Anova exercise class

Sylvain

1st of December 2014

Series 4

Again, I didn't correct it myself. Please send me an email or ask me in person if something is unclear.

There seem to be some problem with the computation of interactions. I noticed them only this morning and couldn't figure out everything yet.

In general, be careful with what contrast is used!

Contrasts

The ANOVA table doesn't change, but the estimated effects and their interpretation do!

In class the contrast we use the most often is the **sum contrast**: $\sum A_i = 0$.

The default in R is to use treatment contrast: $A_1 = 0$.

In R, you can change the default contrast with `options(contrasts=c('contr.sum','contr.poly'))`

Problem with computing interactions comes from here too, but I still need to work out the detail.

Series 5

Exercise 1: Random effects

Experiment:

Study the quantity of *moisture* in pigment pastes. Think about quality testing.

Design: 15 batches, with 2 samples each, analyzed twice.

Load the data

```
paint <- read.table(file="http://stat.ethz.ch/Teaching/Datasets/paint.txt",header=TRUE)
paint$SAMPLE <- as.factor(paint$SAMPLE)
paint$BATCH <- as.factor(paint$BATCH)
head(paint, n=8)
```

```
##   BATCH SAMPLE REP MOISTURE
## 1     1      1   1       40
## 2     1      1   2       39
## 3     1      2   1       30
## 4     1      2   2       30
## 5     2      1   1       26
## 6     2      1   2       28
## 7     2      2   1       25
## 8     2      2   2       26
```

Try to plot MOISTURE vs. BATCH, with the color varying according to SAMPLE.

Random effects

Simple case with one random factor a:

$$Y_{ij} = \mu + a_i + \epsilon_{ij} \text{where: } a_i \sim \mathcal{N}(0, \sigma_a^2)$$

Think of it as a hierarchical model to generate an observation i:

1. Generate a_i normally distributed.
2. Given a_i, generate an additonal noise term ϵ_{ij}.

Two sources of variability!

Random effects

Some reasons to consider an effect as random vs. fixed:

- We are interested in variability and not in the effect of a factor
- We want to generalize to the whole population
- The level was indeed chosen *randomly* (not necessary)

Nested design

$$Y_{ijk} = \mu + a_i + +b_{j(i)} + \epsilon_{k(ij)}$$

Two factors are nested if not all levels of the second factors are tested for each level of the first factor.

Is is the case here?

Would R recognize it automatically?

Mixed-effects models

So far we studied mainly fixed effect model and now random effects.

In practice, what happens most often is a *mix* of both!

1. Some treatment you are interested in the effects
2. Some blocking factors that are considered fixed
3. Some factors that are considered random

Question: what is the difference between considering an effect (like let's say BATCH), as random or as block?

With R:

Two possibility:

1. *By hand* with `aov` and manipulation of the output: See hint and solution of the exercise and script p.67-68

2. Straight to what you want with `lme4`:

```r
library(lme4)
## one effect:
mod1 <- lmer(Y ~ 1 + (1 | a), data=dat)
## b nested in a:
mod2 <- lmer(Y ~ 1 + (1 | a/b), data=dat)
## mixed effect: a fixed, b (nested in a) is random:
mod3 <- lmer(Y ~ a + (1 | a:b), data=dat)
```

Exercise 2: latin squares

Experiment

Compare three new varieties of peanuts to a standard one.

Because the experimental conditions vary in the terrain, we have to account for it. We create a factor east-west (Row) and a factor north-south (Column) with 4 levels each.

Latin square vs. randomized block design

Why not simply randomized?

Latin square allows to do blocking of 2 factors at once, even when there are physical constraints (like here: you can have only one plant in one spot…).

Load the data

```r
peanut <- read.table(file="http://stat.ethz.ch/Teaching/Datasets/Peanut.txt",header=TRUE)
peanut$Row <- as.factor(peanut$Row)
peanut$Column <- as.factor(peanut$Column)
```

```
peanut$Treatment <- as.factor(peanut$Treatment)
head(peanut)
```

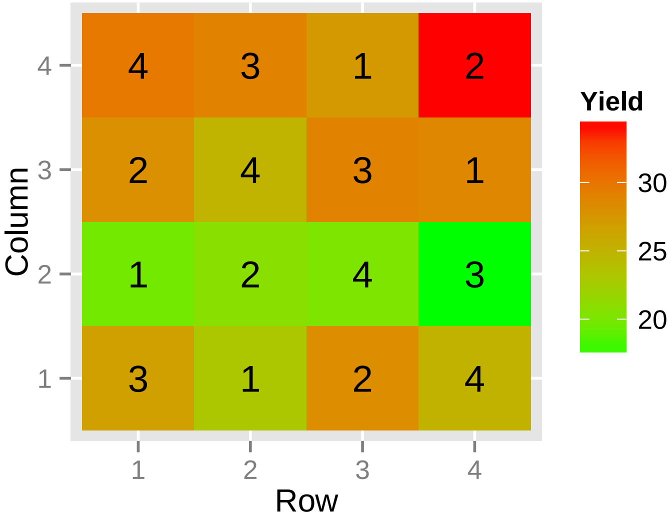
```
##   Row Column Treatment Yield
## 1   1      1         3  26.7
## 2   2      1         1  23.1
## 3   3      1         2  28.3
## 4   4      1         4  25.1
## 5   1      2         1  19.7
## 6   2      2         2  20.7
```

Plot the data

Try the usual boxplot variable by variable.

Also this might be interesting:

```
library(ggplot2)
qplot(x=Row, y=Column, fill=Yield, label=Treatment, data=peanut, geom='tile') +
  scale_fill_gradient(low="green", high="red") +
  geom_text()
```



ANOVA

Perform the analysis of variance in the usual way. Don't forget to add all the factors in the formula!

To test if any particular treatment has a significantly higher yield than the reference one, use the function `TukeyHSD` and look at the confidence intervals for the treatment differences that we are interested in.

Second experiment

The experiment is replicated in three different locations with the same latin square design. We have a new factor `Rep`.

Which factors are nested?

Load the data

```
peanut2 <- read.table(file="http://stat.ethz.ch/Teaching/Datasets/Peanut2.txt",header=TRUE)
peanut2$Row <- as.factor(peanut2$Row)
peanut2$Column <- as.factor(peanut2$Column)
peanut2$Treatment <- as.factor(peanut2$Treatment)
peanut2$Rep <- as.factor(peanut2$Rep)
```

Try this plot:

```
qplot(x=Row, y=Column, fill=Yield, label=Treatment, facets=.~Rep, data=peanut2, geom='tile') +
  scale_fill_gradient(low="green", high="red") +
  geom_text()
```

ANOVA

Fit your model. Be careful to use the correct formula: for example if `c` is nested in `d` use the function call `aov(y ~ a + c/d, data=dat)`.

To test pairwise differences, again use the function `TukeyHSD`.

Exercise 3: Crossover design

Experiment

We want to test the effect of a drug (Mortrin) against tennis elbow.

Two group of patients: A and B

Group A: Mortrin-washout-Placebo Group B: Placebo-washout-Mortrin

Question: what are the advantages/disadvantages of such a design?

Outcome

We measure 4 different outcomes, all in term of degree of pain relief compared to the begining (1-6):

1. Maximum activity pain relief
2. 12 hours after max activity pain relief
3. average activity pain relief
4. overall feeling

Remark 1: we consider each outcome separately, but in practice it might be better to look at all of them together (MANOVA).

Remark 2: we consider the outcome as continuous, even if in practice it was only measured on a discrete scale from 1-6. What do we implicitely assumed by doing so?

Load the data

This is a messy dataset, not to my taste at all…

```r
tennis <- read.table(file="http://stat.ethz.ch/Teaching/Datasets/TENNIS.dat")
names(tennis)=c("id","age","sex","order","max1","twelve1","ave1",
       "overall1","max2","twelve2","ave2","overall2","max3","twelve3","ave3","overall3")
## replace invalid values with NA:
for (i in 3:16)
  tennis[,i][tennis[,i]==9 | tennis[,i]==0]=NA
tennis$sex[tennis$sex==1] <- 'male'
tennis$sex[tennis$sex==2] <- 'female'
```

Remark: beware that e.g. max1 doesn't mean the same if you were in group 1 or 2!

Reorganize the data

Data are in a messy format in which it is very difficult to work properly.

There are different ways to rearrange the data in a better format, here I propose the more recent way to do it with `tidyr` and `dplyr`:

```r
library(dplyr)
library(tidyr)
tennis.nice <- tennis %>%
  gather(ytype_time, pain, -c(id, age, sex, order)) %>%
  separate(ytype_time, c('ytype', 'period'), sep=-2)
tennis.nice$Treatment[tennis.nice$order==1 & tennis.nice$period==1]="Motrin"
tennis.nice$Treatment[tennis.nice$period==2]="Washout"
tennis.nice$Treatment[tennis.nice$order==1 & tennis.nice$period==3]="Placebo"
tennis.nice$Treatment[tennis.nice$order==2 & tennis.nice$period==3]="Motrin"
tennis.nice$Treatment[tennis.nice$order==2 & tennis.nice$period==1]="Placebo"
tennis.nice[,c(1,3,4,5,6,8)] <- lapply(tennis.nice[,c(1,3,4,5,6,8)], as.factor)
```

Nicer data

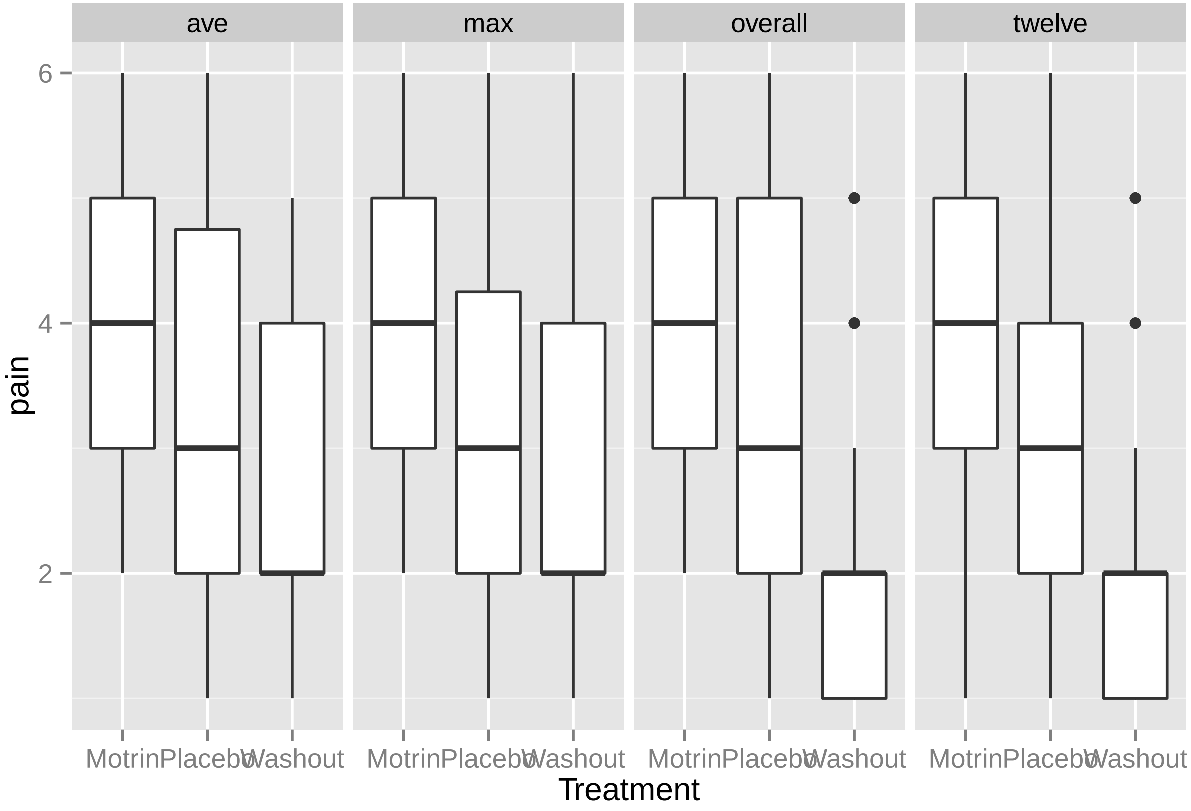
The data looks now like that:

```
head(tennis.nice, 10)
```

```
##      id age    sex order ytype period pain Treatment
## 1   701  42 female     1   max      1    5    Motrin
## 2   725  45   male     2   max      1    2   Placebo
## 3   729  43   male     1   max      1    4    Motrin
## 4   732  48   male     2   max      1    1   Placebo
## 5   733  56 female     1   max      1    5    Motrin
## 6   734  44 female     1   max      1    5    Motrin
## 7   736  31 female     2   max      1    3   Placebo
## 8   740  49 female     2   max      1    2   Placebo
## 9   741  44 female     2   max      1    2   Placebo
## 10  742  38 female     1   max      1    2    Motrin
```

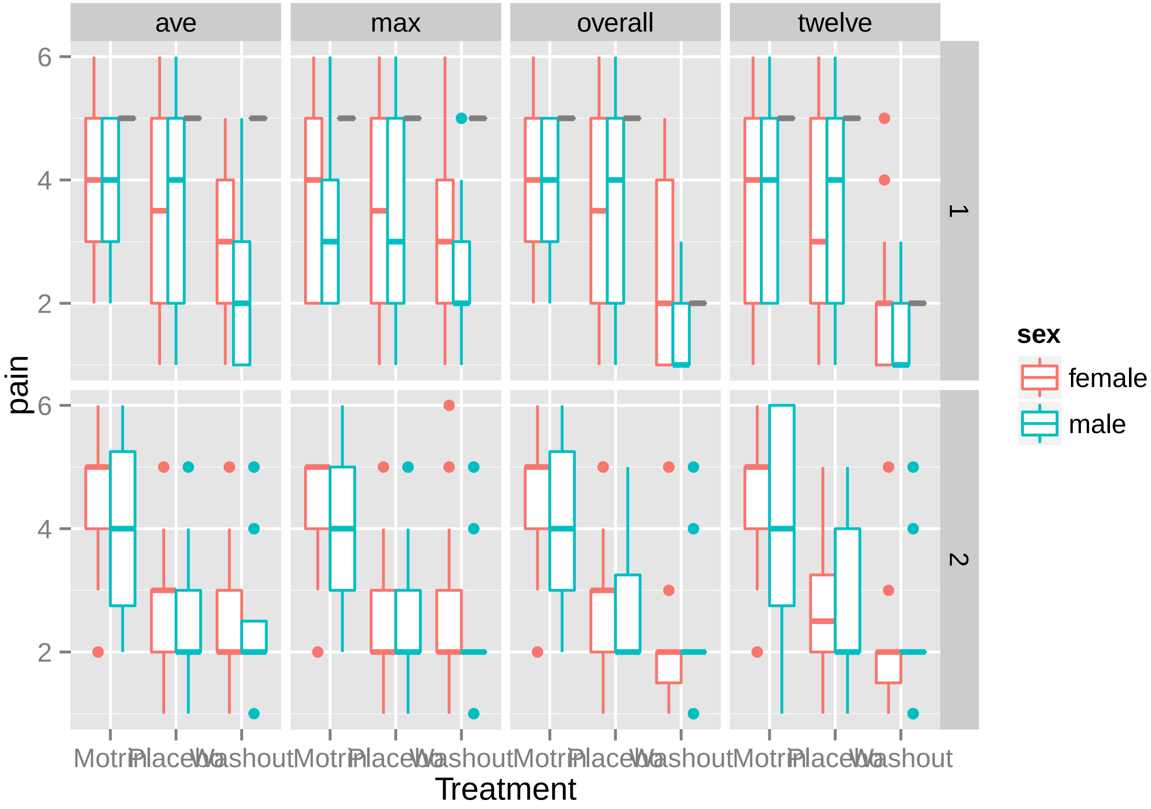
Basic plot

```
qplot(x=Treatment, y=pain, facets=.~ytype, data=tennis.nice, geom='boxplot')
```



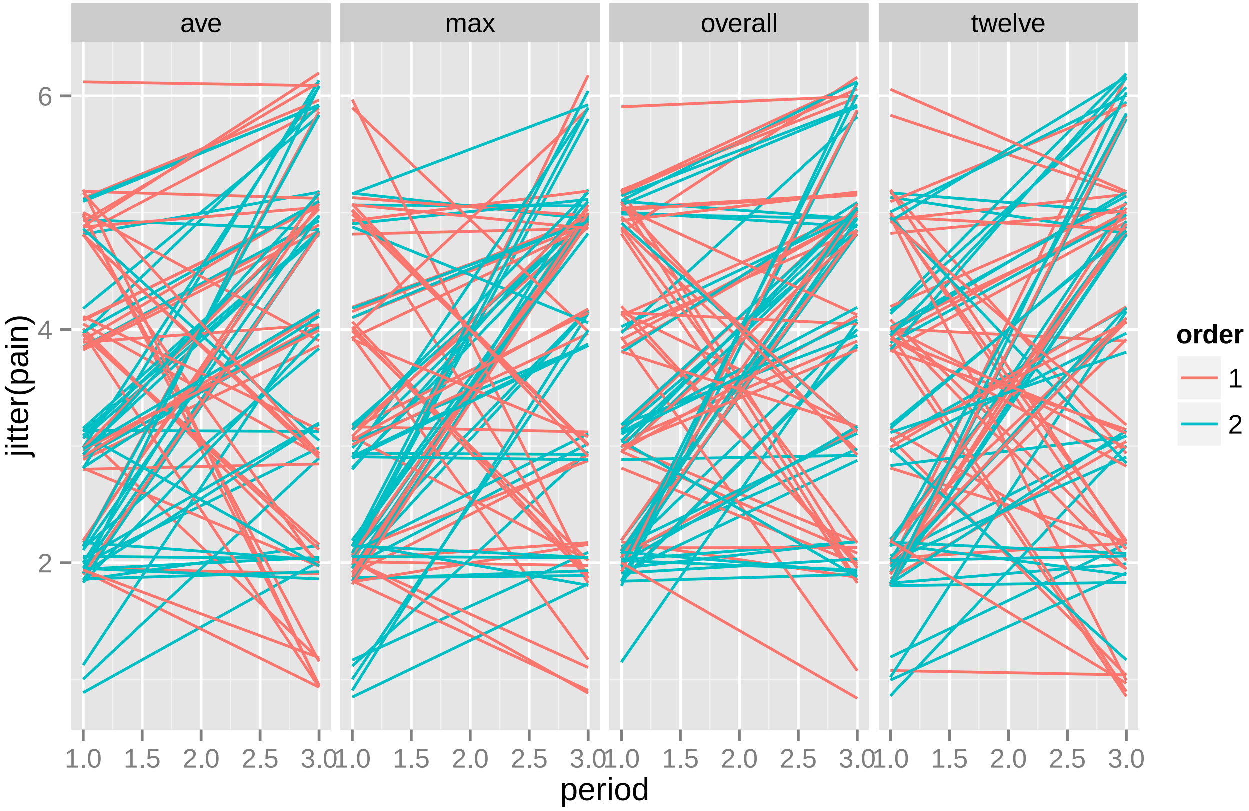
More plots

```
qplot(x=Treatment, y=pain, colour=sex, facets=order~ytype, data=tennis.nice, geom='boxplot')
```



More plots

```
tennis.nice$period <- as.numeric(tennis.nice$period)
ggplot(data=filter(tennis.nice,  period!=2),
       aes(x=period, y=jitter(pain), group=id, colour=order))+geom_line() +facet_grid(.~ytype)
```



Compare means for the different outcomes of interest

Use t-test and wilcoxon test (nonparametric alternative to t-test)

With R: `t.test` and `wilcox.test`

ANOVA

At the end, try to fit a full anova model for one of the outcome (`max`). Add all variables that make sense (e.g. also gender, even if not done in the exercise).

Carry-over effect

Check if there is any carry-over effect. How to do that?

If carry-over effect, the washout period is not long enough. Two way to test it:

- The treatment effect would be different depending on the order you take it.
- Or: there would be some differences between the two group at the end of the washout period.