Interpretation of Proficiency Test Results

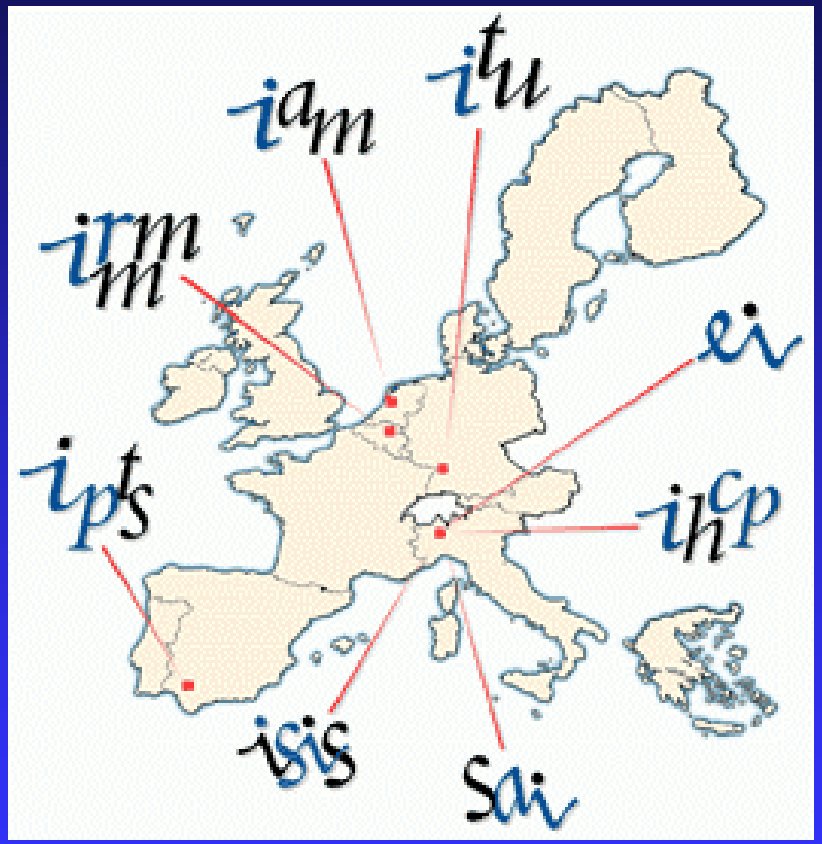
Christoph von Holst

European Pesticide Residue Workshop
30 May 2002

Some words about who we are…..

- Joint Research Centre (JRC) is a service European Commission
- Provides scientific and technical support for the conception, implementation and monitoring of EU policies
- Serves the common interest of the Member States, while being independent of special interests, whether private or national

The institutes of the JRC



- Two institutes deal with food issues in the JRC
- Institute for Health and Consumer Protection (IHCP) in Ispra, Italy
- Institute for Reference Materials and Measurements (IRMM) in Geel, Belgium

The Food Products Unit of the IHCP

- To harmonise and validate analytical procedures
  - Method development (e.g. mycotoxins, PCBs)
  - Conducting validations studies via collaborative trials
- To monitor substances at European level (e.g. bisphenol A in coatings of food cans)
- To establish and maintain data bases (authenticity of wine)
- To act as a “help desk” in urgent cases (Belgian PCB episode 1999)

Our work on pesticides

- Scientific support to DG Health and Consumer protection (Brussels and Dublin)
  - Evaluation of sampling designs applied to the European coordinated monitoring program of pesticides in certain food matrices
  - Contributing to the evaluation of the results from European proficiency tests

Why talking about proficiency tests (PTs) ?

- Many laboratories participate regularly in PTs
- *Example:* Laboratories contributing to the European Monitoring Program for pesticides *have* to participate in a dedicated PT analysing between 10 and 20 compounds
- How to evaluate all these values in order to assess the proficiency of the laboratory ?
- Information needed for
  - The participating laboratories
  - Accreditation body
- Aim of the lecture: To give addition information as to the best use of data from PTs

What is the lecture about ?

- Basic aspects of PTs.
- Some questions when looking at PT results:
  - Is a laboratory well performing when gaining one sufficient PT score ?
  - How to evaluate the proficiency of a laboratory when more than one z-score are available ?
  - Which statistics should be applied when the data are not normal distributed ?
- Examples were taken from the last European Commission's Proficiency Test on Pesticides in fruits and vegetables (EU PT 3) in 1999

Why participating in PTs ?

Participate in proficiency tests

Use validated method

How to assure sufficent quality of my data ?

ISO 17025 requires statement on uncertainty

Follow specific quality guidelines

Analyse certified reference material

Basic aspects of PTs.

- Is the result of a laboratory within a specified limit ?
- The participants analyse test material using their own method. Statistical assessment of the laboratories' results delivers the z-score (z) for each laboratory:

Result of the laboratory

$$z = \frac{x_{lab} - X}{\sigma}$$

Assigned values

Target standard deviation

- Relating the difference ($x_{lab} - X$) to the target standard deviation allows comparison of results from different PTs
- z-scores allow for easy evaluation
  - $z < |2|$ acceptable
  - $z < |3|$ and $> |2|$ questionable
  - $z > |3|$ unacceptable

Important effect on the outcome of a study: The target standard deviation

- The target standard deviation (SD) is the "yardstick" regarding the laboratories' performance
- The target SD defines the range in which the results of the laboratories should be
- Defining the target SD influences the evaluation of the results:
  - A low target SD makes the PT more rigid for the laboratories

How to establish the target SD ?

- Adjusting to the required accuracy of results considering the later user of the data (fitness for purpose)
  - Using the Horwitz-equation
  - Derived from specific quality guidelines the laboratories have to follow (European monitoring program)
  - This SD describes the user requirements, not the data
- Advantage
  - The performance criterion (“yardstick”) is related to needs
- Disadvantage
  - Difficult to establish when required precision is not clear
  - Laboratories may be treated to harshly when analysing difficult compounds (target SD to low)

How to establish the target SD ?

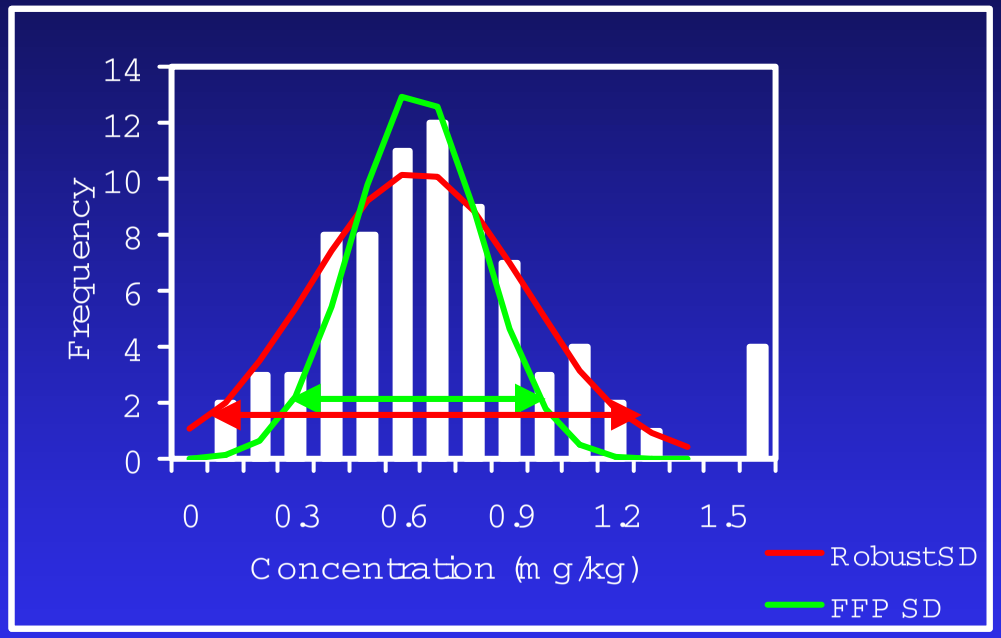
- Robust standard deviation from the submitted results
- Advantages
  - Applying robust statistics the influence of results from poor performing laboratories are minimised
  - Compound specific characteristics are accounted for
  - Evaluation of combined z-scores possible
- Disadvantages
  - The target SD can be to high when portion of poor performing labs is to high
  - At target SDs above 30 %: z-scores are to generous to low results

Robust versus Fitness For Purpose Standard deviation (FFP SD)

- Example:
  - Methamidophos (incurred) in cucumber
  - Assigned value: 0.63 mg/kg
  - Number of laboratories: 84
  - FFP SD: 30 %
  - Robust SD: 47 %
  - Main characteristics: Extremely high standard deviation of the data

Robust versus Fitness For Purpose Standard Deviation (FFP SD)

Example: Methamidophos in cucumber (EU PT 3)



	SD	z-score	
		2	-2
RobustSD	47%	1.24	0.04
FFP SD	30%	1.02	0.26

- Evaluation based on the robust SD accepts results close to 0
- FFP SD avoids this problem but might treat laboratories to harshly (compound specific analytical problems)

Robust versus Fitness For Purpose Standard Deviation (FFP SD)

Selected examples from EU PT 3

	Assigned value (mg/kg)	Robust Statistics		FFP Statistics	
		SD %	Acceptable Results (%)	SD %	Acceptable Results (%)
Propoxur	0.26	24	77	20	72
Carbendazim	0.49	24	83	20	83
Diazinon	0.14	22	86	20	85
Methamidphos	0.63	49	94	30	72
Imazazil	4.23	60	98	30	71
Acephate	0.15	49	93	30	66

- Robust SD and FFP SD are similar: Portion of acceptable results does not differ
- Robust SD is very high: A too (!) high portion of laboratories gains acceptable z-scores

Conclusions from this discussion:

- The coordinator of a PT has to justify the selection of the target standard deviation.
- Analytes showing high variability (Robust SD > 30 %) are difficult to evaluate using the currently available approaches.

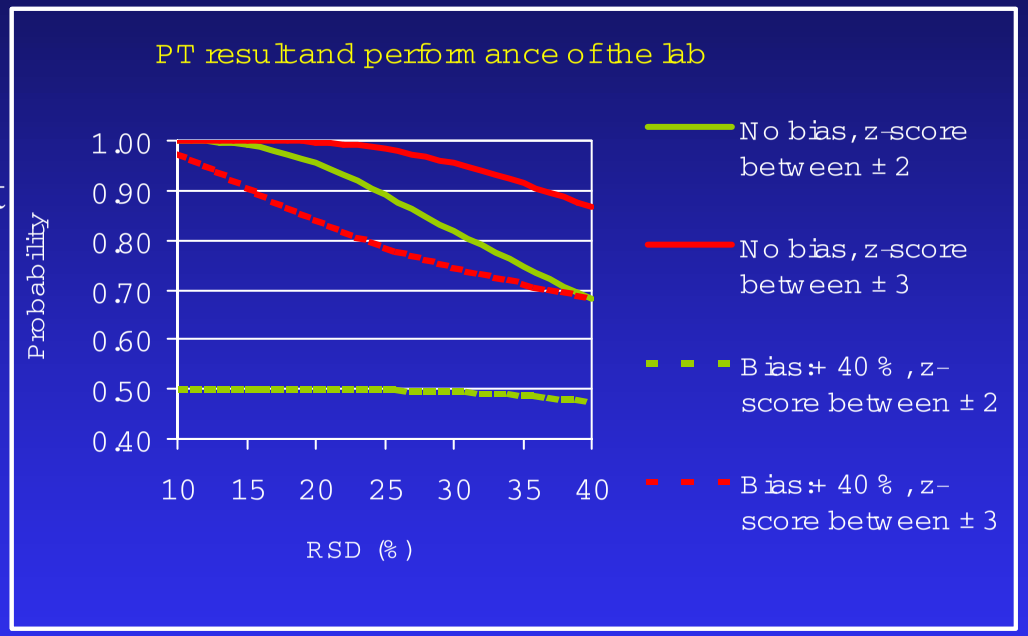
Assumptions about the quality of the laboratory

- The organizer of the PT does not know the uncertainty of the laboratories' results
- It is assumed that the uncertainty of the results of all laboratories are characterized by the target standard deviation
- Therefore a z-score of 2 (= Twice the SD) is still acceptable in order to avoid wrong "punishment" of good laboratories (risk: 5 %)
- Justification: In many fields laboratories need to be accredited according to ISO 17025

Power of a PT based on a single z-score

- Power: Capability of the PT to identify a poor working laboratory.

- Determination of a single compound in a PT.
- What is the probability of a laboratory to gain a sufficient (± 2) or questionable (± 3) z-score ?
- The probability depends on the performance of the laboratory (variability and bias).
- Target standard deviation is 20 %.



- Conclusion: Proficiency of a laboratory can **not** be evaluated, when looking only at **one** z-score. Evaluation needs **more** z-scores.

How to interpret more than one z-scores of one laboratory ?

- Laboratories gain more than one z-score when
  - analysing different analytes in the test material
  - participating in different PTs
- Is the proficiency of a laboratory sufficient when
  - in two PTs both z-scores of an analyte were 2 ?
  - in two PTs one z-score of an analyte was 2 and the other was –2 ?
  - in five PTs all z-score of an analyte varied from 1.3 and 1.5 ?

Probability of certain z-score combinations

- We assume that the laboratory's performance is equivalent to the target standard deviation
- Frequency distribution of low and high z-scores when evaluating the results from various PTs

Total number of z-scores	Between -1 and 1	Between 1 and 2 or -1 and -2	Below -2 and above +2	Combined frequency
1	z-scores 1	z-scores 0	z-scores 0	68%
1	0	1	0	27%
1	0	0	1	5%
3	0	3	0	2%
3	1	2	0	15%
3	1	1	1	5.5%

How to combine z-scores ?

- By combining these z-scores, the power of the PTs can be increased, but still limiting the risk of erroneous rejection of a well performing laboratory to 5 %
- The harmonised protocoll[1*] for PTs suggests:
  - Sum of Squared Scores (SSZ)

    $SSZ = \sum z\text{-score}^2$
- Characteristics of the SSZ
  - Random and systematic are treated equally
  - Variance of the z-scores following the χ^2 distribution
  - Numerical interpretation difficult, since the value depend on the number z-scores.

[*] *M. Thompson and R. Wood, J AOAC International, 76 (1993) 926*

How to combine z-scores ? cont.

- The SSZ combines several z-scores, but it is difficult to understand the meaning.
- Uhlig and Lischer* proposed the "Relative Laboratory Performance" (RLP) derived from the SSZ.

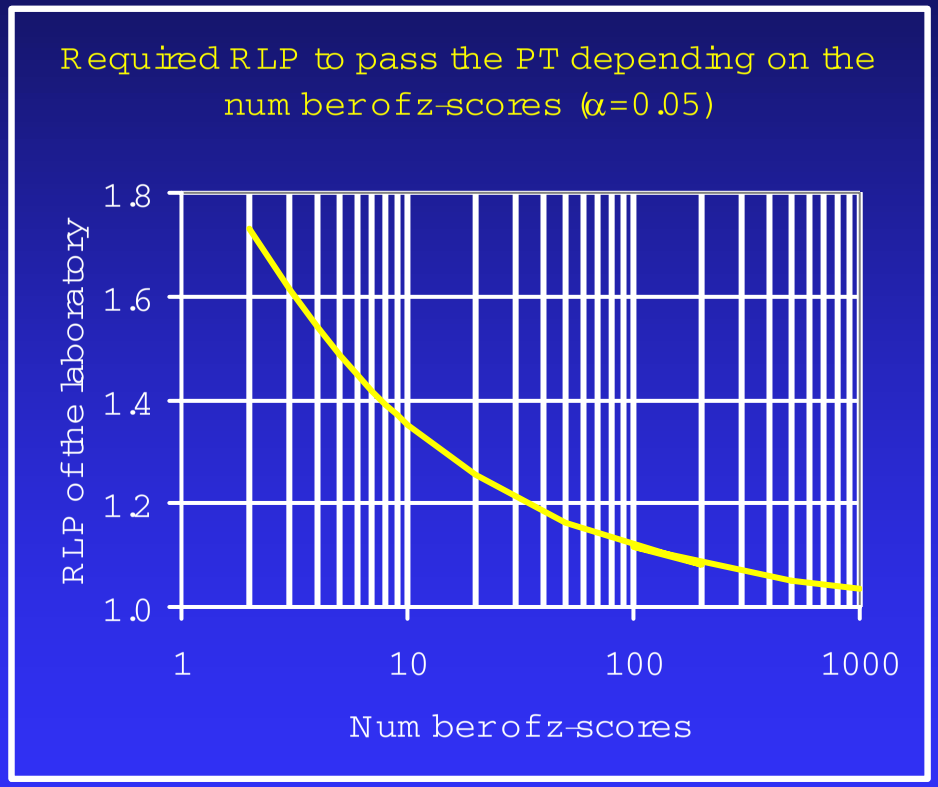
$$RLP = \sqrt{\frac{SSZ}{n}}$$

n = number of z-scores

- The RLP describes the laboratory performance in terms of a factor:
  - Factor of 1: The lab delivers results within the target SD
  - Factor of 2: The lab delivers results within *twice* the target SD

* *S. Uhlig and P. Lischer, The Analyst (1998), 123 (167-172)*

How does a combined z-score improve the power of the PT ?



- Required RLP to pass the test with a probability of 95 % increases when participating repeatedly in PTs
- Example: When 10 z-scores are available the RLP limit is about 1.4

Example: RLP from the EU PT 3

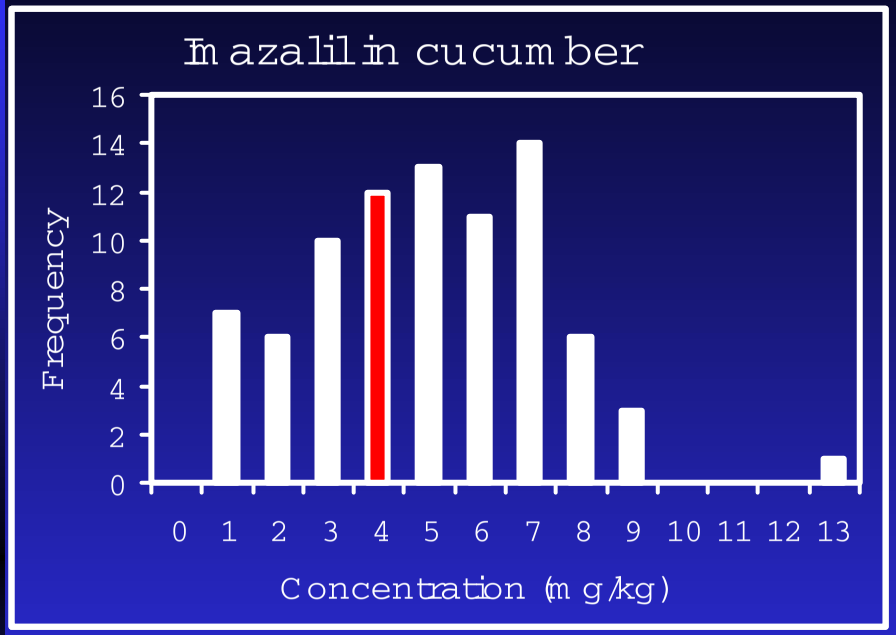
Laboratory Code	Acephate	Aldicarb	Carbendazim	Deltamethrin	Diazinon	Endosulfan	Imazalil	Metalaxyl	Methamidophos	Methomyl	Permethrin	Pirimiphos-methyl	Propoxur	Vinclozolin	Number of z-scores	RLP (critical)	RLP (obtained)	Evaluation
1	0.2		-0.2	0.6	0.6	0.7	1.3	0.9	0.0	2.7	0.5	1.0	-0.6	0.6	13	1.3	1.0	good
2	-0.6		-0.1	-0.6	-0.5	-0.6	-1.3	0.3	0.2		-0.5	-0.4	-0.5	0.0	12	1.3	0.6	good
3	-0.8	0.5	-2.2	0.4	-0.3	0.3	-1.1	3.2	-0.7	0.4	0.2	-0.8	-1.1	0.8	14	1.3	1.2	acceptable
4	1.1	0.2	0.0	1.4	0.6	-0.7	-0.6	-0.4	1.2	1.1	-0.4	0.6	-0.2	0.3	14	1.3	0.8	good
5	3.5	-0.7	0.2	-0.2	0.3	-0.5	0.2	1.5	0.4	-0.4	-0.8	0.3	-0.3	0.9	14	1.3	1.1	good
6	-0.7		1.1	0.8	0.8	-0.2	-0.3	1.1	-0.9		1.3	0.3	2.3	0.5	12	1.3	1.0	good
7	-0.1	-3.5	-1.2	2.7	3.2	0.0	-1.2	-0.4	-0.6	-3.4	-0.7	-0.4	2.9	-0.3	14	1.3	2.0	unacceptable

- The laboratory's proficiency is good when the RLP does not exceed the target standard deviation (RLP=1)
- The laboratory's proficiency is good when the RLP does not exceed the critical RLP value ($\alpha = 0.05$)
- When using the FFP SD, the risk of erroneous rejection of good laboratories can not be estimated

Deviation from normal distribution: A problem ???

- Underlying assumption of all approaches: Data are normally distributed
- Statistical evaluation of the results from EU PT 3 revealed that this assumption is not always fulfilled
- Conclusion: When applying “normal” statistics to results with a skewed distribution laboratories with low results are treated to generous whereas other are treated to harshly

Example: Imazalil (EU PT 3)

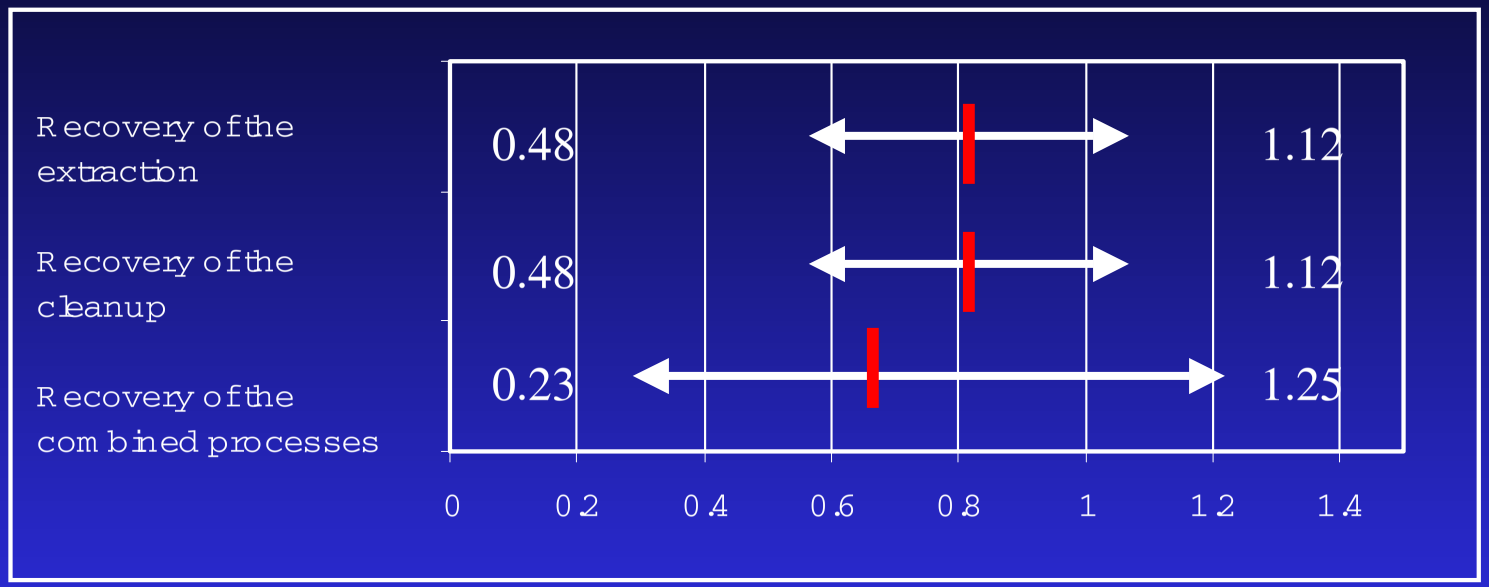


	SD	z-score	
		-2	2
		Concentration	
RobustSD	60%	-0.8	9.2
FFP SD	30%	1.7	6.7
Assigned value (mg/kg)	4.2		

- Distribution of the results is skewed
- Robust SD as target SD: Also negative values are still acceptable
- FFP SD: Low values are handled to leniently compared to high values
  - A low value involving a 2.5 fold error is acceptable (4.2/2.5 =1.7)
  - A high value involving a 2.5 fold error is **not** acceptable (4.2*2.5 = 10.5)

Factor statistics*

- Propagation of errors in analytical chemistry can lead to a skewed distribution



- Recovery (R) of both steps are normally distributed (R: 0.8, RSD: 20 %)
- Due to the multiplicative (not additive) combination of errors the combined distribution becomes skewed

* *Hill A. and von Holst C.,The Analyst (2001) 126 2044-2052*

Hill A. and von Holst C., The Analyst (2001) 126 2053-2060

Factor statistics cont.

- Current methods for analytical uncertainty:
  - *Symmetric* error ranges defined by the SD
  - Disadvantage: Results close to the LOD are acceptable when the target SD is very large
- Alternative approach: Factor statistics
  - Defining *asymmetric* error range by using a factor
  - Example:
    - Assigned value: 1mg/kg
    - Factor describing the acceptable error: 2
    - Range: 0.5 mg/kg (1/2) to 2 mg/kg (1*2)
  - The error range does not include negative values

Factor statistics cont.

- Calculating a robust factor standard deviation (FSD)
  - Determine assigned value (e.g. median)
  - Conversion of the results to factors of the assigned value
    - Concentrations above the assigned value: The values are divided by the assigned value
    - Concentrations below the assigned value : The median is divided by the concentration
    - All factors are above 1
  - Example
    - Assigned value: 1 mg/kg
    - Result: 2 mg/kg → Factor = 2/1 = 2
    - Result: 0.5 mg/kg → Factor = 1/ 0.5 = 2

Factor statistics cont.

- Calculating a robust factor standard deviation (FSD)
  - All factors are sorted and the FSD was set taking the 68 percentile of the factors (corresponds to normal SD).
  - Based on the FSD, factors corresponding to a z-score of 2 and 3 can be calculated.

Examples

Analyte		SD	z-score		Median (mg/kg)	z-score	
			-3	-2		2	3
Imazazil	RobustSD (%)	60	-3.36	-0.8	4.2	9.2	12
(n=84)	FFP SD (%)	30	0.42	1.7	4.2	6.7	8
	FSD	1.7	1.4	1.8	4.2	7.0	12.5
Diazinon							
(n=106)	RobustSD (%)	22	0.05	0.08	0.15	0.22	0.25
	FFP SD (%)	20	0.06	0.09	0.15	0.21	0.24
	FSD	1.3	0.08	0.10	0.15	0.19	0.27

Conclusions

- A **single** result from a PT does not give information about the proficiency of a laboratory
- Repeated participation in PTs is necessary for the evaluation of the laboratories' performance
- Applying normal statistics can lead to a distorted figure of an acceptable error range
- Factor statistics gives more reliable results when evaluating analytes with high variability of the data
- Ongoing improvement of statistical approaches is required

Acknowledgement

- I am grateful to
  - Alan Hill (factor statistics) and Lutz Alder (combined z-scores) for the cooperation on these topics
  - Arne Anderson, André de Kok and Amadeo Fernandez-Alba for the exciting discussions during the evaluation of the results from EU PT 3