Experimental Design and Data Analysis: Assignment 3

This assignment consists of 7 exercises. Throughout this assignment tests should be performed using a level of 0.05. You will need to install the R-packages multcomp and lme4, which are not included in the standard distribution of R. Installing is a one-time task, loading the package is necessary at every R session. To use the package either choose multcomp and lme4 in menu Packages or type

```
> library(multcomp)
> library(lme4)
```

EXERCISE 1

If left alone bread will become moldy, rot or decay otherwise. To investigate the influence of temperature and humidity on this process, the time to decay was measured for 18 slices of white bread, which were placed in 3 different environments and humidified or not. The data are given in the file bread.txt, with the first column time to decay in hours, the second column the environment (cold, warm or intermediate temperature) and the third column the humidity.

1. The 18 slices came from a single loaf, but were randomized to the 6 combinations of conditions. Present an R-code for this randomization process.
2. Make two boxplots of hours versus the two factors and two interaction plots (keeping the two factors fixed in turn).
3. Perform an analysis of variance to test for effect of the factors temperature, humidity, and their interaction. Describe the interaction effect in words.
4. Which of the two factors has the greatest (numerical) influence on the decay? Is this a good question?
5. Check the model assumptions by using relevant diagnostic tools. Are there any outliers?

EXERCISE 2

A researcher is interested in the time it takes a student to find a certain product on the internet using a search engine. There are three different types of interfaces with the search engine and especially the effect of these interfaces is of importance. There are five different types of students, indicating their level of computer handling (the lower the value of this indicator, the better the computer handling of the corresponding student). Fifteen students are selected; three from each group with a certain level of computer handling.

1. Number the selected students 1 to 15 and show how the students could be randomized to the interfaces in a randomized block design, by using R.

The experiment was run according to a randomized block design, as described. The data is given in the file search.txt.

2. Make some graphical summaries of the data. Are any interactions between interface and skill apparent?

3. Test the null hypothesis that the search time is the same for all interfaces. (Beware that the levels of the factors are coded by numbers!)

4. Estimate the time it takes a typical user of skill level 4 to find the product on the website if the website uses interface 3.

5. Make diagnostic plots to test the assumptions for the analysis. Comments?

6. Perform the non-parametric Friedman test to test whether there is an effect of interface.

7. Test the null hypothesis that the search time is the same for all interfaces by a one-way analysis of variance test, ignoring the variable `skill`. Is it right/wrong or useful/not useful to perform this test on this dataset? What assumption on the way the data were obtained is necessary for this test to be valid, and was this assumption met?

EXERCISE 3

The file `cream.txt` contains data on an experiment to produce sour cream. Yogurt was placed in sweet cream, and yogurt bacteria were allowed to develop. Interest was in their number. Bacteria produce lactic acid, and as a surrogate for their number, the acidity of the cream was measured. Interest was in the effect of the type of yogurt used as a `starter`. The mixtures of yogurt and sweet cream were kept at constant temperature in a yogurt maker, in which five different `positions` could be used. The experiment was carried out with five `batches` of sweet cream, which were meant to have the same composition. With each batch each of five types of `starter` was used, with the yogurt placed in one of the five positions. The combinations of levels of three factors formed a three-dimensional latin square.

1. Analyse the data in a three-way experiment without interactions: use the model formula `acidity~starter+batch+position`. (Beware to include the explanatory variables as *factors* in the analysis.) Formulate the conclusions.

2. Produce a table of p-values for testing all hypotheses $H_0 : \alpha_i = \alpha_{i'}$ on equality of differences of the main effects for `starter` simultaneously ($i, i' \in \{1, 2, \ldots, 5\}$). Which starters lead to significantly different acidity? Interpret.

3. A p-value for the test $H_0 : \alpha_2 = \alpha_1$ is also in the output of `summary` in 1). What is it? Is it coincidence that it is smaller than the simultaneous p-value?

4. Produce a table of confidence intervals for testing all differences $\alpha_i - \alpha_{i'}$ of the main effects for `starter` with simultaneous confidence level 95% ($i, i' \in \{1, 2, \ldots, 5\}$). Which intervals do not contain the number 0? Comment.

EXERCISE 4
In a study on the effect of feedingstuffs on lactation a sample of nine cows were fed with two types of food, and their milk production was measured. All cows were fed both types of food, during two periods, with a neutral period in-between to try and wash out carry-over effects. The order of the types of food was randomized over the cows. The observed data can be found in the file `cow.txt`, where `A` and `B` refer to the types of feedingstuffs.

1. Test whether the type of feedingstuffs influences milk production using an ordinary "fixed effects" model, fitted with `lm`.

2. Estimate the difference in milk production.

3. Repeat 1. and 2. by performing a mixed effects analysis, modelling the cow effect as a random effect (use the function `lmer`). Compare your results to the results found by using a fixed effects model.

4. Study the commands:

```
> attach(cow)
> t.test(milk[treatment=="A"],milk[treatment=="B"],paired=TRUE)
```

   Does this produce a valid test for a difference in milk production? Is its conclusion compatible with the one obtained in 1.? Why?

EXERCISE 5
The file `nauseatable.txt` contains data about post-operative nausea after medication against nausea. Two different medicines were administered to patients that complained about post-operative nausea. One of the medicines, Pentobarbital, was administered in two different doses.

1. Set up a `data.frame` in `R` existing of two columns and 304 rows. One column should contain an indicator whether or not the patient in that row suffered from nausea, and the other column should indicate the medicin. (Use `nausea.frame=data.frame(nausea,medicin)` where `nausea` is a vector 0's and 1's and `medicin` is the vector containing the medicin labels for each patient. Make sure these columns match correctly.)

2. Study the outcome of `xtabs(~medicin+naus)`.

3. Perform a permutation test in order to test whether the different medicins work equally well against nausea. Permute the medicin labels for this purpose. Use as test statistic the chisquare test statistic for contingency tables, which can be extracted from the output of `chisq.test`: `chisq.test(xtabs(~medicin+nausea))[[1]]`.

4. Compare the p-value found by the permutation test with the p-value found from the chisquare test for contingency tables. Explain the difference/equality of the two p-values.

EXERCISE 6

This exercise concerns the data in the file `airpollution.txt`. Investigate which explanatory variables need to be included into a linear regression model with `oxidant` as the response variable. Do this as follows.

1. Make scatter plots of the candidate explanatory variables against each other and against the response variable (see the `R`-function `pairs()`). Interpret the plots. Do you judge a linear model to be useful here?

2. Determine for each of the explanatory variables the simple linear regression model. Choose the best among these models, and stepwise extend this model by adding one explanatory variable per step on the basis of the determination coefficient. Use a test to investigate whether the extensions are useful. Determine in this way an appropriate linear regression model for these data.

3. Estimate the parameters in the full linear regression model with all explanatory variables in it. Now stepwise decrease this full model with the aid of tests of the form $H_0 : \beta_i = 0$. Determine in this way an appropriate linear regression model for the data.

4. Present the estimates of the parameters of the final model of your choice.

5. Investigate the normality of the residuals of the chosen model. Do you think, in view of all results, that the chosen linear model is appropriate?

EXERCISE 7

The data in `expensescrime.txt` were obtained to determine factors related to state expenditures on criminal activities (courts, police, etc.) The variables are: `state` (indicating the state in the USA), `expend` (state expenditures on criminal activities in $1000), `bad` (number of persons under criminal supervision), `crime` (crime rate per 100000), `lawyers` (number of lawyers in the state), `employ` (number of persons employed in the state) and `pop` (population of the state in 1000). Perform a regression analysis using `expend` as response variable and `bad`, `crime`, `lawyers`, `employ` and `pop` as independent variables. Present your final model. Your analysis should at least include:

a. investigation of potential or influence points

b. investigation of problems due to collinearity

c. investigation of residuals.

You may use all techniques mentioned on the lecture slides. State clearly all the choices you make during the regression analysis, including arguments for all your choices. (Note that there are several strategies possible!)