7

Analysis of variance and the Kruskal–Wallis test

In this section, we consider comparisons among more than two groups parametrically, using analysis of variance, as well as nonparametrically, using the Kruskal–Wallis test. Furthermore, we look at two-way analysis of variance in the case of one observation per cell.

7.1 One-way analysis of variance

We start this section with a brief sketch of the theory underlying the one-way analysis of variance. A little bit of notation is necessary. Let x_{ij} denote observation no. j in group i, so that x_{35} is the fifth observation in group 3; $\bar{x}_i$ is the mean for group i, and $\bar{x}_.$ is the grand mean (average of all observations).

We can decompose the observations as

$$x_{ij} = \bar{x}_. + \underbrace{(\bar{x}_i - \bar{x}_.)}_{\substack{\text{deviation of} \\ \text{group mean from} \\ \text{grand mean}}} + \underbrace{(x_{ij} - \bar{x}_i)}_{\substack{\text{deviation of} \\ \text{observation from} \\ \text{group mean}}}$$

informally corresponding to the model

$$X_{ij} = \mu + \alpha_i + \epsilon_{ij}, \qquad \epsilon_{ij} \sim N(0, \sigma^2)$$

P. Dalgaard, *Introductory Statistics with R*,
DOI: 10.1007/978-0-387-79054-1_7,

in which the hypothesis that all the groups are the same implies that all α_i are zero. Notice that the error terms ϵ_{ij} are assumed to be independent and have the same variance.

Now consider the sums of squares of the underbraced terms, known as *variation within groups*

$$\mathrm{SSD}_W = \sum_i \sum_j (x_{ij} - \bar{x}_i)^2$$

and *variation between groups*

$$\mathrm{SSD}_B = \sum_i \sum_j (\bar{x}_i - \bar{x}_.)^2 = \sum_i n_i (\bar{x}_i - \bar{x}_.)^2$$

It is possible to prove that

$$\mathrm{SSD}_B + \mathrm{SSD}_W = \mathrm{SSD}_{\mathrm{total}} = \sum_i \sum_j (x_{ij} - \bar{x}_.)^2$$

That is, the total variation is split into a term describing differences between group means and a term describing differences between individual measurements within the groups. One says that the grouping explains part of the total variation, and obviously an informative grouping will explain a large part of the variation.

However, the sums of squares can only be positive, so even a completely irrelevant grouping will always "explain" some part of the variation. The question is how small an amount of explained variation can be before it might as well be due to chance. It turns out that in the absence of any systematic differences between the groups, you should expect the sum of squares to be partitioned according to the degrees of freedom for each term, $k-1$ for SSD_B and $N-k$ for SSD_W, where k is the number of groups and N is the total number of observations.

Accordingly, you can normalize the sums of squares by calculating *mean squares*:

$$\mathrm{MS}_W = \mathrm{SSD}_W / (N-k)$$
$$\mathrm{MS}_B = \mathrm{SSD}_B / (k-1)$$

MS_W is the pooled variance obtained by combining the individual group variances and thus an estimate of σ^2. In the absence of a true group effect, MS_B will also be an estimate of σ^2, but if there *is* a group effect, then the differences between group means and hence MS_B will tend to be larger. Thus, a test for significant differences between the group means can be performed by comparing two variance estimates. This is why the procedure is called *analysis of variance* even though the objective is to compare the group means.

A formal test needs to account for the fact that random variation will cause some difference in the mean squares. You calculate

$$F = \mathrm{MS}_B / \mathrm{MS}_W$$

so that F is ideally 1, but some variation around that value is expected. The distribution of F under the null hypothesis is an F distribution with $k-1$ and $N-k$ degrees of freedom. You reject the hypothesis of identical means if F is larger than the 95% quantile in that F distribution (if the significance level is 5%). Notice that this test is one-sided; a very small F would occur if the group means were very similar, and that will of course not signify a difference between the groups.

Simple analyses of variance can be performed in R using the function `lm`, which is also used for regression analysis. For more elaborate analyses, there are also the functions `aov` and `lme` (linear mixed effects models, from the `nlme` package). An implementation of Welch's procedure, relaxing the assumption of equal variances and generalizing the unequal-variance t test, is implemented in `oneway.test` (see Section 7.1.2).

The main example in this section is the "red cell folate" data from Altman (1991, p. 208). To use `lm`, it is necessary to have the data values in one vector and a factor variable (see Section 1.2.8) describing the division into groups. The `red.cell.folate` data set contains a data frame in the proper format.

```
> attach(red.cell.folate)
> summary(red.cell.folate)
     folate          ventilation
 Min.   :206.0   N2O+O2,24h:8
 1st Qu.:249.5   N2O+O2,op :9
 Median :274.0   O2,24h    :5
 Mean   :283.2
 3rd Qu.:305.5
 Max.   :392.0
```

Recall that `summary` applied to a data frame gives a short summary of the distribution of each of the variables contained in it. The format of the summary is different for numeric vectors and factors, so that provides a check that the variables are defined correctly.

The category names for `ventilation` mean "N_2O and O_2 for 24 hours", "N_2O and O_2 during operation", and "only O_2 for 24 hours".

In the following, the analysis of variance is demonstrated first and then a couple of useful techniques for the presentation of grouped data as tables and graphs are shown.

The specification of a one-way analysis of variance is analogous to a regression analysis. The only difference is that the descriptive variable needs to be a factor and not a numeric variable. We calculate a model object using `lm` and extract the analysis of variance table with `anova`.

```
> anova(lm(folate~ventilation))
Analysis of Variance Table

Response: folate
            Df Sum Sq Mean Sq F value  Pr(>F)
ventilation  2  15516    7758  3.7113 0.04359 *
Residuals   19  39716    2090
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here we have SSD_B and MS_B in the top line and SSD_W and MS_W in the second line.

In statistics textbooks, the sums of squares are most often labelled "between groups" and "within groups". Like most other statistical software, R uses slightly different labelling. Variation between groups is labelled by the name of the grouping factor (`ventilation`), and variation within groups is labelled `Residual`. ANOVA tables can be used for a wide range of statistical models, and it is convenient to use a format that is less linked to the particular problem of comparing groups.

For a further example, consider the data set `juul`, introduced in Section 4.1. Notice that the `tanner` variable in this data set is a numeric vector and not a factor. For purposes of tabulation, this makes little difference, but it would be a serious error to use it in this form in an analysis of variance:

```
> attach(juul)
> anova(lm(igf1~tanner))  ## WRONG!
Analysis of Variance Table

Response: igf1
           Df   Sum Sq  Mean Sq F value    Pr(>F)
tanner      1 10985605 10985605  686.07 < 2.2e-16 ***
Residuals 790 12649728    16012
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This does not describe a grouping of data but a linear regression on the group number! Notice the telltale 1 DF for the effect of `tanner`.

Things can be fixed as follows:

```
> juul$tanner <- factor(juul$tanner,
+                       labels=c("I","II","III","IV","V"))
```

```
> detach(juul)
> attach(juul)
> summary(tanner)
   I   II  III   IV    V NA's
 515  103   72   81  328  240
> anova(lm(igf1~tanner))
Analysis of Variance Table

Response: igf1
           Df   Sum Sq  Mean Sq F value    Pr(>F)
tanner      4 12696217  3174054  228.35 < 2.2e-16 ***
Residuals 787 10939116    13900
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We needed to reattach the `juul` data frame in order to use the changed definition. An attached data frame is effectively a separate copy of it (although it does not take up extra space as long as the original is unchanged). The `Df` column now has an entry of 4 for `tanner`, as it should.

7.1.1 *Pairwise comparisons and multiple testing*

If the F test shows that there is a difference between groups, the question quickly arises of where the difference lies. It becomes necessary to compare the individual groups.

Part of this information can be found in the regression coefficients. You can use `summary` to extract regression coefficients with standard errors and t tests. These coefficients do not have their usual meaning as the slope of a regression line but have a special interpretation, which is described below.

```
> summary(lm(folate~ventilation))
Call:
lm(formula = folate ~ ventilation)
Residuals:
    Min      1Q  Median      3Q     Max
-73.625 -35.361  -4.444  35.625  75.375
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)           316.62      16.16  19.588 4.65e-14 ***
ventilationN2O+O2,op  -60.18      22.22  -2.709   0.0139 *
ventilationO2,24h     -38.62      26.06  -1.482   0.1548
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 45.72 on 19 degrees of freedom
Multiple R-squared: 0.2809,    Adjusted R-squared: 0.2052
F-statistic: 3.711 on 2 and 19 DF,  p-value: 0.04359
```

The interpretation of the estimates is that the intercept is the mean in the first group (`N2O+O2,24h`), whereas the two others describe the *difference* between the relevant group and the first one.

There are multiple ways of representing the effect of a factor variable in linear models (and one-way analysis of variance is the simplest example of a linear model with a factor variable). The representations are in terms of *contrasts*, the choice of which can be controlled either by global options or as part of the model formula. We do not go deeply into this but just mention that the contrasts used by default are the so-called *treatment contrasts*, in which the first group is treated as a baseline and the other groups are given relative to that. Concretely, the analysis is performed as a multiple regression analysis (see Chapter 11) by introducing two *dummy variables*, which are 1 for observations in the relevant group and 0 elsewhere.

Among the t tests in the table, you can immediately find a test for the hypothesis that the first two groups have the same true mean ($p = 0.0139$) and also whether the first and the third might be identical ($p = 0.1548$). However, a comparison of the last two groups cannot be found. This can be overcome by modifying the factor definition (see the help page for `relevel`), but that gets tedious when there are more than a few groups.

If we want to compare all groups, we ought to correct for *multiple testing*. Performing many tests will increase the probability of finding one of them to be significant; that is, the p-values tend to be exaggerated. A common adjustment method is the *Bonferroni correction*, which is based on the fact that the probability of observing at least one of n events is less than the sum of the probabilities for each event. Thus, by dividing the significance level by the number of tests or, equivalently, multiplying the p-values, we obtain a *conservative* test where the probability of a significant result is less than or equal to the formal significance level.

A function called `pairwise.t.test` computes all possible two-group comparisons. It is also capable of making adjustments for multiple comparisons and works like this:

```
> pairwise.t.test(folate, ventilation, p.adj="bonferroni")

        Pairwise comparisons using t tests with pooled SD

data:  folate and ventilation

           N2O+O2,24h N2O+O2,op
N2O+O2,op 0.042      -
O2,24h    0.464      1.000

P value adjustment method: bonferroni
```

The output is a table of p-values for the pairwise comparisons. Here, the p-values have been adjusted by the Bonferroni method, where the unadjusted values have been multiplied by the number of comparisons, namely 3. If that results in a value bigger than 1, then the adjustment procedure sets the adjusted p-value to 1.

The default method for `pairwise.t.test` is actually not the Bonferroni correction but a variant due to Holm. In this method, only the smallest p needs to be corrected by the full number of tests, the second smallest is corrected by $n - 1$, etc., unless that would make it smaller than the previous one, since the order of the p-values should be unaffected by the adjustment.

```
> pairwise.t.test(folate,ventilation)

       Pairwise comparisons using t tests with pooled SD

data:  folate and ventilation

          N2O+O2,24h N2O+O2,op
N2O+O2,op 0.042      -
O2,24h    0.310      0.408

P value adjustment method: holm
```

7.1.2 *Relaxing the variance assumption*

The traditional one-way ANOVA requires an assumption of equal variances for all groups. There is, however, an alternative procedure that does not require that assumption. It is due to Welch and similar to the unequal-variances t test. This has been implemented in the `oneway.test` function:

```
> oneway.test(folate~ventilation)

        One-way analysis of means (not assuming equal variances)

data:  folate and ventilation
F = 2.9704, num df =  2.000, denom df = 11.065, p-value = 0.09277
```

In this case, the p-value increased to a nonsignificant value, presumably related to the fact that the group that seems to differ from the two others also has the largest variance.

It is also possible to perform the pairwise t tests so that they do not use a common pooled standard deviation. This is controlled by the argument `pool.sd`.

```
>  pairwise.t.test(folate,ventilation,pool.sd=F)

        Pairwise comparisons using t tests with non-pooled SD

data:  folate and ventilation

          N2O+O2,24h N2O+O2,op
N2O+O2,op 0.087      -
O2,24h    0.321      0.321

P value adjustment method: holm
```

Again, it is seen that the significance disappears as we remove the constraint on the variances.

7.1.3 *Graphical presentation*

Of course, there are many ways to present grouped data. Here we create a somewhat elaborate plot where the raw data are plotted as a stripchart and overlaid with an indication of means and SEMs (Figure 7.1):

```
> xbar <- tapply(folate, ventilation, mean)
> s <- tapply(folate, ventilation, sd)
> n <- tapply(folate, ventilation, length)
> sem <- s/sqrt(n)
> stripchart(folate~ventilation, method="jitter",
+    jitter=0.05, pch=16, vert=T)
> arrows(1:3,xbar+sem,1:3,xbar-sem,angle=90,code=3,length=.1)
> lines(1:3,xbar,pch=4,type="b",cex=2)
```

Here we used `pch=16` (small plotting dots) in `stripchart` and put `vertical=T` to make the "strips" vertical.

The error bars have been made with `arrows`, which adds arrows to a plot. We slightly abuse the fact that the angle of the arrowhead is adjustable to create the little crossbars at either end. The first four arguments specify the endpoints, (x_1, y_1, x_2, y_2); the `angle` argument gives the angle between the lines of the arrowhead and shaft, here set to 90°; and `length` is the length of the arrowhead (in inches on a printout). Finally, `code=3` means that the arrow should have a head at both ends. Note that the x-coordinates of the stripcharts are simply the group numbers.

The indication of averages and the connecting lines are done with `lines`, where `type="b"` (*b*oth) means that both points and lines are printed, leaving gaps in the lines to make room for the symbols. `pch=4` is a cross, and `cex=2` requests that the symbols be drawn in double size.

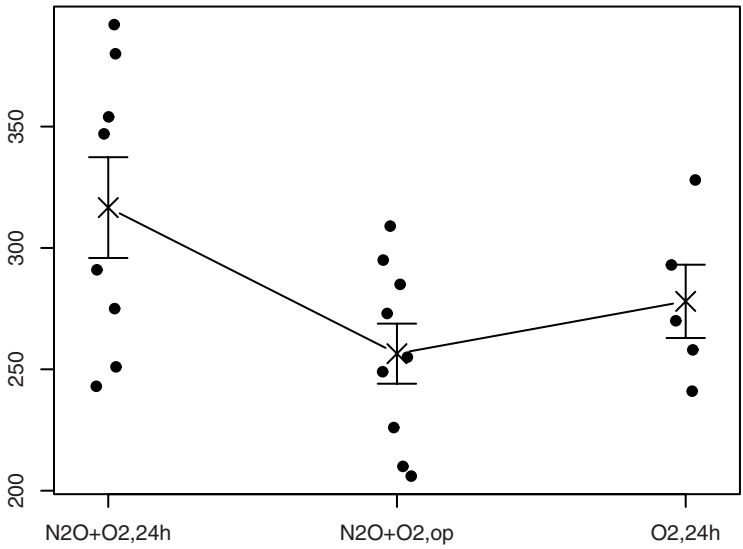

Figure 7.1. "Red cell folate" data with $\bar{x} \pm 1$ SEM.

It is debatable whether you should draw the plot using 1 SEM as is done here or whether perhaps it is better to draw proper confidence intervals for the means (approximately 2 SEM), or maybe even SD instead of SEM. The latter point has to do with whether the plot is to be used in a descriptive or an analytical manner. Standard errors of the mean are not useful for describing the distributions in the groups; they only say how precisely the mean is determined. On the other hand, SDs do not enable the reader to see at a glance which groups are significantly different.

In many fields it appears to have become the tradition to use 1 SEM "because they are the smallest"; that is, it makes differences look more dramatic. Probably, the best thing to do is to follow the traditions in the relevant field and "calibrate your eyeballs" accordingly.

One word of warning, though: At small group sizes, the rule of thumb that the confidence interval is the mean $\pm$ 2 SEM becomes badly misleading. At a group size of 2, it actually has to be 12.7 SEM! That is a correction heavily dependent on data having the normal distribution. If you have such small groups, it may be advisable to use a pooled SD for the entire data set rather than the group-specific SDs. This does, of course, require

that you can reasonably assume that the true standard deviation actually is the same in all groups.

7.1.4 *Bartlett's test*

Testing whether the distribution of a variable has the same variance in all groups can be done using Bartlett's test, although like the F test for comparing two variances, it is rather nonrobust against departures from the assumption of normal distributions. As in `var.test`, it is assumed that the data are from independent groups. The procedure is performed as follows:

```
> bartlett.test(folate~ventilation)

        Bartlett test of homogeneity of variances

data:  folate by ventilation
Bartlett's K-squared = 2.0951, df = 2, p-value = 0.3508
```

That is, in this case, nothing in the data contradicts the assumption of equal variances in the three groups.

7.2 Kruskal–Wallis test

A nonparametric counterpart of a one-way analysis of variance is the Kruskal–Wallis test. As in the Wilcoxon two-sample test (see Section 5.5), data are replaced with their ranks without regard to the grouping, only this time the test is based on the between-group sum of squares calculated from the average ranks. Again, the distribution of the test statistic can be worked out based on the idea that, under the hypothesis of irrelevant grouping, the problem reduces to a combinatorial one of sampling the within-group ranks from a fixed set of numbers.

You can make R calculate the Kruskal–Wallis test as follows:

```
> kruskal.test(folate~ventilation)

        Kruskal-Wallis rank sum test

data:  folate by ventilation
Kruskal-Wallis chi-squared = 4.1852, df = 2, p-value = 0.1234
```

It is seen that there is no significant difference using this test. This should not be too surprising in view of the fact that the F test in the one-way analysis of variance was only borderline significant. Also, the Kruskal–Wallis

test is less efficient than its parametric counterpart if the assumptions hold, although it does not invariably give a larger p-value.

7.3 Two-way analysis of variance

One-way analysis of variance deals with one-way classifications of data. It is also possible to analyze data that are cross-classified according to several criteria. When a cross-classified design is *balanced*, then you can almost read the entire statistical analysis from a single analysis of variance table, and that table generally consists of items that are simple to compute, which was very important before the computer era. Balancedness is a concept that is hard to define exactly; for a two-way classification, a sufficient condition is that the cell counts be equal, but there are other balanced designs.

Here we restrict ourselves to the case of a single observation per cell. This typically arises from having multiple measurements on the same experimental unit and in this sense generalizes the paired t test.

Let x_{ij} denote the observation in row i and column j of the $m \times n$ table. This is similar to the notation used for one-way analysis of variance, but notice that there is now a connection between observations with the same j, so that it makes sense to look at both row averages $\bar{x}_{i\cdot}$ and column averages $\bar{x}_{\cdot j}$.

Consequently, it now makes sense to look at both *variation between rows*

$$\mathrm{SSD}_R = n \sum_i (\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})^2$$

and *variation between columns*

$$\mathrm{SSD}_C = m \sum_j (\bar{x}_{\cdot j} - \bar{x}_{\cdot\cdot})^2$$

Subtracting these two from the total variation leaves the *residual variation*, which works out as

$$\mathrm{SSD}_{\mathrm{res}} = \sum_i \sum_j (x_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{x}_{\cdot\cdot})^2$$

This corresponds to a statistical model in which it is assumed that the observations are composed of a general level, a row effect, and a column effect plus a noise term:

$$X_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \qquad \epsilon_{ij} \sim N(0, \sigma^2)$$

The parameters of this model are not uniquely defined unless we impose some restriction on the parameters. If we impose $\sum \alpha_i = 0$ and $\sum \beta_j = 0$, then the estimates of α_i, β_j, and μ turn out to be $\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot}$, $\bar{x}_{\cdot j} - \bar{x}_{\cdot\cdot}$, and $\bar{x}_{\cdot\cdot}$.

Dividing the sums of squares by their respective degrees of freedom $m-1$ for SSD_R, $n-1$ for SSD_C, and $(m-1)(n-1)$ for $\mathrm{SSD}_{\mathrm{res}}$, we get a set of mean squares. F tests for no row and column effect can be carried out by dividing the respective mean squares by the residual mean square.

It is important to notice that this works out so nicely only because of the balanced design. If you have a table with "holes" in it, the analysis is considerably more complicated. The simple formulas for the sum of squares are no longer valid and, in particular, the order independence is lost, so that there is no longer a single SSD_C but ones with and without adjusting for row effects.

To perform a two-way ANOVA, it is necessary to have data in one vector, with the two classifying factors parallel to it. We consider an example concerning heart rate after administration of enalaprilate (Altman, 1991, p. 327). Data are found in this form in the `heart.rate` data set:

```
> attach(heart.rate)
> heart.rate
    hr subj time
1   96    1    0
2  110    2    0
3   89    3    0
4   95    4    0
5  128    5    0
6  100    6    0
7   72    7    0
8   79    8    0
9  100    9    0
10  92    1   30
11 106    2   30
12  86    3   30
13  78    4   30
14 124    5   30
15  98    6   30
16  68    7   30
17  75    8   30
18 106    9   30
19  86    1   60
20 108    2   60
21  85    3   60
22  78    4   60
23 118    5   60
24 100    6   60
25  67    7   60
26  74    8   60
27 104    9   60
```

```
28  92    1  120
29 114    2  120
30  83    3  120
31  83    4  120
32 118    5  120
33  94    6  120
34  71    7  120
35  74    8  120
36 102    9  120
```

If you look inside the `heart.rate.R` file in the `data` directory of the `ISwR` package, you will see that the actual definition of the data frame is

```
heart.rate <- data.frame(hr = c(96,110,89,95,128,100,72,79,100,
                                92,106,86,78,124,98,68,75,106,
                                86,108,85,78,118,100,67,74,104,
                                92,114,83,83,118,94,71,74,102),
                         subj=gl(9,1,36),
                         time=gl(4,9,36,labels=c(0,30,60,120)))
```

The `gl` (*g*enerate *l*evels) function is specially designed for generating patterned factors for balanced experimental designs. It has three arguments: the number of levels, the block length (how many times each level should repeat), and the total length of the result. The two patterns in the data frame are thus

```
> gl(9,1,36)
 [1] 1 2 3 4 5 6 7 8 9 1 2 3 4 5 6 7 8 9 1 2 3 4 5 6 7 8 9 1 2 3 4
[32] 5 6 7 8 9
Levels:  1 2 3 4 5 6 7 8 9
> gl(4,9,36,labels=c(0,30,60,120))
 [1] 0   0   0   0   0   0   0   0   0   30  30  30  30  30  30
[16] 30  30  30  60  60  60  60  60  60  60  60  60  120 120 120
[31] 120 120 120 120 120 120
Levels:  0 30 60 120
```

Once the variables have been defined, the two-way analysis of variance is specified simply by

```
> anova(lm(hr~subj+time))
Analysis of Variance Table

Response: hr
          Df Sum Sq Mean Sq F value    Pr(>F)
subj       8 8966.6  1120.8 90.6391 4.863e-16 ***
time       3  151.0    50.3  4.0696   0.01802 *
Residuals 24  296.8    12.4
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

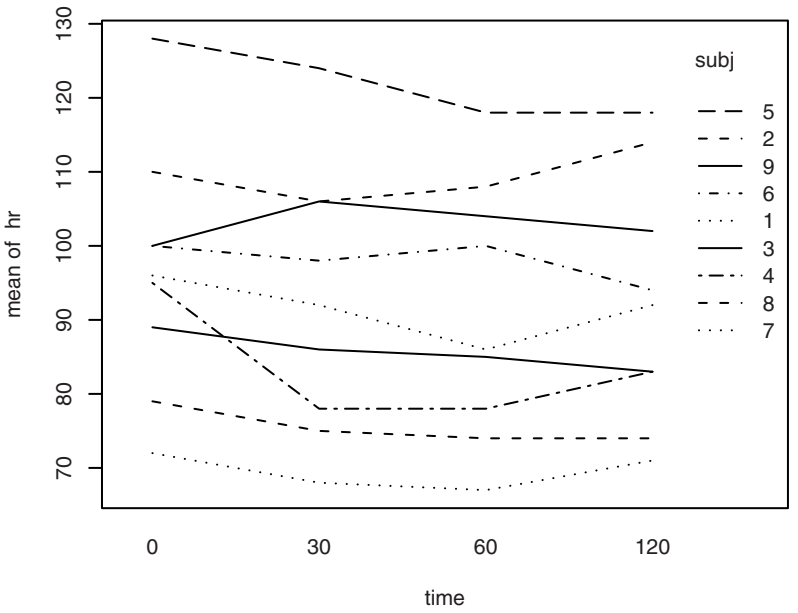

Figure 7.2. Interaction plot of heart-rate data.

Interchanging `subj` and `time` in the model formula (`hr~time+subj`) yields exactly the same analysis except for the order of the rows of the ANOVA table. This is because we are dealing with a balanced design (a complete two-way table with no missing values). In unbalanced cases, the factor order will matter.

7.3.1 *Graphics for repeated measurements*

At least for your own use, it is useful to plot a "spaghettigram" of the data; that is, a plot where data from the same subject are connected with lines. To this end, you can use the function `interaction.plot`, which graphs the values against one factor while connecting data for the other factor with line segments to form traces.

```
> interaction.plot(time, subj, hr)
```

In fact there is a fourth argument, which specifies what should be done in case, there is more than one observation per cell. By default, the mean is taken, which is the reason why the *y*-axis in Figure 7.2 reads "mean of hr".

If you prefer to have the values plotted according to the times of measurement (which are not equidistant in this example), you could instead write (resulting plot not shown)

```
> interaction.plot(ordered(time),subj,hr)
```

7.4 The Friedman test

A nonparametric counterpart of two-way analysis of variance exists for the case with one observation per cell. Friedman's test is based on ranking observations *within each row* assuming that if there is no column effect then all orderings should be equally likely. A test statistic based on the column sum of squares can be calculated and normalized to give a χ^2-distributed test statistic.

In the case of two columns, the Friedman test is equivalent to the *sign test*, in which one uses the binomial distribution to test for equal probabilities of positive and negative differences within pairs. This is a rather less sensitive test than the Wilcoxon signed-rank test discussed in Section 5.2.

Practical application of the Friedman test is as follows:

```
> friedman.test(hr~time|subj,data=heart.rate)

        Friedman rank sum test

data:  hr and time and subj
Friedman chi-squared = 8.5059, df = 3, p-value = 0.03664
```

Notice that the blocking factor is specified in a model formula using the vertical bar, which may be read as "time within `subj`". It is seen that the test is not quite as strongly significant as the parametric counterpart. This is unsurprising since the latter test is more powerful when its assumptions are met.

7.5 The ANOVA table in regression analysis

We have seen the use of analysis of variance tables in grouped and cross-classified experimental designs. However, their use is not restricted to these designs but applies to the whole class of *linear models* (more on this in Chapter 12).

The variation between and within groups for a one-way analysis of variance generalizes to *model variation* and *residual variation*

$$\mathrm{SSD}_{\mathrm{model}} = \sum_i (\hat{y}_i - \bar{y}_\cdot)^2$$

$$\mathrm{SSD}_{\mathrm{res}} = \sum_i (y_i - \hat{y}_i)^2$$

which partition the total variation $\sum_i (y_i - \bar{y}_\cdot)^2$. This applies only when the model contains an intercept; see Section 12.2. The role of the group means in the one-way classification is taken over by the fitted values $\hat{y}_i$ in the more general linear model.

An F test for significance of the model is available in direct analogy with Section 7.1. In simple linear regression, this test is equivalent to testing that the regression coefficient is zero.

The analysis of variance table corresponding to a regression analysis can be extracted with the function `anova`, just as for one- and two-way analyses of variance. For the `thuesen` example, it will look like this:

```
> attach(thuesen)
> lm.velo <- lm(short.velocity~blood.glucose)
> anova(lm.velo)
Analysis of Variance Table

Response: short.velocity
              Df  Sum Sq Mean Sq F value Pr(>F)
blood.glucose  1 0.20727 0.20727   4.414 0.0479 *
Residuals     21 0.98610 0.04696
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Notice that the F test gives the same p-value as the t test for a zero slope from Section 6.1. It is the same F test that gets printed at the end of the `summary` output:

```
...
Residual standard error: 0.2167 on 21 degrees of freedom
Multiple R-Squared: 0.1737,     Adjusted R-squared: 0.1343
F-statistic: 4.414 on 1 and 21 DF, p-value: 0.0479
```

The remaining elements of the three output lines above may also be derived from the ANOVA table. "Residual standard error" is the square root of "Residual mean squares", namely $0.2167 = \sqrt{0.04696}$. R^2 is the proportion of the total sum of squares explained by the regression line, $0.1737 = 0.2073/(0.2073 + 0.9861)$; and, finally, the adjusted R^2 is the relative improvement of the residual variance, $0.1343 = (v - 0.04696)/v$, where $v = (0.2073 + 0.9861)/22 = 0.05425$ is the variance of `short.velocity` if the glucose values are not taken into account.

7.6 Exercises

7.1 The `zelazo` data are in the form of a list of vectors, one for each of the four groups. Convert the data to a form suitable for the use of `lm`, and calculate the relevant test. Consider t tests comparing selected subgroups or obtained by combining groups.

7.2 In the `lung` data, do the three measurement methods give systematically different results? If so, which ones appear to be different?

7.3 Repeat the previous exercises using the `zelazo` and `lung` data with the relevant nonparametric tests.

7.4 The `igf1` variable in the `juul` data set is arguably skewed and has different variances across Tanner groups. Try to compensate for this using logarithmic and square-root transformations, and use the Welch test. However, the analysis is still problematic — why?