Swabs

To study the effect of household conditions on the risk of infectuous deseases 18 families with 3 children each (in pre-specified age intervals) were followed over a period of time, during which repeated swabs were taken.

Data from the study is given in the table below. The outcome variable `swabs` indicate how many times the swab was positive for *pneumococcus*. We are particularly interested in the effect of `crowding`, which describes the space available for the household.

Table 8.13. Numbers of swabs positive for pneumococcus during fixed periods

Crowding category	Family serial number	Family status					
				Child			
		Father	Mother	1	2	3	Total
Overcrowded	1	5	7	6	25	19	62
	2	11	8	11	33	35	98
	3	3	12	19	6	21	61
	4	3	19	12	17	17	68
	5	10	9	15	11	17	62
	6	9	0	6	9	5	29
		41	55	69	101	114	380
Crowded	7	11	7	7	15	13	53
	8	10	5	8	13	17	53
	9	5	4	3	18	10	40
	10	1	9	4	16	8	38
	11	5	5	10	16	20	56
	12	7	3	13	17	18	58
		39	33	45	95	86	298
Uncrowded	13	6	3	5	7	3	24
	14	9	6	6	14	10	45
	15	2	2	6	15	8	33
	16	0	2	10	16	21	49
	17	3	2	0	3	14	22
	18	6	2	4	7	20	39
		26	17	31	62	76	212
Total		106	105	145	258	276	890

Data is contained in the file `swabs.txt` which can be downloaded from the course webpage.

Questions

1. Read in the data from the file. Beside the outcome `swabs` and the variable `crowding` what other variables does the dataset contain? What impact do you expect these variables to have on the outcome?

2. Firstly, we will analyse data using an appropriate **two-level model**.

   - What should the two levels of the model be? And which covariates belong to each level? In other words which of the variables in the data should be consideres as systematic effects on pneumococcus, and which are to be considered as random effects in the sense that they represent a sample from a population?
   - Which effects do you find? In particular, find the 95% confidence interval for the difference between 'overcrowded' and 'uncrowded' and for the difference between mother and youngest child (child3)?
   - Is there any evidence that some family members are more affected by crowding than others?
   - Try to make an illustration of the model, e.g. by plotting predicted values from the model for different family members and varying degrees of crowding.

   The following questions consider alternative analyses of the data, which should be compared to the two-level model:

3. What happens if we forget to take into account that data is clustered (the five family members belong to the same family)? I.e. we ignore the effect of `family` and do a two-way analysis of variance to estimate the effects of `name` and `crowding`.

4. Make a new dataset called `averages` containing the average number of positive swabs (`mswabs`) for each family. In SAS this can be obtained by writing the following code:

```
PROC MEANS NWAY NOPRINT DATA=swabdata;
CLASS family;
VAR swabs;
OUTPUT OUT=averages MEAN=mswabs;
ID crowding;
RUN;
```

Next we look at the new dataset `averages`. We want to evaluate the effect of `crowding` on the variable `mswabs`.

- What kind of statistical analysis is this?
- Give an estimate with corresponding 95% confidence interval for the difference between 'overcrowded' and 'uncrowded' households.
- Try to make a suitable illustration of the data.

5. We now return to the original data set `swabdata`. If we disregard the effect of `crowding`, we can evaluate the differences between family members while adjusting for `family`-differences is an ordinary two-way analysis of variance.

   - Is there any overall differences between the five family members?
   - Find an estimate with corresponding 95% confidence interval for the difference between mother and youngest child.
   - Try to make a suitable illustration for the analysis.

One final question:

6. Which considerations do you think lie behind the choice of design in this investigation? An alternative could be to just take a random sample of individuals.