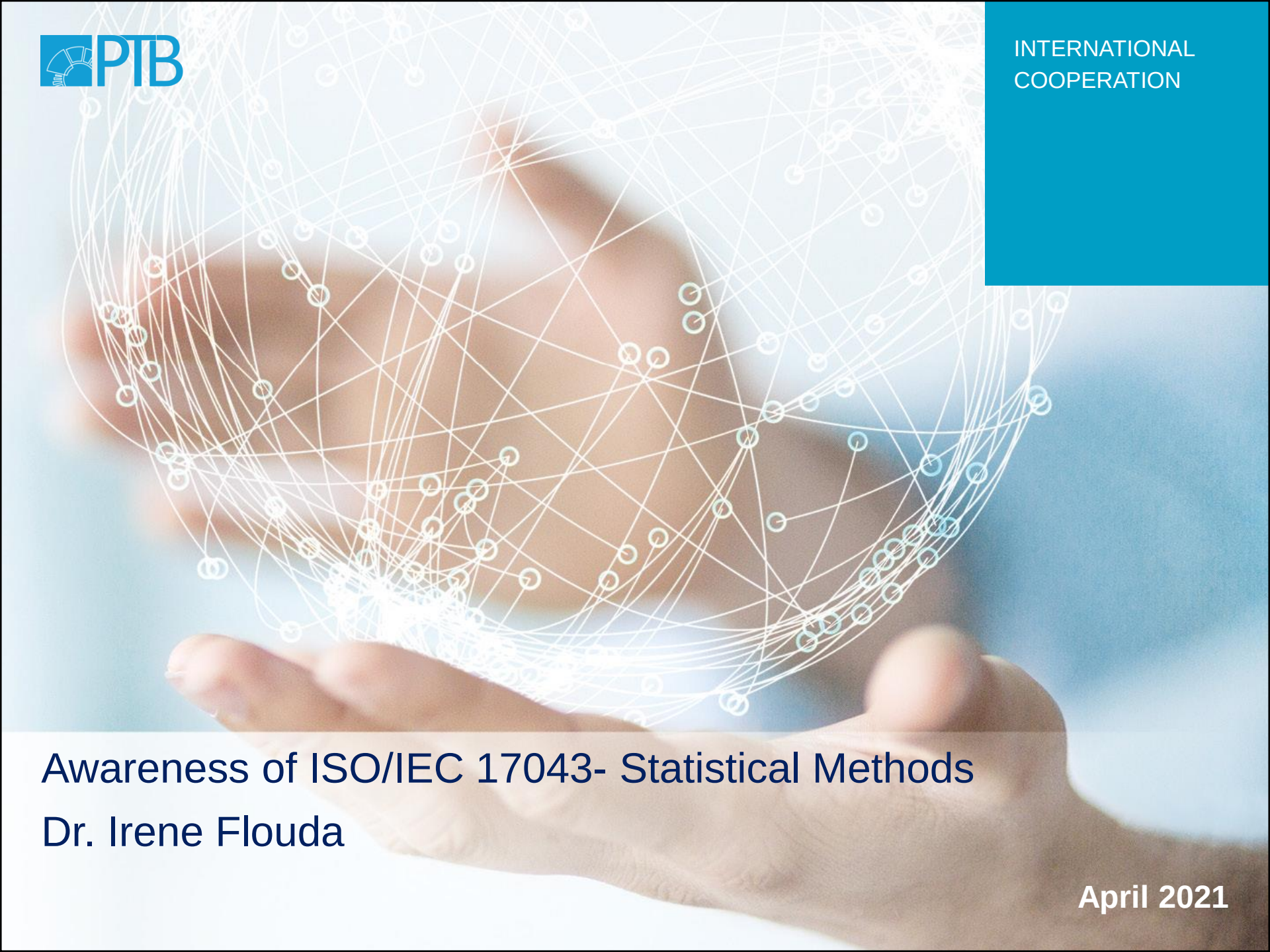


INTERNATIONAL COOPERATION

Awareness of ISO/IEC 17043- Statistical Methods

Dr. Irene Flouda

April 2021

ISO/IEC 13528:2015- Contents

- Introduction
- Scope, Normative references, Terms and definitions
- General Principles –basic requirements–statistical model and design considerations
- Initial review of PT items and results
- Determination an the assigned value
- Determination of evaluation criteria
- Performance scores

ISO/IEC 13528:2015- Contents

- Graphical techniques
- Qualitative data
- Normative Annexes
  - ✓ A. Symbols
  - ✓ B. Homogeneity and Stability
  - ✓ C. Robust analysis
- Informative Annexes
  - ✓ D. Additional guidance
  - ✓ E. Illustrative examples

Introduction to ISO/IEC 13528:2015

Sections 0.1-0.5

- ✓ Purposes of proficiency testing
  - ➢ Same as ISO/IEC 17043 Introduction
- Rationale for scoring
  - ✓Use participant results or independent criteria
- ISO 13528 and ISO/IEC 17043
  - ✓New sections in 13528, new topics in ISO 17043
- Statistical expertise – Important
- Computer software
  - ✓Must be validated

ISO/IEC 13528:2015- Section 1: Scope

- For Providers of PT schemes:
  - ✓ Detailed descriptions of statistical methods for design of scheme, and
  - ✓ For analysis of data from the PT scheme
- For participants and accreditation bodies:
  - ✓ Statistical methods to interpret PT data.
- General:
  - ✓ Can be used to demonstrate acceptable performance relative to specific criteria
  - ✓ Procedures for quantitative & qualitative data
  - ✓ Can be applied in Inspection

ISO/IEC 13528:2015- Section 2: Normative Documents

- Essential Documents for application
  - ✓ ISO/IEC 17043
  - ✓ ISO Guide 30: Terms and definitions for RM
  - ✓ ISO 3534 : Statistics vocabulary and symbols
  - ✓ ISO 5725: Accuracy of measurement methods and results
  - ✓ ISO Guide 99: VIM

Section 3: Terms and Definitions

Most terms and definitions from normative references, some repeated for clarity

- PT, PT item, PT provider, PT scheme, ILC, participant, measurement error

Some terms modified slightly

- SDPA- Standard Deviation for Proficiency Assessment, assigned value, outlier, PT item

Some new definitions

- Consensus value, action signal

Section 3: Terms and Definitions

SDPA

- Measure of dispersion used in the evaluation of results of proficiency testing
- NOTE 1 This can be interpreted as the population standard deviation of results from a hypothetical population of laboratories performing exactly in accordance with requirements.
- NOTE 2 The standard deviation for proficiency assessment applies only to ratio and interval scale results.
- NOTE 3 Not all proficiency testing schemes evaluate performance based on the dispersion of results. [revised from ISO/IEC 17043]

Section 3: Terms and Definitions

Assigned Value

- value attributed to a particular property of a proficiency test item

Consensus Value

- value derived from a collection of results in an interlaboratory comparison
- NOTE The phrase ‘consensus value’ is typically used to describe estimates of location and dispersion derived from participant results in a proficiency test round, but may also be used to refer to values derived from results of a specified subset of such results or, for example, from a number of expert laboratories

Section 3: Terms and Definitions

Outlier

- a member of a set of values which is inconsistent with other members of that set.
- NOTE 1 An outlier can arise by chance from the expected population, originate from a different population, or be the result of an incorrect recording or other blunder.
- NOTE 2 Many schemes use the term outlier to designate a result that generates an action signal. This is not the intended use of the term. While outliers will usually generate action signals, it is possible to have action signals from results that are not outliers.

Section 3: Terms and Definitions

Action Signal

- indication of a need for action arising from a proficiency test result

proficiency test item

- sample, product, artefact, reference material, piece of equipment, measurement standard, data set or other information used to assess participant performance in proficiency testing

4 General Principles

- 4.1 General Requirements

4.1.1 The statistical methods used shall be fit for purpose and statistically valid.

- Any statistical assumptions shall be stated in the design (or any other written description

4.1.3 The proficiency testing provider shall provide participants with:

- a description of the calculation methods used
- an explanation of the general interpretation of results,
- a statement of any limitations relating to interpretation

4 General Principles

4.2 Basic Model

- **4.2.1** For quantitative results in proficiency testing schemes where a single result is reported for a given PT item:

$$x_i = \mu + \varepsilon_i$$

Where

- ✓ x_i = proficiency test result for participant i
- ✓ μ = true value for the measurand
- ✓ ε_i = measurement error for participant i

4 General Principles

- **NOTE 1** Common models for ε include:

✓ the normal distribution $\varepsilon_i \sim N(0, \sigma^2)$ with mean 0 and variance either constant or different for each laboratory;

✓ or more commonly, an 'outlier-contaminated normal' distribution consisting of a mixture of a normal distribution with a wider distribution representing the population of erroneous results.

4.3 General Approached for the Evaluation of Performance

4.3.1 Three different general approaches for evaluating performance in a proficiency testing scheme.

Performance evaluated by comparison:

- a) with externally derived criteria
- b) with other participants
- c) with claimed measurement uncertainty

Might have a mix of consensus statistics and reference criteria

- Reference mean and consensus σ_{pt} ; or consensus mean and reference σ_{pt}
- For c), the AV is typically an appropriate RV (difficult to do with a consensus assigned value)

5 Statistical Design of PT schemes

- 5.1 Introduction to the statistical design of PT schemes
- 5.2 Basis of a statistical design
- 5.3 Considerations for the statistical distribution of results
- 5.4 Considerations for small numbers of participants
- 5.5 Guidelines for choosing the reporting format

5.1 Introduction to the Statistical Design

- PT does not generally evaluate lab bias or precision (but could if that is an objective)
- Evaluates fitness of a result as it would be submitted to a customer
  - ✓Based on difference from the best estimate of "correct"
- Examination over several rounds can indicate bias and poor precision

5.1 Introduction to the Statistical Design

PT does not generally evaluate lab bias or precision (but could if that is an objective)

Evaluates fitness of a result as it would be submitted to a customer

- ✓ Based on difference from the best estimate of “correct”

Examination over several rounds can indicate bias and poor precision

5.2 Basis of Statistical Design

- Design must be appropriate for the stated objectives for the scheme.
- Objectives and sources of error: inputs to statistical design
- Quantitative or qualitative data
  - ✓Quantitative: interval or relative/ratio scale
  - ✓Nominal (categorical) / Ordinal scale
- Statistical assumptions
- Nature of errors
- Expected number of results

5.2 Basis of Statistical Design

- Design considerations for common objectives:

If participants results are compared:	The design will require to consider:
against a pre-determined RV and within limits that are specified before the round begins	a method for obtaining an externally defined RV, a method of setting limits, and a scoring method
with combined results from a group in the same round, and limits that are specified before the round begins	how the AV will be determined from the combined results as well as methods for setting limits and scoring
with combined results from a group in the same round, and limits determined by the variability of participant results	the calculation of an AV and an appropriate measure of dispersion as well as the method of scoring
with the AV, using the participant's own measurement uncertainty	how the AV and its uncertainty are to be obtained and how participant measurement uncertainties are to be used in scoring

5.4 Considerations for Small Numbers

- ISO/IEC 17043 requires consideration of what to do with fewer results than expected
  - ✓ IUPAC/CITAC : use CRMs
  - ✓ EA-4/21: Three scenarios, some guidance
- Minimum number depends on several factors
  - ✓ Other solutions can be found, as well
- Annex D provides further guidance on small numbers of results

5.5 Guidelines for Report Format

- Provider could ask specified format, but should request results are generated and reported the same as for customers
- If replicate results are requested, record all
  - ✓ Not just the mean or SD
- Have design consideration for “<” and “>”
- Rounding error should be negligible
- If participants can report different formats, need to take that into consideration

6 Initial review of proficiency testing items and results

- 6.1 Homogeneity and stability of PT items
- 6.2 Different measurement methods
- 6.3 Blunder removal
- 6.4 Visual review of data
- 6.5 Robust statistical methods
- 6.6 Outlier techniques for individual results

6.1 Homogeneity and stability of PT items

- Three alternatives:
  - ✓ Experimental studies as in Annex B
  - ✓ Use of experience on “closely similar” items in previous rounds
  - ✓ Assess participant results in current round- compare with previous rounds, compare SDPA
- For calibration PTs: Assure stability or take into account drift
  - ✓ Multiple shipments of the same artefact?
- Need for same or different assigned values must be considered in the design
  - ✓ Design could allow flexibility

6.2 Different measurement methods

- Should normally have the same assigned value for all methods that have the same measurand
  - ✓ Not always possible
- Need for same or different assigned values must be considered in the design
  - ✓ *Design could allow flexibility*

6.3 Blunders

- Remove blunders prior to data analysis
  - ✓Based on technical judgment and experience
  - ✓They are usually easily identified
  - ✓Would be treated separately (i.e. contact the participant)
  - ✓Possible to correct some of them, but according to a relevant policy and procedure
  - ✓Can affect robust techniques and outlier detection routines
- When in doubt, do not discard
  - ✓Robust techniques will minimize the effect

6.4 Visual Review of Data

- Arrange for visual review of data. Conducted by a person with technical competence.
- Expect unimodal and symmetric for most techniques.
- Look for bimodal, asymmetric, or a large set of statistical outliers (minor modes).
  - ✓ Bimodal: Different methods, contaminated samples, poor instructions
  - ✓ Histograms
  - ✓ Kernel density plots
- Might have different procedure for first time PT than for well established schemes

6.5 Robust Statistical Methods

Robust techniques preferred to outlier removal.

Better to retain all results that were not obvious blunders

Most techniques are based (in the first step) on the median and the range of the central 50% of results

- ✓Simple estimators: Median, scaled Median Absolute Deviation ($MADe$) and normalized IQR ($nIQR$).
- Algorithm A (and Algorithm S for precision).
- ✓Qn and Q methods for estimating SD.

6.5 Robust Statistical Methods

Robust techniques pre
to outlier remova

Transforms the original data to provide alternative estimators of mean and standard deviation for near-normal data and is most useful where the expected proportion of outliers is below 20%.

Most techniques are based (in the first step) on the median and the range of the central 50% of results

- ✓Simple estimators: Median, scaled Median Absolute Deviation ($MADe$) and normalized IQR ($nIQR$).
- Algorithm A (and Algorithm S for precision).
- ✓Qn and Q methods for estimating SD.

6.5 Robust Statistical Methods

Robust techniques preferred to outlier removal.

Better to retain all results that

For situations where a large proportion (>20%) of results can be discrepant, or where data cannot be reliably reviewed by experts

Most techniques are ba
and the range of the central 50% of results

- ✓Simple estimators: Median, scaled Median Absolute Deviation ($MADe$) and normalized IQR ($nIQR$).
- Algorithm A (and Algorithm S for precision).
- ✓Qn and Q methods for estimating SD.

6.6 Outlier techniques for individual results

- Can be useful to support visual review for blunders, but not optimal for extreme values
  - ✓ Assumptions underlying the test should be demonstrated to be appropriate and apply sufficiently
- Rejection strategies are permitted when robust methods are not applicable.
- If a result is removed, it should still be evaluated according to criteria used for all participants.

7 The assigned value and its standard uncertainty

7.1 Five different alternatives for determining the assigned value.

7.2 Uncertainty of the assigned value.

- ✓ A measurement is incomplete without its uncertainty

7.3-7.7 Different approaches that are allowed

7.8 Comparison of the assigned value with a reference value. Be aware that:

- ✓ A consensus value might be biased
- ✓ A reference value might be unachievable

7 T

- Alternative methods may be used if they have a sound statistical basis and the method is described in the plan for the scheme.
  - ✓ Regardless of the method chosen, it must be checked for every round
- The method used must be fully described to participants in every report (or referenced)

7.2 Uncertainty of the assigned value:

✓ A measurement is incomplete without its uncertainty

7.3-7.7 Different approaches that are allowed

7.8 Comparison of the assigned value with a reference value. Be aware that:

✓ A consensus value might be biased

✓ A reference value might be unachievable

7.2 Determining the uncertainty of the assigned value

- Model for the assigned value:

$$x_{pt} = x_{char} + \delta_{hom} + \delta_{trans} + \delta_{stab}$$

- x_{pt} : assigned value;
- x_{char} : the property value obtained from the characterization (determination of assigned value);
- δ_{nom} :error due to the difference between proficiency test items;
- δ_{trans} : error due to instability under transport conditions;
- δ_{stab} : error due to instability during the period of proficiency testing.

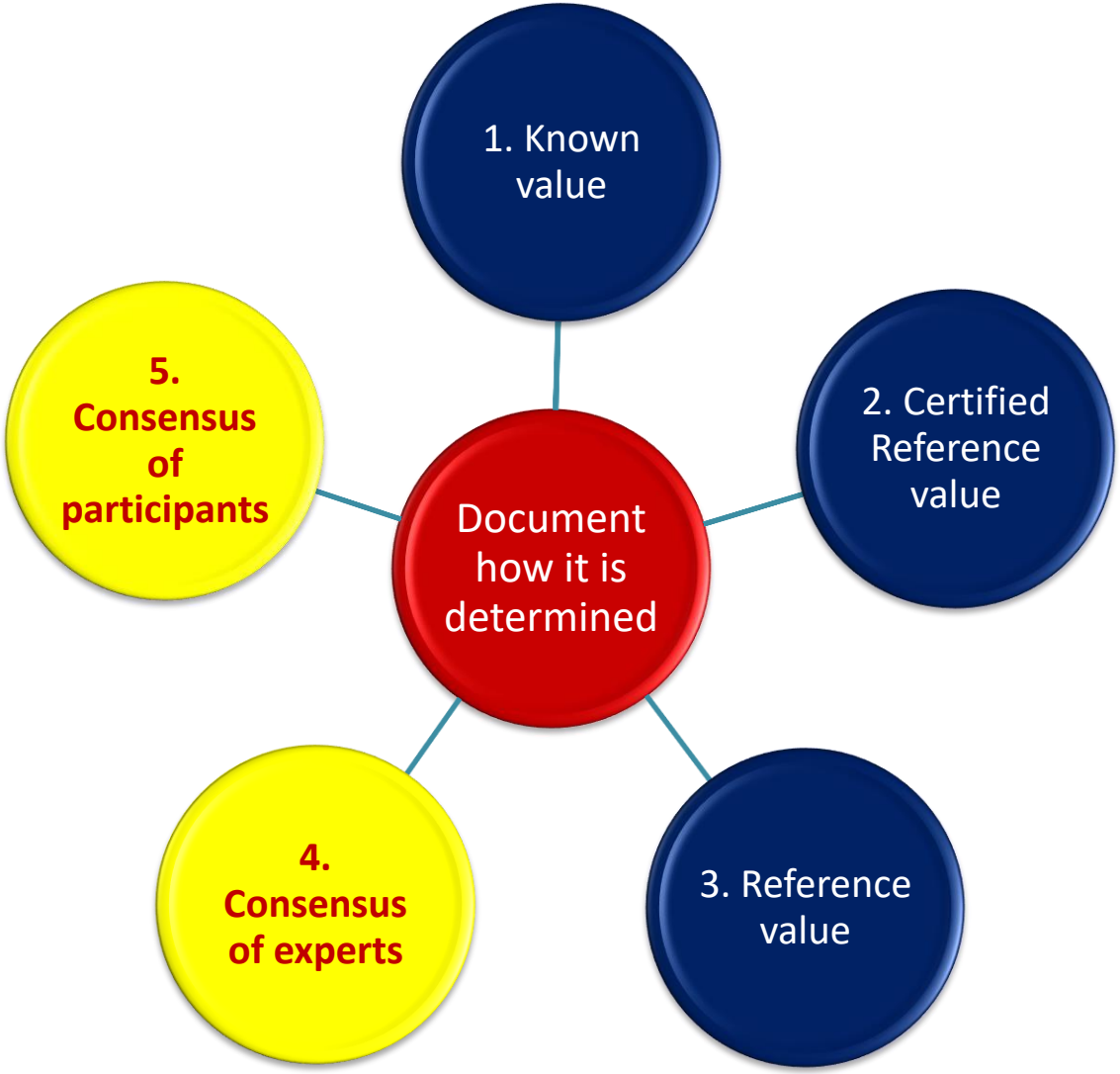
7.2 Determining the uncertainty of the assigned value

- Uncertainty: Reference to GUM and ISO Guide 35

$$u(x_{pt}) = \sqrt{u_{char}^2 + u_{hom}^2 + u_{trans}^2 + u_{stab}^2}$$

- Some components can reasonably be expected to be zero, based on experience.
- Changes related to instability or incurred in transport are expected to be negligible compared to the standard deviation for proficiency assessment
- **Bias in assigned value is not accounted for**.

Determination of assigned value and its uncertainty



7.3 Formulation

Mixing constituents in specified proportions, or by adding a specified proportion of a constituent to a base material.

Assigned value by calculation from the masses of properties used

- **Advantages**:

✓ Individual samples may be prepared

✓ The proportion of the constituents or of the addition has to be determined: there is then no need to prepare a bulk quantity and ensure that it is homogeneous.

7.3 Formulation

- **Concerns**:

✓ The base material must be effectively free of the added constituent, or its proportion is accurately known.

✓ The constituents are mixed together homogeneously (when required)

✓ All significant sources of error are identified

✓ There is no adverse interaction between the constituents and the matrix.

✓ The behaviour of PT items containing added material is similar to customer samples that are routinely tested.

7.3 Formulation

Concerns:

- The base material must be effectively free of the added constituent, or its proportion is accurately known.
- The constituents are mixed together homogeneously (when required).
- All significant sources of error are identified.
- There is no adverse interaction between the constituents and the matrix.
- The behaviour of PT items containing added material is similar to customer samples that are routinely tested.

7.4 Certified reference material

The test material used is a CRM.

Assigned value: x_{CRM}

- The standard uncertainty is derived from the information on uncertainty provided on the certificate for the CRM.

- **Advantages:**

✓ The certified value(s) and associated uncertainty can be used directly.

✓ Quick and simple to implement, and usually provides a value independent of the participant results.

✓ Appropriate traceability for the reference value is also automatically provided (by definition).

7.4 Certified reference material

Concerns:

- CRMs are not usually available in sufficient amounts.
- Cost!
- They may be easily recognizable by the participants.
- Often processed quite heavily to ensure long-term stability.

7.5 Results from one laboratory

- 7.5.1 Single laboratory using a reference method (such a primary one)
  - ✓The reference method used should be completely described and understood.
  - ✓Complete uncertainty statement and appropriate documented metrological traceability.
  - ✓The reference method should be commutable for all measurement methods used by participants.
- 7.5.2 Value obtained by calibration against a CRM
  - ✓Value from result and difference, with uncertainty from the result and the difference

7.5 Res

- 7.5.1 Single laborat
primary one)
  - ✓ The reference m
and understood
  - ✓ Complete uncer
documented me
  - ✓ The reference method should be commutable for all measurement methods used by participants.
- 7.5.2 Value obtained by calibration against a CRM
  - ✓ Value from result and difference, with uncertainty from the result and the difference

$$x_{pt} = x_{CRM} + \bar{d} \qquad u(x_{pt}) = \sqrt{u_{CRM}^2 + u_d^2}$$

x_{CRM} is the assigned value for the CRM
x_{pt} is the assigned value for the PT item
d_i is the difference between the average results for the PT item and the CRM on the i^{th} samples
$\bar{d}$ is the average of the differences d_i

7.5 Results from one laboratory

- 7.5.3 Check for metrological compatibility of results from before scheme and after it.

When a reference value is assigned before and after a round of a sequential proficiency testing scheme, the difference between the values shall be less than two times the uncertainty of that difference.

If not, the PTP may choose to use an average of the measurements as the assigned value, with the appropriate uncertainty.

If the results are not metrologically compatible, the reason should be investigated

7.6 Consensus value from expert laboratories

- Using a design for an interlaboratory study for characterization of CRMs, as described in ISO Guide 35
  - ✓ Each participant must provide their uncertainty
  - ✓ PTP must have a procedure to combine uncertainties

If expert labs provide:

- single results and no uncertainty, follow procedures in clause 7.7 (This also applies if there is evidence that some uncertainties are not correctly determined
- more than one result each, the PTP shall establish an alternative method, statistically valid

7.7 Consensus value from participants results

Specific techniques described in Annex C

- Careful application of techniques in clauses 6.2-6.6 to assure that adequate agreement exists and assumptions are demonstrated to be reasonable
- May wish to use a subset of participants
- Can use other calculation methods with sound statistical basis

Advantages

- No additional measurements needed
- May be necessary with a standardized operationally-defined method

7.7 Consensus value from participants results

- **Concerns:**

Insufficient agreement among the participants.

The consensus value may include unknown bias (faulty methodology) which is not reflected in the standard uncertainty of the assigned value

the consensus value could be biased due to the effect of bias in methods that are used to determine the assigned value.

It may be difficult to determine the metrological traceability of the consensus value. The PTP must have complete information about the calibration standards used and control of other relevant method conditions by all of the participants contributing to the consensus value.

7.7 Consensus value from participants results

Uncertainty of characterization from the method used.

- For some robust methods:

$$u(x_{pt}) = 1{,}25 \times \frac{s^*}{\sqrt{p}}$$

- s* is the robust standard deviation of the results. (“result” for a participant is the average of all their measurements on the PT item.)
- *p* = number of participants
- 1,25: Quite high factor, because PT results might be not strictly normally distributed. In some cases lower values can be justified

7.8 Compare assigned value with independent reference value

- When consensus value is used as x_{pt}, then PTP should try to obtain independent and reliable reference value (formulation, expert, etc), denoted x_{ref}
- When reference value is x_{pt}, then should compare with consensus mean.
- Calculate $x_{diff} = (x_{ref} - x_{pt})$
- The standard uncertainty of the difference is

$$u_{diff} = \sqrt{u^2(x_{ref}) + u^2(x_{pt})}$$

- **Criterion for acceptance: $| x_{diff} | < 2u_{diff}$**

7.8 Compare assigned value with independent reference value

- **If not, investigate the reason-** Possible reasons:

- bias in the reference measurement method
- a common bias in the results of the participants
- failure to appreciate the limitations of the measurement method when using the formulation method described in 7.3
- bias in the results of the “experts” when using the approaches in sections 7.5 or 7.6
- the comparison value and assigned value are not traceable to the same metrological reference

8. Determination of criteria for evaluation of performance

- 8.1 Approaches for determining evaluation criteria
- 8.2 By perception of experts
- 8.3 By experience from previous rounds of a proficiency testing scheme
- 8.4 By use of a general model
- 8.5 Repeatability and reproducibility SD from a collaborative study of precision
- 8.6 From data obtained in the same round of a proficiency testing scheme

8.1 Approaches for determining evaluation criteria

- Basic approach is to compare participant result with x_{pt} and compare the difference with an allowance for measurement error
  - ✓ Using a standardized performance statistic

    z score, z' score, zeta ζ score, En,
  - ✓ Using a defined criterion (e.g., regulation)

    D and D%, or "within limits" / "not within limits"
  - ✓ Using a participant's claim for uncertainty, combined with $u(x_{pt})$
- Can be useful to have standardized scope to compare across rounds
- If a regulatory requirement or a fitness for purpose goal is given as a standard deviation it may be used directly as $\sigma_{pt.}$

Standard deviation for proficiency assessment

σ_{pt}: is a parameter that is used to provide a scaling for the laboratory deviations from the assigned value.

σ_{pt} does not represent a general idea of how laboratories are performing, but how they **should** be performing to fulfill their commitment to their clients.

8.2 Perception of experts

- Allowance for error can be determined by technical experts, accreditation bodies, or regulatory bodies.
  - ✓ Can be expressed as SDPA, σ_{pt}
  - ✓ Can be expressed as Maximum Permissible Error: δ_E
- If criterion for acceptable performance is **z < 3.0**, then

$$\delta_E = 3\sigma_{pt} \quad \text{and} \quad \sigma_{pt} = \delta_E / 3$$

8.3 Experience with previous rounds of PT

When a PTP has experience over several rounds with similar PT items, measurands and methods, then σ_{pt} can be anticipated (Annex E.8)

- **Several advantages:**
  - ✓ Evaluations based on reasonable criteria
  - ✓ Criteria will not vary from round to round due to random error or changing participant base
  - ✓ Criteria will not vary by PT provider
- Previous round data need to be checked for consistency and perhaps for performance by competent participants (not all participants)

8.4 Use of a general model

- Can use a general model for reproducibility σ_R to be used as σ_{pt}
- Results must be reasonable (don't use without check)
- Only one example is described (Horwitz Curve modified by Thompson), but others might be possible.
- With c = mass fraction of chemical species to be determined and $0 \leq c \leq 1$:

✓ $\sigma_R = 0.22c$	when $c < 1.2 \times 10^{-7}$
✓ $\sigma_R = 0.2\ c^{0.8495}$	when $1.2 \times 10^{-7} \leq c \leq 0.138$
✓ $\sigma_R = 0.1\ c^{0.5}$	when $c > 0.138$

8.4 Use of a general model-Horwitz curve

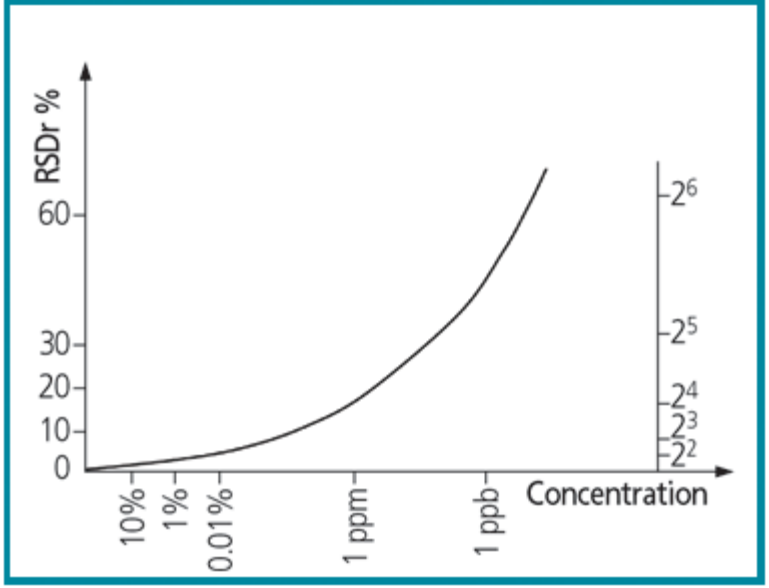
$$\sigma_R = 0{,}02c^{0{,}8495}$$

The results of collaborative trials seem to obey this law regardless

- ✓ of the nature of the analyte
- ✓ the test material
- ✓ The physical principle

underlying the measurement method.

Figure 1: Relative reproducibility standard deviation RSD_R as a function of concentration.



So far, nobody has managed to explain the strange empirical exponent from basic principles, although several people have made conjectures.

8.5 Use σ_r and σ_R from previous collaborative precision study

- If a previous collaborative study followed principles of ISO 5725-2 (standardized method), repeatability σ_r and reproducibility σ_R estimates can be used to determine σ_{pt}
- With m the number of replicate values:

$$\sigma_{pt} = \sqrt{\sigma_R^2 - \sigma_r^2(1 - 1/m)}$$

- When m = 1 then $\sigma_{pt} = \sigma_R$

8.6 From data obtained in the same round of PT

- Consensus SD from results of participants in the same round can be used as σ_{pt}
  - ✓Should use robust technique from Annex C
- Caution about SDPA being inappropriate for evaluation of performance – can be too large or can be too small for “fitness for use” :
  - ✓**Should have limits:**
    - for smallest SDPA to be used
    - for largest SDPA that can be used
    - on range of values that can be evaluated as “acceptable”, based on fitness for use (for example, a minimum acceptable recovery of a formulated level)

8.6 From data obtained in the same round of PT

- **Advantages**
  - ✓ Easy, commonly used, may be the only feasible approach
- **Disadvantages**
  - ✓ SD can vary widely from round to round
  - ✓ Can be unreliable with small number of labs
  - ✓ Can lead to approximately same proportion of “action signals” (unacceptable)
  - ✓ There is no useful interpretation of suitability of a result based on intended use (shows only that a lab agrees with others in the scheme). This can be important when the measurand involves health or safety.

8.7 Monitoring Interlaboratory agreement

PTP should use a procedure to monitor interlaboratory agreement (robust SD) of participants across rounds

- Useful for PTP to show benefits of participation
- Useful to check suitability of statistical methods
- Useful to check for unexpected increase or decrease in agreement

9. Calculation of performance statistics

- 9.1 General considerations
- 9.2 Limiting the uncertainty of the assigned value
- 9.3 Estimates of deviation (measurement error)
- 9.4-9.7 Scores
- 9. 8 Evaluation of participant measurement uncertainties
- 9.9 Combined performance scores

9.1 General considerations

- Statistics used for determining performance shall be consistent with the objectives for the PT scheme.
- Performance scores should be easily reviewed across measurand levels and different rounds of a PT scheme
- Results should be reviewed and determined to be consistent with the assumptions in the design
  - ✓ Approximate normality (unimodal, symmetric)
  - ✓ No signs of instability or inhomogeneity
  - ✓ Signs of mixed population

9.2 Limiting the uncertainty of the assigned value

If $u(x_{pt})$ is large relative to the performance criterion, there is risk of adverse evaluations due to factors other than poor measurement technique

If $u(x_{pt}) < 0.3\sigma_{pt}$ or $u(x_{pt}) < 0.1\delta_E$:

$u(x_{pt})$ may be considered to be negligible and need not be included in the interpretation of the results of the round of the PT scheme

9.2 Limiting the uncertainty of the assigned value

- This can be a difficult criterion. If it is exceeded:

Use a different assigned value

Accommodate the uncertainty in the evaluation (z', ζ, E_n)

Report different x_{pt} for different methods

Inform participants of the potential impacts on evaluations

If none of the above apply, do not evaluate performance

9.3 Estimates of deviation (measurement error)

- All performance measures start with measurement error – deviation from the assigned value, expressed in units or %
- This deviation can be compared to a criterion δ_E expressed in units or as a percentage of x_{pt} :

$$D_i = (x_i - x_{pt}) < \delta_E$$

$$D_i\,\% = (x_i - x_{pt})/\, x_{pt} < \delta_E \qquad \text{if } \delta_E \text{ is a \%}$$

  - ✓ If $-\delta_E < D < \delta_E$ then the performance is considered to be 'acceptable'
  - ✓ δ_E can be a regulatory limit, analytical goal, expert opinion, etc.
- Main disadvantage: D is not standardized

9.4 z score

- The most commonly used statistic for PTs

$$z_i = \frac{(x_i - x_{pt})}{\sigma_{pt}}$$

- if |z| < 2, the performance is satisfactory;
- if 2 <|z| < 3, the performance is questionable -“warning signal”
- if |z| > 3, the performance is unsatisfactory –“action signal”
- Assumption: the individual z scores have a Gaussian or normal distribution with a mean of zero and a standard deviation of one. On this basis analytical results can be described as 'well-behaved’.

9.5 z' score

- A slight variation to z score, to allow consideration of uncertainty of x_{pt}

$$z_i' = \frac{x_i - x_{pt}}{\sqrt{\sigma_{pt}^2 + u^2(x_{pt})}}$$

- When criterion in clause 9.2 is met ($u(x_{pt}) < 0.3\sigma_{pt}$):

  $0.96 < z'/z < 1.00$ (uncertainty of the assigned value is negligible) and z' and z scores are almost identical.

- z' score is evaluated same as z score

9.6 ζ score

- If an objective of the scheme is to evaluate a result compared to the participant's claim for uncertainty, ζ (zeta) is used:

$$\zeta_i = \frac{x_i - x_{pt}}{\sqrt{u^2(x_i) + u^2(x_{pt})}}$$

- Generally ζ can be interpreted the same as z score (2-warning signal, 3-action signal)
- They can be used in conjunction with *z* scores

9.7 E_n scores

- E_n (Error, normalized) is a conventional score for PT in calibration, but can be applied anywhere

$$E_{ni} = \frac{x_i - x_{pt}}{\sqrt{U^2(x_i) + U^2(x_{pt})}}$$

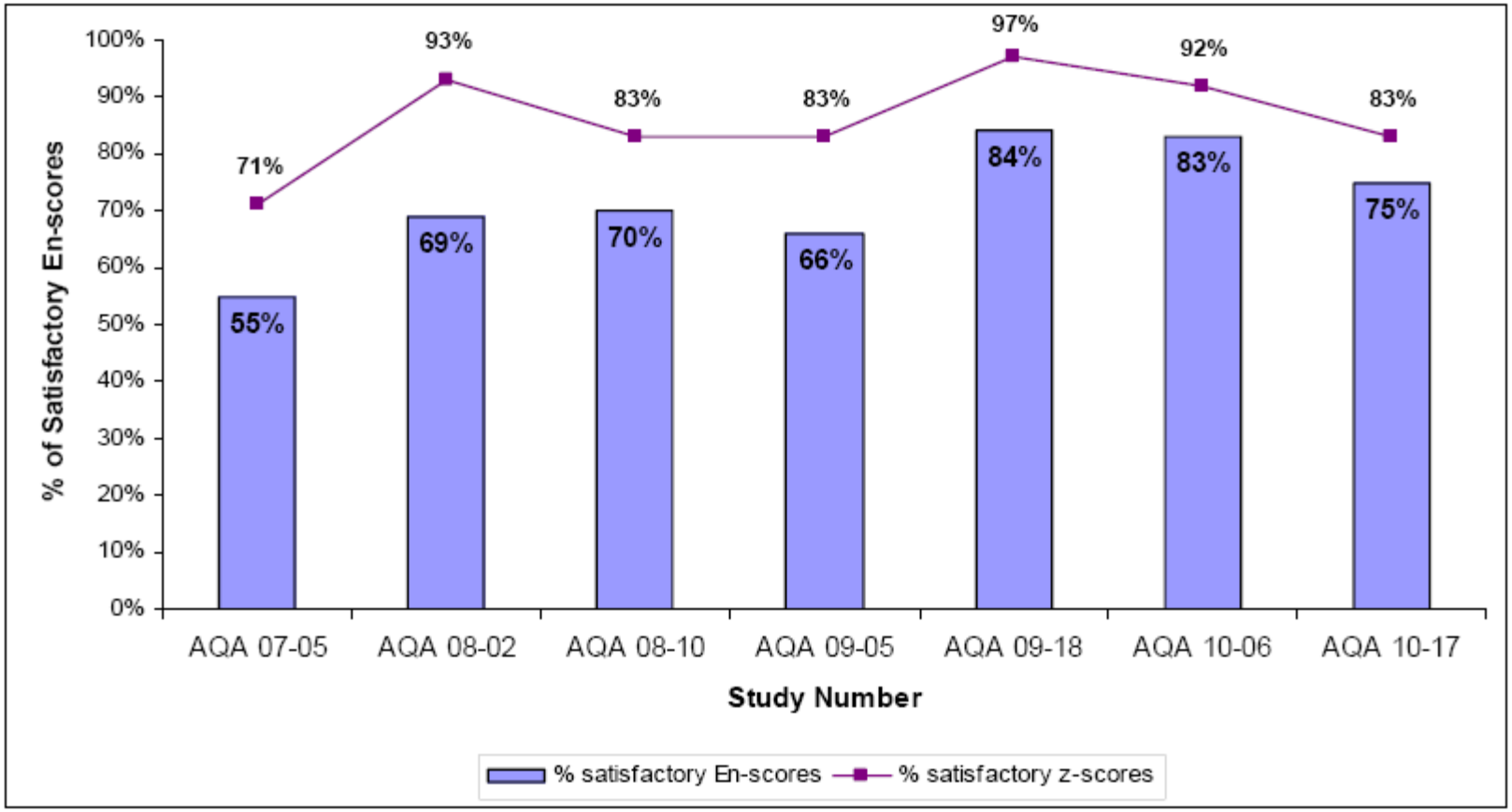
- $U(x_i)$ is the expanded uncertainty of a participant's result x_i
- if $|E_n| < 1.0$, the performance is satisfactory;
- if $|E_n| > 1.0$ the performance is unsatisfactory.

E_n and ζ scores

- Scores that evaluate performance compared to claimed uncertainty must be interpreted with caution, because some participants might not calculate uncertainty correctly (GUM), or report them correctly.
- A large uncertainty leads to lower scores; small uncertainty leads to higher scores
- Often useful to report E_n and ζ in addition to a conventional score (e.g., z z' D D%)

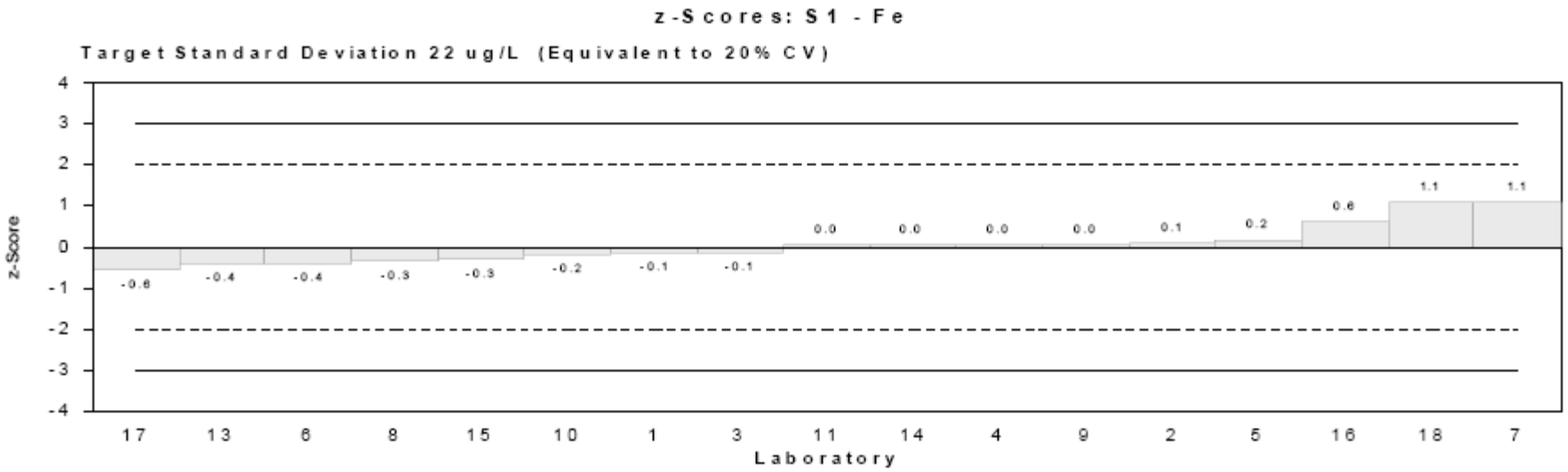
Performance statistics and criteria

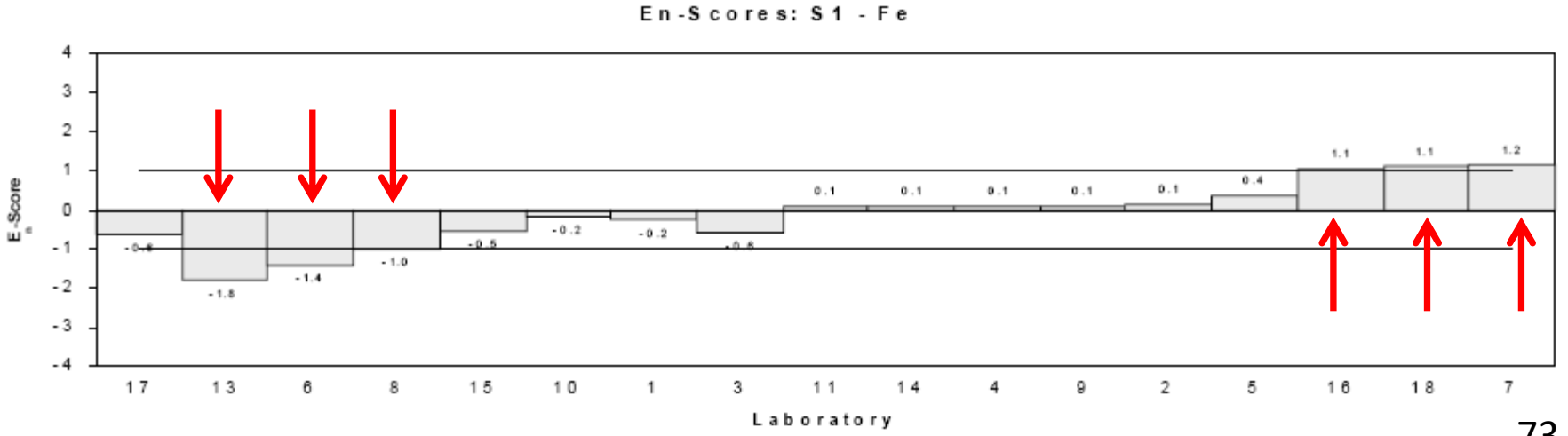
Satisfactory z and E_n scores
Trace metals in water



|z| score was satisfactory

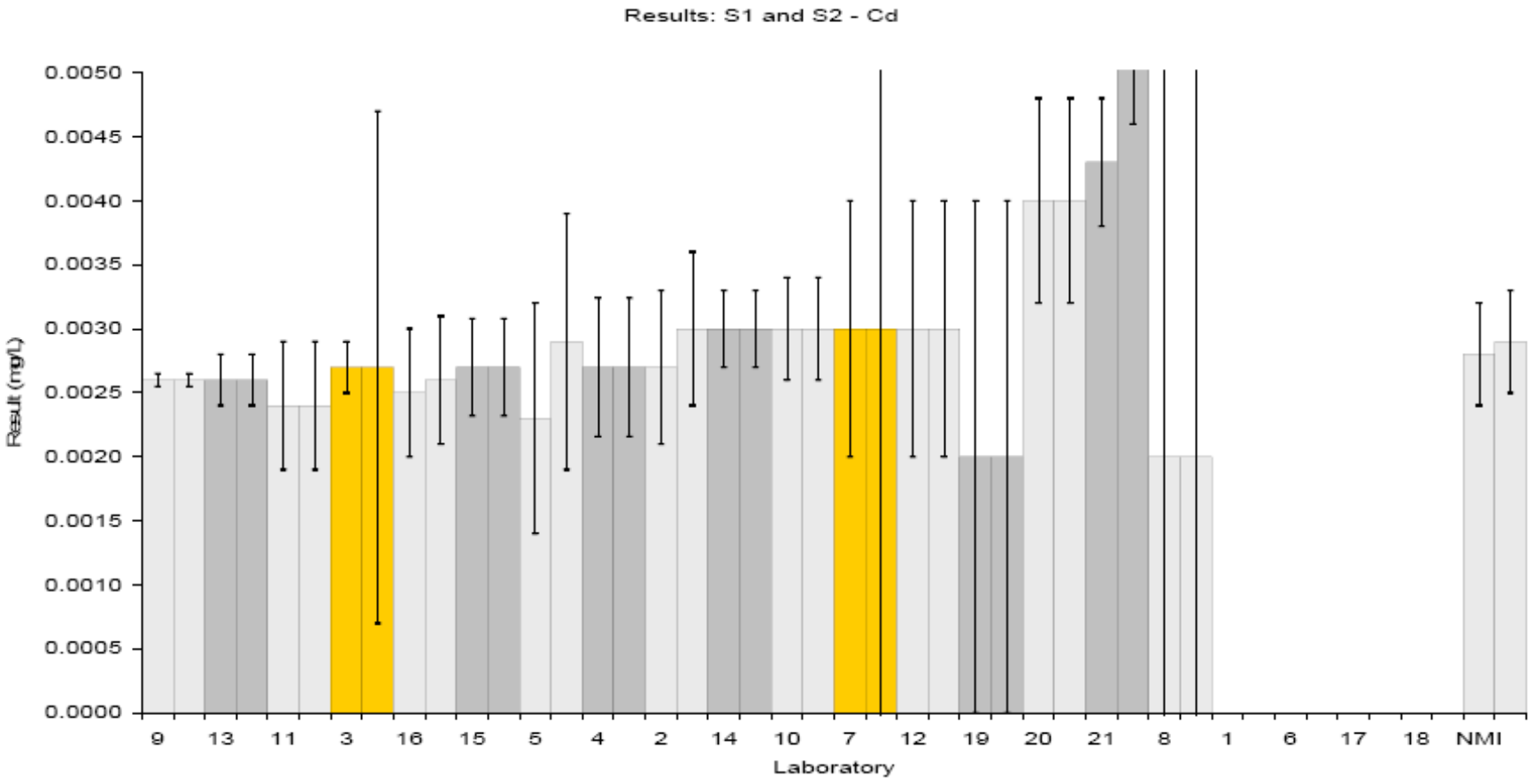
|En| value indicates that some results are not in agreement with the assigned values when the uncertainties are taken into consideration.





Performance statistics and criteria

Lab 3 and 7 have reported different uncertainties for identical samples.



9.8 Evaluation of participant measurement uncertainties

- Proficiency testing is a useful tool for showing differences between laboratory measurements. This includes estimates of measurement uncertainty.
- Many laboratories could benefit from seeing that their estimates are much different than those of other laboratories using the same method
- **ISO 13528 recommends informational ‘flags’ of questionable uncertainties**

9.8 Evaluation of participant measurement uncertainties

- Reasonable criteria for MU:

$$u_{min} = u_{ref}$$

$$u_{max} = 1.5\ \sigma_{pt}$$

- ✓ $u_{min} \leq u_{lab} \leq u_{max}$ then u_{lab} may be OK
- ✓ $u_{lab} < u_{min}$ then u_{lab} may be too small
- ✓ $u_{lab} > u_{max}$ then u_{lab} may be too large

9.8 Evaluation of participant measurement uncertainties

- Reasonable criteria for MU:

$$u_{min} = u_{ref}$$
$$u_{max} = 1.5\ \sigma_{pt}$$

} To be explained to the participants

- ✓ $u_{min} \leq u_{lab} \leq u_{max}$ then u_{lab} may be OK
- ✓ $u_{lab} < u_{min}$ then u_{lab} may be too small
- ✓ $u_{lab} > u_{max}$ then u_{lab} may be too large

9.9 Combined performance scores

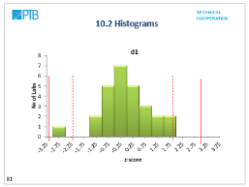
- Some PT schemes combine scores for different PT items in the same round (e.g., average z scores)
  - ✓ Useful when there are many samples or measurands
- Sometimes this is part of the design:
  - ✓ e.g., evaluation of repeatability, systematic error or linearity
- Combined scores have unknown statistical properties, so should be used with caution
- Graphical techniques are preferred

10 Graphical techniques

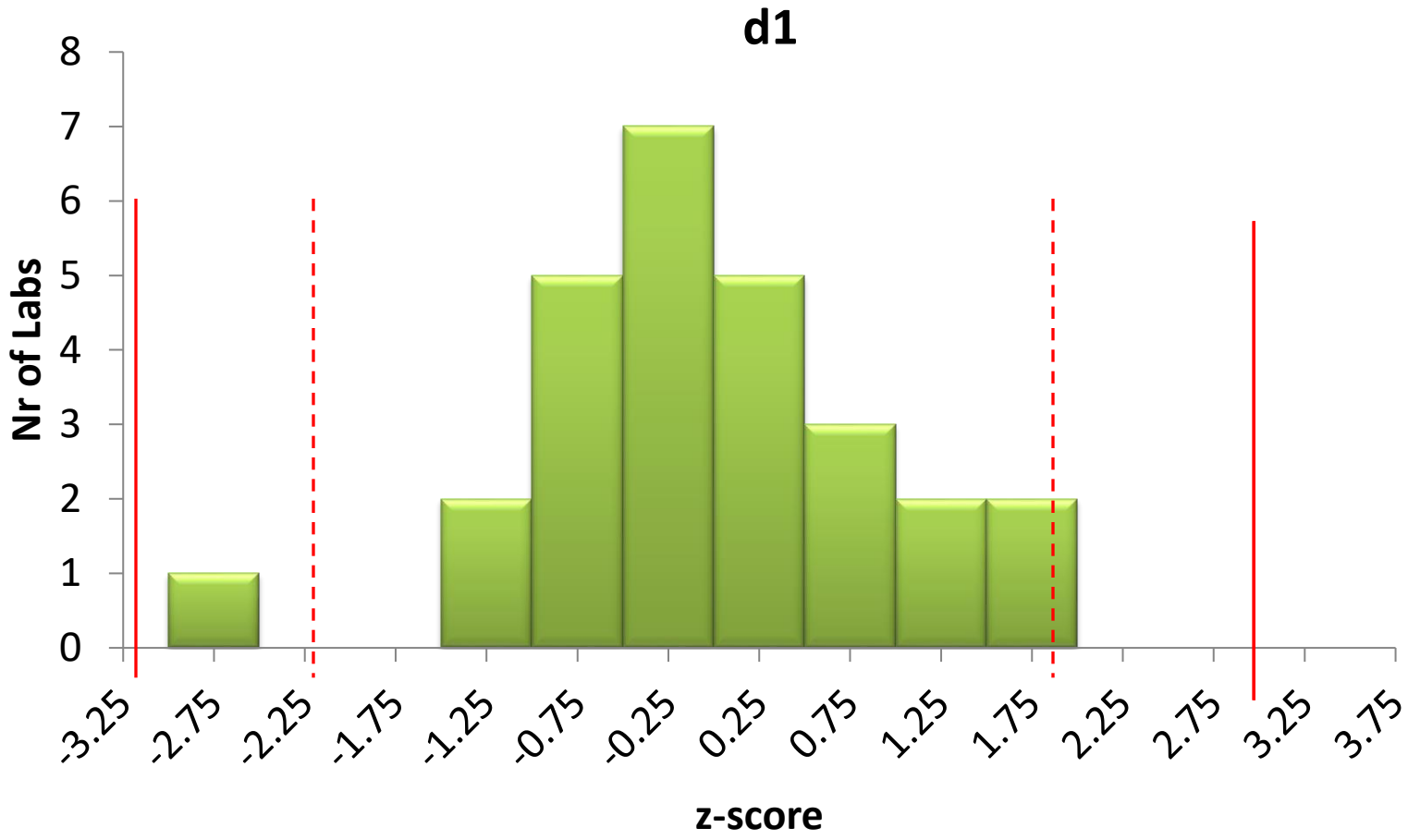
- Graphs are encouraged, and are required in ISO/IEC 17043 (reports)
- Histograms are most common, for preliminary data checks and for reporting
  - ✓Kernal density plots are similar and easy
- Other techniques
  - ✓Bar-plots of standardized performance scores
  - ✓Youden Plot
  - ✓Plots of repeatability standard deviations
  - ✓Graphical methods for combining performance scores over several rounds of a proficiency testing scheme

10.2 Histograms

- z-score vs. Nr of labs
- Participants:
  - Identify the position of their scores and assess their performance and the need to investigate their methods (how exceptional they are).
- Coordinator:
  - How frequently the participants fail to satisfy the PT assessment criterion.
  - Outside the ±3 limits, probably method fault.



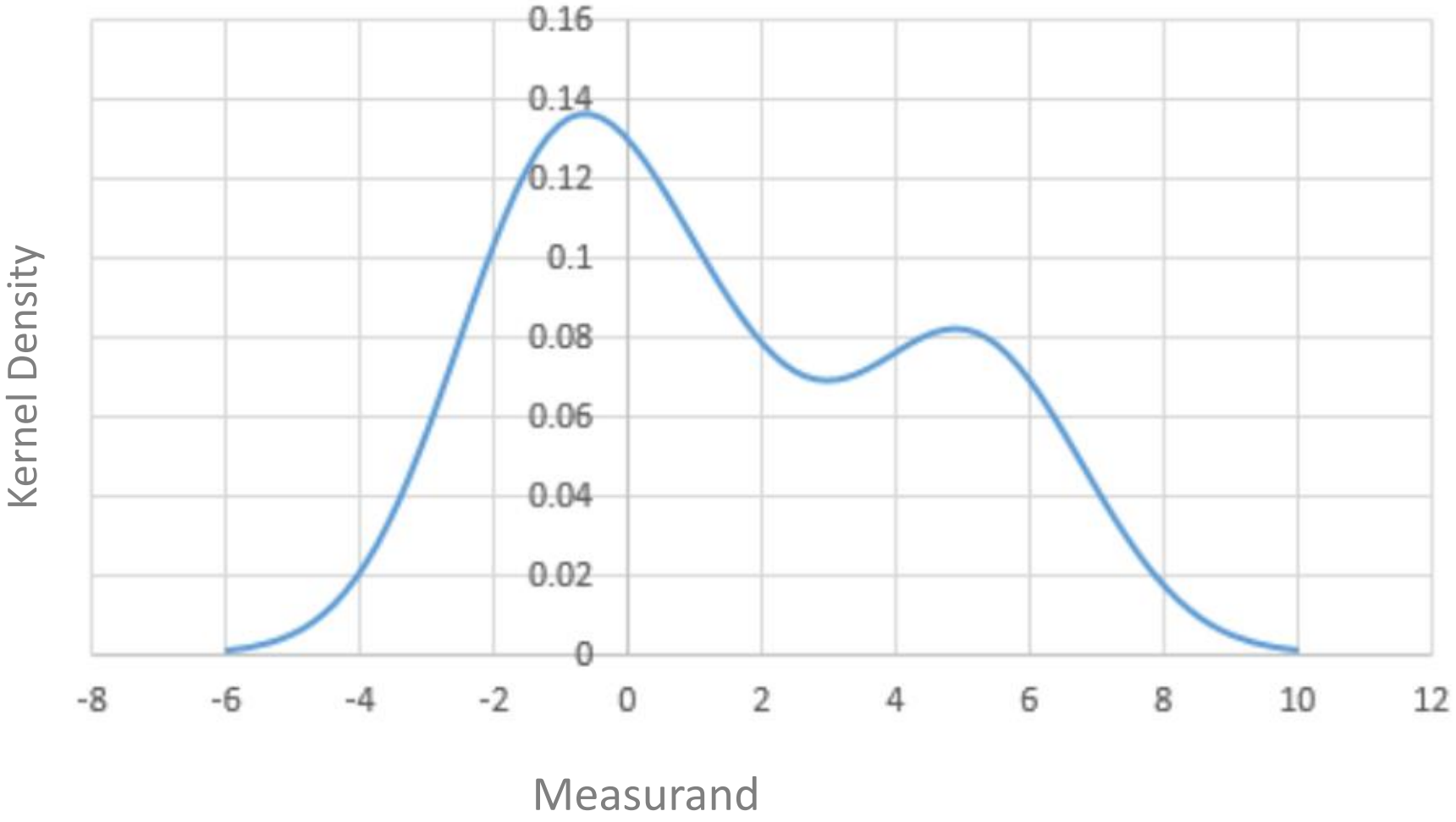
10.2 Histograms



10.3 Kernel Density Plots

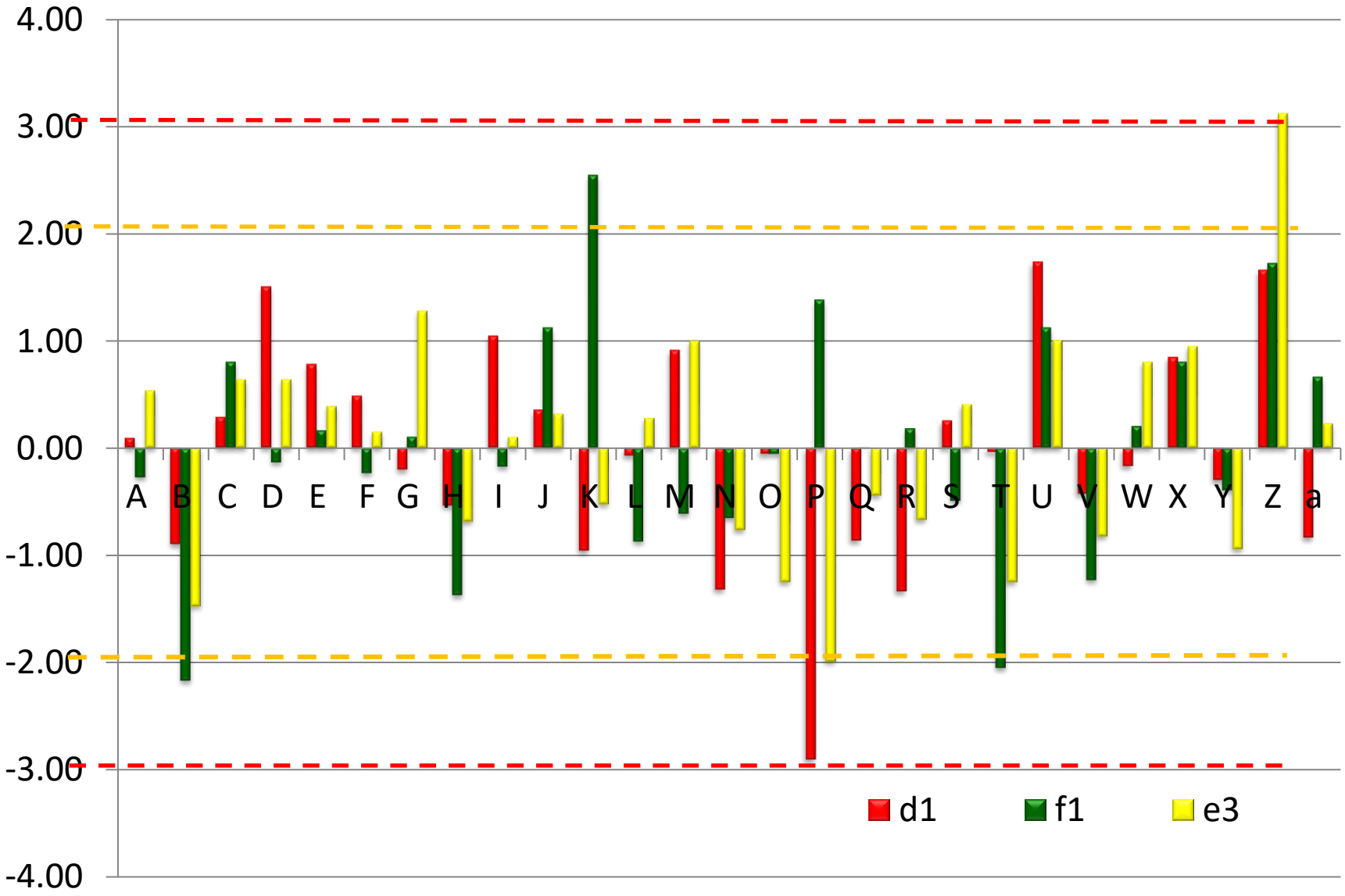
- Provide a smooth curve describing the general shape of the distribution of a data set.
- Each data point is replaced by a specified distribution (typically normal), centred on the point and with a standard deviation σ_k ('bandwidth').
- These distributions are added together and the resulting distribution, scaled to have a unit area, gives a 'density estimate' which can be plotted as a smooth curve

10.3 Kernel Density Plots



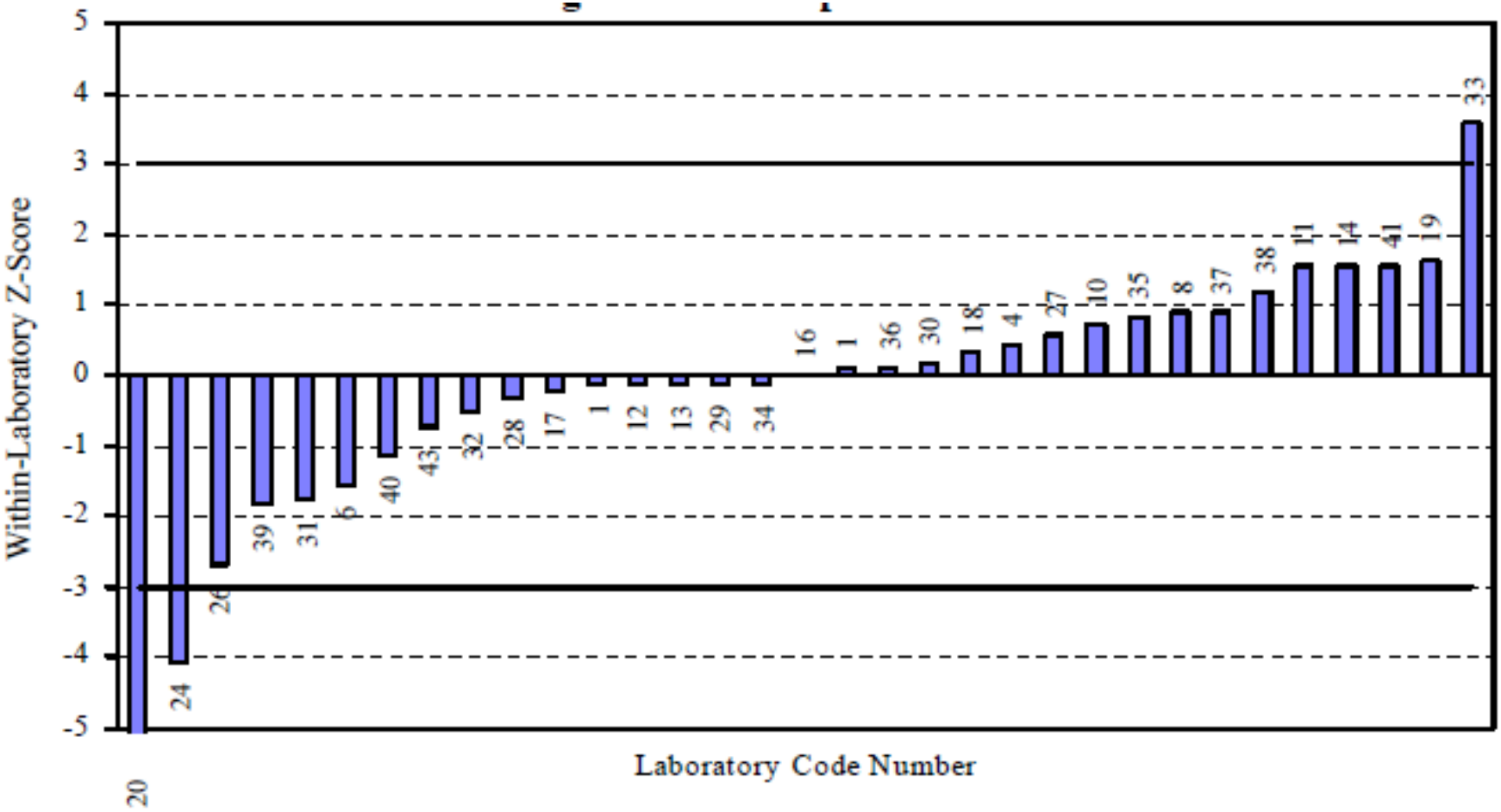
10.4 Bar Plots

- Bar Plots- z-scores of each participant are grouped together
- A suitable method of presenting the z-scores for a number of similar characteristics.
- Identify any common feature in the z-scores of a participant



Bar-chart of z-scores (4,0 to –4,0) for one round of a proficiency test

10.4 Bar Plots



10.5 Youden Plot

- They are generated for pairs of results for duplicate samples, and for duplicate results requested from the same sample.
- They are presented to highlight laboratory systematic differences. They are based on a plot of each laboratory's pair of results, represented by a black spot •.

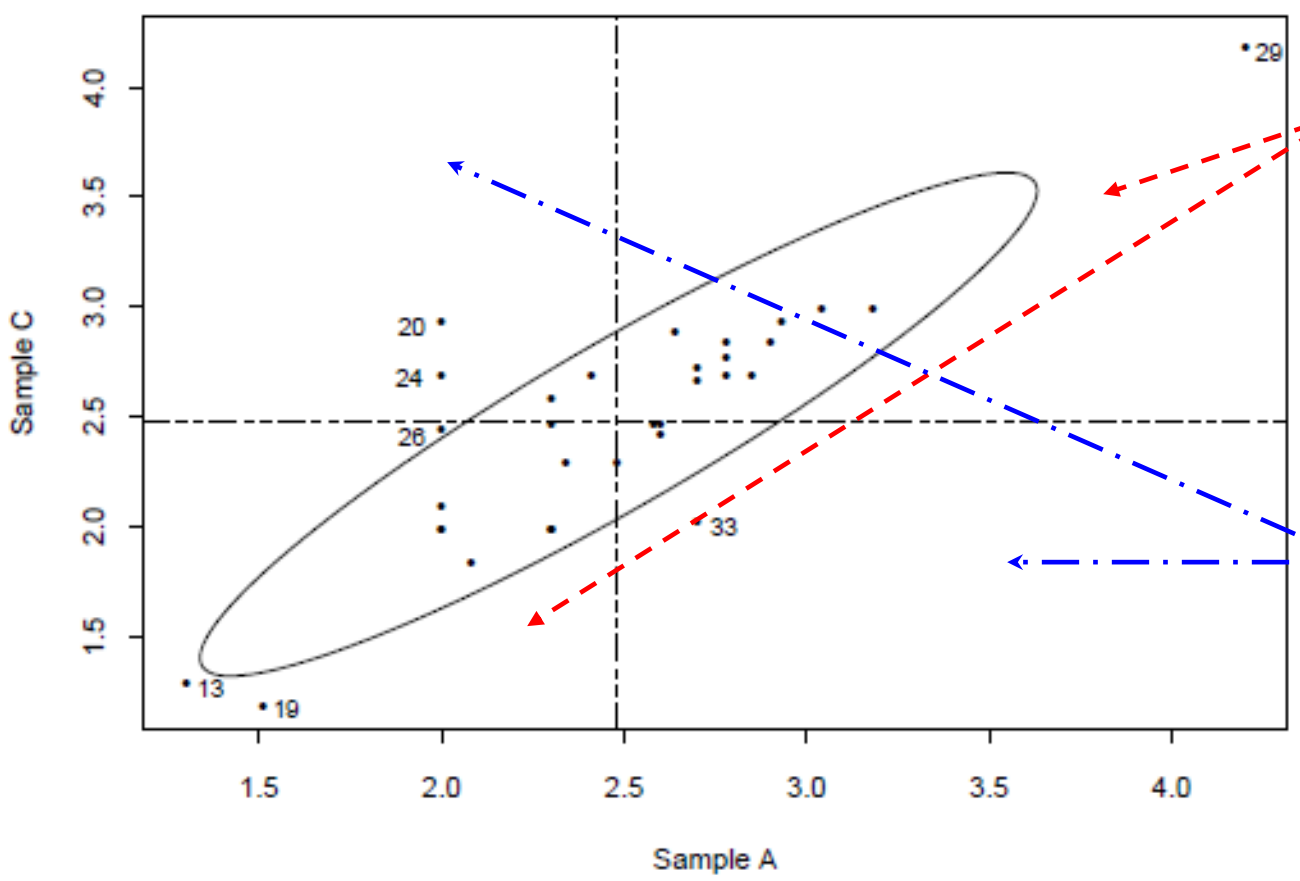
10.5 Youden Plot

Points well-separated from the rest of the data:

- The laboratory is not following the test method correctly (results subject to bias- a point far out along the major axis of the ellipse).
- The laboratory suffers a large variation from time to time in the level of its results.
- Points far away from the major axis represent participants whose repeatability is poor.

10.5 Youden Plot- Interpretation

95% confidence ellipse with dashed lines indicating median values for each of the samples



significant systematic error components

random error components

10.5 Youden Plot- Interpretation

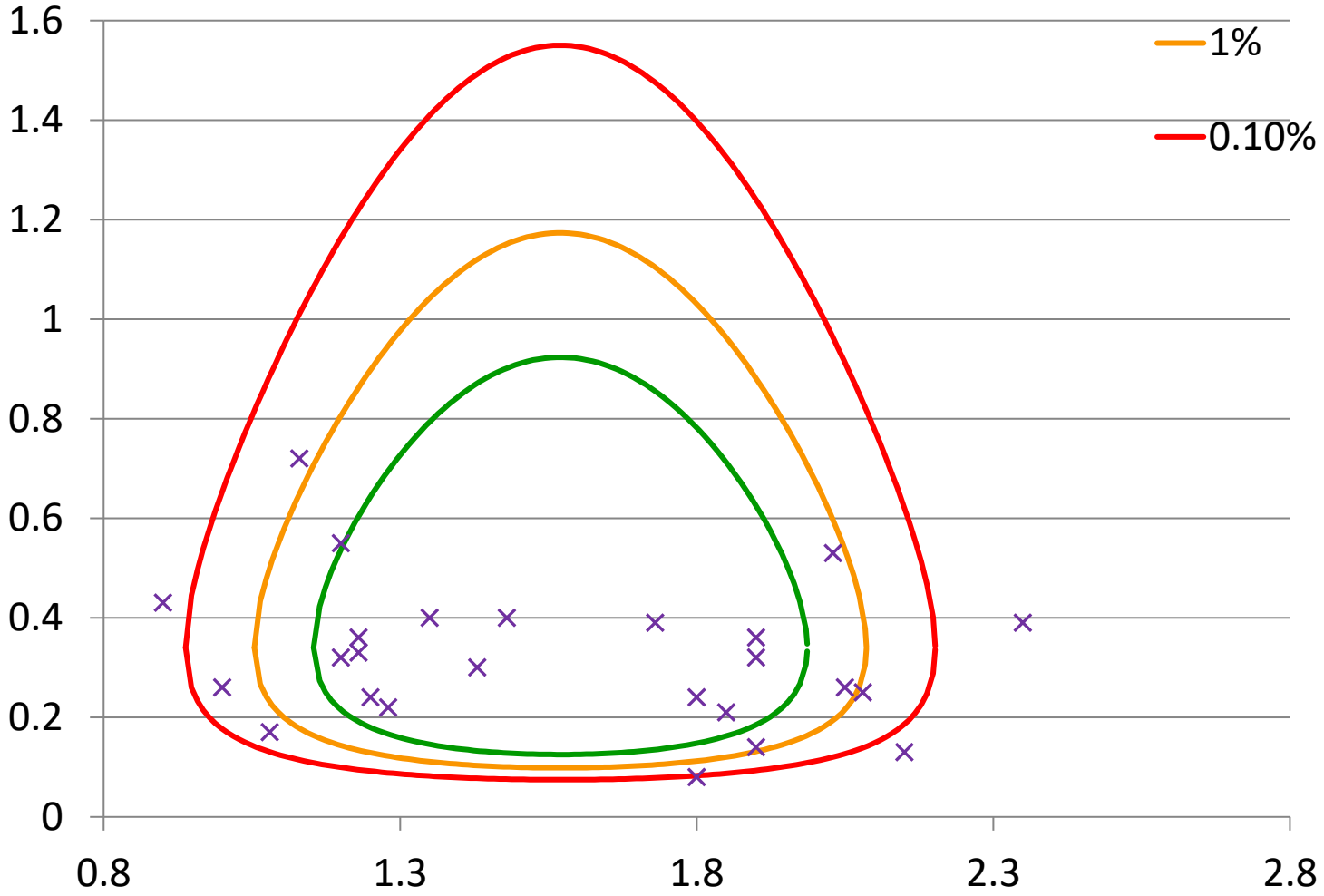
Youden plot is applied to identify labs with unusual high systematic error as well as labs have unusually high random errors.

- When a lab has both large total error, the data point will be far away from the center. They are the first group of labs that should be closely investigated.
- It may happen that a material is less sensitive to different environment than the other. When this happens, the data points will tend to be parallel to X- axis or Y-axis, with a large variation due to a material. This can also be quickly identified.

10.6 Repeatability Standard Deviations Plots

- They are used to identify any laboratories whose average and SD are unusual.
- Plot the within-laboratory SD for each lab against the corresponding average

10.6 Repeatability Standard Deviations Plots



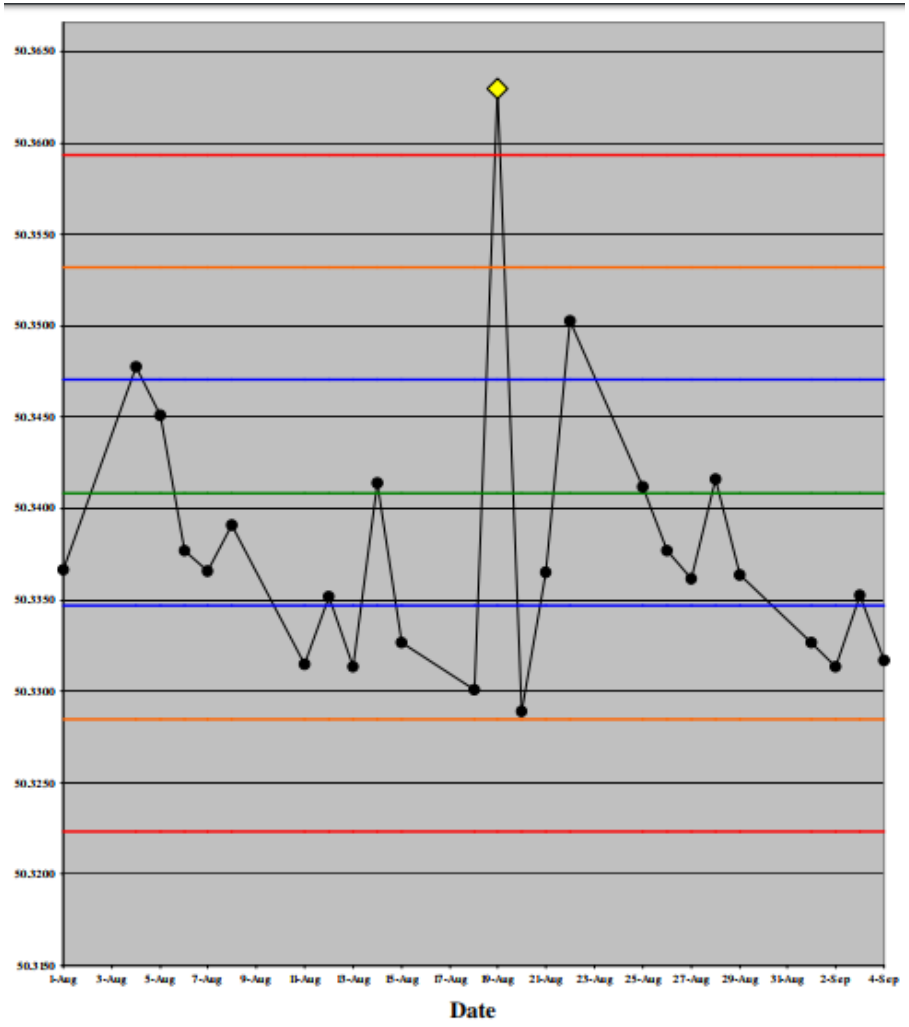
Plot of standard deviations against averages for 25 laboratories

10.8.2 Shewhart control charts

- Effective method of identifying problems that cause large erratic values of *z* scores.
  - ✓ Standardized scores, (such as *z scores)*, for a participant are plotted as individual points.
  - ✓ Action and warning limits set consistent with the design for the proficiency testing scheme.
  - ✓ When several characteristics are measured in each round, the performance scores for different characteristics may be plotted on the same graph, using different plotting symbols and/or different colours.

10.8.2 Interpretation of Shewhart control charts

- a) a single point falls outside the action limits (± 3,0 for *z scores,* or 100 % for P_A*);*
- b) two out of three successive points outside either warning limit (± 2,0 for *z scores* or 70 % for P_A*);*
- c) six consecutive results either positive or negative.



11 Design and analysis for qualitative schemes

- 11.1 Types of qualitative data
- 11.2 Statistical design
- 11.3 Assigned values for qualitative schemes
- 11.4 Performance evaluation and scoring for qualitative PT schemes

11.1 Types of qualitative data

- ✓Nominal or categorical scale
- ✓Presence or absence (including above or below threshold)
- ✓Ordinal (response has magnitude, but no mathematical relationship between levels)

- Require special consideration for the design, value assignment and performance evaluation

11.2 Statistical design

- Homogeneity
  - ✓Test suitable number of items
  - ✓All results should be the same
- Stability
  - ✓Should not be a factor in identity
  - ✓Concern for presence if not stable
- Performance criterion based on expert judgment, often after review of results
  - ✓Preferred to have a panel of experts, and defined criteria for their agreement

11.3 Assigned values for qualitative schemes

- Assigned value usually determined by:
  - ✓ a) by expert judgement
  - ✓ b) by use of reference materials as PT items
  - ✓ c) from knowledge of the origin or preparation of the PT item(s)
  - ✓ d) using the mode or median of participant results.
  - ✓ Categorical: can use participant mode (most common observation)
  - ✓ Ordinal: can use participant median or mode
- Often need to document origin or source of PT item
- Uncertainty should not be a factor

11.4 Performance evaluation

- Evaluation criteria must meet objectives and be fit for the purpose of the test.
- Criteria are usually determined by expert opinion
  - ✓ Might be individual, based on expert review of each participant's results
- Can have weighted performance score
  - ✓ **"perfect"** **score 0**
  - ✓ **"not perfect, but not bad"** **score 1**
  - ✓ **"bad"** **score 3**

Annexes

Informational annexes

- Annex A: Symbols
- Annex D: Additional guidance on statistical procedures
- Annex E: Illustrative examples

Normative annexes

- Annex B: Homogeneity and stability of PT items
- Annex C: Robust analysis

Annex B.4- Stability Testing

Annex B.4-Stability

- Tested periodically over a range of storage conditions prior to distribution (for held samples, or remedial PT).
- Simulate effects of handling and shipping
- Compare with Homogeneity Test results

Annex B.4-Stability

- If experience or technical reasons show stability can be expected for the time of PT study, then a limited stability study is adequate to show measurands were stable.
- Should check all measurands , unless…(6.1.3)
- Two PT items are adequate if homogeneity is assured, else use >2 items.
- Use more items or replicates if $\sigma_r > 0.5\ \sigma_{pt}$

Annex B.4-Stability

- Simple experiment is to check mean of results on stability measurements ($\bar{y}_2$) versus mean of results from before shipment ($\bar{y}_1$ e.g., homogeneity check)
- Criterion for acceptance:

$$|\bar{y}_1 - \bar{y}_2| \leq 0{,}3\sigma_{pt} \text{ or } |\bar{y}_1 - \bar{y}_2| \leq 0{,}1\delta_E$$

- If criterion is met, instability will not affect evaluations

B.5 Assessment criterion for a stability check

- If criterion is not met, consider if intermediate precision is source of difference $|\bar{y}_1 - \bar{y}_2|$. In such a case:
  - ✓ If possible, use isochronus* stability study, or a different method
  - ✓ Increase $u(x_{pt})$ to include instability
  - ✓ Expand criterion for acceptance:

$$|\bar{y}_1 - \bar{y}_2| \leq 0{,}3\sigma_{pt} + 2\sqrt{u^2(\bar{y}_1) + u^2(\bar{y}_2)}$$

* Based on a storage design where the samples are stored at different temperatures for different time periods

B.5 Assessment criterion for a stability check

- If new criteria are still not met:
  - ✓ Quantify the effect of instability and include it in the evaluation (e.g. z' scores)
  - ✓ Examine production, storage, shipment to see if improvements are possible
  - ✓ Do not evaluate performance

B.6 Transport stability

- PT provider should check the effects of transport, at least initially (newly developed PT schemes).
  - ✓Compare results on shipped items vs. stored items
  - ✓Criterion for acceptance same as in B.5
- Any known effects should be considered in evaluation of performance, and included in $u(x_{pt})$
- If consensus mean and SD are used, then all samples may have same effect, so not an issue

Annex C: Robust analysis

- PT providers need to mitigate the effect of extreme results, because not all participants are competent, and extreme results are always possible.
- These results have a strong effect on consensus statistics
- There are two choices:
  - ✓Remove statistical outliers
  - ✓ Use statistical techniques that are robust to these values
- Robust techniques are preferred

Annex C: Robust analysis

Simple techniques:

- Median for x_{pt}
- $nIQR$ for σ_{pt}
- $MADe$ for σ_{pt}

Conventional

- Algorithm A: for x^* and s^*
- Algorithm S: for s_r

Computationally intense techniques:

- Q_n for σ_{pt}
- Q/Hampel for σ_{pt} and x_{pt}

Annex C.2.1 The median

- Median of a continuous variate is defined as the value where ½ of the observations are below and the other ½ above.
- For a sample of n ordered variables x_1, x_2, ..., x_n, the sample median is denoted as $med(x)$ and calculated as:

$$med(x) = \begin{cases} x_{\{(p+1)/2\}} & for\ odd\ p \\ \dfrac{[x_{\{p/2\}} + x_{\{1+p/2\}}]}{2} & for\ even\ p \end{cases}$$

Annex C.2.2 Scaled median absolute deviation *MADe*

- *MADe(x)* provides an estimate of the population standard deviation for normally distributed data and is highly resistant to outliers.
- Calculations (after sorting the data):

$$d_i = |x_i - med(x)|$$

$$MAD_e(x) = 1{,}483\ med(d)$$

- If 50 % or more of the participant results are the same, then $MADe(x)=0$, and it may be necessary to use other estimators

Annex C.2.3 Normalized interquartile range $nIQR$

- Is a measure of the variability of the results- robust estimator of the SD
- Calculations (after sorting the data):
  - ✓ $IQR = Q3 - Q1$, Where
  - ✓ $Q1$= Is the value below which ¼ of the results lie=(N+1)/4
  - ✓ $Q3$= Is the value above which ¼ of the results lie=3*(N+1)/4

  (N is the total number of results received for a particular test/sample)

$$nIQR = 0.7413 \times IQR$$

- In most cases $Q1$ and $Q3$ are obtained by interpolating between the data values.

Annex C.3 Robust analysis: Algorithm A

Yields robust average x^* and robust standard deviation s^*

n items of data: $x_1, x_2, \ldots, x_p$

1) Initial values: x^* = median of x_i (i = 1, 2, ..., p)

s^* = 1,483 median of $x_i - x^*$ (i = 1, 2, ..., p)

2) Update the values of x^* and s^* :

2a) $\delta = 1{,}5s^*$

For each x_i calculate:

$$x_i^* = \begin{cases} x^* - \delta & if\ x_i < x^* - \delta \\ x^* + \delta & if\ x_i > x^* + \delta \\ x_i & otherwise \end{cases}$$

2b)

$$x^* = \frac{\sum_{i=1}^{p} x_i^*}{p}$$

$$s^* = 1{,}134\sqrt{\frac{\sum \left(x_i^* - x^*\right)^2}{p-1}}$$

Annex C.3 Robust analysis: Algorithm A

Iterative calculations (updating x^* and s^*) until the process converges

Convergence

There is no change from one iteration to the next in the third significant figures of x^* and s^*

Uncertainty of assigned value:

$$u_x = \frac{1{,}25 \times s^*}{\sqrt{p}}$$

Annex D Additional Guidance

- **Few participants, or comparison groups with small numbers of participants, such as when participants are grouped and scored by method**
- Ideally:
  - ✓ The assigned value is determined using a metrologically valid procedure, independent of the participants, such as by formulation or from a reference laboratory.
  - ✓ Performance evaluation criteria is based on external criteria, such as expert judgement or criteria based on fitness for purpose.

Bilateral comparison, or measurement audit

Annex D Additional Guidance

- **Few participants, but the assigned value or the dispersion, or both, need to be derived from participant results.**
- Identify outliers
  - ✓ Robust statistics are recommended when populations are outlier-contaminated- Not recommended for very small data sets.
  - ✓ Possible scenario: Identify outliers, reject them, calculate mean and SD

Annex D Additional Guidance

- **Few participants, but the assigned value or the dispersion, or both, need to be derived from participant results.**
- Considerations for estimates of location (mean) and dispersion (SD)
  - ✓ Efficiency and breakdown points for robust procedures for the criterion on limiting the uncertainty ($u(x_{pt}) < 0.3\sigma_{pt}$ **or** $u(x_{pt}) < 0.1\delta_E$)
    - Simple mean: p=12
    - Median: p=18
    - Algorithm A: p=12
- Breakdown point is the proportion of values in the data set that can be outliers without the estimate being adversely affected.

Annex D Additional Guidance

- **Few participants, but the assigned value or the dispersion, or both, need to be derived from participant results.**
- Efficiency and breakdown points for robust procedures for resisting to outliers
  - ✓ Breakdown point is the proportion of values in the data set that can be outliers without the estimate being adversely affected.

Annex D Additional Guidance

- Efficiency and breakdown points for robust procedures for resisting to outliers

Statistical estimator	Population parameter estimated	Breakdown Point	Resistance to Minor Modes
Sample mean	Mean	0%	Poor
Sample standard deviation	Standard deviation	0 %	Poor
Median	Mean	50%	Good
$nIQR$	Standard deviation	25%	Moderate
$MADe$	Standard deviation	50%	Moderate - Good
Algorithm A	Mean and Standard deviation	25%	Moderate
Q_n and Q / Hampel	Mean and Standard deviation	50%	Moderate (Very Good for minor modes more distant than 6 s*)

Other Methods for Assigned Value

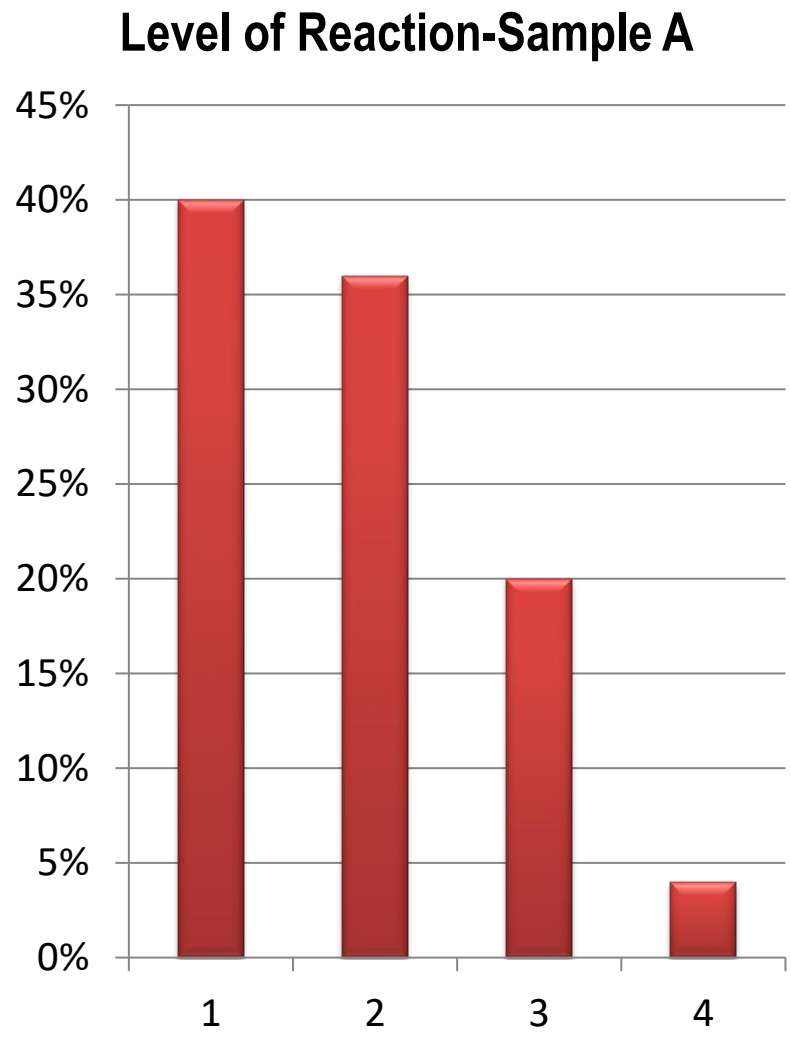
Semi-Quantitative-Example

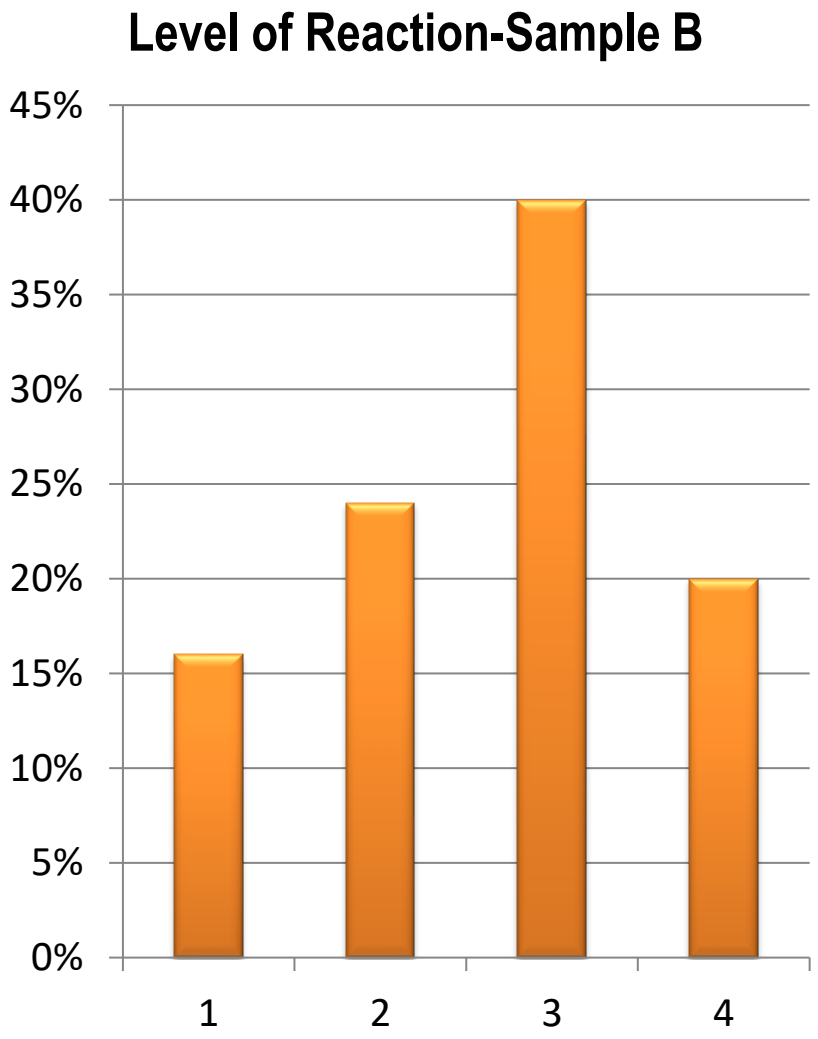
- Measurand: Level of reaction, by category:
  - 1= no reaction, normal
  - 2= mild reaction
  - 3= moderate reaction
  - 4= severe reaction
- 2 PT samples, A and B
- 50 participants

Semi-Quantitative-Example

Sample A	Sample B
1= 20 results (40%)	1= 8 results (16%)
2= 18 results (36%)	2= 12 results (24%)
3= 10 results (20%)	3= 20 results (40%)
4= 2 results (4%)	4= 10 results (20%)

Responses for Samples A and B





Determination of assigned value and its uncertainty

Example 1: One measuring result as the reference value

- The reference value can be the measuring result of the pilot laboratory.
- The pilot laboratory should have the best measuring capability of the participants.
- In this case the pilot laboratory must be an accredited calibration laboratory or an NMI with CMC registration and long-term experience in the area of the measured physical quantity.

Determination of assigned value and its uncertainty

Example 2: NMI´s or calibration laboratories with equivalent measuring capabilities

- In this case the reference value can be the arithmetic mean value.

$$X_R = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad U = 2 \times \sqrt{\frac{1}{n(n-1)}} \times \sqrt{\sum_{i=1}^{n}(X_i - X_R)^2}$$

- x_R arithmetic mean value
- x_i measuring result of a participant
- n number of participants
- U expanded measuring uncertainty *(k = 2)*

Determination of assigned value and its uncertainty

Example 2: NMI´s or calibration laboratories with equivalent measuring capabilities

- Reference value can be the arithmetic mean value.
- Expanded measuring uncertainty (*k =2):*

$$U = 2 \times \sqrt{\frac{1}{n(n-1)}} \times \sqrt{\sum_{i=1}^{n} (X_i - X_R)^2}$$

- X_R arithmetic mean value
- X_i measuring result of a participant
- n number of participants
- U expanded measuring uncertainty *(k = 2)*

Determination of assigned value and its uncertainty- Degree of equivalence

The difference between the value of a participating institute and the reference value is also called “offset”

$$D_i = X_i - \frac{1}{n}\sum_{i=1}^{n} X_i$$

D_i offset
X_i measuring result of a participant
n number of participants
i count index

Determination of assigned value and its uncertainty- Degree of equivalence

The uncertainty of the offset is calculated according to the equation:

$$u^2(D_i) = \frac{1}{n^2}\sum_{i=1}^{n} u^2 + \left(1 - \frac{2}{n}\right) u_i^2$$

D_i offset
u standard measuring uncertainty of the offset
u_j standard measuring uncertainty of a participant
n number of participants
i count index

Determination of assigned value and its uncertainty

Example 3: Use of a weighted mean value

$$X_R = \frac{\sum_{i=1}^{n} \frac{X_i}{U_i^2}}{\sum_{i=1}^{n} \frac{1}{U_i^2}}$$

X_R weighted mean value
X_i measuring result of a participant
U_i expanded measuring uncertainty (k=2) of a participant
n number of participants
i count index

Determination of assigned value and its uncertainty

Example 3: Use of a weighted mean value (cont'd)

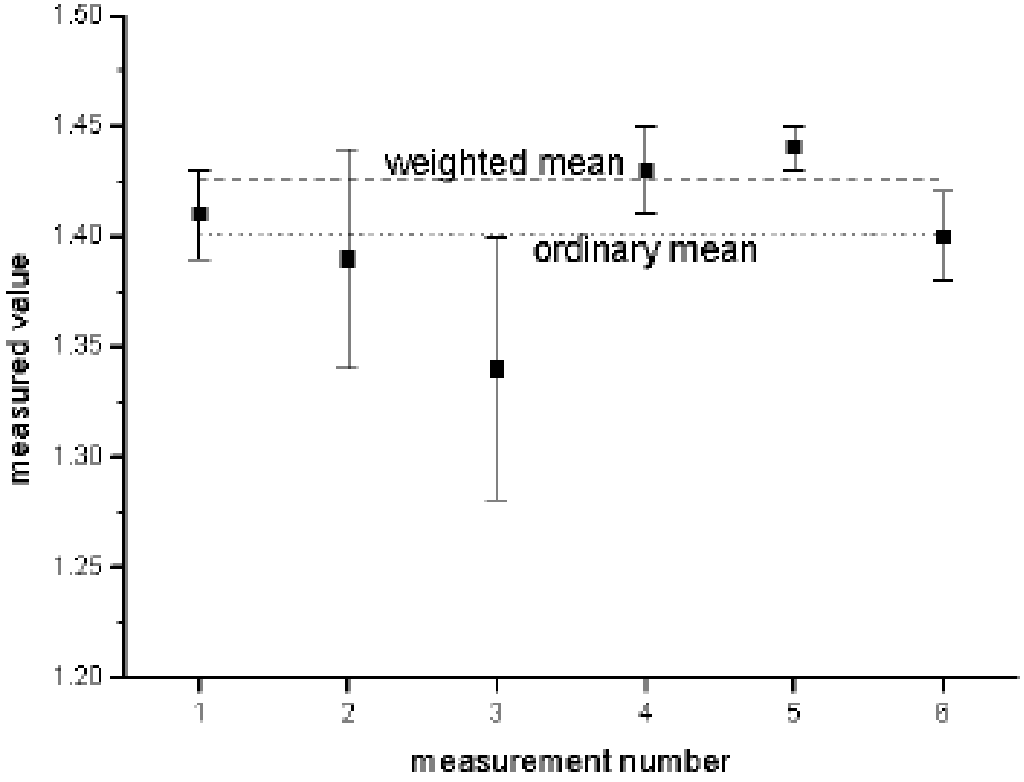
Measuring uncertainty:

$$U_R = 2 \times \sum_{i=1}^{n} \left(\frac{\partial X_R}{\partial X_i} \right)^2 \times u_i^2$$

U_R expanded measuring uncertainty (k=2) of the mean value
X_R weighted mean value
X_i measuring result of a a participant
u_i standard measuring uncertainty of a participant
n number of participants
i count index

Determination of assigned value and its uncertainty

Best estimated value

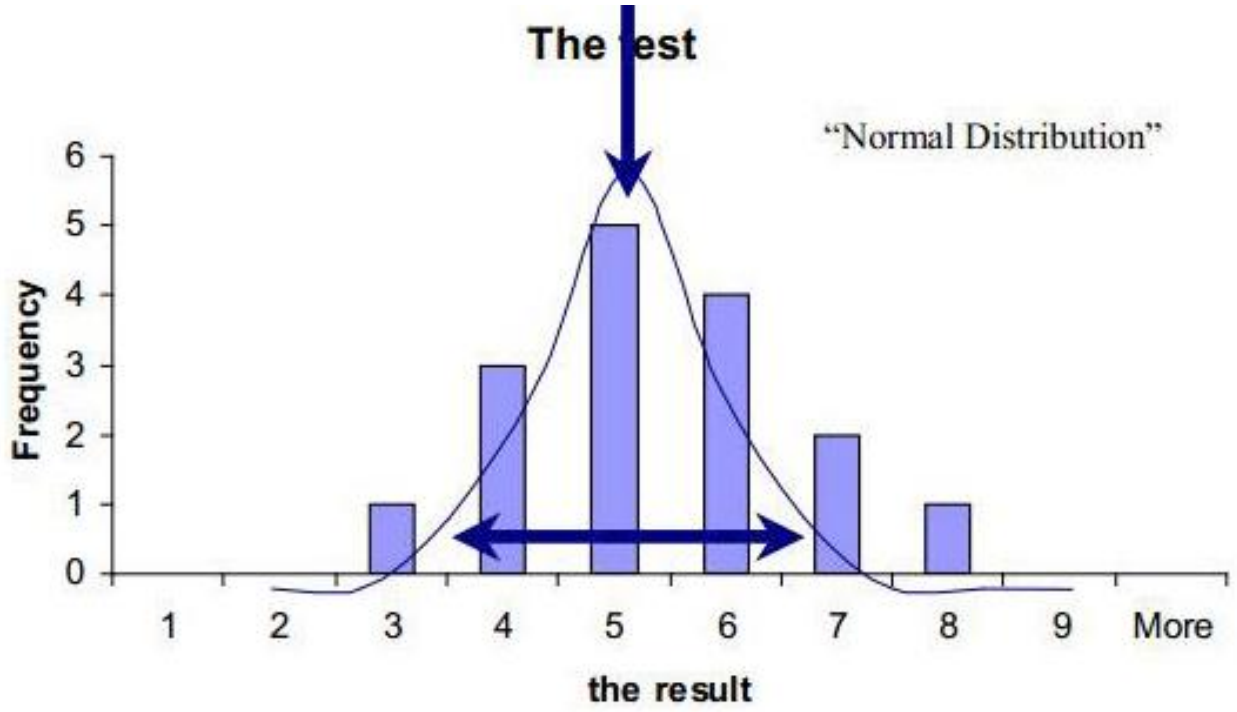


Ordinary mean compared with weighted mean

Assessing the Results

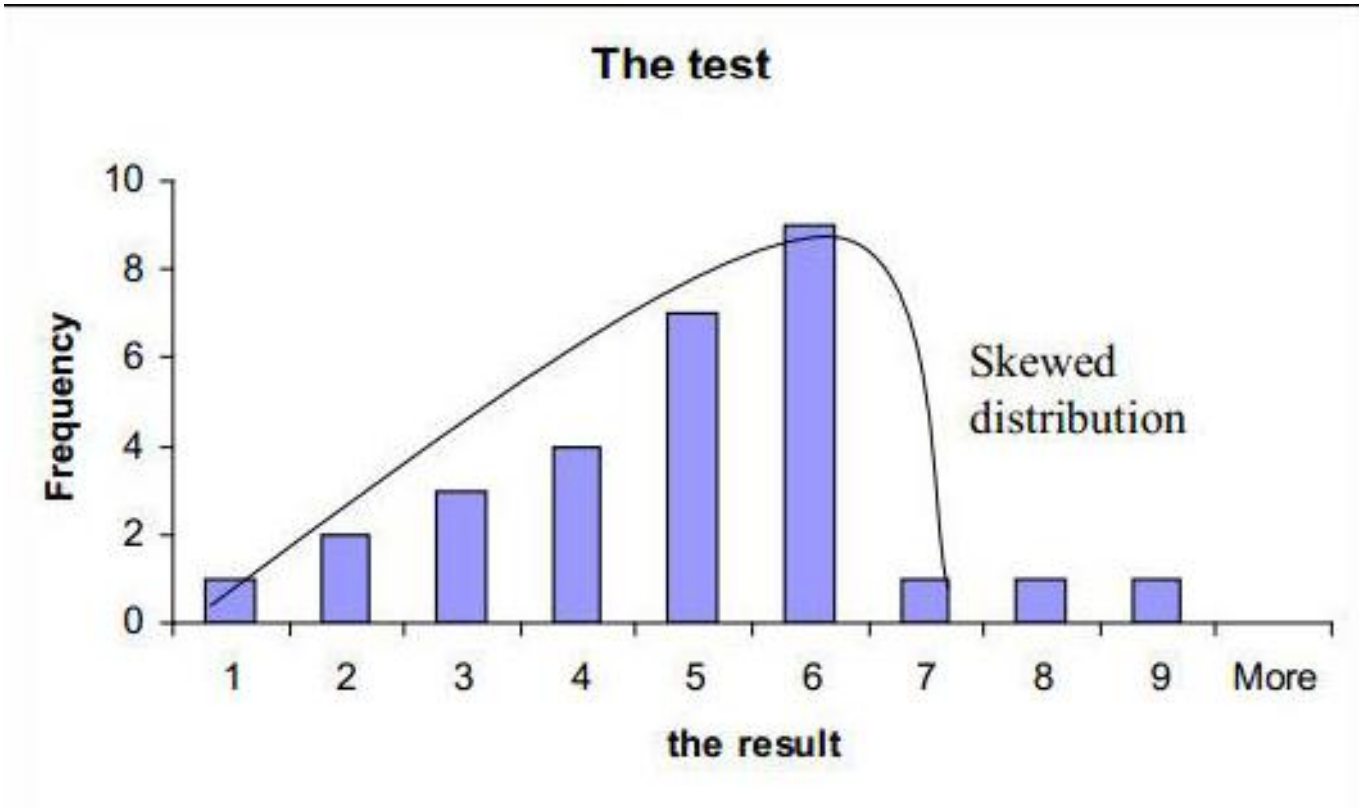
Assessing the Results

All PT stats search for the consensus (or true) result and the distance



Assessing the Results

If the distribution is skewed, perhaps some factor has limited the evaluation of the result (f.i. an aspect of test methodology may not be able to detect readings higher than a certain value)

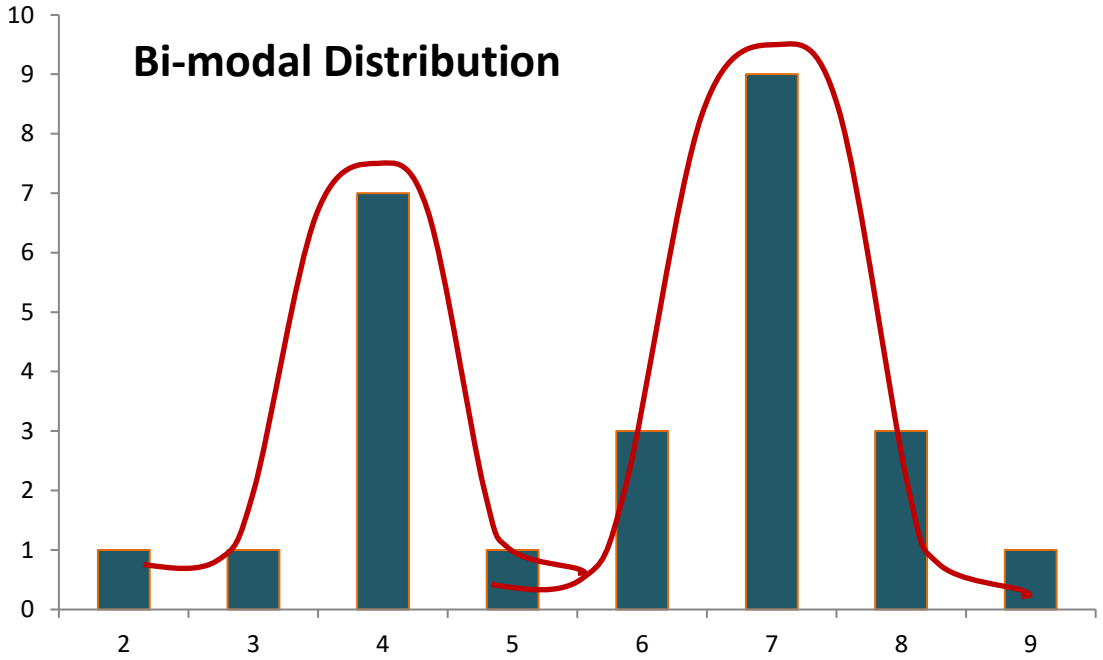


Assessing the Results
Skewed/multimodal distributions

- Skews and extra modes can arise when the participants' results come from two or more inconsistent methods (different bias).
- Rarely, skews can arise when the distribution is truly lognormal (*e.g., in GMO determinations).*

Assessing the Results
Skewed/multimodal distributions

- Perhaps two methods? Contaminated samples or poorly worded instructions?



Other Documents for PT-ILC Statistics

- L. Nielsen, DFM Report 99-R39 " Evaluation of measurement intercomparisons by the method of least squares".
- J. Mueller, "Possible advantages of a robust evaluation of comparisons", J. Res. NIST **105**(4) (2000), 551
- M.G. Cox "The evaluation of key comparison data", Metrologia **39** (2000), 589
- H.S. Nielsen, 2003 NCSL Intern. Workshop & Symposium, "Determining consensus values in interlaboratory comparisons and proficiency testing".
- E. Filipe, XIX IMEKO World Congress (2009) "Laboratories best measurement capability validation".

Thank you for your attention!