Approaches for assessing lab performance from nonbinary qualitative PT data

Eurachem PT 2023
Windsor, UK
28 September 2023

Steffen Uhlig & Bertrand Colson

About QuoData

Berlin and Dresden (Germany)

Team of mathematicians, physicists, biologists, biotechnologists, bioinformaticians, data scientists, computer scientists, software engineers etc.

Developer and operator of web portal for proficiency testing
Licenses PROLAB
PT provider

Design and evaluation of validation studies for CEN/ISO standards, official methods, test kits and in-house methods

Statistical QA helpdesks (e.g. for German Federal Office of Consumer Protection and Food Safety and for US FDA)

Contributions to numerous ISO standards and CODEX guidelines on validation, measurement uncertainty, acceptance sampling and proficiency testing

Introduction

Types of qualitative data

- Binary data
  - Presence/absence of a pathogen
    Presence = 1, absence = 0
  - Identification of bacterial species
    Correct identification = 1, incorrect identification = 0
- Ordinal data
  - Wine quality
    Ordinal scale from 1 (worst) to 10 (best)
- Nominal data
  - Ethnicity
  - Blood type: A, B, AB, O

Evaluation of binary proficiency test data

L-score

- Labs perform a certain number of tasks with positive or negative outcome
- Basic idea for a statistical model of the success probability for Lab *i* and Task *j*:

$$Y_{ij} = \ln\left(\frac{p_{ij}}{1-p_{ij}}\right) = C_i - D_j \text{ where}$$

Y_{ij} represents the Logit of the probability of success

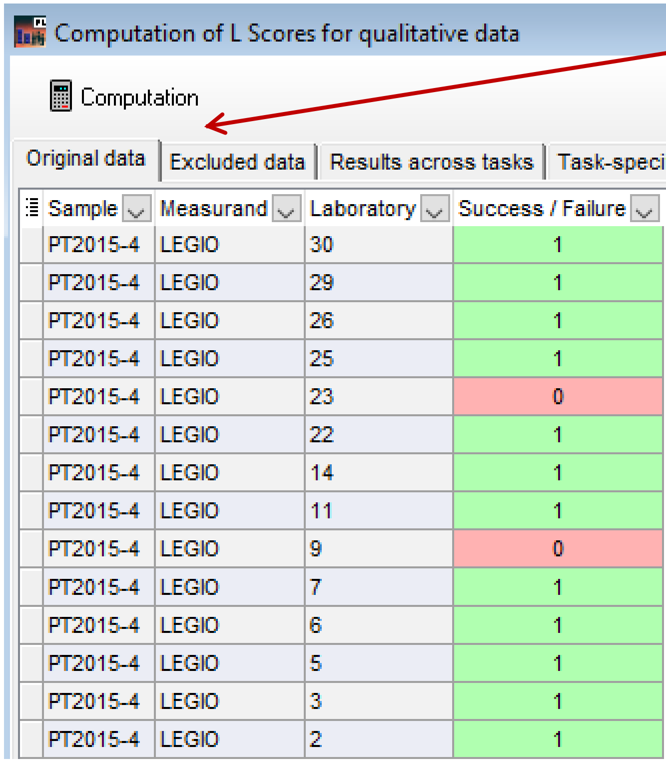
C_i denotes the Competence of Lab *i*

D_j denotes the Level of difficulty of Task *j*

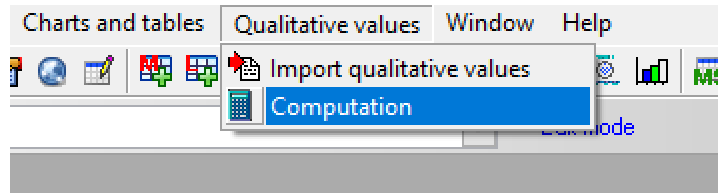
Sample	Laboratories																														
	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
HPB 1	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	No result	No result	No result	+	+	+	+	+	+	+	+
HPB 2	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+				+	+	+	+	-	+	+	+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

Example in PROLab – Legionella in drinking water

Data overview in PROLab

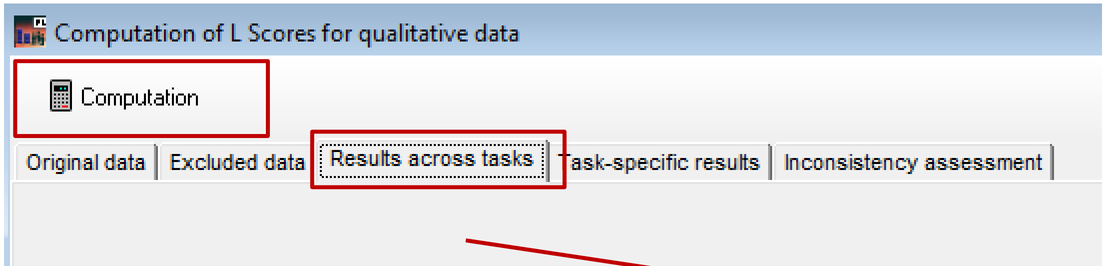


Sample	Measurand	Laboratory	Success / Failure
PT2015-4	LEGIO	30	1
PT2015-4	LEGIO	29	1
PT2015-4	LEGIO	26	1
PT2015-4	LEGIO	25	1
PT2015-4	LEGIO	23	0
PT2015-4	LEGIO	22	1
PT2015-4	LEGIO	14	1
PT2015-4	LEGIO	11	1
PT2015-4	LEGIO	9	0
PT2015-4	LEGIO	7	1
PT2015-4	LEGIO	6	1
PT2015-4	LEGIO	5	1
PT2015-4	LEGIO	3	1
PT2015-4	LEGIO	2	1



Example in PROLab – Legionella in drinking water

Laboratory-specific L-scores (i.e. across tasks)



Laboratory	L score
1	0,427
2	-0,450
3	-0,268
4	-1,037
5	1,229
6	-0,576
7	1,209
8	-1,288
9	-2,179
10	1,280
11	0,269
12	
13	-0,169
14	-0,157
15	-0,515
16	-1,501
17	0,122
18	0,770
19	-0,797
20	0,427
21	-0,169
22	2,660
23	-0,186
24	-0,720

Example in PROLab – Legionella in drinking water

Laboratory- and task-specific L-scores

Computation of L Scores for qualitative data

Computation

Original data | Excluded data | Results across tasks | Task-specific results | Inconsistency assessment

Laboratory △	LEGIO_PT20	LEGIO_PT20	LEGIO_PT20	LEGIO_PT20	LEGIO_PT20	LEGIO_PT20	LEGIO_PT20	LEGIO_PT20	LEGIO_PT20	LEGIO_PT20	LEGIO_PT20	LEGIO_PT20	LEGIO_PT20	LEGIO_PT20	LEGIO_PT20	LEGIO_PT20
1	0,183	0,055	0,183	0,176	0,109	0,054	0,064	0,064	-1,377	0,065	0,065	0,137	0,060	0,125	-0,048	0,119
2										0,065	0,065	0,137	0,060	0,125	-0,048	-1,673
3	-1,458	0,055	0,183	0,176	0,109	0,054	-1,948	-1,948	0,213	0,065	0,065	0,137	0,060	0,125	-0,048	0,119
4	0,183	0,055	0,183	0,176	-1,712	-2,020										
5																
6																
7																
8	0,183	0,055	-1,458	0,176	0,109	0,054	0,064	0,064	0,213	0,065	-1,947	-1,603	-1,982	-1,646	-0,048	0,119
9																
10	0,183	0,055	0,183	0,176	0,109	0,054	0,064	0,064	0,213							
11	0,183	0,055	0,183	-1,478	0,109	0,054	0,064	0,064	0,213	0,065	0,065	0,137	0,060	0,125	-0,048	-1,673
12																
13	0,183	0,055	-1,458	0,176	0,109	0,054	0,064	0,064	0,213							
14																
15																
16	0,183	0,055	-1,458	-1,478	-1,712	0,054	0,064	0,064	-1,377	0,065	0,065	-1,603	0,060	0,125	-0,048	0,119
17	-1,458	0,055	0,183	0,176	0,109	0,054	0,064	0,064	0,213	0,065	0,065	0,137	0,060	0,125	-0,048	0,119
18																
19																
20	0,183	0,055	0,183	0,176	0,109	0,054	0,064	0,064	0,213	0,065	0,065	0,137	0,060	0,125	-0,048	0,119
21	0,183	-2,014	0,183	0,176	0,109	0,054	0,064	0,064	0,213							
22	0,183	0,055	0,183	0,176	0,109	0,054	0,064	0,064	0,213	0,065	0,065	0,137	0,060	0,125	2,074	0,119
23																
24													0,060	-1,646	-0,048	0,119
25	0,183	0,055	0,183	0,176	0,109	0,054	0,064	0,064	0,213	-1,947	0,065	0,137				0,119
26				0,176	0,109	0,054	0,064	0,064	0,213	0,065	0,065	0,137	0,060	0,125	-0,048	0,119
27																
28	0,183	0,055	0,183	0,176	0,109	0,054	0,064	0,064	0,213							
29	-1,458	0,055	0,183	-1,478	0,109	0,054	0,064	0,064	-1,377	0,065	0,065	0,137	0,060	0,125	-0,048	0,119
30																

Scores for ordinal qualitative data

Application of z-scores

- Basic idea: transform the class labels into numerical values.
- For instance, if there are 12 classes, number them 1 through 12
- The z-scores are then calculated on the basis of these numerical values
- The assigned value x_{pt} is the numerical value corresponding to the correct class
- The result x_i is the numerical value corresponding to the class chosen by laboratory i
- The reproducibility standard deviation σ_{pt} is best calculated by means of a robust algorithm (e.g. the Q method) in order to take into account the discrete nature of the numerical values and to minimize the effect of outliers.
- Note that the transformation of class labels described above corresponds to a Euclidian metric in a one-dimensional space with equidistant “distances” between the classes. One could implement a similar approach where the distances are not equidistant.
- This approach is only applicable if the correct class lies somewhere near the middle of the ordered classes.

Scores for nonbinary qualitative data

Ordinal data – Correct class is at either end - application of L_1 scores

If the “correct class” lies at either end of the ordered classes, the z-score approach cannot be applied. For such cases, the L1 approach can be applied.

An added degree of sophistication: the level of difficulty/penalty for error can be mapped/controlled via difference scores.

Example: Identification of firearms

- 5 levels of conclusion (classes), labelled A, B, C, D, E
- A = “the cartridge matches the firearm”
  B = “similar”
  C = “possible match”
  D = “clear differences”
  E = “all but certain that the cartridge does not match the firearm”
- For a given task, the correct class (here: either A or E) is known.

L_1-scores

- A Probit model can be fitted that takes in account the actual distribution of *Difference scores*

$$\mathbf{L_1 = \theta_0+\theta_1+\ldots+\theta_j - \beta_i}$$

  where

  – θ_0, θ_1, θ_2, θ_3 and θ_4 are the estimated weights of the *Difference scores* 0, 0.5, 1, 3 and 4

  – The index *j* represents the *difference score* corresponding to the submitted Conclusion Level

  – $\boldsymbol{\beta}_i$ denotes the estimated level of difficulty of Test set i (the higher this coefficient, the greater the difficulty)

- **Interpretation:**

  – $\mathbf{L_1}$ < 2 → acceptable
  – $\mathbf{L_1}$ > 2 → questionable performance
  – $\mathbf{L_1}$ > 3 → unsatisfactory performance

L_1-scores

Theta values and controlled penalization via difference scores

Test set	Number of laboratories having submitted Conclusion Level…				
	A	B	C	D	E
2	49	4	0	1	0
6	38	10	4	2	0

Test set	Result	Difference
2	A	0
	B	0
	C	0.5
	D	3
	E	4
6	A	0
	B	0.5
	C	1
	D	3
	E	4

		Difference score			
		0	1	2	3
Set		1.3492	1.8133	1.9091	2.0509
1	0.00				
2	-0.70	98.0%	99.4%	99.5%	99.7%
3	-0.32				
4	-0.76				
5	-0.64				
6	0.68	74.9%	87.2%	89.1%	91.5%
7	0.50				
8	0.55				
9	-0.48				
10	-0.40				

L1-Scores	Difference score				
	0	0.5	1	3	4
1					
2	0	2.22	2.55	2.67	2.97
3					
4					
5					
6	0	0.88	1.18	1.30	1.72
7					
8					
9					
10					

L_1-scores corresponding to the *Difference scores*

	Difference score				
Test set	0	0.5	1	3	4
1	0	1.54	1.86	1.98	2.32
2	0	2.22	2.55	2.67	2.97
3	0	1.85	2.18	2.29	2.62
4	0	2.29	2.62	2.74	3.03
5	0	2.16	2.50	2.61	2.91
6	0	0.88	1.18	1.30	1.72
7	0	1.05	1.36	1.48	1.88
8	0	1.01	1.31	1.43	1.84
9	0	2.01	2.34	2.45	2.76
10	0	1.93	2.25	2.37	2.69

L_1-scores corresponding to the *Conclusion Levels*

Test set	Correct answer	Percentage wrong	A	B	C	D	E
1	A	24.1 %	N=41 **L=0**	N=8 **L=0**	N=4 **L=1.54**	N=0 **L=1.98**	N=1 **L=2.32**
2	A	9.3 %	N=49 **L=0**	N=4 **L=0**	N=0 **L=2.22**	N=1 **L=2.67**	N=0 **L=2.97**
3	D/E	13.0 %	N=2 **L=2.62**	N=0 **L=2.29**	N=5 **L=0**	N=19 **L=0**	N=28 **L=0**
4	A	20.4 %	N=43 **L=0**	N=10 **L=0**	N=1 **L=2.29**	N=0 **L=2.74**	N=0 **L=3.03**
5	D/E	3.7 %	N=1 **L=2.91**	N=0 **L=2.61**	N=1 **L=0**	N=10 **L=0**	N=42 **L=0**
6	A	29.6 %	N=38 **L=0**	N=10 **L=0.88**	N=4 **L=1.18**	N=2 **L=1.30**	N=0 **L=1.72**
7	A	53.7 %	N=25 **L=0**	N=19 **L=0**	N=5 **L=1.05**	N=0 **L=1.48**	N=5 **L=1.88**
8	E	57.4 %	N=4 **L=1.84**	N=2 **L=1.43**	N=5 **L=1.01**	N=20 **L=0**	N=23 **L=0**
9	A	3.7 %	N=52 **L=0**	N=2 **L=2.01**	N=0 **L=2.34**	N=0 **L=2.45**	N=0 **L=2.76**
10	E	38.9 %	N=1 **L=2.69**	N=0 **L=2.37**	N=1 **L=1.93**	N=19 **L=0**	N=33 **L=0**

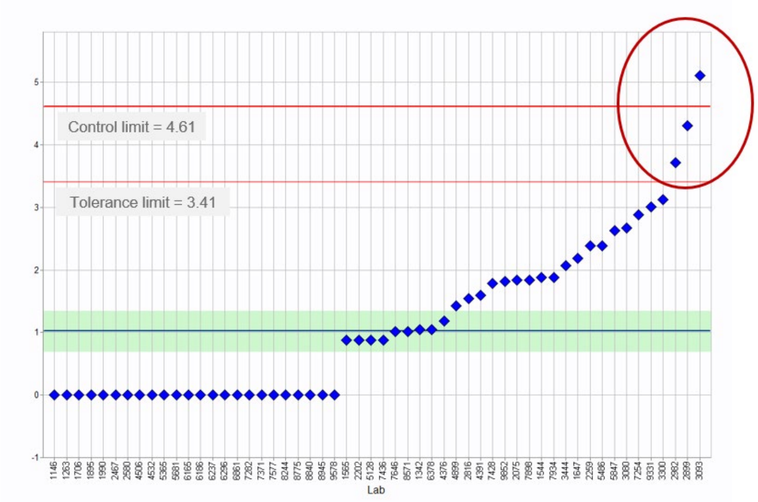
L_1-scores

Lab evaluation across test sets (overall scores)

- Combined probabilities corresponding to the invidivual (test set-specific) scores

  The overall scores can be obtained by multiplying the individual probability values. For instance, an L_1-score of 2 corresponds to a probability of around 5 % and an L_1-score of 1 corresponds to a probability of around 32 %. Accordingly, the combined probability is around $0.05 \cdot 0.32 \approx 0.014$, that is 1.4 %. This, in turn, would correspond to a combined L_1-score of 2.4.

- The overall assessment of laboratory performance is therefore performed by computing "robust" **tolerance** and **control limits** for the overall L_1-scores.



Summary

- The z-score approach is relatively simple
- Constraint: the “correct class” should lie near the middle of the range of classes
- Separate evaluation per test set (sample) – i.e. no combined evaluation of “level of difficulty” and “lab competence”

Advantages of L_1 scores

- Flexibility regarding the position of the “correct class”
- Combined evaluation of task difficulty and lab performance
- Map level of difficulty via difference scores