Qualitative PT data analysis with easy-to-interpret scores

Christian Bläul and Steffen Uhlig

QuoData in short

- Since 1995, in Dresden (main office) and in Munich/Freising (Germany)
- Staff: 31
  - Scientific staff: 12, mainly Mathematics, Physics and Bioinformatics (9)
  - IT staff: 8

Extracting Value from Data

Our statistical services:

- Sampling and Extrapolation
- Statistically Advanced Experimental Design
- Validation and certification of measurement methods, bioassays and biosensors
- Interlaboratory Studies
- Meta Studies

Our products (software development):

- Software for optimization, validation and PT („PROLab Plus“)

Our Main Application Areas:

- Food Safety, Consumer Protection, Environmental Science, Forensics, Medical Diagnostics

How to derive tolerance limits?

Introduction

- Idea: Calculate the laboratory specific ROS (over all samples) and use Binomial distribution to derive tolerance limit.

- Example (PT on the Detection of Highly Infectious Pathogens)
  - n=9 replicates/samples
  - ROS=0.901 (227 out of 252 tests were successful) across laboratories
  - BINOM.INV(9;0.901;0.05)=6 (in Excel)
  - In other words: As long as a laboratory has at least 6 positive results, there is no reason to believe that this laboratory has lower competence than the average.
  - Or put it this way: the lower 95 % tolerance limit for the number of positive results for a participant with average competence is 6.
  - Therefore the assessment criterion: at least 6 positive results.

Example: Identification of bacteria species

PT on the Detection of Highly Infectious Pathogens

Bacteria species that have been correctly (+) and incorrectly (-) identified by the laboratories

Sample (Species)	Laboratories																													
	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
HPB 1	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	no performance	no performance	no performance	+	+	+	+	+	+	+	+
HPB 2	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+				+	+	+	+	-	+	+	+
HPB 3	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-				+	+	+	+	+	+	+	+
HPB 4	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+				+	+	+	+	-	-	+	-
HPB 5	-	+	-	+	-	+	+	-	+	+	+	-	-	+	-	-	+	+	+	+				+	-	+	+	+	+	+	+
HPB 6	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+				+	+	+	+	+	+	+	+
HPB 7	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+				+	+	+	+	+	-	+	-
HPB 8	+	+	-	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+				+	-	+	+	+	-	+	+
HPB 9	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+				+	-	-	+	+	+	+	-

How to derive tolerance limits?

Introduction

- Performance assessment: Based on Rate Of Success (ROS) over all samples

Sample	Laboratory		
	15	16	17
HPB 1	+	+	+
HPB 2	+	+	+
HPB 3	+	+	+
HPB 4	+	+	+
HPB 5	-	-	+
HPB 6	+	+	+
HPB 7	-	+	+
HPB 8	-	+	+
HPB 9	-	+	+
ROS	**56%**	**89%**	**100%**

- Assessment criterion?

How to derive tolerance limits?

Introduction

- Idea: Calculate the laboratory specific ROS (over all samples) and use Binomial distribution to derive tolerance limit.

- Example (PT on the Detection of Highly Infectious Pathogens)
  - n=9 replicates/samples
  - ROS=0.901 (227 out of 252 tests were successful) across laboratories
  - BINOM.INV(9;0.901;0.05)=6 (in Excel)
  - In other words: As long as a laboratory has at least 6 positive results, there is no reason to believe that this laboratory has lower competence than the average.
  - Or put it this way: the lower 95 % tolerance limit for the number of positive results for a participant with average competence is 6.
  - Therefore the assessment criterion: at least 6 positive results.

How to derive tolerance limits?

Example: PT on the Detection of Highly Infectious Pathogens

Almost all participants fulfill the performance criterion of at least 6 successful samples.

Not successful: Lab 15

Sample (Species)	Laboratories																													
	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
HPB 1	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	no performance	no performance	no performance	+	+	+	+	+	+	+	+
HPB 2	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+				+	+	+	+	-	+	+	+
HPB 3	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-				+	+	+	+	+	+	+	+
HPB 4	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+				+	+	+	+	-	-	+	-
HPB 5	-	+	-	+	-	+	+	-	+	+	+	-	-	+	-	-	+	+	+	+				+	-	+	+	+	+	+	+
HPB 6	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+				+	+	+	+	+	+	+	+
HPB 7	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+				+	+	+	+	+	-	+	-
HPB 8	+	+	-	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+				+	-	+	+	+	-	+	+
HPB 9	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+				+	-	-	+	+	+	+	-

How to derive tolerance limits?

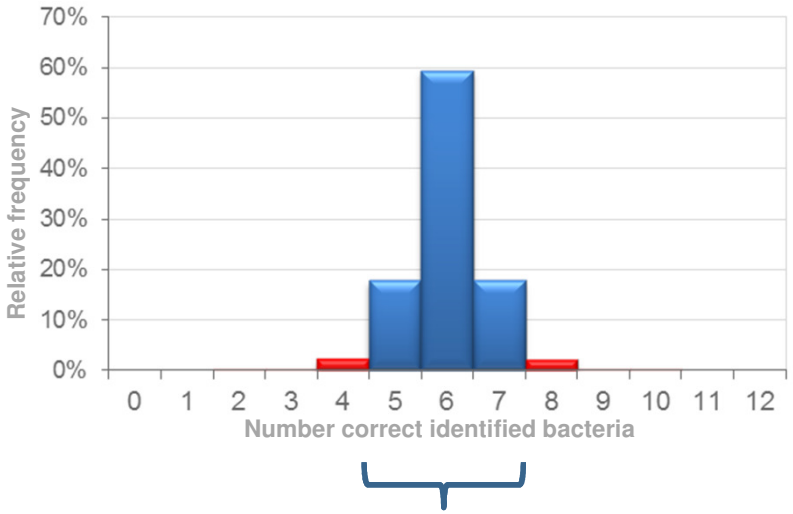
- What are the prerequisites of the Binomial criterion?
  - Binomial distribution applies in case of n independent Bernoulli experiments with constant probability POS per round, e.g. throwing a dice n times.
  - Therefore constant success probabilities are required for each sample.
  - This requirement is not fulfilled (PT on the Detection of Highly Infectious Pathogens):

Sample	Negative results (out of 28 participants)
HPB 1	0
HPB 2	1
HPB 3	1
HPB 4	3
HPB 5	9
HPB 6	0
HPB 7	3
HPB 8	4
HPB 9	4

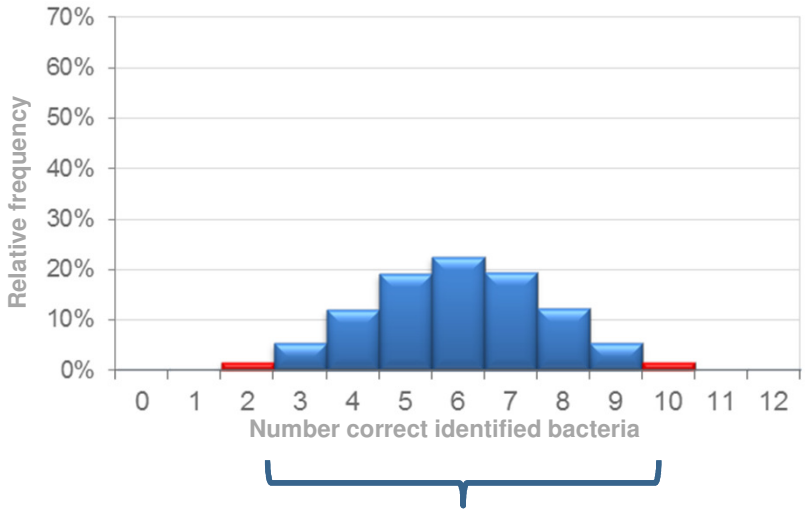
How to derive tolerance limits?

What can happen in case of unequal success probabilities?

Assumption: the probability to identify a bacterium correctly is…	
▪ 5 % for each of 6 samples (very difficult) and ▪ 95 % for each of 6 samples (very easy)	▪ 50 % for each of 12 samples



95 % of all laboratories identify
5 – 7 bacteria correctly (ROS: 42–58 %)



96 % of all laboratories identify
3 – 9 bacteria correctly (ROS: 25–75 %)

Not plausible? Then consider the situation that POS is 0 % for 6 samples and 100 % for the other 6 samples. Result: 6 bacteria will be identified correctly …

How to derive tolerance limits?

Conclusion of simulation study

- If the average probability of success across laboratories and samples is 0.5 and if all samples have an identical level of difficulty, then a ROS of 3/12 = 0.25 is unremarkable.
- But if the level of difficulties differs between the samples, then a ROS of 0.25 is significantly different from the average of 0.5.
- In other words: If the level of difficulties differs between the samples, the 95 % assessment criterion for the minimum number of positive results per participants (which is equivalent to Z=-2) can be stricter than with equal probabilities.
- Or put this paradox in another way: the more variability in LDT, the less variability in ROS (as long as laboratories with constant LCL are considered)

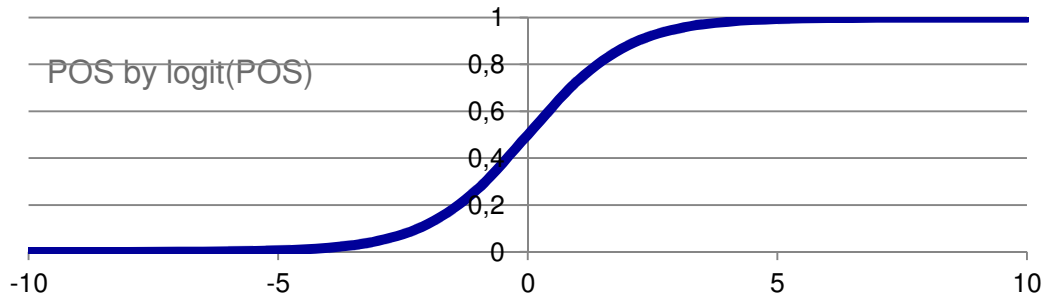
- Conclusion: Level of Competence of the Laboratory (LCL) cannot be considered without considering the Level of Difficulty of the Task (LDT)

The Logit approach

Modeling success rates

- The probability p=POS (Prob. Of Success) for a positive results depends on Level of Competence of the Laboratory (LCL) and Level of Difficulty of the Task (LDT):

  ln[POS/(1-POS)] = logit(POS) = LCL – LDT



- The higher LCL, the higher POS. The higher LDT, the lower POS
- If LCL is tending to –infinity, POS is tending to 0
- If LCL is tending to +infinity, POS is tending to 1
- Ref.: Schilling, Powilleit, Uhlig: Macrozoobenthos interlaboratory comparison on taxonomical identification and counting of marine invertebrates in artificial sediment samples including testing various statistical methods of data evaluation. ACQUAL 2006, 422–429

The Logit approach

Logit approach

Parameter	Explanation
Probability POS	Probability of fulfilling a task correctly (e.g. correct identification)
Chance (odds)	Ratio of the probability for being successful to the probability for not being successful Chance (odds) = exp(mean + level of competence - level of difficulty)
LCL **Level of** **Competence**	Depending on the relative knowledge, experience and practise of the laboratory – laboratory with average competence → level of competence is set to 0 – laboratories with higher competence → positive level of competence – laboratories with lower competence → negative level of competence
LDT **Level of** **Difficulty**	Depending on the relative difficulty of the task – depends on e.g. sample or species (so the probability of correct identification for an average laboratory can vary from species to species)

- LCL and LDT are estimated by means of Maximum-Likelihood

The Logit approach

Example continued (Identification of bacteria species)

	Overall	Z score	HPB 1	HPB 2	HPB 3	HPB 4	HPB 5	HPB 6	HPB 7	HPB 8	HPB 9
No. of laboratories that submitted results	28		28	28	28	28	28	28	28	28	28
No. of participants (according to design)	28		28	28	28	28	28	28	28	28	28
LPOD	0,963		0,994	0,980	0,980	0,943	0,740	0,994	0,943	0,919	0,919
Lower confidence limit of LPOD	0,65		0,876	0,683	0,683	0,423	0,112	0,876	0,423	0,336	0,336
Upper confidence limit of LPOD	0,997		1	0,999	0,999	0,997	0,985	1	0,997	0,996	0,996
01	0,889	-0,725	0,986 +	0,956 +	0,956 +	0,880 +	0,560 -	0,986 +	0,880 +	0,835 +	0,835 +
02	1,000	0,573	0,998 +	0,995 +	0,995 +	0,986 +	0,922 +	0,998 +	0,986 +	0,979 +	0,979 +
03	0,778	-1,836	0,966 +	0,897 +	0,897 +	0,749 +	0,340 -	0,966 +	0,749 +	0,673 -	0,673 +
04	1,000	0,573	0,998 +	0,995 +	0,995 +	0,986 +	0,922 +	0,998 +	0,986 +	0,979 +	0,979 +
05	0,889	-0,725	0,986 +	0,956 +	0,956 +	0,880 +	0,560 -	0,986 +	0,880 +	0,835 +	0,835 +
06	1,000	0,573	0,998 +	0,995 +	0,995 +	0,986 +	0,922 +	0,998 +	0,986 +	0,979 +	0,979 +
07	1,000	0,573	0,998 +	0,995 +	0,995 +	0,986 +	0,922 +	0,998 +	0,986 +	0,979 +	0,979 +
08	0,889	-0,725	0,986 +	0,956 +	0,956 +	0,880 +	0,560 -	0,986 +	0,880 +	0,835 +	0,835 +
09	1,000	0,573	0,998 +	0,995 +	0,995 +	0,986 +	0,922 +	0,998 +	0,986 +	0,979 +	0,979 +
10	1,000	0,573	0,998 +	0,995 +	0,995 +	0,986 +	0,922 +	0,998 +	0,986 +	0,979 +	0,979 +
11	1,000	0,573	0,998 +	0,995 +	0,995 +	0,986 +	0,922 +	0,998 +	0,986 +	0,979 +	0,979 +
12	0,889	-0,725	0,986 +	0,956 +	0,956 +	0,880 +	0,560 -	0,986 +	0,880 +	0,835 +	0,835 +
13	0,889	-0,725	0,986 +	0,956 +	0,956 +	0,880 +	0,560 -	0,986 +	0,880 +	0,835 +	0,835 +
14	1,000	0,573	0,998 +	0,995 +	0,995 +	0,986 +	0,922 +	0,998 +	0,986 +	0,979 +	0,979 +
15	0,556	-3,472	0,891 +	0,712 +	0,712 +	0,457 +	0,127 -	0,891 +	0,457 -	0,368 -	0,368 -
16	0,889	-0,725	0,986 +	0,956 +	0,956 +	0,880 +	0,560 -	0,986 +	0,880 +	0,835 +	0,835 +
17	1,000	0,573	0,998 +	0,995 +	0,995 +	0,986 +	0,922 +	0,998 +	0,986 +	0,979 +	0,979 +
18	1,000	0,573	0,998 +	0,995 +	0,995 +	0,986 +	0,922 +	0,998 +	0,986 +	0,979 +	0,979 +
19	1,000	0,573	0,998 +	0,995 +	0,995 +	0,986 +	0,922 +	0,998 +	0,986 +	0,979 +	0,979 +
20	0,889	-0,725	0,986 +	0,956 +	0,956 -	0,880 +	0,560 +	0,986 +	0,880 +	0,835 +	0,835 +
24	1,000	0,573	0,998 +	0,995 +	0,995 +	0,986 +	0,922 +	0,998 +	0,986 +	0,979 +	0,979 +
25	0,667	-2,739	0,936 +	0,817 +	0,817 +	0,603 +	0,208 -	0,936 +	0,603 +	0,512 -	0,512 -
26	0,889	-0,725	0,986 +	0,956 +	0,956 +	0,880 +	0,560 +	0,986 +	0,880 +	0,835 +	0,835 -
27	1,000	0,573	0,998 +	0,995 +	0,995 +	0,986 +	0,922 +	0,998 +	0,986 +	0,979 +	0,979 +
28	0,778	-1,836	0,966 +	0,897 -	0,897 +	0,749 -	0,340 +	0,966 +	0,749 +	0,673 +	0,673 +
29	0,667	-2,739	0,936 +	0,817 +	0,817 +	0,603 -	0,208 +	0,936 +	0,603 -	0,512 -	0,512 +
30	1,000	0,573	0,998 +	0,995 +	0,995 +	0,986 +	0,922 +	0,998 +	0,986 +	0,979 +	0,979 +

POS across laboratories

Laboratory specific POS

(+) correctly or (-) uncorrectly identified bacterium

Calculation of z scores

Normalised LCL

$$\text{z score} = \frac{\text{LCL} - \text{Average LCL}}{\text{Standard error}} = \frac{\text{LCL}}{\text{Standard error}}$$

- LCL = level of competence of the laboratory
- Average LCL = average level of competence over all laboratories = 0
- Standard error = standard error of the estimated level of competence of the laboratory (derived from Maximum Likelihood estimation; no explicit formula available)

- PT on the Detection of Highly Infectious Pathogens:
  - Four participants are below z=-2
  - Therefore the Logit approach is stricter than the Binomial approach (where only one participant is not succeeding). This is in line with the simulation study presented before.

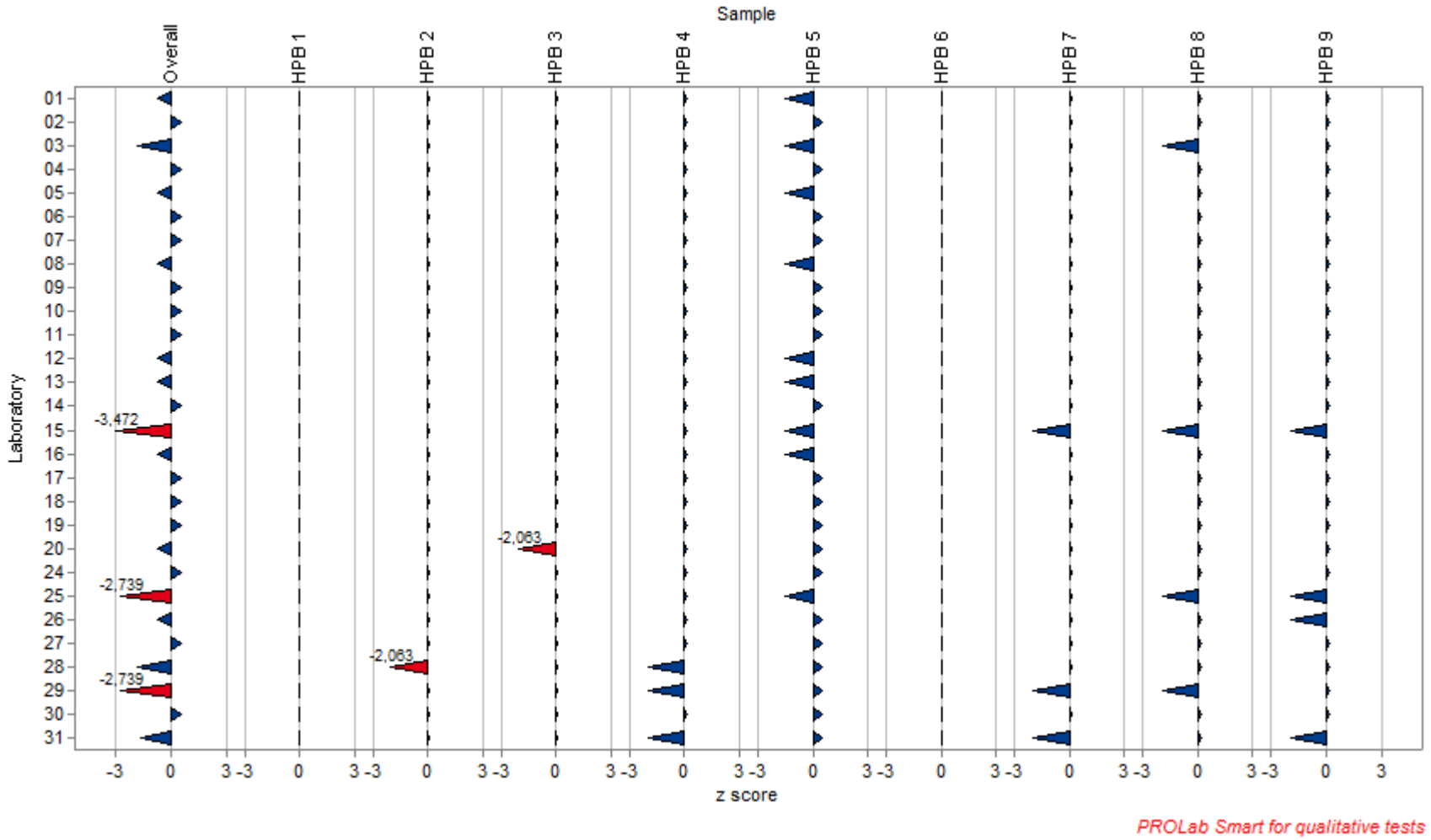
Calculation of z scores

Interpretation is different from z scores of quantitative methods

z score	Interpretation	
< -2	Competence is significantly lower	
-2 ... +2	Laboratory result is not significant different from average	
> 2	Competence is significantly higher	

Calculation of z scores

Example continued (Identification of bacteria species)



Interpretation of results

- Z scores across samples (left column) measure relative competence of the laboratories.

- Z scores for specific samples (columns 2…10):
  - only two outcomes per sample
  - Significant deviations (z <-2) possible only when the probability of a negative result is less than 5 % (only for two tasks with very high LDT, HBP 3+4)

Discussion

- z score is equivalent to LCL, normalized by standard error
- Interpretation of z scores for qualitative methods is **not equivalent** to z scores for quantitative methods
- If LDT is equal for all samples, the Binomial approach and the Maximum Likelihood method provide similar results.
- However, both easy and difficult tasks are required to differentiate between laboratories with lower and higher competence
- If LDT varies considerably between samples, the Maximum Likelihood method provides stricter assessment criteria (allows better identification of lower competence)
- Maximum Likelihood method is available in several software packages and in PROLab POD (www.quodata.de)
- Similar procedures are available for repeated tests (method validation) – also in PROLab POD

Get in touch with us

Meet QuoData at the exhibition hall

We'd like to welcome you at our booth

Attend a live presentation

- Find out about the variety of evaluation methods and PROLab's compelling charts and reports – free and non-binding.

Get a trial version

- Get your free trial version to give PROLab a try.

Get to know QuoData

- We offer a wide range of services and software tools for analytical quality assurance.

Attend a PROLab workshop this fall:

- St. Louis, Missouri (NIST, 19-21 Oct) and
- Dresden, Germany (QuoData, 12-14 Nov)

Let's talk. You are welcome.

Thank you for your kind attention.