```
D-A    0  -4.05604   4.0560 1.00000
C-B    2  -1.82407   5.8241 0.47660
D-B   -5  -8.57709 -1.4229 0.00441
D-C   -7 -10.57709 -3.4229 0.00013
```

We find that only the $A-D$ and $B-C$ differences are not significant as the corresponding intervals contain zero. The duality between confidence intervals and hypothesis testing can also be used to produce an adjusted p-value as seen in the final column of the output. This merely confirms our impression of which differences are significant. The intervals can be plotted as seen in Figure 15.3.

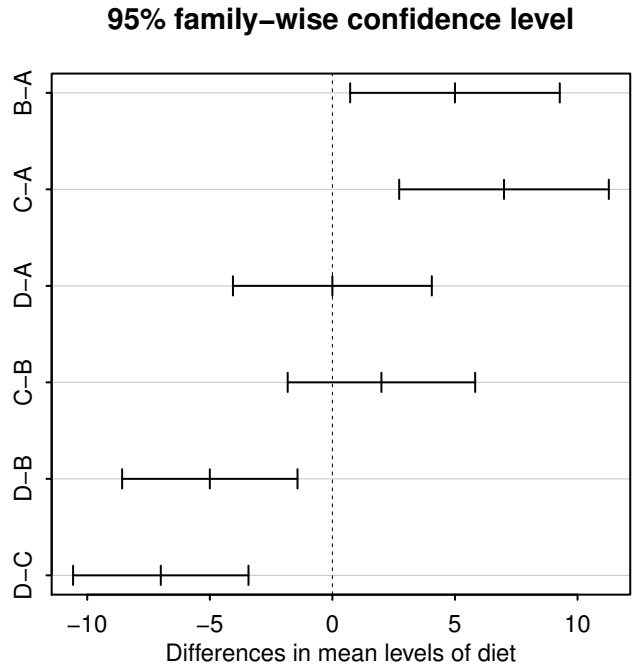

Figure 15.3 Tukey HSD 95% confidence intervals for the pairwise differences.

```
> plot(tci)
```

The Tukey method assumes the worst by focusing on the largest difference. There are other competitors like the Newman–Keuls, Duncan's multiple range and the Waller–Duncan procedure, which are less pessimistic or do not consider all possible pairwise comparisons. For a detailed description of the many available alternatives see Bretz, Hothorn, and Westfall (2010) and the R package `multcomp`.

15.5 False Discovery Rate

Consider some data taken from the Junior School Project collected from primary (US term is elementary) schools in inner London. The data is described in detail in

Mortimore et al. (1988). We focus on just two of the variables in the data — the `school`, of which there are 49, and the `math`ematics test scores for students from these schools. Suppose we are interested in deviations from the average and so we center the scores:

```
> data(jsp, package="faraway")
> jsp$mathcent <- jsp$math - mean(jsp$math)
```

A more pleasing plot of the data can be obtained using the `ggplot2` package as seen in Figure 15.4. We have to rotate the school labels so they can be distinguished.

```
> require(ggplot2)
> ggplot(aes(x=school,y=mathcent),data=jsp)+geom_boxplot() + theme(
    axis.text.x = element_text(angle = 90))+ylab("Centered Math
    Scores")
```

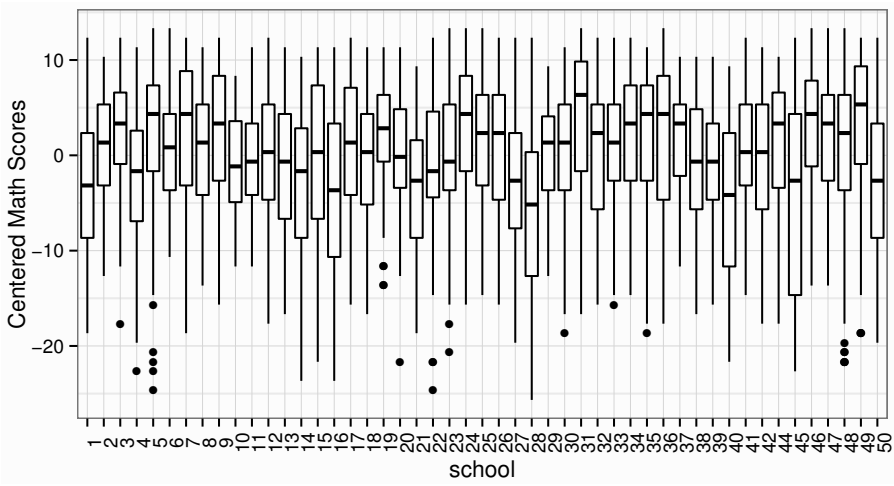

Figure 15.4 Variation-centered math scores by school.

Let's choose the parameterization that omits the intercept term:

```
> lmod <- lm(mathcent ~ school-1, jsp)
> sumary(lmod)
          Estimate Std. Error t value Pr(>|t|)
school1    -3.3685     0.7686   -4.38  1.2e-05
school2     0.6714     1.2287    0.55  0.58481
...
school50   -2.6520     0.7336   -3.62  0.00030

n = 3236, p = 49, Residual SE = 7.372, R-Squared = 0.08
```

Since we have centered the response, the t-tests, which check for differences from zero, are meaningful. We can see there is good evidence that schools 1 and 50 are significantly below average while the evidence that school 2 is above average is not statistically significant. We can test for a difference between the schools:

```
> anova(lm(mathcent ~ school, jsp))
Analysis of Variance Table

Response: mathcent
            Df Sum Sq Mean Sq F value Pr(>F)
school      48  15484     323    5.94 <2e-16
Residuals 3187 173212      54
```

We find a strongly significant difference. This comes as little surprise as the sample size is large, giving us the power to detect quite small differences. Furthermore, we may have strong prior reasons to expect differences between the schools. A more interesting question is which schools show clear evidence of under- or over-performance?

There are too many pairwise comparisons on which to focus our interest. Instead let us ask which schools have means significantly different from the average. The parameterization we have chosen makes these comparisons easy but we would expect about 5% of these differences to be significant even if the null hypothesis held.

Some adjustment is necessary. One approach is to control the *familywise error rate* (FWER) which is the overall probability of falsely declaring a difference (where none exists). The *Bonferroni correction* is a simple way to do this. We just multiply the unadjusted p-values by the number of comparisons. Any probability computed above one is truncated to one. Let's see which schools have adjusted p-values less than 5%:

```
> pvals <- summary(lmod)$coef[,4]
> padj <- p.adjust(pvals, method="bonferroni")
> coef(lmod)[padj < 0.05]
 school1 school16 school21 school28 school31 school40 school45
 -3.3685  -3.7374  -3.6185  -5.8286   3.8241  -4.9855  -4.6392
school50
 -2.6520
```

We see that eight schools are identified with all except school 31 marked as significantly below average.

The Bonferroni correction is known to be conservative but even were we to use one of the more generous alternatives, the familywise error rate restriction imposes a high bar on the identification of significant effects. As the number of levels being compared increases, this requirement becomes ever more stringent.

An alternative approach is to control the *false discovery rate* (FDR) which is the proportion of effects identified as significant which are not real. The best known method of doing this is due to Benjamini and Hochberg (1995).

Given sorted p-values $p_{(i)}$ for $i = 1,\ldots,m$ the procedure finds the largest i for which $p_{(i)} \leq \alpha i/m$. All tests corresponding to $p_{(i)}$ up to and including this i are declared significant. We compute this for our example:

```
> names(which(sort(pvals) < (1:49)*0.05/49))
 [1] "school28" "school31" "school21" "school1"  "school45" "school40"
 [7] "school16" "school50" "school47" "school49" "school4"  "school36"
[13] "school46" "school14" "school24" "school27" "school34" "school9"
```

We see that 18 schools are identified compared to the 8 by the previous procedure. FDR is less stringent than FWER in identifying significant effects. A more convenient method of computing the adjusted p-values is:

```
> padj <- p.adjust(pvals, method="fdr")
> coef(lmod)[padj < 0.05]
 school1  school4  school9 school14 school16 school21 school24
 -3.3685  -2.6619   2.2458  -3.2898  -3.7374  -3.6185   2.6238
school27 school28 school31 school34 school36 school40 school45
 -2.3058  -5.8286   3.8241   1.9359   2.3605  -4.9855  -4.6392
school46 school47 school49 school50
  3.0636   2.3969   2.7964  -2.6520
```

FDR methods are more commonly used where large numbers of comparisons are necessary as often found in imaging or bioinformatics applications. In examples such as these, we expect to find some significant effects and FDR is a useful tool in reliably identifying them.

Exercises

1. Using the `pulp` data, determine whether there are any differences between the operators. What is the nature of these differences?
2. Determine whether there are differences in the weights of chickens according to their feed in the `chickwts` data. Perform all necessary model diagnostics.
3. Using the `PlantGrowth` data, determine whether there are any differences between the groups. What is the nature of these differences? Test for a difference between the average of the two treatments and the control.
4. Using the `infmort` data, perform a one-way ANOVA with `income` as the response and `region` as the predictor. Which pairs of regions are different? Now check for a good transformation on the response and repeat the comparison.
5. The `anaesthetic` data provides the time to restart breathing unassisted in recovering from general anaesthetic for four treatment groups.
   (a) Produce a boxplot depicting the data. Comment on any features of interest.
   (b) Make an appropriate stripchart of the data.
   (c) Produce versions of the previous two plots using the `ggplot2` package. Show how you can overlay these two plots.
   (d) Fit a one-factor model for the recovery times and test for a difference between the two groups.
   (e) Try the Box-Cox transformation method. Explain what went wrong.
   (f) Try a square root transformation on the response. Are the diagnostics satisfactory? Is there a significant difference among the treatment groups?
6. Data on the butterfat content of milk from Canadian cows of five different breeds can be found in the `butterfat` dataset. Consider only mature cows.
   (a) Plot the data and interpret what you see.
   (b) Test for a difference between the breeds.