The First
International Proficiency Testing Conference

Sinaia, România
11th – 13th October, 2007

EVALUATION OF PROFICIENCY TEST DATA BY DIFFERENT STATISTICAL METHODS COMPARISON

Pedro Rosario, José Luis Martínez and José Miguel Silván

RPS-Qualitas S.L., Marqués de Corbera, 62 , 7º D – 28017 Madrid - Spain

[1] pedro.rosario@rpsqualitas.es; rps@rpsqualitas.es

Abstract

In analytical chemistry, proficiency testing usually consists in tests that laboratories conduct under routine conditions and report the result to PT provider who then converts the result to a score which helps the participant to assess the accuracy of theresult.
The aim of this work is to show PT providers, accreditations bodies and participating laboratories that different scoring results can be achieved depending on the selected evaluation system.
The influence of different evaluation techniques on the results of an interlaboratory comparison for the determination of gold in precious metals alloys was investigated. The results of 19 participating laboratories were evaluated by means of the following procedures:
1.- ISO 5725 / Harmonized Protocol / ISO Guide 43-1
2.- Robust Methods:2.A. Robust procedure; 2.B. ISO 13528
3.- Fitness for purpose
The evaluation of the same PT data showed very interesting issues depending on the different scoring systems that were used as well as robustness of statistical methods for detecting outliers. As a general rule, laboratories with scoring Z > 2 offered clearly poorer performance in robust approaches than classical ones.
The selection and comparison of different scoring systems have to be done very carefully because sometimes they are not the best approach for studying the data population or the more appropriate one for evaluating the distribution of the data. Finally it should be taking into account that sometimes the robust scoring systems are not always suitable for evaluating the results of certain PT schemes.

Key words

Proficiency Test Data - Statistical Methods - Evaluation - ISO 5725 - Harmonized Protocol - ISO 13528 - Robust statistics - Fitness for purpose

1 INTRODUCTION

In the framework of quality management system, proficiency tests schemes (PTS) become an important and useful tool for comparing results and so verifying technical competence of laboratories objectively. Moreover, in the accreditation requirements of ISO/IEC 17025 standard [1], laboratories are asked for participating in proficiency tests if they are available and suitable, regardless of other quality assurance activities for checking validity of test results. The use of a specific statistical protocol of the data set coming from a proficiency testing scheme in which the assigned value and the assessment of performance is estimated from the consensus of participant results, has been largely addressed as a matter of discussion in an attempt of establish a certain degree of comparability among the different approaches that can be tackled. Since the vast majority of proficiency tests are performed by establishing a standardised result, often in terms of a Z-score value for each participant laboratory, it is essential to achieve a significant measurement not only for the assigned value but also for the estimation of the interlaboratory precision. Both parameters should be expected to adequately describe the analytical procedure involved in order to enable a sort of comparability among the parameters applied in the statistic procedure. The underlying assumptions of this parametric statistics lie in the fact of normality in data distribution, so in some cases a further step should be considered in terms of statistical approach. This is perhaps the main interest to do not apply conventional estimators in favour of robust statistical methods in the evaluation of proficiency testing scheme data, which provide a powerful tool by means of the performance of calculations widely undertaken in ISO 13528 [2], that deals with the statistical methods for use in proficiency testing by interlaboratory comparisons. Apart from this reason, the use of mean values in the evaluation of a proficiency test calculations are known to have a poor stability when considering the effect of outliers data, so the application of robust methods of estimation should be considered. This fact could not be of extreme importance in many of the fields of application covered by intercomparison statistics estimated by classical approach, but in certain cases should be taken into account, particularly when the determination involves complex material or when the diversity in the origin of participant laboratories is not negligible.
Accordingly, in this paper it is proposed to perform the intercomparison parameters by calculating both statistical approaches; thus, the PT-provider has a strong basis to provide information on the best estimation in order to fulfil the participant claims [3].
Prior to the exposition of the calculations carried out in this round, it is worthy to state that the revision of the Harmonised Protocol for the Proficiency Testing in Analytical Chemistry issued by the IUPAC [4], encourages to base the scoring methods on *fitness for purpose criterion*, envisaged by the PT-provider in the specific application according to the particular circumstances of the determination.

2 EXPERIMENTAL

2.1 Materials

RPS-Qualitas S.L.has carried out a Proficiency testing by an interlaboratory program for comparison on precious metals alloys. These tests were carried out on gold samples following cupellation method.

The test samples consisted of ten encoded pieces of gold alloy and fineness (585 ‰ aprox.), coming each one from the same ingot.

2.2 Procedures

The laboratories carried out the tests taking into account the instructions given by organizer and processing the samples as routine ones. They were asked to test eight samples at least.
This Program was designed for evaluating quality tests and technical competence of participants laboratories. [5]
The participant identity was kept under secret, thus guaranteeing data confidentiality. The coordinator had the right to eject from the programme any laboratory that unjustifiably did not meet the terms for sending the results.
Each participant had an assigned laboratory identification code number, in order to maintain the relevant information on a confidential basis.

2.2.1 Homogeneity

Although the variability in the production run of the ingots was not exactly known, the degree of homogeneity of the test samples was determined and checked by using the statistical criteria based on analysis of variance (ISO 13528 [2] and IUPAC [4]).
Once successfully performed this verification, the suitability of the samples was considered satisfactory for the purpose of this proficiency programme, so the test samples to be delivered were both homogeneous and stable enough.

2.2.2 Statistical design

The statistical handling has been planned in compliance with ISO 13528 [2] , ISO Guide 43 [6] and ISO 5725 standards [7], with a number of sequential steps:

- Detection and removal of outlier data
- Calculation of the summary parameters
- Calculation of the performance indicators
- Graphical presentation of the obtained results.

2.2.2.1 Outlier Test. ISO Guide 43

2.2.2.1.1 Cochran´s Test

The Cochran outlier test [7] is used to check the assumption that between laboratories only small differences exist in the within-laboratory variances.

C is calculated by using the maximum variance of all results and the sum of all variances:

$$C = \frac{S_{máx}^2}{\sum_{i=1}^{p} S_i^2} \qquad (1)$$

Where:

p = total number of standard deviations.
S_i = Standard deviations
$S_{máx}$ = Maximum standard deviation of all results.

The result of this test was: *There was one outlier. (Lab. 31)*

2.2.2.1.2 Grubbs` Test

Once performed the verification of no outliers in laboratories variances, Grubbs' test [7] is used to determine whether the largest or smallest observation in a set of data is an outlier.

For this aim, Grubb's statistic G_h is calculated for largest value as follows:

$$G_h = \left(x_h - \bar{x} \right) / s \qquad (2)$$

For the smallest value Grubb's statistic G_l is calculated as follows:

$$G_l = \left(\bar{x} - x_l \right) / s \qquad (3)$$

These obtained values are compared with critical ones of Grubbs' test tables respectively.

The result of this test was: *There was one outlier. (Lab. 31)*

In order to achieve more confidence for coping with outliers, Double Grubbs' test [7] was used to determine whether the two largest or two smallest values might be outliers.

In our statistical study we evaluated the two largest and the two smallest values from participating laboratories.

The result of this test was: *There was no outliers.*

2.2.3 Assigned value

To determine the assigned value, this Program took into account the following criteria:

- Centre Value or Grand Mean, given to each sample that is assumed to be conventionally true. It is a value obtained from *the average of the whole laboratories results after excluding those which have been found anomalous* under the Grubbs'/Cochran's Tests. (ISO Guide 43) [6]

- The value obtained from *the robust mean of the whole laboratories results* without previous exclusion. (ISO 13528) [2]

- Summary parameters, including the standard deviation of the repeatability and the reproducibility, relative standard deviation (RSD), or other robust measures. [2], [7].

2.2.4 Data evaluation

2.2.4.1 Z-score values

Standardised Value (Z-score), is an indication of the performance of a participant, depending on its interpretation, showing as satisfactory, questionable or unsatisfactory.

For each test, the Z-score of laboratory is calculated as:

$$Z_i = (X_i - V_c) / \sigma \qquad (4)$$

where:

X_i = each individual laboratory mean result
V_c = centre value as an estimation of the assigned value
σ = the reproducibility standard deviation

2.2.4.2 Results evaluation judgement

According to *ISO Guide 43* the following judgement is made:

- Results leading to $|Z| \leq 2$ are satisfactory.
- Results leading to $|Z| > 3$ are unsatisfactory.
- All other results ($2 < |Z| < 3$) are questionable.

According to *ISO 13528* the following judgement is made:

- Results leading to $|Z| \leq 2$ are satisfactory.
- Results leading to $|Z| > 3$ are considered to give "action signal".
- All other results ($2 < |Z| < 3$) are considered to give "warning signal".

Note: A single "action signal", or " warning signals" in two successive rounds, shall be taken as evidence that an anomaly has occurred that requires investigation.

2.3 Participants

19 Laboratories from thirteen countries have participated in this PTS.

3 RESULTS AND DISCUSSIONS

3.1 Laboratory results

The results of participating laboratories are shown in table 1. The Z-score values are indicated in tables and appropriate figures as well as the general statistical parameters of this Program.

Table 1 – Participant laboratories results (‰)

Laboratory Code	replicates	mean	Std. dev.	C.V.
Lab.03	4	586,7	0,08	0,01
Lab.05	4	586,6	0,05	0,01
Lab.06	5	585,5	0,05	0,01
Lab.07	4	586,9	0,08	0,01
Lab.08	10	585,6	0,10	0,02
Lab.10	4	586,9	0,06	0,01
Lab.12	6	586,8	0,08	0,01
Lab.13	5	585,8	0,08	0,01
Lab.14	5	585,6	0,13	0,02
Lab.15	7	586,8	0,07	0,01
Lab.16	6	586,9	0,10	0,02
Lab.20	5	586,7	0,25	0,04
Lab.21	5	586,8	0,07	0,01
Lab.22	10	586,6	0,27	0,05
Lab.24	6	587,0	0,30	0,05
Lab.27	5	586,4	0,09	0,02
Lab.29	5	586,8	0,10	0,02
Lab.30	5	587,1	0,19	0,03
Lab.31	4	584,1	0,92	0,16

3.2 Calculation of the intercomparison parameters.

In the development of this interlaboratory comparison, the statistical protocol is conducted to achieve the estimation of both the true value and the target standard deviation of the proficiency test. According to that, a couple of statistical approaches have been considered in order to evaluate the analytical results received from the participants: one conventional method based on ISO 5725 calculations and a second approach based on the application of robust statistics by means of the algorithms A and S as described in ISO 13528.

With regard to the assigned value, no relevant differences have been found between the statistical methods applied, concluding in small discrepancies less than 0,2 ‰, so the robust average of the whole participants without outlier detection was chosen as the best estimation of the true value of the concentration of gold in this scheme (robust average = 586,7 ‰).

However, as for the estimation of the interlaboratory precision, the two above-mentioned protocols were compared, leading to some relevant differences in terms of reproducibility standard deviation. As a result of that, the statistical treatment of the results revealed that following classical methods the %RSD value is roughly two times larger than the one obtained according to robust estimation with the same consideration about outliers that were expressed previously.

Then, the reproducibility standard deviation in this interlaboratory comparison was chosen to be determined from robust estimation by iterative calculation as expressed in algorithm S in ISO 13528 (SD robust = 0,31 ‰).

Table 2 - Comparative results of assigned value and interlaboratory precision between classical vs. robust statistical methods

	ISO 5725	ISO 13528
Assigned value	586,5 ‰	586,7 ‰
Reproducibility std. deviation	0,56 ‰	0,31 ‰
RSD	0,10	0,05

3.3 Discussion on the assessment of laboratories performance.

This topic is intended to explain a wider range of possibilities that the PT-scheme provider might consider to evaluate the data submitted by the participants, so a number of cases with a different approach were estimated to cover these statistical principles. [8].

On the whole, in order to evaluate the assessment of each laboratory performance by interpreting Z-score values, the results were compared using four methods: the traditional approach according to ISO 5725 including outlier detection; two robust statistical methods, based on median and NIQR, and based on Huber test and the algorithms detailed in ISO 13528, respectively; finally was considered a practical approach that sets up a specified target value taking into account a fit-for-purpose criterion.

The values considered in the expression of Z-score (assigned value and standard deviation) have been calculated as follows in each one of the four statistical approaches:

1. ISO 5725: general mean and reproducibility standard deviation, with outlier detection [7]

2. Median and NIQR method: median of the whole data and normalized interquartile range [9]
3. ISO 13528: robust average and robust standard deviation calculated according to algorithms A and S, without outlier detection [10]
4. Fit-for-purpose criterion: robust average and a target reproducibility standard deviation value according to a fixed %RSD from appropriate past PT-rounds at this level of concentration [11]

As a result of that, in view of the analytical laboratory performances following each one of the four protocols considered (Table 3), it can be stated that fourteen participants show Z-score values considered as acceptable (|Z| [2) regardless the statistical method applied. The reason for this behaviour lies in the fact that they are the laboratories that provide the most balanced results with less deviations in data spread due to common analytical techniques, so the statistical protocol has not relevant influence in their performance.

Furthermore, in terms of distribution of Z-score values corresponding to the other five laboratories, a certain trend in the spread is revealed. Thus, according to ISO 5725, Z-score values comply with the acceptance criteria, particularly because one laboratory data have been rejected as outlier.

Table –3. Summary of the overall Z-score results obtained by participant laboratories reported following the different statistical protocols

Participant	ISO 5725	Median & NIQR	ISO 13528	Fit-for-purpose
Lab.31	outlier	-4,76	-8,27	-5,68
Lab.06	-1,84	-2,28	-3,86	-2,65
Lab.14	-1,63	-2,06	-3,48	-2,39
Lab.08	-1,52	-1,95	-3,28	-2,25
Lab.13	-1,30	-1,73	-2,89	-1,98
Lab.27	-0,09	-0,50	-0,69	-0,48
Lab.22	0,15	-0,26	-0,27	-0,19
Lab.05	0,16	-0,25	-0,26	-0,18
Lab.03	0,34	-0,06	0,08	0,06
Lab.20	0,40	0,00	0,19	0,13
Lab.15	0,53	0,13	0,43	0,29
Lab.21	0,56	0,16	0,47	0,32
Lab.29	0,56	0,16	0,47	0,32
Lab.12	0,62	0,22	0,58	0,40
Lab.10	0,65	0,25	0,64	0,44
Lab.07	0,74	0,34	0,80	0,55
Lab.16	0,81	0,41	0,92	0,63
Lab.24	0,86	0,46	1,01	0,70
Lab.30	1,10	0,71	1,45	1,00

On the other hand, in order to avoid the influence of extreme results, the application of robust statistical methods [9], [10] brings about significant larger Z-score values since no outlier elimination is applied. In this line, when the calculation is performed by using median & NIQR method, Z-score values are slightly smaller than the corresponding ones estimated following the robust method based on ISO 13528. In this case, one laboratory is given a warning signal whereas four laboratories shows Z-score values considered to give action signals, so that special investigation is required for the laboratory previously considered as outlier in the parametric approach.

At last, when applying a fit-for-purpose criterion [11] according to end-user requirement where the performance ratio is determined by the PT-provider itself and no outlier rejection of data has been considered due to the own statistic protocol formulation, it is showed that Z-score values offer results considered as satisfactory in fifteen laboratories, whilst is considered to give a warning signal in three participants and there is one single laboratory that widely exhibit an action signal that requires further investigation.

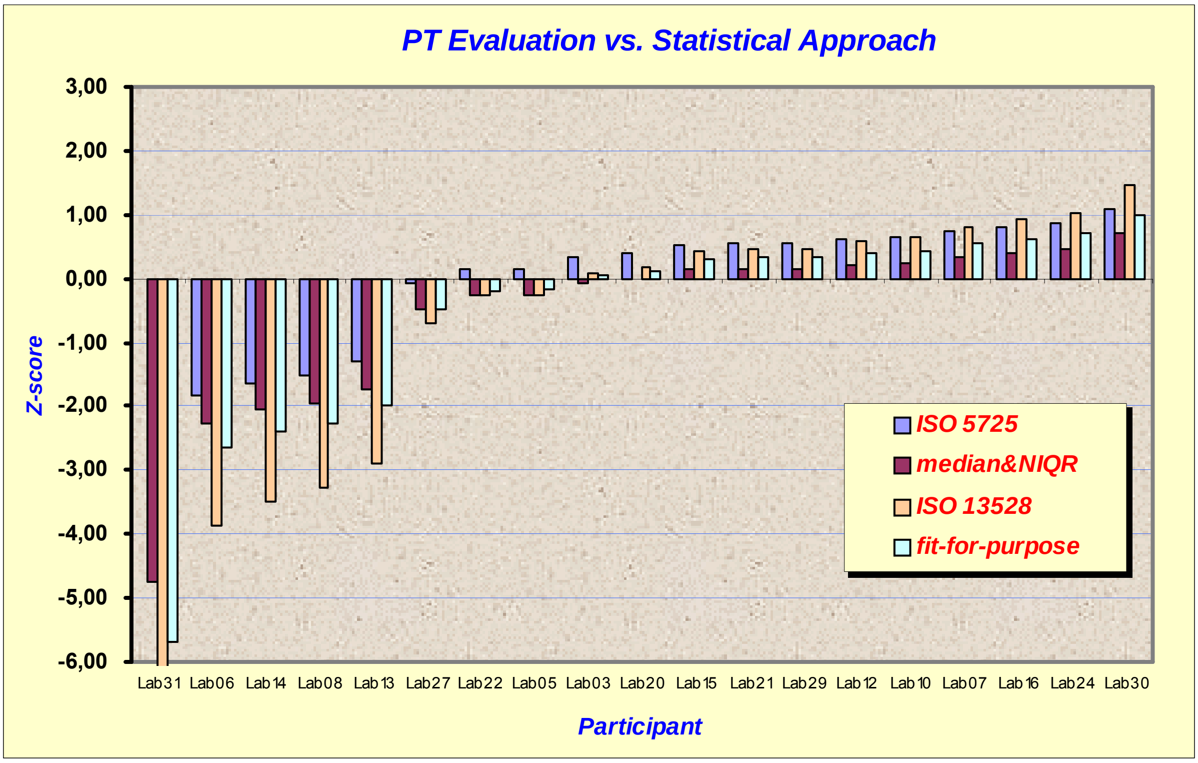

Figure 1- Graphic summary of the overall Z-score results obtained by participant laboratories reported following the different statistical protocols.

4 CONCLUSIONS

On balance, due to the fact that proficiency testing participant data are usually heavy-tailed and the presence of outlier is very common, these arguments can lead to overestimations of the standard deviation value in which more Z-score values are considered as satisfactory, unless robust statistical methods were applied since they are more insensitive to anomalies.

Then, this kind of robust protocols are particularly applicable to look-like normal distribution data with no more than 10% of outliers, unimodal and roughly symmetric, apart from cases when it is assumed that all participants do not have the same analytical performance. However, it is observed that median & NIQR method is more robust for asymmetry data, while in case of multimodal or skewed distribution of data, the application of mixture models and kernel density functions should be considered.

In this report, the application of both classical and robust statistical methods when dealing with proficiency test data clearly shows that mean values are quite similar, whereas significant differences in standard deviation value have been found, in some cases too large for fitting the objective of this interlaboratory programme.

Furthermore, it is quite important to obtain an appropriate estimation of the overall standard deviation parameter in a suitable way, that allows not only to describe the analytical method in terms of precision but also to provide performance assessment compatible with the intercomparison requirements.

Finally, in this line, the application of a fit-for-purpose criterion should describe the end-user requirement and must be consistent from round to round, so that scores in successive rounds might be comparable. The specification of a target value in terms of relative standard deviation involves more a quality goal that the data should meet to reflect fitness for purpose, rather than a simple description of the data results.

REFERENCES

[1] UNE-EN ISO/IEC 17025:2005. Requisitos generales relativos a la competencia de los laboratorios de ensayo y calibración

[2] ISO 13528:2005. Statistical methods for use in proficiency testing by interlaboratory comparisons

[3] Van der Venn, A.M.H. and Hafkenscheid, T.L.: Harmonisation of proficiency testing schemes. Accred. Qual. Assur. 9, 657-661, (2003)

[4] The International Harmonized Protocol for the Proficiency Testing of Analytical Chemistry Laboratories. Pure Appl. Chem., 78, 1; 145-196*. (*2006)

[5] UNE 66543-1 IN:1999. Ensayos de aptitud por intercomparación de laboratorios. Parte 1: Desarrollo y aplicación de programas de ensayo de aptitud

[6] ISO/IEC Guide 43:1997. Proficiency Testing by Interlaboratory Comparisons

[7] ISO 5725-2:1994. Accuracy (trueness and precision) of measurement methods and results. Part 2: Basic method for the determination of repeatability and reproducibility of a standard measurement method

[8] Uhlig, S. and Lischer, P: Statistically-based performance characteristics in laboratory performance studies. Analyst, 123 , 167-172 (1998)

[9] Rousseeuw P.J. and Leroy A.M. Robust regression and outlier detection. Wiley, New York. (1987)

[10] Analytical Methods Committee. Robust Statistic Part I & II. Analyst 114, 1693-1702 (1989)

[11] Thompson, M. and Ellison, S.L.R. : Fitness for purpose – the integrating theme of the revised Harmonised Protocol for Proficiency Testing in Analytical Chemistry Laboratories. Accred. Qual. Assur. 11, 373-378, (2006)