Exercises for Day 5

Used datasets and R scripts can be downloaded from the course homepage:

`https://phillipmogensen.github.io/SmB-I-2020/courseplan.html`

Exercise 5.1: *Hypertension in diabetic patients.*

In this exercise we will analyse the dataset from Exercise 1.5 using an ANCOVA with a random effect. This is an alternative to the four T-tests done in Exercise 1.5 (two of these tests did the comparison of the two drugs, and gave the p-values 0.0932 and 0.1108). But first let us recap the description of the dataset:

An experiment on 19 diabetic patients was conducted in order to compare the effects of two drugs called *Drug E* and *Drug N* on the treatment of high blood pressure. The study is a cross-over study. This means that all patients try both drugs in two different study periods. Both study periods lasted for 14 days. In between the two study periods was a wash-out period, which also lasted for 14 days. The patients were randomly assigned to two groups called *E/N* and *N/E*. The patients in the E/N-group received drug E in the first study period and drug N in the second study period. The patients in the N/E-group received drug N in the first study period and drug E in the second study period.

The systolic and the diastolic blood pressure was measured for all the patients at the beginning and the end of both study periods. In this exercise we will only analyse the observations of the systolic blood pressure. These observations are shown in Table 1. The observations for the diastolic blood pressure may be found on the internet (search on the reference given at the end of this exercise).

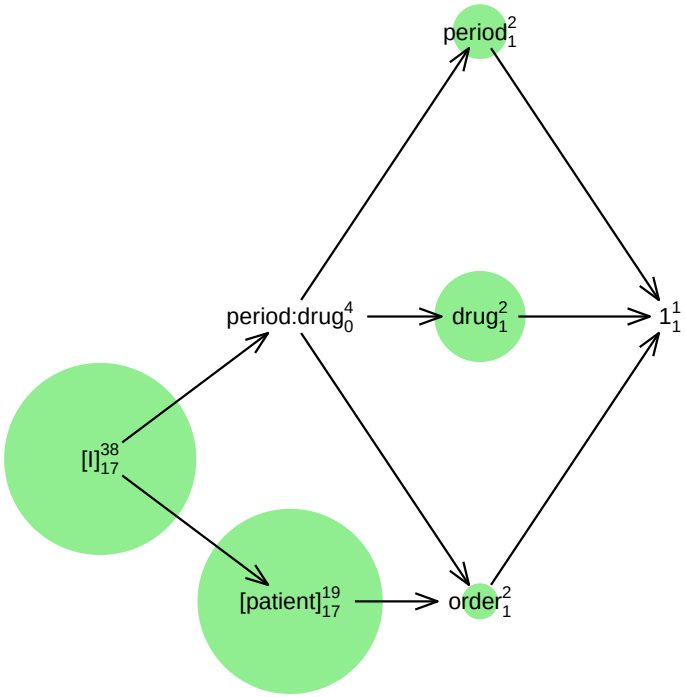
The dataset may be found in the text file `hypertension.txt`. Here the dataset is on the so-called *long form* (also called the *vertical form*), i.e. with one response measurement in each row. The dataset contains the following variables:

Variable	**Type**	**Range**	**Usage**
drug	nominal	E, N	??
period	nominal	1, 2	??
order	nominal	E/N, N/E	??
patient	nominal	8,. . . ,26	??
baseline	continuous	$[107; 156]$	??
end	continuous	$[91; 154]$	??
change	continuous	$[-25; 25]$	??

Please answer the following 5 questions:

1. In the table above replace the *"??"*'s by either *"fixed effect"*, *"random effect"*, *"response"*, or *"not used"*.

2. In Exercise 1.5 two of the T-tests were done to see if there was an effect of `order` or an interaction between `period` and `drug`. For the cross-over study design to be successful neither of these should be significant. To investigate these potential problems in an ANOVA model we would incorporate the main effect of `order` and the interaction `period:drug` in the model. Hence a model with the following design diagram[1]:



Inspecting the design diagram we see that the interaction `period:drug` has zero degrees of freedom. This means that this interaction does not contribute to the model at all. Can you explain why? Or posed differently; why does it not make sense to include both the main effect of `order` and the interaction `period:drug` at the same time?

Hint: What is the relation between `period:drug` and `order`?

[1] In this design diagram the area of the green circles visualizes the proportion of total variance explained by the different terms in the model.

Patient id	Treatment order	Systolic blood pressure			
		Baseline 1	End 1	Baseline 2	End 2
9	Drug E, Drug N	124	136	120	145
21	Drug E, Drug N	120	132	138	126
8	Drug E, Drug N	115	96	111	91
12	Drug E, Drug N	134	118	123	123
16	Drug E, Drug N	131	106	111	123
19	Drug E, Drug N	119	108	113	112
20	Drug E, Drug N	124	112	108	112
24	Drug E, Drug N	127	113	121	143
13	Drug N, Drug E	113	113	107	97
17	Drug N, Drug E	132	109	122	119
18	Drug N, Drug E	129	133	139	130
23	Drug N, Drug E	124	120	127	118
25	Drug N, Drug E	112	103	112	121
10	Drug N, Drug E	124	112	128	122
11	Drug N, Drug E	144	154	156	137
14	Drug N, Drug E	134	118	122	109
15	Drug N, Drug E	119	118	115	114
22	Drug N, Drug E	123	123	114	108
26	Drug N, Drug E	122	123	124	120

Table 1: Treatment of hypertension in diabetic patients

3. Analyse the dataset. Possibly also using `baseline` as a fixed effect.

   Hint: If you want to include `baseline` in the model you simply add it to the right hand side of the "$\sim$" in the model equation. Furthermore, remember that `period` and `patient` should be used as categorical factors in the model.

4. Is this analysis more powerful than the analysis done in Exercise 1.5?

5. In your opinion is this analysis more easy to communicate compared to the analysis done in Exercise 1.5?

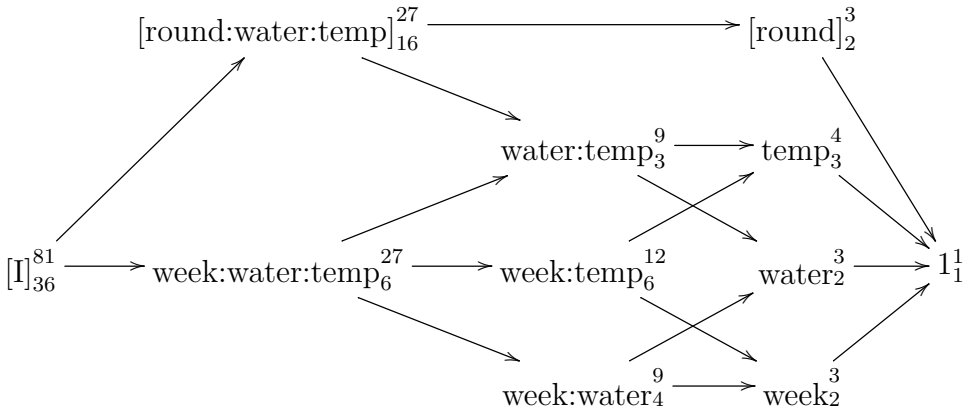
(Reference: Bradstreet, T.E. (1994) "Favorite Data Sets from Early Phases of Drug Research - Part 3." *Proceedings of the Section on Statistical Education of the American Statistical Association.*)

Exercise 5.2: *Random effects model.*

In an experiment about production of milk powder two factors were varied: water activity on three levels (`water`=1, 2, 3) and temperature while drying on 4 levels (`temp`=100, 110, 120, 140 degrees Celsius). Only 9 of the 12 combinations were tested in the experiment. There were three replications in the experiment, in the sense that milk powder was prepared in three

rounds (`round`=1, 2, 3). This gives 3*9=27 samples of milk powder in total. Each of these was stored and measurements were taken after 4, 6 and 8 weeks. Each time the concentration of maillard reaction products as well as a sensory taste score (high values means good taste) were measured. The dataset consists of 81 observations, which are listed in Table 2 and can be found in the text file `milk.txt`. In this exercise we will analyse the response `maillard`.

The factor `round` can be considered a block factor, and should be used as a random factor in this exercise[2]. The 3-way interaction `round:water:temp` corresponds to the grouping of the 81 observations into the 27 different samples. Since the samples are measured 3 times (i.e. repeated measurements) it is also standard to make this 3-way interaction a random factor. Thus, the design diagram for the experiment is as follows:



Recall, that the superscript designate the number of levels, and that the subscripts designate the corresponding degrees of freedom. Please have a closer look at the design diagram, which contains the random effects of `round:water:temp` and of `round`, and the full factorial design of the fixed effects `temp`, `water` and `week`. Can you see how it relates to the description of the experiment given above?

The model described above is fitted in R using the following call[3] to lmer():

```
lmer(maillard~week*water*temp+(1|round:water:temp)+(1|round),data=milk)
```

[2]This violates the "rules of thumb" that factors with less than 5 levels can be used with fixed effect. However, this is also a matter of taste. Moreover, using `round` as a random effect also let you try a model with more than one random effect.

[3]Remark: If your version of lme4 is prior to 1.1-6, then this call might result in an error message like this ``Error in lme4::lFormula(formula = maillard ~ week * water * temp + (1 | : rank of X = 27 < ncol(X) = 36''. If you encounter this error message, then you should update your installation of the lme4-package.

Round	Week	Maillard	Taste	Water	Temp	Round	Week	Maillard	Taste	Water	Temp
1	4	2.90	10.1	1	100	2	6	2.11	11.2	3	100
1	4	2.13	11.0	1	110	2	6	1.98	11.8	3	110
1	4	2.00	11.1	1	120	2	6	2.20	11.0	3	140
1	4	2.13	11.1	2	100	3	6	2.20	7.0	1	100
1	4	2.38	11.9	2	120	3	6	2.34	10.7	1	110
1	4	2.56	10.7	2	140	3	6	2.49	10.3	1	120
1	4	2.60	10.8	3	100	3	6	2.63	9.7	2	100
1	4	1.91	11.0	3	110	3	6	3.06	9.0	2	120
1	4	2.27	10.8	3	140	3	6	3.28	9.6	2	140
2	4	2.19	11.0	1	100	3	6	2.34	10.2	3	100
2	4	2.32	11.0	1	110	3	6	2.51	9.2	3	110
2	4	2.41	11.6	1	120	3	6	2.77	10.2	3	140
2	4	2.49	11.1	2	100	1	8	2.39	9.6	1	100
2	4	2.61	11.7	2	120	1	8	2.41	9.8	1	110
2	4	2.63	10.8	2	140	1	8	2.71	11.4	1	120
2	4	2.06	11.0	3	100	1	8	2.49	11.2	2	100
2	4	1.98	10.0	3	110	1	8	2.06	11.2	2	120
2	4	2.27	11.2	3	140	1	8	3.10	9.8	2	140
3	4	2.13	10.1	1	100	1	8	2.32	10.8	3	100
3	4	2.13	9.4	1	110	1	8	2.29	9.4	3	110
3	4	2.22	10.7	1	120	1	8	2.72	12.0	3	140
3	4	2.80	8.3	2	100	2	8	2.27	11.0	1	100
3	4	2.77	10.9	2	120	2	8	2.25	11.2	1	110
3	4	2.99	9.2	2	140	2	8	2.46	9.6	1	120
3	4	1.98	10.3	3	100	2	8	2.53	9.2	2	100
3	4	1.98	9.3	3	110	2	8	2.70	11.0	2	120
3	4	2.20	10.5	3	140	2	8	2.81	11.6	2	140
1	6	2.13	10.0	1	100	2	8	2.20	11.8	3	100
1	6	2.34	10.5	1	110	2	8	2.15	10.6	3	110
1	6	2.49	11.2	1	120	2	8	2.41	11.4	3	140
1	6	2.41	10.8	2	100	3	8	2.41	9.6	1	100
1	6	2.85	11.2	2	120	3	8	2.42	9.0	1	110
1	6	2.84	11.2	2	140	3	8	2.73	10.2	1	120
1	6	2.24	8.4	3	100	3	8	3.33	7.8	2	100
1	6	2.06	11.4	3	110	3	8	3.25	9.4	2	120
1	6	2.42	11.6	3	140	3	8	3.75	9.6	2	140
2	6	2.20	9.3	1	100	3	8	2.80	10.6	3	100
2	6	2.27	11.3	1	110	3	8	2.81	10.2	3	110
2	6	2.49	11.7	1	120	3	8	3.06	10.0	3	140
2	6	2.34	11.2	2	100						
2	6	2.70	10.8	2	120						
2	6	2.61	11.0	2	140						

Table 2: The milk powder data

Here we assume that the data is available in a data frame called `milk`, where the explanatory variables are encoded as factors.

Now analyse the relation between maillard and the 3 explanatory factors using the following 4 steps:

1. Fit the initial model using `lmer()`.
2. Validate the initial model as proposed in the lectures.
3. Do backward model reduction of the fixed effects using `drop1(.,test="Chisq")` and `update()` as exemplified in the lectures.
4. Report em-means for the final model.

(Reference: Exercise 8.7 in Sørensen & Tolver: *Lecture Notes for Applied Statistics.*)

Exercise 5.3: *Logistic regression with overdispersion.*

20 persons have participated in an experiment where two different diets are to be compared. By randomization 10 persons have been assigned to each diet and every week a *weight gain* or *weight loss* has been observed. The observations are the number of weeks where the diet resulted in a weight loss for each of the 20 persons in the experiment. The table below displays the results for a period of eight weeks showing the number of persons for each combination of diet and weeks with weight loss:

No. of weeks with weight loss	0	1	2	3	4	5	6	7	8
Diet 1	1	0	2	0	1	1	2	0	3
Diet 2	2	1	0	1	2	1	2	1	0

The dataset available in the text file `diet.txt` table contains four variables:

Variable	**Type**	**Range**	**Usage**
person	nominal	20 levels	random effect
diet	nominal	1, 2	fixed effect
gain	count	0,. . . ,8	response
loss	count	0,. . . ,8	response

Fit a logistic regression to the dataset and answer the following questions:

- Is there indication of overdispersion?
- What is the p-value for the effect of diet on the probability of weight loss in each week?
- What is the odds ratio for weight loss between the two diets? Please answer this question even if the effect of diet is non-significant.

Above you should find strong evidence for overdispersion. Actually this is visible by the naked eye when looking at the data table (if you know what to look for).

(Reference: Exercise 8.20 in Sørensen & Jensen: *Lecture Notes for Applied Statistics.*)

End of exercises.