Analytical Chemistry, and Chemistry in General requires Constancy and Consistency

We as scientist are governed by immutable Laws.

1. Laws of Nature

2. Laws of Logic

1

Laws of Nature

- General:
  - Gravity
  - Electromagnetism
  - Strong Nuclear Forces
  - Weak Nuclear Forces
  - Quantum Mechanics
  - Relativity

2

Laws of Nature

Specific:

Symbol	Definitions
I	Quantity of radiation
I_o	Quantity of incident radiation
I_R	Quantity of reflected radiation
I_T	Quantity of transmitted radiation
I_A	Quantity of absorbed radiation
n_1	Index of refraction of the surrounding medium
n_2	Index of refraction of the object
θ_i	Angle of incidence
θ_r	Angle of transmitted ray
K	Absorption coefficient as used in Beer-Lambert's Law
L	Distance a light ray is propogated in a medium
$R\perp$	Perpendicular component of reflected light
$R_{\parallel}$	Parallel component of reflected light

3

Laws of Logic

- Identity
- Law on Non-Contradiction

4

Identity

Definitions: Object Identity.

$$x = y \ =_{df}$$

$$(O!x \ \& \ O!y \ \& \ \Box\forall F(Fx \equiv Fy)) \ \vee \ (A!x \ \& \ A!y \ \& \ \Box\forall F(xF \equiv yF))$$

Definitions: Relation Identity.

.1) $F = G \ =_{df} \ \Box\forall x(xF \equiv xG)$

.2) $F^n = G^n \ =_{df}$ (where $n > 1$)

$$\forall x_1 \ldots \forall x_{n-1}([\lambda y \ F^n y x_1 \ldots x_{n-1}] = [\lambda y \ G^n y x_1 \ldots x_{n-1}] \ \& \ [\lambda y \ F^n x_1 y x_2 \ldots x_{n-1}] = [\lambda y \ G^n x_1 y x_2 \ldots x_{n-1}] \ \& \ \ldots \ \& \ [\lambda y \ F^n x_1 \ldots x_{n-1} y] = [\lambda y \ G^n x_1 \ldots x_{n-1} y])$$

.3) $p = q \ =_{df} \ [\lambda y \ p] = [\lambda y \ q]$

Axioms: Logic of Identity. Where α, β are both individual variables, or both n-place relation variables ($n \geq 0$), and $\phi(\alpha, \beta)$ is the result of substituting one or more occurrences of β for α in $\phi(\alpha, \alpha)$ (provided β is substitutable for α in the occurrences of α it replaces), the modal closures of the following are logical axioms of our system:

$$\alpha = \beta \rightarrow [\phi(\alpha, \alpha) \equiv \phi(\alpha, \beta)]$$

Aristotle's Law of Identity

Everything that exists has a specific nature. ... Each entity exists as something specific, its identity is particular, and it cannot exist as something else. An entity can have more than one characteristic, but any characteristic it has is a part of its identity.

5

Law of Non-Contradiction

Aristotle's contribution

The traditional source of the law of non-contradiction is Aristotle's *Metaphysics* where he gives three different versions.

1.ontological: "It is impossible that the same thing belong and not belong to the same thing at the same time and in the same respect." (1005b19-20)

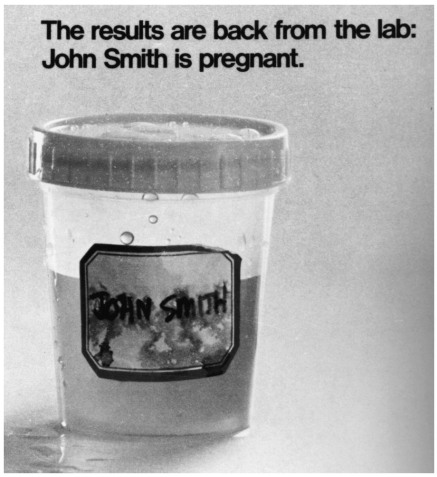
2.psychological: "No one can believe that the same thing can (at the same time) be and not be." (1005b23-24)

3.logical (aka the medieval Lex Contradictoriarum): "The most certain of all basic principles is that contradictory propositions are not true simultaneously." (1011b13-14)

6

7

The importance of reliable data!



Exploring Chemical Analysis, 3th Ed., D.C. Harris, WH Freeman, NY, 2005

8

8

Error in chemical analyses

- Mean (average, arithmetic mean)
  - Serves as the central "best" value for a set of replicate measurements
  - For a set of measurements, x_i, where i = 1,2,3…

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

$\bar{x}$ = mean

N = number of measurements

- Median

Is the middle result when the data are arranged in order of size

- Ideally, Mean = Median, but this is rarely true

9

9

Definitions

- Replicates:
  - Samples from the same system that are carried through an analysis in an identical manner
  - Forensic samples (e.g. blood) often taken in duplicates (think Olympics)
- Error:
  - Deviation of a result from the accepted true value
- Precision:
  - Deviation of a result from the average of a set of results:

$$d_i = |x_i - \bar{x}|$$

d_i = deviation of result i

x_i = result i

$\bar{x}$ = mean

- Accuracy:
  - Deviation of a result from the accepted true value

$$E_i = x_i - x_t$$

E_i = error of result i

x_i = result i

x_t = "true" value

- Error can be determined in a single measurement; precision cannot

10

10

Definitions

- Absolute error: has units & sign (±)

$$E_i = x_i - x_t$$

E_i = error of result i

x_i = result i

x_t = "true" value

- Relative error: error expressed as a percent & has a sign (±)

$$E_r = \frac{x_i - x_t}{x_t} \times 100\% = \text{percent error}$$

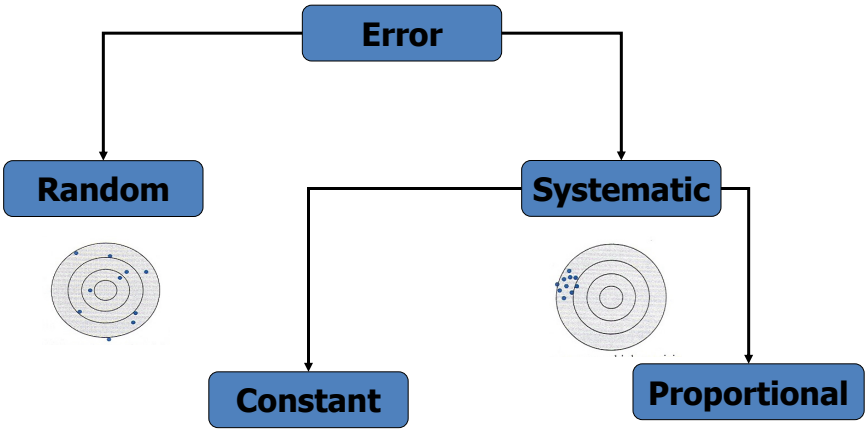
- Relative error can be expressed in ppt or ppm:

$$E_r = \frac{x_i - x_t}{x_t} \times 10^6 = \text{ppm error}$$

11

11

Types of errors



E.g. balance reads 0.03 g too heavy

1.00 g measured = 0.97 g actual (+3% error)

100.00 g measured = 99.97 g actual (0.03 % error)

E.g. balance always -0.8% off

1.00 g measured = 1.008 g actual

100.00 g measured = 100.8 g actual

12

12

Systematic error

- Has definite value, assignable cause, and affects all replicates similarly
- Origins of systematic errors:
  1) Instrumental errors
     - faulty calibration
     - use outside appropriate range
     - instabilities in components
     - etc.
  2) Method error
     - unrepresentative sampling
     - incomplete chemistry
     - impure/unstable reagents

13

13

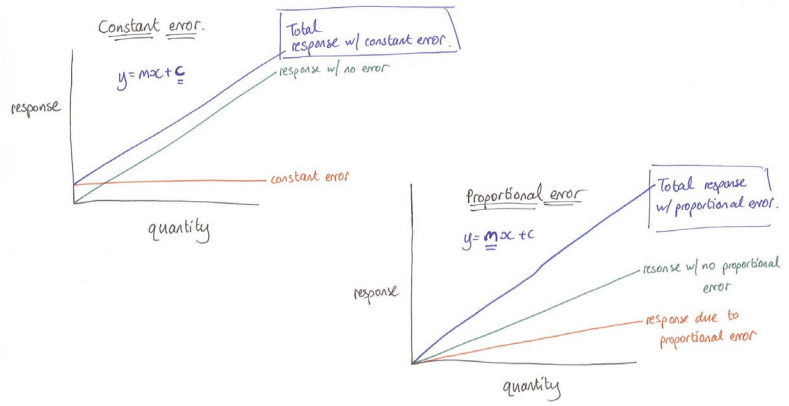
Systematic error

- Has definite value, assignable cause, and affects all replicates similarly
- Origins of systematic errors:
  3) Operator errors
     - Misreading dials, scales, menisci
     - Poor control of conditions (temp, pH etc.)
     - Poor technique (e.g. blowout pipeting)
     - etc.

14

14

Constant v. proportional errors

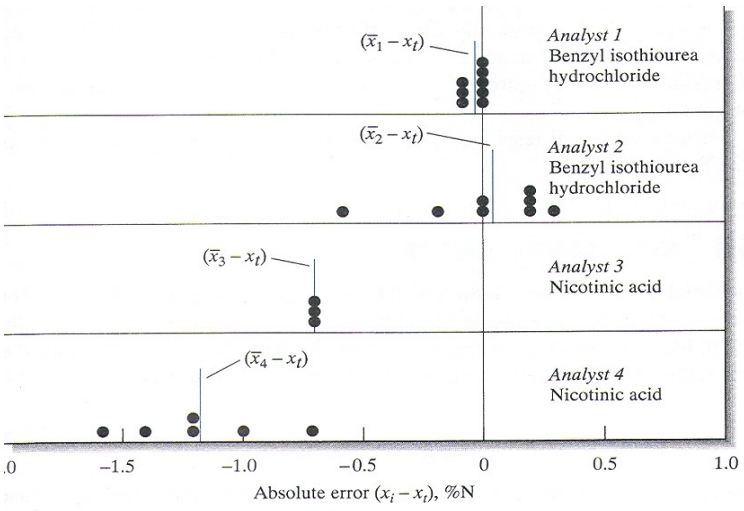


15

15

Accuracy and Precision

Analysis of nitrogen in two compounds by two different methods



16

16

Samples & populations

$\mu = \text{mean value of the population (system)}$

$\bar{x} = \text{estimate of } \mu, \text{because we only take small samples from the system}$

$\text{As } n_{sample} \to N_{population},\ \bar{x} \to \mu$

17

Samples & populations

Precision of a population of data is described by population standard deviation, σ:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)}{N}}$$

Precision of a sample of data is described by sample standard deviation, s:

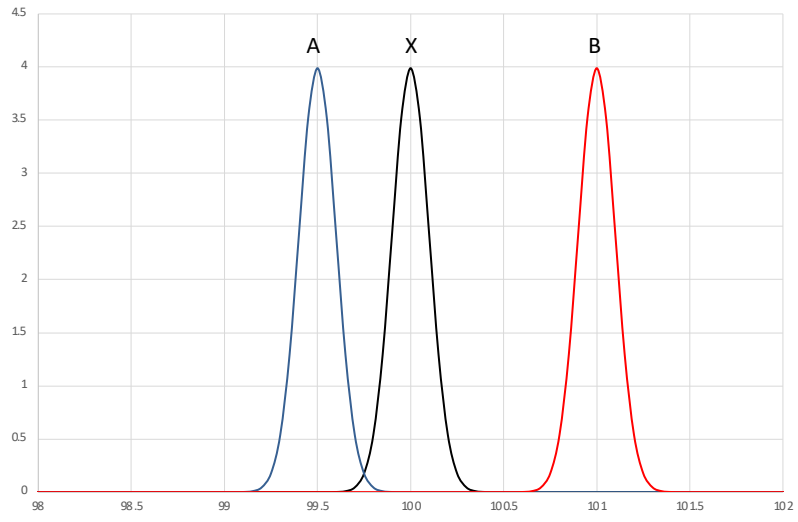
$$s = \sqrt{\frac{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{N}\left(d_i\right)^2}{n-1}}$$

$\text{As } n_{sample} \to N_{population},\ s \to \sigma \text{ just as } \bar{x} \to \mu$

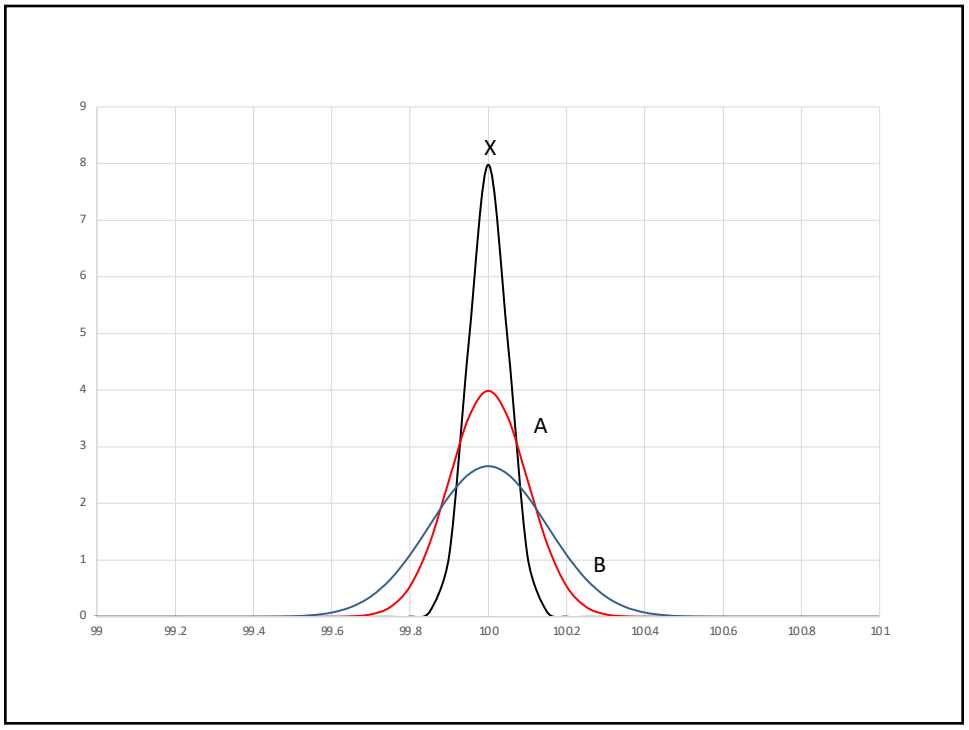
18

Description	Equation	Variable
Arithmetic Mean	$x = \sum_i \frac{x_i}{n}$	$x_i = the\ ith\ value$ $n = total\ number\ of\ values$
Degrees of Freedom	$v = (n-1)$	
Variance	$V(x) = \frac{\sum_i (x_i - x)^2}{v}$	
Standard Deviation	$s = \sqrt{V(x)} = \sqrt{\frac{\sum_i (x_i - x)^2}{v}}$	
Normal Distribution	$y(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$	$\mu = population\ mean$ $\sigma = population\ standard\ deviation$
Confidence Limits	$\mu = x \pm t(\frac{s}{\sqrt{n}})$	$t = confidence\ interval\ value$

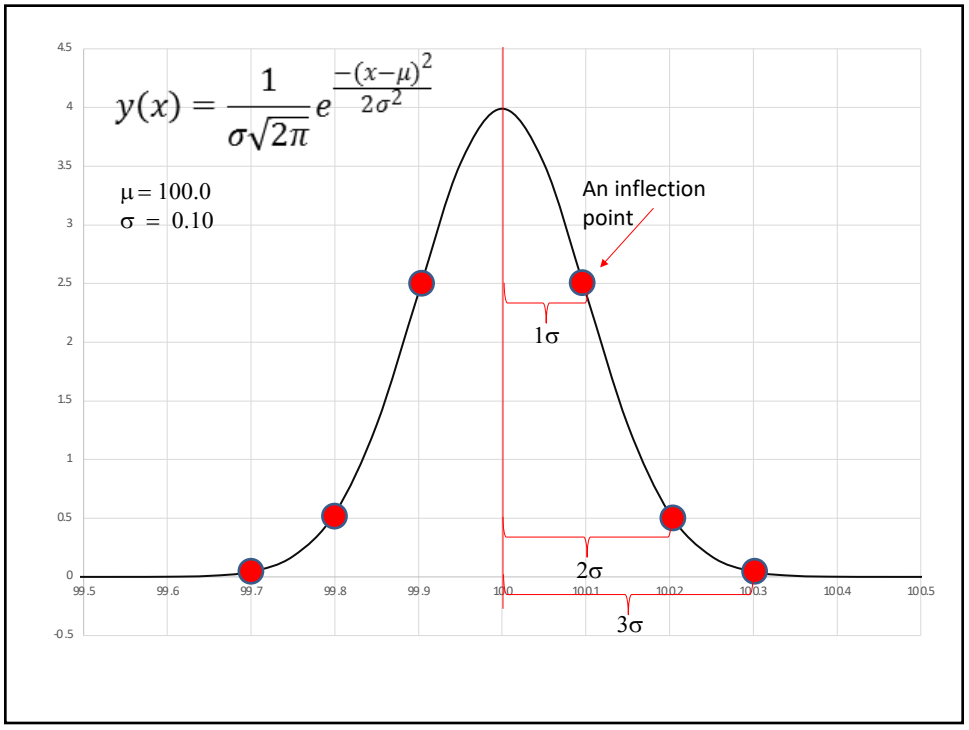
19



20


X
A
B
9
8
7
6
5
4
3
2
1
0
99
99.2
99.4
99.6
99.8
100
100.2
100.4
100.6
100.8
101

21


$y(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$
μ = 100.0
σ = 0.10
An inflection point
1σ
2σ
3σ
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
-0.5
99.5
99.6
99.7
99.8
99.9
100
100.1
100.2
100.3
100.4
100.5

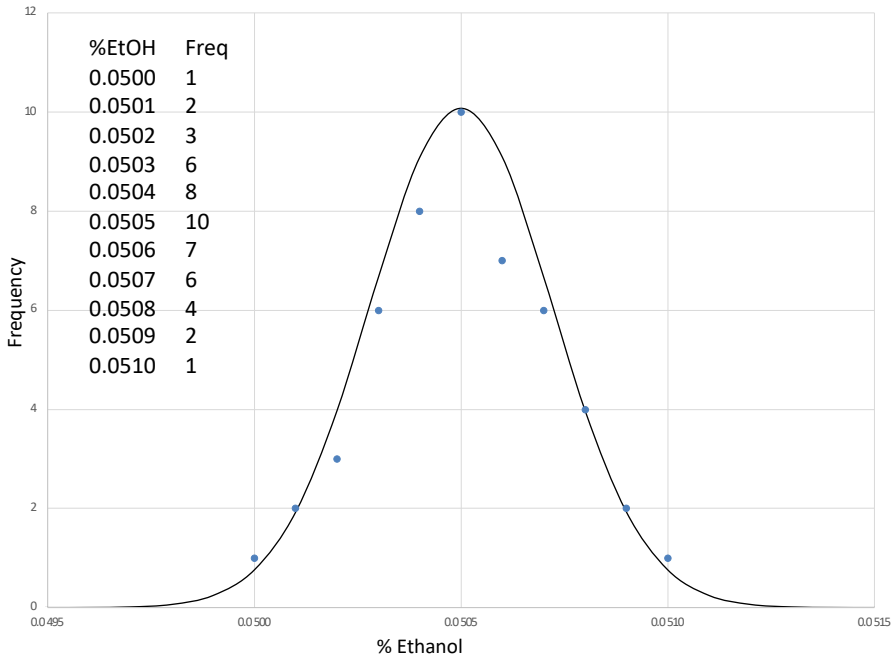
22

t values for confidence intervals

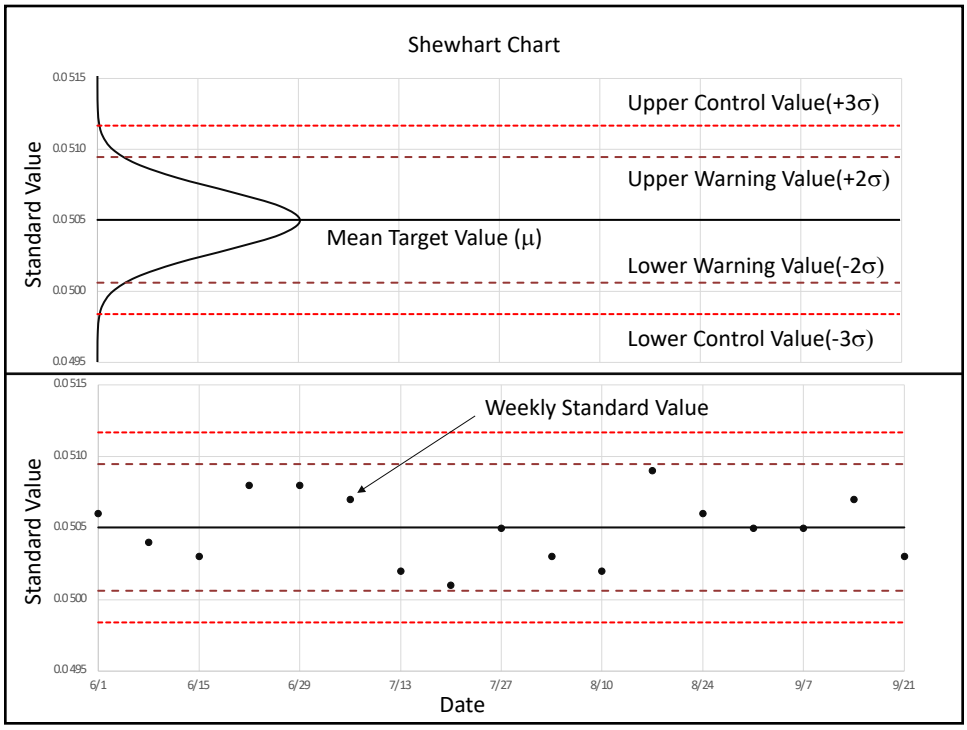
V	95%	99%
1	12.71	63.66
2	4.3	9.92
3	3.18	5.84
4	2.78	4.6
5	2.57	4.03
10	2.23	3.17
20	2.09	2.85
30	2.04	2.75
50	2.01	2.68
100	1.98	2.63

$$\mu = x \pm t(\frac{s}{\sqrt{n}})$$

23



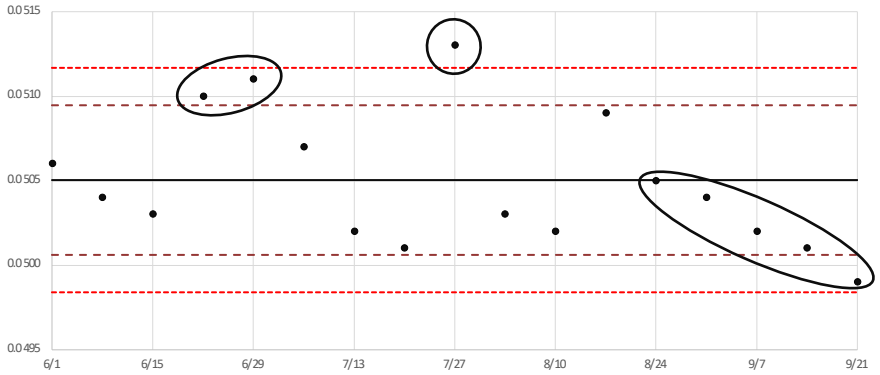
24



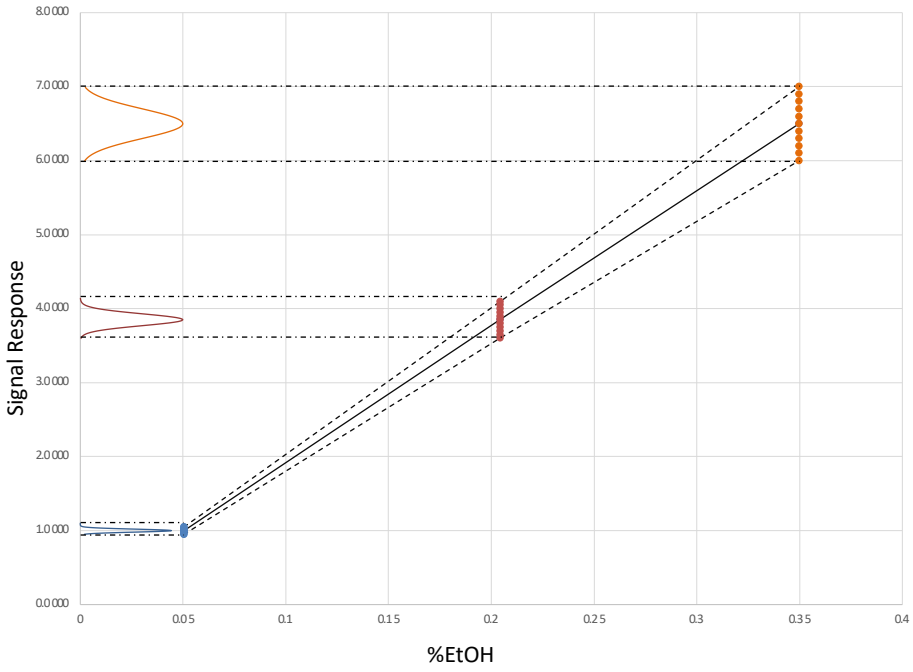
25

Run rules *are rules that are used to indicate out-of-statistical control or determinate trend situations.*

- a point lying beyond the upper and lower control limits
- 2 consecutive points lying beyond the upper or lower warning values (0.06% chance)
- 7 or more consecutive points lying on one side of the mean (0.8% chance of occurring)
- 6 consecutive points going in the same direction



26



27

Comparing two experimental means

- This can be used to test
  - 1) whether two samples are significantly different
  - 2) whether two methods are significantly different
- We still use the T-test (and the t-tables)

- Different samples and different methods might have different standard deviations
  - This has to be accounted for by pooling the deviations of the means

28

28

Comparing two experimental means

$$\bar{x}_a - \bar{x}_b = \pm t s_{pooled}\sqrt{\frac{N_1 + N_2}{N_1 N_2}}$$

$$\text{or}, t = \frac{\bar{x}_a - \bar{x}_b}{s_{pooled}\sqrt{\frac{N_1 + N_2}{N_1 N_2}}}$$

29

29

Comparing two experimental means

$$\bar{x}_a - \bar{x}_b = \pm t s_{pooled}\sqrt{\frac{N_1 + N_2}{N_1 N_2}}$$

$$\text{or}, t = \frac{\bar{x}_a - \bar{x}_b}{s_{pooled}\sqrt{\frac{N_1 + N_2}{N_1 N_2}}}$$

$$s_{pooled} = \sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1) + \ldots}{n_1 + n_2 \ldots - \text{number data sets}}}$$

If $t > t_{crit}$ then reject the hypothesis that the means are the same.

30

30

Data analysis and interpretation

- Data collected represented in a matrix

Variables going across in different columns

Objects going down in different rows	X-var 1	X-var 2	X-var 3	Y-var 1	Y-var 2
Sample 1					
Sample 2…					

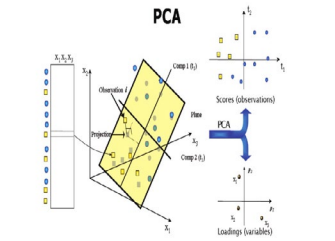
A propositional approach to describing and using metabolomics data (the x-data) for analyzing complex systems. These may have other specific properties (the y-data) which one may also wish to "explain" in terms of the x-data.

- Chemometric Approach
  - Principle Component Analysis (PCA)
  - Soft Independent Modeling of Class Analogy (SIMCA)
  - Partial Least-Squares (PLS) Method by Projections to Latent Structures
  - Orthogonal PLS (OPLS)
- Targeted Profiling

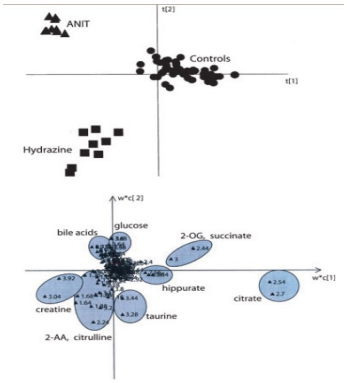
31

PCA

- Unsupervised
- Multivariate analysis based on projection methods
- Main tool used in chemometrics
- Extract and display the systematic variation in the data
- Each Principle Component (PC) is a linear combination of the original data parameters
- Each successive PC explains the maximum amount of variance possible, not accounted for by the previous PCs
- PCs Orthogonal to each other
- Conversion of original data leads to two matrices, known as scores and loadings
- The scores(T) represent a low-dimensional plane that closely approximates X. Linear combinations of the original variables. Each point represents a single sample spectrum.
- A loading plot/scatter plot(P) shows the influence (weight) of the individual X-variables in the model. Each point represents a different spectral intensity.
- The part of X that is not explained by the model forms the residuals(E)
- $X = TP^T = t_1p_1^T + t_2p_2^T + ... + E$



A principal component analysis (PCA) model approximates the variation in a data table by a low dimensional model plane. This model plane provides a score plot, where the relation among the observations or samples in the model plane is visualized, for example, if there are any groupings, trends, or outliers. The loading plot describes the influence of the variables in the model plane, and the relation among them. An important feature is that directions in the score plot correspond to directions in the loading plot, and vice versa.



32

SIMCA

- Supervised learning method based on PCA
- Construct a seperate PCA model for each known class of observations
- PCA models used to assign the class belonging to observations of unknown class origin
- Boundaries defined by 95% class interval
- Recommended for use in one class case or for classification if no interpretation is needed

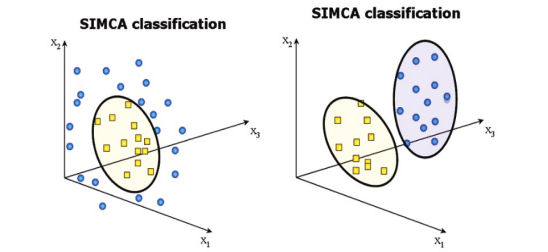


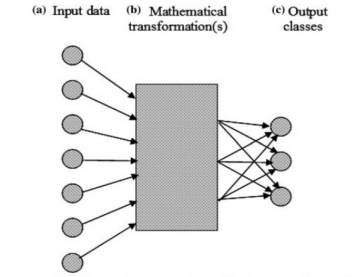
Illustration of SIMCA classification. In the left panel, the one-class classifier is shown, referred to as the asymmetric case. In the right panel, the SIMCA classification is shown with two classes, separately modeled by PCA.

- One-class problem: Only disease observations define a class; control samples are too heterogeneous, for example, due to other variations caused by diseases, gender, age, diet, lifestyle, etc.
- Two-class problem: Disease and control observations define two seperate classes

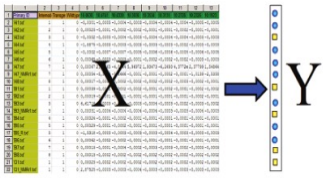
33

PLS

- Supervised learning method.
- Recommended for two-class cases instead of using SIMCA.
- Principles that of PCA. But in PLS, a second piece of information is used, namely, the labeled set of class identities.
- Two data tables considered namely X (input data from samples) and Y (containing qualitative values, such as class belonging, treatment of samples)
- The quantitive relationship between the two tables is sought.
- $X = TP^T + E$
- $Y = TC^T + E$
- The PLS algorithm maximizes the covariance between the X variables and the Y variables
- PLS models negatively affected by systematic variation in the X matrix not related to the Y matrix (not part of the joint correlation structure between X-Y.

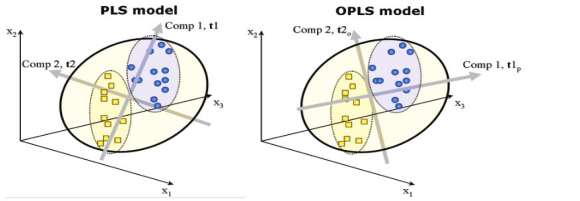


The class assignment problem. The inputs can be considered, and are referred to, as the "explanatory variables" or "x-data" whereas the functional or the other classes of interest, which are still variables associated with the samples, are referred to as "dependent variables" or "y-data" and are to be obtained as the outputs.



Class information can also be used to construct an additional matrix, hereinafter called the **Y** matrix, consisting of a discrete 'dummy' variable where [1]/[0] indicate the class belonging.

34



A geometrical illustration of the difference between the PLS-DA and OPLS-DA models. In the left panel, the PLS components cannot separate the between-class variation from the within-class variation, and the resulting PLS component loadings mixes both types of variations. In the right panel, the OPLS components are able to separate these two different variations. Component 1 ($t1_p$) is the predictive component and displays the between-class ((blue circles), (yellow squares)) variation of the samples. The corresponding loading profile can be used for identifying variables important for the class separation. Component 2 ($t2_o$) is the **Y**-orthogonal component and models the within group (within-class) variation.

- OPLS method is a recent modification of the PLS method to help overcome pitfalls
- Main idea to seperate systematic variation in X into two parts, one linearly related to Y and one unrelated (orthogonal).
- Comprises two modeled variations, the Y-predictive ($T_pP_p^T$) and the Y-orthogonal ($T_oP_o^T$) compononents.
- Only Y-predictive variation used for modeling of Y.
- $X = T_pP_p^T + T_oP_o^T + E$
- $Y = T_pC_p^T + F$
- E and F are the residual matrices of X and Y
- OPLS-DA compared to PLS-DA

35

Remarks on pattern classification

- Intent in using these classification techniques not to identify specific compound
- Classify in specific categories, conditions or disease status
- Traditional clinical chemistry depended on identifying and quantifying specific compounds
- Chemometric profiling interested in looking at all metabolites at once and making a phenotypic classification of diagnosis

36

Targeted profiling

- Targeted metabolomic profiling is fundamentally different than most chemometric approaches.
- In targeted metabolomic profiling the compounds in a given biofluid or tissue extract identified and quantified by comparing the spectrum of interest to a library of reference spectra of pure compounds.
- Key advantage: Does not require collection of identical sets = More amenable to human studies or studies that require less day-to-day monitoring.
- Disadvantage: Relatively limited size of most current spectral libraries = bias metabolite identification and interpretation.
- A growing trend towards combining the best features of both chemometric and targeted methods.

37

Databases

- Large amount of data
- Need for databases that can be easily searched
- Better databases will help in combining chemometric and targeted profiling methods
- Newly emerging databases
- HMDB good model for other databases
- Challenge of standardisation

38