

ASCA in one hour

Morten Arendt Rasmussen

Heads up!

We are recording this webinar



UNIVERSITY OF
COPENHAGEN

Monday webinars from ODIN
Chemometrics and Machine Learning



Kudos to



• Age Smilde



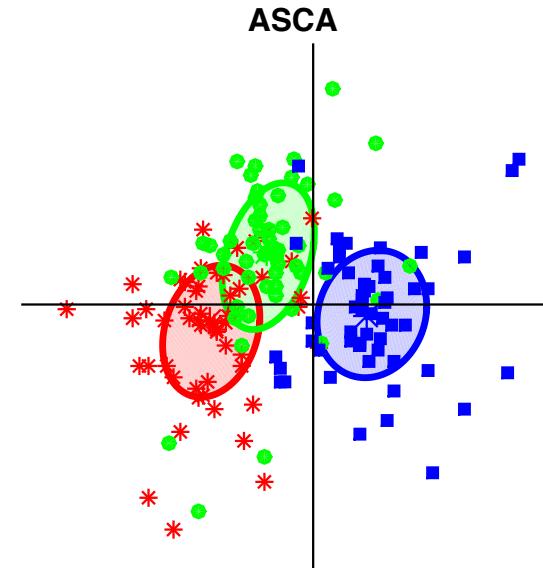
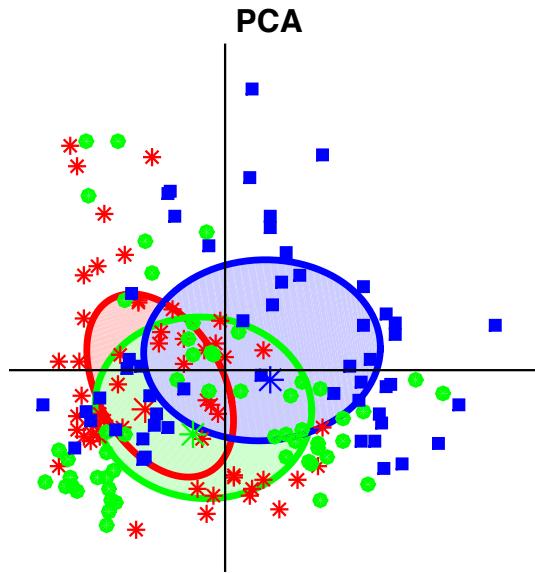
Jeroen Jansen



Take home

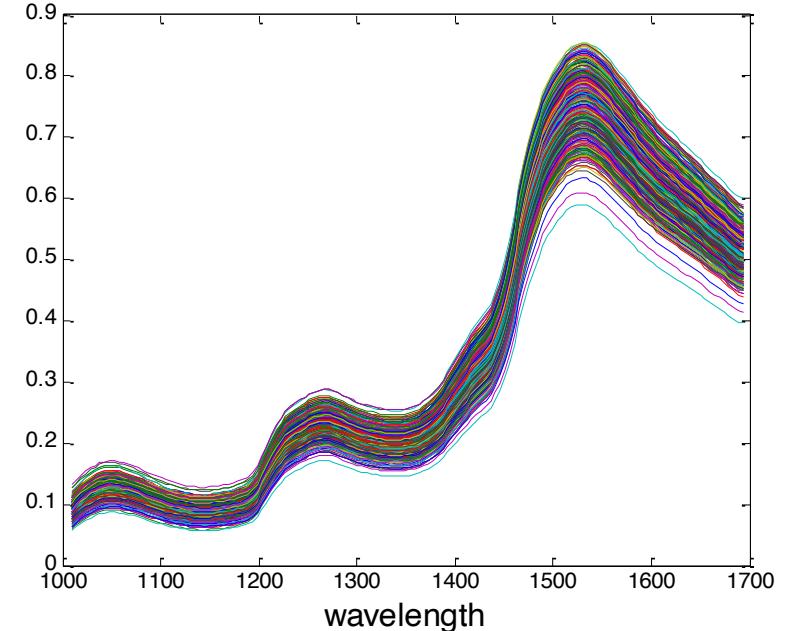
- ASCA combines the best of two worlds: **ANOVA** from statistics and **PCA** from chemometrics
- The **dominating variance** in data is often **trivial**, while our **questions of interest** are way more **subtle**.
- ASCA can *dig* such information out

A Motivating Example



Design of Experiments

- Example – Bread
- Three factors:
 - - **Type** (different enzymes): A, B, C
 - - **Days** after production: 1,..,6
 - - **Position**: Top, middle, bottom
- ***Full factorial with 6 replicates***
- **$N = 324$**

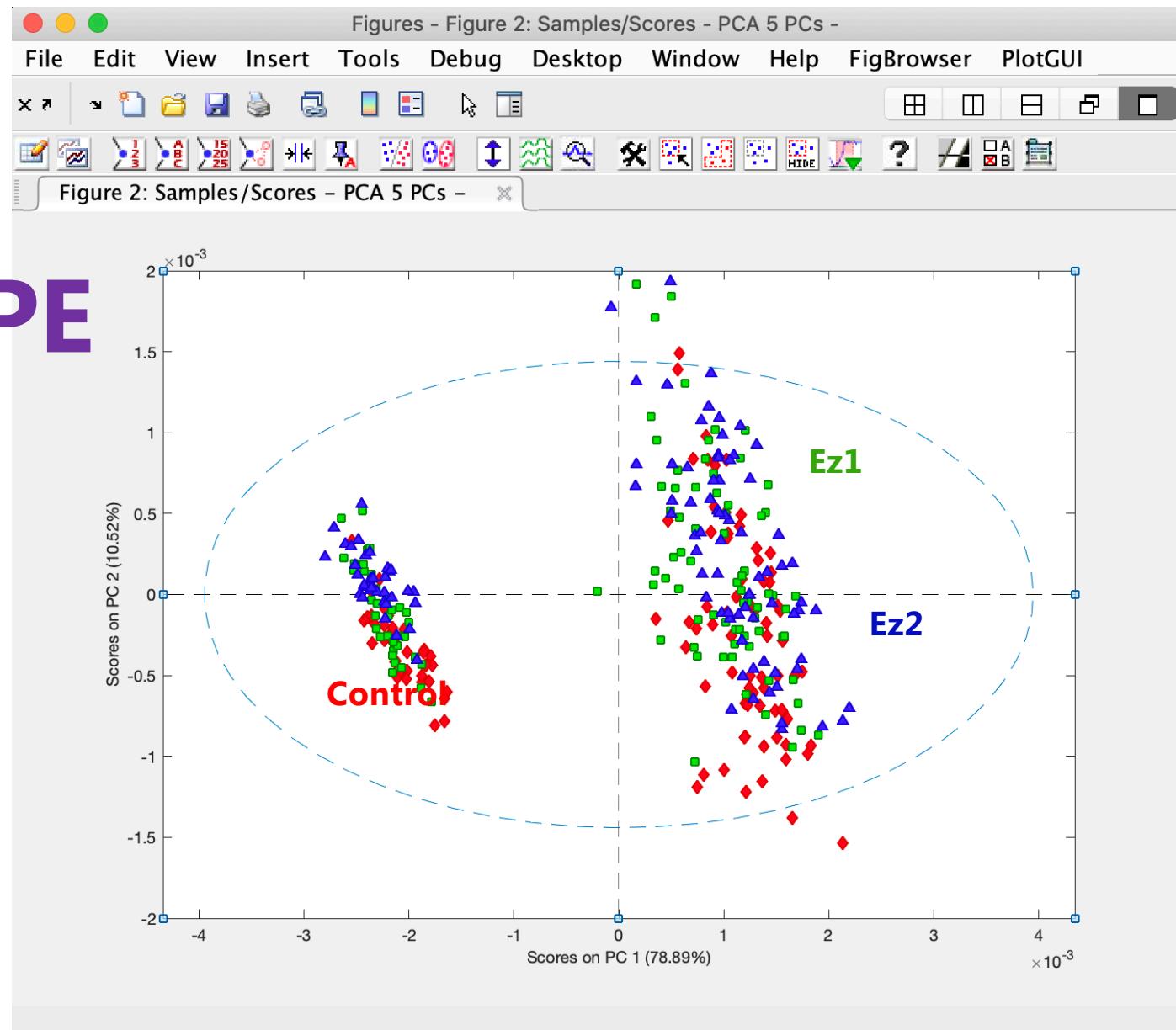


Data can be found here:
<https://github.com/mortenarendt/ASCA>

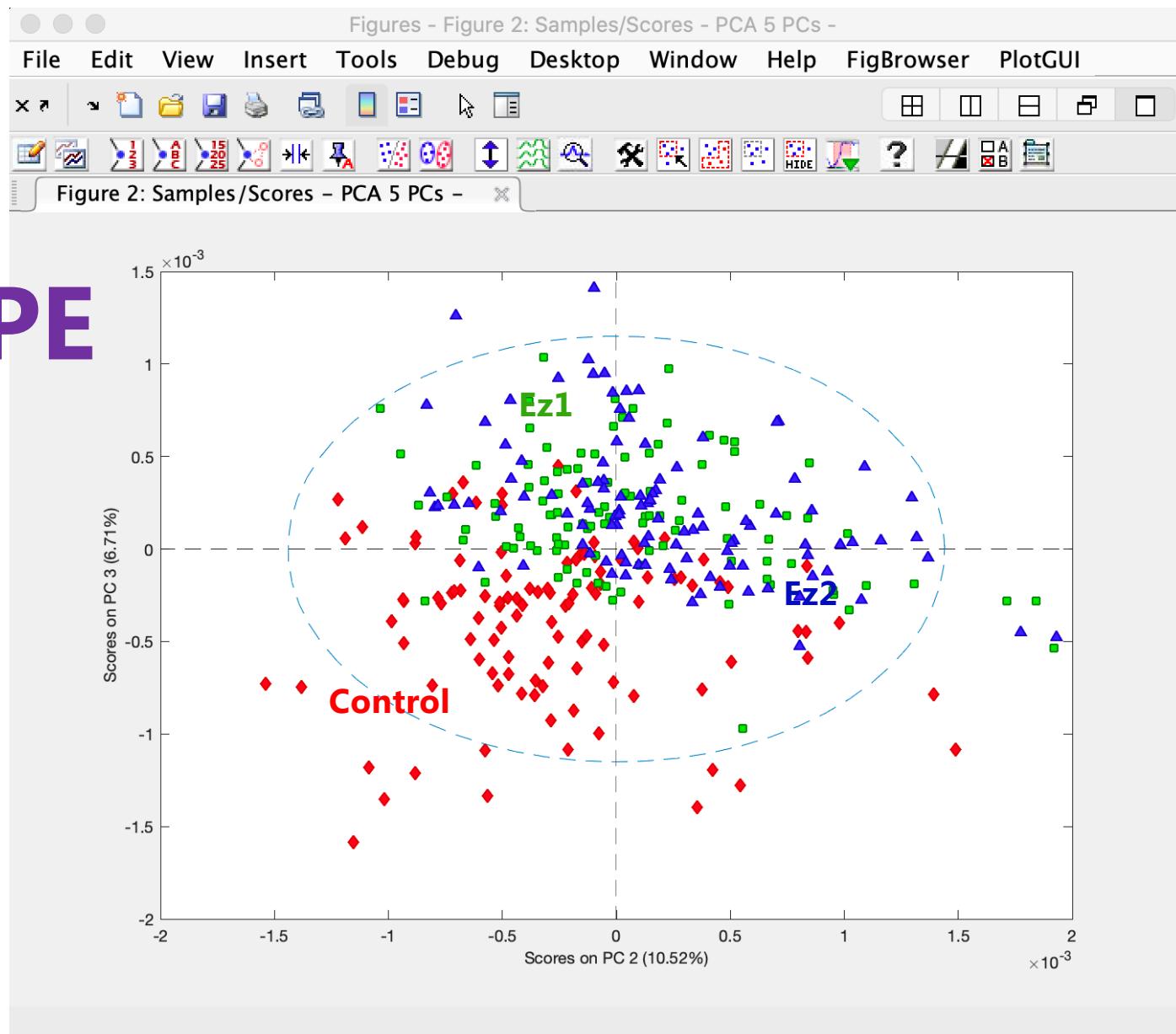
Staling of Bread



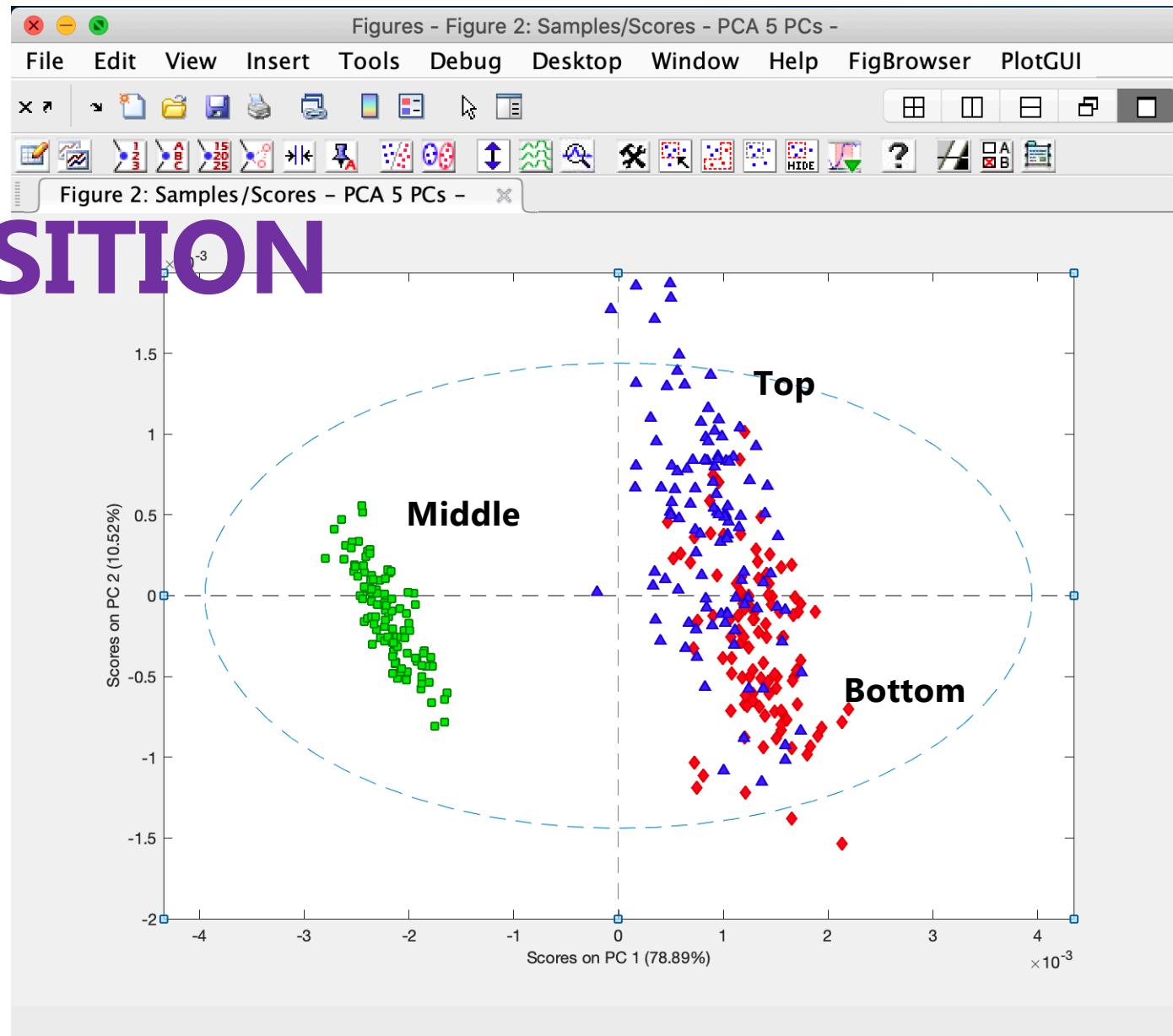
TYPE



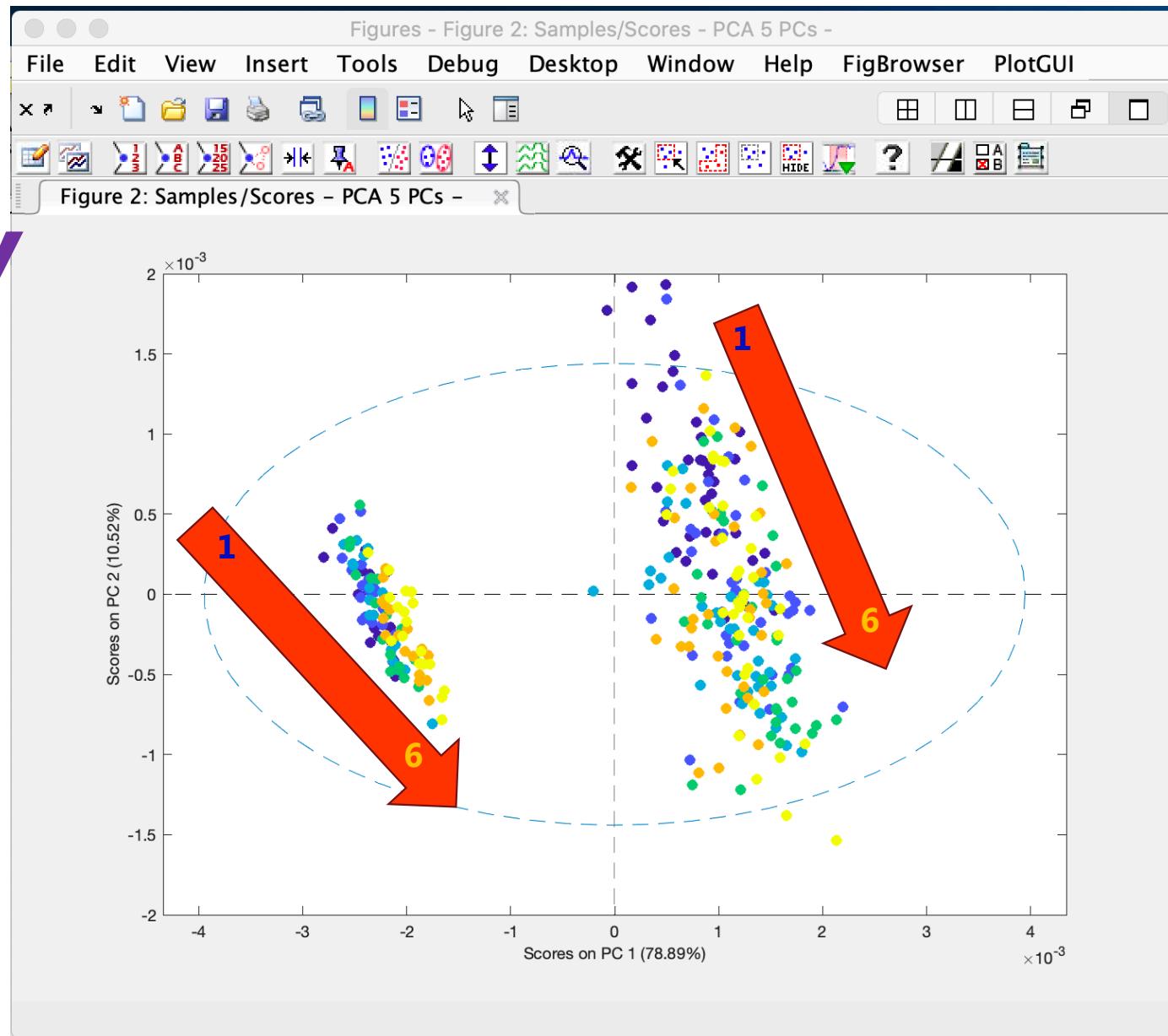
TYPE



POSITION



DAY



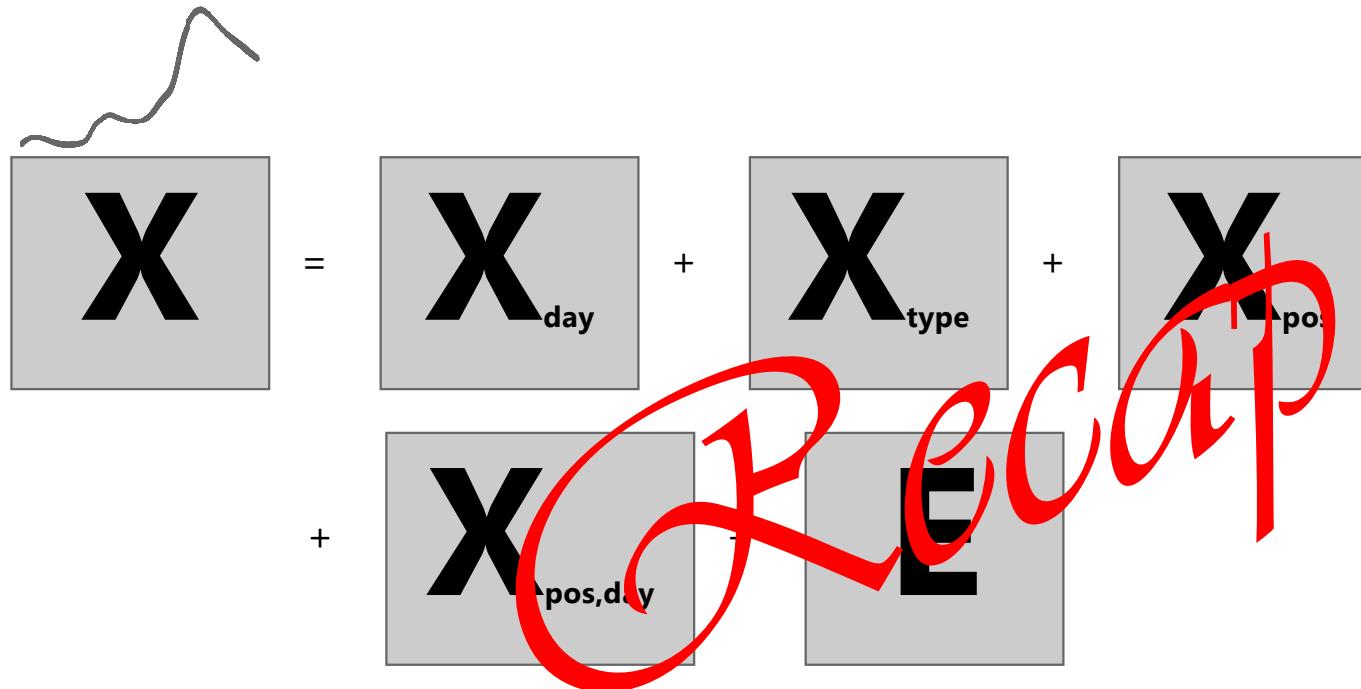
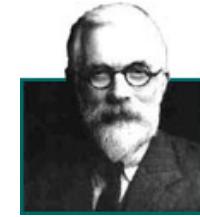
Where is the information?

Three factors:

- **Type** PC2 & PC3
- **Days** PC1 & PC2
- **Position** PC1 & PC2

- **PCA is powerful in extracting information**
- ***But...***
- **In a normal PCA we do not at all utilize that there were an experimental design**

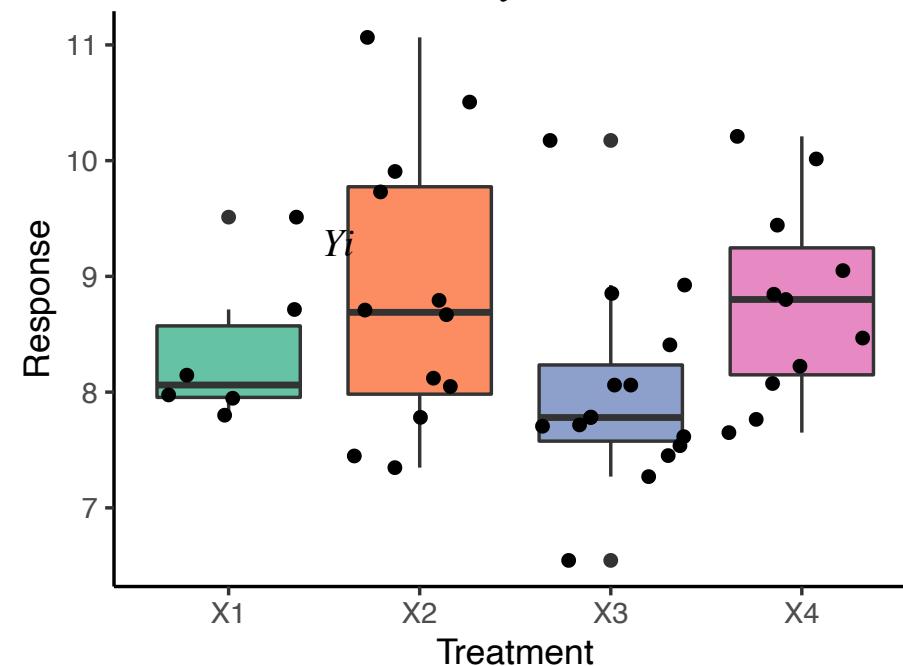
ANOVA



ANOVA *(oneway)*

$$Y_i = \mu + \alpha_{Treatment_i} + e_i$$

e_1, \dots, e_n
Comes from the same
distribution



Monday webinars from ODIN
Chemometrics and Machine Learning

ANOVA

$$Y_i = \mu$$

$$+ \alpha_{Treatment_i}$$

$$+ \beta_{Time_i}$$

$$+ \gamma_{Treatment_i, Time_i}$$

$$+ e_i$$

Treatment	Time	Y
1	1	12,2
1	2	14,5
1	3	22,4
1	1	9,2
1	2	13,0
1	3	27,1
1	1	6,9
1	2	19,2
1	3	21,5
2	1	12,5
2	2	16,6
2	3	24,7
2	1	10,7
2	2	24,0
2	3	28,9
2	1	10,9
2	2	18,8
2	3	24,6

Monday webinars from ODIN
Chemometrics and Machine Learning

ANOVA – *some features*

- Particular useful for analysis of **experimental data**.
- Analysis of comparative experiments—those in which only the **difference** in outcomes is of interest.
- The calculations of ANOVA can be characterized as computing a number of means and variances.
- ANOVA can be performed for a single factor or multiple factors.

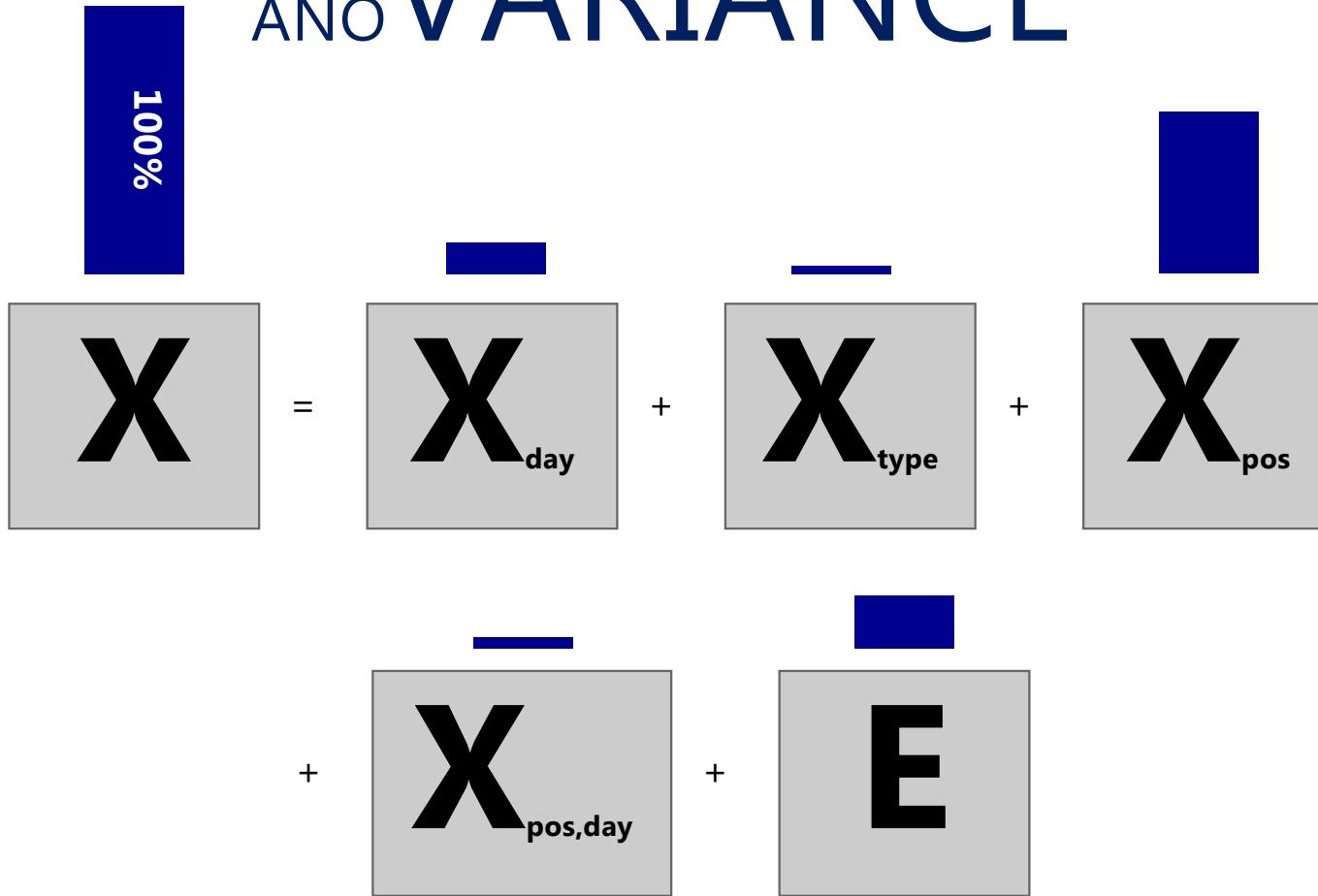
- **Assumptions**

- Response variables are normally distributed (or approximately normally distributed).
- Samples are independent.
- Variances of populations (different groups in the data) are equal.
- This is in statistics known as the **iid assumption**

iid = independent and identically distributed



ANO VARIANCE

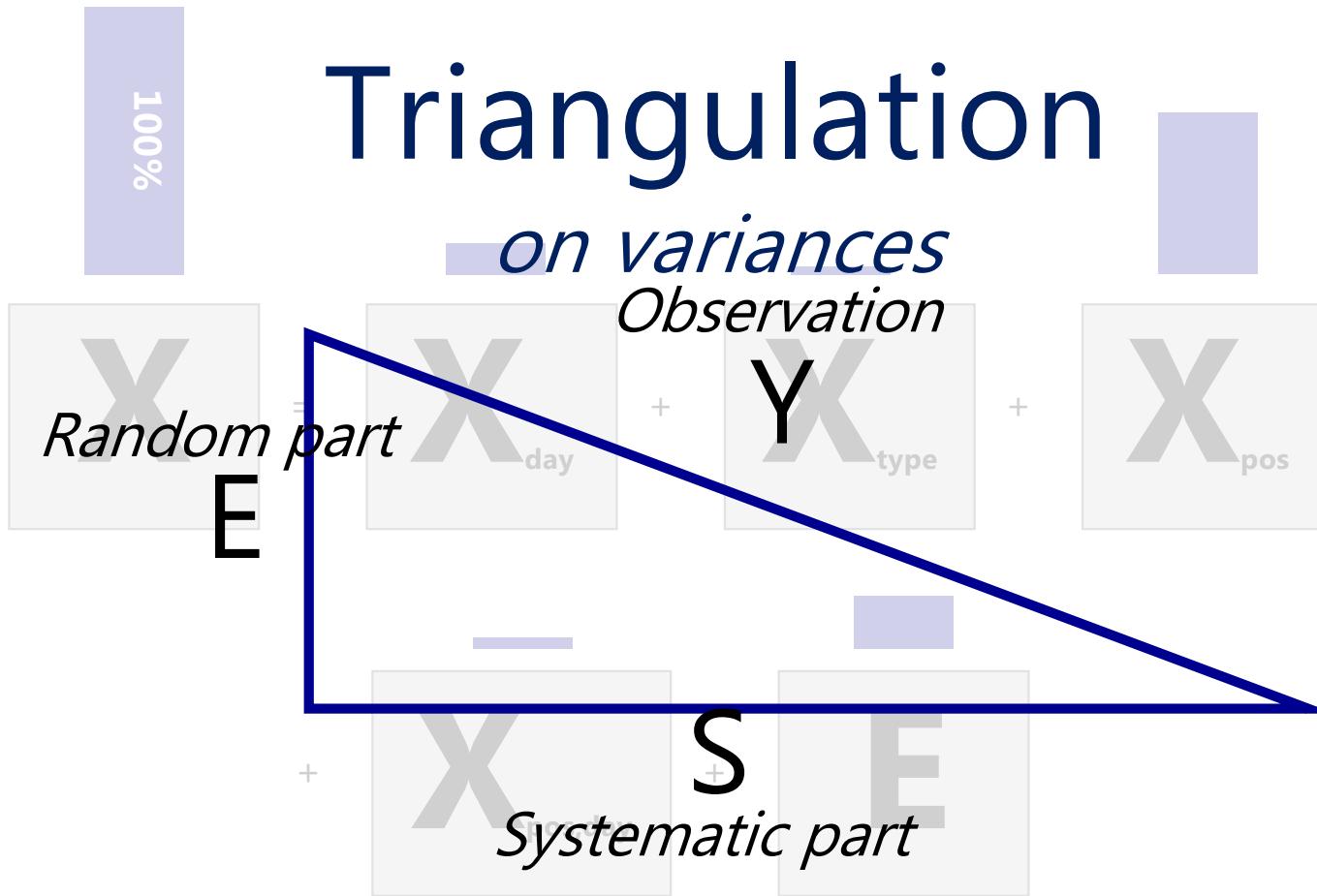


Monday webinars from ODIN
Chemometrics and Machine Learning

Triangulation

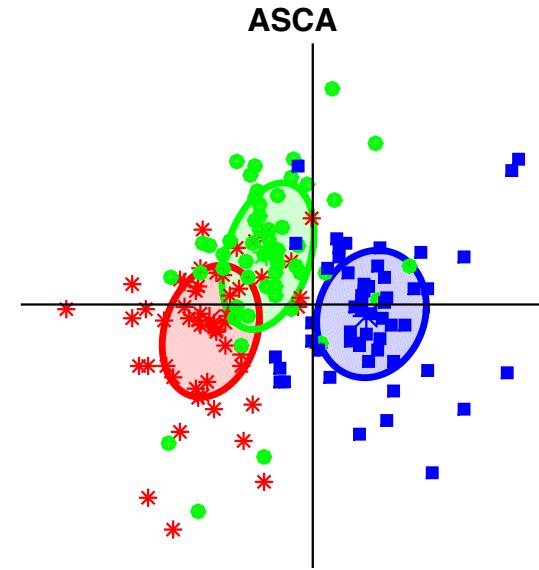
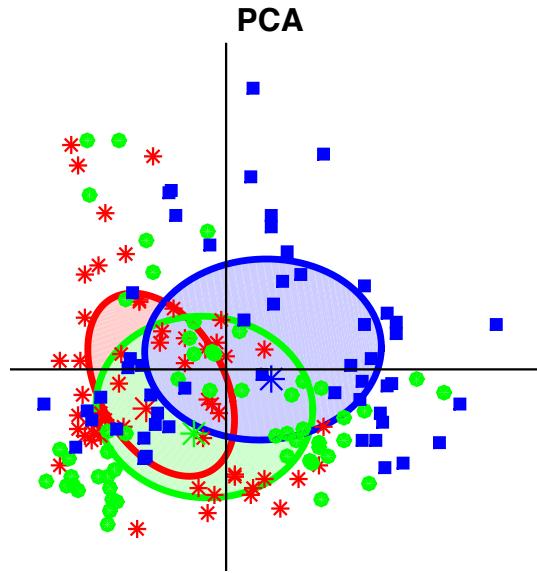
on variances

Observation



Monday webinars from ODIN
Chemometrics and Machine Learning

Example in PLStb



ASCA

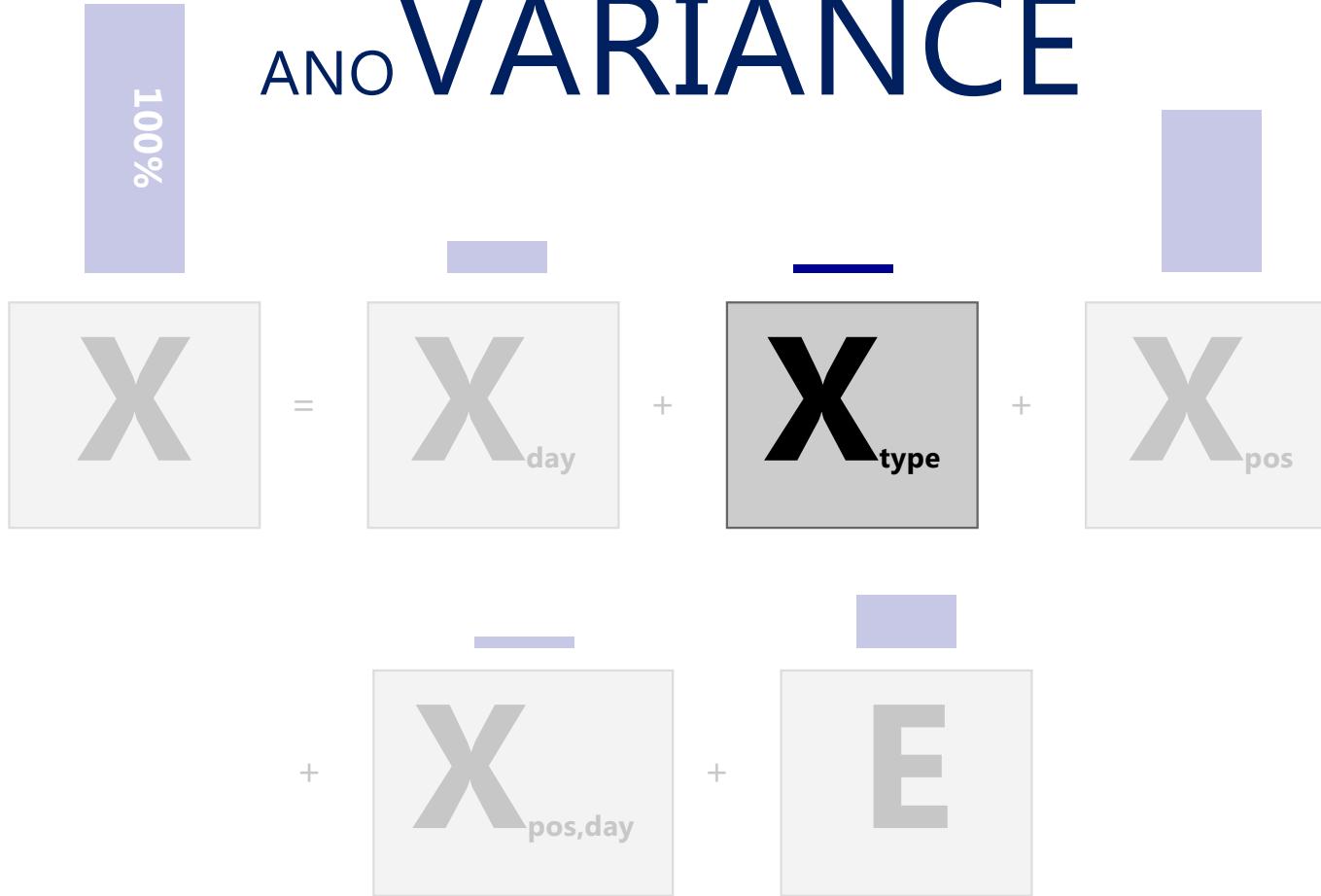
A two step procedure

1. **Split** the data into design related matrices and a random term
 - This procedure is basically univariate ANOVA for each variable.
 - Testing of the individual design effects is a straight-forward extension of the normal ANOVA test.
2. Use **PCA** on the effect of interest.
 - A bilinear model is calculated directly on the systematic effect matrix
 - In order to get an idea about the spread (uncertainty) the random term is projected on this part.

Monday webinars from ODIN
Chemometrics and Machine Learning



ANOVARIENCE



Monday webinars from ODIN
Chemometrics and Machine Learning

Simultaneous Component Analysis

"A PCA model on the part of interest"

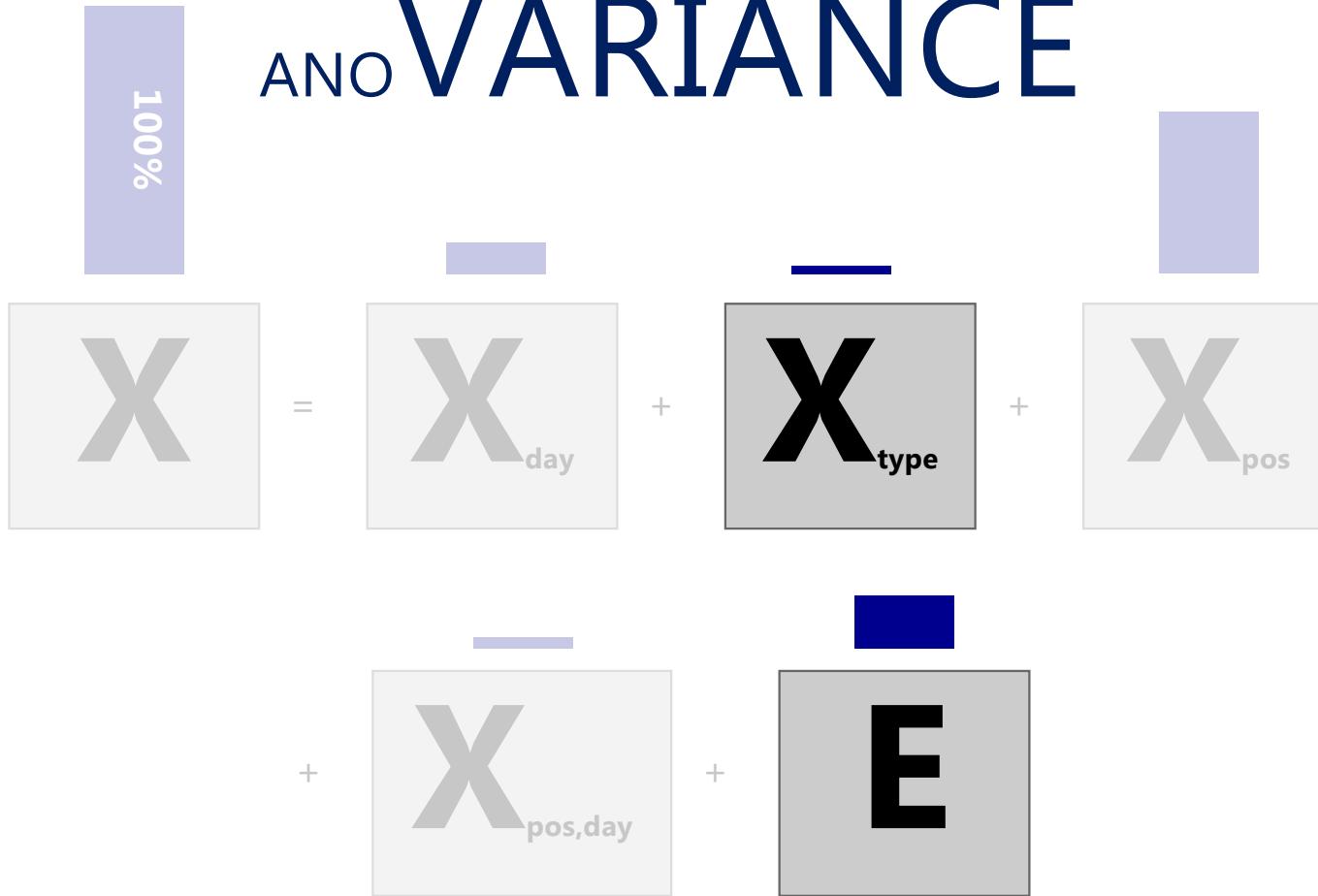
$$\mathbf{X}_{\text{type}} = \mathbf{T} \mathbf{P}'$$

But!... What about the error?

This has rank =
number of classes - 1

Smilde, Age K., et al. "ANOVA-simultaneous component analysis (**ASCA**): a new tool for analyzing designed metabolomics data." *Bioinformatics* 21.13 (2005): 3043-3048.

ANOVARIANCE



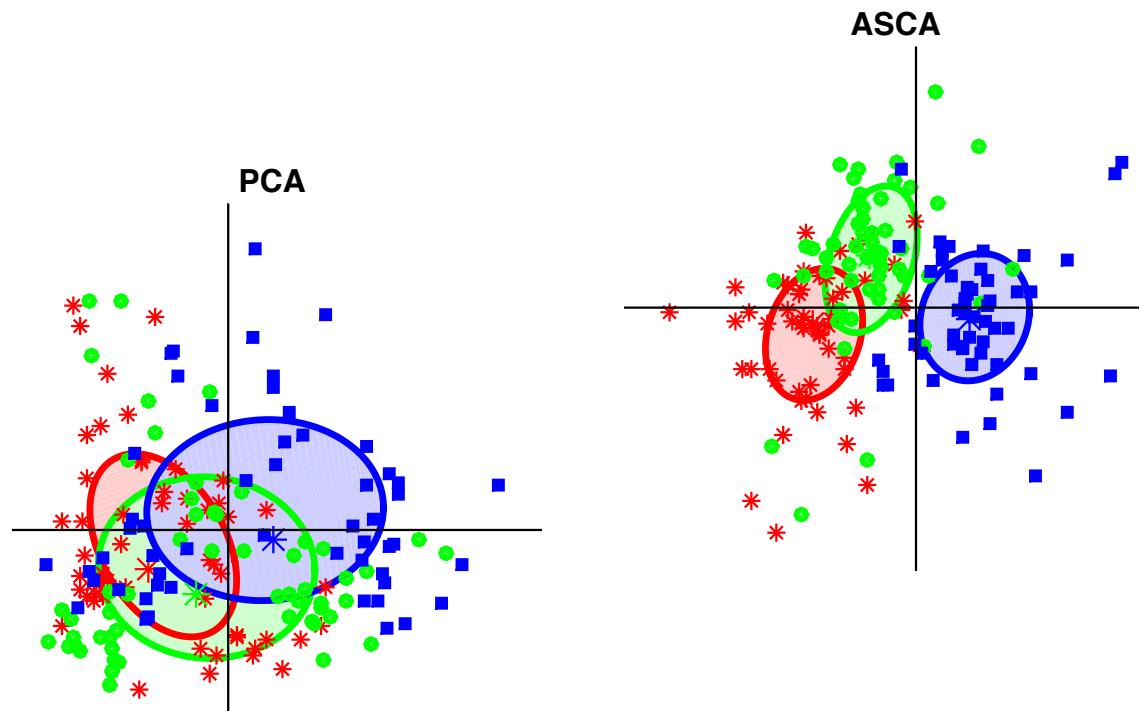
Monday webinars from ODIN
Chemometrics and Machine Learning

Simultaneous Component Analysis Score plots

$$\mathbf{T} = * \left(\begin{array}{c} \mathbf{X}_{\text{type}} \\ + \\ \mathbf{E} \end{array} \right) \mathbf{P}$$

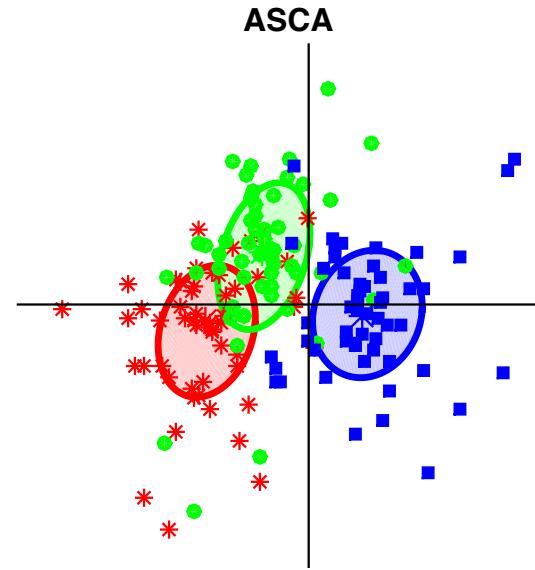
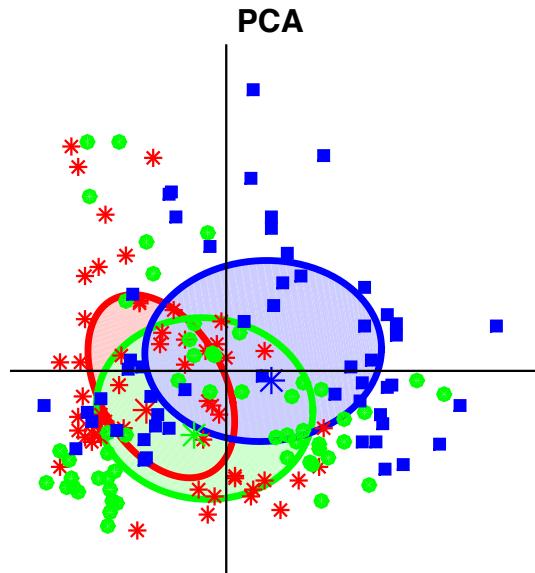
T, **T*** and **P** are used for interpretation

Example continued



Monday webinars from ODIN
Chemometrics and Machine Learning

Hypothesis Testing



Null Hypothesis Testing

The general idea

- **State a question**

- ...in relation to the design.

- **Central statistics**

- Figure out which number that encapsulate the differences observed in the data (with respect to the design variable of interest)

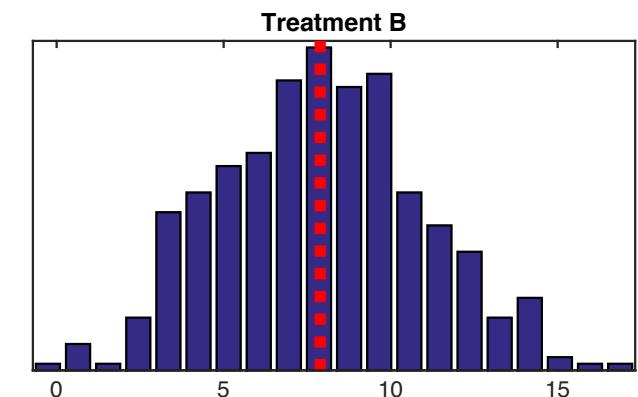
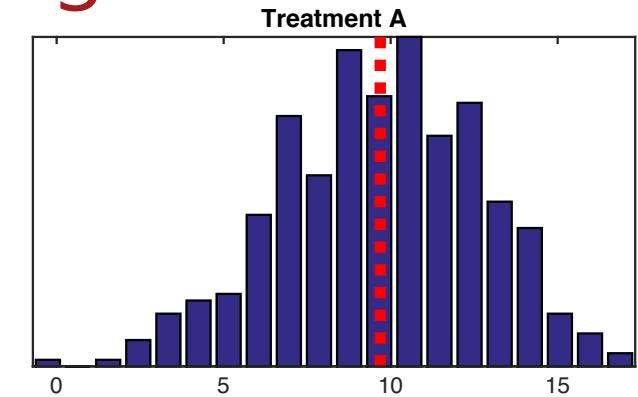
- *e.g. The difference between means*

- **Test**

- The central statistics measures the ***distance*** to the null hypothesis.

- Find out whether this number is a big or small.

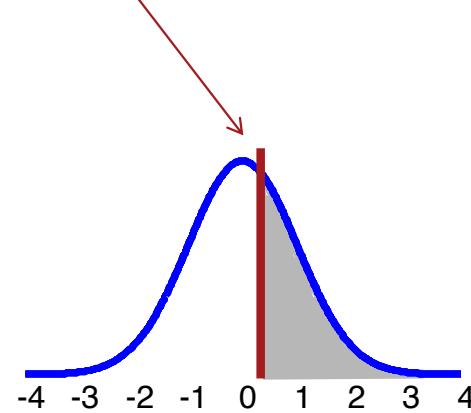
- We can use analytical *null-distributions* – if this one is known! Or we can create it from data.



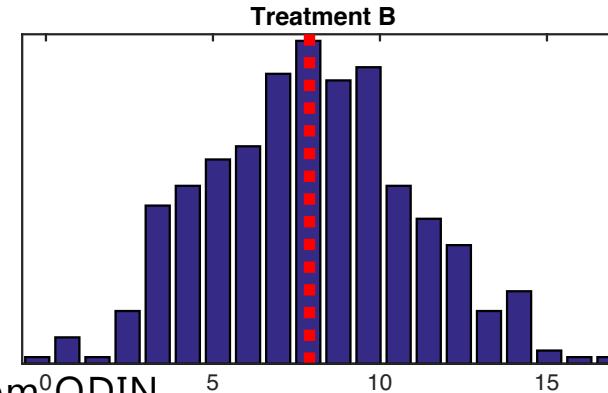
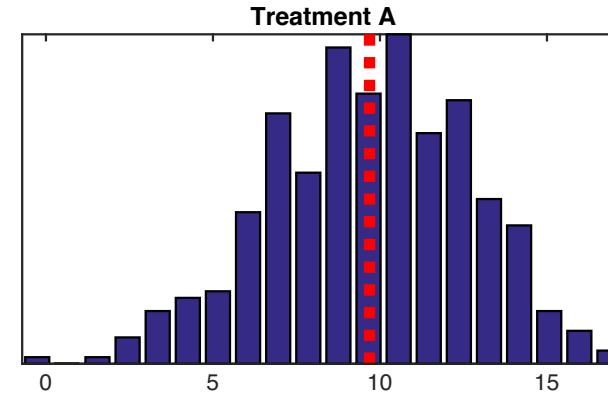
Null Hypothesis Testing

T-test

$$t = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$



t_{df}



Testing individual effects

Big model

$$\mathbf{X} = \mathbf{X}_{\text{day}} + \mathbf{X}_{\text{type}} + \mathbf{X}_{\text{pos}} + \mathbf{X}_{\text{pos,day}} + \mathbf{E}$$

Small model

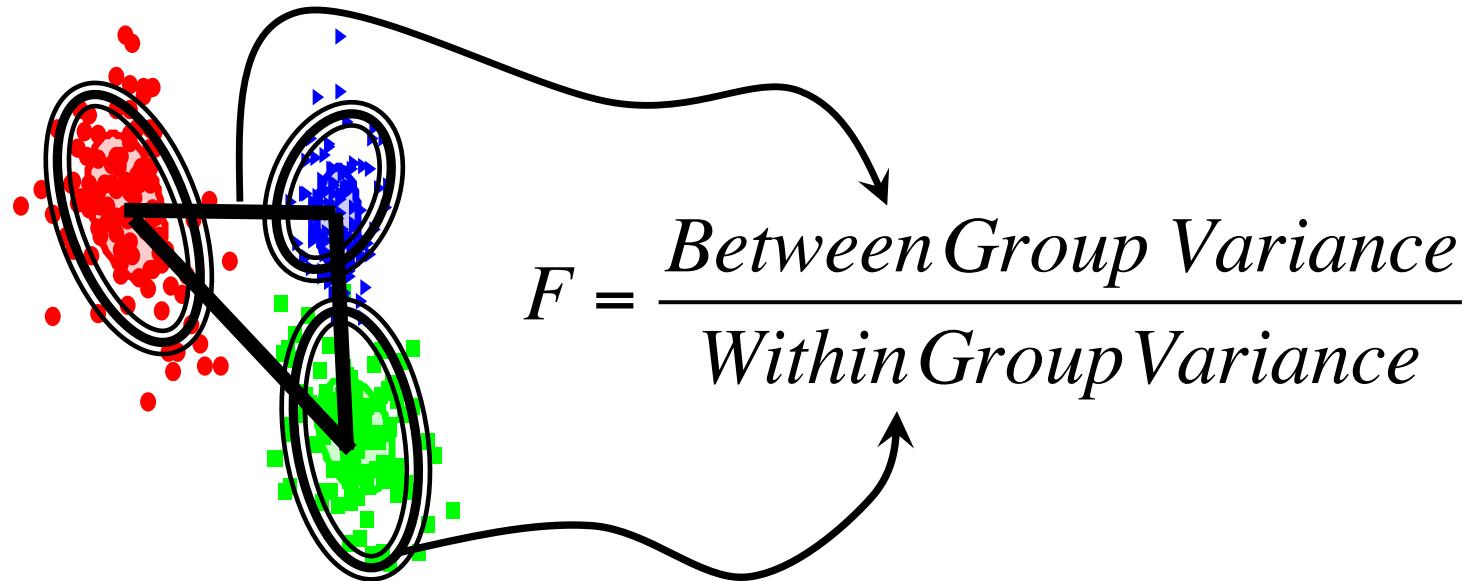
$$\mathbf{X} = \mathbf{X}_{\text{day}} + \mathbf{X}_{\text{pos}} + \mathbf{X}_{\text{pos,day}} + \mathbf{E}$$

Central Statistics

$$F = \frac{(SSE_{small\ m} - SSE_{big\ m}) / (dfe_{small\ m} - dfe_{big\ m})}{SSE_{big\ m} / dfe_{big\ m}}$$



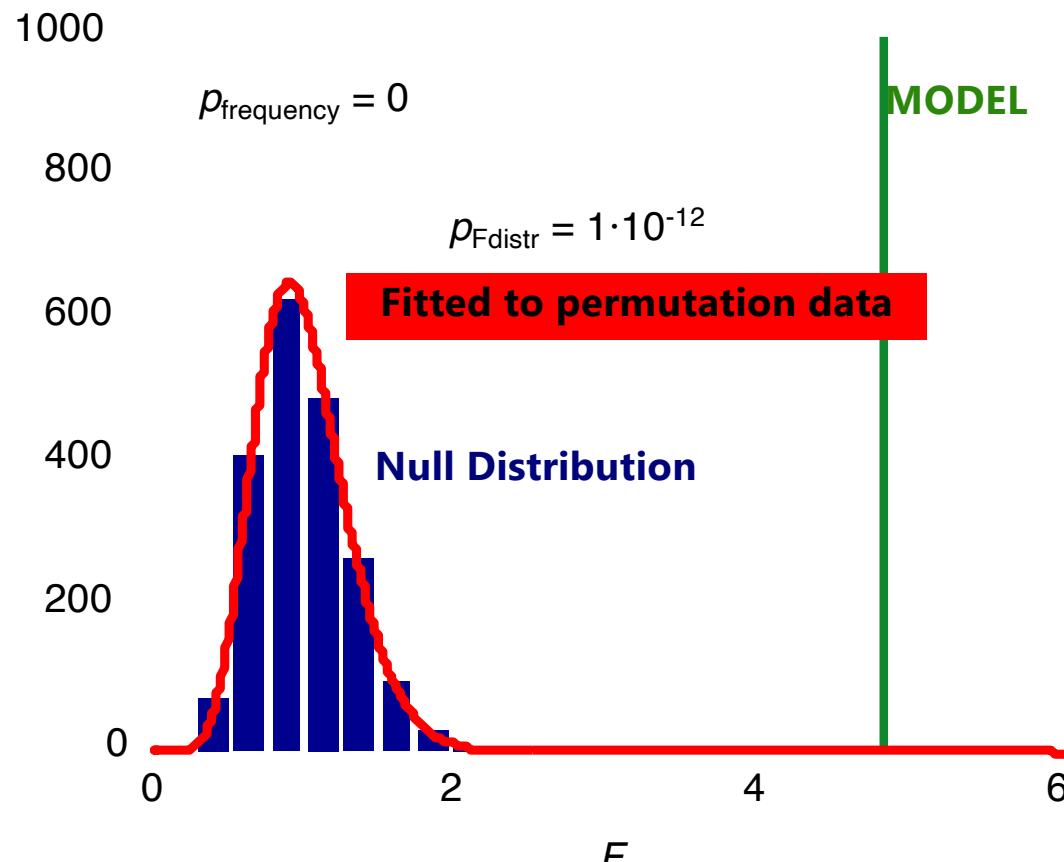
Testing individual effects



Central Statistics

$$F = \frac{(SSE_{small\ m} - SSE_{big\ m}) / (dfe_{small\ m} - dfe_{big\ m})}{SSE_{big\ m} / dfe_{big\ m}}$$

Permutation testing



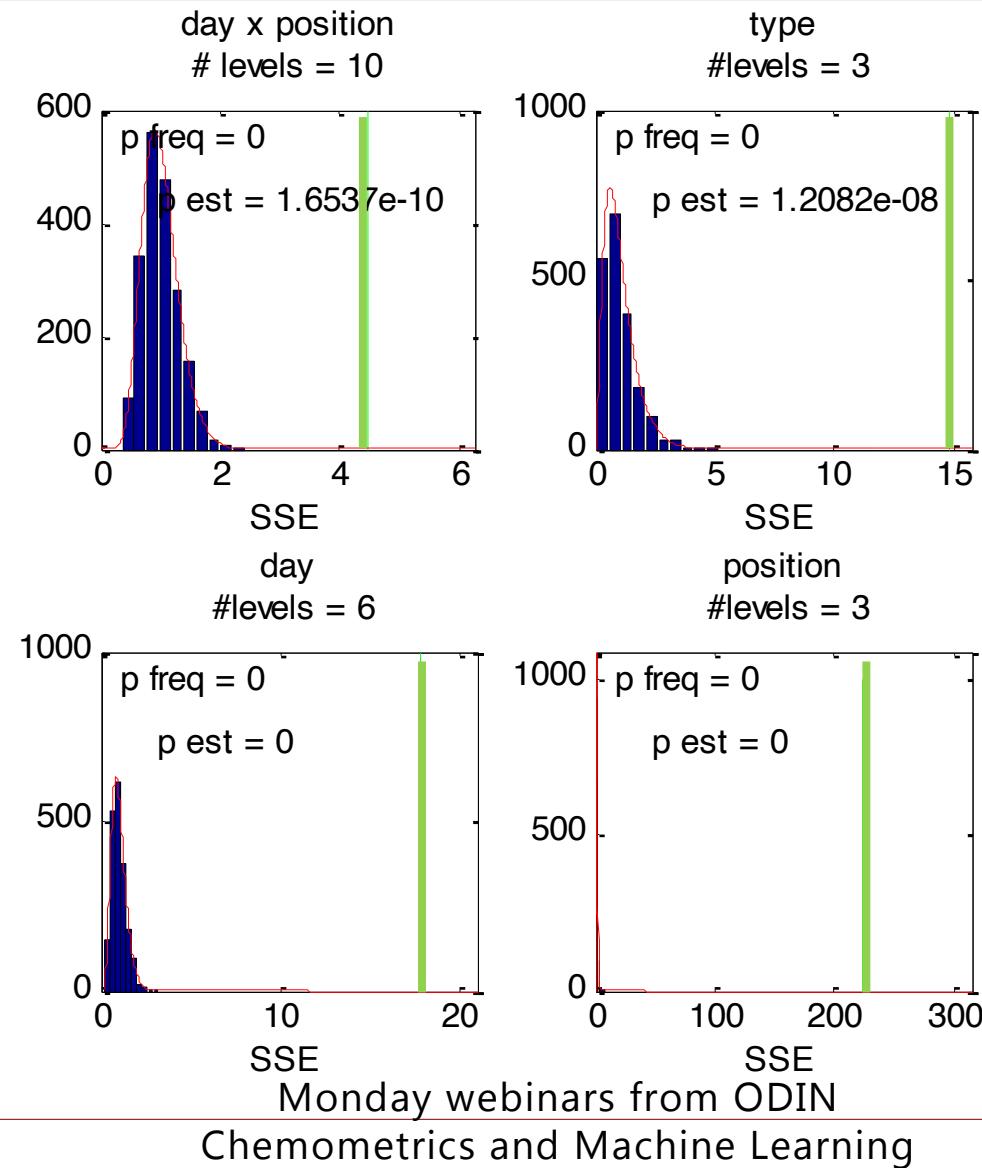
Monday webinars from ODIN
Chemometrics and Machine Learning



UNIVERSITY OF
COPENHAGEN



data • information • control



Summing up

ASCA is

- Powerful in disentangling subtle variation from multivariate data
- Basically ANOVA extended to multivariate data.
 - That is:
 - Splitting of each variable into design related variance and error
 - Effects are tested by comparing the individual design variance terms with the error variance. This is done by permutation testing in lack of an analytical distribution.
- Design related variance is interpreted by a *focused* PCA model, with uncertainty imposed by the residuals.

Do you want to know more?
Do you have projects with tricky data?

Contact us:

Frans van den Berg

fb@food.ku.dk



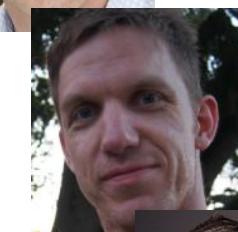
Rasmus Bro

rb@food.ku.dk



Åsmund Rinnan

aar@food.ku.dk



Morten Arendt
Rasmussen

mortenr@food.ku.dk





Thank you
for the attention

一十六景 神奈川沖
浪裏

葛飾北斎