3rd European-American Workshop on NDE Reliability

TUTORIAL No 2

How to determine repeatability and reproducibility (R&R)

Dr Damir Markucic

damir.markucic@fsb.hr
dmark@fsb.hr

University of Zagreb, Department of Quality,
Faculty of Mechanical Engineering & Naval Architecture, Croatia

Tutorial topics

1. Introduction
2. Basics, terminology / definitions
3. R&R experiment
   (how to perform, layout, DoE)
4. R&R analysis
   (statistical analysis, examples)
5. Conclusion remarks & discussion

Literature

1. Introduction

Concept of R&R as a measure of precision in measurement laboratories, was introduced & developed during 1978-1983 [2].

It was widely accepted but also criticised, and finally resulted with ISO 5725:1986.

As in any other field, first essential prerequisite is to exchange information, data and knowledge and to understand each other properly – so, convention about definitions is needed. .....

1. Introduction

.....

"Each term must have the same meaning for all of its users; it must therefore at the same time express a well defined concept and not to be in conflict with everyday language" [1].

First, we'll discuss two terms;
accuracy and **precision**;
so we'll come to R&R.

1. Introduction

It is known from common everyday use that when we **repeat** some measurement or want to **reproduce** measurement result in some other occasion, we will almost never get exactly the same result.

We could observe same phenomena if we want to measure some quantity with different **systems !** .....

1. Introduction

..... now, we can carry out simple measurement:

What time is .....

just now !

1. Introduction

..... Is your watch an accurate and/or precise one ?

True time	*My watch*	*Christine's watch*
10:**00**	10:**00**	10:**02**
11:**00**	11:**01**	11:**02**
12:**00**	12:**02**	12:**02**
13:**00**	13:**03**	13:**02**
14:**00**	14:**02**	14:**02**
15:**00**	15:**01**	15:**02**

Which one you more prefer ?

2. Definitions / terminology ...

- Accuracy, trueness, precision
- R&R conditions and limits
- Elements of test system
- Systematic and random errors
- Accepted reference value

Accuracy

designates closeness of agreement between
a test result (*measurement result*)
and the accepted reference value (*true value*) [1,3].

Notes
The term precision should not be used for accuracy [1].

Trueness

designates closeness of agreement between
the average value obtained from
a large series of test result
and an accepted reference value [3].

Notes

The measure of trueness is usually expressed in terms of bias [3].

Trueness has been referred to as “accuracy of the mean”, what is not recommended [3].

Precision

designates variability of results, or
closeness of the agreement between test results of repetitive measurements [3].

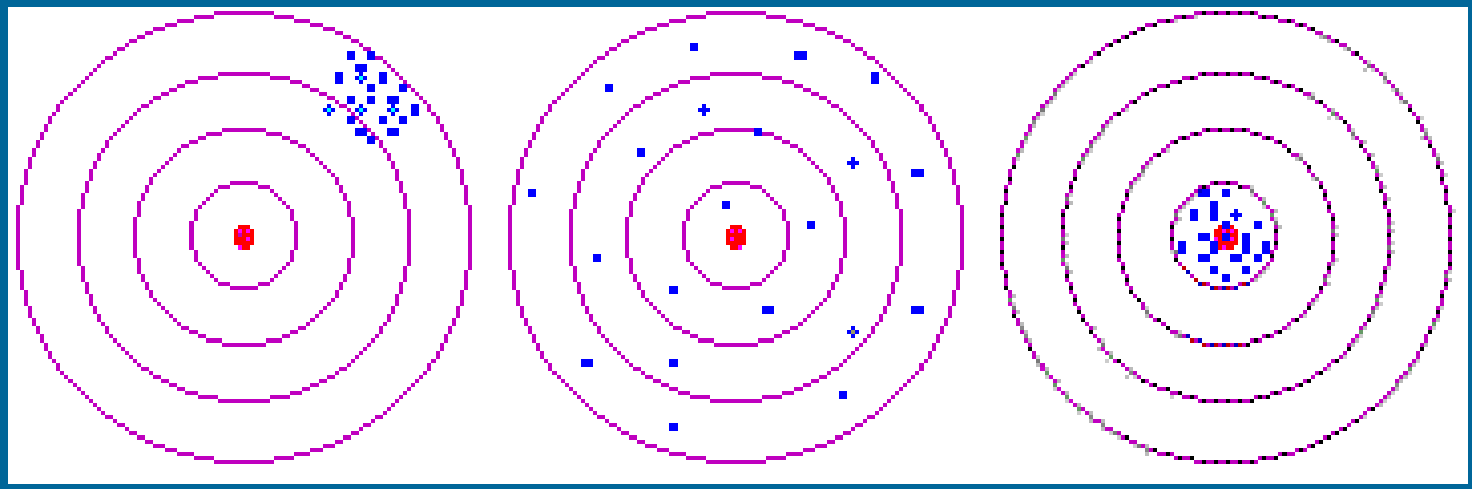
In total, the results of particular set of measurements will be (randomly) distributed around expected value.

The measure of dispersion of results gives us information about precision.

Accuracy (trueness and/or precision) .....

..... variability of results

A) B) C)



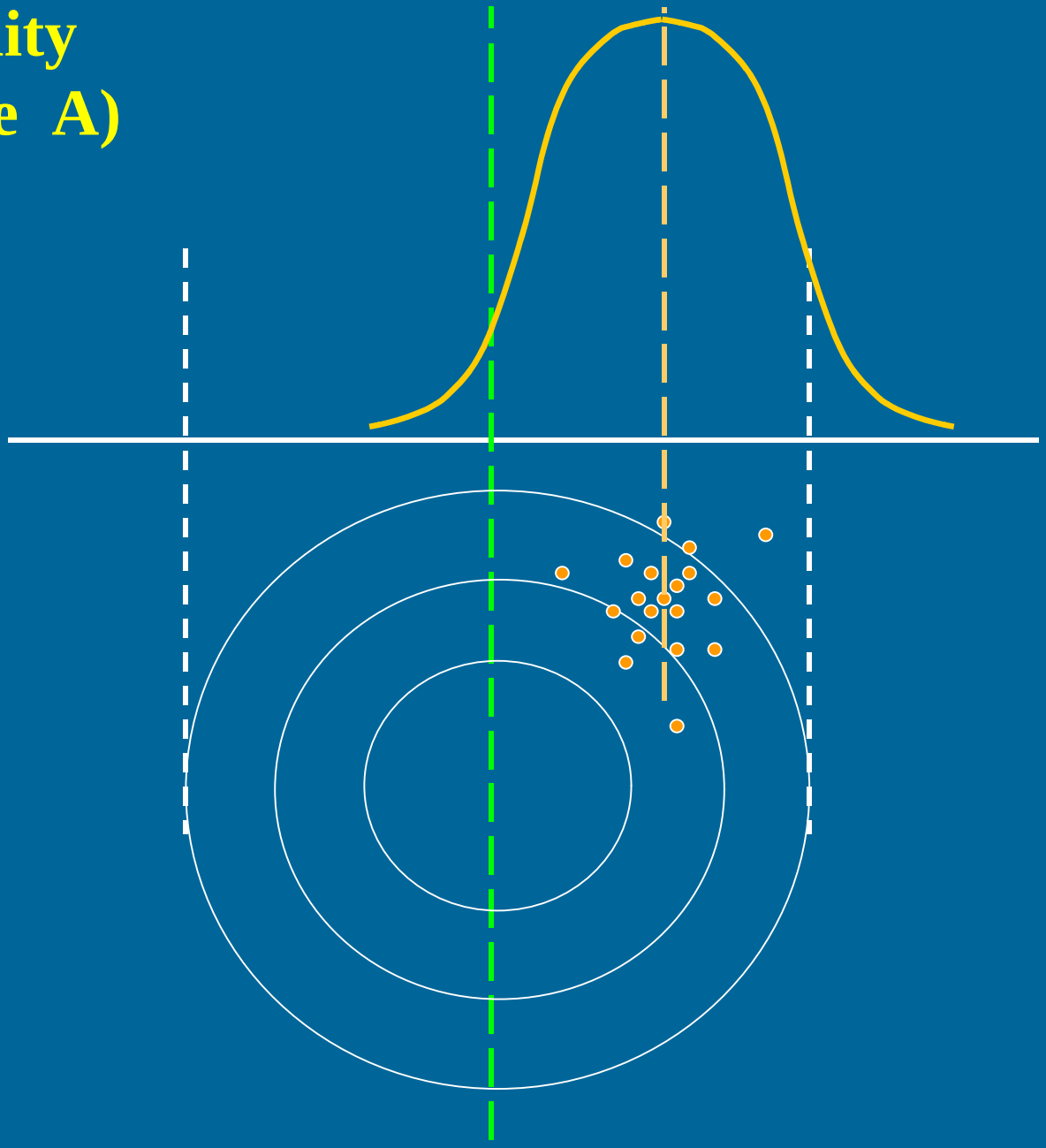
..... variability of results

Both, trueness and precision are quantitative parameters.

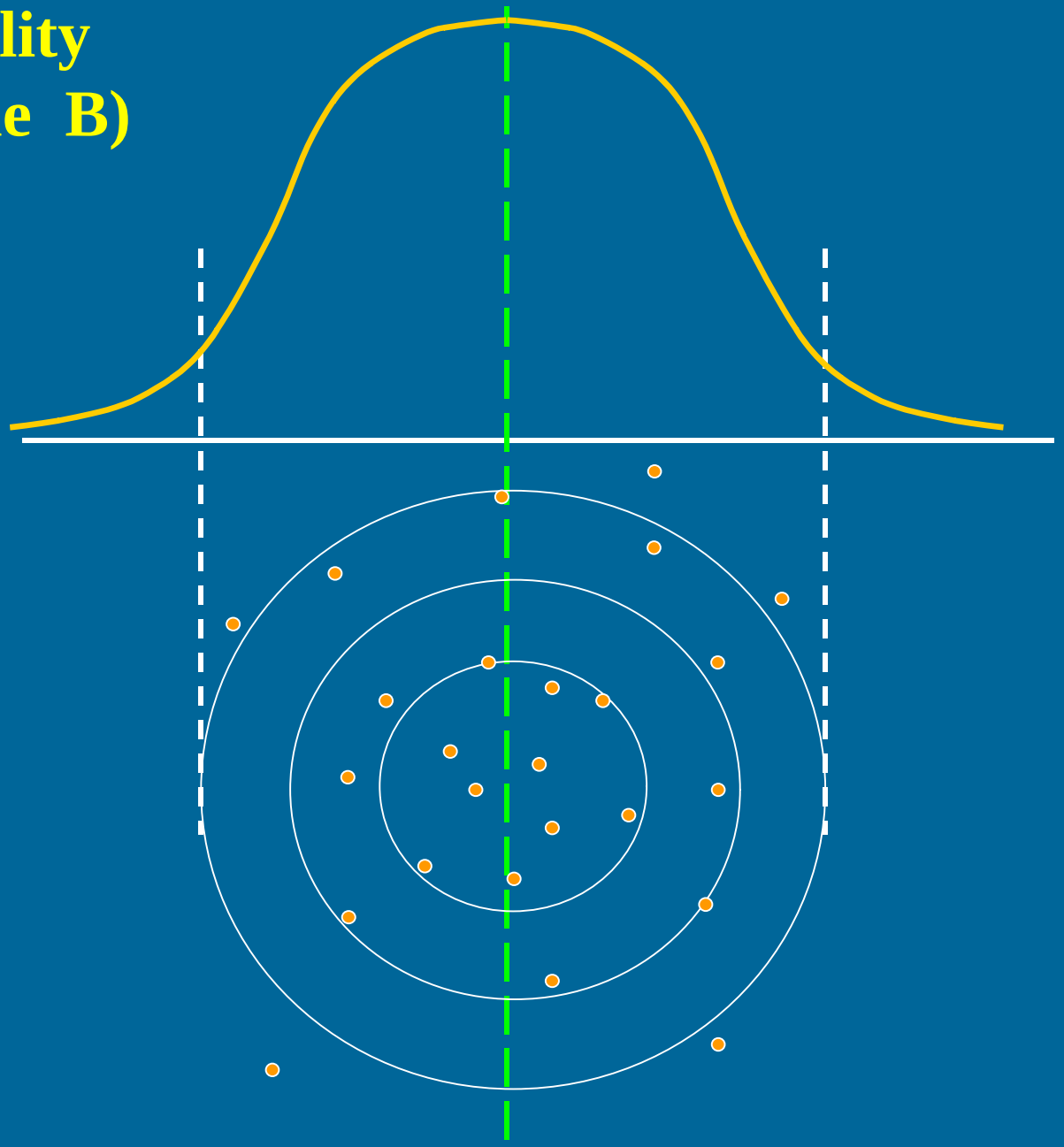
We can estimate dispersion (*precision*) by calculation of **standard deviation**.

Also, we can estimate so-called "*true value*" by calculation of a **mean**.

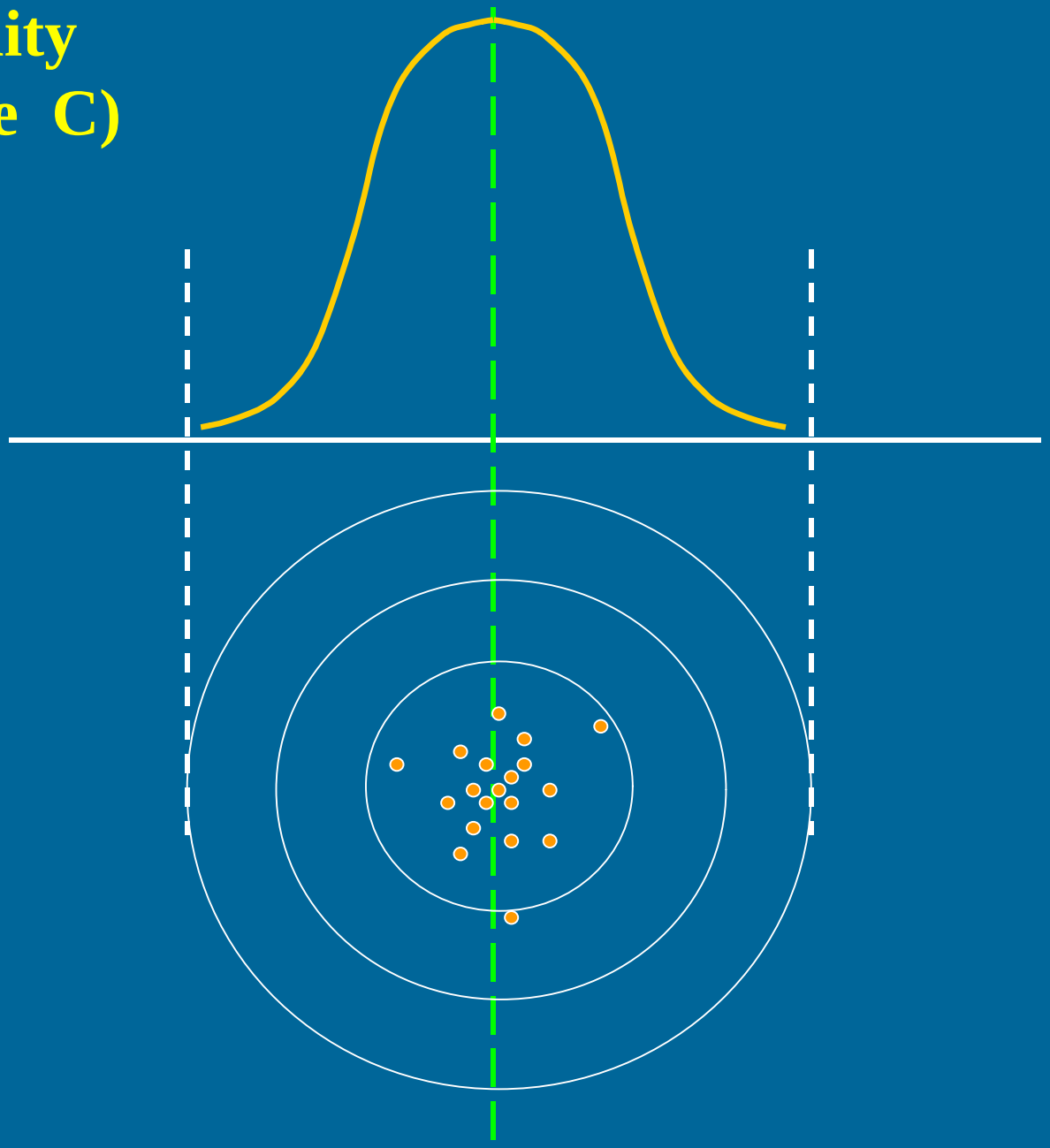
variability
example A)



variability
example B)



variability
example C)



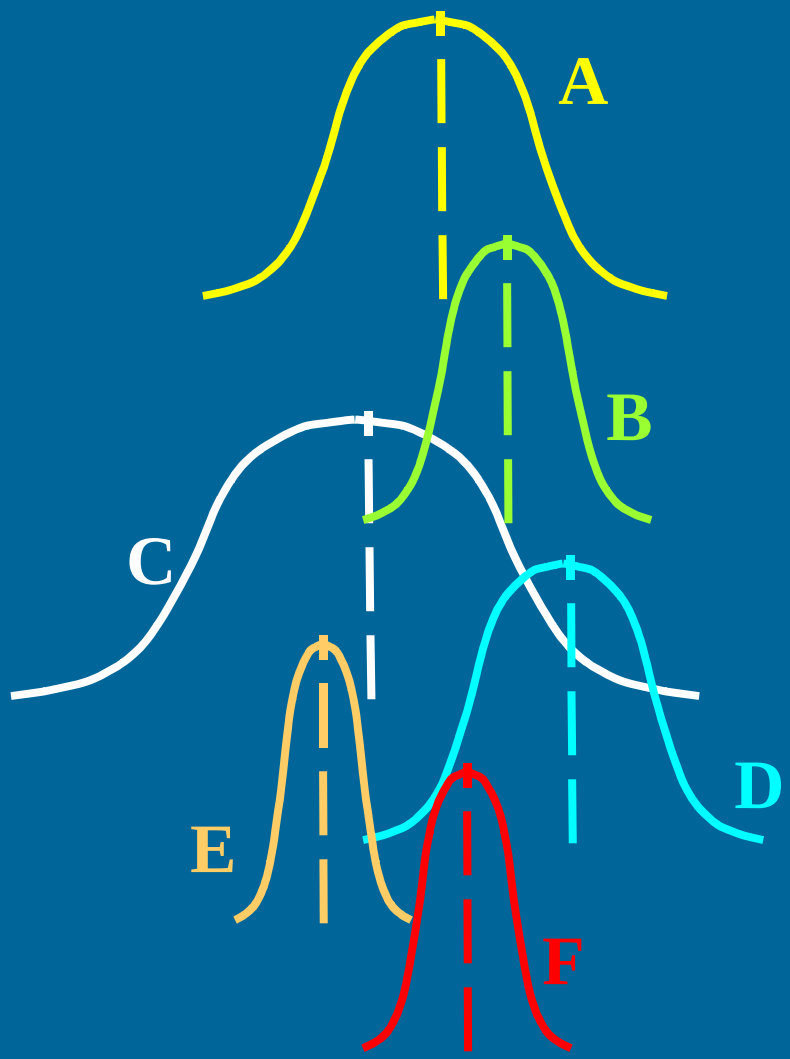
..... variability of results

More important is that we can compare and analyse differences between particular sets of test results.

This directs us to the subject of
repeatability and reproducibility.

Different variabilities

between different sets of test results regarding to different sources of variabilities.



..... definitions of R&R .....

Regarding factors (*elements*) of the system which contribute to the variability of the test results,

two **conditions of precision**,
termed repeatability and reproducibility conditions,

have been found necessary and useful for describing the variability of a measurement method [3].

Repeatability conditions

Conditions where independent test results are obtained
with the same method
on identical test items
in the same laboratory
by the same operator
using the same equipment
within short intervals of time [3].

Repeatability is precision under repeatability conditions.

Reproducibility conditions

Conditions where test results are obtained
with the same method
on identical test items
in different laboratories
with different operators
using different equipment [3].

Reproducibility is precision under reproducibility conditions.

..... extremes of variability

Thus repeatability and reproducibility are
the two extremes of precision.

Repeatability describes **minimum** and
reproducibility the **maximum variability** in results.

Repeatability limit, *r*

The value less than or equal to
which the absolute difference
between two test results
obtained under repeatability conditions
may be expected to be with a probability of 95 % [3].

$$r = f \cdot s_r \sqrt{2}$$

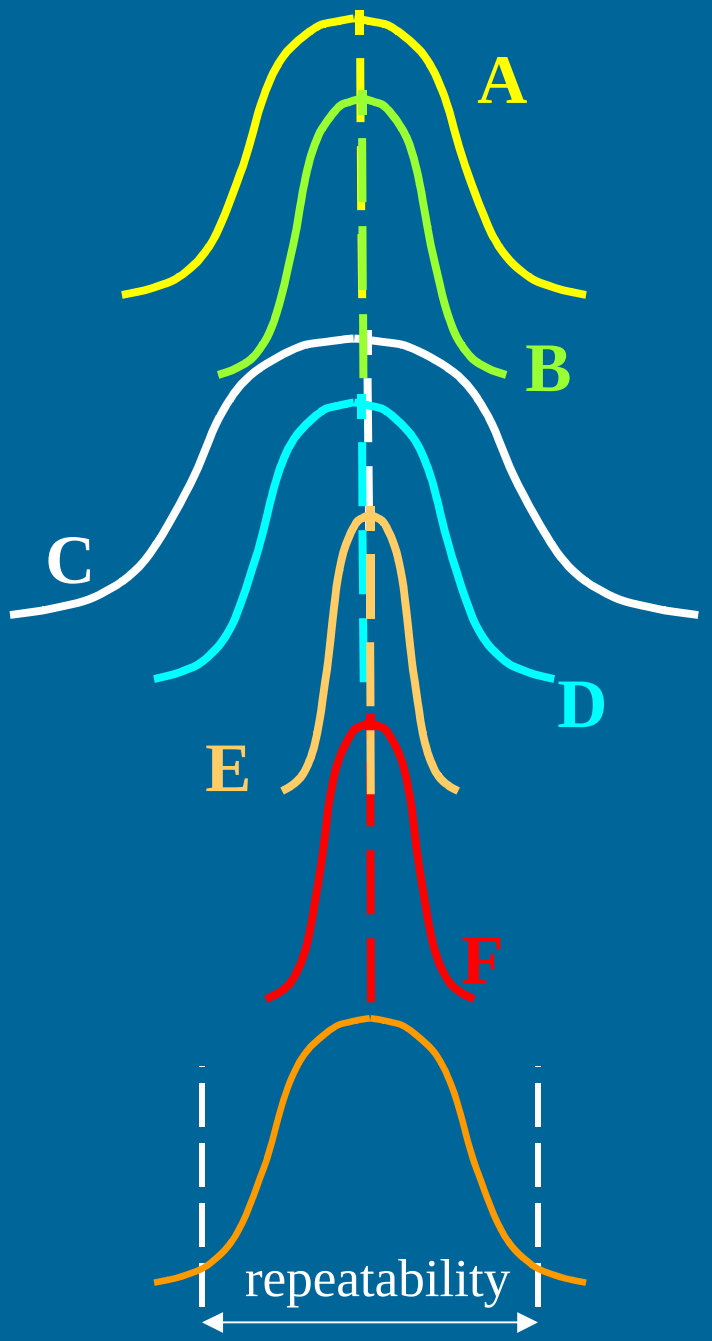
Reproducibility limit, *R*

The value less than or equal to
which the absolute difference
between two test results
obtained under reproducibility conditions
may be expected to be with a probability of 95 % [3].

$$R = f \cdot s_R \sqrt{2}$$

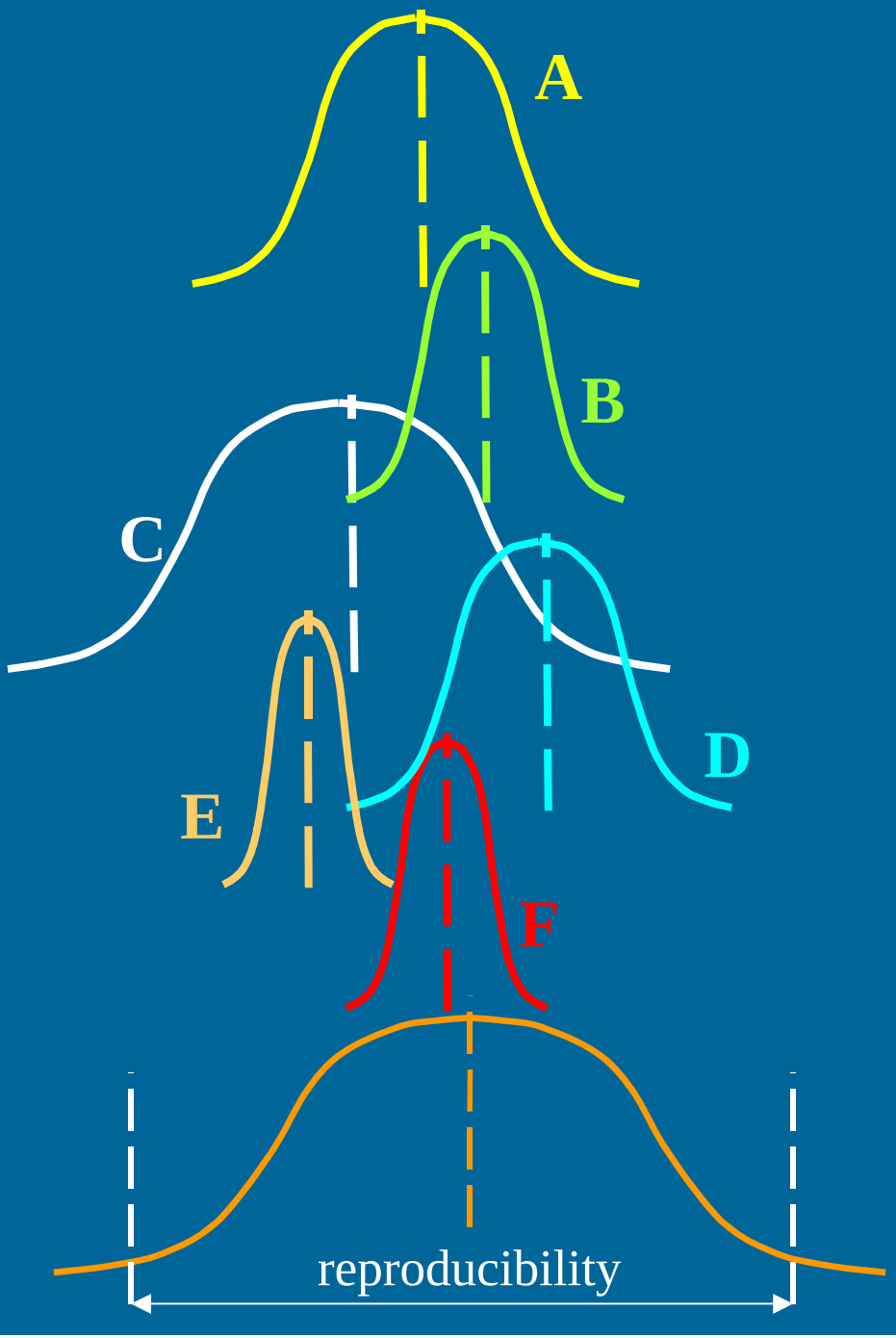
Repeatability

Illustration of the meaninig of repeatability standard deviation and limit.



Reproducibility

Illustration of the meaninig of reproducibility standard deviation and limit.



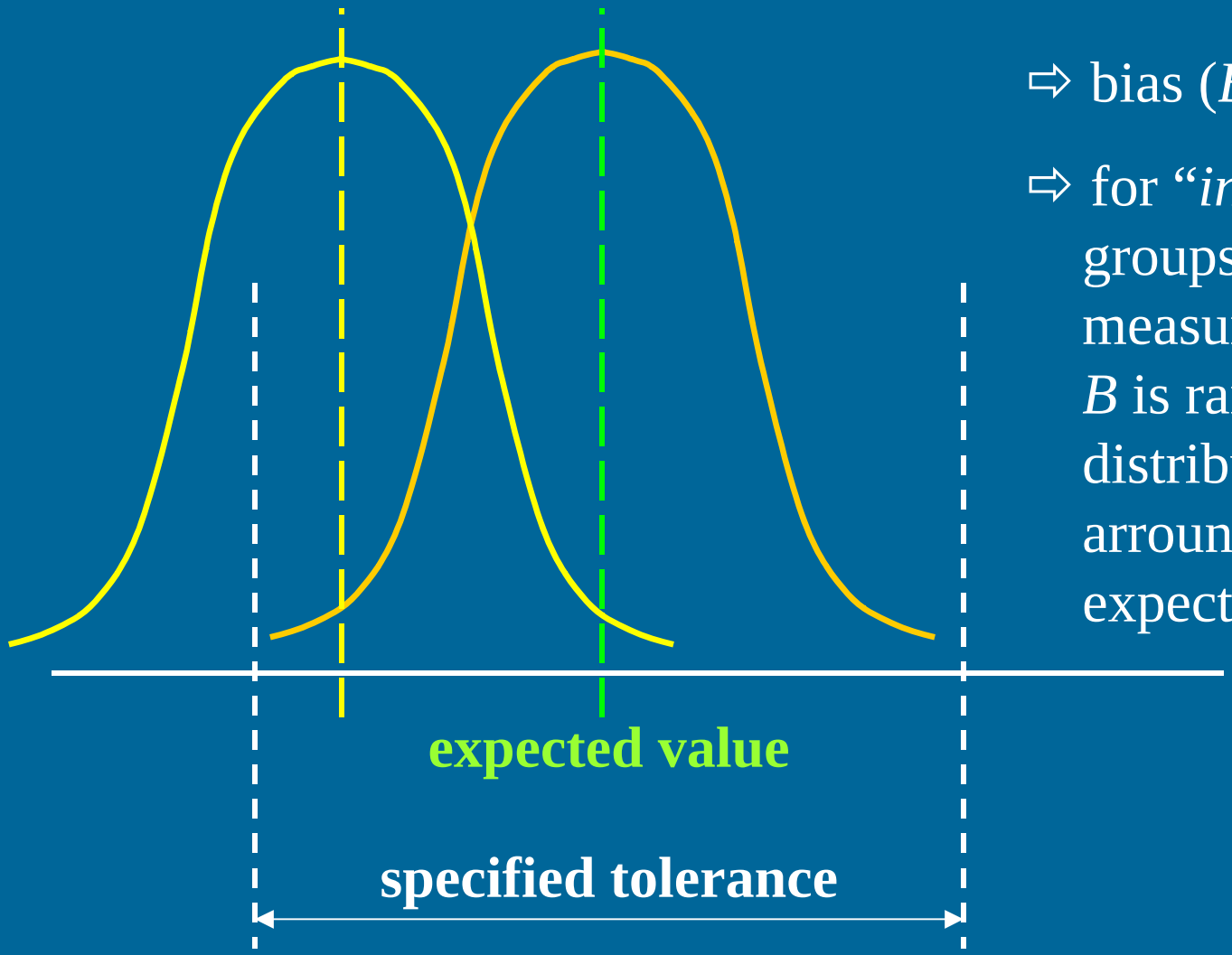
Contributing elements of “test system” [4]

- Operator
- Equipment
- Calibration
- Time

Elements of "test system" regarding NDT/mine_det

- **Personnel** (operator/technician ⇨ human factor)
- **Equipment** (instruments, tools, utensils, probes/sensors, manipulators, spare parts & materials, reference blocks, ...)
- **Procedure**
  (measurement/test method, calibration)
- **Test object / measured entity**
  (any characteristic: microstructure, geometry, surface conditions, orientation, etc. ....)
- Environment
  (temperature, humidity, viewing conditions ....)

Trueness and systematic error



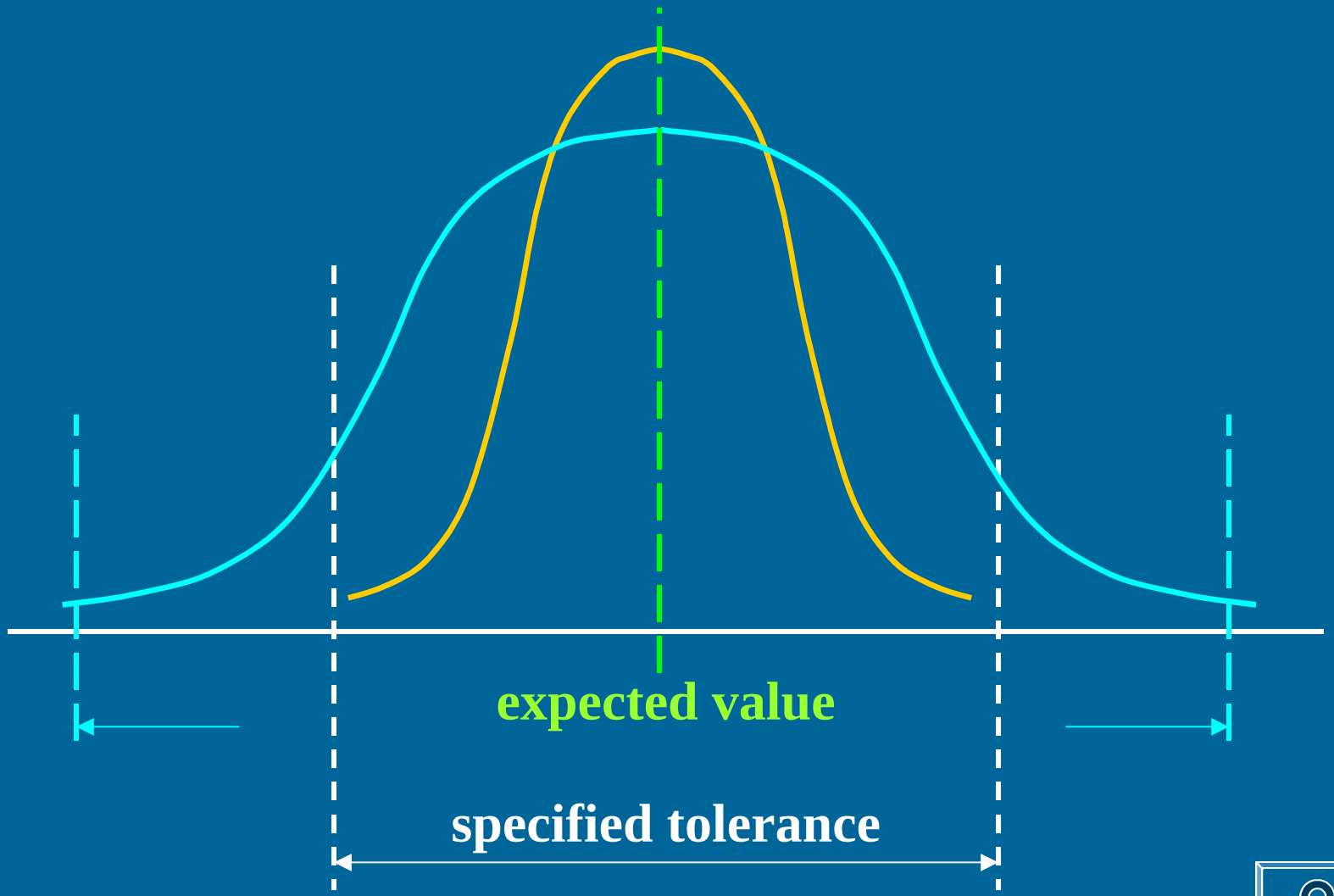
⇨ bias (*B*)

⇨ for "*infinite*" groups of measurements *B* is randomly distributed arround expected value

Systematic error

Is a **mean** that would result from
an infinite number of measurements of
the same measurand
carried out under repeatability conditions
minus
a **true value** of the measurand [1].

Precision and random error



Random error

Is a **result** of a measurement minus the **mean** that would result from an infinite number of measurements of the same measurand
carried out under repeatability conditions [1].

Because only a finite number of measurements can be made, it is possible to determine only an estimate of random error [1].

Accepted reference value

A value that serves as an agreed-upon reference for comparison [3].

Could be derived as *theoretical or established value, an assigned or certified value, a consensus*.

When not available, the expectation of the quantity, i.e. the mean of a specified population of measurements [3].

3. R&R experiment

- Experiment layout
- How to organise and carry out the experiment
- DoE

Experiment layout

- The objective of the experiment is
  to determine the precision
  obtainable by the general population of operators
  performing standard measurement (test) method.

- Each combination of
  a laboratory and a level
  is called **cell**.
  Each cell contains ***n*** repetitive test results.

Experiment layout

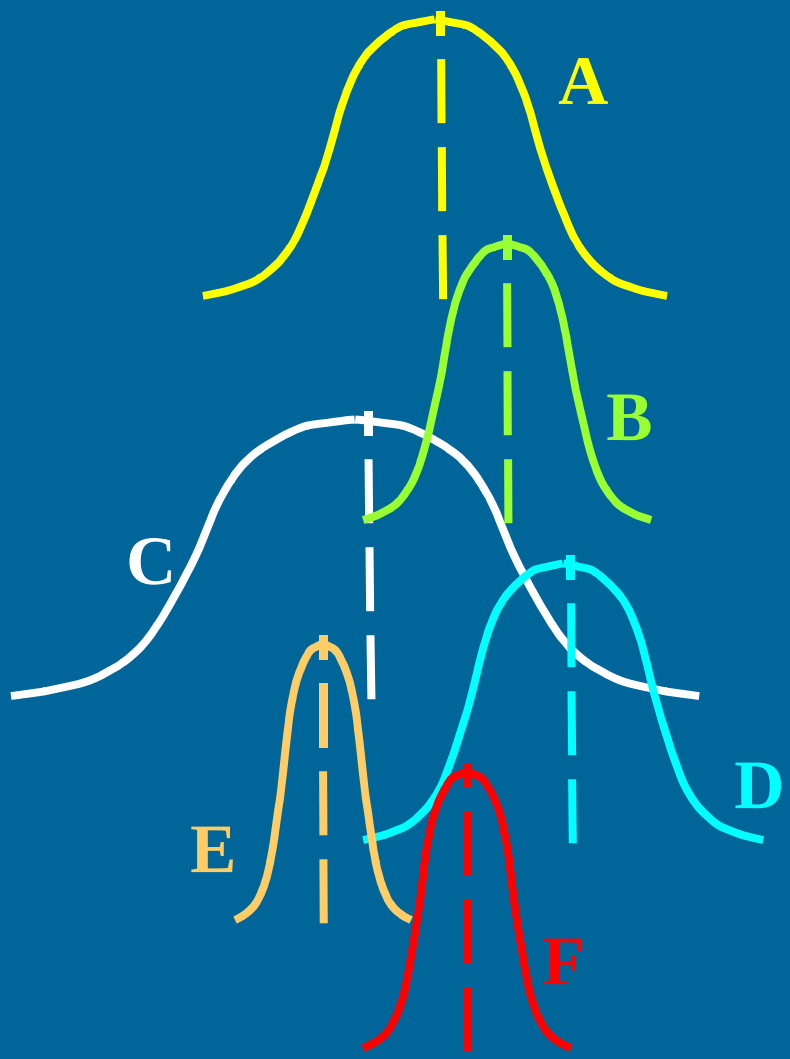
p Lab.	n	level 1	level 2	level 3
	1	25,3	37,5	43,9
A	2	25,2	37,7	44,4
	3	25,3	37,7	44,0
	1	25,4	37,4	44,4
B	2	25,4	37,5	44,5
	3	25,3	37,7	44,9
	1	25,1	37,7	44,3
C	2	25,3	37,3	44,9
	3	25,0	37,6	44,0
	1	25,1	37,4	44,6
D	2	25,5	37,2	44,6
	3	25,5	37,5	44,0
	1	25,3	37,5	44,3
E	2	25,2	37,4	44,5
	3	25,3	37,7	44,7

Experiment layout

q levels

p laboratories

n replicate test results



Performance of the experiment (1)

- Perform the measurements
  according to the specified standard method !

- Group of ***n*** measurements
  on each level
  shall be carried out under repeatability conditions
  (*short time,*
  without intermediate recalibration of the apparatus
  unless this is an integral part of standard procedure).

Performance of the experiment (2)

- It is essential that a group of ***n*** measurements on each level be performed independently (*independent repetitive test results*).

- Sometimes, despite instructions to the operator, during obtaining group of ***n*** measurements at specific level, previous results may influence subsequent test results, and thus the repeatability variance. Then the codification of samples should be considered in such way that the operator will not know which are the replicates for a given level. (*repeatability conditions?*)

Performance of the experiment (3)

- ***It is not essential*** that all
  the ***q*** groups of ***n*** measurements each
  be ***performed strictly within short interval***;
  different groups of measurements
  may be carried out on different days.

- Measurements of all ***q*** levels
  shall be performed ***by one and the same operator***.
  In addition, the ***n*** measurements at a given level
  shall be performed ***using the same equipment*** throughout.

Performance of the experiment (4)

- If unexpectedly operator becomes unavailable, another operator may complete measurements but this can only occur ***between two of the q groups***, and it has to be reported.

- A time limit shall be given....

- All samples shall be clearly and properly labeled. (***identification***)

Performance - Tips (1)

- For some measurements,
  there may be in fact a ***team*** of operators, each of whom performs some specific part of the procedure.
  In such case, the team shall be regarded as "***the operator***".
  Any change in team indicates "different operator".

- In precision experiment test results shall be reported to at least ***one more digit*** than
  specified in the standard procedure.

Performance - Tips (2)

- For the purposes of this ISO,
  a ***laboratory*** is considered to be a combination
  of the operator,
  the equipment and
  the test site [3].

  One test site (or laboratory) may thus provide
  several "laboratories" (several operators
  each with the independent sets of equipment
  and situations).

Performance - Tips (3)

- It should be pointed out to the operators that the purpose of the exercise is to ***discover the extent to which results can vary in practice***, so that there will be less temptation for them to discard or rework results that they feel are inconsistent.

- Operators shall report any anomalies or difficulties, rather than adjusting test result in order to avoid missing data. (*There is R&R analysis specifically for experiments with missing data, and also for redundant data and outliers.*)

4. R&R analysis

- Statistical analysis (*step by step*)
- Calculation of R&R values
- Examples

R&R step by step "procedure" (1)

.... to determine R&R values for a standard measurement/test method.

1) Premises [3, 4]:

- repeatability conditions within labs (*n replicate test results @ q levels*)
- reproducibility conditions between *p* labs
- it is assumed that distribution of the test results is approximately normal
- the model for test result is $y = m + B + e$, ***m*** is a general mean (expectation), ***B*** bias, ***e*** error.
- measurements on a continuous scale

R&R step by step "procedure" (2)

2) Calculation of R&R values (parameters) [4]:

↳ s_W^2 is the estimate of the **within-lab** variance $\text{var}(e) = s_W^2$

for each cell s_i^2

↳ s_r^2 is the estimate of **repeatability** variance,

formally, it is
the arithmetic mean of s_W^2 …. $s_r^2 = \dfrac{\sum_{i=1}^{p}(n_i - 1)s_i^2}{\sum_{i=1}^{p}(n_i - 1)}$

R&R step by step "procedure" (3)

2) Calculation of R&R values (parameters) [4]:

- s_L^2 is the estimate of the **between-lab** variance $\mathrm{var}(B) = s_L^2$

  $$s_L^2 = \frac{s_d^2 - s_r^2}{n}$$ at each of q levels,

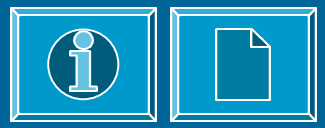
- s_R^2 is the estimate of **reproducibility** variance, $s_R^2 = s_r^2 + s_L^2$

- M_i are cells' means, m is a "general mean" @ certain level

R&R step by step "procedure" (4)

3) To check data for stragglers and outliers by means of statistical tests [4]:

- Cohran's test.
  It checks the homogeneity of variances
  at certain level.
- Grubbs' tests.
  It checks the "consistency" of the cells' means.
- These tests check existance of stragglers and outliers
  at 5 % and 1 % significance level.



R&R step by step "procedure" (5)

4) Determination of R&R limits [5]:

- when a quantity is based on sums (or differences) of n independent estimates, then the resultant quantity has standard deviation $s\sqrt{n}$

- R&R limits are for differences between two results, therefore, $s\sqrt{2}$

repetability limit $r = f \cdot s_r \sqrt{2}$

reproducibility limit $R = f \cdot s_R \sqrt{2}$

according to J.Mandel $f = 1{,}96$ (*confidence level 95 %*)

R&R step by step "procedure" (6)

5) **Checking the acceptability of test results** obtained under repeatability conditions.

When examining :

- two single test results the comparison shall be made with repeatability limit, ***r***

$$r = 2.8 \cdot s_r$$

- two groups of measurements in **one** laboratory, n_i measurements giving arithmetic mean, comparison shall be made with critical difference, ***CD***

$$CD = r\sqrt{\frac{1}{2n_1} + \frac{1}{2n_2}}$$

R&R step by step "procedure" (7)

5) **Checking the acceptability of test results** obtained under reproducibility conditions.

When examining :

- two single test results the comparison shall be made with reproducibility limit, ***R***

$$R = 2.8 \cdot s_R$$

- two groups of measurements in **two** laboratories, n_i measurements giving arithmetic mean, comparison shall be made with critical difference, ***CD***

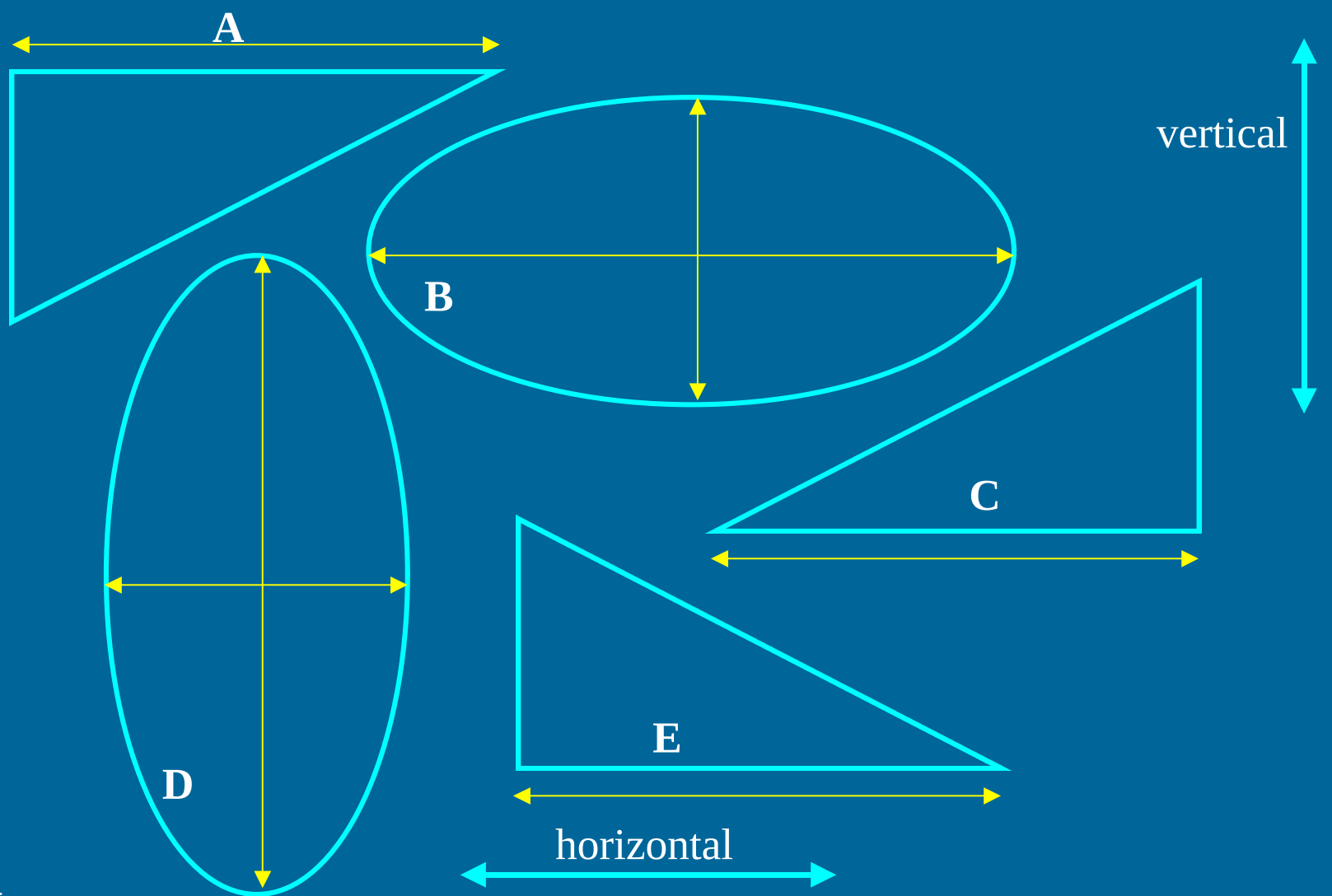
$$CD = \sqrt{R^2 - r^2\left(1 - \frac{1}{2n_1} - \frac{1}{2n_2}\right)}$$

Example #3 - our R&R experiment !

"Procedure" for measurements:

- sheet of paper with "specimens" A, B, C, D, E, F,
- you are measuring length of dimension
  in direction (vert. and hor.)
  as it is required in the columns of the table for test results,
- only one (*repetitive*) measurement at this moment, n=1,
- all measurements in mm !
- please estimate up to the one tenth of the mm.

Example #3 - our R&R experiment !



Example #1

Ultrasonic thickness measurements
were carried out by 8 teams consisting of:

- 8 operators,
- 8 different digital ultrasonic instruments with A-scan (flaw detectors),
- (8) different TR and single transducer probes,
- using different couplants (greas, oil, gel),

where 8 operators with different equipment represent
8 differently equipped laboratories **(*p*=8)**.

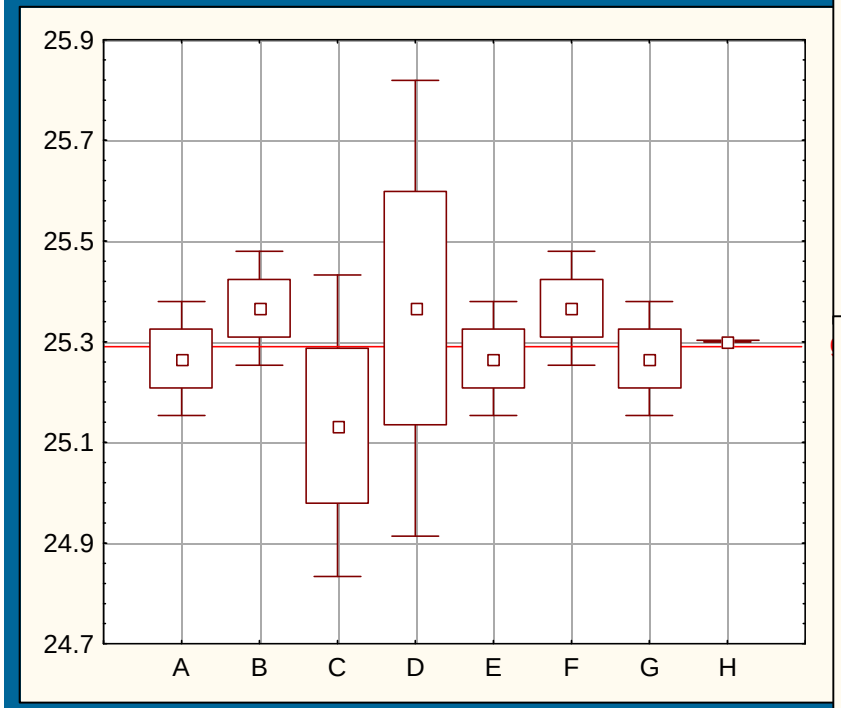
DoE 8x3 (*n*=3)

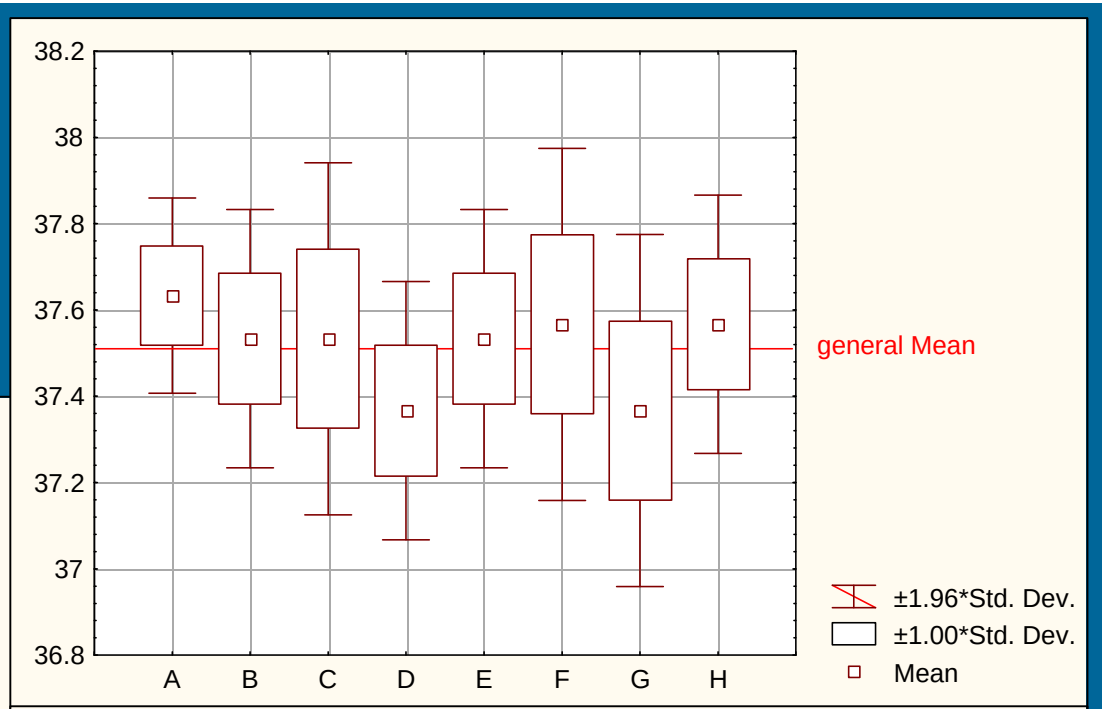
Example #1 DoE 8 x 3, q=3

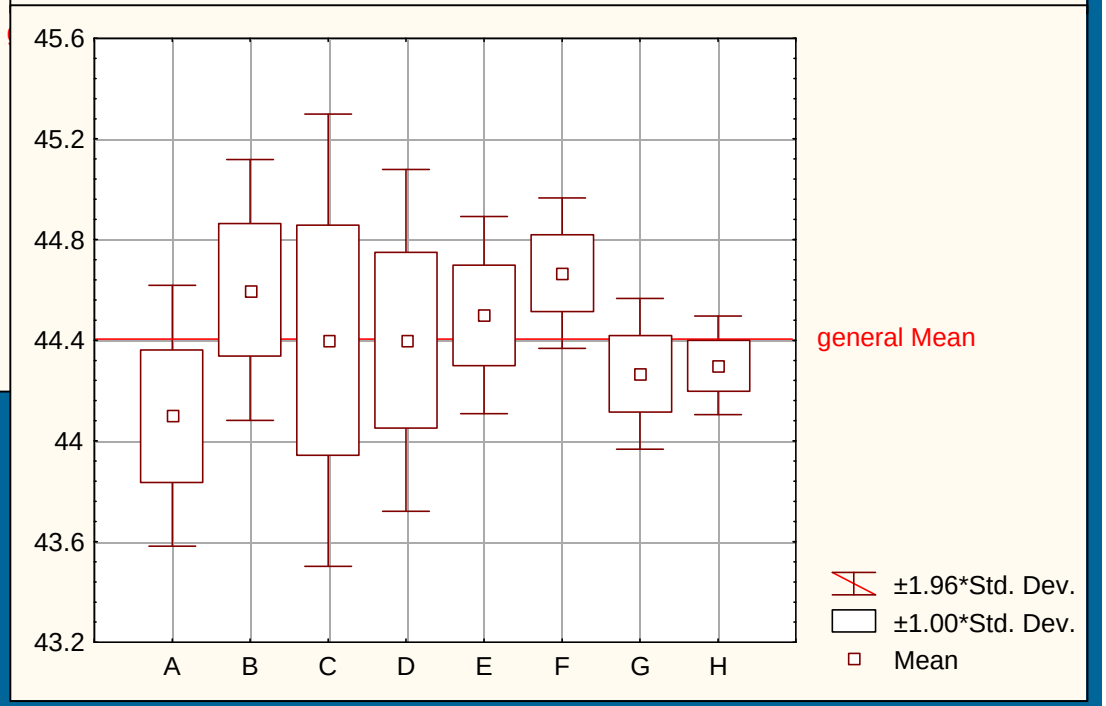
Furthermore:

- 3 repetitive measurements, ***n*=3**,
- 3 specimens (3 levels of thickness, ***q*=3**),
- digital readout; appearance was set to one decimal place,
- no (re)calibration during the measurements,
- approx. 1 hour for all measurements,
- all operators were in the same room ☺.....

Example #1


25.9
25.7
25.5
25.3
25.1
24.9
24.7
A
B
C
D
E
F
G
H


38.2
38
37.8
37.6
37.4
37.2
37
36.8
A
B
C
D
E
F
G
H
general Mean
±1.96*Std. Dev.
±1.00*Std. Dev.
Mean


45.6
45.2
44.8
44.4
44
43.6
43.2
A
B
C
D
E
F
G
H
general Mean
±1.96*Std. Dev.
±1.00*Std. Dev.
Mean

Example #2

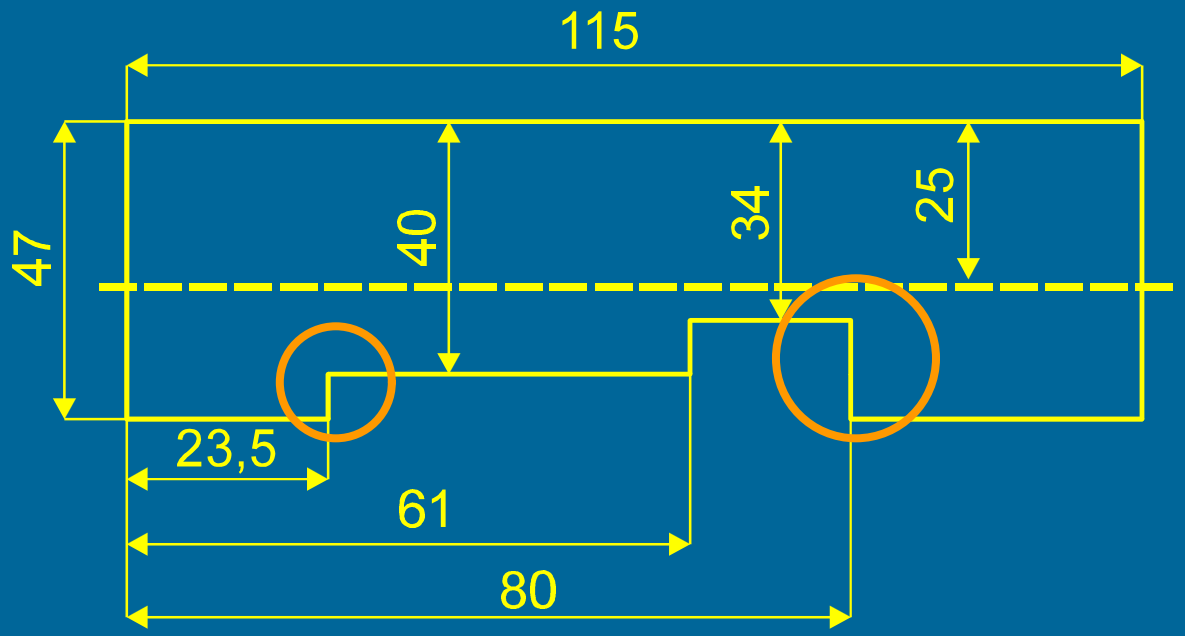
Ultrasonic testing; A scan instruments; straight beam probes.

Measurement of the location of the indication (2 circled edges). Results are given as subtraction to the “true value”. Results are in mm, measured with one decimal place.

DoE 6x6

n=6

q=2



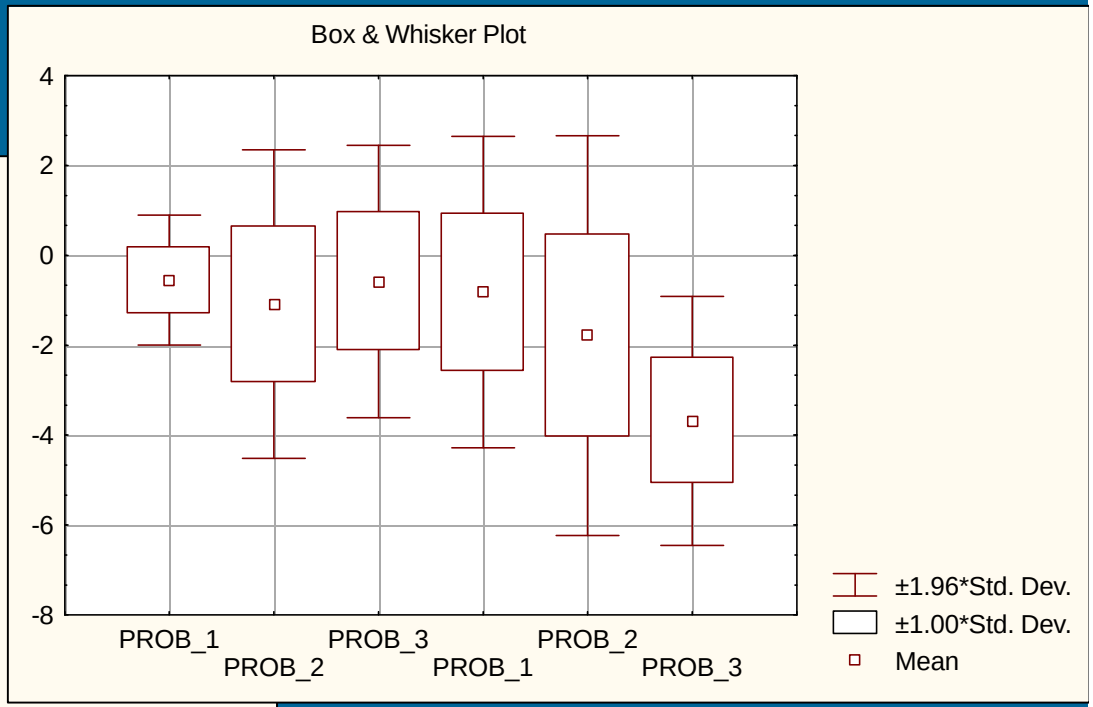
Example #2 DoE 6 x 6, q=2

Furthermore:

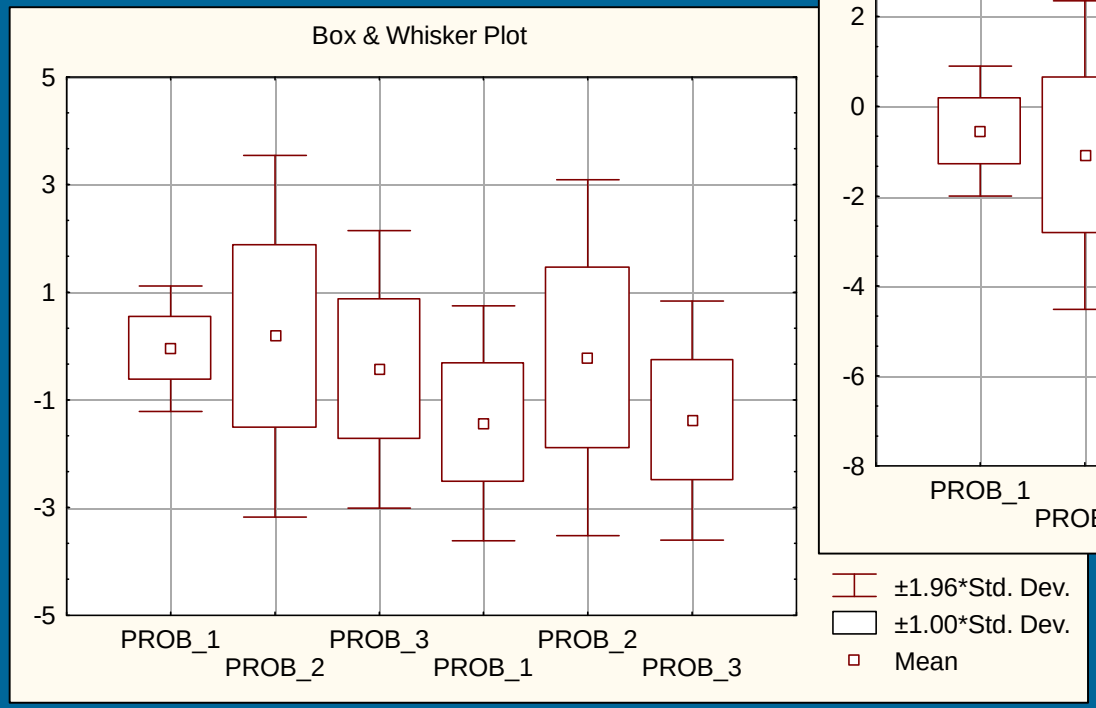
- ⇨ 6 labs were combinations of 3 probes and 2 instruments, **p=6**
- ⇨ only one operator !?!
- ⇨ test results were taken in 3 days, each day 2 diverse combination (probe-instrument)

Example #2

Level 2, @40mm

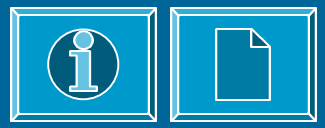


Level 1, @34mm



.... back to: R&R step by step "procedure"

When R&R analysis has been successfully performed,
and if the number of involved laboratories with “valid” test results is in accordance with ISO recommendations [3],
then,
one could come up to the publication of the conclusions,
that is,
to the publication of results in form of statements.



Publication of analysis results / statements

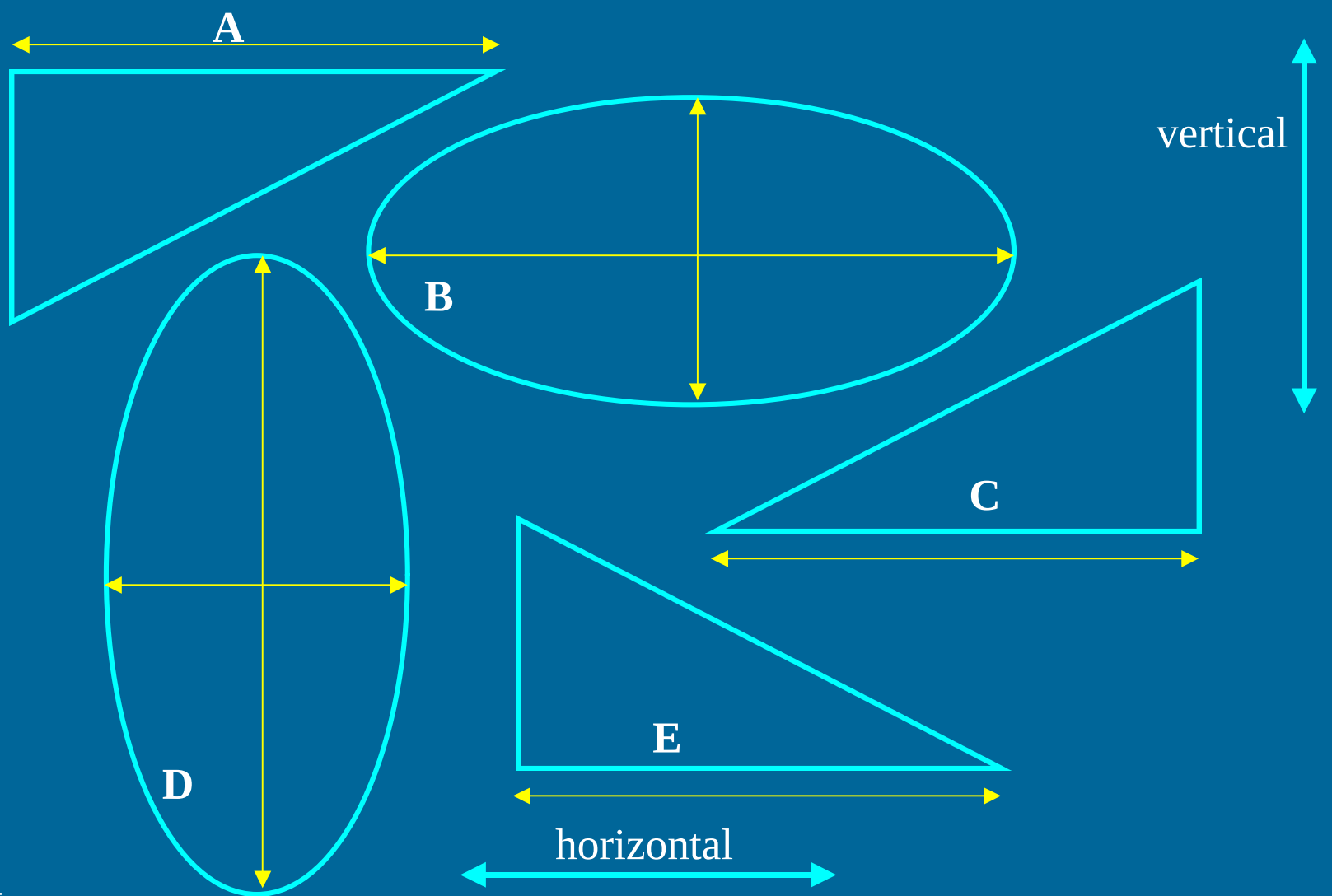
The difference between two test results
found on identical test material
by one operator
using the same apparatus
within the shortest feasible time interval
will exceed the repeatability limit (r)
on average not more than once in 20 cases (5%)
in the normal and correct operation of the method. [5]

Publication of analysis results / statements

Test results

on identical test material
reported by two laboratories
will differ by more than the reproducibility limit (R)
on average not more than once in 20 cases (5%)
in the normal and correct operation of the method. [5]

Example #3 - our R&R experiment !



5. Conclusion remarks & discussion

- Discussion of examples proposed by participants
- Conclusions

Discussion - your ideas for R&R analysis

- If you have ideas how to organize
  R&R experiment (DoE) please give us an example

  (*examples from your professional occupation*)

Discussion - your ideas for R&R analysis

- Perhaps R&R experiment for Mine Detection
- which method, system, conditions,
- consider certain “population” of applications from practice,
- which implicates ranges of operators, equipment, levels....

Conclusion remarks

We want to carry out R&R analysis,
with focuse on NDT/demining applications/methods.

First we have to design R&R experiment. To design R&R experiment we have to be aware about:

- inherent characteristics of the (test/measurement) ***process*** under the consideration, and
- that the results are obtained by the particular (test/measurement) ***system***
- under the certain ***conditions***.

Conclusion remarks

- The goal is to get a ***realistic demonstration*** (good representation) of the measurement/test method, not an "artificial" one.
  ("... to represent a realistic cross-section with the R&R experiment.“ [3]).
- Finally, by quantifying R&R we can ***determine process/system capabilities***.
  For this we need competent interpretation of R&R parameters and analysis.
- The conclusions are valid over the range of the concerned levels' values.

Conclusion remarks

Additionally, R&R analysis opens up further possibilities:

- check of the acceptability of test results,
- assessment of stability of test results over a period of time within a laboratory,
- assessment of the laboratory performance,
- comparison of alternative measurement methods.

Conclusion remarks

Furthermore, R&R also provides:

- identification and differentiation of influencing factors,
- quantification of the specific influence and
- better understanding of the behavior/response of particular test system.

@ The end ...

Why we were listening all this ..... ??!?

We already know

how to calculate standard deviations, arithmetic means,
how to perform statistical tests (ANOVA, t-test),
how to determine system or process capability, etc.....

Well,

I like the way how test results determine **limits** for themselves - **giving quantitative criteria**.

Using proposed ISO model I keep my NDT system under control !

maybe you will like it too ! ☺

@ The end ...

... imperfections still exist ! ....

... they need to be corrected in the future [1].

Thank you for your attention and co-operation !

Literature

[1] ..... *International Vocabulary of basic and general terms in metrology*, BIPM, IEC, IFCC, ISO, IUPAC, IUPAP, OIML, 1984,1987.

[2] D.Benèiæ, F.Dusman. *Concept and significance of repeatability and reproducibility*, Geodetski list, Zagreb, 1995.

[3] ISO 5725-1:1994. *Accuracy (trueness and precision) of measurement methods and results -- Part 1: General principles and definitions*

[4] ISO 5725-2:1994. -- *Part 2: Basic method for the determination of repeatability and reproducibility of a standard measurement method*

[5] ISO 5725-6:1994 -- *Part 6: Use in practice of accuracy values*