Computer Exercise 5 (R)

1. Read the data in **Girls.txt**. The data consist of the lower face width (in cm) for 15 girls, measured at 5 and again at 6 years old. You job is to determine if the faces of girls are bigger at 6 than at 5 years old.
(a) First, perform a paired *t*-test.

```
t.test(dat$FaceW5Y, dat$FaceW6Y, paired=T)
```

Do the faces grow?

(b) To instead perform an analysis of variance with age as one factor and girl as another factor (girl is used as block), read the data in **Girls2.txt**. You need to make the variables Girl and Age into factors. To use block in the model you can load the `lme4` package: `library(lme4)` (you first need to install it). The package `lme4` in R specializes in mixed models. This package contains the main functions `lmer` and `glmer`, the mixed model equivalents of `lm` and `glm` (`glm` you will use later in this course). To perform the blocked analysis of variance with this package, you would write:

```
fm1 <- lmer(FaceW ~ Age + (1| Girl), data=dat)
```

Since we now have some additional assumptions such as the normally distributed random factor-leves (random variation in face-width among girls in this case), we should check whether our data could reasonably meet these assumptions or not. Try the package `sjPlot` (don't forget to install it using `install.package` and load it using `library`).

```
plot_model(fm1, type = "diag")
```

Use the arrows above the plot window to scroll through the diagnostic plots. What do you think, could the assumptions be met to a reasonable extent? Perhaps the residuals look a little trimodal, but that could also happen by chance. Perhaps check for significant deviations from normality in the residuals if you are unsure:

```
shapiro.test(resid(fm1))
```

Also look for nonlinearities using the following code (see slide 15from lecture 7):

```
plot_model (fm1, type = "slope", show.data = TRUE)
plot_model (fm1, type = "resid", show.data = TRUE)
```

To look at the results of the statistical test, write:

```
Anova(fm1)
```

What do you conclude and how does the conclusion compare with the paired *t*-test?

(c) Finally, do a one-way analysis of variance, with only age as factor, for instance

```
fm2 <- lm(FaceW ~ Age, data=dat)
Anova(fm2)
```

What would one have concluded from this analysis? What is the explanation for the different results?

2. Read the data in **Mcinxia3.txt**. The file contains *M. cinxia* female pupal weights. The females derive from two sites on Öland: Hildeborg and Littorinavallen. The individuals were reared in the lab from larvae collected in the wild. *M. cinxia* larvae are group-living (a larval group consists of one or several sibling groups) and those collected from a group in the wild were raised together in the lab. The variable Lgroup identifies the larval group (you need to make it into a factor). Your job is to determine if females from the two sites differ in pupal weight.

(a) First, perform a one-way analysis, using site as a factor but ignoring the larval groups, e.g.

```
    fm1 <- lm(PupW ~ Site, data=dat)
```

Do you find evidence for differences between localities in female pupal weight?

(b) Next, investigate the same question using a nested analysis of variance. To perform the nested analysis of variance using `lme4`, you would write:

```
          fm2 <- lmer(PupW ~ Site + ( 1 | Lgroup), data=dat)
```

Don't forget to check the assumptions like in the previous question!

To get the results of your model you can either just write its name or use the `summary()` function. To get F-statistics for the main effect, try the `anova()` function. However, to get the p-values, you need to use the `Anova()` function, or produce a reduced model and then compare this model to the full model. You could do it like this:

```
          fm2b <- lmer(PupW ~ 1 + ( 1 | Lgroup), data=dat)

          anova(fm2,fm2b)
```

Try to interpret the output you get using the `summary()` command. Do you find evidence for differences between localities in female pupal weight? What has happened in the analysis? It might help to produce a graph showing pupal weights of the different larval group belonging to the different locations. First load the `lattice-package` that offers great plotting functions for nested data:

```
                library(lattice)
```

Then try this plot:

```
                bwplot(PupW ~ Lgroup | Site, data=dat)
```

Last, you might want to see how well your model estimates actually reflect the observed data. Here you can yet again use the `sjPlot` package. The following line of code shows model estimates (with confidence intervals) as well as the real data:

```
plot_model(fm2, type = "eff", show.data = TRUE)
```

Do you think the model fits well? If there were more than 1 fixed effect (i.e. multiple predictors variables modelled as fixed effects), this command shows the effect of each of the predictors separately while averaging the effects of the other predictors (so-called marginal effects).

c) In the old days, people used the `aov()` function for such models. However, this is not ideal for unbalanced designs, for which the `lmer()` function may be better. However, many statisticians prefer to estimate significance of fixed effects (or any model parameter) in mixed effect models in a Bayesian analysis. You can attempt to estimate confidence limits for the fixed effect `Site` using **M**arkov **C**hain **M**onte **C**arlo simulations. The `MCMCglmm` package allows you to do just this. The way of writing the model is slightly different, but there is a need to specify fixed and random effects in the model just as for `lmer()`.

```
mcmcfm <- MCMCglmm(PupW ~ Site, random = ~Lgroup, data=dat)
```

and to look at your data:

```
summary(mcmcfm)
```

The summary output gives you a p-value for the fixed effects. How does this compare to the effect of site you got through the `lmer` method? For a Bayesian analysis the focus is on confidence intervals rather parameter estimates, look at your model by typing:

```
plot(mcmcfm)
```

Type `summary(mcmcfm$VCV)` to look at random effects more closely, and

type `summary(mcmcfm$Sol)` to look at fixed effects more closely.

We will deal more with Bayesian methods later, but please ask your teacher for help if you feel unsure about this exercise.

3. Read the data in **Fluoride.txt**. Three water samples were taken from each of three localities (lakes; the localities are numbered from 1 to 3 and water samples from 1 to 9). Two determinations of fluoride content were performed on each of the nine water samples. First, test if fluoride content varies between localities. Use the commands

```
     fm <- aov(Fluoride ~ Loc + Error(WtrSmpl), data=dat)
```

```
summary(fm)
```

Second, find estimates of the *true variance components*, i.e. the within water sample variance σ^2 (using the output from the summary command) and the among water samples within locality variance $\sigma^2_{\beta(\alpha)}$ (using the output from the summary command and a formula). The formula to use for the latter estimate is

$$s^2_{\beta(\alpha)} = \frac{1}{n}\left(MS_{among\ samples\ within\ locality} - MS_{within\ sample}\right)$$

where n is the number of observations per water sample.

For the first variance component (σ^2 , within water sample variance), we can just look at the model-output's "within" residuals. That is the variance at the "deepest" level of the model, which in this case is technical replicates (these give us an idea of the within water sample variation). For the second ($\sigma^2_{\beta(\alpha)}$), we need to look at the variance among water samples (the residuals under "Error: WtrSmpl") *in light of the variance among technical replicates*. In other words, the model gives an estimate of variation among water samples within locality, but to know the true variation caused by water sample variability, we need to account for the fact that some of that variability could be caused by, for example, measurement inaccuracy.

Now compare the among water sample variance with the estimate you get from `lmer`:

```
fm2 <- lmer(Fluoride ~ Loc + (1|WtrSmpl), data=dat)
summary(fm2)
```

Note that you can also formally test for significance of the random effects:

```
library(lmerTest)
ranova(fm2)
```

Third, do you have an opinion about how one ought to allocate sampling effort in this case? Assuming that it is fairly “cheap” to get additional determinations from the same water sample, how many should one get? On the other hand, assuming that the main effort or cost lies in determining fluoride content, how should one proceed?

4. Keep using the data in **Fluoride.txt**. In the previous exercise, you looked for differences between localities in fluoride content using a nested analysis of variance. Your job now is to investigate the same question using a nonparametric test. How do you go about it and which test do you use? (Hint: maybe you could use an average for each water sample. You can derive water sample means using the `tapply()` function. Try writing `?tapply` to find out more about this function, or look at the previous exercise 3 if you can’t remember). (**a**) What is your conclusion about differences between localities and how does it compare with your previous analysis using the nested analysis of variance? (**b)** Finally, perform an anova on the new data on means using `lm()`. How do the variance components compare?