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

Big Data Cybernetics Gang

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

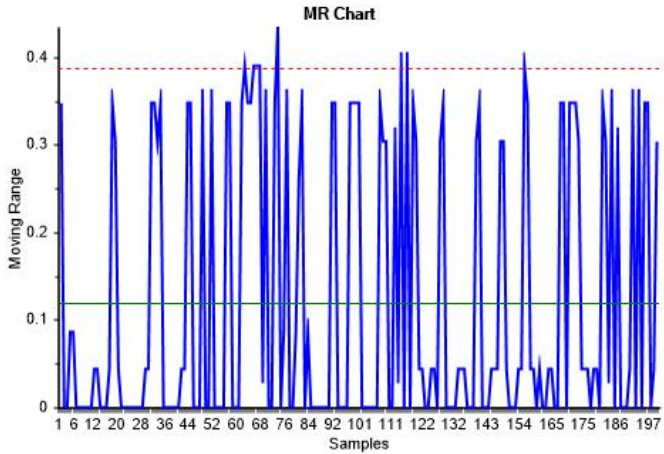
The upper and lower control limits for the moving range are given as:

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

$$LCL = 0$$

D_4 is an unbiasing constant equal to 3.267.

Moving Range - example



Plotting individual values with limits: I Chart

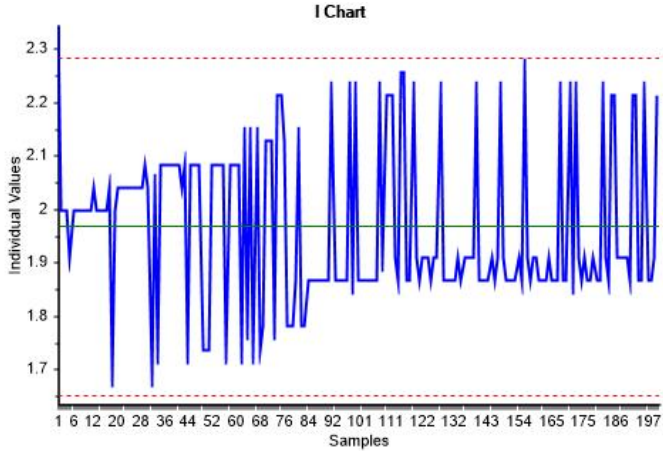
Calculations:

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

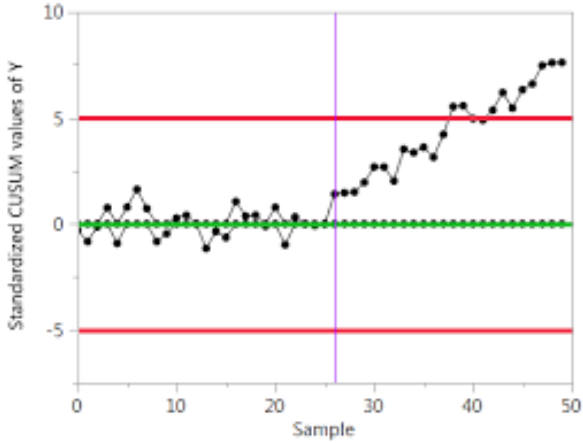


CuSum

- 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:

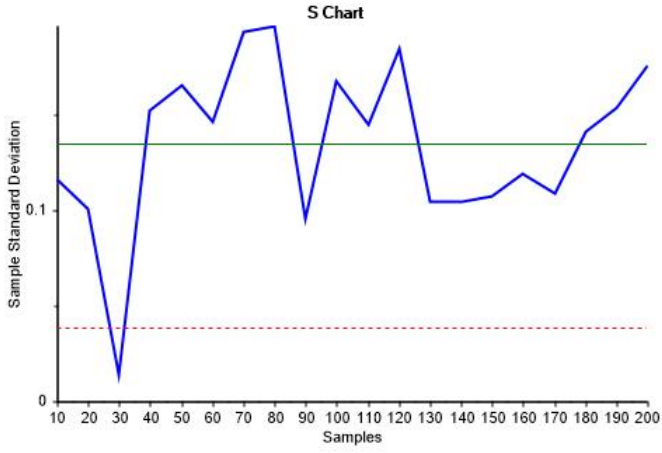
$$UCL = B_4 \cdot \bar{\sigma}$$

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

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

S Chart - example

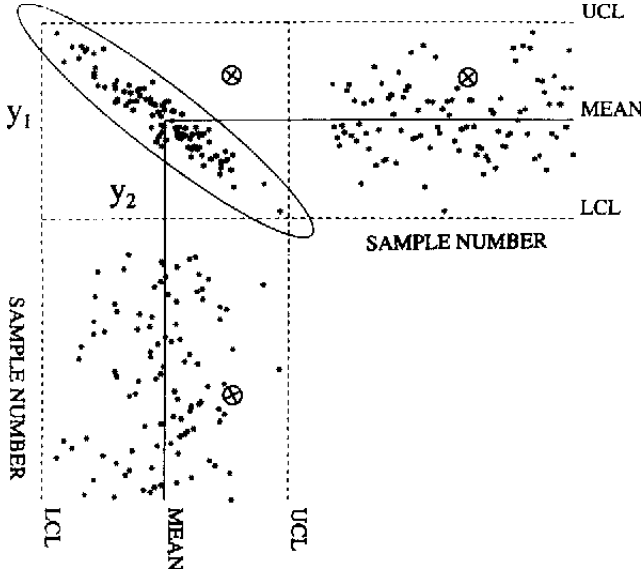
Segment size = 10



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:

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

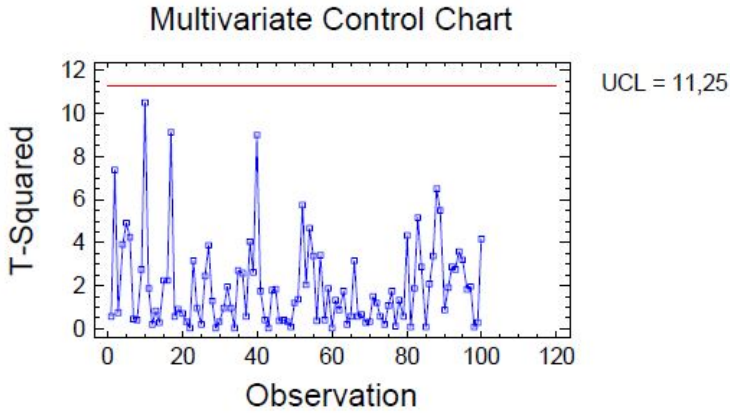
There exists a relationship between the Hotelling's T^2 and the F-distribution:

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

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

This is the expression used to estimate the critical limit

Illustration of a multivariate control chart

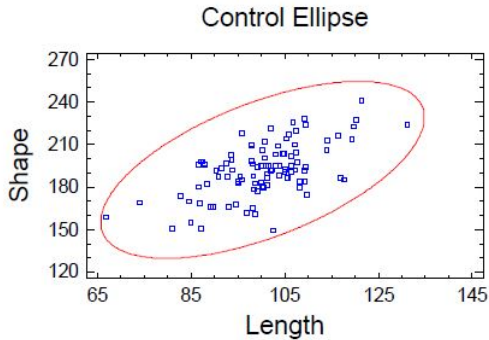


Representing the critical limit as a confidence ellipse

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

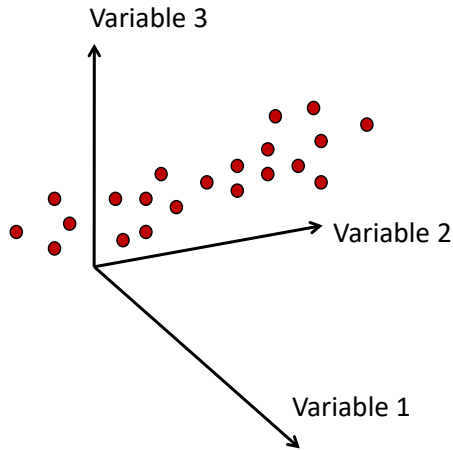
- Pattern recognition
- Dimensionality reduction
- Clustering and classification
- Condition monitoring and predictive maintenance, e.g. of wind turbines
- Outlier detection
- Denoising
- Data imputation
- Reduce model complexity
- Speed up training of machine learning models

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

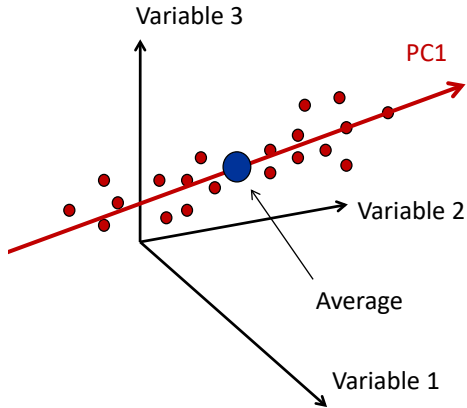
PCA by maximizing the variance - I

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

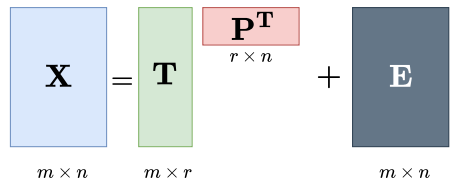
$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathbf{T}} + \mathbf{E}$

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

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

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

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

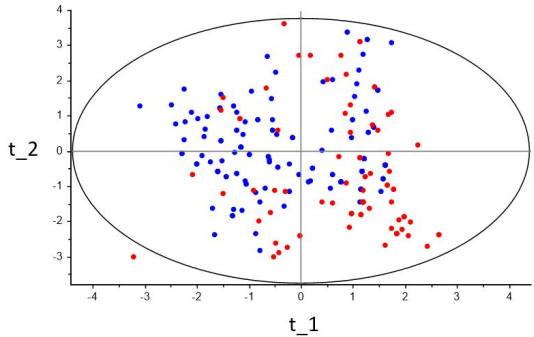


Projecting new samples onto an existing PCA model for MSPC

- 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

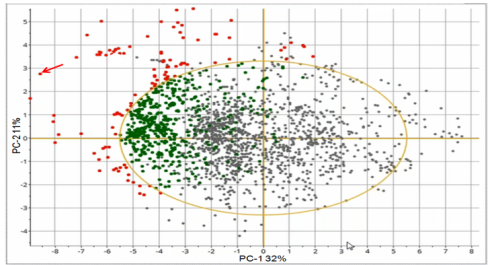
Blue: Model

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

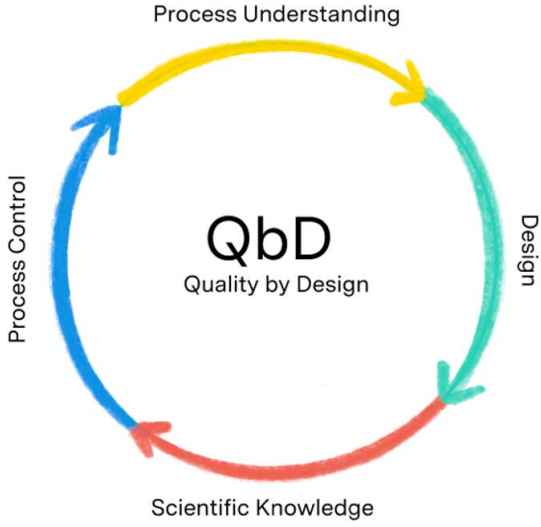
- QbD: Quality by Design
- PAT: Process Analytical Technology
- QbD and PAT complement each other and are methodologies for ensuring the production of goods and services that
  1. have the desired quality
  2. the planned cost
  3. the planned production time
  4. with controlled variation in the attributes 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

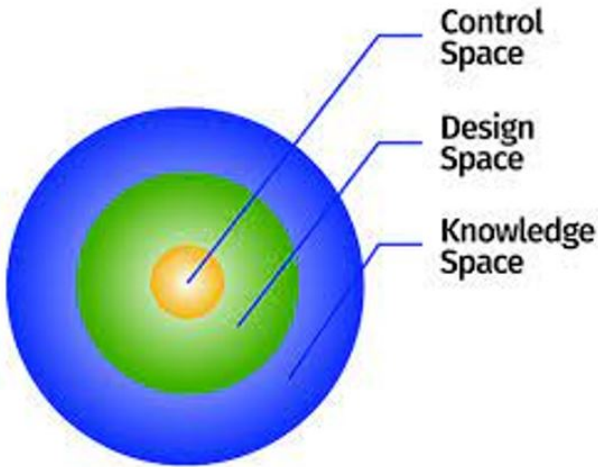
QbD and continuous improvement



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.

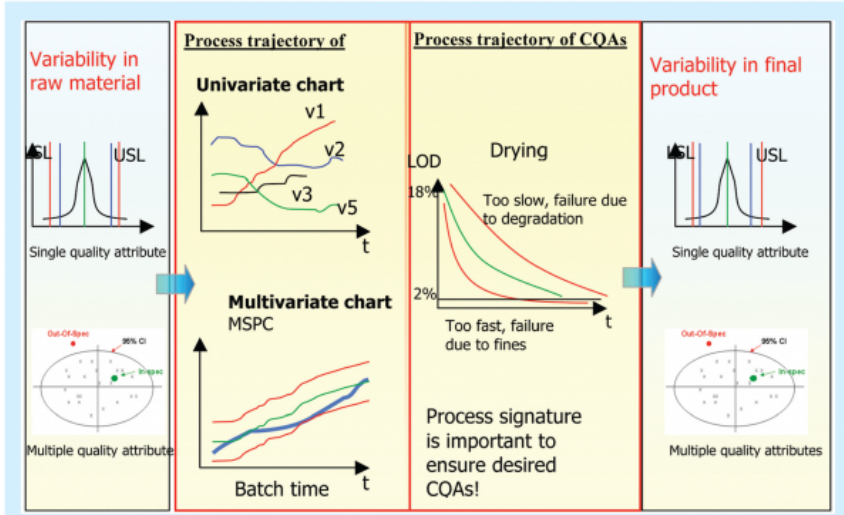


Figure 1 PAT to understand and manage raw material and process variability, and ensure final product quality