sciensano

healthy all life long

HOMOGENEITY AND STABILITY TESTING IN EQA PROGRAMMES

Eqalm sympostium, Ljubljana, 8 October 2019

Wim Coucke – wim.Coucke@sciensano.be

.be

Alternatives for homogeneity and stability testing

Why do we need homogeneity and stability testing ?

To demonstrate that laboratory can, if it works well, obtain a value very close to the assigned value

If a laboratory does not obtain a value close to the assigned value, does this indicate a weak performance of the laboratory ?

True value deviates from sample to sample → homogeneity

True value increased or decreased between beginning and end of EQA round → stability

sciensano

.be

Homogeneity according to ISO 13528

$$s_s < 0.3\sigma_{pt}$$

What is σ_{pt} ? Why 0.3 ?

- ISO 13528: sample heterogeneity contributes less than 10% to the variance for performance evaluation
- Fearn&Thompson*: Z-scores do not increase by 5%
- It is NOT the standard deviation of the reported results
  It is the fixed, known-on-beforehand standard deviation that can be used to calculated Z-scores

  - If using fixed limits, like analytical performance specifications (APS):

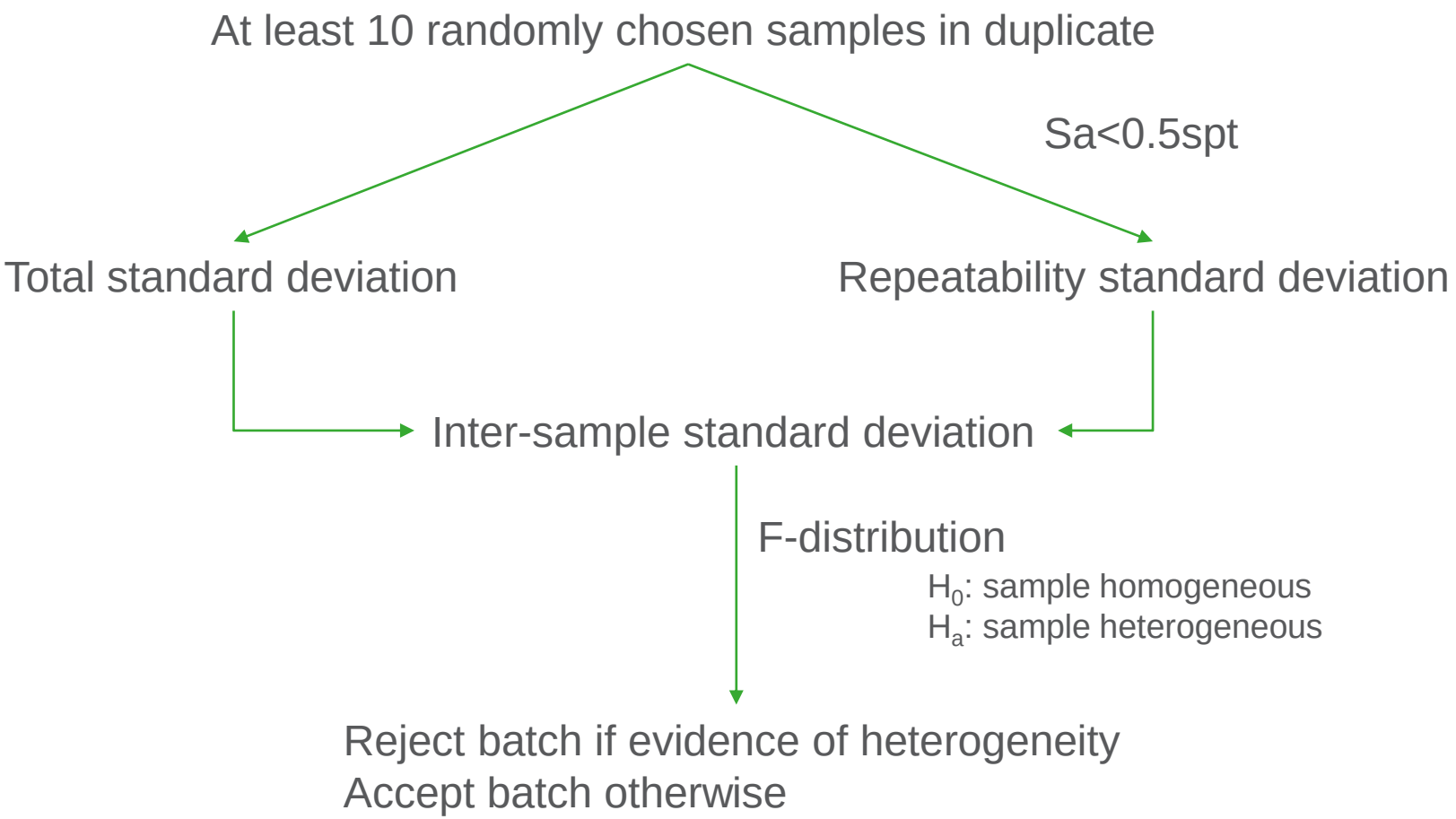
    σ_{pt}= assigned value*APS/2 or assigned value*APS/3

*FEARN, Tom; THOMPSON, Michael. A new test for 'sufficient homogeneity'. *Analyst*, 2001, 126.8: 1414-1417.

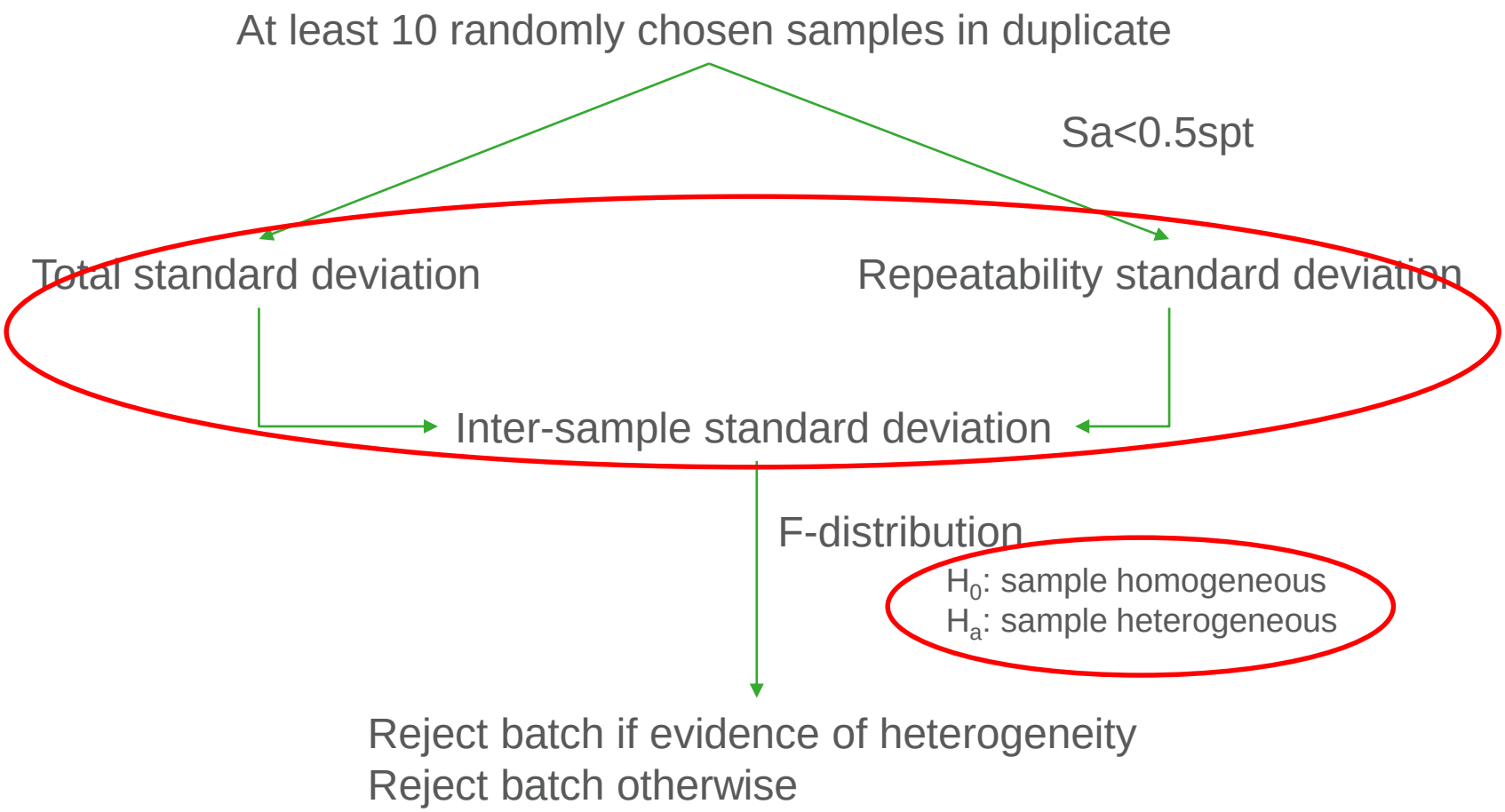
sciensano

.be

Homogeneity testing according to ISO 13528

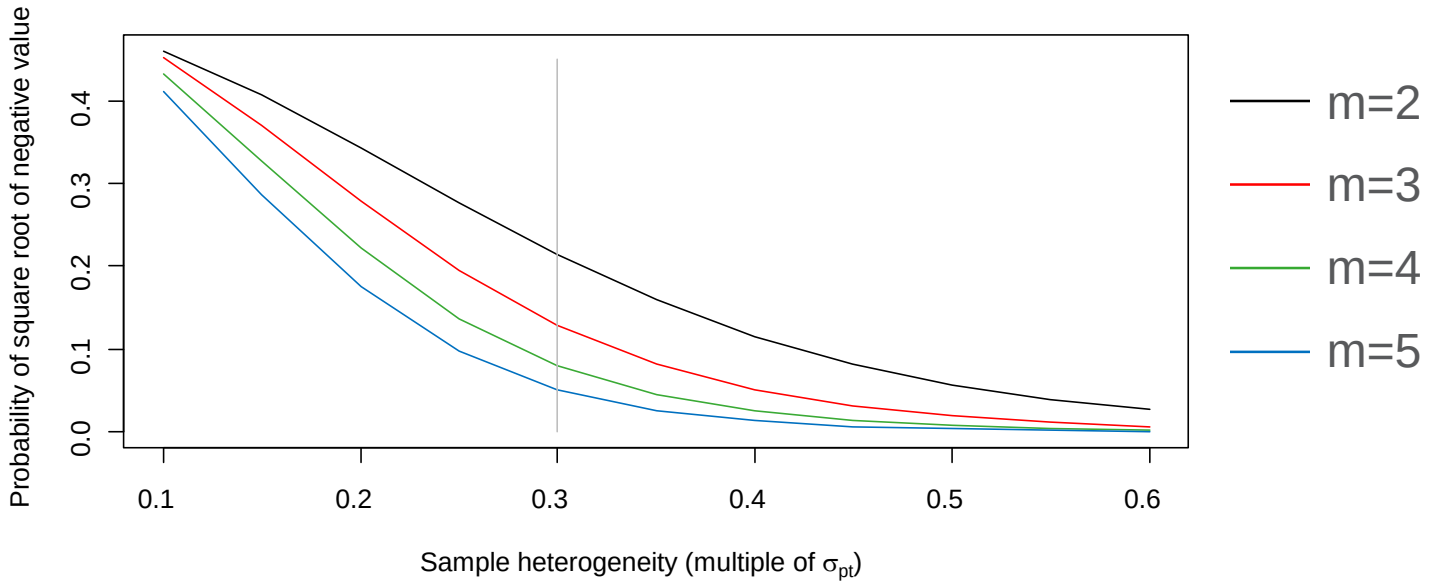


Homogeneity testing according to ISO 13528



Homogeneity testing according to ISO 13528

- Calculation of s_s involves square root of difference
- If difference is negative: $\sigma_s = 0$
  - Probability:



sciensano

.be

Homogeneity testing according to ISO 13528

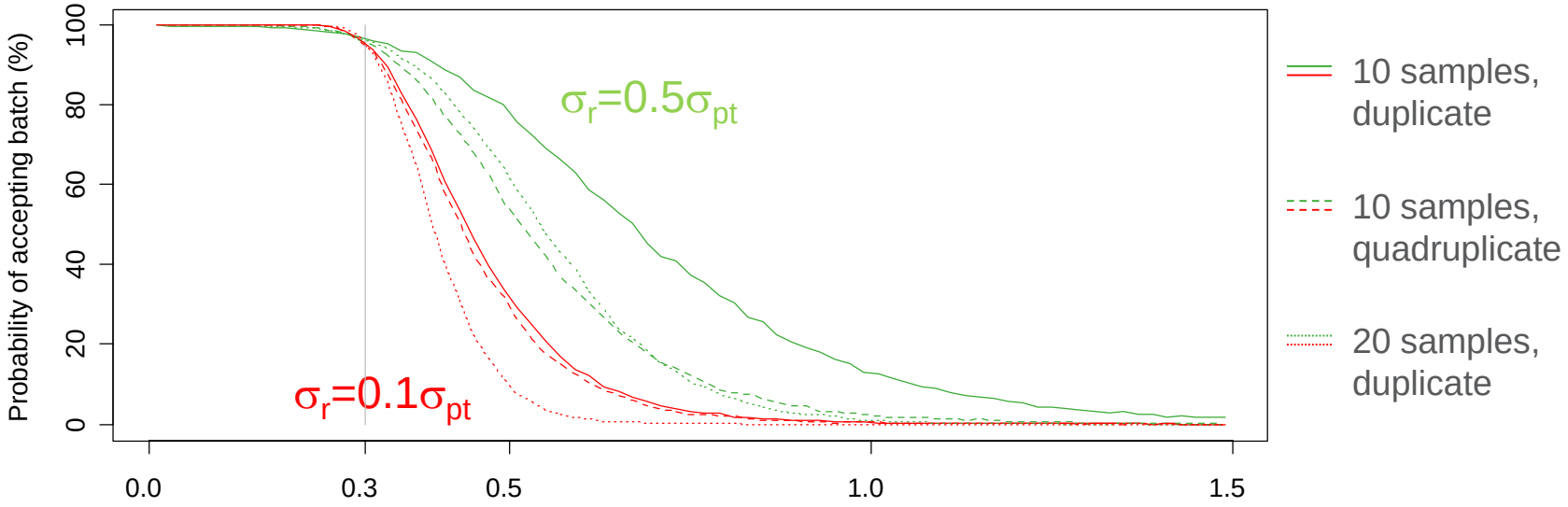
- Hypothesis testing:
- General:
  - state null hypothesis and alternative hypothesis.
  - Alternative hypothesis should be desired outcome.
  - Collect enough evidence to reject null hypothesis

- ISO 13528:
  - H_0: sample homogeneous --- H_a: sample heterogeneous

  - Only if there is enough evidence of heterogeneity, sample will be rejected
  - The higher the analytical variability, the higher the probability that sample will be accepted for homogeneity

Homogeneity testing according to ISO 13528

- Pervert situation:
- Try to have an analytical variability that is as high as possible, but still within limits and do not think of analyzing in triplicate or more:
  - Higher chance of forcing σ_s=0 by square root of negative value
  - Higher chance of accepting batch, even when true sample standard deviation does no meet limits



sciensano

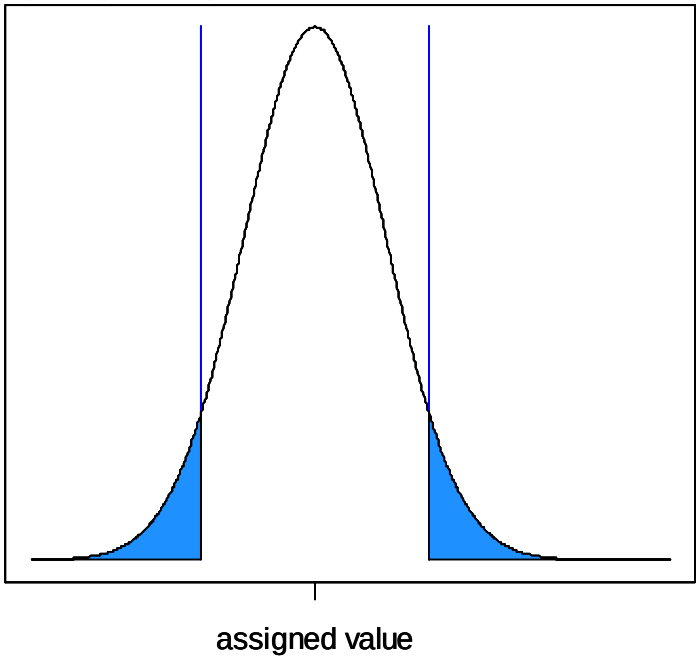
.be

Problems with sample homogeneity assessment according to ISO 13528

- Criterion only applicable in case of fixed-limits evaluation
- Drawbacks in estimating inter-sample standard deviation
  - Too high chance of forced to presume that $\sigma_s=0$
- Hypothesis test favors accepting batches of which homogeneity is doubtful
- Approach valid if using fixed limits-evaluation, samples are analysed with very high precision and at least 20 samples in duplicate

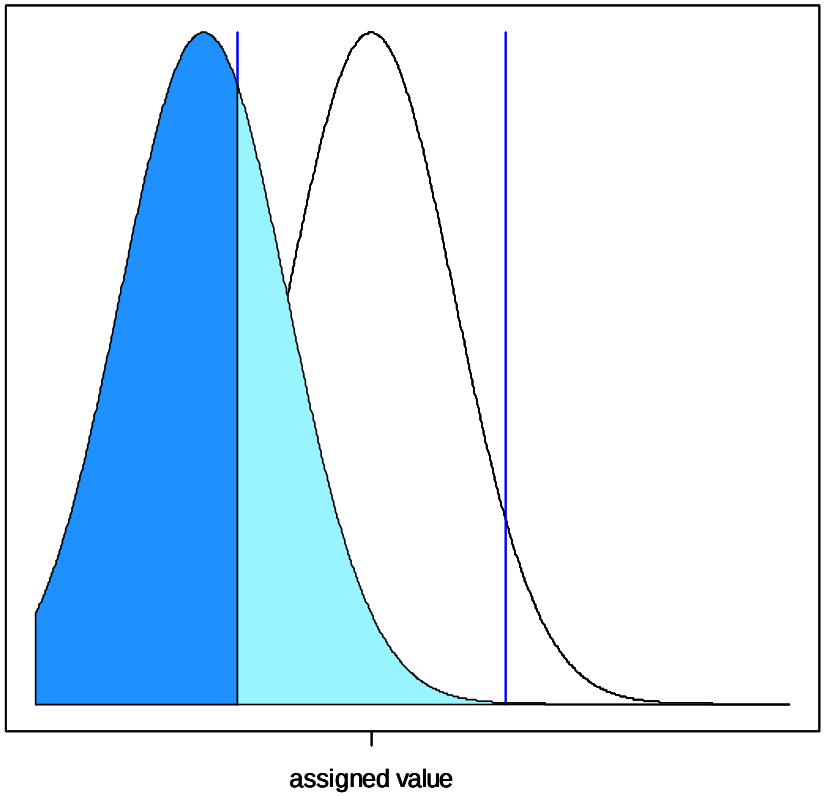
Some concepts (I)

- Flagging of laboratories: Indication of poor performance
  - |Z-score | > 2 or |Z-score | > 3
  - |Q-score| > Analytical Performance Specification (APS)
- Probability of flagging of good result



Some concepts (I)

- Flagging of laboratories: Indication of poor performance
  - |Z-score | > 2 or |Z-score | > 3
  - |Q-score| > Analytical Performance Specification (APS)
- Probability of not flagging of bad result

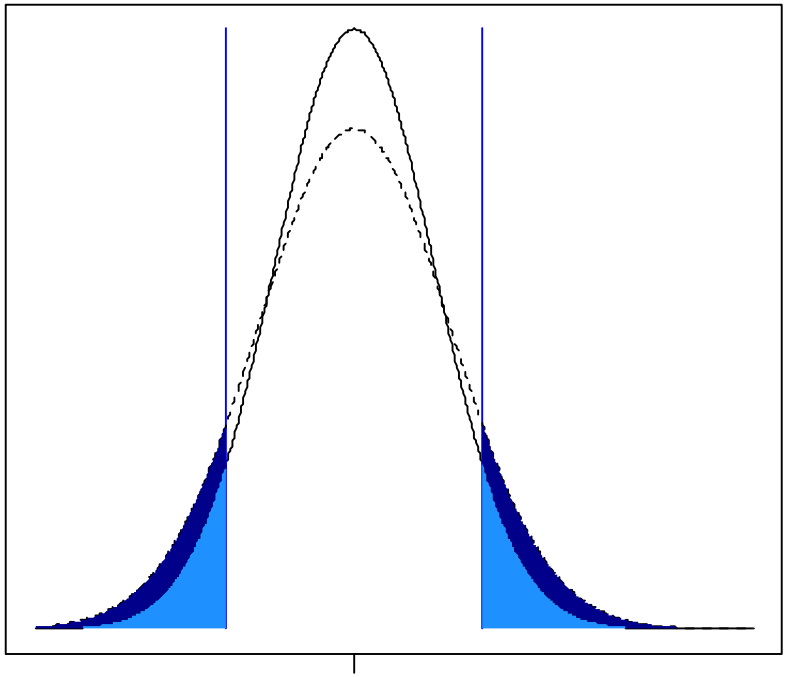


sciensano

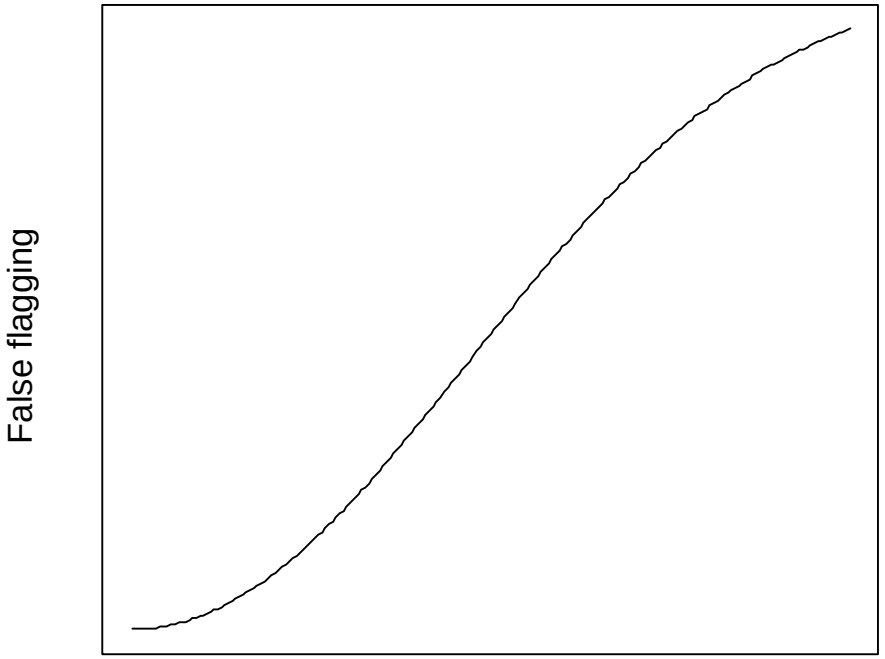
.be

Falsely flagging laboratories*

- Falsely flagging of results by anomaly in data
  - Fixed limits evaluation:



assigned value

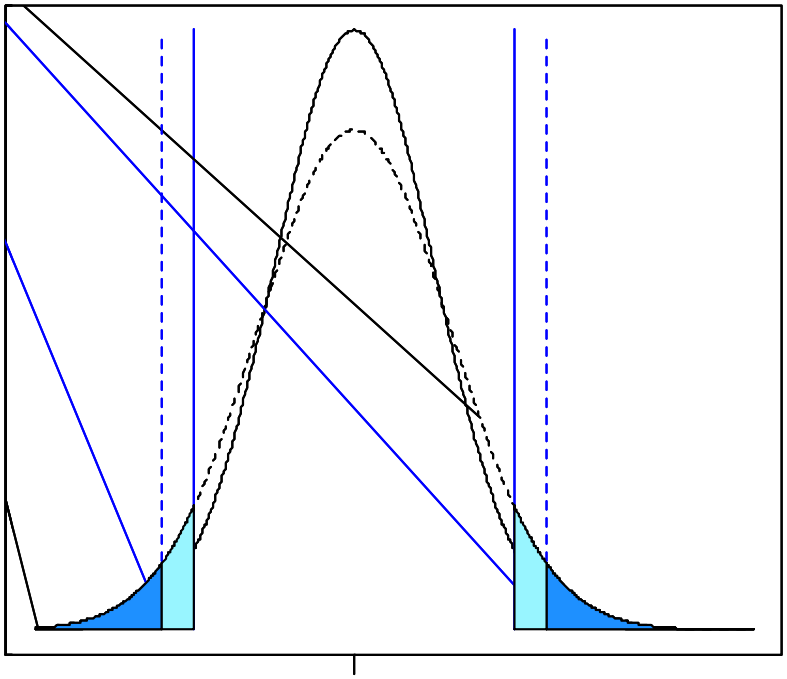


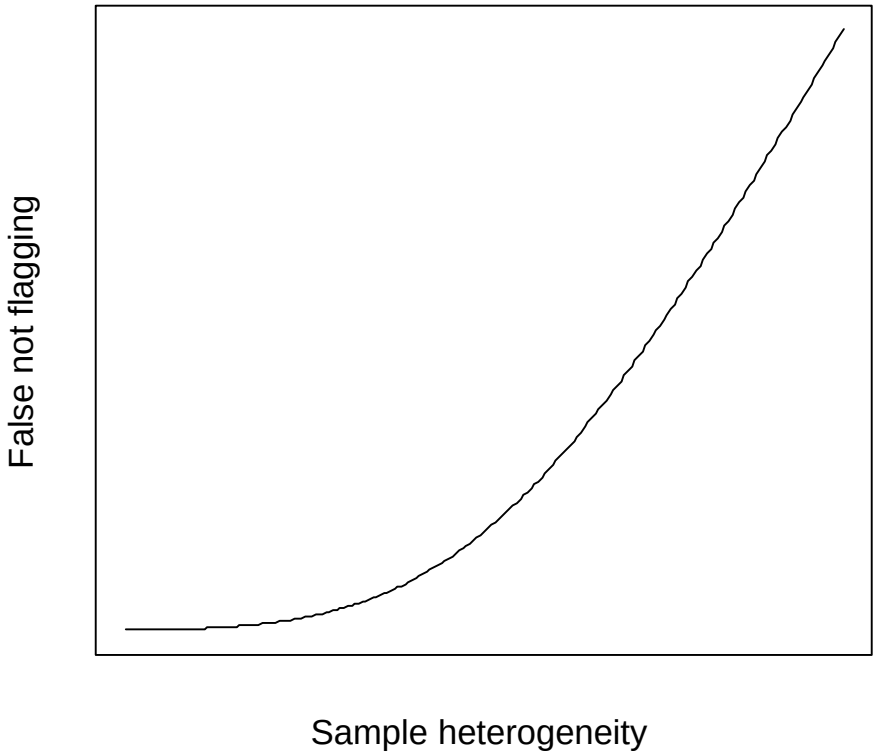
Sample heterogeneity

*Coucke, Wim, et al. "Alternative Sample-Homogeneity Test for Quantitative and Qualitative Proficiency Testing Schemes." *Analytical chemistry* 91.3 (2019): 1847-1854.

Falsely not flagging laboratories

- Falsely flagging of results by heterogeneity





Alternative criterion for homogeneity check

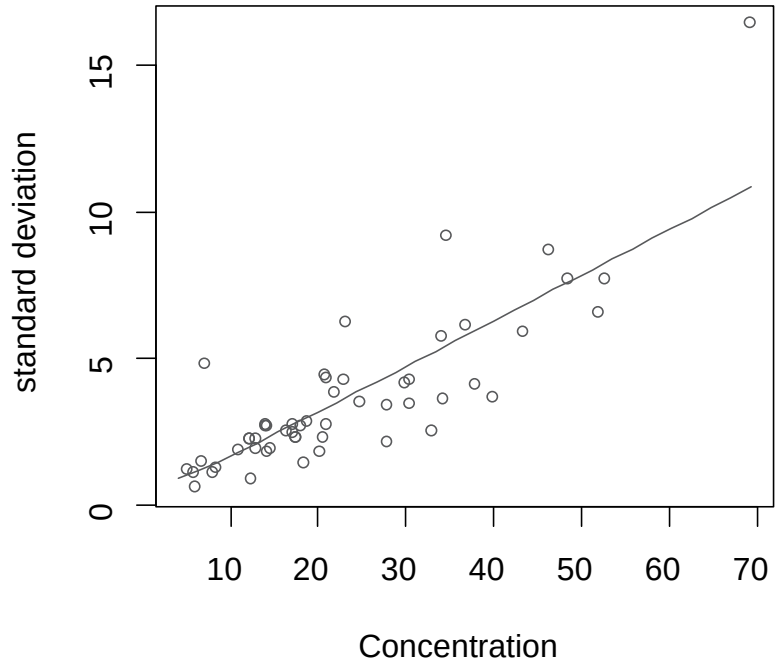
- Maximal s_s should be small enough such that:
  - Probability of falsely flagging well performing laboratories by Q-scores is small
  - Probability of falsely not-flagging badly performing laboratories by standard deviation-based limits is small

- Maximal s_s depends on expected variability of EQA results and evaluation criteria
  - Limits are peer group-dependent
  - Maximum s_s should be calculated for every peer group
  - Smallest maximum s_s of all peer groups should be chosen

How to estimate expected variability of EQA results ?

- Characteristic function*:

  Draws a relation between assigned value and expected variability of EQA results



$$SD = \sqrt{a + b * concentration^2}$$

Coucke, Wim, et al. "Application of the characteristic function to evaluate and compare analytical variability in an external quality assessment scheme for serum ethanol." *Clinical chemistry* 61.7 (2015): 948-954.

Alternative criterion for homogeneity check

- Example of new limits:

- Ethanol: Sample of 0.75 g/L
- APS: 6.5%
- Increase of 0.02 in false flagging rate
- False non-flagging rate equivalent to Z-score of 4

Method	sEQA	Limit Q-scores	Limit Z-scores	Final limit
Headspace chromatography (capillary-column)	0.0252	0.0081	0.022	0.0081
ADH- Abbott (Aeroset-Architect-Alinity)	0.0231	0.0079	0.0204	**0.0079**
ADH- Dade (Emit)	0.0354	0.0104	0.0312	0.0104
ADH- Roche	0.0238	0.0080	0.021	0.0080
ADH- Vitros	0.0335	0.0099	0.0295	0.0099

Alternative assessment of homogeneity criterion

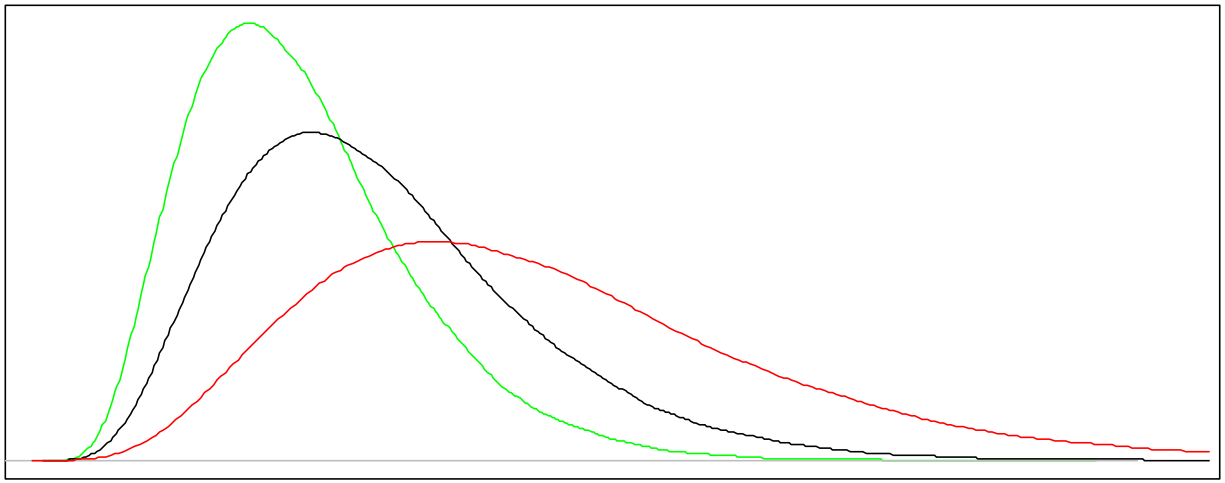
Basis of assessment:

If we measure a set of randomly chosen vials:

Measured variance = Inter-sample variance + Analytical variance

$$s^2 = s_s^2 + s_r^2$$

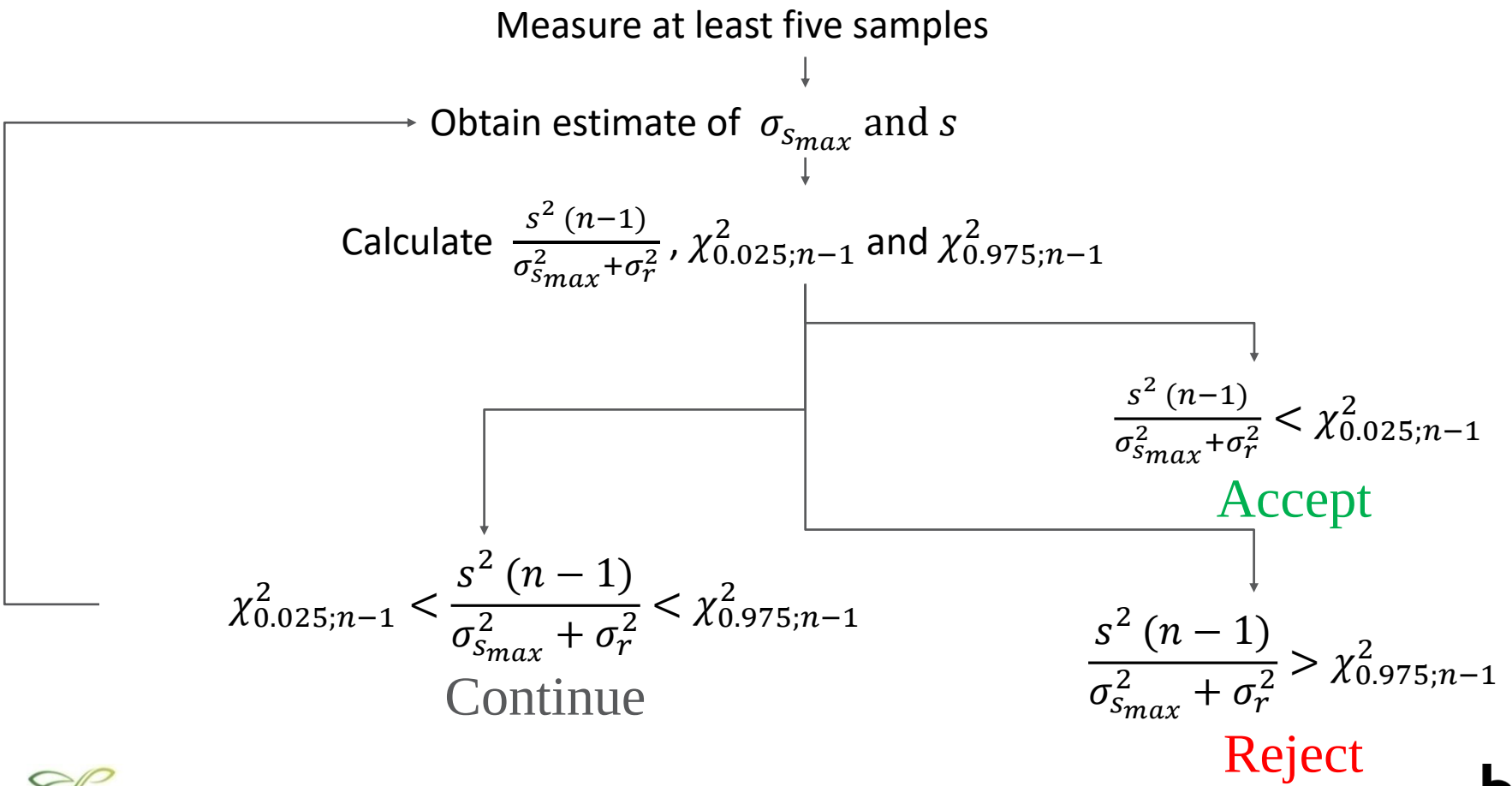
$$\frac{s^2\,(n-1)}{\sigma_{s_{max}}^2 + \sigma_r^2}$$



sciensano

.be

Alternative assessment of homogeneity criterion

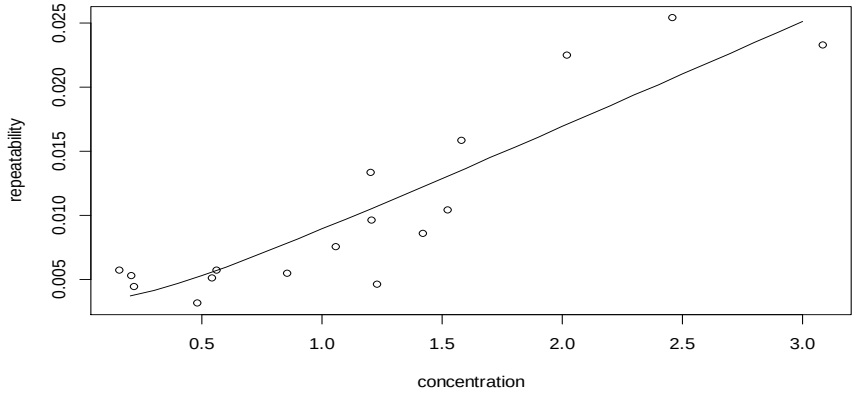


sciensano

.be

Alternative assessment of homogeneity criterion

- Practically:
  - Obtain estimate of sr (repeatability standard deviation)
    - Initially: take 20 consecutive measurements in one vial
    - After a while: take less measurements, use data from past and apply characteristic function



  - If sample by sample is too complicated: measure batches of 5 or 10 samples

Alternative assessment of homogeneity criterion: repeatability

Procedure in Excel:

	A	B	C	D	E	F
1	allowed sample heterogeneity	0.0079	g/L			
2	repeatibility	0.006	g/L			
3						
4	Sample number	Sample identification	Concentration			
5	1	Sample 1	0.7550			
6	2	Sample 2	0.7498			
7	3	Sample 3	0.7540			
8	4	Sample 4	0.7447			
9	5	Sample 5	0.7427			
10	6	Sample 6	0.7525		Accept batch	
11	7	Sample 7	0.7471			
12	8	Sample 8	0.7481			
13	9	Sample 9				
14	10	Sample 10				
15	11	Sample 11				
16	12	Sample 12				
17	13	Sample 13				
18	14	Sample 14				
19	15					
20	16					
21	17					
22	18					
23	19					
24	20					
25						
26						
27		Standard deviation:	0.004408			
28						

sciensano

.be

Homogeneity testing for non-continuous data

ISO 13528: appropriate number of samples that should all have the desired property

What is appropriate ?

10 ? We may end up with 23.8% nonconforming

100 ? We may end up with .9% nonconforming

Some concepts (II)*

Producer's quality level (PQL): proportion of defective items below which the producer doesn't want the batch to be rejected (read: wants very low probability that batch is rejected)

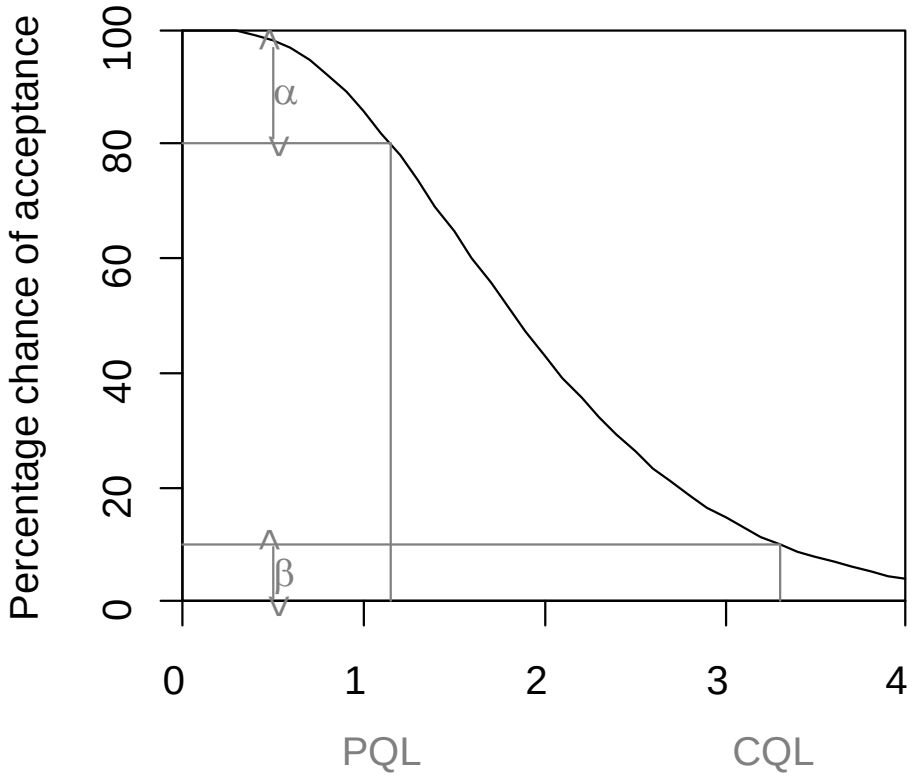
Consumer's quality level (CQL): proportion of defective items above which the consumer doesn't want the batch to be accepted (read: wants the batch to be accepted with very low probability)

The better the sampling plan, the lower the chance that batch would be falsely accepted or falsely rejected

*Schilling, Edward G., and Dean V. Neubauer. *Acceptance sampling in quality control*. Chapman and Hall/CRC, 2009.

Some concept (II)

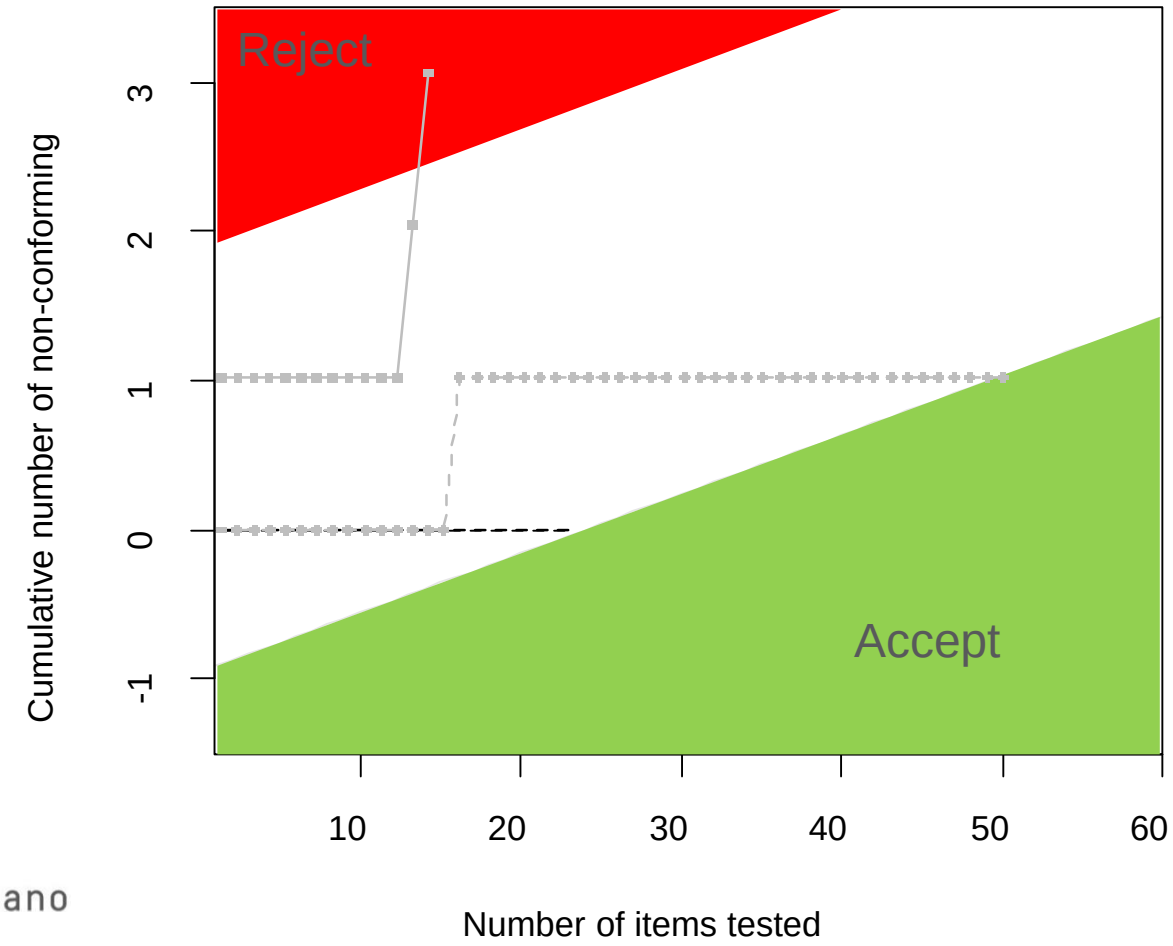
Operating-characteristic function



α: probability of falsely rejecting
β: probability of falsely accepting

Acceptance sampling for non-continuous data

PQL=1%
CQL=10%
α=1%
β=10%



sciensano

.be

Stability testing

- ISO 13528:

  Take at least 2 samples, measure them in the beginning and at the end.

- If |difference of means| $<0.3\sigma_{pt}$, accept batch for stability

- Comment:

  For a perfectly stable batch, and analytical variability of $0.5\sigma_{pt}$, the chance of accepting the batch is 45%

Stability testing: critics

- Why not an inferential test ?
  - Stability is assured if we have enough evidence of stability
    - H_0: no evidence of stability
    - H_a: evidence of stability

- What is beginning, what is end ?
  - Different sources of possible instability:
    - Transportation conditions
    - Duration till analysis

How to define stability ?

- How NOT to define stability ?
  - Some measurements in beginning, some measurements at end
  - Accept stability if t-test of comparison between beginning and end is not significant

- How to define stability ?
  - Stability is assured of falsely flagging of laboratories due to instability is lower than a predefined limit
  - Stability assessed by comparing two groups

Stability assured	Stability to be verified
Early analysis	Delayed analysis
Simple/short transportation	Complicated/long transportation

Instability and falsely flagging laboratories

- Instability: assigned value changes
- Effect of instability on evaluation depends on:
  - Using reference value or data-based assigned value;
  - Using fixed limits or data variability-based limit.
- Proving stability by Two One-sided Test Statistics (TOST)*
  - Limits of confidence interval of difference of mean are, in absolute value, smaller than allowed difference

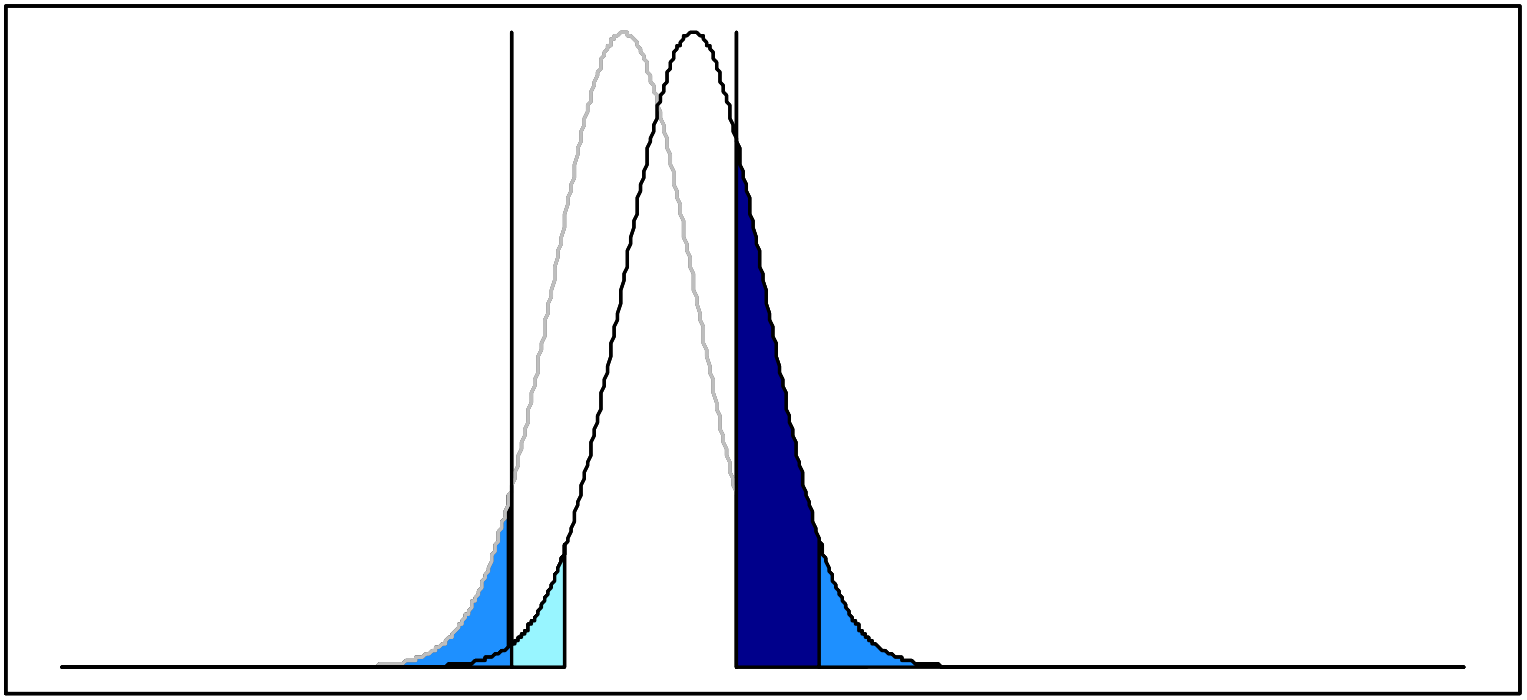
sciensano

* Ialongo, Cristiano. "The logic of equivalence testing and its use in laboratory medicine." *Biochemia medica: Biochemia medica* 27.1 (2017): 5-13.

.be

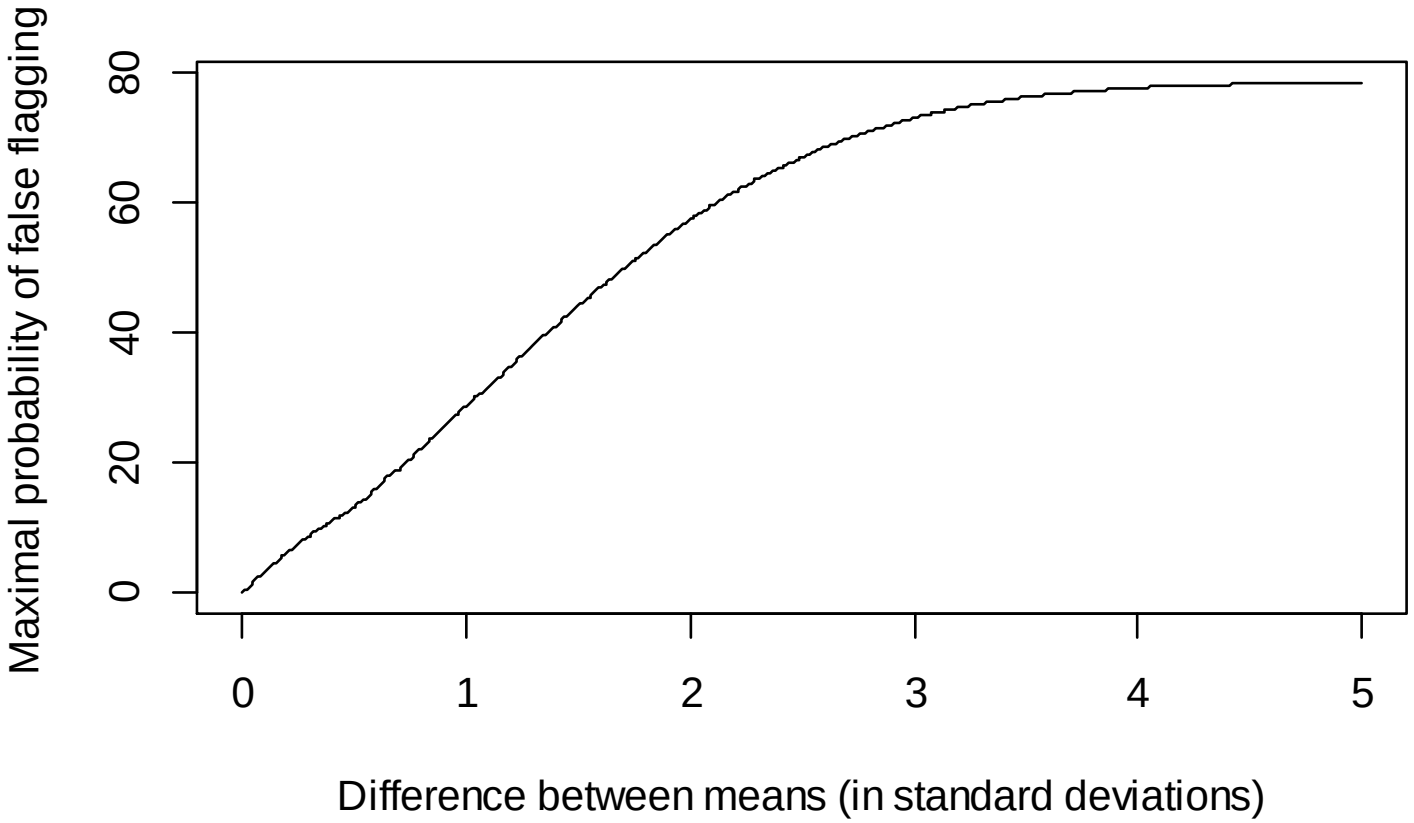
Effect of instability on data evaluation

Case 1: reference value, fixed limits



Effect of instability on data evaluation

Case 1: reference value, fixed limits

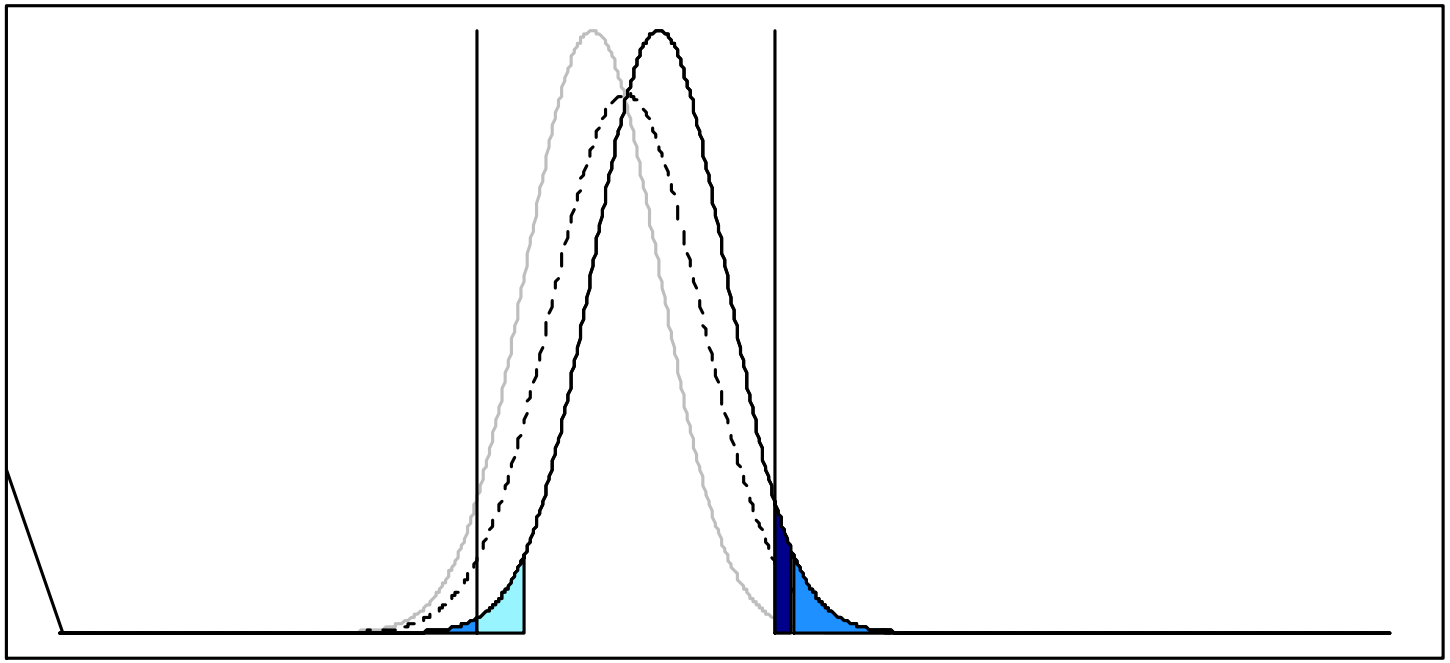


sciensano

.be

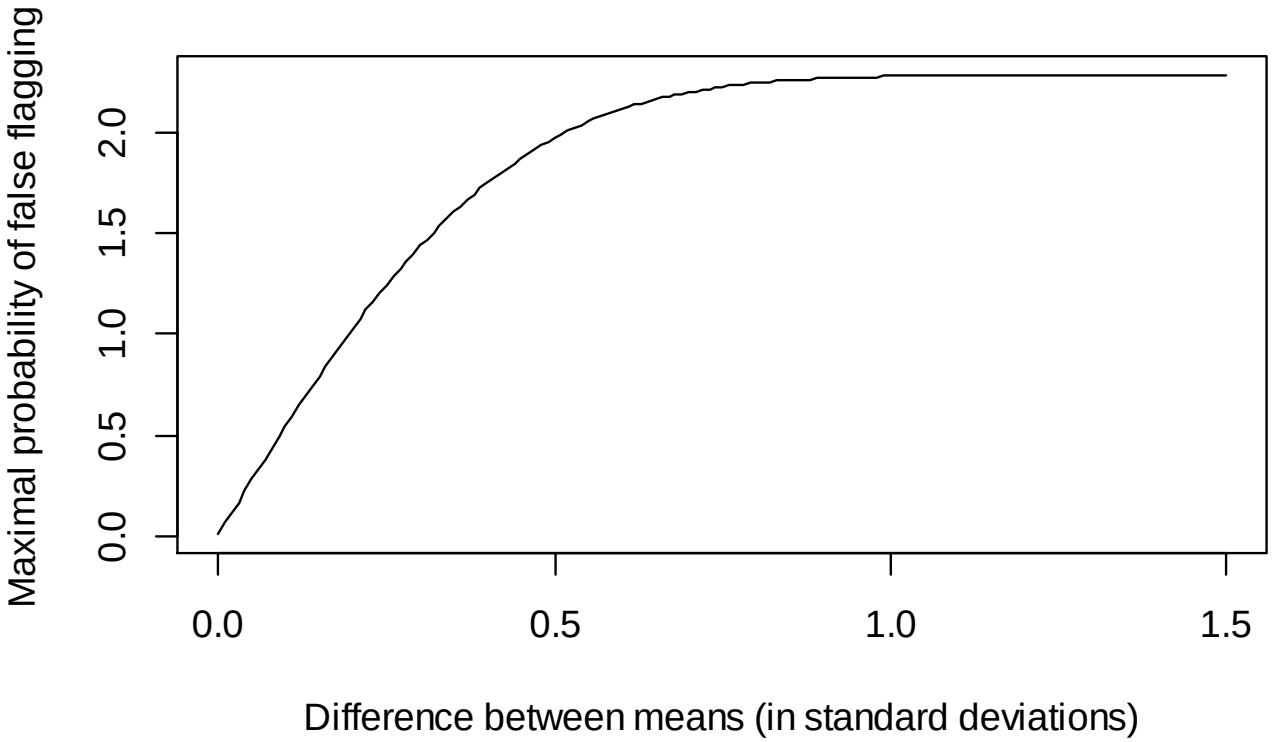
Effect of instability on data evaluation

Case 2: data-based assigned value and Z-scores



Effect of instability on data evaluation

Case 2: data-based assigned value and Z-score



sciensano

.be

Effect of instability on data evaluation

- Irrespective of calculation of assigned value and evaluation limits:
  - There is always a zone of extra flagging
  - There is always a zone of extra non flagging

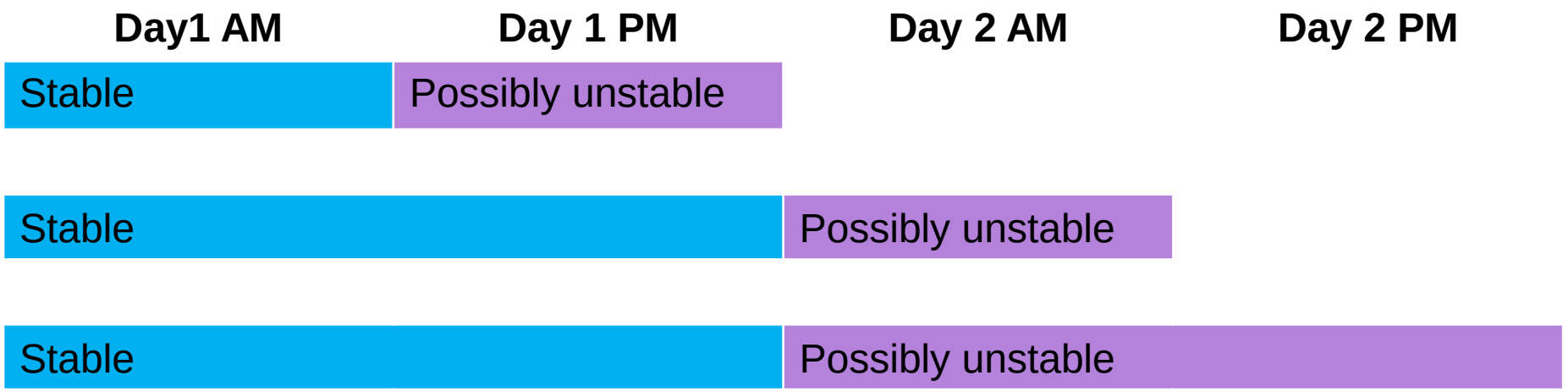
- Solution:
  - Identify all the zones where flagging change
  - Calculate all areas where flagging change and identify largest area
  - Allow difference between groups such that largest area remains smaller than a predefined level

Calculating limit for stability

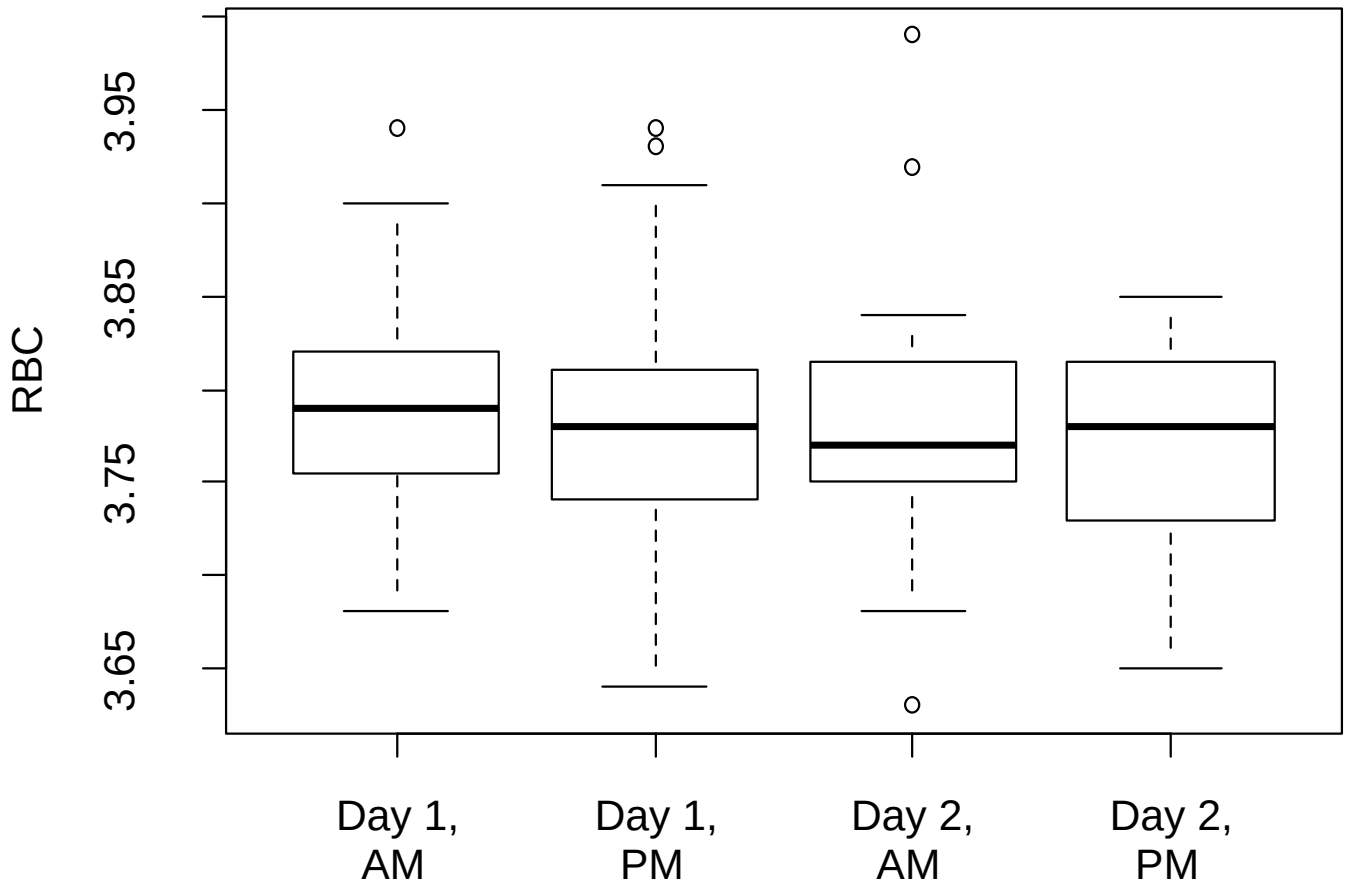
- Example: Haematology
- Fresh blood is sent via express mail
    - Day 1: 85% of the laboratories receive sample
    - Day 2: 15% of the laboratories receive sample
- Samples are analysed immediately after reception
- Laboratories are asked to report date and hour of analysis
- Analysed parameters: White Blood cells, Hematocrite, Hemoglobin, Red Blood Cells, Platelets
- False flagging rate for Q-scores should not be higher than 2%p
- False flagging rate for Z-scores should not be higher than 0.5%p

Calculating limits for stability

Various definitions of stable and possibly unstable group:



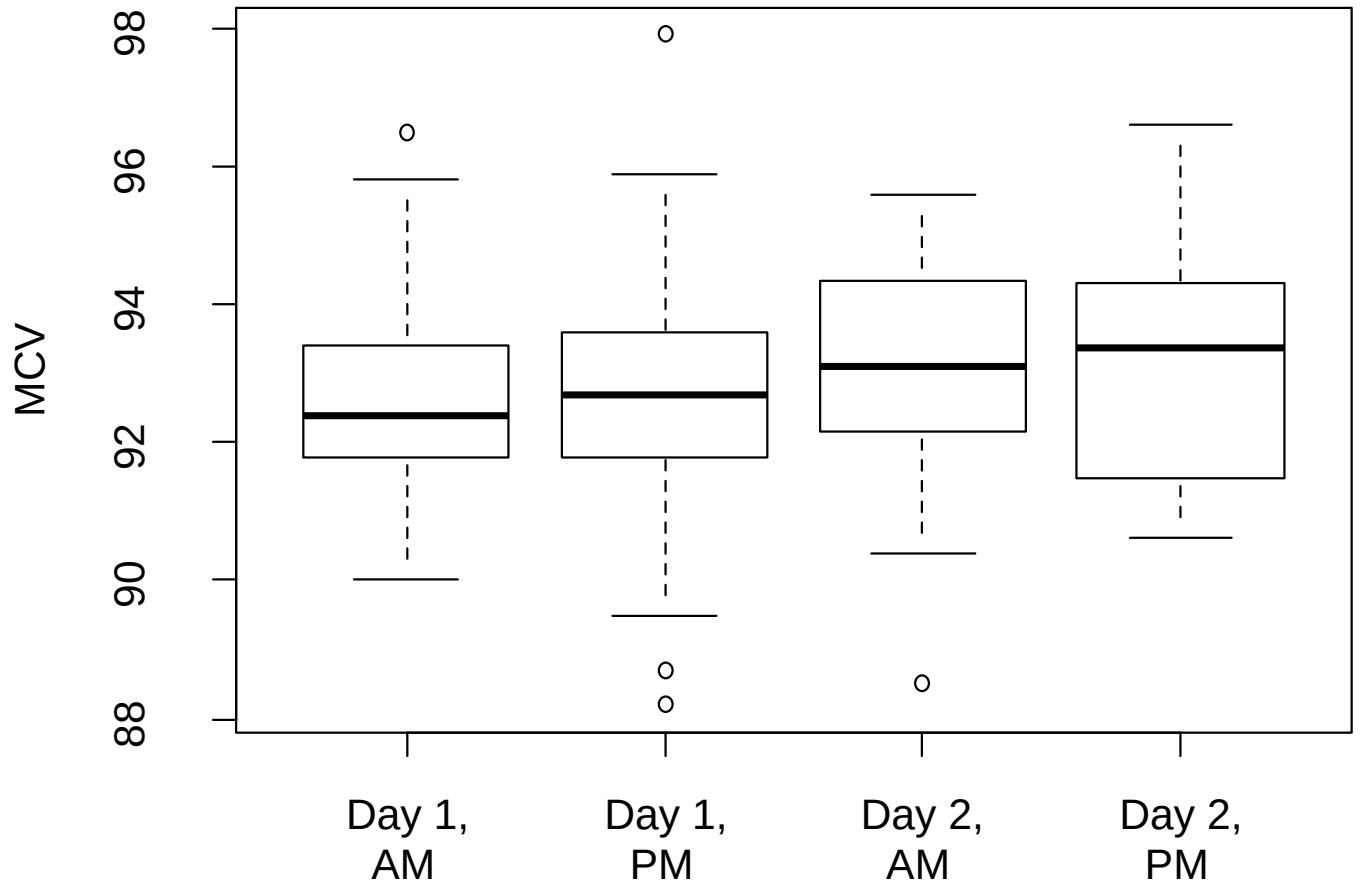
RBC: graphical representation



RBC: results

Comparison	Limit stability, Q-scores	Limit for stability, Z-scores	Confidence interval of diference	
Day 1 AM-PM	0.0773	0.1172	[-0.0097; 0.0387]	Proof of stability
Day 1-Day 2 (AM)	0.0675	0.0734	[-0.0418; 0.0581]	Proof of stability
Day 1 -Day 2	0.0676	0.0754	[-0.0414; 0.041]	Proof of stability

MCV: overview of results



sciensano

.be

MCV: results

Comparison	Limit for stability, Q-scores	Limit for stability, Z-scores	Confidence interval of diference	Evaluation
Day 1 AM-PM	2.2454	3.9544	[-0.8035; 0.5071]	Proof of stability
Day 1 -Day 2 (AM)	1.8322	2.578	[-2.6932; 0.0298]	No proof of stability
Day 1 -Day 2	1.8344	2.6505	[-2.8783; -0.6184]	No proof of stability

Discussion

- Homogeneity testing
  - Performed prior to sending, batch can be rejected if necessary
  - Sample size depends on difference between actual heterogeneity and maximum limit
  - Limit for falsely (non) flagging is proposed, but to be confirmed
  - Limit could depend on category of EQA scheme
- Stability testing
  - Performed during EQA round itself, batch cannot be rejected if necessary
  - Sample size determines power
    - The more data, the sooner proof of stability
    - Comfortable if 50 data or more
  - Extra information needed: way of transportation, hour of analysis

Discussion

Homogeneity and stability testing for one parameter or all parameters ?

Alternative, post-hoc evaluation:

1. Calculate classic averages and standard deviations after outlier exclusion
2. Calculate characteristic function

$$SD = \sqrt{a + b * concentration}$$

3. Calculate confidence interval of each standard deviation

$$\frac{\sqrt{(n-1) * SD^2}}{\sqrt{\chi^2_{0.01;n-1}}}$$

4. Check points outside confidence interval

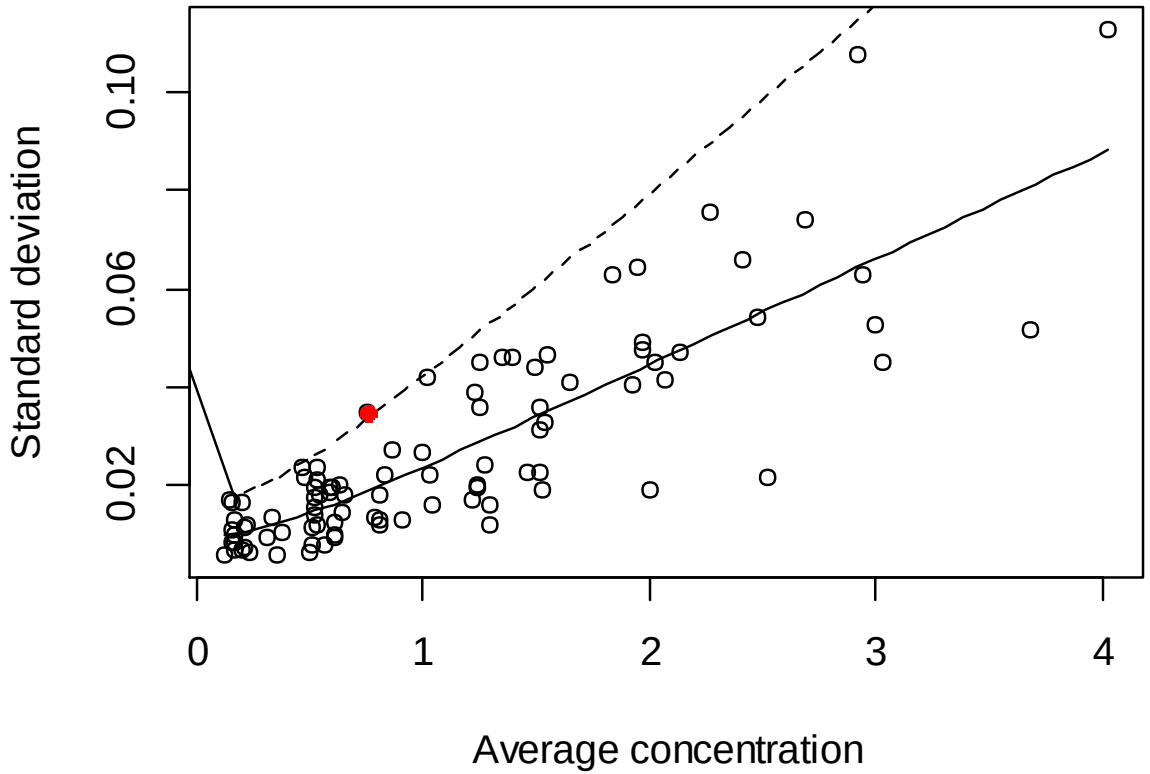
sciensano

.be

Discussion

Homogeneity and stability testing for one parameter or all parameters ?

Alternative, post-hoc evaluation:



sciensano

.be

Future prospects

- Homogeneity evaluation when order of processing is known
  - Based on evolution of parameter through distribution
- Bayesian approach of homogeneity and stability

  Information from prior distributions can be included
- Combining information of multiple samples for homogeneity and stability testing
  - If power of stability testing is not enough
  - If homogeneity cannot be assured
- Stability testing for non-continuous data
  - Don’t know how