6 One-Way ANOVA and Treatment Effects

6.1 An example - drug development

A pharmaceutical company experiments with four different processes for making its hay fever allergy eye drops. One of the key active ingredients of the product is Sodium Cromoglycate (SC). There are a variety of processes available for trying to introduce the active ingredient into the solution, none of which are totally exact. Interest lies in determining whether the processes differ significantly in the amount of Sodium Cromoglycate that can be dissolved in 20ml of solution.

In this experiment, twenty-four 20ml batches of purified water were prepared, divided at random into four groups of six batches. To each group of six batches, a different process (Process 1, Process 2, Process 3 or Process 4) was used and each batch was prepared and processed in random order. The data below are the amounts of the solute (SC) in milligrams contained in each batch upon completion of the experiment.

Milligrams of solute contained in each batch

Process 1	Process 2	Process 3	Process 4
276	376	412	310
208	426	296	220
298	430	344	240
278	248	265	289
222	287	309	243
224	290	337	229

This is an example of a *completely randomized design*, one of the simplest types of experimental design. In general, an arbitrary number of experimental conditions, treatments or levels of a factor may be compared, with any number of observations at each level. In this example, there are four treatments (the four processes), with a sample of six observations for each treatment.

An *experimental unit* is one of the individuals or one of the aggregates of material to which a treatment is applied in a single trial of the experiment. In our example the experimental units are the twenty-four 20ml batches of purified water.

Note the fundamental role of *variation* in experimental data. Any differences among the four processes in the amounts of solute found are to some extent masked by the inherent variability of the data.

Replication

- the taking of repeated observations for each treatment, which enables (i) an assessment of the size of the experimental errors to be made and, in particular, an estimate of their variance to be obtained and (ii) more accurate estimates of the effects of the various treatments to be found.

Randomization

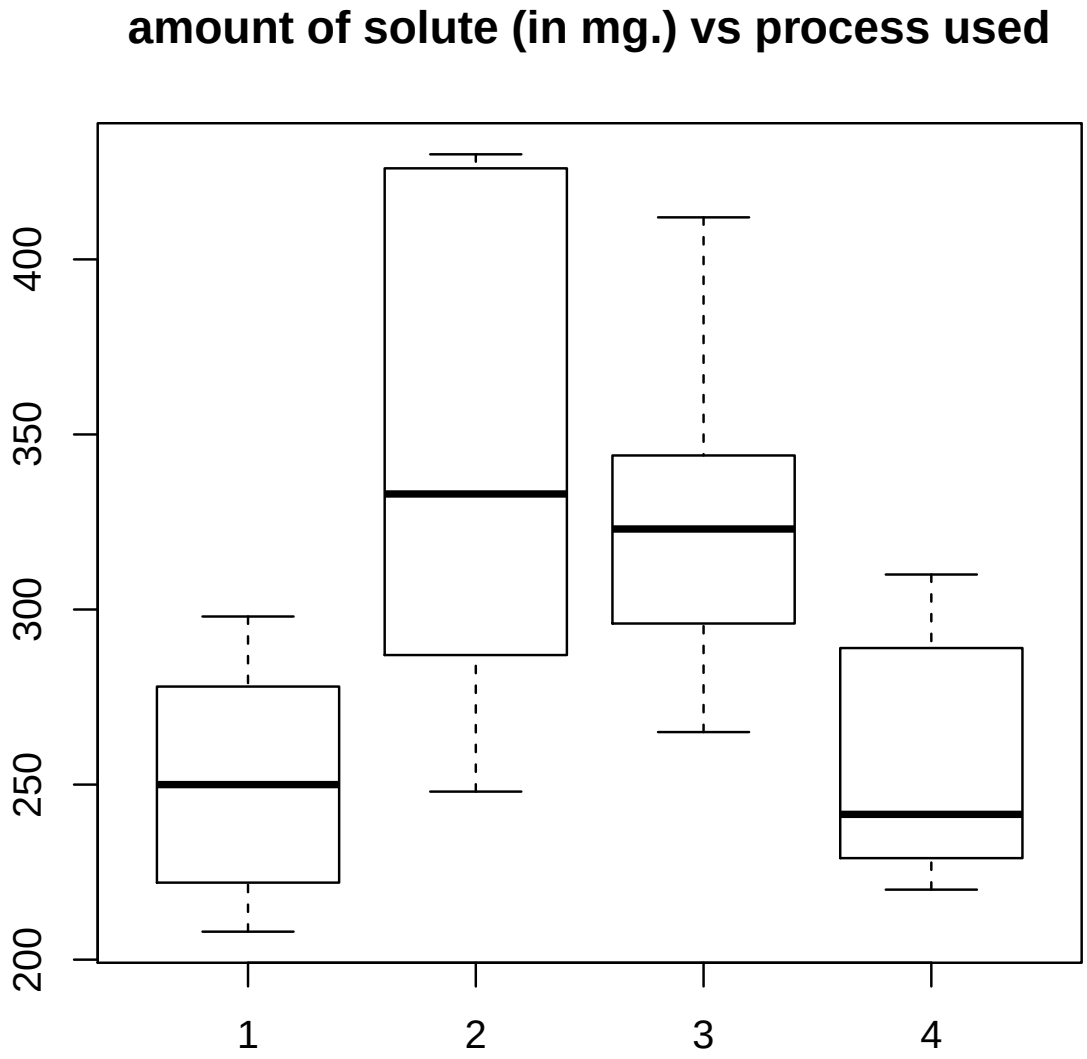
- the random allocation of the experimental units to the various treatments and/or the random ordering of the individual experimental trials in order to avoid biasing the results through any extraneous factors which might be present. (Note that the readings as presented in the table above are not in the order in which they were observed.)

In the following output, the data matrix for our example is created in the form of an R data frame `drug.batches`:

```
> sc <- c(276, 208, 298, 278, 222, 224, 376, 426, 430, 248, 287, 290,
          412, 296, 344, 265, 309, 337, 310, 220, 240, 289, 243, 229)
> process <- factor(rep(1:4, rep(6, 4)))
> drug.batches <- data.frame(sc, process)
> rm(sc, process)
> drug.batches[c(1:3, 7:9, 13:15, 19:21),  ]
   sc process
 1 276       1
 2 208       1
 3 298       1

 7 376       2
 8 426       2
 9 430       2

13 412       3
14 296       3
15 344       3

19 310       4
20 220       4
21 240       4
```

We may carry out a preliminary investigation of the data, by looking at the basic descriptive statistics and plotting the data. To obtain summary statistics of the variable `sc` separately for each level of the factor `process`, we use the function `by`. The `plot` function in this case produces a separate boxplot of `sc` for each level of the factor `process`.

```
> attach(drug.batches)
> plot(process, sc)
> by(sc, process, summary)

process: 1
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  208.0   222.5   250.0   251.0   277.5   298.0
------------------------------------------------------------
process: 2
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  248.0   287.8   333.0   342.8   413.5   430.0
------------------------------------------------------------
process: 3
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  265.0   299.2   323.0   327.2   342.2   412.0
------------------------------------------------------------
process: 4
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  220.0   231.8   241.5   255.2   277.5   310.0
```



There are observable differences among the processes in the mean amount of solute in the solution. Thus Process 2 has the greatest mean amount of solute and Process 1 the least (although Process 4 has the smallest median, as evidenced by the plots). But are these differences due to genuine underlying differences amongst the processes or are they just the result of random fluctuations? A detailed analysis of variance will be carried out to determine whether these apparent differences are statistically significant. We might also note that the data for Process 2 has markedly greater dispersion than the data for the other processes.

As the boxplots show, although there may appear to be differences amongst the processes, these apparent differences are obscured by the inherent variability in the data.

6.2 The statistical model for a completely randomized design

Consider a general situation, in which the effects of a treatments (or levels of a single factor) on a *response variable* are being compared, with n_i observations for Treatment i, $i=1,\ldots,a$. The total number of observations, N, is given by

$$N = \sum_{i=1}^{a} n_i.$$

We set up a *linear statistical model* – a mathematical abstraction, a set of equations which provides a simple conceptual framework for our analysis. Let y_{ij} represent the jth observed value of the response variable for Treatment i, $i=1,\ldots,a$, $j=1,\ldots,n_i$. The model expresses each y_{ij} as the sum of a *systematic* component $\mu+\tau_i$ and a *random* component ϵ_{ij}:

$$y_{ij} = \mu + \tau_i + \epsilon_{ij} \qquad i=1,\ldots,a,\ j=1,\ldots,n_i. \tag{6.1}$$

The parameter μ is an overall mean, the parameters τ_i, $i=1,\ldots,a$ are the *treatment effects* and the ϵ_{ij}, $i=1,\ldots,a$, $j=1,\ldots,n_i$, are *random errors*, the parameters being unknown and to be estimated. It will be assumed that the ϵ_{ij} are independently and identically distributed random variables (i.i.d. r.v.s), each having a normal distribution with mean zero and variance σ^2, where the *error variance* σ^2 is unknown and to be estimated. The shorthand notation that we shall use is that the ϵ_{ij} are NID$(0,\sigma^2)$. Thus the y_{ij} are independently distributed and, for each i, the y_{ij}, $j=1,\ldots,n_i$ are NID$(\mu+\tau_i,\sigma^2)$. Note the assumptions that

(1) the errors are normally distributed, and

(2) the error variance is the same for each treatment.

The model (6.1) is the statistical model for a completely randomized design, the data from which will be analysed using a *one-way analysis of variance* (ANOVA).

There is, in a sense, one redundant parameter in the model, since we could write equation (6.1) as

$$y_{ij} = \mu_i + \epsilon_{ij} \qquad i=1,\ldots,a,\ j=1,\ldots,n_i$$

where μ_i is the ith *treatment mean*,

$$\mu_i = \mu + \tau_i \qquad i=1,\ldots,a, \tag{6.2}$$

thus replacing the set of $a+1$ parameters μ and τ_i, $i=1,\ldots,a$, by the set of a parameters μ_i, $i=1,\ldots,a$.

However, to simplify the interpretation and presentation of the results and the mathematical analysis of the model and to make generalizations to more complex experimental designs more natural, it is usual to retain the model in the form (6.1), but to impose a constraint such as

$$\sum_{i=1}^{a} n_i\tau_i = 0 \tag{6.3}$$

on the treatment effects, interpreting μ as an overall mean and τ_i as a *deviation* from the overall mean. This constraint is often termed the *'sum to zero'* constraint (or `sum` *contrasts* in R).

- It is perfectly possible to impose a linear constraint other than the one specified by equation (6.3), but the latter turns out to be mathematically one of the most convenient.
- From equations (6.2) and (6.3) it follows that $\mu = \sum_i n_i\mu_i/N$. However, μ is not of particular interest.
- In the special and quite common case of a *balanced* design, where the number of observations is the same for each treatment, so that $n_i = n$, $i = 1, \ldots, a$, say, the constraint (6.3) reduces to $\sum_i \tau_i = 0$.

Recall that the default constraint in R for linear models involving *factors* is the *corner constraint* (`treatment` contrasts),

$$\tau_1 = 0 \tag{6.4}$$

so that τ_i, $i = 2, \ldots a$ measures the difference from the baseline or control level μ_1, leading to the dummy variable coding that we have seen.

- Both constraints (coding schemes) are equivalent in that they provide the same fit to the data, producing the same regression and residual sum of squares and hence the same F-test for the differences among group means.

We shall be interested in problems of estimating the parameters or testing hypotheses about them. For carrying out the relevant calculations, it is useful to introduce the following *dot notation*, which readily lends itself to later generalizations, to represent sums and averages that can be calculated from the sample data. Let $y_{i\cdot}$ represent the total of the observations on Treatment i,

$$y_{i\cdot} = \sum_{j=1}^{n_i} y_{ij} \qquad i = 1, \ldots, a,$$

and $\bar{y}_{i\cdot}$ the sample mean of the observations on Treatment i,

$$\bar{y}_{i\cdot} = \frac{y_{i\cdot}}{n_i} = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i} \qquad i = 1, \ldots, a.$$

The observed values of the $\bar{y}_{i\cdot}$ are the sample means of the variable `sc` for the different levels of the factor `process`, as shown in the following output.

```
> by(sc, process, mean)
process: 1
[1] 251
------------------------------------------------------------
process: 2
[1] 342.8333
------------------------------------------------------------
process: 3
[1] 327.1667
------------------------------------------------------------
process: 4
[1] 255.1667
```

Since, for each i, the y_{ij} are NID$(\mu + \tau_i, \sigma^2)$, it follows that

$$\bar{y}_{i\cdot} \sim \mathrm{N}\left(\mu + \tau_i, \frac{\sigma^2}{n_i}\right) \qquad i = 1, \ldots, a. \tag{6.5}$$

Where a dot replaces a subscript, it represents summation (or averaging) over that subscript. A bar over a letter represents averaging. The grand total of all the observations is represented by $y_{..}$,

$$y_{..} = \sum_{i=1}^{a} \sum_{j=1}^{n_i} y_{ij} = \sum_{i=1}^{a} n_i \bar{y}_{i.},$$

and the average over all N observations by $\bar{y}_{..}$,

$$\bar{y}_{..} = \frac{y_{..}}{N} = \frac{\sum_{i=1}^{a} n_i \bar{y}_{i.}}{N}. \tag{6.6}$$

6.3 Estimation of the parameters

Circumflexes over parameters denote their estimates. Given estimates $\hat{\mu}$ and $\hat{\tau}_i$, $i = 1, \ldots, a$, of μ and the τ_i, respectively, the corresponding *fitted value* $\hat{y}_{ij}$ is given by substituting the estimates of the parameters into the *systematic* part of the model (6.1) , i.e.,

$$\hat{y}_{ij} = \hat{\mu} + \hat{\tau}_i.$$

According to the *method of least squares*, given the observed values y_{ij}, $i = 1, \ldots, a$, $j = 1, \ldots, n_i$, we choose as our estimates, $\hat{\mu}$ and $\hat{\tau}_i$, $i = 1, \ldots, a$, those values of μ and τ_i which jointly minimize

$$L = \sum_{i=1}^{a} \sum_{j=1}^{n_i} (y_{ij} - \mu - \tau_i)^2,$$

Setting the partial derivative of L with respect to μ equal to zero, we deduce that the parameter estimates satisfy

$$N\hat{\mu} + n_1\hat{\tau}_1 + n_2\hat{\tau}_2 + \ldots + n_a\hat{\tau}_a = y_{..} \tag{6.7}$$

Also, by setting the partial derivative of L with respect to the τ_i equal to zero, we deduce that the parameter estimates satisfy

$$n_i\hat{\mu} + n_i\hat{\tau}_i = y_{i.} \qquad i = 1, \ldots, a. \tag{6.8}$$

The $a+1$ equations (6.7) and (6.8) for the $a+1$ least squares estimates $\hat{\mu}$ and $\hat{\tau}_i$, $i = 1, \ldots, a$ are known as the *(least squares) normal equations*. However, the equations (6.7) and (6.8) are not linearly independent, since adding together the a equations (6.8) we obtain equation (6.7). To obtain uniquely defined estimates, corresponding to the constraint $\sum_i n_i\tau_i = 0$ on the model parameters, we apply the constraint

$$\sum_{i=1}^{a} n_i\hat{\tau}_i = 0. \tag{6.9}$$

From the equations (6.8), without yet using the constraint (6.9), we obtain

$$\hat{\tau}_i = \bar{y}_{i.} - \hat{\mu} \qquad i = 1, \ldots, a.$$

Substituting the constraint (6.9) into equation (6.7), we obtain $\hat{\mu} = \bar{y}_{..}$. Thus the equations (6.7), (6.8) and (6.9) taken together have the unique solution

$$\hat{\mu} = \bar{y}_{..} \tag{6.10}$$

and

$$\hat{\tau}_i = \bar{y}_{i.} - \bar{y}_{..} \qquad i=1,\ldots,a. \tag{6.11}$$

Equations (6.10) and (6.11) give the *least squares estimates* of the model parameters. Note the convenience of the particular constraint (6.9) adopted in obtaining simple expressions for the estimates.

From the expressions (6.10) and (6.11) (or using equation (6.8)) we obtain for the fitted values $\hat{y}_{ij} \equiv \hat{\mu} + \hat{\tau}_i$ the simple expression

$$\hat{y}_{ij} = \bar{y}_{i.} \qquad i=1,\ldots,a,\ j=1,\ldots,n_i.$$

The corresponding minimized value of L is, therefore,

$$\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2.$$

It is worth noting the unbiasedness properties of our estimators. Firstly, from (6.5) it follows that

$$E[\bar{y}_{i.}] = \mu + \tau_i \equiv \mu_i \qquad i=1,\ldots,a.$$

Thus $\bar{y}_{i.}$ is an unbiased estimator of the treatment mean μ_i, $i = 1,\ldots,a$, a result which, incidentally, does not depend upon our choice of constraint for the treatment effects. Hence, using (6.6) and now also the constraint of equation (6.3),

$$E[\bar{y}_{..}] = \frac{\sum_{i=1}^{a} n_i E[\bar{y}_{i.}]}{N} = \frac{\sum_{i=1}^{a} n_i(\mu+\tau_i)}{N} = \mu.$$

Thus, given our choice of constraint, $\bar{y}_{..}$ is an unbiased estimator of μ. Finally,

$$E[\hat{\tau}_i] = E[\bar{y}_{i.} - \bar{y}_{..}] = E[\bar{y}_{i.}] - E[\bar{y}_{..}] = (\mu+\tau_i) - \mu = \tau_i \qquad i=1,\ldots,a.$$

Thus $\hat{\tau}_i$ is an unbiased estimator of τ_i, $i=1,\ldots,a$.

6.4 The partition of the total corrected sum of squares

The basic *null hypothesis*, H_0, to be tested is that the treatment means $\mu_i \equiv \mu + \tau_i$ are all equal, which is equivalent to the hypothesis that the treatment effects τ_i are all zero, i.e.,

$$H_0 : \tau_1 = \tau_2 = \ldots = \tau_a = 0.$$

The *alternative hypothesis* is

$$H_1 : \tau_i \neq 0 \text{ for at least one } i.$$

The test procedure is based upon a partition of the total sum of squares.

The *total (corrected) sum of squares* SS_T (the sum of squares about the overall mean) for the experiment is defined by

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2.$$

It is a measure of the overall variability of the data. (SS_T, when divided by $N-1$, is just the sample variance of the y_{ij}.) A little algebraic manipulation shows that the total sum of squares may be partitioned into two components,

$$SS_T = SS_{Treatments} + SS_R, \tag{6.12}$$

where

$$SS_{Treatments} = \sum_{i=1}^{a} n_i(\bar{y}_{i.} - \bar{y}_{..})^2$$

and

$$SS_R = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2.$$

$SS_{Treatments}$, the *sum of squares for treatments* or the *sum of squares between treatments* (i.e., between levels of the factor), which may also be written as $\sum_i n_i\hat{\tau}_i^2$, reflects the differences among the treatment means; SS_R, the *residual sum of squares* or the *sum of squares within treatments*, reflects the variation among observations within treatments and is also the minimized value of L.

6.5 The distributional basis for the one-way ANOVA

Consider first in more detail the formula for SS_R. For any given i, whether or not H_0 is true,

$$s_i^2 \equiv \frac{\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2}{n_i - 1},$$

is the sample variance for the NID$(\mu + \tau_i, \sigma^2)$ observations from Treatment i, so that

$$\frac{(n_i - 1)s_i^2}{\sigma^2} \sim \chi^2_{n_i - 1}.$$

The observed values of the s_i^2 for our example are the sample variances of the variable `sc` for the different levels of the factor `process`, as shown in the following output.

```
> by(sc, process, var)
process: 1
[1] 1396.4
------------------------------------------------------------
process: 2
[1] 6103.367
------------------------------------------------------------
process: 3
[1] 2548.567
------------------------------------------------------------
process: 4
[1] 1290.167
```

Now

$$\frac{SS_R}{\sigma^2} = \sum_{i=1}^{a}\frac{(n_i - 1)s_i^2}{\sigma^2}$$

and

$$\sum_{i=1}^{a}(n_i - 1) = N - a.$$

Since the s_i^2 are independently distributed of each other, it follows that

$$\frac{SS_R}{\sigma^2} \sim \chi^2_{N-a} \tag{6.13}$$

and, in particular,

$$E\left[\frac{SS_R}{\sigma^2}\right] = N - a.$$

Hence

$$E\left[\frac{SS_R}{N-a}\right] = \sigma^2. \tag{6.14}$$

Thus the *pooled estimator* s^2,

$$s^2 = \frac{SS_R}{N-a} = \sum_{i=1}^{a} \frac{(n_i - 1)s_i^2}{N-a},$$

a weighted average of the s_i^2, is an unbiased estimator of the error variance σ^2 (in fact, the *minimum variance unbiased estimator* of σ^2).

- An unbiased estimate of σ^2 is provided by the sample variance s_i^2 for any of the treatments, but by pooling the information from all the treatments we obtain a more accurate estimate.

Under $\mathbf{H_0}$, all the y_{ij}, $i = 1, \ldots, a$, $j = 1, \ldots, n_i$ are NID(μ, σ^2) and hence

$$\frac{SS_T}{\sigma^2} \sim \chi^2_{N-1}.$$

It can further be shown that the two terms on the right hand side of the partition (6.12) are independently distributed, from which it follows that, under H_0,

$$\frac{SS_{Treatments}}{\sigma^2} \sim \chi^2_{a-1}, \tag{6.15}$$

where the $a - 1$ degrees of freedom are obtained as the difference $(N - 1) - (N - a)$.

6.6 The ANOVA table

The partition of the total sum of squares and the corresponding partition of the total chi-square degrees of freedom are presented in the first three columns of the Analysis of Variance (ANOVA) table.

ANOVA Table

Source	DF	SS	MS	F
Treatments	$a-1$	$\sum_{i=1}^{a} n_i(\bar{y}_{i.}-\bar{y}_{..})^2$	$SS_{Treat's}/(a-1)$	$MS_{Treat's}/MS_R$
Residuals	$N-a$	$\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i.})^2$	$s^2 \equiv SS_R/(N-a)$	
Total	$N-1$	$\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{..})^2$		

The fourth column in the ANOVA table is for *mean squares* (MS), which are always given by $MS = SS/DF$.

From equations (6.13) and (6.15), under H_0, the F-statistic,

$$F = \frac{MS_{Treatments}}{MS_R}, \tag{6.16}$$

has the $F_{a-1,N-a}$ distribution. This statistic is used to test H_0.

Recall that, as shown in (6.14), whether or not H_0 is true, $s^2 \equiv MS_R$ is an unbiased estimator of σ^2, i.e.,

$$E[MS_R] = \sigma^2.$$

It may further be shown that

$$E[MS_{Treatments}] = \sigma^2 + \frac{1}{a-1}\sum_{i=1}^{a} n_i\tau_i^2.$$

It follows that, under $H_0 : \tau_1 = \tau_2 = \ldots = \tau_a = 0$, $E[MS_{Treatments}] = \sigma^2$, but if H_0 does not hold then $E[MS_{Treatments}] > \sigma^2$.

If H_0 is true then we expect to see values of the F-statistic in the neighbourhood of 1. But if H_0 is false then the larger the departure from H_0, the larger does the value of the F-statistic (6.16) tend to be.

Given N and a, the greater the value of the F-statistic obtained the stronger is the weight of evidence against H_0. Thus we use a one-tail test based on the F-statistic (6.16) to test H_0 – we reject H_0 if the value of the F-statistic is large enough.

In the following output, the ANOVA is carried out using the `aov` function to create an object `drug.batches.aov` that contains the results of the analysis. The model to be used for the ANOVA is specified as `sc` $\sim$ `process`, i.e., the response variable `sc` is to be modelled in terms of the single factor `process`. The object `drug.batches.aov` can be examined in various ways, but the function `summary` gives the basic ANOVA table.

```
> drug.batches.aov <- aov(sc ~ process, data = drug.batches)
```

```
> summary(drug.batches.aov)
             Df Sum Sq Mean Sq F value Pr(>F)
process       3  41050   13683   4.827 0.0109 *
Residuals    20  56693    2835
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

The value of the F-statistic is 4.827265 with p-value 0.01094938, so that we reject H_0 at the 5% significance level (but not quite at the 1% level). We deduce that there is very strong evidence that the process used influences the amount of Sodium Cromoglycate that is introduced into the solution.

6.7 Confidence intervals for treatment means

Recall from (6.5) that

$$\text{var}(\bar{y}_{i\cdot}) = \frac{\sigma^2}{n_i}.$$

The *standard error* of $\bar{y}_{i\cdot}$ is obtained by replacing σ^2 in the above expression by its estimate $s^2 \equiv MS_R$ and then taking the square root to give

$$\frac{s}{\sqrt{n_i}},$$

with $N - a$ degrees of freedom.

- Note that the above standard error is not the same as the one that would be used if the ith treatment were being considered in isolation, in which case we would use $s_i/\sqrt{n_i}$ with $n_i - 1$ degrees of freedom as the standard error. However, we make use of the information on the error variance that is available from all the treatments to obtain a more accurate standard error for any individual treatment.

- In R, $s \equiv \sqrt{MS_R}$ is referred to as the *residual standard error*.

An unbiased estimator of the population mean $\mu_i \equiv \mu + \tau_i$ for Treatment i is given by $\hat{\mu}_i = \bar{y}_{i\cdot}$ and a 95% confidence interval is given by

$$\bar{y}_{i\cdot} \pm t_{N-a,0.025}\frac{s}{\sqrt{n_i}},$$

where $t_{N-a,0.025}$ is the (upper) 2.5% point of the t-distribution with $N - a$ degrees of freedom.

As an illustration, in the present example the standard error of each of the means is $\sqrt{2834.63/6} = 21.73565$ and a 95% confidence interval for the mean amount of solute found in the solution under process 1 is given by

$$251 \pm t_{20,0.025} \times 21.73565, \qquad i.e., \qquad (205.6602, 296.3398).$$

The calculations may be carried out in R as follows.

```
> m <- mean(sc[process == 1])
> m
[1] 251
> tt <- qt(0.975, 20)
> tt
[1] 2.085963
> SE <- sqrt(2834.63/6)
> SE
[1] 21.73565
> CI <- c(m - tt * SE, m + tt * SE)
> CI
[1] 205.6602 296.3398
```

- Note that s can be obtained in R directly using `summary.lm` function, which retains an estimate of the model standard deviation as an *attribute*.

```
> s2 <- summary.lm(drug.batches.aov)$sigma^2
> s2
[1] 2834.625
```

6.8 Estimable functions and contrasts

The estimates of the model parameters as given in equations (6.10) and (6.11) depend upon the particular constraint (6.9) that is adopted. Those functions of the parameters which have unbiased estimators that do not depend upon the constraint that is adopted are known as *estimable functions*. Such functions turn out to be of particular importance in more detailed investigation of the treatment effects.

Since, as we have noted earlier, $E[\bar{y}_{i\cdot}] = \mu + \tau_i$, the treatment mean $\mu_i \equiv \mu + \tau_i$ is an estimable function for $i = 1, \ldots, a$, having $\bar{y}_{i\cdot}$ as an unbiased estimator. Another important class of estimable functions is the family of *contrasts* in the treatment effects. A contrast ψ is a linear combination of the treatment effects τ_i,

$$\psi = \sum_{i=1}^{a} c_i \tau_i \tag{6.17}$$

such that the coefficients c_i satisfy

$$\sum_{i=1}^{a} c_i = 0. \tag{6.18}$$

We shall discuss the construction and interpretation of contrasts later. The simplest example of a contrast is the difference between any two treatment effects, $\tau_i - \tau_{i'}$, which is zero if and only if there is no difference between the effects of Treatment i and Treatment i'.

To check that the contrast ψ is indeed estimable, note that

$$\begin{aligned} \mathrm{E}\left[\sum_{i=1}^{a} c_i \bar{y}_{i\cdot}\right] &= \sum_{i=1}^{a} c_i \mathrm{E}[\bar{y}_{i\cdot}] = \sum_{i=1}^{a} c_i(\mu + \tau_i) \\ &= \sum_{i=1}^{a} c_i \mu + \sum_{i=1}^{a} c_i \tau_i = \sum_{i=1}^{a} c_i \tau_i = \psi, \end{aligned}$$

using Equations (6.17) and (6.18). Thus the contrast ψ is estimable with unbiased estimator $\hat{\psi} = \sum_{i=1}^{a} c_i \bar{y}_{i\cdot}$, which is indeed the natural and best estimator to use.

6.9 The estimates of the model parameters in R

Recall that the model parameters, with the constraint $\sum_{i=1}^{a} n_i\tau_i = 0$ that we have adopted, are estimated by

$$\hat{\mu} = \bar{y}_{..}$$

and

$$\hat{\tau}_i = \bar{y}_{i.} - \bar{y}_{..} \qquad i=1,\ldots,a.$$

The calculations are carried out in the following output. The function `by` creates a type of object known as a *list*. The function `unlist` converts the list into a vector. In R, when a scalar is subtracted from a vector, it is subtracted from each element of the vector. Here, the scalar `mean(sc)` is subtracted from each element of the vector `unlist(by(sc,process,mean))`.

```
> attach(drug.batches)
> mean(sc)
[1] 294.0417
> unlist(by(sc, process, mean)) - mean(sc)
process: 1
[1] -43.04167
------------------------------------------------------------
process: 2
[1] 48.79167
------------------------------------------------------------
process: 3
[1] 33.125
------------------------------------------------------------
process: 4
[1] -38.875
```

Thus

$$\hat{\mu} = 294.0417$$

and

$$(\hat{\tau}_1, \hat{\tau}_2, \hat{\tau}_3, \hat{\tau}_4) = (-43.04167, 48.79167, 33.125, -38.875).$$

The parameter estimates and treatment means can be accessed directly by repeated application of the function `model.tables` to the object that contains the results of the analysis. In the present case we obtain the following.

```
> model.tables(drug.batches.aov)

Tables of effects

process
     1      2      3      4
-43.04  48.79  33.12 -38.88

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

Tables of means
Grand mean

294.0417

process
    1     2     3     4
251.0 342.8 327.2 255.2
```

Recall that we have so far assumed the use of the `sum` constraint ($\sum_i \tau_i = 0$), but that the default in R is the 'corner' or `treatment` constraint ($\tau_1 = 0$). Hence, looking at the associated coefficients from the associated `aov` object we find

```
> coef(drug.batches.aov)
 (Intercept)    process2    process3    process4
 251.000000   91.833333   76.166667    4.166667

> dummy.coef(drug.batches.aov)
Full coefficients are

(Intercept):          251
process:                1         2         3         4
                0.000000 91.833333 76.166667  4.166667
```

- You should satisfy yourself that these parameters lead to the same fitted values we have described above.

The same estimates can be obtained by using the R function `lm`:

```
drug.batches.lm <- lm(sc~process,data=drug.batches)
summary(drug.batches.lm)

Call:
lm(formula = sc ~ process, data = drug.batches)

Residuals:
   Min     1Q Median     3Q    Max
-94.83 -32.17 -13.67  33.33  87.17

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  251.000     21.736  11.548 2.67e-10 ***
process2      91.833     30.739   2.988  0.00728 **
process3      76.167     30.739   2.478  0.02226 *
process4       4.167     30.739   0.136  0.89353
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 53.24 on 20 degrees of freedom
Multiple R-squared:   0.42,     Adjusted R-squared:  0.333
F-statistic: 4.827 on 3 and 20 DF,  p-value: 0.01095
```