Series 2

1. A plant scientist wants to test the effect of a new herbicide on lentils. He considers the following treatments:

1. Control (without weeding, no fertilizer) \( TR = 1 \)
2. weeding by hand \( TR = 2 \)
3. spraying with herbicide before \( TR = 3 \)
4. spraying with herbicide after \( TR = 4 \)
5. weeding by hand + fertilizer \( TR = 5 \)
6. spraying with herbicide before + fertilizer \( TR = 6 \)
7. spraying with herbicide after + fertilizer \( TR = 7 \)

The data is available at ("http://stat.ethz.ch/Teaching/Datasets/WBL/lentil.dat"). In this exercise, we only consider the variable \( TR \) (treatment) and as response we consider the variable \( Y \) (the harvesting weight).

a) Check with an analysis of variance if there are differences between treatments. In addition, check the model assumptions by analysing the residuals (Tukey-Anscombe plot, QQ-plot).

**R hints:**
- \( d\.len \leftarrow \text{read.table}("http://stat.ethz.ch/Teaching/Datasets/WBL/lentil.dat", header = TRUE) \)
- \( d\.len\$TR \leftarrow \text{factor}(d\.len\$TR) \)
- Use the function \( \text{stripchart}() \) to plot the data (\( TR \) on the x-axis and \( Y \) on the y-axis).

b) In order to detect existing differences between treatments, we consider the following contrasts:

<table>
<thead>
<tr>
<th>Contrast</th>
<th>( c_1 )</th>
<th>( c_2 )</th>
<th>( c_3 )</th>
<th>( c_4 )</th>
<th>( c_5 )</th>
<th>( c_6 )</th>
<th>( c_7 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>-6</td>
<td>+1</td>
<td>+1</td>
<td>+1</td>
<td>+1</td>
<td>+1</td>
<td>+1</td>
</tr>
<tr>
<td>L2</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
<td>+1</td>
<td>+1</td>
<td>+1</td>
<td>+1</td>
</tr>
<tr>
<td>L3</td>
<td>0</td>
<td>+2</td>
<td>-1</td>
<td>-1</td>
<td>+2</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>L4</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>+1</td>
<td>0</td>
<td>-1</td>
<td>+1</td>
</tr>
<tr>
<td>L5</td>
<td>0</td>
<td>-2</td>
<td>+1</td>
<td>+1</td>
<td>+2</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>L6</td>
<td>0</td>
<td>0</td>
<td>+1</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td>+1</td>
</tr>
</tbody>
</table>

Are these contrasts orthogonal? Which questions do they address?

c) Test the contrasts.

**R hints:**
- **Contrasts:** Generate a matrix \( \text{mat.contr} \) with the contrasts L1 to L6 in the rows (either use \( \text{rbind()} \) or \( \text{matrix}() \)). Then use the following R code:
  - \( \text{library(multcomp)} \)
  - \( \text{fit.mc <- glht(fit.len, linfct = mcp(TR = mat.contr))} \)
- **ANOVA tables for contrasts:**
  - \( \text{summary(fit.mc, test = adjusted("none"))} \)

(Source: R.G. Peterson, *Agricultural Field experiments - Design and Analysis*, 1994, p. 113)

2. The response time (in milliseconds) was determined for three types of electrical circuits. The results were:

<table>
<thead>
<tr>
<th>Type</th>
<th>9</th>
<th>12</th>
<th>10</th>
<th>8</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type 2</td>
<td>20</td>
<td>21</td>
<td>23</td>
<td>17</td>
<td>30</td>
</tr>
<tr>
<td>Type 3</td>
<td>6</td>
<td>5</td>
<td>8</td>
<td>16</td>
<td>7</td>
</tr>
</tbody>
</table>
a) Test the hypothesis, that all types have the same expected response time.
   
   **R hints:**
   We want to construct a data.frame with two columns, one for the factor and one for the response.
   ```
   y <- c(9, 12, 10, 8, 15,
          20, 21, 23, 17, 30,
          6, 5, 8, 16, 7)
   type <- c(...)
   circ <- data.frame(Type = type, Y = y)
   circ$Type <- as.factor(circ$Type)
   ```

b) Use Tukey’s method to compare all pairs of treatment means.
   
   **R hint:**
   TukeyHSD(circ.fit, which = "Type", conf.level = ...)

c) Construct a set of orthogonal contrasts, starting with the comparison of circuit Type 2 with the other two.

d) Test the contrasts using a Bonferroni correction.
   
   **R hints:**
   ```
   library(multcomp)
   mat.contr <- rbind(...)
   circ.mc <- glht(circ.fit, linfct = mcp(Type = mat.contr))
   summary(circ.mc, test = adjusted("bonferroni"))
   ```

3. Fuel consumption of five types of cars (variable AUTO) has been measured in three cities of the US (variable STADT, values between 1 and 3). For each combination the investigators carried out 3 test drives. The response variable (variable KMP4L) is the distance in km that can be covered by 4 liters of gasoline. The data set is available online:

a) Visualize the relationship of variable KMP4L and variable AUTO by a joint scatterplot. Use different symbols or colors for the various cities.
   
   **R hints:**
   The function plot() has options for different symbols and colors. In our case, you can use the arguments pch = as.numeric(...$STADT) or col = as.numeric(...$STADT) to plot the cities with different symbols or colors.

b) Perform a two-way ANOVA on the target variable KMP4L. Use the full model with interaction.
   
   **Hint:** First make sure that all factors in the data.frame are indeed encoded as factors.

c) Analyse the residuals.

d) Visualize the cell means in an interaction plot. How could we explain the significant interaction?
   
   **R hints:**
   An interaction plot can be generated with interaction.plot. Look at the help page to find out how to use this function. The arguments x.factor, trace.factor and response are the most important ones.

e) Perform an individual one-way ANOVA for each city.
   
   **R hints:**
   With the function subset you can select a subset of a data.frame.

f) Repeat exercise b) without the values of the city San Francisco.


4. To study the effect of cigarette smoking on blood platelet aggregation, scientists drew blood samples from 11 individuals before and after they smoked a cigarette and measured the extent to which the blood platelets aggregated. Platelets are involved in the formation of blood clots, and it is known that smokers suffer from disorders involving blood clots more often than nonsmokers do. The data set smoking.dat can be found at http://stat.ethz.ch/Teaching/Datasets/WBL/smoking.dat and contains the following variables:

   **PERSON** person 1, …, 11
   **PERIODE** before (= 1), after (= 2)
   **AGGREG** maximum percentage of all platelets that aggregated after being exposed to a stimulus.
a) Get a first overview of the data by drawing an interaction plot. Then, perform a two-way ANOVA to investigate whether there is a significant difference in the clotting of platelets before and after smoking.
Which formula is suitable here?
AGGREG ∼ PERSON * PERIODE or AGGREG ∼ PERSON + PERIODE?

b) Test the same question with a t-test for paired samples.

c) Compare the results of a) and b).

d) (optional, for math interested people) Show the equivalence of the two-way ANOVA with single replicates and the t-test for paired samples.

(Source: John A. Rice, Mathematical Statistics and Data Analysis, Duxbury Press, 1995)

5. (optional)
We use F-distributions (named after Ronald Aylmer Fisher (1890-1962)) to perform significance tests for ANOVA. The F-distribution is the distribution of a quotient of two random variables. The numerator is the mean square of n independent standard normal distributed random variables. The denominator is the mean square of m independent standard normal distributed random variables. We call the parameter n “degrees of freedom of the numerator” and m “degrees of freedom of the denominator”. To develop some intuition for both parameters of the F-distribution, we will investigate some of the (infinitely many) F-distributions.

a) Plot the probability density functions of the following F-distributions:
$F_{3,1} \quad F_{3,5} \quad F_{3,10} \quad F_{3,20}$
How does increasing the degrees of freedom of the denominator affect the shape of the F-distribution?

R hint:
• The command `df(x, df1 = 3, df2 = 5)` returns the value of the probability density function at value x with 3 and 5 as degrees of freedom.
• You can plot the graph of a function by using `curve(function of x, xlim = ..., ylim = ...)`, where x is the input variable.
• To compare several functions graphically, use `par(mfrow = c(..., ...))` or `curve(..., add = TRUE, lty = ..., lwd = ..., col = ...)`.

b) Calculate the 95% quantile of every F-distribution in a). What does this mean for “achieving significance”?

R hint:
`pf()` returns the cumulative distribution function and `qf()` the quantiles of the corresponding F-distribution. This is analogous to the commands `dnorm()`, `pnorm()` and `qnorm()` for the standard normal distribution.

c) Analogously as in a), investigate the following F-distributions:
$F_{1,20} \quad F_{5,20} \quad F_{10,20} \quad F_{20,20}$
How does increasing the degrees of freedom of the numerator change the shape of the F-distribution?

d) Calculate the 95% quantile for every distribution in c).

e) Calculate the p-value for every distribution in c), when the F-value is 2.37.

R hint:
`pf(q, ..., lower.tail = FALSE)` returns the integral of the probability density function on the right hand side of q, i.e. the p-value.

f) In ANOVA, what influences the degrees of freedom of the numerator? What about the degrees of freedom of the denominator?

Preliminary discussion: Monday, October 23.
Deadline: Thursday, October 26 by 12:00.