldots, K$), where

$$\bar{F}_{..}(x) = \frac{1}{JK}\sum_{j=1}^{J}\sum_{k=1}^{K} F_{jk}(x).$$

As usual, let X_{ijk} represent the ith observation for level j of Factor A and level k of Factor B. Here, $i = 1, \ldots, n_j$. That is, the jth level of Factor A has n_j vectors of observations, each vector containing K values. So for the jth level of Factor A there are a total of $n_j K$ observations, and among all the groups, the total number of observations is denoted by N. So the total number of vectors among the J groups is $n = \sum n_j$, and the total number of observations is $N = K\sum n_j = Kn$.

Pool all N observations and assign ranks. As usual, midranks are used if there are tied values. Let R_{ijk} represent the rank associated with X_{ijk}. Let

$$\bar{R}_{.jk} = \frac{1}{n_j}\sum_{i=1}^{n_j} R_{ijk},$$

$$\bar{R}_{.j.} = \frac{1}{K}\sum_{k=1}^{K} \bar{R}_{.jk},$$

$$\bar{R}_{ij.} = \frac{1}{K}\sum_{k=1}^{K} R_{ijk},$$

$$\hat{\sigma}_j^2 = \frac{1}{n_j - 1}\sum_{i=1}^{n_j}(\bar{R}_{ij.} - \bar{R}_{.j.})^2,$$

$$S = \sum_{j=1}^{J}\frac{\hat{\sigma}_j^2}{n_j},$$

$$U = \sum_{j=1}^{J}\left(\frac{\hat{\sigma}_j^2}{n_j}\right)^2,$$

$$D = \sum_{j=1}^{J}\frac{1}{n_j - 1}\left(\frac{\hat{\sigma}_j^2}{n_j}\right)^2.$$

Factor A: The test statistic is

$$F_A = \frac{J}{(J-1)S}\sum_{j=1}^{J}(\bar{R}_{.j.} - \bar{R}_{...})^2,$$

where $\bar{R}_{...} = \sum \bar{R}_{.j.}/J$. The degrees of freedom are

$$\nu_1 = \frac{(J-1)^2}{1 + J(J-2)U/S^2},$$

and

$$\nu_2 = \frac{S^2}{D}.$$

Reject if $F_A \geq f$, where f is the $1-\alpha$ quantile of an F distribution with ν_1 and ν_2 degrees of freedom.

Factor B: Let

$$\mathbf{R}_{ij} = (R_{ij1}, \ldots, R_{ijK})',$$

$$\bar{\mathbf{R}}_{\cdot j} = \frac{1}{n_j}\sum_{i=1}^{n_j}\mathbf{R}_{ij},\ \bar{\mathbf{R}}_{\cdot\cdot} = \frac{1}{J}\sum_{j=1}^{J}\bar{\mathbf{R}}_{\cdot j},$$

$n = \sum n_j$ (so $N = nK$),

$$\mathbf{V}_j = \frac{n}{N^2 n_j(n_j-1)}\sum_{i=1}^{n_j}(\mathbf{R}_{ij} - \bar{\mathbf{R}}_{\cdot j})(\mathbf{R}_{ij} - \bar{\mathbf{R}}_{\cdot j})'.$$

So $\mathbf{V}_j$ is a K-by-K matrix of covariances based on the ranks. Let

$$\mathbf{S} = \frac{1}{J^2}\sum_{j=1}^{J}\mathbf{V}_j$$

and let $\mathbf{P}_K$ be defined as in Section 7.9. The test statistic is

$$F_B = \frac{n}{N^2\text{tr}(\mathbf{P}_K\mathbf{S})}\sum_{k=1}^{K}(\bar{R}_{\cdot\cdot k} - \bar{R}_{\cdot\cdot\cdot})^2.$$

The degrees of freedom are

$$\nu_1 = \frac{(\text{tr}(\mathbf{P}_K\mathbf{S}))^2}{\text{tr}(\mathbf{P}_K\mathbf{S}\mathbf{P}_K\mathbf{S})},\ \nu_2 = \infty,$$

and H_0 is rejected if $F_B \geq f$, where f is the $1-\alpha$ quantile of an F distribution with ν_1 and ν_2 degrees of freedom.

Interactions: Let $\mathbf{V}$ be the block diagonal matrix based on the matrices $\mathbf{V}_j$, $j = 1, \ldots, J$. Letting $\mathbf{M}_{AB}$ be defined as in Section 7.9, the test statistic is

$$F_{AB} = \frac{n}{N^2\text{tr}(\mathbf{M}_{AB}\mathbf{V})}\sum_{j=1}^{J}\sum_{k=1}^{K}(\bar{R}_{\cdot jk} - \bar{R}_{\cdot j\cdot} - \bar{R}_{\cdot\cdot k} + \bar{R}_{\cdot\cdot\cdot})^2.$$

The degrees of freedom are

$$\nu_1 = \frac{(\text{tr}(\mathbf{M}_{AB}\mathbf{V}))^2}{\text{tr}(\mathbf{M}_{AB}\mathbf{V}\mathbf{M}_{AB}\mathbf{V})},\ \nu_2 = \infty.$$

Reject if $F_A \geq f$ (or if $F_{AB} \geq f$), where f is the $1-\alpha$ quantile of an F distribution with ν_1 and ν_2 degrees of freedom.

8.6.13 R Function bwrank

The R function

bwrank(J,K,x)

performs a between-by-within ANOVA based on ranks using the method just described. In addition to testing hypotheses as just indicated, the function returns the average ranks ($\bar{R}_{.jk}$) associated with all JK groups as well as the relative effects, $(\bar{R}_{.jk} - 0.5)/N$.

Example

Lumley (1996) reports data on shoulder pain after surgery; the data are from a study by Jorgensen, Gilles, Hunt, Caplehorn, and Lumley (1995). Table 8.7 shows a portion of

Table 8.7: Shoulder Pain Data (1=low, 5=high).

Active Treatment			No Active Treatment		
Time 1	Time 2	Time 3	Time 1	Time 2	Time 3
1	1	1	5	2	3
3	2	1	1	5	3
3	2	2	4	4	4
1	1	1	4	4	4
1	1	1	2	3	4
1	2	1	3	4	3
3	2	1	3	3	4
2	2	1	1	1	1
1	1	1	1	1	1
3	1	1	1	5	5
1	1	1	1	3	2
2	1	1	2	2	3
1	2	2	2	2	1
3	1	1	1	1	1
2	1	1	1	1	1
1	1	1	5	5	5
1	1	1	3	3	3
2	1	1	5	4	4
4	4	2	1	3	3
4	4	4			
1	1	1			
1	1	1			

the results where two treatment methods are used and measures of pain are taken at three different times. The output from bwrank is

```
$test.A:
[1] 12.87017

$sig.A:
[1] 0.001043705

$test.B:
[1] 0.4604075

$sig.B:
[1] 0.5759393

$test.AB:
[1] 8.621151

$sig.AB:
[1] 0.0007548441

$avg.ranks:
         [,1]     [,2]     [,3]
[1,] 58.29545 48.40909 39.45455
[2,] 66.70455 82.36364 83.04545

$rel.effects:
          [,1]      [,2]      [,3]
[1,] 0.4698817 0.3895048 0.3167036
[2,] 0.5382483 0.6655580 0.6711013
```

So at approximately the 0.001 level, treatment methods are significantly different and there is a significant interaction, but no significant difference is found over time. Note that the average ranks and relative effects suggest that a disordinal interaction might exist. In particular, for group 1 (the active treatment group), time 1 has higher average ranks versus time 2, and the reverse is true for the second group. However, the Wilcoxon signed rank test fails to reject at the 0.05 level when comparing time 1 to time 2 for both groups. When comparing time 1 versus time 3 for the first group, again using the Wilcoxon signed rank test, we reject at the 0.05 level, but a nonsignificant result is obtained for group 2. So again a disordinal interaction appears to be a possibility, but the empirical evidence is not compelling.

■

■ Example

Section 6.11 illustrated a method for comparing multivariate data corresponding to two independent groups based on the extent that points from one group are nested within

the other. For the data in Table 6.6, it was found that schizophrenics differed from the control group; also see Figure 6.10. If the two groups are compared based on the OP estimator (using the function smean2), again the two groups are found to differ. Comparing the groups with the method for means and trimmed means described in this section, no difference between the schizophrenics and control group is found at the 0.05 level. Using the rank-based method in this section, again no difference is found. (The p-value is .11.) The only point is that how we compare groups can make a practical difference about the conclusions reached.

■

8.6.14 Rank-Based Multiple Comparisons

Multiple comparisons based on the rank-based methods covered here can be performed using simple combinations of methods already considered. When dealing with Factor A, for example, one can simply compare level j to level j', ignoring the other levels. When comparing all pairs of groups, FWE can be controlled with Rom's method or the Benjamini–Hochberg technique. Factor B and the collection of all interactions (corresponding to any two rows and any two columns) can be handled in a similar manner.

8.6.15 R Function bwrmcp

The R function

bwrmcp(J,K,x,grp=NA,alpha=0.05,bhop=F)

performs all pairwise multiple comparisons using the method of Section 8.6.14 with the FWE (familywise error) rate controlled using Rom's method of the Benjamini–Hochberg method. For example, when dealing with Factor A, the function simply compares level j to level j' ignoring the other levels. All pairwise comparisons among the J levels of Factor A are performed and the same is done for Factor B and all relevant interactions.

8.6.16 Multiple Comparisons when Using a Patel–Hoel Approach to Interactions

Rather than compare distributions when dealing with a between-by-within design, one could use a simple analog of the Patel–Hoel approach instead. First consider a 2-by-2 design and focus on level one of Factor A. Then the two levels of Factor B are dependent and can be compared with the sign test. In essence, inferences are being made about p_1, the probability that for a randomly sampled pair of observations, the observation from level one of Factor B is less than the corresponding observation from level two. Of course, for level two of Factor A, we can again compare levels one and two of Factor B with the sign test. Now we let p_2 be

the probability that for a randomly sampled pair of observations, the observation from level one of Factor B is less than the corresponding observation from level two. Then no interaction can be defined as $p_1 = p_2$.

The hypothesis of no interaction,

$$H_0 : p_1 = p_2,$$

is just the hypothesis that two independent binomials have equal probabilities of success, which can be tested using one of the methods described in Section 5.8. Here, Beal's method is used rather than the Storer–Kim method because it currently seems that Beal's method provides more accurate control over FWE for the problem at hand, execution time can be much lower when sample sizes are large, and unlike the Storer–Kim procedure, Beal's method provides confidence intervals. Method KMS in Section 5.8.3 might be used as well, but there are no published results on how it performs for the situation at hand.

There are various ways FWE might be controlled. Among a collection of techniques considered by Wilcox (2001c), the following method was found to be relatively effective. Let q be the $1-\alpha$ quantile of a C-variate Studentized maximum modulus distribution with degrees of freedom $\nu = \infty$, where C is the total number of hypotheses to be tested. Assuming that all pairs of rows and columns are to be considered when testing the hypothesis of no interactions,

$$C = \frac{J^2 - J}{2} \times \frac{K^2 - K}{2}.$$

(For a formal definition of a Studentized maximum modulus distribution, see Miller, 1966, p. 71. Some quantiles are reported in Wilcox, 2003a.) Let Z be a standard normal random variable. Then if FWE is to be α, test each of the C hypotheses at the α_a level where

- If $(J,\ K) = (5,\ 2)$, then $\alpha_a = 2[1 - P(Z \leq q)]$.
- If $(J,\ K) = (3, 2)$, $(4, 2)$ or $(2, 3)$, then $\alpha_a = 3[1 - P(Z \leq q)]$.
- For all other J and K values, $\alpha_a = 4[1 - P(Z \leq q)]$.

These adjusted α values appear to work well when the goal is to achieve FWE less than or equal to 0.05. Whether this remains the case with FWE equal to 0.01 is unknown. For $C > 28$ and FWE equal to 0.05, use

$$q = 2.383904C^{1/10} - 0.202.$$

(Of course, for $C = 1$, no adjustment is necessary; simply use Beal's method.)

Tied values are handled in the same manner as with the signed rank test: pairs of observations with identical values are simply discarded. So among the remaining observations, for every pair of observations, the observation from level one of Factor B, for example, is either less than or greater than the corresponding value from level two.

A criticism of this method is that power can be relatively low. However, it directly addresses an issue that might be deemed interesting and useful that is not directly addressed by other methods in this chapter.

A variation of the approach in this section is where, for level one of Factor B, p_1 is the probability that an observation from level one of Factor A is less than an observation from level 2. Similarly, p_2 is now defined in terms of the two levels of Factor A when working with level two of Factor B. However, the details of how to implement this approach have not been studied.

8.6.17 R Function sisplit

The method just described for interactions can be applied with the R function

sisplit(J,K,x)

This function assumes $\alpha = 0.05$; other values are not allowed. As usual, x is any R variable containing the data that is an n-by-JK matrix or has list mode.

8.7 Some Rank-Based Multivariate Methods

This section describes two rank-based methods for comparing J independent groups with K measures associated with each group.

8.7.1 The Munzel–Brunner Method

The first method was derived by Munzel and Brunner (2000). (For recent results regarding how the Munzel–Brunner method compares to several techniques not covered here, see Bathke, Solomon, & Madden, 2008. A variation of the Munzel–Brunner method can be used in place of the Agresti–Pendergast method, but the relative merits of these two techniques have not been explored.) Let n_j represent the number of randomly sampled vectors from the jth group, each vector containing K measures. Let $F_{jk}(x)$ be the distribution associated with the jth group and kth measure. So for example, $F_{32}(6)$ is the probability that for the third group, the second variable will be less than or equal to 6 for a randomly sampled individual. For the kth measure, the goal is to test the hypothesis that all J groups have identical distributions. And the more general goal is to test the hypothesis that simultaneously, all groups have identical distributions for each of the K measures under consideration. That is, the goal is to test

$$H_0 : F_{1k}(x) = \cdots = F_{Jk}(x) \text{ for all } k = 1, \ldots, K. \tag{8.9}$$

To apply the method, begin with the first of the K measures, pool all the observations among the J groups and assign ranks. Ties are handled in the manner described in Section 5.7.2. Repeat this process for all K measures and label the results R_{ijk}. That is, R_{ijk} is the rank of the ith observation in the jth group and for the kth measure. Let

$$\bar{R}_{jk} = \frac{1}{n_j}\sum_{i=1}^{n_j} R_{ijk},$$

be the average rank for the jth group corresponding to the kth measure. Set

$$\hat{Q}_{jk} = \frac{\bar{R}_{jk} - 0.5}{n},$$

where $n = \sum n_j$ is the total number of randomly sampled vectors among the J groups. The remaining calculations are summarized in Table 8.8. The $\hat{Q}$ values are called the *relative effects* and reflect the ordering of the average ranks. If, for example, $\hat{Q}_{11} < \hat{Q}_{21}$, the typical rank for variable one in group one is less than the typical rank for variable one in group two. More generally, if $\hat{Q}_{jk} < \hat{Q}_{j'k}$, then based on the kth measure, the typical rank (or observed value) for group j is less than the typical rank for group j'.

Table 8.8: The Munzel–Brunner One-Way Multivariate Method.

Let

$$\hat{\mathbf{Q}} = (\hat{Q}_{11}, \hat{Q}_{12}, \ldots, \hat{Q}_{1K}, \hat{Q}_{21}, \ldots, \hat{Q}_{JK})',$$

$\mathbf{R}_{ij} = (R_{ij1}, \ldots, R_{ijK})'$, $\bar{\mathbf{R}}_j = (\bar{R}_{j1}, \ldots, \bar{R}_{jK})'$,

$$\mathbf{V}_j = \frac{1}{nn_j(n_j-1)} = \sum_{i=1}^{n_j}(\mathbf{R}_{ij} - \bar{\mathbf{R}}_j)(\mathbf{R}_{ij} - \bar{\mathbf{R}}_j)',$$

$n = \sum n_j$ and let

$$\mathbf{V} = \text{diag}\{\mathbf{V}_1, \ldots, \mathbf{V}_J\}.$$

Compute the matrix $\mathbf{M}_A$ as described in Section 7.9. The test statistic is

$$F = \frac{n}{\text{tr}(\mathbf{M}_A\mathbf{V})}\hat{\mathbf{Q}}'M_A\hat{\mathbf{Q}}.$$

Decision Rule: Reject if $F \geq f$, where f is the $1-\alpha$ quantile of an F distribution with

$$\nu_1 = \frac{(\text{tr}(\mathbf{M}_A\mathbf{V}))^2}{\text{tr}(\mathbf{M}_A\mathbf{V}\mathbf{M}_A\mathbf{V})},$$

and $\nu_2 = \infty$ degrees of freedom.

8.7.2 R Function mulrank

The R function

mulrank(J, K, x)

performs the one-way multivariate method in Table 8.8. The data are stored in x which can be a matrix or have list mode. If x is a matrix, the first K columns correspond to the K measures for group 1, the second K correspond to group 2, and so forth. If stored in list mode, x[[1]], . . . , x[[K]] contain the data for group 1, x[[K+1]], . . ., x[[2K]] contain the data for group 2, and so on.

Example

Table 8.9 summarizes data (reported by Munzel & Brunner, 2000) from a psychiatric clinical trial where three methods are compared for treating individuals with panic disorder. The three methods are exercise, clomipramine and a placebo. The two measures of effectiveness were a clinical global impression (CGI) and the patient's global impression (PGI). The test statistic is $F = 12.7$ with $\nu_1 = 2.83$ and a significance level less than 0.001. The relative effects are:

```
$q.hat:
          [,1]      [,2]
[1,] 0.5074074 0.5096296
[2,] 0.2859259 0.2837037
[3,] 0.7066667 0.7066667
```

Table 8.9: CGI and PGI Scores After Four Weeks of Treatment.

Exercise		Clomipramine		Placebo	
CGI	PGI	CGI	PGI	CGI	PGI
4	3	1	2	5	4
1	1	1	1	5	5
2	2	2	0	5	6
2	3	2	1	5	4
2	3	2	3	2	6
1	2	2	3	4	6
3	3	3	4	1	1
2	3	1	4	4	5
5	5	1	1	2	1
2	2	2	0	4	4
5	5	2	3	5	5
2	4	1	0	4	4
2	1	1	1	5	4
2	4	1	1	5	4
6	5	2	1	3	4

So among the three groups, the second group, clomipramine, has the lowest relative effects. That is, the typical ranks were lowest for this group, and the placebo group had the highest ranks on average.

■

8.7.3 The Choi–Marden Multivariate Rank Test

This section describes a multivariate analog of the Kruskal–Wallis test derived by Choi and Marden (1997). There are actually many variations of the approach they considered, but here attention is restricted to the version they focused on. As with the method in Section 8.7.1, we have K measures for each individual and there are J independent groups. For the jth group and any vector of constants $\mathbf{x} = (x_1, \ldots, x_K)$, let

$$F_j(\mathbf{x}) = P(X_{j1} \leq x_1, \ldots, X_{jK} \leq x_K).$$

So for example, $F_1(\mathbf{x})$ is the probability that for the first group, the first of the K measures is less than or equal to x_1, the second of the K measures is less than or equal to x_2, and so forth. The null hypothesis is that for any $\mathbf{x}$,

$$H_0 : F_1(\mathbf{x}) = \cdots = F_J(\mathbf{x}), \tag{8.10}$$

which is sometimes called the *multivariate hypothesis* to distinguish it from Eq. (8.8), which is called the *marginal hypothesis*. The multivariate hypothesis is a stronger hypothesis in the sense that if it is true, then by implication the marginal hypothesis is true as well. For example, if the marginal distributions for both groups are standard normal distributions, the marginal hypothesis is true, but if the groups have different correlations, the multivariate hypothesis is false.

The Choi–Marden method represents an extension of a technique derived by Möttönen & Oja (1995) and is based on a generalization of the notion of a rank to multivariate data which was also used by Chaudhuri (1996, Section 4). First consider a random sample of n observations with K measures for each individual or thing and denote the ith vector of observations by

$$\mathbf{X}_i = (X_{i1}, \ldots, X_{iK}).$$

Let

$$A_{ii'} = \sqrt{\sum_{k=1}^{K} (X_{ik} - X_{i',k})^2},$$

Table 8.10: The Choi–Marden Method.

Pool the data from all J groups and compute rank vectors as just described in the text. The resulting rank vectors are denoted by $\mathbf{R}_1, \ldots, \mathbf{R}_n$, where $n = \sum n_j$ is the total number of vectors among the J groups. For each of the J groups, average the rank vectors and denote the average of these vectors for the jth group by $\bar{\mathbf{R}}_j$.

Next, assign ranks to the vectors in the jth group, ignoring all other groups. We let $\mathbf{V}_{ij}$ (a column vector of length K) represent the rank vector corresponding to the ith vector of the jth group ($i = 1, \ldots, n_j$; $j = 1, \ldots, J$) to make a clear distinction with the ranks based on the pooled data. Compute

$$\mathbf{S} = \frac{1}{n - J} \sum_{j=1}^{J} \sum_{i=1}^{n_j} \mathbf{V}_{ij} \mathbf{V}_{ij}',$$

where $\mathbf{V}_{ij}'$ is the transpose of $\mathbf{V}_{ij}$ (so $\mathbf{S}$ is a K-by-K matrix). The test statistic is

$$H = \sum_{j=1}^{J} n_j \bar{\mathbf{R}}_j' \mathbf{S}^{-1} \bar{\mathbf{R}}_j. \tag{8.11}$$

(For $K = 1$, H does not quite reduce to the Kruskal–Wallis test statistic. In fact, H avoids a certain technical problem that is not addressed by the Kruskal–Wallis method.)

Decisions Rule: Reject if $H \geq c$, where c is the $1 - \alpha$ quantile of a chi-squared distribution with degrees of freedom $K(J - 1)$.

Here, the "rank" of the ith vector is itself a vector (having length K) given by

$$\mathbf{R}_i = \frac{1}{n} \sum_{i'=1}^{n} \frac{\mathbf{X}_i - \mathbf{X}_{i'}}{A_{ii'}},$$

where

$$\mathbf{X}_i - \mathbf{X}_{i'} = (X_{i1} - X_{i'1}, \ldots, X_{iK} - X_{i'K}).$$

The remaining calculations are summarized in Table 8.10. All indications are that this method provides good control over the probability of a type I error when ties never occur. There are no known problems when there are tied values, but this issue is in need of more research.

8.7.4 R Function cmanova

The R function

```
cmanova(J,K,x)
```

performs the Choi–Marden method just described. The data are assumed to be stored in x as described in Section 8.6.2.

8.8 Three-Way Designs

Generally, two-way designs can be extended to a three-way design where one or more factors involve dependent groups. This section outlines some of the methods that might be used. It is stressed, however, that simulation studies reporting the relative merits of the methods considered are extremely limited.

8.8.1 Global Tests Based on Trimmed Means

The method in Section 8.6.1 is readily extended to a three-way design. Note that the matrix $\mathbf{V}$ used in Eq. (8.4) reflects the variances and covariances among the trimmed means where the covariances are taken to be zero if the groups are independent. Here, $\mathbf{V}$ is computed in a similar manner. That is, for a J-by-K-by-L design, $\mathbf{V}$ is a JKL square matrix that contains the squared standard errors and covariances among the sample trimmed means, with independent trimmed means having a covariance of zero. Once $\mathbf{V}$ is available, compute the test statistic Q given by Eq. (8.4), where now the matrix $\mathbf{C}$ is computed as described in Table 7.5.

When dealing with situations where one or more factors involve dependent groups, comments should be made regarding the approximation of the null distribution using an F distribution. Unlike the method in Section 8.6.1 where the second degree of freedom, ν_2, is estimated based on the data, the strategy here is to simply set $\nu_2 = 999$. The reason is that even with $\nu_2 = 999$, the actual level of the method can drop well below the nominal level when the sample size is small and 20% trimmed means are used. For example, when dealing with a between-by-between-by-within design, under normality with all correlations equal to zero and $J = 2$, $K = L = 3$, and $n = 25$, the actual type I error probability is approximately .003 when dealing with the A-by-C interaction and testing at the 0.05 level. Increasing n to 50, now the actual level is approximately 0.013, for $n = 100$ it is 0.037, and for $n = 900$ it is 0.04. For the main effect associated with factor A, the estimated level is 0.050 with $n = 25$ and 0.052 with $n = 900$. A bootstrap-t method appears to suffer from the same problem, but an extensive study of this issue has not been conducted. A similar problem occurs when dealing with a between-by-within-by-within design. Again $\nu_2 = 999$ is used, but even with $n = 100$, the actual level can drop as low as 0.025 under normality. Using instead a bootstrap-t method (via the R function bbwtrimbt), with $n = 25$, the actual level was estimated to be 0.067.

Evidently there are no published studies comparing methods, in terms of type I errors, when dealing with three-way designs with one or more within group factors. Very limited results suggest that perhaps a better approach, compared to the methods described here, is to use the R functions in Section 8.8.6. They test hypotheses about all of the usual linear contrasts associated with a three-way design using a percentile bootstrap method in conjunction with a trimmed mean. In terms of controlling the probability of one or more type I errors, performing

one of the global tests described here is not required when using the percentile bootstrap methods via the R functions in Section 8.8.6. Moreover, limited results suggest that control over the type I error probability is more satisfactory when the amount of trimming is 20%. For the situation considered here, where $n = 25$ and the actual type I error is approximately .003, the probability of one or more type I errors was estimated to be .039 when using a percentile bootstrap method via the R function bbwmcppb, based on a simulation with 1000 replications. (And execution time can be substantially less when using a percentile bootstrap method rather than the bootstrap-t method.) Using instead the R function bwwmcppb, the probability of one or more type I errors was estimated to be .049. But again, a more comprehensive study is needed.

8.8.2 R Functions bbwtrim, bwwtrim, wwwtrim, bbwtrimbt, bwwtrimbt, and wwwtrimbt

The R function

bbwtrim(J,K,L,x,grp=c(1:p),tr=0.2)

tests all omnibus main effects and interactions associated with a between-by-between-by-within design. The data are assumed to be stored as described in Section 7.3.1. For a between-by-within-by-within design use

bwwtrim(J,K,L,x,grp=c(1:p),tr=0.2).

And for a within-by-within-by-within design use

wwwtrim(J,K,L,x,grp=c(1:p),tr=0.2).

The R functions

bbwtrimbt(J,K,L,x,grp=c(1:p),tr=0.2, nboot = 599, SEED = T)

bwwtrim(J,K,L,x,grp=c(1:p),tr=0.2, nboot = 599, SEED = T).

and

wwwtrimbt(J,K,L,x,grp=c(1:p),tr=0.2, nboot = 599, SEED = T).

are the same as the functions bbwtrim, bwwtrim, and wwwtrim, respectively, only a bootstrap-t method is used.

8.8.3 Data Management: R Functions bw2list and bbw2list

For a between-by-within-by-within design, the R function

bw2list(x, grp.col, lev.col),

which was introduced in Section 8.6.3, can be used when dealing with data that are stored in a matrix or a data frame with one column indicating the levels of the independent groups and other columns containing data corresponding to within group levels. For example, setting the argument grp.col=c(5) would indicate that the levels for Factor A are stored in column 5 and lev.col=c(3,9,10,12) indicates that the within levels data are stored in columns 3, 9, 10, and 12. Note that it must be the case that KL is equal to the number of values stored in lev.col. So lev.col=c(3,9,10,12) would be appropriate if the within factors have two levels each, with the data for two levels of Factor C being stored in columns 10 and 12.

The R function

bbw2list(x, grp.col, lev.col),

deals with a between-by-between-by-within design and assumes that the argument grp.col contains two values that indicate the columns of x that indicate the levels of Factors A and B. Now the argument lev.col indicates the columns containing the within data.

Example

Imagine that for a between-by-between-by-within design, column 14 of the R variable dis contains values indicating the levels of Factor A, column 10 has values that contain the levels of Factor B, and columns 2, 4, and 9 contain the outcomes values at times 1, 2, and 3, respectively. Then

z=bbw2list(dis, grp.col=c(14,10), lev.col=c(2,4,9))

would store the data in list mode in z, after which the command

bbwtrim(3,4,3,z)

would test the usual hypotheses, assuming that Factors A and B have 3 and 4 levels, respectively. The values in column 14 and 10 would be sorted in ascending order, or in alphabetical order if the values in these columns are character data.

■

8.8.4 Multiple Comparisons

Multiple comparisons in a three-way design can be performed using a straightforward extension of methods described in previous sections. The R function con3way, described in Section 7.4.4, can be used to generate the linear contrast coefficients that are often used. Here, when computing A in Section 8.1.3, we set $d_{jk} = 0$ whenever j and k correspond to independent groups. Otherwise, this term is computed as described in Section 8.1.3. The next two sections summarize some R functions aimed at facilitating the analysis.

8.8.5 R Function rm3mcp

When dealing with a within-by-within-by-within design, a nonbootstrap method can be used to test the hypotheses associated with all of the linear contrasts generated by the R function con3way. This can be done with the R function

rm3mcp(J, K, L, x, tr = 0.2, alpha = 0.05, dif = T, grp = NA).

(That is, it uses the R function con3way to generate the linear contrast coefficients and then it tests the corresponding hypotheses.) When dealing with designs where there are both between and within factors, use a bootstrap method via one of the R functions described in the next section. Another approach is to use the R function rmmcp in Section 8.1.5 in conjunction with the R function con3way. (For an illustration of how to interpret three-way interactions based on the contrast coefficients returned by con3way, see the example at the end of Section 7.4.4.)

8.8.6 R Functions bbwmcp, bwwmcp, bbwmcppb, bwwmcppb, and wwwmcppb

Bootstrap-t Methods

The R function

bbwmcp(J, K, L, x, tr = 0.2, JKL = J * K * L, con = 0, alpha = 0.05, grp = c(1:JKL), nboot = 599, SEED = T, ...)

performs all multiple comparisons associated with main effects and interactions using a bootstrap-t method in conjunction with trimmed means when analyzing a between-by-between-by-within design. The function uses con3way to generate all of the relevant linear contrasts and then uses the function lindep to test the hypotheses. The critical value is designed to control the probability of at least one type I error among all the linear contrasts associated with factor A. The same is done for factor B and factor C.

The R function

bwwmcp(J, K, L, x, tr = 0.2, JKL = J * K * L, con = 0, alpha = 0.05, grp = c(1:JKL), nboot = 599, SEED = T, ...)

handles a between-by-within-by-within design.

Percentile Bootstrap Methods

For a between-by-between-by-within design, the R function

bbwmcppb(J, K, L, x, tr = 0.2, JKL = J * K * L, con = 0, alpha = 0.05, grp = c(1:JKL), nboot = 599, SEED = T, ...)

tests hypotheses using a percentile bootstrap method. As for a between-by-within-by-within and within-by-within-by-within design, use the functions

bwwmcppb(J, K, L, x, tr = 0.2, JKL = J * K * L, con = 0, alpha = 0.05, grp = c(1:JKL), nboot = 599, SEED = T, ...)

and

wwwmcppb(J, K, L, x, tr = 0.2, JKL = J * K * L, con = 0, alpha = 0.05, grp = c(1:JKL), nboot = 599, SEED = T, ...),

respectively.

8.9 Exercises

1. Section 8.6.2 reports data on hangover symptoms. For group 2, use the R function rmanova to compare the trimmed means corresponding to times 1, 2, and 3.
2. For the data used in Exercise 1, compute confidence intervals for all pairs of trimmed means using the R function pairdepb.
3. Analyze the data for the control group reported in Table 6.1 using the methods in Sections 8.1 and 8.2. Compare and contrast the results.
4. Repeat Exercise 3 using the rank-based method in Section 8.5. How do the results compare to using a measure of location?
5. Repeat Exercises 3 and 4 using the data for the murderers in Table 6.1.
6. Analyze the data in Table 6.1 using the methods in Sections 8.6.1 and 8.6.4.
7. Repeat Exercise 6, only now use the rank-based method in Section 8.6.12.