TTK31 - Design of Experiments (DoE), metamodelling and Quality by Design (QbD)
Autumn 2021

Big Data Cybernetics Gang

General course information

- Lecturers:
  - Frank Westad
  - Øivind Riis
- Lecture time: Thursdays 12:15-14:00
- 10 lectures, see Blackboard
- 2-4 Exercises
- hands-on analysis

Reference group - VERY IMPORTANT

- at least 3 students
- will do 4 meetings (1 after the exam)
- shall represent the whole class $\Longrightarrow$ you will have meetings among yourselves too
- shall lead to a referansegrupperapport containing suggestions for improvements

Design of Experiments

Objectives with this course (specialization topic):

- Understand the principles of Design of Experiments (DoE)
- Understand the use of ANOVA in analysing results from DoE (and in general)
- Be able to decide on the right design given the problem at hand
- How to apply DoE for metamodelling
- How DoE falls into a framework of Quality by Design (QbD) and Process Analytical Technology (PAT)

Lecture overview

1. DoE: Introduction and motivation
2. ANalysis Of VAriance (ANOVA)
3. Factorial designs
4. Fractional factorial designs
5. Response surface designs
6. Optimal designs
7. Metamodelling
8. Combining DoE with multivariate analysis/machine learning
9. QbD – PAT
10. Practical examples of DoE related to cybernetics

Exercises/assignments and hands-on analysis

The Design-Expert®software from the company Stat-Ease is available for your use during this course
The classroom serial number is 8300-5935-5835-CLAS

`http://www.stat-ease.com/`

Follow these steps to register, download, and install the program to your personal computer:

- Create an account on the Stat-Ease website register this license to your account
- Download and install the software on your computer

Introduction

Design of Experiments

- We claim: Everybody working in quantitative sciences should know about the principles of DoE
- DoE is useful for:
  - discovering what are the important parameters in a system/process
  - identify if there are interaction effects and higher order relationships between input and output
  - metamodelling based on physical models
  - speeding up simulation systems
  - finding the best process settings given changes in raw material, equipment, sensors etc.

Why is not DoE widely used?

A survey showed:

The main barriers that hinder the widespread use of DoE are

- low managerial commitment
- engineers' general weakness in statistics.

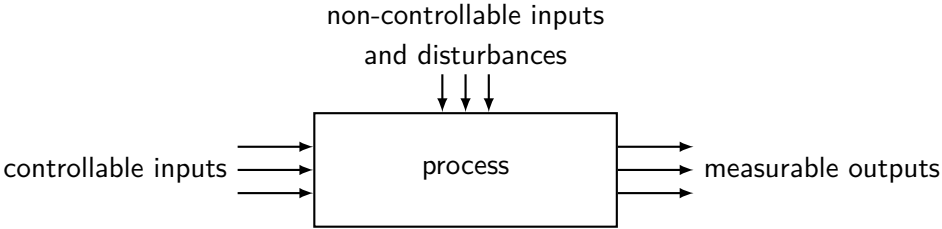
The overall 16 barriers were classified in three groups:

- business barriers
- **educational barriers**
- technical barriers

Although DoE is commonly found in statistics and quality literature, it is clearly underused in industry.

Standing assumption:

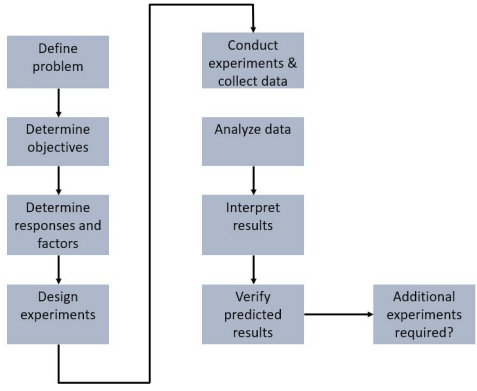
We want to model a system, and we can control some of its inputs / decide how to measure



What is Design of Experiments?

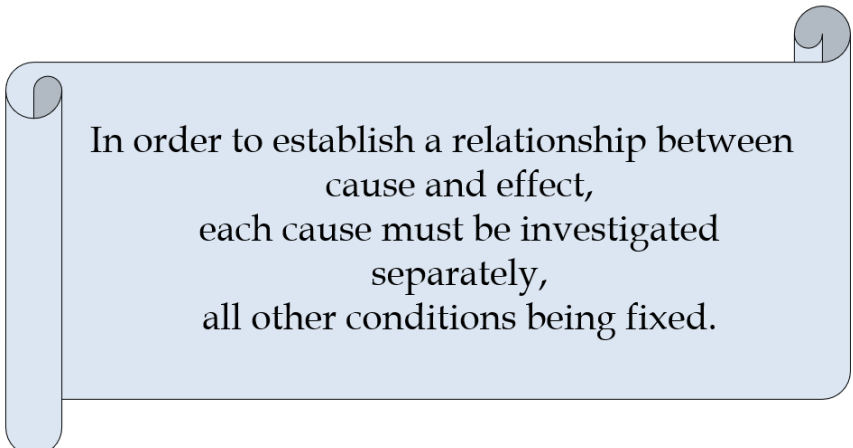
Design of Experiments (DoE) is the pre-planned, **systematic variation of controllable** experimental factors that **induce a response** in a system. The factors are measured in such a way that the **minimum effort** is required to gain a **maximum amount of information**

The experimental design cycle



If needed: Iterate by starting from top or at a certain step in the cycle

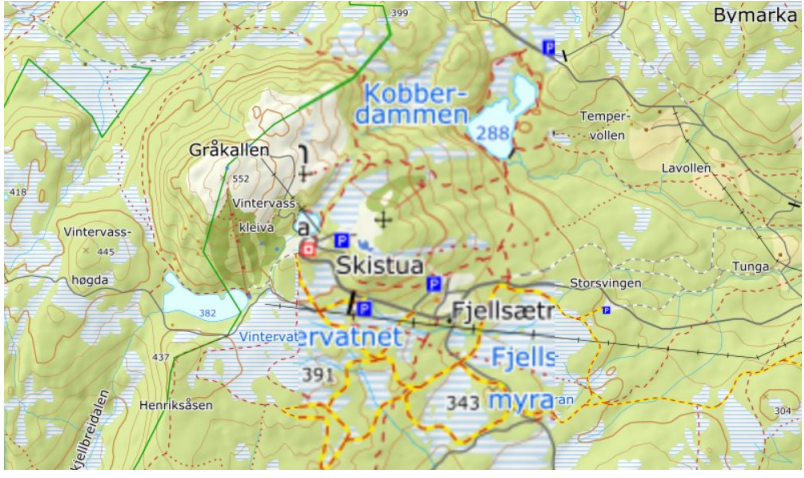
One variable at the time (OVAT)



An excursion to Gråkallen

Assume you want to hike to the top of Gråkallen

How would you reach the top?



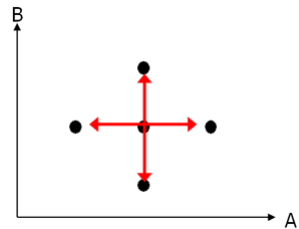
Experimental design versus OVAT

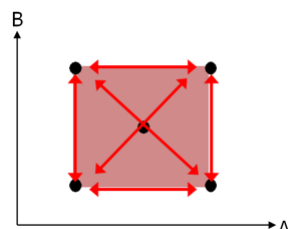
Traditional approach: One Variable At a Time

- One variable at a time (OVAT)
- Cannot detect interactions
- Inefficient (serial processing)

Experimental design

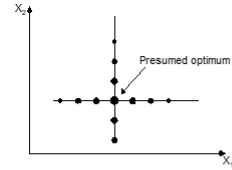
- All experiments are used to estimate effects of A and B
- Interactions can be estimated
- Precision can be estimated
- Maximizes information with minimum runs



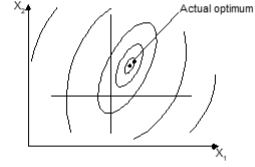


How to span the experimental space

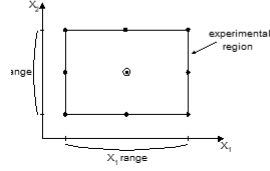
The Classical Approach:



(OVAT) What can go wrong?



How can we do it better?



The DoE process

- Identify opportunity and define objective
- State objective in terms of measurable responses
  1. Define the minimal change (Δy) that is important to detect for each response (signal)
  2. Estimate experimental error (σ) for each response (noise)
  3. Use the signal to noise ratio (Δy / σ) to estimate power
- Select the input factors to study.
- Select an appropriate design

DoE is in most cases a sequential process

In most cases the experiments must be performed as a sequence of trials

- Screening: A design with 6-12 factors with the purpose of identifying the important ones
- Advanced screening: Investigate possible interaction effects of a smaller number of factors
- Optimization: Find the optimum settings with a more precise model

NB! With proper DoE one can re-use experiments from the previous steps!

Advantages of Experimental Design

	OVAT	**Experimental Design**
Main effects	Not estimated	Estimated
Interactions	Not detected	Detected and estimated
Experimental variability	100% impact	Reduced impact
Number of experiments	Unknown	Known per step
Best solution	???	Spotted
If no solution	???	Detected
Several responses	Difficult	As easy as one response
New objectives	Start all over again	Re-using existing results

Main types of designs

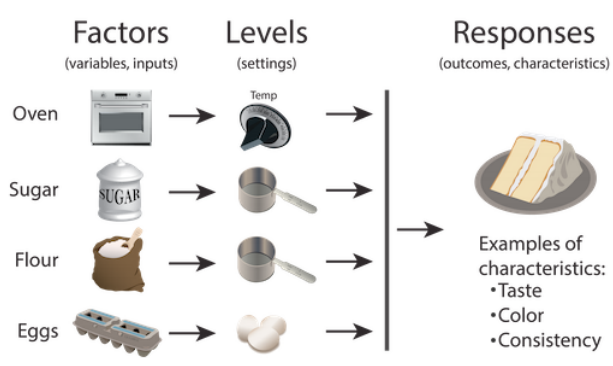
Type of design	*Objective*
Fractional factorial	Find main effects
Full factorial	Find main effects and interactions
Optimization designs	Find optimal settings for a response surface
Mixture designs	Find the optimal recipe of a mixture
Optimal designs	Designs with constraints

A small example

Assume you want to bake the best cake ever

Which are the factors you can change?

What characterize the quality?



Lecture overview

1. DoE: Introduction and motivation
2. **ANalysis Of VAriance (ANOVA)**
3. Factorial designs
4. Fractional factorial designs
5. Response surface designs
6. Optimal designs
7. Metamodelling
8. Combining DoE with multivariate analysis/machine learning
9. QbD – PAT
10. Practical examples of DoE related to cybernetics

Reminder: Reference group - VERY IMPORTANT

- at least 3 students
- will do 4 meetings (1 after the exam)
- shall represent the whole class $\Longrightarrow$ you will have meetings among yourselves too
- shall lead to a referansegrupperapport containing suggestions for improvements

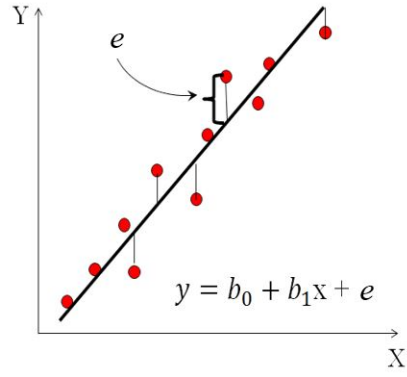
Multiple Linear Regression (MLR) and ANOVA

ANalysis of VAriance (ANOVA)

- ANOVA is the most frequently used way to analyse results from design of experiments
- The main purpose is to estimate the variance in the responses due to the various model terms and assess if the model terms are significant
- Extensions:
  - ANCOVA (additional variables; covariates)
  - MANOVA (simultaneous analysis of several responses)
  - MANCOVA (several responses and covariates)

Linear Regression - the univariate case

- Fit a straight line to the data
- The parameters $\mathbf{b_0}$ and $\mathbf{b_1}$ need to be estimated
- The aim of least squares is to minimize the squared sum of the error terms, $\mathbf{e}$
- Thus, the assumption is that there are no errors in $\mathbf{X}$
- MLR require more objects than variables (influences the design matrix as it depends on the choice of model complexity)

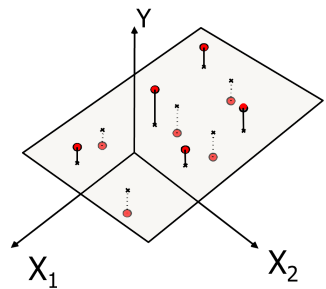


Multiple Linear Regression (MLR) - general case

The model equation, which relates a response variable to several predictors by means of regression coefficients, has the following structure:

$$y = b_0 + b_1 x_1 + b_2 x_2 + ... + b_k x_k + e$$

- Least Squares criterion: the plane should lie where it minimizes the sum of squares of all residuals.
- The method of choice for orthogonal experimental designs
- The ANOVA table is calculated from the regression coefficients in most implementations; in the earlier days by means of square sums (programming it yourself should take 10-15 minutes :-)



Modelling statistics and diagnostics

Modelling statistics

Residual sum of squares

$$SS_{residuals} = \sum_{i=1}^{n} e_i^2$$

R-squared: The amount of variance explained by the model

$$R^2 = 1 - \frac{SS_{residuals}}{SS_{residuals} + SS_{model}}$$

Adjusted R-squared:R-squared adjusted for the number of terms in the model

$$R^2 = 1 - \left(\frac{SS_{residuals}}{df_{residuals}}\right) \Big/ \left(\frac{SS_{residuals} + SS_{model}}{df_{residuals} + SS_{model}}\right)$$

Model diagnostics

- Before concluding on the ANOVA results, one needs to investigate various model diagnostics
  - The structure of the residuals
  - The impact on the model for the individual samples
  - The goodness of the model
- Some statistical figures of merit
  - Leverage (hat matrix)
  - Cook's distance

Model diagnostics - Leverage

The hat matrix $\mathbf{H}$ is given by:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}$$

Leverage for sample *i*:

$h_i =$ diagonal element of $\mathbf{H}$:

$$h_i = \mathbf{x_i}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x_i^T}$$

Leverage is a value between 0 and 1

Model diagnostics - residuals

Estimate of the standard deviation, MSE:

$$\hat{\sigma} = \sqrt{\sum_{i}^{I} e_{i=1}^{2}/(I - K - 1)}$$

Internally Studentized Residual: The residual divided by the estimated standard deviation of that residual. It measures the number of standard deviations separating the actual and predicted values.

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_i}}$$

More modelling statistics

PRESS: Predicted Residual Error Sum of Squares. The error when predicting a sample with a model of which the sample was not included

$$e_{-i} = y_i - \hat{y_{-i}} = \frac{e_{-i}}{1 - h_{ii}}$$

$$PRESS = \sum_{i=1}^{n} e_{-i}^2$$

where

h_{ii} is the leverage of sample *i*

$$\text{RMSE (Root Mean Square)} = \sqrt{\frac{1}{I}\sum_{i=1}^{I}(y_i - \hat{y}_i)^2} = \sqrt{\frac{1}{I}\sum_{i=1}^{I} e_i^2}$$

Model diagnostics - Cook's distance

Cook's distance is a diagnostic that takes into account the residual as well as the leverage of one sample

It represents the change in the regression line when the sample *i* is removed

Cook's distance $= D_i = \frac{r_i^2}{k+1}\left(\frac{h_i}{(1-h_i)}\right)$

The F-distribution and relationship to the t-test

- F-tests are named after Sir Ronald Fisher, who developed the theory in the 1920's
- The F-statistic is simply a ratio of two variances estimated as mean squares corrected for degrees of freedom (DF)
- Rule of thumb: If the variance due to changing a design factor is three times the noise/error, it is most likely not due to chance
- If you have only two groups/factor levels, the F-test statistic is the square of the t-test statistic, and the F-test is equivalent to the two-sided t-test.
- The calculated test statistic F_0 is compared to an F-table for a specified number of degrees of freedom. The form of the test statistic is as follows: $F_{\alpha,n_{1-1},n_{2-1}}$
- α is the significance level that you will decide upon *a priori*

ANalysis Of VAriance (ANOVA)

- ANOVA separates data into contributions from structure and noise

- Data = Structure + Noise
- $SS_{Total} = SS_{Model} + SS_{residuals}$
- Total variation = Modelled + Not modelled

$$\sum_{i=1}^{I}(y_i - \bar{y_i})^2 = \sum_{i=1}^{I}(y_i - \hat{y_i})^2 + \sum_{i=1}^{I}(\hat{y_i} - \bar{y_i})^2$$

ANOVA output (1/2)

Summary-section:

- Model (SS_{Model}):
  - Contribution form all terms in the model
  - Degrees of Freedom (DF) is given by the number of estimated model parameters
- Error ($SS_{residuals}$):
  - Non-modelled variation or noise
  - DF given by number of (runs - number of terms - 1)
- Significance of model is estimated from

$$\text{F-ratio} = \frac{\text{MS}_{\text{Model}}}{\text{MS}_{\text{residuals}}}$$

ANOVA output (2/2)

- Variables-section:
  - The significance of each model parameter is estimated
- Model check selection
  - Sums the contribution from linear terms, interaction terms etc.
- Lack-of-fit section
  - Total error may be divided into
    - Pure error: Spread between replicates
    - Lack-of-fit: Modelled values vs. Mean of replicates

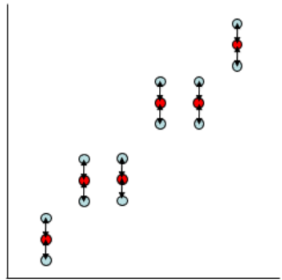
Lack of fit

Testing Lack of fit is important for evaluating the goodness of the model

$SS_{residuals} = SS_{pure\ error} + SS_{lack\ of\ fit}$

$SS_{pure\ error}$ = SS of the replicates about their means

$SS_{lack\ of\ fit}$ = SS of the means about the fitted model.



$$F = \frac{MS_{lack\ of\ fit}}{MS_{pure\ error}}$$

Is the variation about the model greater than what is expected given the variation of the replicates about their means?

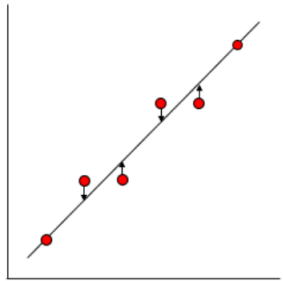
Residual distance

Lack of fit is compared to the residual distance to the model

$SS_{residuals} = SS_{pure\ error} + SS_{lack\ of\ fit}$

$SS_{pure\ error}$ = SS of the replicates about their means

$SS_{lack\ of\ fit}$ = SS of the means about the fitted model.



$$F = \frac{MS_{lack\ of\ fit}}{MS_{pure\ error}}$$

Is the variation about the model greater than what is expected given the variation of the replicates about their means?

ANOVA table

The ANOVA table for the overall model has the following structure:

Table: The structure of the ANOVA table for the overall model

Source	**SS**	**df**	**MS**	**F-ratio**	p-value
Model	SS_{Model}	k	MSR $= SS_{Model}/$(k)	MSR/MSE	p
Error	SS_{Error}	I-k-1	MSE $= SS_{Residual}/$(I-k-1)		
Total	SS_{Total}	I-1	MST $= SS_{Total}/$(I-1)		

ANOVA for the individual model terms

The ANOVA table for the individual model terms has the following structure, here shown for a model with two variables at two levels:

Table: The structure of the ANOVA table for the individual model terms

Source	**SS**	**df**	**MS**	**F-ratio**	p-value
Intercept	$SS_{Intercept}$	1	MS	MS/MSE	*p*
Variable1	$SS_{Variable1}$	1	MS	MS/MSE	*p*
Variable2	$SS_{Variable2}$	1	MS	MS/MSE	*p*

ANOVA in case of unbalanced designs and non-orthogonal designs

- When the design is orthogonal, the way the square sums is estimated will not alter the ANOVA table
- In other cases there is no "truth" in how to estimate the square sums
- In short, the options are:
  - Type I: Estimate the square sums sequentially: First assign a maximum of variation to variable A; in the remaining variation, assign the maximum of variation to variable B.
  - Type II: This type tests for each main effect after the other main effect SS(A|B), SS(B|A). The common option if you are mostly interested in the main effects.
  - Type III: Estimate square sums while taking into account all other model terms; For models with interactions terms

A small example to illustrate MLR and ANOVA

- House prices in the Boston area, 434 samples
- Model: Median value of property = f(No. of rooms, age)
- Demo in Design-Expert®
  - Look at raw data
  - ANOVA
  - Diagnostics plots
  - Model statistics

Multiple Linear Regression details

MLR details - I

Estimation of regression coefficients $\mathbf{b}$

$$\hat{\mathbf{b}} = (\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\mathbf{y} \tag{1}$$

NB! This requires full rank in $\mathbf{X}$

The fitted values of $\mathbf{y}$ (prediction from the calibration):

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{b}} = \mathbf{X}(\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\mathbf{y} \tag{2}$$

The Y-residuals, e_i are given by

$$e_i = y_i - \hat{y}_i \tag{3}$$

MLR details - II

The error standard deviation is estimated as

$$\hat{\sigma} = \sqrt{\sum_{i=1}^{I} e_i^2/(I-K-1)} \tag{4}$$

The variance of b_0:

$$\hat{\sigma}_{b_0} = \hat{\sigma}\left[\frac{1}{I} + \frac{\bar{\mathbf{X}}^{\mathbf{2}}}{\mathbf{X}^{\mathbf{T}}\mathbf{X}}\right] \tag{5}$$

The variances of $b_1, ..., b_K$:

$$\hat{\sigma}_b = \hat{\sigma}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1} \tag{6}$$

MLR details - III

The t-statistic for the beta coefficients is:

$$t = \frac{\hat{b}}{\hat{b}_\sigma} \tag{7}$$

The critical t-value is given by the t-distribution with (I-K-1) degrees of freedom.
The confidence interval for b:

$$\hat{b} \pm t_{\alpha/2}\hat{\sigma}_b \tag{8}$$

Lecture overview

1. DoE: Introduction and motivation
2. ANalysis Of VAriance (ANOVA)
3. **Factorial designs and Fractional factorial designs**
4. Response surface designs
5. Optimal designs
6. Metamodelling
7. Combining DoE with multivariate analysis/machine learning
8. QbD – PAT
9. Practical examples of DoE related to cybernetics

Reminder: Reference group - VERY IMPORTANT

- at least 3 students (two or more seats free!)
- will do 4 meetings (1 after the exam)
- shall represent the whole class $\Longrightarrow$ you will have meetings among yourselves too
- shall lead to a referansegrupperapport containing suggestions for improvements

Experimental Work: The Basic Questions - recap

- Which design factors in my system are the most important?
- Are there any interactions?
- How can I get maximum information at minimum cost?
- Where is the optimal region?
- Where is the stable region?
- How can I span the variation of my calibration variables?
- How can I build a good calibration / validation data set?

Factorial designs

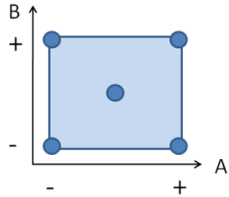
The full factorial design

Motivation for use

- Simplest design situation
- Basis for many other designs
- Optimal for detecting main effects and their interactions

2-level full factorial designs

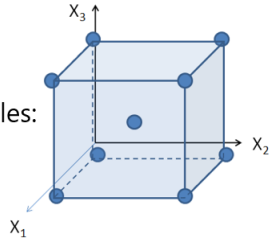
2 X-variables:



run #	X_1	X_2
2	-	-
4	-	+
6	+	-
1	+	+
3	0	0
5	0	0

2^2 experiments (+ centre samples)

3 X-variables:

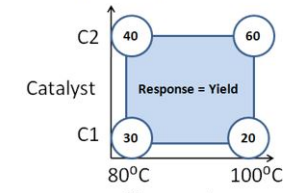


run #	X_1	X_2	X_3
2	-	-	-
11	-	-	+
5	-	+	-
8	-	+	+
4	+	-	-
1	+	-	+
9	+	+	-
7	+	+	+
3	0	0	0
6	0	0	0
10	0	0	0

2^3 experiments (+ centre samples)

Calculating effects - main effects

A simple experimental design:



Average of C at High Level = (40+60)/2 =**50**

Average of C at Low Level = (30+20)/2 =**25**

Main Effect of Catalyst

=50-25

= +25

The Main Effects indicate what the overall effect a single design factor has on the overall response

Average of T at Low Level = (30+40)/2 =**35**

Average of C at High Level = (20+60)/2 =**40**

Main Effect of Temperature =40-35 = +5

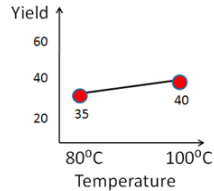
Calculating effects - interaction effects

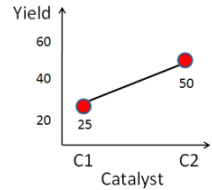
The simplest way to find how to calculate the various effects is to use the design table: An example for 2^3 full factorial design

Std	A	B	C	AB	AC	BC	ABC	
1	–	–	–	+	+	+	–	y_1
2	+	–	–	–	–	+	+	y_2
3	–	+	–	–	+	–	+	y_3
4	+	+	–	+	–	–	–	y_4
5	–	–	+	+	–	–	+	y_5
6	+	–	+	–	+	–	–	y_6
7	–	+	+	–	–	+	–	y_7
8	+	+	+	+	+	+	+	y_8

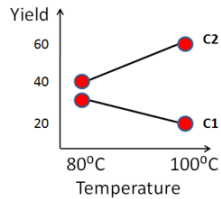
Interpreting effects - overview

Main effects



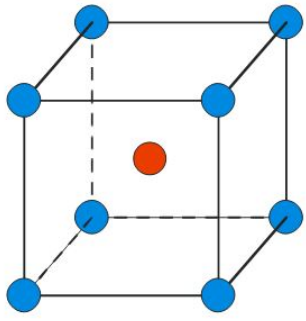


Interactions

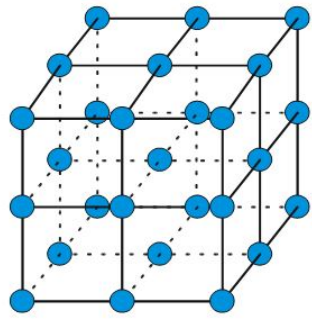


But what about adding more levels to the design factors?

- Adding more levels will rapidly increase the number of runs
- If the goal is to have a more precise description of the design space, other designs are more economical



2^3 factorial with center point
(8 runs plus 4 cp's = 12 pts)



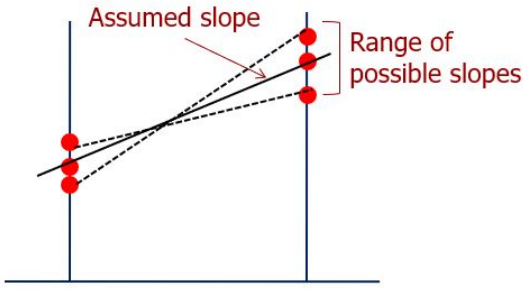
3^3 Three-level factorial
(27 runs + 5 cp's = 32 pts)

Additional experiments

- Center samples
  - To detect curvature
  - To estimate error variance
  - For category variables need one experiment for each level
- Replicated samples
  - Replication of factorial points
  - More precise estimate of error variance
- Remember the assumptions about the residuals, $N(0, \sigma^2)$:
  - Normally distributed
  - Mean of zero
  - Constant variance

Replicates and center samples

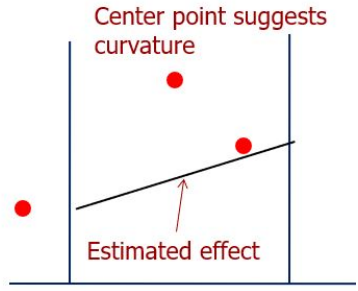
Replicates :



Precision

$SD_{\text{repl. samples}} << SD_{\text{whole design}}$?

Center samples :



Curvature check

$\overline{Y}_{\text{center samples}} = \overline{Y}_{\text{design}}$?

Power of an experimental design

- The power of a design *must* be calculated prior to performing any experiments!
- The calculation of power takes four inputs:
  1. Δ, what *you* regard as a significance difference for a given response
  2. σ, the noise in the measurement of the response variable
  3. The number of experiments
  4. The significance level α

Estimation of power

Power = (1−β)*100%

Power is the probability of revealing an active effect of size delta (Δ) relative to the noise (σ) as measured by signal to noise ratio (Δ/σ).

It should be high (at least 80%!) for the effect size of interest.

Effect?		ANOVA says:	
		Retain H_0	***Reject H_0***
Truth:	***No***	OK☺	Type I Error (alpha) *False Alarm*
	Yes	Type II Error (beta) *Failure to detect*	OK☺

How to Select Ranges of Variation

- Wide enough to generate response variation
- Narrow enough to avoid huge non-linearities
- Useful tip: Start with two extreme combinations
  - If too extreme results: narrow down
  - If different enough results: OK
  - If too close results: Check center samples

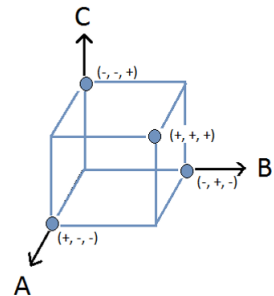
Fractional Factorial designs

Fractional Factorial designs

- Full factorial designs are expensive if many design factors
- Often higher order interactions can be neglected → Fractional factorial design
- Subset of the full factorial design
- Experiments are systematically chosen to cover the widest possible design space

2-level fractional factorial design

- 3 design variables, A, B, C
- 2^{3-1} design, C = AB
- All main effects are estimated in $2^{3-1} = 4$ runs
- The main effects are aliased with the interaction effects



Aliasing/Confounding

- The price to be paid for performing fewer experiments (fractional designs)
- $\rightarrow$ some effects cannot be studied independently of each other
- The degree of confounding is described by the *confounding pattern* and the *resolution*

Constructing a 2-level Fractional Factorial design: Aliasing

- Example: Constructing the 2^{4-1} Design from a 2^3 Design
- Write out the full design
- Let D = ABC (aliasing with the highest interaction term)

A	B	C	AB	AC	BC	ABC
-	-	-	+	+	+	-
+	-	-	-	-	+	+
-	+	-	-	+	-	+
+	+	-	+	-	-	-
-	-	+	+	-	-	+
+	-	+	-	+	-	-
-	+	+	-	-	+	-
+	+	+	+	+	+	+

Aliasing/Confounding: Four design variables

Example: four design variables, can only afford 10 runs
2^{4-1} fractional factorial design + 2 center samples

The column for **D** equals the interaction **ABC**

Defining relation: **I=ABCD**, resolution=**IV**

This gives the following confounding pattern:

A=BCD
B=ACD
C=ABD
D=ABC
AB=CD
AC=BD
AD=BC

These effects cannot be estimated separately from each other

	A	B	C	D=ABC
1	-	-	-	-
2	+	-	-	+
3	-	+	-	+
4	+	+	-	-
5	-	-	+	+
6	+	-	+	-
7	-	+	+	-
8	+	+	+	+
9	0	0	0	0
10	0	0	0	0

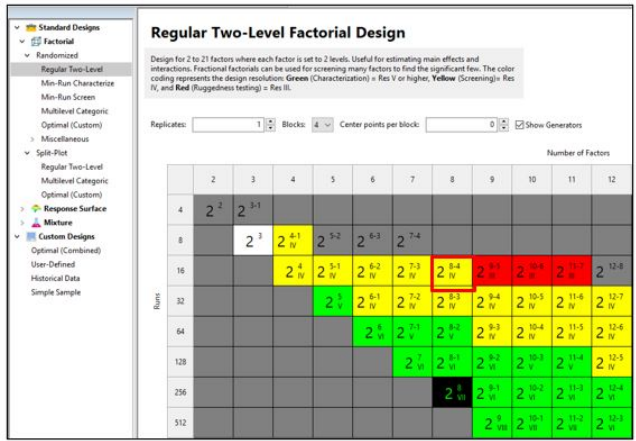
The resolution of a design

- Resolution V: Main effects are not confounded with 2-way interactions
- Resolution IV: 2-way interactions are confounded and main effects are confounded with three-factor interactions
- Resolution III: Main effects are confounded with 2-way interactions

Number of experiments for various designs

	Resolution		
Factors	**Full**	**V**	**Runs**
5	32	16	$^1/_2$
6	64	32	$^1/_2$
7	128	64	$^1/_2$
8	256	64	$^1/_4$
9	512	128	$^1/_4$
10	1,024	128	$^1/_8$
11	2,048	128	$^1/_{16}$
12	4,096	256	$^1/_{16}$
13	8,192	256	$^1/_{32}$
14	16,384	256	$^1/_{64}$
15	32,768	256	$^1/_{128}$

Resolution and confounding patterns for various designs



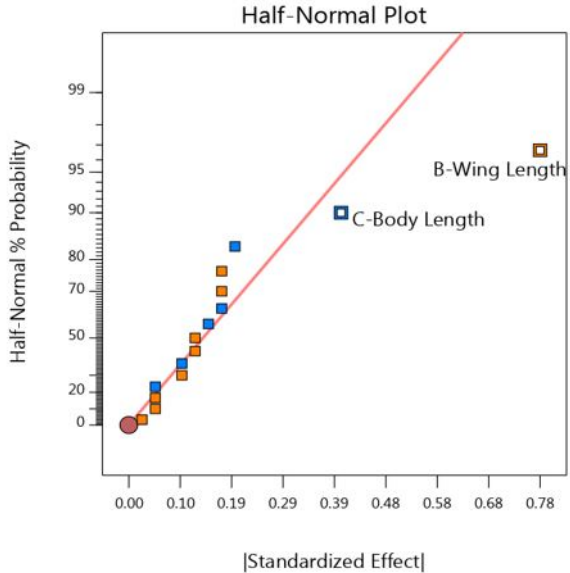
How to identify the important effects?

- If all main and interaction effects are estimated, there are no degrees of freedom for the error, and thus no output for F-ratios and p-values in the ANOVA table.
- However, the effects can be visualized in various plots to make an assessment:
  - Half-Normal Plot
  - Normal Plot
  - Pareto Chart

Half-Normal Plot

- If there are no significant effect, they should follow a normal distribution
- One way of visualizing this is in the normal probability plot (ref. plot of residuals)
  1. Sort the effects
  2. Plot them on a logarithmic scale
- Effects that deviate from a straight line in this plot might be significant
- One can show the effects as absolute values (half-normal) or as is (normal plot)

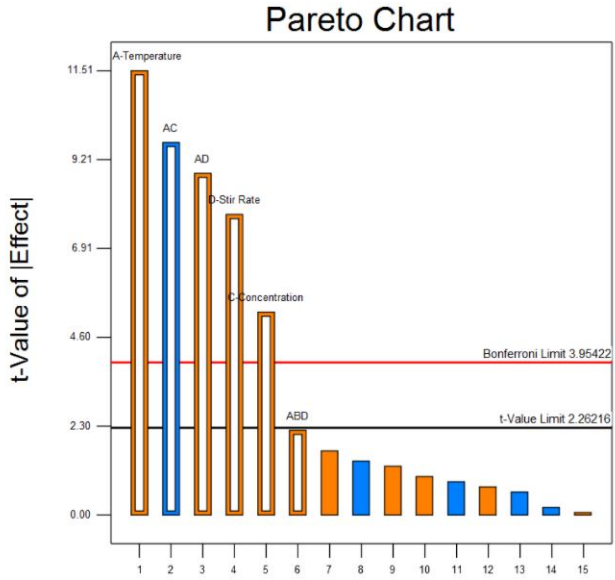
Half-Normal Plot - example



Pareto Chart

- The Pareto Chart is another option to show the importance of the effects
- It has two critical limits:
  - t-limit: A limit based on the t-distribution
  - Bonferroni limit: A conservative limit taking into account the number of terms in the model
- Selected Effects that are above the Bonferroni Limit are almost certainly important and should be left in the model.
- Effects that are above the t-value Limit are possibly important and should be added if they make sense to the experimenter
- Effects that are below the t-value limit should only be selected to support hierarchy. They can also be forced into the model by the analyst.

Pareto Chart - example



Small example Factorial design: Popcorn

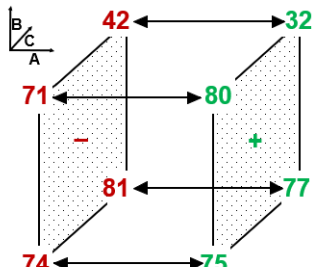
Popcorn example: Finding the best experimental settings

Multi-response optimization

Std	A: Brand expense	B: Time minutes	C: Power percent	R_1: Taste rating	R_2: UPKs oz.
1	Cheap	4.0	75.0	74	3.1
2	Costly	4.0	75.0	75	3.5
3	Cheap	6.0	75.0	71	1.6
4	Costly	6.0	75.0	80	1.2
5	Cheap	4.0	100.0	81	0.7
6	Costly	4.0	100.0	77	0.7
7	Cheap	6.0	100.0	42	0.5
8	Costly	6.0	100.0	32	0.3

Calculations of effects: Popcorn

Popcorn example



A: Brand

$$\text{Effect}(\Delta y) = \frac{\sum y_+}{n_+} - \frac{\sum y_-}{n_-}$$

$$\Delta y_A = \frac{75+80+77+32}{4} - \frac{74+71+81+42}{4} = -1$$

Small example 2: Factorial design

- Optimization of filtration rate in a chemical process
- A 2^4 full factorial design
- Four numerical design factors
  - Temperature
  - Pressure
  - Concentration
  - Stir rate
- Response variable: Filtration rate

Lecture overview

1. DoE: Introduction and motivation
2. ANalysis Of VAriance (ANOVA)
3. Factorial designs and Fractional factorial designs
4. **Response surface designs and model selection**
5. Optimal designs
6. Metamodelling
7. Combining DoE with multivariate analysis/machine learning
8. QbD – PAT
9. Practical examples of DoE related to cybernetics

Topics for today

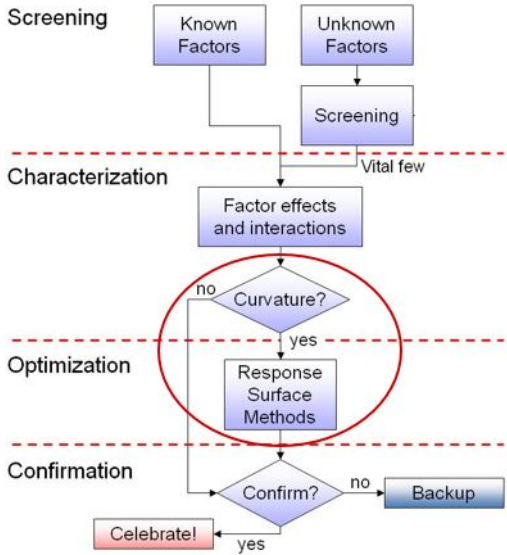
- Response surface designs
- Model selection
- Demo: Response surface design
- If time allows: Rerun of helicopter design from last week

What if you cannot perform all experiments in one day? Some words about blocking

- Blocking is a technique used to mathematically remove the variation caused by some identifiable change during the experimental campaign
- E.g. if you extend a factorial design to an optimization design, and performing the new experiments the week after
- It is assumed that the block variable does not interact with the factors.
- You can look at the alias structure to see which effects have been “lost to blocks".

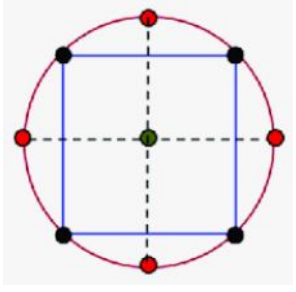
Response surface designs

Strategy for experimentation



Optimization with two design factors

Optimization with two design factors



Optimization designs

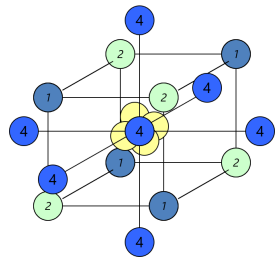
- Objective
  - Model the response surface with *accuracy*, so as to know the *precise shape* of the response surface and find the *optimal* values
- The model terms
  - Include main effects
  - Include interactions
  - Include squared and/or cubic terms
- Design types
  - Central Composite designs
  - Box-Behnken designs
  - Optimal designs (in situations with constraints or to minimize the number of runs)

While a two-level design with center points cannot estimate individual pure quadratic effects, it can detect them effectively

Central Composite Designs (CCD)

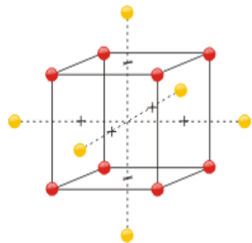
A full factorial 2-level design can be extended to a *Central Composite Design* by adding star points

- Good for modelling a response surface
- 5 levels for each variable
- Can be built as an extension of a full factorial
- Additional points are called star or axial points

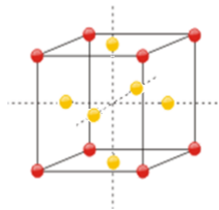


1. Fractional factorial, 2. Full factorial, 3. Centre points, 4. Axial points

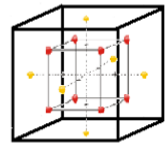
Types of CCD



Circumscribed central composite (CCC)



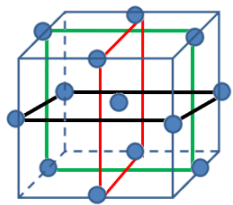
Faced central) composite (CCF)

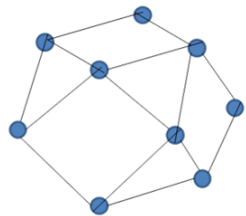


Inscribed central composite (CCI)

Bob-Behnken designs

- 3 levels for each variable
- Slightly fewer runs than CCD
- Extreme combinations are avoided





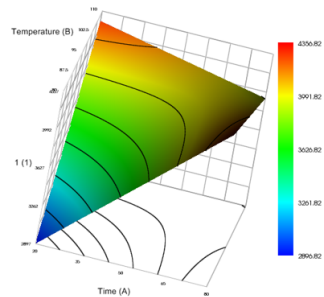
How to select ranges of variation

- Wide enough to generate response variation
- Narrow enough to avoid huge non-linearities
- Useful tip: Start with two extreme combinations
  - If too extreme results: narrow down
  - If different enough results: OK
  - If too close results: Check center sample
    - If close to the others: widen up
    - If different: curvature (narrow down)

Response surface methodology

$$y = b_0 + \sum b_i x_i + \sum b_i x_i^2 + \sum \sum b_{ij} x_i x_j + f$$

- Purpose: Closely approximate the true shape of the response surface
- Quadratic model
- Method: MLR/ANOVA
- Predict the response value(s) for any combination of the design variable settings in the experimental region
- Find the variable settings that give desired response value(s) in the experimental region (optimization)

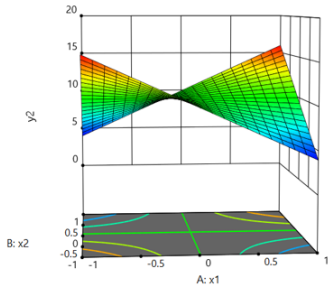


Plots for main results and diagnostics

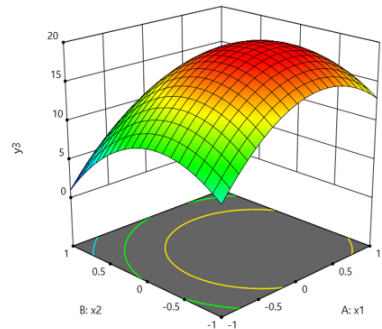
- Plots of main effects interactions (e.g. half-normal plots)
- Response surface (contour, 3D surface)
- Predicted vs. Reference
- Residuals
- Various statistical diagnostics, e.g. R^2

Examples of response surfaces - I

Two-factor interaction

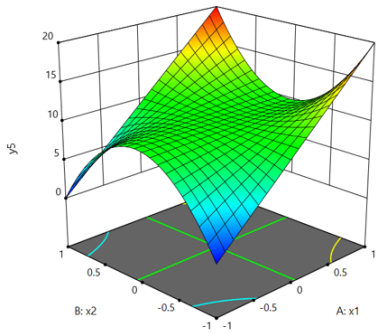


Squared terms

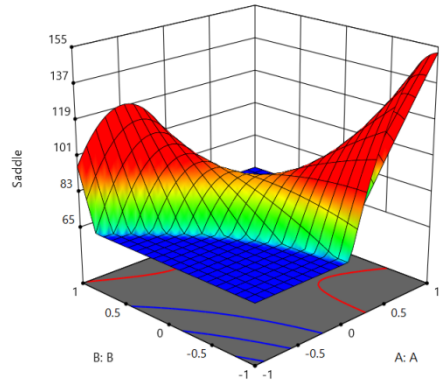


Examples of response surfaces - II

Cubic model: Inflection



Cubic model: Saddle



Example: CCD with three design factors

Std	Run	Factor 1 A:Time min.	Factor 2 B:Temperature deg C	Factor 3 C:Catalyst %	Response 1 Conversion %	Response 2 Activity
1	6	40	80	2	74	53.2
2	20	50	80	2	51	62.9
3	7	40	90	2	88	53.4
4	2	50	90	2	70	62.6
5	4	40	80	3	71	57.3
6	3	50	80	3	90	67.9
7	15	40	90	3	66	59.8
8	14	50	90	3	97	67.8
9	8	36.591	85	2.5	81	59.2
10	11	53.409	85	2.5	75	60.4
11	12	45	76.591	2.5	76	59.1
12	10	45	93.409	2.5	83	60.6
13	9	45	85	1.6591	76	53.6
14	1	45	85	3.3409	79	65.9
15	16	45	85	2.5	85	60
16	17	45	85	2.5	97	60.7
17	18	45	85	2.5	55	57.4
18	13	45	85	2.5	81	63.2
19	5	45	85	2.5	80	60.8
20	19	45	85	2.5	91	58.9

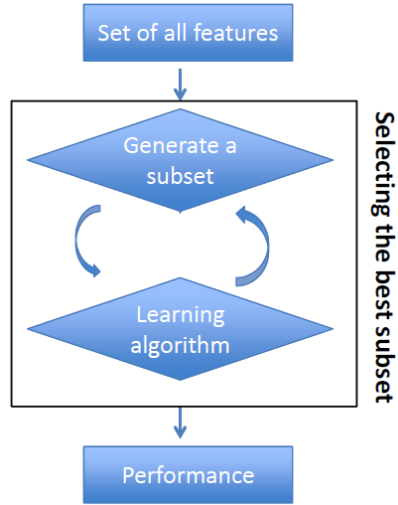
ANOVA for a quadratic model with replicated center samples

Source	Sum of Squares	df	Mean Square	F-value	p-value	
Model	1603.32	9	178.15	1.52	0.2625	not significant
A-Time	0.0871	1	0.0871	0.0007	0.9788	
B-Temperature	160.19	1	160.19	1.36	0.2700	
C-Catalyst	155.25	1	155.25	1.32	0.2771	
AB	36.12	1	36.12	0.3075	0.5914	
AC	1035.12	1	1035.12	8.81	0.0141	
BC	120.12	1	120.12	1.02	0.3358	
A^2	42.42	1	42.42	0.3611	0.5613	
B^2	20.25	1	20.25	0.1723	0.6868	
C^2	51.61	1	51.61	0.4393	0.5224	
Residual	1174.88	10	117.49			
Lack of Fit	127.38	5	25.48	0.1216	0.9814	not significant
Pure Error	1047.50	5	209.50			
Cor Total	2778.20	19				

Model selection

Model selection

- Several subsets of features are searched through, and the data modelling accuracy is evaluated.
- Advantages:
  - Allow detection of possible interactions between variables
  - Often finds a good subset for a specific learning algorithm
- Disadvantages:
  - Risk of overfitting when the number of observations is low
  - Computationally demanding when the number of features is large



Examples of various approaches

- **Forward selection**
  - Starts with having no feature in the mode
  - Iteratively adds the feature which best improves the model
  - Stops when no improvement is observed upon addition of features
- **Backward elimination**
  - Starts with all the features
  - Iteratively removes the least significant feature
  - Stops when no improvement is observed upon removal of features
- Stepwise Selection:
  - Re-considers all dropped/added variables for reintroduction/drop-out in each step
- Randomized wrapper methods:
  - Include some degree of randomness in the selection

NB! Good approach for orthogonal designs, with non-orthogonal designs and with covariates present must be careful with interpretation

Criteria for including or excluding model terms

- Adjusted R-squared
- p-values
- Akaike's Information Criterion (AIC) or AICc
- Bayesian Information Criterion (BIC)

When selecting your model, it is suggested to use multiple methods and criteria

More details

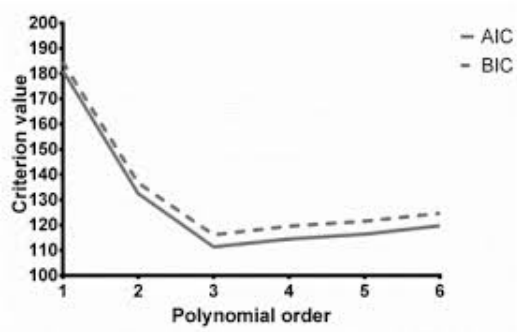
- For all methods:
  - Use Type II or Type III Sum of Squares
  - Select model complexity (main effects, interactions, squares,...)
- For p-values, AIC and BIC: Forward or backward
- for p-values and Adjusted R-squared: Cut-off for excluding model terms

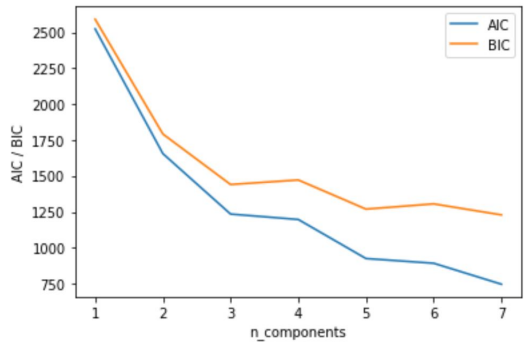
In general: What is the least significant difference in R-squared that should be applied to say that one model is better than another?

AIC and BIC

- AIC rewards goodness of fit, but it also includes a penalty that is an increasing function of the number of estimated parameters
- When the sample size is small, there is a substantial probability that AIC will select models that have too many parameters
- To address such potential overfitting, AICc was introduced: AICc is AIC with a correction for small sample sizes
- BIC is similar to AIC but applies a different penalty for the number of parameters

Comparison of AIC and BIC





Demo: Optimization design

- Objective: Optimize the conversion of a chemical reaction
- A Central Composite Design
- Design factors: Time, temperature, catalyst (%)
- Responses: Conversion (%), activity
- Procedure:
  - Set up the design
  - Inspect raw data
  - Analyze data
  - Numerical optimization

Optimal designs and mixtures (designs with constraints)

Introduction to constrained designs

- Optimal designs
- Mixture designs
- Two examples:
  - Example 1, baking a cake: with long time and high temperature the cake is burned
  - Example 2, cocktail mix: the sum of ingredients is 100% (lime juice, lemon juice, tonic water)

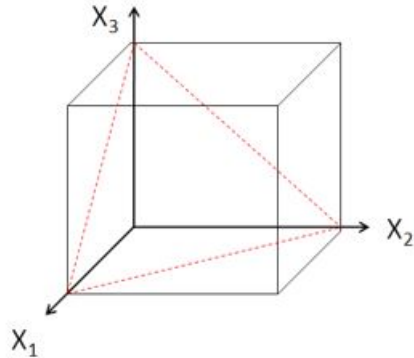
Introduction to constrained designs

- Optimal designs
- Mixture designs
- Two examples:
  - Example 1, baking a cake: with long time and high temperature the cake is burned
  - Example 2, cocktail mix: the sum of ingredients is 100% (lime juice, lemon juice, tonic water)

  Problem: the design variables involved cannot vary completely independently from the others:
  - In case 1, the maximum allowed temperature will depend on the baking time
  - In case 2, the proportions of the cocktail must add up to 100%

Mixture designs

- 3 dimensions collapsed into 2
- For a mixture of 3 ingredients the experimental region becomes flat
- This shape is called a simplex
- The simplex region contains all possible combinations



As there is closure in the design this must be handled in the ANOVA, with so-called Scheffé polynomials

Scheffé polynomials

First order:

$$Y = \sum_{i=1}^{q} \beta_i X_i$$

Second order:

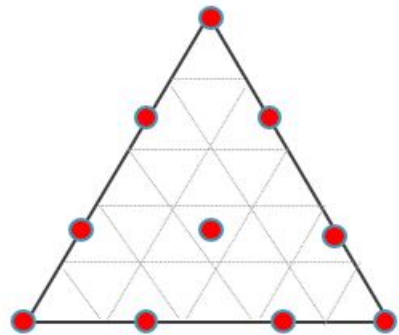
$$Y = \sum_{i=1}^{q} \beta_i X_i + \sum_{i=1}^{q} \sum_{j=i+1}^{q} \beta_{ij} X_i X_j$$

Special cubic:

$$Y = \sum_{i=1}^{q} \beta_i X_i + \sum_{i=1}^{q} \sum_{j=i+1}^{q} \beta_{ij} X_i X_j + \sum^{q} \sum_{i<j<k}^{q} \sum^{q} \beta_{ijk} X_i X_j X_k$$

Simplex Lattice designs

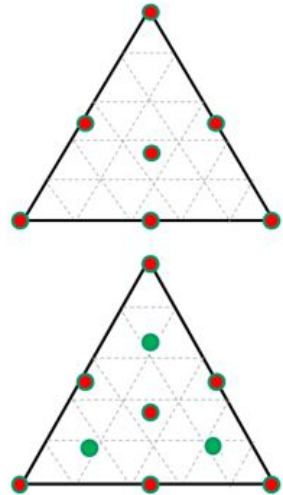
- The Simplex Lattice design has the form $[q,m]$ where q is the number of mixture components and m is the order of the design to be supported.
- Excellent designs for investigating the extremes of a design space and for building prediction models



The [3,3] Simplex Lattice Design

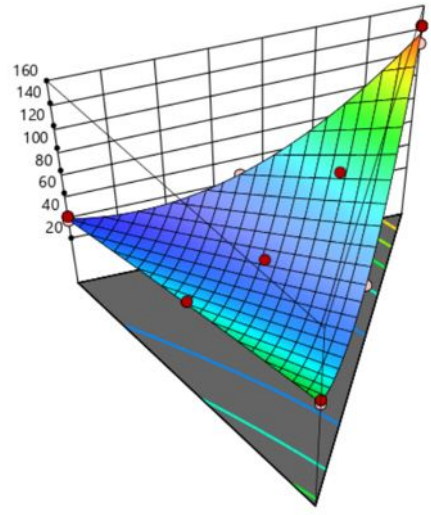
Simplex centroid designs

- In a q component simplex, the number of distinct points is $2q$-1
- Simple design for investigating the entire region economically.
- May be further augmented with axial check blends

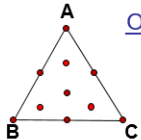


Simplex response surface

- 3D-representation of the mixture triangle
- The three coordinates sum to 100%



Mixture design - Rocket fuel



Objective: Maximize elasticity of solid rocket fuel.

Design: Augmented Simplex Lattice.

	Fuel 1	Fuel 2	Fuel 3	Elasticity	
Pure	1	0	0	323	328
Mixtures	0	1	0	298	430
	0	0	1	674	737
Binary	1/2	1/2	0	210	307
Blends	1/2	0	1/2	657	816
	0	1/2	1/2	940	850
Centroid	1/3	1/3	1/3	888	1068
Check	2/3	1/6	1/6	713	711
Blends	1/6	2/3	1/6	808	739
	1/6	1/6	2/3	1012	972

Optimal designs - Optimality criteria I

There exists a number of different criteria to replace the initial set of points to find the best subset

- I-optimal
  - The I-optimal criteria is recommended to build response surface designs where the goal is to optimize the factor settings, requiring greater precision in the estimated model.
  - Gives a good distribution inside the design space
- D-optimal
  - A D-optimal algorithm chooses runs that minimizes the determinant of the variance-covariance matrix
  - This minimizes the errors in the regression coefficients
  - The D-optimal criteria is recommended to build factorial designs where the goal is to find factors important to the process

Optimal designs - Optimality criteria II

- A-optimal
  - An A-optimal design minimizes the trace of the variance-covariance matrix.
  - This has the effect of minimizing the average prediction variance of the polynomial model coefficients.
- Modified Distance
  - The modified distance-based point selection algorithm selects model points to obtain a maximum spread throughout the design region while ensuring that adequate runs are chosen to fit the polynomial model
  - Thus, this gives the "best of both worlds"
- Distance (Maximin)
  - This design scatter points throughout the experimental region
  - The algorithm starts from a vertex and then adds points that maximize the minimum Euclidean distance from points already in the design

Optimal exchange methods

Two algorithms are frequently used to decide on the best optimal design. The starting point is a pseudo-random set of points.

- Coordinate exchange
- Point exchange
- The methods require a set of initial candidate points:
  - Vertices are at the extreme limits of the design
  - Center of edges are half-way between two vertices
  - Thirds of edges are one-third and two-thirds the way along an edge between two vertices
  - Constraint plane centroids are the center of a hyper-plane defined by three or more coplanar vertices
  - Axial check points are half-way between the overall centroid and the vertices
  - Interior points are half-way between the centroid and center of edges
  - Overall centroid is the geometric center-of-mass of the design

Coordinate exchange algorithm

1. Select a random initial set of p points, where p is the number of terms in the designed for model
   - Start with a random coordinate (point) within the design space
   - Randomly pick each subsequent design point and evaluate if it increases the rank of the matrix. Continue this process until a full rank matrix is obtained.
2. Randomly select any extra model points
3. Start the coordinate exchange algorithm
   - Calculate the current optimality criterion (OC)
   - Starting with the first point, move it along a set of directions in incremental steps
   - If the OC improves, change the point and move on to the next point in the list. If the OC does not improve, retain the point and move on to the next point
   - Repeat the algorithm several times to improve the odds of finding the globally optimal design
4. Lack-of-fit points and replicates are chosen that best support the optimality criterion

Point exchange

1. Define a candidate set of possible factor combinations
2. Select a random initial set of p points from the above candidate set where p is the number of terms as specified in the model
   - Start with a random point from the candidate set
   - Randomly pick each subsequent design point. If a new point increases the rank of the matrix it is added to the set of points
   - Randomly select any extra model points
3. Perform exchange steps by exchanging first 1 point, then continue up to exchanging 10 points until there is no improvement

Evaluation of the designs

- One important property of the design is the condition number
- It is calculated as the ratio between the largest and smallest eigenvalue
- It measures the multicollinearity of the design
- An orthogonal factorial design has condition number of 1
- There's some guidelines of the size of the condition number w.r.t. the subsequent analysis:
  - $1 < 100$ Multicollinearity does not pose a problem
  - $100 < 1000$ Moderate to strong multicollinearity
  - > 1000 Severe multicollinearity

Example: Mixture design

- Objective: Make a system for testing air condition systems in cars
- Technology: Multivariate calibration with six gas sensors that are partly selective for the individual gases
- Gases and their concentration ranges (anonymized):

Component	Name	Minimum	Maximum
A	G1	75	89
B	Air	0	15
C	Propane	0	10
D	G2	0	5
E	G3	0	6
F	G4	0	6

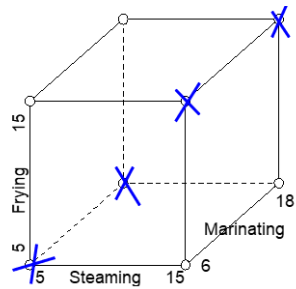
Example of a constrained situation - I

- Optimizing the quality of cooked Meat
  - Design variables: Marinating Time, Steaming Time, Frying Time
  - Responses: Sensory measurements
- Original idea: Full Factorial design
- 8 experiments combining the low and high levels of the variables

Sample	Marinating	Steaming	Frying
1	6	5	5
2	18	5	5
3	6	15	5
4	18	15	5
5	6	5	15
6	18	5	15
7	6	15	15
8	18	15	15

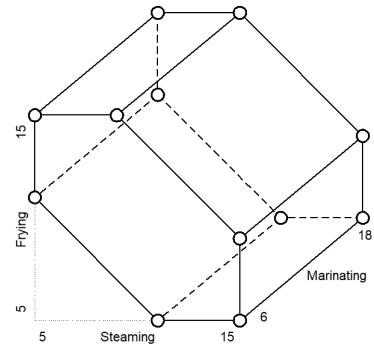
Example of a constrained situation - II

- Extreme combinations should be avoided:
  - Steaming + Frying $< 16 \implies$ raw meat
  - Steaming + Frying $> 24 \implies$ overcooked
- Full factorial does not apply:
  - 4 out of 8 cube samples are excluded
  - The remaining 4 are not enough to explore the region of interest



Visualizing the constrained regions

- Multi-linear constraints:
  - Steaming + Frying $\geq$ 16 min
  - Steaming + Frying $\leq$ 24 min
- The experimental region becomes a polyhedron
- Orthogonal designs are no longer possible
- $\Longrightarrow$ optimal design



How to define constraints

- Multi-factor constraints must be entered as an equation
- Constraints points (CPs) need to be defined
- The equation describes the boundaries for the experimental region
- The following equation applies:

$$1 \leq \frac{A - LL_A}{CP_A - LL_A} + \frac{B - LL_B}{CP_B - LL_B}$$

A small example of setting constraints

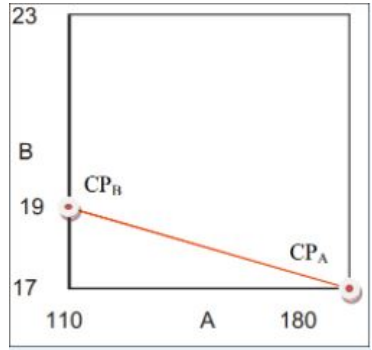
- Assume to design factors, A and B
- When A is at low level, B must be higher than 19

In this case:

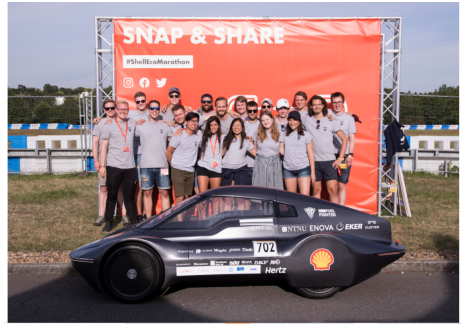
$$1 \leq \frac{A-110}{180-110} + \frac{B-17}{19-17}$$

which leads to:

$$775 \leq A + 35B$$



Fuel fighter - the NTNU team



- Problem: What influences the vertical acceleration on the front and rear of the car when driving over speed bumps?
- ... and how to find the optimal settings given a maximum value for the acceleration?
- Design factors: Height of bump, center of gravity, weight and speed
- Will most likely be presented as an exercise in two weeks' time!

Today's topics

- Statistical Process Control (SPC)
- Multivariate Statistical Process Control (MSPC)
- Crash course in Principal Component Analysis (PCA)
- Quality by Design (QbD)
- Process Analytical Technology (PAT)

(Multivariate) Statistical Process Control

Statistical Process Control

- Statistical Process Control (SPC) is a common methodology used in various industries to monitor and control a process. Various statistical methods are used to ensure that the quality is maintained during the manufacturing process.
- The purpose is to monitor the process and detect once the process becomes out-of-control
- The next slides will shortly present some of the typical outputs and plots from SPC

Moving range (MR)

The moving range is calculated as the absolute difference between each data point x_i and its predecessor x_{i-1}:

$$MR_i = |x_i - x_{i-1}|$$

For $\mathbf{m}$ individual data points, there are $\mathbf{m-1}$ ranges. The average moving range is defined as:

$$\bar{MR} = \frac{\sum_{i=2}^{m} MR_i}{m-1}$$

Moving Range - limits

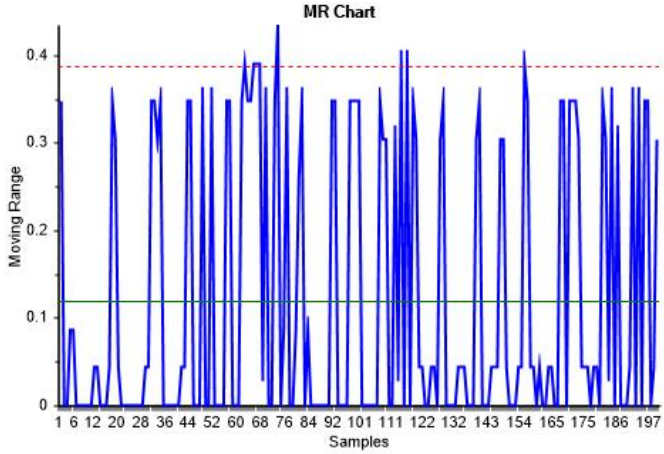
The upper and lower control limits for the moving range are given as:

$$UCL = D_4 \bar{MR}$$

$$LCL = 0$$

D_4 is an unbiasing constant equal to 3.267.

Moving Range - example



Plotting individual values with limits: I Chart

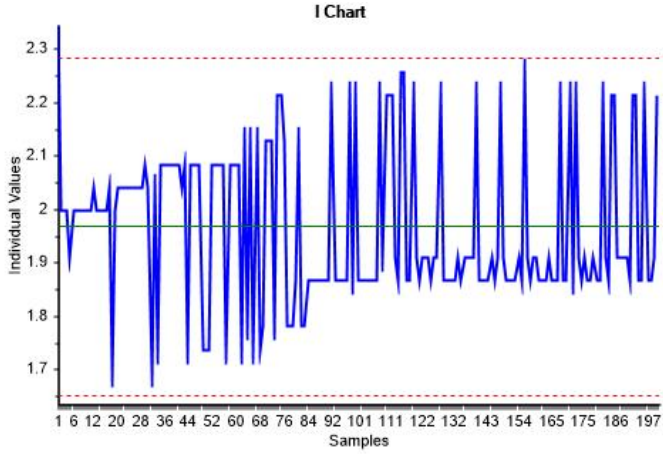
Calculations:

Center Line (CL) $= \bar{X} = \frac{\sum_{i=1}^{m} x_i}{m}$

The Upper and Lower Control Limits (UCL and LCL) are calculated as:

UCL $= \bar{X} + 3\frac{MR}{1.128}$, LCL $= \bar{X} - 3\frac{MR}{1.128}$

I Chart - example

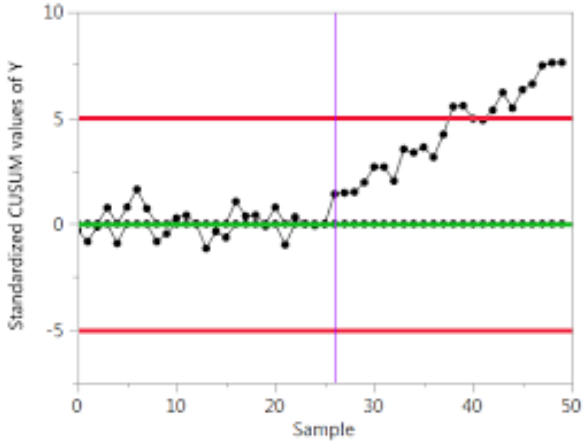


CuSum

- A main disadvantage with the control charts described above is that they use only a single sample and are not based on the sequence of samples. This makes them insensitive towards small shifts in the process and will require a lot of samples before detecting the shift.
- One effective alternative when small shifts are important is the use of CuSum control charts. It plots the cumulative sum of the deviations of the sample values from a target value μ_0:

$$C = \sum_{i=1}^{m} (\bar{x}_i - \mu_0)$$

CuSum chart



S Control chart statistics

The S chart plots the variation of successive samples. The centre line is given as the average of the standard deviation for all segments:

$$CL = \bar{\sigma}$$

The upper and lower control limits are given as:

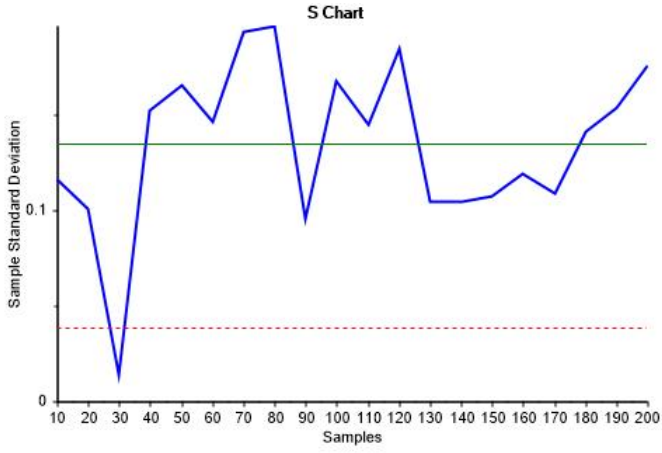
$$UCL = B_4 \cdot \bar{\sigma}$$

$$LCL = B_3 \cdot \bar{\sigma}$$

where B_4 and B_3 are unbiasing constants that depend on the window size.

S Chart - example

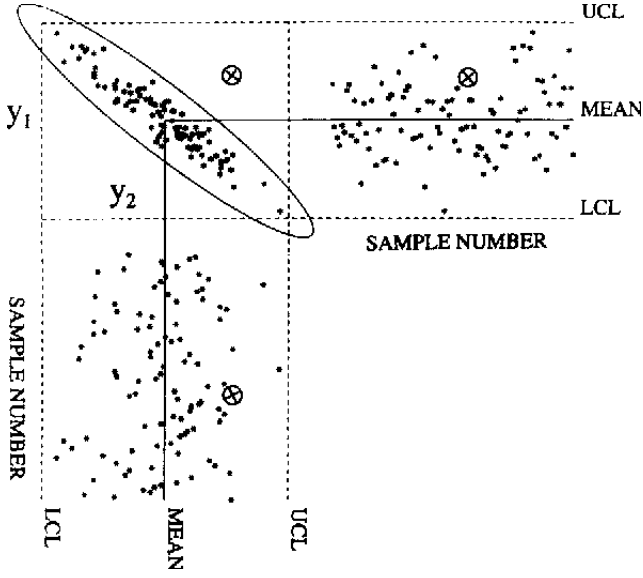
Segment size = 10



Extending SPC to more variables (MSPC)

- Consider two variables with individual critical limits for SPC
- Assume the confidence interval is set to 95% (significance level 0.05)
- Then if one applies this individually for two variables, the significance level will be $1 - 0.95^2 = 0.095$
- Thus one will not detect deviations from the Normal Operational Conditions (NOC) at the desired level

Illustration with two variables



Multivariate SPC

Any multivariate control procedure should fulfill these four conditions:

1. Is the process in control?
2. An overall probability of the Type I error must be specified (saying the process is out-of-control when it is not)
3. The relationships among the variables should be taken into account
4. If the process is out-of-control, what is the problem?

A multivariate normal distribution

Assume two variables that follow a multivariate normal distribution. The means are:

$$\overline{\mathbf{x}} = \frac{1}{m_x}\sum_{i=1}^{m_x}\mathbf{x}_i \ , \qquad \overline{\mathbf{y}} = \frac{1}{m_y}\sum_{i=1}^{m_y}\mathbf{y}_i$$

as the sample means, and

$$\hat{\mathbf{\Sigma}}_{\mathbf{x}} = \frac{1}{m_x - 1}\sum_{i=1}^{m_x}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})'$$

$$\hat{\mathbf{\Sigma}}_{\mathbf{y}} = \frac{1}{m_y - 1}\sum_{i=1}^{m_y}(\mathbf{y}_i - \overline{\mathbf{y}})(\mathbf{y}_i - \overline{\mathbf{y}})'$$

as the respective sample covariance matrices. Then the pooled covariance matrix is

$$\hat{\mathbf{\Sigma}} = \frac{(m_x - 1)\hat{\mathbf{\Sigma}}_{\mathbf{x}} + (m_y - 1)\hat{\mathbf{\Sigma}}_{\mathbf{y}}}{m_x + m_y - 2}$$

Hotelling's T-square distribution (T^2)

The Hotelling's T^2 statistic is a multivariate generalization of the Student t-test. The Hotelling's two-sample t-squared statistic is:

$$t^2 = \frac{m_x m_y}{m_x + m_y}(\overline{\mathbf{x}} - \overline{\mathbf{y}})' \hat{\mathbf{\Sigma}}^{-1} (\overline{\mathbf{x}} - \overline{\mathbf{y}}) \sim T^2(n, m_x + m_y - 2)$$

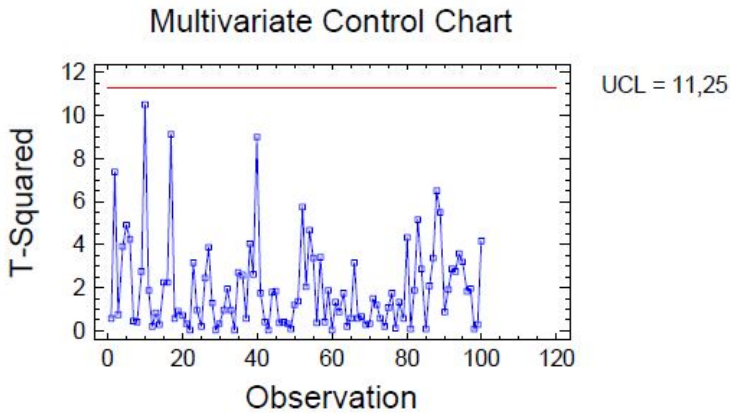
There exists a relationship between the Hotelling's T^2 and the F-distribution:

$$t^2 \sim T^2_{n,m-1} = \frac{n(m^2-1)}{m(m-n)} F_{n,m-n}$$

$m =$ number of samples, $n =$ number of variables

This is the expression used to estimate the critical limit

Illustration of a multivariate control chart

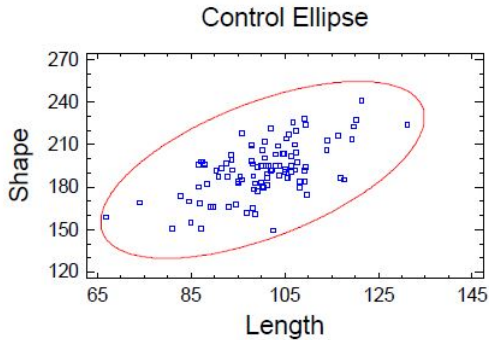


Representing the critical limit as a confidence ellipse

Given the covariance of $\mathbf{X}$, $\hat{\mathbf{\Sigma}}$, and the diagonal elements of $\hat{\mathbf{\Sigma}}$ representing the variance of the variables n; $Var(\mathbf{x}_n)$, the lengths of the axes of a confidence ellipse are:

$$Var(\mathbf{x}_n)\ \frac{2(m^2-1)}{m(m-2)}F_{p,2,m-2}$$

where *p* specifies the significance level, typically 95% or 99%



An incentive for multivariate analysis

- The ellipse (or hyper ellipse i 3D) is an efficient way of visualizing if the process is under control
- ... but what if there are 10 or 100 variables?
- Then the solution is to represent the individual variables as a weighted sum of the original ones
- This grouping and weighting of variables may be done "by hand" based on background knowledge
- ... or by use of multivariate analysis

Principal Component Analysis (PCA)

Principal Component Analysis (PCA)

PCA is a dimension reduction method which has many interesting applications

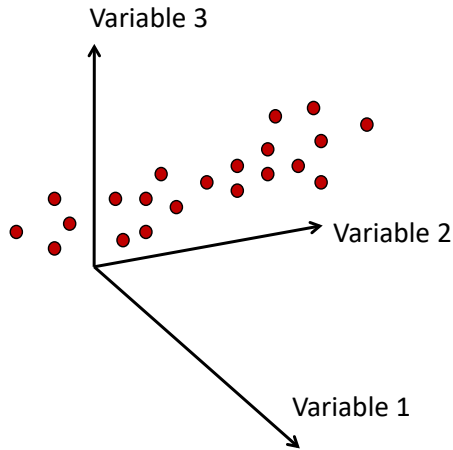
- Pattern recognition
- Dimensionality reduction
- Clustering and classification
- Condition monitoring and predictive maintenance, e.g. of wind turbines
- Outlier detection
- Denoising
- Data imputation
- Reduce model complexity
- Speed up training of machine learning models

The PCA approach

- Assume a data table $X_{m \times n}$ with variables $x_1, x_2, ..., x_n$, each variable sampled m times (and $n < m$), e.g. m time series each containing n variables in time
- Objective of PCA: If n is a large number, we would like to capture the essence of $x_1, x_2, ..., x_n$ by a smaller set of derived variables $\mathbf{t}_1$,.....,$\mathbf{t}_r$ (i.e. $r < n$)
- One criterion for estimating $\mathbf{t}_r$ is to find a line that maximizes the variance, i.e. to minimize the sum of square distances

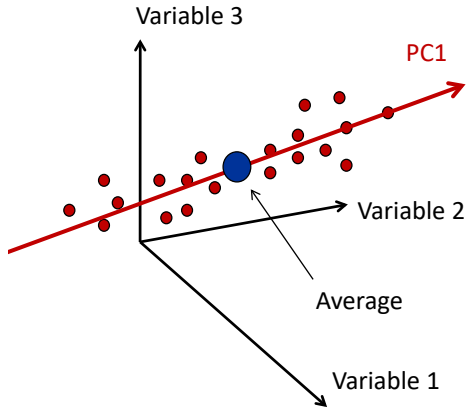
PCA by maximizing the variance - I

The PCA problem formulated as a maximization of the variance



PCA by maximizing the variance - II

The PCA problem formulated as a maximization of the variance



Singular Value Decomposition (SVD)

Theorem: For any matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$, there exist two orthogonal matrices $\mathbf{U} \in \mathbb{R}^{m \times m}, \mathbf{P} \in \mathbb{R}^{n \times n}$ and a nonnegative, diagonal matrix $\mathbf{\Sigma} \in \mathbb{R}^{m \times n}$ such that

$$\mathbf{X}_{m \times n} = \mathbf{U}_{m \times m} \mathbf{\Sigma}_{m \times n} \mathbf{P}^{\mathbf{T}}{}_{n \times n}$$

This is called the Singular Value Decomposition (SVD) of $\mathbf{X}$:

- The diagonals of $\mathbf{\Sigma}$ are called the singular values of $\mathbf{X}$ (often sorted in decreasing order)
- The columns of $\mathbf{P}$ are called the right singular vectors of $\mathbf{X}$
- The columns of $\mathbf{U}$ are called the left singular vectors of $\mathbf{X}$
- Let $\mathbf{T} = \mathbf{U}_{m \times m} \mathbf{\Sigma}_{m \times n} \rightarrow \mathbf{X}_{m \times n} = \mathbf{T}_{m \times n} \mathbf{P}^{\mathbf{T}}{}_{n \times n}$
- The columns in $\mathbf{T}$ are the weighted sums of the original variables (linear combinations), the weights are given in the matrix $\mathbf{P}$

PCA: A linear transformation

$$\mathbf{X}_{m\times n} = \mathbf{T}_{m\times n}\mathbf{P}^{\mathbf{T}}{}_{n\times n} \rightarrow \mathbf{T}_{m\times n} = \mathbf{X}_{m\times n}\mathbf{P}_{n\times n}/(\mathbf{P}^{\mathbf{T}}{}_{n\times n}\mathbf{P}_{n\times n})$$

The columns in $\mathbf{P}$ are orthonormal, thus $\mathbf{P}^{\mathbf{T}}\mathbf{P} = \mathbf{I}$

$$\mathbf{t_1} = p_{11}\mathbf{x_1} + p_{12}\mathbf{x_2} + \ldots + p_{1n}\mathbf{x_n}$$
$$\mathbf{t_2} = p_{21}\mathbf{x_1} + p_{22}\mathbf{x_2} + \ldots + p_{2n}\mathbf{x_n}$$
$$\ldots$$
$$\ldots$$
$$\mathbf{t_n} = p_{n1}\mathbf{x_1} + p_{n2}\mathbf{x_2} + \ldots + p_{nn}\mathbf{x_n}$$

$$\begin{bmatrix}\mathbf{t_1}\\ \mathbf{t_2}\\ \vdots\\ \mathbf{t_n}\end{bmatrix} = \begin{bmatrix}p_{1,1} & p_{1,2} & \cdots & p_{1,n}\\ p_{2,1} & p_{2,2} & \cdots & p_{2,n}\\ \vdots & \vdots & \ddots & \vdots\\ p_{n,1} & p_{n,2} & \cdots & p_{n,n}\end{bmatrix}\begin{bmatrix}\mathbf{x_1}\\ \mathbf{x_2}\\ \vdots\\ \mathbf{x_n}\end{bmatrix}$$

Reducing the dimensionality without losing information

- When some of the variables are correlated, there is redundancy in the system, and one does not need to represent $\mathbf{X}$ by including all n dimensions (assuming $m < n$)
- The remaining part of $\mathbf{X}$ is then represented as an error matrix $\mathbf{E}$
- We name $\mathbf{T}$ and $\mathbf{P}$ for Scores and Loadings, respectively

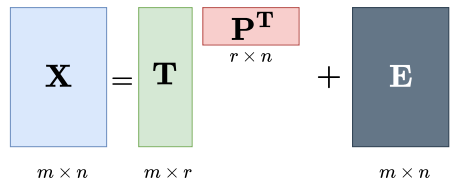
$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathbf{T}} + \mathbf{E}$

$\mathbf{X}$: Original data $(m \times n)$

$\mathbf{T}$: Scores $(m \times r)$

$\mathbf{P}$: Loadings $(r \times n)$

$\mathbf{E}$: Error $(m \times n)$

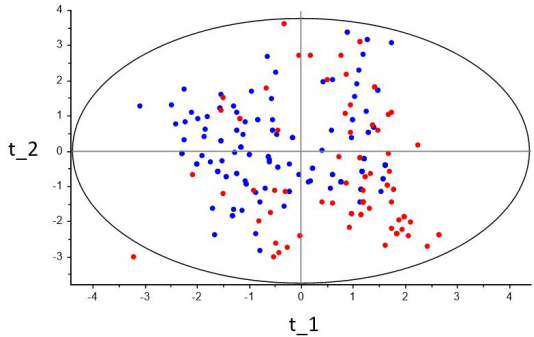


Projecting new samples onto an existing PCA model for MSPC

- Once a PCA model has been established for normal process conditions, new samples can be projected onto the model for real-time MSPC
- Note that the original variables x_n are now represented by their linear combinations t_r

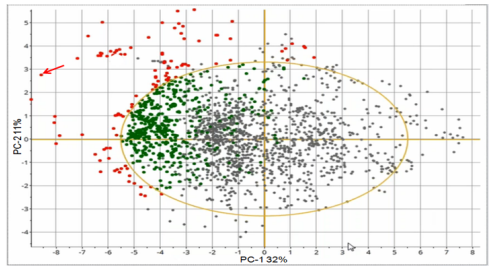
Blue: Model

Red: Projection of new samples



Example of projection: Environmental monitoring

- A PCA model was established with 80 sensors monitoring the conditions on the sea floor outside the Lofoten islands
- The model was based on data collected from April-July (in grey)
- Samples from August were projected onto this model (the ones marked red lie outside the critical limit)



Quality by Design and Process Analytical Technology

QbD - PAT introduction - I

- QbD: Quality by Design
- PAT: Process Analytical Technology

QbD - PAT introduction - II

- Pharmaceutical industries did not optimize/modernise their production e.g. by adopting state-of-the-art process monitoring and control strategies for production as many other industries did due to
  1. costs regarding documentation/certification of quality of products
  2. a perception that they could not utilize innovations that methodology/technology that other industries used for enhancing their production because they differ from other industries

QbD - PAT introduction III

- U.S. Food & Drug Administration encouraged in 2011 pharmaceutical and other industries to use innovations in their production by stating that all new submissions for approval of production processes and products must be based on the QbD approach
- Recent view on PAT and MSPC from FDA (hyperlink)

QbD - PAT - introduction IV

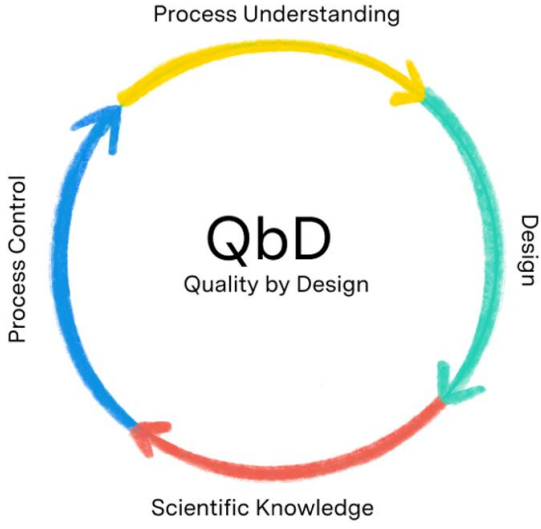
- QbD: Quality by Design
- PAT: Process Analytical Technology
- QbD and PAT complement each other and are methodologies for ensuring the production of goods and services that
  1. have the desired quality
  2. the planned cost
  3. the planned production time
  4. with controlled variation in the attributes specified above
  5. makes testing of finished products redundant
     real time release (testing)

What are the objectives of Quality by Design (QbD)

- *"Quality cannot be tested into products, it should be built in, or by design"*
- Ensure quality of all products not just the ones tested
- Three terms related to the QbD initiative

1. Critical Quality attributes (CQAs)
   A physical, chemical, biological or microbiological property or characteristic that should be within an appropriate limit, range, or distribution to ensure the desired product quality. Response variables in DOE
2. Critical process parameters (CPPs)
   A process parameter whose variability has an impact on a critical quality attribute and therefore should be monitored or controlled to ensure to process produces the desired quality. Identified by DOE
3. Quality target product profile (QTPP)
   "A prospective summary of the quality characteristics of a drug product that ideally will be achieved to ensure the desired quality, taking into account safety and efficacy of the drug product". The target product profile forms the basis of design for development of the product

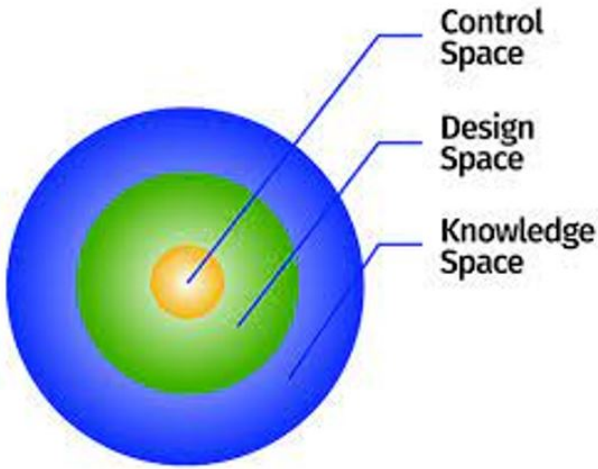
QbD and continuous improvement



Design space

- Definition: *The multidimensional combination and interaction of input variables and process parameters that have been demonstrated to provide assurance of quality*
- That is: The use of methods such as DOE, multivariate analysis and statistical process control that have established the effects and interactions of the CPPs such that the CQAs have been assured at the point of manufacture in real time.

Design space in the context of process control



Process Analytical Technology (PAT)

- Definition *"The Agency considers PAT to be a system for designing, analysing, and controlling manufacturing through timely measurements (i.e., during processing) of critical quality and performance attributes of raw and in-process materials and processes, with the goal of ensuring final product quality."*
- That is: A true PAT platform should provide timely measurements! PAT is Dynamic, Real Time and Process Based, with the capability of process correction and potential open or close loop control. The latter would be the ideal scenario. Close loop process controls are old school to process/control engineers in other industries, but pharma industry has a long way to go.
- The goal of PAT development and implementation should be aligned with the scope of a QbD plan. A PAT system is developed to measure critical process parameters and critical quality attributes, understand product and process variability, and thus control manufacturing processes to help achieve a predefined target product profile and/or bring robustness to the process.

PAT Tools

- The PAT toolbox
  1. Multivariate tools for design, data acquisition and analysis
  2. Process analyzers
  3. Process control tools
  4. Continuous improvement and knowledge management tools
- PAT is one of the many tools or enablers of QbD.

A successfully implemented PAT system should be able to:

1. Identify, understand and manage the sources of variability
2. Establish relationship between raw material, process parameters and final product quality attributes
3. Control raw material/processes to ensure CQAs as specified

Variability existing from raw materials, processes, intermediates to final product.

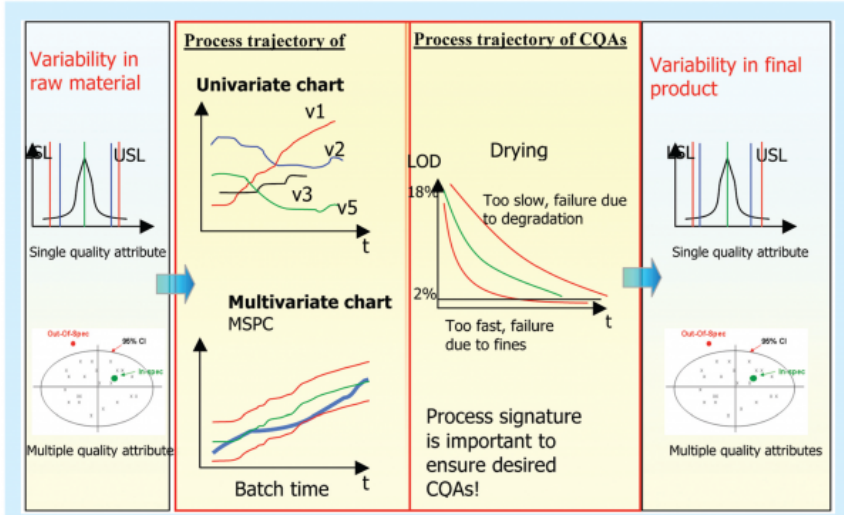
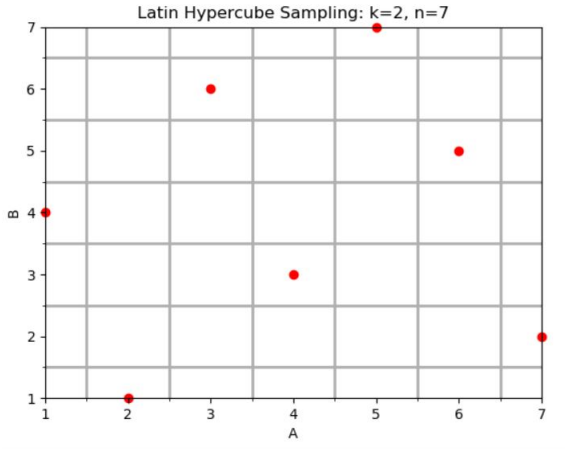


Figure 1 PAT to understand and manage raw material and process variability, and ensure final product quality

Latin hypercube designs

- Hypercube designs are model independent, space filling designs often used in computer experiments
- Each dimension space is cut into n sections where n is the number of sampling points → only one point is put in each section
- The number of experiments is n
- These designs are optimized by distance in order to fill out the factor space.
- Criteria for construction of the design:
  1. Center the points within the sampling intervals
  2. Maximize the minimum distance between points and place the point in a randomized location within its interval
  3. Maximize the minimum distance between points and center the point within its interval
  4. Preferably construct orthogonal designs, also among interaction terms

Latin hypercube design for 2 factors and 7 experiments



Paper on DoE and metamodelling

Syberfeldt, A., Grimm, H., Ng, A. (2008) Design of Experiments for Training Metamodels in Simulation-Based Optimisation of Manufacturing Systems.

- Applying metamodels in simulation-based optimisation of manufacturing systems
- Comparing results for two DoE designs and randomly sampled data
- Goal: Using Artificial Neural Networks (ANNs) to reduce the computational burden in optimization
- Procedure:
  1. construct an experimental plan of the different simulation input parameter settings that are to be tested using DoE
  2. simulate the system with the given input parameter values
  3. train the ANN using the simulated input-output samples
  4. optimize using an evolutionary algorithm

Application details

- Modelling objective: Optimise the schedule for the most efficient utilisation of the production
- Design factors: Lead time for five components
- Response variables: Utilisation, component shortage and total tardiness (delay)
- Evaluating the model results for these three cases:
  1. A 3^5 full factorial design
  2. A Latin hypercube design
  3. Randomly selected samples
- The objective function is specified from the three response variables:
  $\sum_{i \in C} = (w_s i_{shortage} + w_t i_{tardiness}) - w_u utilisation$
- The w's are user-defined weights for importance
- 1000 samples where randomly generated as a test set
- Evaluation criterion: Minimise the Mean Square Error (MSE)

Points of discussion

- Where the ANN models comparable in terms of complexity and architecture (e.g. number of hidden layers)

Points of discussion

- Where the ANN models comparable in terms of complexity and architecture (e.g. number of hidden layers)
- Comparison of MSE without the standard deviation of the optimisation objective is rather meaningless (are the values 0.0132, 0.0307, 0.0097 significantly different?)

Points of discussion

- Where the ANN models comparable in terms of complexity and architecture (e.g. number of hidden layers)
- Comparison of MSE without the standard deviation of the optimisation objective is rather meaningless (are the values 0.0132, 0.0307, 0.0097 significantly different?)
- For the two designs, no ANOVA results are reported

Points of discussion

- Where the ANN models comparable in terms of complexity and architecture (e.g. number of hidden layers)
- Comparison of MSE without the standard deviation of the optimisation objective is rather meaningless (are the values 0.0132, 0.0307, 0.0097 significantly different?)
- For the two designs, no ANOVA results are reported
- No interpretation of the effect of the design factors

Points of discussion

- Where the ANN models comparable in terms of complexity and architecture (e.g. number of hidden layers)
- Comparison of MSE without the standard deviation of the optimisation objective is rather meaningless (are the values 0.0132, 0.0307, 0.0097 significantly different?)
- For the two designs, no ANOVA results are reported
- No interpretation of the effect of the design factors
- Also, no Predicted vs. Actual plot or other diagnostics, e.g. residuals

Points of discussion

- Where the ANN models comparable in terms of complexity and architecture (e.g. number of hidden layers)
- Comparison of MSE without the standard deviation of the optimisation objective is rather meaningless (are the values 0.0132, 0.0307, 0.0097 significantly different?)
- For the two designs, no ANOVA results are reported
- No interpretation of the effect of the design factors
- Also, no Predicted vs. Actual plot or other diagnostics, e.g. residuals
- No response surface or contour plots

Points of discussion

- Where the ANN models comparable in terms of complexity and architecture (e.g. number of hidden layers)
- Comparison of MSE without the standard deviation of the optimisation objective is rather meaningless (are the values 0.0132, 0.0307, 0.0097 significantly different?)
- For the two designs, no ANOVA results are reported
- No interpretation of the effect of the design factors
- Also, no Predicted vs. Actual plot or other diagnostics, e.g. residuals
- No response surface or contour plots
- The optimisation could have been performed directly from the DoE model, also including constraints