

# Anova exercise class

*Sylvain*

*15th of December 2014*

## Organization

- Question hour: Th. 15.01.2015 from 2 to 3pm in HG G26.3
- Exam date (no guarantee): Sat. 31.01.2015 from 9 to 11am (Höngg)
- Exam review: We. 25.02.2015 from 12 to 1pm
- Remark: please fill in the [TA evaluation form](#) and hand it in at the end!

## Series 5

As usual, not corrected by myself... just ask me if you have questions!

## Series 6

### Exercise 1: Blocking

#### Blocking

- Goal: deal with *nuisance factor* (reduce extra variability and avoid confounding)
- Origin: agriculture
- Examples: batches, subjects, hospital, etc.

#### Blocking or not blocking?

- Key element:  $SS_{tot} = SS_{treat} + SS_{block} + SS_{res}$
- What happens if you don't block?
  - remember  $F_{stat} = MS_{treat}/MS_{res} \dots$  What happens if  $H_0$  is true? think about it.

#### Blocking: remarks

- What if you know, but can't *control* a nuisance factor (ex: people choose their hospital, you forgot to control, etc.)?
- Analysis of *Covariance*. Why is this different? What is preferable and why?
- What about nuisance factor that you don't know and can't control for? *randomization*
- How to interpret p-values of block factor?
- Model with fixed vs. random effects.

## RCBD and BIBD

Ideally, every treatment is tested in every block. The experimental units are assigned at random to each block, and each treatment is tested in each block:

- RCBD: Randomized complete block design

Sometimes it is impossible to have each treatment in each block (e.g. wine tasting with 100 wines and 20 people)

How would you solve that?

1. Ignore blocking and assign at random
2. BIBD: Balanced incomplete block design

## BIBD

- Typically, you are given the number of treatments  $n$  to test and the size of a block  $k$ . These are constraints of the experiment.
- For example:  $k$  is 4 wheels on a car, maximum number of wines to taste, etc.
- Rule: any two treatments occur together the same number of times:  $\lambda$ .
- Goal: find such a design, while minimizing the number of blocks needed. . .
- Solution: play with the equations on p.86 of the lecture notes

## Exercise 2: split-plot design

### Cirrhosis treatment

- Treatment: two types of surgery.
- Each patient is given one of two surgery.  $Y$  is measured before and after.



## Load the data

```
Sh <- read.table("http://stat.ethz.ch/Teaching/Datasets/Shunt.txt", header=TRUE)
Sh$Subject <- as.factor(Sh$Subject)
Sh$Treatment <- as.factor(Sh$Treatment)
Sh$Time <- as.factor(Sh$Time)
head(Sh, 10)
```

```
##      Subject Treatment Time  Y
## 1         1 Selective  Pre 51
## 2         1 Selective Post 48
## 3         2 Selective  Pre 35
## 4         2 Selective Post 55
## 5         3 Selective  Pre 66
## 6         3 Selective Post 60
## 7         4 Selective  Pre 40
## 8         4 Selective Post 35
## 9         5 Selective  Pre 39
## 10        5 Selective Post 36
```

## What is the design?

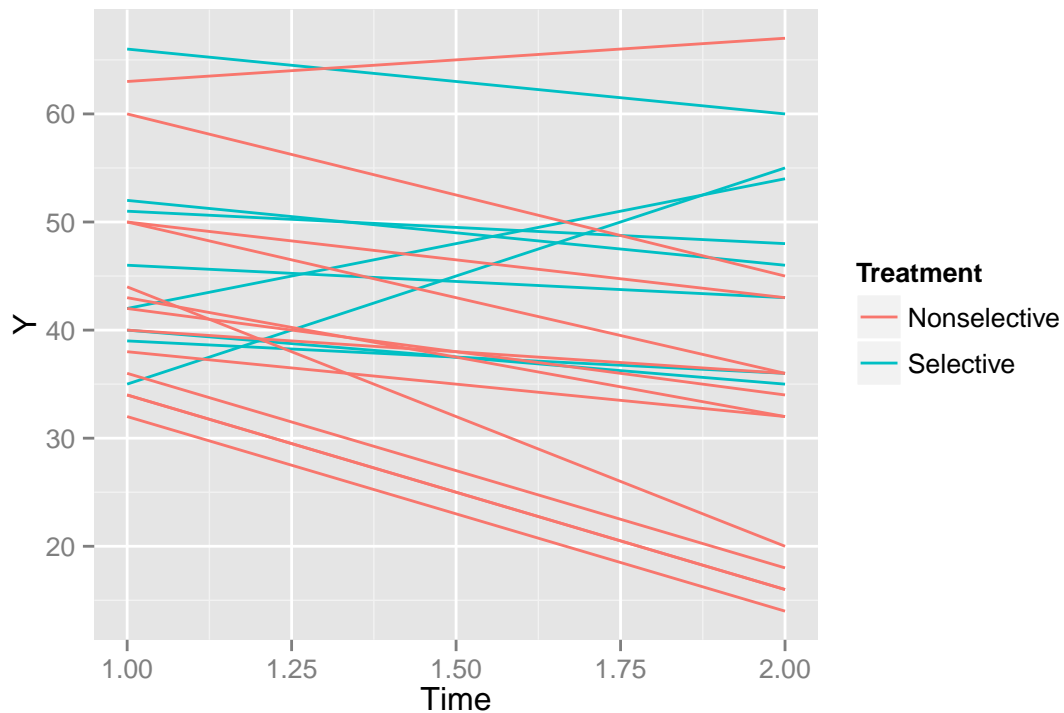
- How to recognize a split-plot design?
- More info in this [pdf](#)
- Fix one factor (mainplot), vary a second factor (subplot)
- Example: mainplot=land with irrigation system, subplot=fertilizer type
- With R, use formula with nested factors:
- `aov(Y ~ main*nested + Error(grouping/nested))`
- or, with lme4: `lmer(Y~ main*nested + (1|grouping), data=Sh)`

## Split-plot

- What is wrong if you don't take into account the design?
- If ignore Subject: you ignore the fact that observations are not true replicate, but come from the same person => wrong inference
- If Subject as blocking factor: you cannot estimate your model (over-parametrized).
- That's why Subject appears as a random effect!

## Plot the data

Does the type of surgery have an effect?



### **Exercise 3: Pizza**

**What type of design is it?**

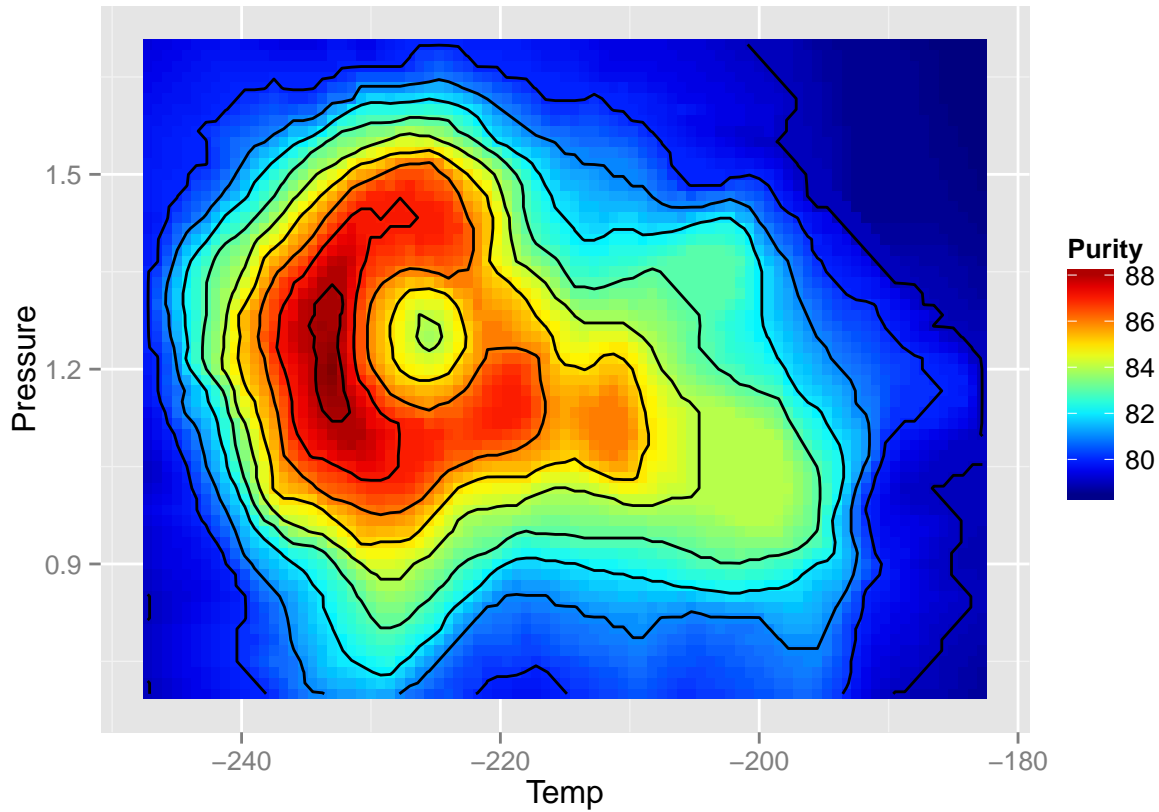
3 types of pizzas in 6 different packaging. What kind of experiment could you make?



### **Exercise 4: oxygen**

**Optimization of a process**

A chemical plant produces oxygen with some process. We want to find the optimal pressure and temperature.

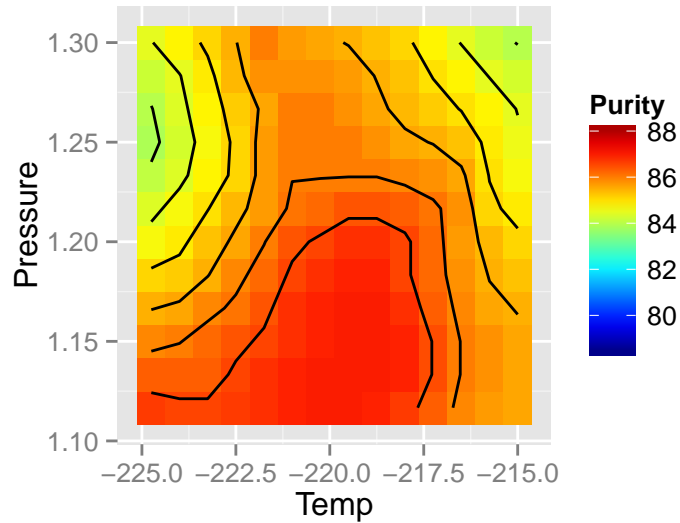


### Response Surface Method

- Like a factorial design, but we want to apply it sequentially to find the maximum.
- Steepest ascent model (First-order):  $Y = \mu + A_{temp} + B_{pressure} + \epsilon$
- From your point, find the direction in which the response changes the most. Move in this direction and repeat.

### Response Surface Method

We zoom on a zone and approximate it with a plane. Then move in the direction of steepest ascent, etc.



## Data and fit

Your browser does not support the HTML5 canvas element.

You must enable Javascript to view this page properly.

## Remarks

### Orthogonal vs. non-orthogonal designs

- Often we assume that data are balanced. But in reality often you have missing values, or incomplete design, etc.
- Such design are said to be non-orthogonal. What does it mean? Why does it matter?
- $X_1$  and  $X_2$  are orthogonal if:  $\sum_i X_{1i}X_{2i} = 0$ .

$X_1$	$X_2$	Y
1	1	3
1	-1	4
-1	1	2
-1	-1	3

### Non-orthogonal

- When the design is non-orthogonal, the effects cannot be estimated independently...
- Think about linear regression:
- If all X are independent, multiple regression is equivalent to  $p$  univariate regression.
- If not, then things change...

## Non-orthogonal design and SS

- $SS$  type I and III are the same with orthogonal design
- with non-orthogonal design not anymore.
- In particular, type I  $SS$  depends on the order:  $\text{aov}(y \sim A+B) \neq \text{aov}(y \sim B+A)$
- type III  $SS$  doesn't depend on the order, but the usual  $SS$  decomposition is lost... with R: `drop1(fit)`.
- See this [link](#) for a detailed example

## Some common p-values traps

Which of the following is True/False?

- If  $P=.05$ , the null hypothesis has only a 5% chance of being true
- A nonsignificant difference (eg,  $P=.05$ ) means there is no difference between groups
- If  $P=0.05$ , even if there is no difference (the null is true), there is a 5% chance to observe a difference such as the one in the data.
- If an effect has a  $P$  of 0.00001 it is very important for the problem at hand

Reference: (A dirty dozen: twelve p-value misconceptions)[<http://www.ncbi.nlm.nih.gov/pubmed/18582619>]

## Exam

- open book
- everything is tested: ANOVA and experimental design
- Some basic calculations by hand
- Don't stress out: most subtasks are independent of each other. If you're stuck at one of them, continue to the next one. If you need a result that you didn't obtain, assume a fictive value and move on.
- Good luck!
- Question hour: Th. 15.01.2015 from 2 to 3pm in HG G26.3
- Remark: please remember the [TA evaluation form](#)



Merry Christmas and Happy New Year!!!

