7 Comparison of Treatment Effects

7.1 Pairwise comparisons

When in the ANOVA we find a significant value of F, so that $H_0 : \tau_i = 0,\ i = 1, \ldots, a$, is rejected, we should investigate in more detail what are the differences between the treatment effects/means.

Whether or not H_0 is true, for any given i, $\bar{y}_{i.}$ has the NID$(\mu + \tau_i, \sigma^2/n_i)$ distribution, as seen in equation (6.5) of Section 6.2. For a given pair i, i' with $i \neq i'$, $\bar{y}_{i.} - \bar{y}_{i'.}$ has the N$(\tau_i - \tau_{i'}, \sigma^2(1/n_i + 1/n_{i'}))$ distribution. Hence the standard error of $\bar{y}_{i.} - \bar{y}_{i'.}$ as an estimator of $\tau_i - \tau_{i'}$ is $s\sqrt{1/n_i + 1/n_{i'}}$, where s^2 is the pooled estimator of the error variance σ^2, obtained from the ANOVA.

Using a test procedure similar to that of a two-sample t-test, for the given pair i, i', we reject the null hypothesis $\tau_i = \tau_{i'}$ against the two-sided alternative, $\tau_i \neq \tau_{i'}$, at the $100\alpha\%$ significance level if and only if the absolute value of the appropriate t-statistic,

$$t = \frac{\bar{y}_{i.} - \bar{y}_{i'.}}{s\sqrt{1/n_i + 1/n_{i'}}} \ ,$$

exceeds the *critical point*, the upper $100 \times \frac{\alpha}{2}\%$ point of the t_{N-a} distribution, i.e.,

$$|t| > t_{N-a,\alpha/2},$$

or, equivalently,

$$|\bar{y}_{i.} - \bar{y}_{i'.}| > t_{N-a,\alpha/2} s\sqrt{1/n_i + 1/n_{i'}} \ . \tag{7.1}$$

The quantity

$$t_{N-a,\alpha/2} s\sqrt{1/n_i + 1/n_{i'}}$$

is the *least significant difference*. A $100(1-\alpha)\%$ confidence interval for $\tau_i - \tau_{i'}$ is given by

$$\bar{y}_{i.} - \bar{y}_{i'.} \pm t_{N-a,\alpha/2} s\sqrt{1/n_i + 1/n_{i'}} \ . \tag{7.2}$$

It is easy to see that the inequality (7.1) is satisfied if and only if the confidence interval (7.2) for $\tau_i - \tau_{i'}$ does not contain the origin.

In our drug development example, the standard error of the difference of two treatment means is

$$\sqrt{s^2(1/6 + 1/6)} = \sqrt{s^2/3}.$$

In R, the least significant difference, `lsd`, at the 5% level can be calculated from first principles as follows.

```
> SED <- sqrt(2834.63/3)
> SED
[1] 30.73885
```

```
> tt <- qt(0.975, 20)
> tt
[1] 2.085963
> lsd <- tt * SED
> lsd
[1] 64.12011
```

Hence the 95% confidence intervals for $\tau_i - \tau_{i'}$ are at

$$\bar{y}_{i.} - \bar{y}_{i'.} \pm 64.1.$$

The multiple comparisons function `pairwise.t.test`, may be used to calculate p-values associated with the confidence intervals for all the pairwise differences, which by default are 95% confidence intervals.

```
> pairwise.t.test(sc,process, p.adj="none")

        Pairwise comparisons using t tests with pooled SD

data:  sc and process

  1      2      3
2 0.0073 -      -
3 0.0223 0.6159 -
4 0.8935 0.0099 0.0296

P value adjustment method: none
```

In our example, processes 2 and 3 both yield higher amounts of solute in the solution than either process 1 or process 4. None of the other pairwise differences are significant at the 5% level.

7.2 The problem of multiple comparisons

The problem with carrying out the $\frac{1}{2}a(a-1)$ such pairwise comparisons simultaneously is that the overall significance level, the error probability of finding at least one significant result when H_0 is true, will be much greater than the significance level α specified for a single test.

The significance level for a single test is often referred to as the *comparison-wise error rate* and the overall significance level is referred to as the *family-wise error rate*.

Correspondingly, the overall confidence level that all the $\frac{1}{2}a(a-1)$ confidence intervals contain the true values of the differences in treatment effects is not $100(1-\alpha)\%$ but substantially less.

It can easily happen then that the F-test from the ANOVA does not give a significant result at the α significance level but one or more significant pairwise differences are found at the same significance level. This apparent inconsistency can be dealt with by the procedure of first carrying out the F-test and then only carrying out the pairwise t-tests if the F-test yields a significant result.

However, there is no clearly "best" way of dealing with the problem of multiple comparisons. A number of alternative approaches may be adopted. The approach used above is straightforward and, despite its simplicity, can be a sensible one to adopt. This use of pairwise t-tests is sometimes referred to as the *Fisher test* or the *Fisher LSD method*.

7.3 The Studentized range

The precise values of the overall significance level of the above approach are determined by the distribution of what is known as the Studentized range. Let $y_1, y_2, \ldots, y_k$ be a random sample from a $N(\mu, \sigma^2)$ distribution, that is, $y_1, y_2, \ldots, y_k$ are NID(μ, σ^2) random variables. The *range* R is defined by

$$R = y_{(k)} - y_{(1)},$$

i.e.,

$$R = \max(y_1, y_2, \ldots, y_k) - \min(y_1, y_2, \ldots, y_k).$$

Let s^2 be an estimate of σ^2 based upon ν degrees of freedom and independently distributed of R. More specifically, assume that

$$\frac{\nu s^2}{\sigma^2} \sim \chi^2_\nu.$$

- Note that because of the independence assumption, s^2 could not be the sample variance of the random sample $y_1, y_2, \ldots, y_k$.
- The distribution of R depends upon k and σ^2, but not upon μ. Taking $\mu = 0$ and dividing the y_i by σ turns them into standard normal variates. Hence the distribution of R/σ depends only upon k and not upon μ or σ.

Define the *Studentized range* q by

$$q = \frac{R}{s}.$$

The distribution of q depends upon k and ν (but not upon μ or σ). Denote the upper $100\alpha\%$ point of this distribution by $q_{k,\nu}(\alpha)$. Thus

$$\mathbb{P}(q > q_{k,\nu}(\alpha)) = \alpha.$$

From the definition of the range and the Studentized range, this probability statement is equivalent to

$$\mathbb{P}\left(\frac{|y_i - y_{i'}|}{s} > q_{k,\nu}(\alpha) \text{ for some } i \neq i'\right) = \alpha.$$

7.4 Tukey's test

In the completely randomized design and the corresponding ANOVA, assume for the present that the number of observations for each treatment is the same,

$$n_i = n \qquad i = 1, \ldots, a.$$

Then under $H_0 : \tau_i = 0,\ i = 1, \ldots, a$, the treatment means, $\bar{y}_{1.}, \bar{y}_{2.}, \ldots, \bar{y}_{a.}$, are NID$(\mu, \sigma^2/n)$. The variance σ^2/n is estimated by s^2/n, where s^2, again the pooled estimate of the error

variance σ^2 obtained from the ANOVA, is based on $N-a$ degrees of freedom and is distributed independently of $\bar{y}_{1.}, \bar{y}_{2.}, \ldots \bar{y}_{a.}$. It follows that

$$q = \frac{\sqrt{n}(\max(\bar{y}_{1.}, \bar{y}_{2.}, \ldots \bar{y}_{a.}) - \min(\bar{y}_{1.}, \bar{y}_{2.}, \ldots \bar{y}_{a.}))}{s}$$

has the distribution of the Studentized range with $k = a$ and $\nu = N - a$.

Tukey's test with overall significance level α rejects the null hypothesis that $\tau_i = \tau_{i'}$ if and only if

$$\frac{\sqrt{n}\ |\bar{y}_{i.} - \bar{y}_{i'.}|}{s} > q_{a,N-a}(\alpha)$$

or, equivalently,

$$|\bar{y}_{i.} - \bar{y}_{i'.}| > q_{a,N-a}(\alpha)\frac{s}{\sqrt{n}}. \tag{7.3}$$

It follows from our discussion of the Studentized range that, under $H_0 : \tau_i = 0,\ i{=}1, \ldots, a$,

$$\mathbb{P}\left(\frac{\sqrt{n}\ |\bar{y}_{i.} - \bar{y}_{i'.}|}{s} > q_{k,\nu}(\alpha) \text{ for some } i \neq i'\right) = \alpha.$$

This observation confirms the fact that for Tukey's test α is the overall significance level, the probability of obtaining at least one significant result when H_0 is true. Correspondingly, confidence intervals for the $\tau_i - \tau_{i'}$ with overall confidence level $100(1-\alpha)\%$ are given by

$$\bar{y}_{i.} - \bar{y}_{i'.} \pm q_{a,N-a}(\alpha)\frac{s}{\sqrt{n}}. \tag{7.4}$$

As in the Fisher case, it is easy to see that the inequality (7.3) is satisfied if and only if the confidence interval (7.4) for $\tau_i - \tau_{i'}$ does not contain the origin. To obtain an explicit comparison with the Fisher confidence interval (7.2), the confidence interval (7.4) may be rewritten as

$$\bar{y}_{i.} - \bar{y}_{i'.} \pm \frac{q_{a,N-a}(\alpha)}{\sqrt{2}}\ s\sqrt{1/n + 1/n}\ . \tag{7.5}$$

In the special case when $n_i = n_{i'} = n$, Fisher's confidence interval (7.2) is seen to be identical in form to the Tukey interval (7.4)/(7.5), only with the quantity $t_{N-a,\alpha/2}$ replaced by $q_{a,N-a}(\alpha)/\sqrt{2}$.

Using R to calculate from first principles the least significant difference according to the Tukey method, $q_{a,N-a}(\alpha)s/\sqrt{n}$, we obtain the following.

```
> SE <- sqrt(2834.63/6)
> SE
[1] 21.73565
> tky <- qtukey(0.95, 4, 20)
> tky
[1] 3.958293
> tsd <- tky * SE
> tsd
[1] 86.03607
```

Hence the Tukey 95% confidence intervals for $\tau_i - \tau_{i'}$ are at

$$\bar{y}_{i.} - \bar{y}_{i'.} \pm 86.0.$$

The Tukey 95% simultaneous confidence intervals for all the pairwise differences can be obtained as follows:

```
> Tukey.drug.batches <-TukeyHSD(drug.batches.aov)
> Tukey.drug.batches
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = sc ~ process, data = drug.batches)

$process
          diff         lwr        upr     p adj
2-1  91.833333    5.797341 177.869325 0.0338079
3-1  76.166667   -9.869325 162.202659 0.0944175
4-1   4.166667  -81.869325  90.202659 0.9990797
3-2 -15.666667 -101.702659  70.369325 0.9558282
4-2 -87.666667 -173.702659  -1.630675 0.0448416
4-3 -72.000000 -158.035992  14.035992 0.1217709
```

while using

```
> plot(Tukey.drug.batches)
```

the corresponding plot can be produced (see Figure 1).
The confidence intervals corresponding to Tukey's test are longer and the test criterion more stringent than for Fisher's test with the same specified significance level. The differences for the pairs (1,3) and (3,4), which previously came out to be significant at the 5% level, are not significant at the 5% level according to the Tukey test. Only the difference between processes 1 and 2, and processes 2 and 4, are still significant.

There is no clear answer as to whether Tukey's test is to be preferred to Fisher's or vice-versa. Tukey's test is more conservative in that it guards more securely against the error of wrongly rejecting a null hypothesis, but it is also less powerful in that it is less likely to reject a null hypothesis when it is false. As stated earlier, one recommended procedure is to first carry out the F-test and then only carry out pairwise t-tests if the F-test yields a significant result.

Strictly speaking, use of the Studentized range distribution is valid only when the number of observations is the same for each treatment. However, as an approximation, for arbitrary sample sizes, $100(1-\alpha)\%$ Tukey confidence intervals for $\tau_i - \tau_{i'}$ are given by

$$\bar{y}_{i.} - \bar{y}_{i'.} \pm \frac{q_{a,N-a}(\alpha)}{\sqrt{2}}\ s\sqrt{1/n_i + 1/n_{i'}}\ .$$

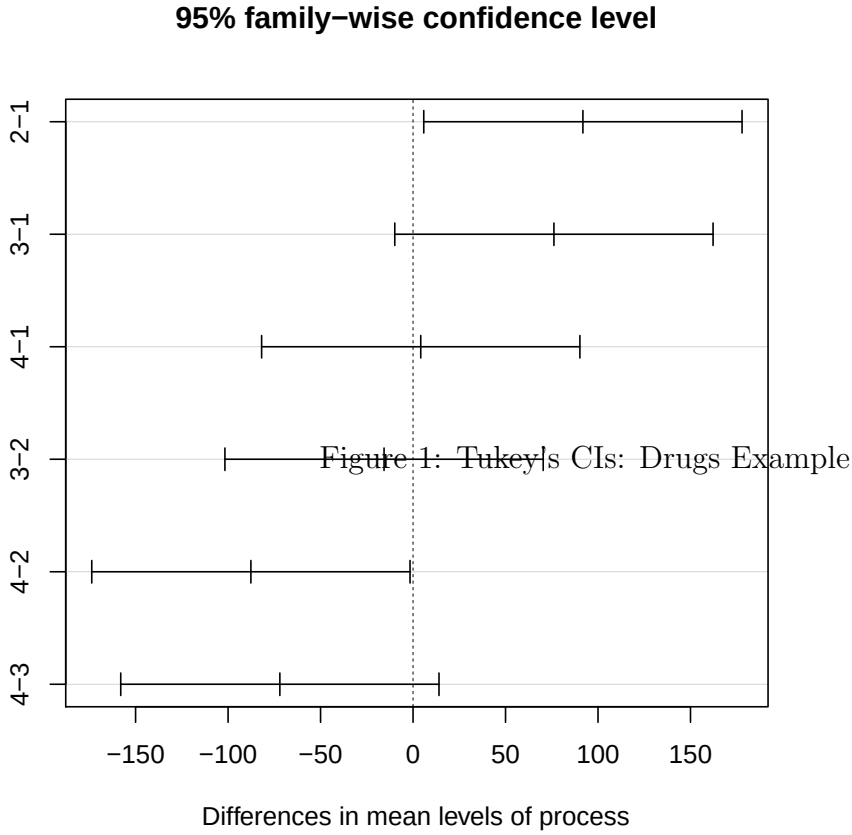

Figure 1: Tukey's CIs: Drugs Example

7.5 Orthogonal contrasts

Recall that a contrast ψ is a linear combination of the treatment effects,

$$\psi = \sum_{i=1}^{a} c_i \tau_i, \tag{7.6}$$

such that

$$\sum_{i=1}^{a} c_i = 0 \tag{7.7}$$

and that it is estimated by

$$\hat{\psi} = \sum_{i=1}^{a} c_i \bar{y}_{i.}, \tag{7.8}$$

which is an unbiased estimator of ψ. Recall again equation (6.5) of Section 6.2 that

$$\bar{y}_{i.} \sim \mathrm{N}\left(\mu + \tau_i, \frac{\sigma^2}{n_i}\right) \qquad i = 1, \ldots, a. \tag{7.9}$$

Furthermore, the $\bar{y}_{i.}$ are independently distributed. Hence from equations (7.8) and (7.9)

$$\hat{\psi} \sim \mathrm{N}(\psi, \sum_{i=1}^{a} \frac{c_i^2}{n_i} \sigma^2). \tag{7.10}$$

We shall be interested in testing the hypothesis $H_0^{\psi} : \psi = 0$. Under H_0^{ψ},

$$\frac{\hat{\psi}}{\sigma \sqrt{\sum_{i=1}^{a} c_i^2 / n_i}} \sim \mathrm{N}(0, 1).$$

Define a sum of squares $SS(\psi)$ for the contrast ψ by

$$SS(\psi) = \frac{\hat{\psi}^2}{\sum_{i=1}^{a} c_i^2 / n_i} \tag{7.11}$$

It follows that, under H_0^{ψ},

$$\frac{SS(\psi)}{\sigma^2} \sim \chi_1^2.$$

There is thus one degree of freedom associated with $SS(\psi)$. To test H_0^{ψ} we may use the test statistic

$$F = \frac{SS(\psi)}{MS_R}, \tag{7.12}$$

which, under H_0^{ψ}, has the $F_{1,N-a}$ distribution.

Now consider a pair of contrasts,

$$\psi = \sum_{i=1}^{a} c_i \tau_i \ , \qquad \psi' = \sum_{i=1}^{a} d_i \tau_i,$$

and the corresponding pair of estimators,

$$\hat{\psi} = \sum_{i=1}^{a} c_i \bar{y}_{i.} \ , \qquad \hat{\psi}' = \sum_{i=1}^{a} d_i \bar{y}_{i.}.$$

We then have

$$\mathrm{cov}(\hat{\psi}, \hat{\psi}') = \sum_{i=1}^{a} \frac{c_i d_i}{n_i} \ \sigma^2.$$

It follows that $\hat{\psi}$ and $\hat{\psi}'$ are independently distributed if and only if the condition

$$\sum_{i=1}^{a} \frac{c_i d_i}{n_i} = 0 \tag{7.13}$$

is satisfied. If the condition of equation (7.13) is satisfied then the contrasts ψ and ψ' are said to be *orthogonal*. In such a case, the corresponding estimators, $\hat{\psi}$ and $\hat{\psi}'$, are independently distributed and so are the corresponding sums of squares, $SS(\psi)$ and $SS(\psi')$.

Note that, in the common special case of the same number of observations for each treatment, the orthogonality condition (7.13) reduces to the simpler condition

$$\sum_{i=1}^{a} c_i d_i = 0. \tag{7.14}$$

It can be shown that, given any set of $a-1$ mutually orthogonal contrasts ψ_1, ψ_2, ..., ψ_{a-1}, the sum of squares for treatments in the ANOVA can be partitioned as

$$SS_{Treatments} = SS(\psi_1) + SS(\psi_2) + \ldots + SS(\psi_{a-1}), \tag{7.15}$$

the terms on the right hand side of equation (7.15) being independently distributed. Thus $SS_{Treatments}$ is partitioned by the $a-1$ individual sums of squares for the contrasts into $a-1$ independently distributed components, each having one degree of freedom. For each of the contrasts, an F-statistic of the form (7.12) can then be used to test whether the contrast is equal to zero.

7.6 An example from plant physiology

In our earlier discussion of multiple comparisons, all the treatments were regarded as being on the same footing. After the ANOVA had been carried out, pairwise comparisons of the treatments were made.

In many experimental situations, however, the treatments are such that, *a priori*, certain particular comparisons of the treatments are regarded as being of special importance. In such situations these comparisons can be cast into the form of examining sets of mutually orthogonal contrasts, which are usually constructed so that each contrast may be interpreted as representing some particular aspect of the differences among treatment effects. They should in principle be constructed before examination of the data to reflect the experimenter's or statistician's views about what aspects of the possible differences are of importance.

If the construction of the contrasts were carried out after examination of the data, there would be the danger of deliberately constructing contrasts to yield large values, which would bias the conclusions.

The data in the following table record the length in coded units of pea sections grown in tissue culture. The purpose of the experiment was to test the effects of various sugars on growth as measured by length. Four experimental groups, representing three different sugars and one mixture of sugars, were used, plus one control without sugar. Ten observations were made for each treatment.

Lengths in ocular units

control	2% glucose	2% fructose	1% gl. +1% fr.	2% sucrose
75	57	58	58	62
67	58	61	59	66
70	60	56	58	65
75	59	58	61	63
65	62	57	57	64
71	60	56	56	62
67	60	61	58	65
67	57	60	57	65
76	59	57	57	62
68	61	58	59	67

A data frame `growth` was created in R with two variables. The first variable `sugar` is a factor which specifies the treatment/group, coded 1 to 5. The second variable is the response variable `length`. The following ANOVA shows that, overall, there are highly significant differences among the five treatments. The table of treatment means gives an initial impression of the relative effects of the treatments.

```
> growth.aov <- aov(length ~ sugar, data = growth)
> summary(growth.aov)
            Df Sum Sq Mean Sq F value   Pr(>F)
sugar        4 1077.3  269.33   49.37 6.74e-16 ***
Residuals   45  245.5    5.46
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

> model.tables(growth.aov)

Tables of effects

 sugar
sugar
    1     2     3     4     5
 8.16 -2.64 -3.74 -3.94  2.16

> model.tables(growth.aov, type = "means")

Tables of means
Grand mean

61.94
```

```
 sugar
sugar
   1    2    3    4    5
70.1 59.3 58.2 58.0 64.1

> coef(growth.aov)
(Intercept)      sugar2      sugar3      sugar4      sugar5
       70.1       -10.8       -11.9       -12.1        -6.0
```

In R, sets of contrasts are specified by matrices, where each column of the matrix is the vector of coefficients of a contrast. The functions `contr.helmert` and `contr.sum` are two of the functions that are available for automatically producing certain contrast matrices in addition to the default `contr.treatment`. In the case of the function `contr.helmert`, if the number of observations is the same for each treatment, the contrasts are mutually orthogonal. In the following output, the number of specified treatments is 5, which results in a 5×4 matrix, that specifies a set of 4 mutually orthogonal contrasts. The *Helmert* set of contrasts successively compare the effect of each treatment with the average of all the previous ones. The set of contrasts produced by the function `contr.sum` compares the final treatment with each of the others. These contrasts are *not* mutually orthogonal.

```
> contr.treatment(5)
  2 3 4 5
1 0 0 0 0
2 1 0 0 0
3 0 1 0 0
4 0 0 1 0
5 0 0 0 1

> contr.sum(5)
  [,1] [,2] [,3] [,4]
1    1    0    0    0
2    0    1    0    0
3    0    0    1    0
4    0    0    0    1
5   -1   -1   -1   -1

> contr.helmert(5)
  [,1] [,2] [,3] [,4]
1   -1   -1   -1   -1
2    1   -1   -1   -1
3    0    2   -1   -1
4    0    0    3   -1
5    0    0    0    4
```

Next we construct our own set of contrasts for the particular problem in hand. (If we did not do so then the default set of *treatment* contrasts, shown above, would be used in the subsequent analysis). The individual vectors of contrast coefficients are specified by the vectors `c1`, `c2`, `c3`, `c4`, and it is readily checked that the coefficients for any pair of contrasts do satisfy the orthogonality condition of equation (7.14). The first is a contrast of the control against the sugars. The second is a contrast of sucrose against glucose and fructose. The third is a simple comparison of glucose with fructose. The fourth contrast is a measure of interaction between glucose and fructose.

```
>  c1 <- c(4,-1,-1,-1,-1)
>  c2 <- c(0,-1,-1,-1,3)
>  c3 <- c(0,1,-1,0,0)
>  c4 <- c(0,-1,-1,2,0)
>  ctr <- matrix(c(c1,c2,c3,c4), nrow=5)
>  ctr
     [,1] [,2] [,3] [,4]
[1,]    4    0    0    0
[2,]   -1   -1    1   -1
[3,]   -1   -1   -1   -1
[4,]   -1   -1    0    2
[5,]   -1    3    0    0
```

The function `matrix` is used to construct an object of the *matrix* type, in this case `ctr`. The first argument, here `c(c1,c2,c3,c4)`, of the function `matrix` is a vector that contains the elements of the matrix, to be entered column by column. The second argument here specifies the number of rows of the matrix.

We now set the contrasts specified by the matrix `ctr` to be associated with the factor `sugar`, using the function `contrasts`. After the ANOVA is carried out again using `aov`, the `split` argument in the function `summary` is used to exhibit an analysis of variance with the sum of squares for `sugar` partitioned into four components, one for each of the contrasts. The function `coef` shows estimated values, $\hat{\psi}/\sum c_i^2$, of scaled versions of each of the contrasts.

```
> contrasts(growth$sugar) <- ctr
> growth.aov <- aov(length ~ sugar, data = growth)
> summary(growth.aov)
            Df Sum Sq Mean Sq F value   Pr(>F)
sugar        4 1077.3  269.33   49.37 6.74e-16 ***
Residuals   45  245.5    5.46
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

> summary(growth.aov, split = list(sugar = list("control v sugars" = 1,
                                    "sucrose v gl,fr" = 2,
                                    "gl v fr" = 3, "gl,fr interaction" = 4)))
                            Df Sum Sq Mean Sq F value   Pr(>F)
sugar                        4 1077.3   269.3  49.368 6.74e-16 ***
  sugar: control v sugars    1  832.3   832.3 152.564 4.68e-16 ***
  sugar: sucrose v gl,fr     1  235.2   235.2  43.112 4.50e-08 ***
  sugar: gl v fr             1    6.0     6.0   1.109    0.298
  sugar: gl,fr interaction   1    3.7     3.7   0.687    0.411
Residuals                   45  245.5     5.5
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1


> model.tables(growth.aov, type = "means")
Tables of means
Grand mean

61.94
```

```
 sugar
sugar
   1    2    3    4    5
70.1 59.3 58.2 58.0 64.1

> coef(growth.aov)
(Intercept)      sugar1      sugar2      sugar3      sugar4
      61.94        2.04        1.40        0.55       -0.25
```

We see that there is very strong evidence that the presence of sugars reduces growth and also that glucose and fructose reduce growth more than sucrose. However, there is no evidence for any difference between the effects of glucose and fructose, nor for any interaction between them.