Evaluation of proficiency testing in the area of seeds by the Z-score method and cluster analysis

Kataoka, Verônica Yumi (1st Author)
Bandeirante University of São Paulo, Post-Graduate in Mathematics Education
Av Braz Leme, 3029, Santana, São Paulo (02022-011), Brazil
E-mail: veronicayumi@terra.com.br

Oliveira, Marcelo Silva de (2nd Author)
Federal University of Lavras, Departament of Exact Sciences
E-mail: marcelo.oliveira@ufla.br

Carvalho, Maria Laene Moreira de (3rd Author)
Federal University of Lavas, Departament of Agriculture
Caixa Postal 3037, Lavars (37200-000), Brazil
E-mail: mlaenemc@ufla.br

INTRODUCTION

The participation of laboratories in interlaboratory programmes is one of the mechanisms of quality control of their expected results in NBR ISO/IEC 17025 (Association of Analytical Communities, 1993). These types of comparative testing laboratories can assist in detecting and, therefore, correcting possible sources of error. According to the statement of that rule, to maintain the quality of a laboratory's results, they must be obtained by a validated method, followed by the estimation of measurements of uncertainty, that is, an assessment of accuracy and precision.

Proficiency testing is one of the types of interlaboratory programmes that allow the performance of a laboratory to be determined by comparing the use of measurements in materials that are homogeneous or similar in at least two laboratories under predetermined conditions. For the process of maintaining accreditation, the International Seed Testing Association (ISTA) requires the participation of seed laboratories in proficiency tests. The purpose of the ISTA proficiency tests is not to identify the best seeds analysis laboratories in the world, but rather to highlight those that do not meet the minimum performance that is reasonably expected of an accredited laboratory and to determine whether the laboratory is actually taking corrective actions so that its performance will reach at least the minimum level (ISTA, 2007).

The statistical technique used by ISTA in evaluating the results of a proficiency test is the Z-score method. In this context, the objective of this study was to evaluate whether the technique of cluster analysis could also be appropriate for analysing the performance of twelve laboratories in a proficiency test in the area of seeds, more specifically for the germination test.

PROFICIENCY TEST

The service provided by a seed laboratory is the assessment of seed quality through analytical procedures, which according to Brazilian law must follow the Rules for Seed Analysis (RAS; Brazil, 1992). These rules are based on uniformity of procedures and specify the different methods of analysis that should be employed, the maximum sizes for seed lots, the minimum weight of the sample, and the use of tables of tolerance for different types of tests on seeds (Zorato, 2005).

Rennie and Tomlin (1984) observed in their studies that even following the RAS there is a difference in results between laboratories which is greater than the variation between repetitions or between tests in the same laboratory. For the germination test, Miles (1963, cited in Oliveira and Cicero, 1996), states that the causes of significant differences in test results or between germination tests may be: chance, lack of equipment (including change of environment within the incubator), deficiency of the method, technical failure, error or inconsistency in distinguishing between normal and abnormal seedlings, fungi and bacteria, chemicals in the seed, inaccurate counts or records, non-random sampling, and change in germination percentage between tests.

Then, the participation of laboratories in, for example, proficiency testing for comparison can assist in finding sources of error and therefore in the internal quality control of the laboratory. In proficiency testing, the materials sent to the laboratories should be stable and homogeneous so that any significant variation observed is not attributed to them. In the case of the area of seeds, the ISTA promotes proficiency tests that are performed per scope. Accredited laboratories in the area are required to participate, and the voluntary adherence of accredited laboratories in other scopes as well as non-accredited ones that want to assess themselves is also permitted (ISTA, 2007). Another feature of the ISTA proficiency tests is that at least two annual evaluations are conducted, each of which is called a round. For the germination test, three rounds per year are usually held and the final assessment is made within three years.

Z-SCORE METHOD

The Z-score technique is adopted by ISTA to evaluate the technical performance of laboratories, mainly in proficiency tests, giving a good indication of the analytical competence of the laboratory. The principle of this method is the calculation of Z-scores based on the determination of the accepted reference value, represented by the estimate of the average, after removing the outliers (ISTA, 2007). Performance evaluation of the laboratory, for example, for the germination test, is done according to the following criteria: $|Z| \leq 2$ indicates satisfactory performance; if $2 < |Z| < 3$ the performance is questionable; and if $|Z| \geq 3$ performance is considered unsatisfactory. In the ISTA proficiency test programme, these criteria are adjusted because three samples are sent to each laboratory in each round, and thus the Z-scores calculated for each one are summed and assigned the following ratings: A: sum of absolute Z-scores ≤ 3.5; B: 3.5 < sum of absolute Z-scores ≤ 5.3; C: 5.3 < sum of absolute Z-scores ≤ 7.0; and BMP (*below minimum performance*): sum of absolute Z-scores > 7.0. These ratings are converted to "notes" in each round and the classification at the end of six rounds is defined as described in Table 1.

***Table 1 Overall ranking of all the tests, based on scores after six rounds* (ISTA, 2007)**

Round rating	Grade given	Interval	General Classification
A	5	28–30	A
B	4	21–27	B
C	3	16–20	C
BMP	0	Below 16	BMP

CLUSTER ANALYSIS

Another technique that can be used to evaluate the performance of laboratories in proficiency testing is cluster analysis. For this procedure, it is essential to first define the dissimilarity measure to be used. The smaller its value, the more similar are the elements being compared. There are

several types of dissimilarity measures, and the choice depends on several factors, including the type of variable under study. According to Ferreira (2008), the Euclidean distance is appropriate for cases in which groups of variables have similar scales because otherwise variables with greater variability will dominate the ranking of distances. In the context of a proficiency test, measurements of variables (samples) always have the same scale.

The second step in using this technique is the determination of the clustering method. Clustering methods are divided into hierarchical and non-hierarchical. According to Rencher (2002), in non-hierarchical methods, the number of groups should be defined beforehand and the elements are allocated optimally; hierarchical methods are those in which the elements are sorted into groups at different stages, that is hierarchically, producing the final result, a graph called a dendrogram, which shows the elements and their points of merging or dividing the groups formed in each stage (Ferreira, 2008). In addition, a hierarchical method is referred to as agglomerative when the clustering process starts with g groups, each containing one of the elements, and ends with a single group comprising all the elements. Agglomerative techniques commonly used are: nearest neighbour methods, the farthest neighbour, average linkage, centroid, median, and Ward's.

METHOD

Three simulations were carried out to generate the database. Each simulation represented the results of a germination test with twelve laboratories participating in proficiency testing. These twelve laboratories formed a composite of two discrepant groups in relation to some statistical parameter(s). Group 1 was composed of ten laboratories and group 2 of only two. Determination of the number of laboratories per group was established by an empirical decision, because in a proficiency test it is expected that the results of accredited laboratories will not be too discordant. Thus, only two laboratories were included with some measure of variance to assess whether the techniques would be adequate to detect such differences.

For the definition of the parameters (mean, matrices of variance and covariance) of the three simulations, the actual results of germination tests were used. With the initial parameters set, the three simulations were generated under the following conditions: simulation 1: means were equal between groups while matrices of variances and covariances differed; simulation 2: means differed while matrices of variance and covariance were equal between groups; simulation 3, means and matrices of variance and covariance differed.

Despite the recommendation by the ISTA (2007) that six rounds of proficiency tests should be held, this study considered only one round, since it would not make practical sense to use, in the simulations, the same statistical parameters to generate the results of the remaining rounds.

To calculate the indices of bias, accuracy, and precision of each laboratory using the Z-score method, the methodology adopted in the ISTA proficiency tests (2002) was followed: identification of outliers, determination of the Z-score, one round of classification, calculation of bias, calculation of precision, and calculation of accuracy. In the stage of identification of outliers, the laboratory was considered an outlier if the estimate of its average in a given sample (a-th sample) was not contained in a confidence interval for the median. Then, considering only non-discrepant laboratories, $\overline{Y}_a$ (estimator of the mean percentage germination of the a-th sample) and S_a (standard deviation of the estimator of the a-th sample) were calculated for each sample. Then we calculated the Z-score for each laboratory in a given sample (Z_{al}), including the outliers, according to the

expression $Z_{al} = \left(\overline{Y}_{al} - \overline{Y}_{a}\right) / S_a$, where $\overline{Y}_{al}$ is the estimator of average germination percentage of the l-th laboratory in the sample. The classification of laboratories with the designations A, B, C, or BMP was done according to ISTA rules (Table 1), summing up the absolute Z score (Z_{al}) for all q samples. The notes 5, 4, 3, or 0 were attributed with this value for each laboratory, also according to the ISTA rules for just one round.

To calculate the bias (V_l) of each laboratory the average Z-score was calculated from the sum divided by the number of samples (q). Since the measure of reliability $\left(P_l^{(I)}\right)$ for the l-th laboratory was determined by the bias and the Z-score, the square root of the sum of squared differences between Z_{al} and V_l divided by q was calculated. The measurement accuracy $\left(E_l^{(I)}\right)$ of the l-th laboratory was obtained from the square root of the sum of the square of the bias and the square of precision. Laboratory accuracy was considered acceptable for $0.000 < E_l^{(I)} < 1.499$, critical for $1.500 < E_l^{(I)} < 1.999$, and unacceptable for $E_l^{(I)} \geq 2.000$ (ISTA, 2002).

To use the technique of cluster analysis, only the means of each laboratory for each sample were considered. The methodology used was: choosing the method of dissimilarity, choosing the clustering algorithm, determining the number of groups, and group validation. In this study the Euclidean distance was adopted as a measure of dissimilarity between two laboratories k and j, which determines the physical distance between two objects, considering the Euclidean space. The average linkage method and medium distances were selected, where the distance between groups is calculated by averaging the distances between all pairs of laboratories in the two groups being compared.

To determine the number of groups g, the criterion proposed by Mojena (1977) was used, which improves the quality of the fit to the data grouping. The number of groups given by the first stage in the dendrogram was selected, in which $\alpha_m > \overline{\alpha} + \phi S_a$, with $m = 1,2,...,p$, being the maximum number of groups (in this context the number of laboratories was equal to 12); α_m is the distance value for the stage of the junction corresponding to $p - m + 1$ groups, $\overline{\alpha}$ and S_a are estimators of mean and standard deviation of $\alpha's$ distances, and ϕ is a constant. Millingan and Cooper (1985) suggest using $\phi = 1,25$, based on simulation results. To validate the cluster the actually observed distances between objects (Euclidean distances) were compared with the distances predicted by the clustering process using the Pearson correlation, but in this context it is called the cophenetic correlation. The closer to 1 it is, the better the fit of the cluster.

RESULTS

For the first simulation of the Z-score values of the twelve laboratories for the three samples, the third laboratory had two values in the range of performance seen as unsatisfactory ($|Z| \geq 3$), while laboratories 6, 7, 10, 11, and 12 each had one value in the unsatisfactory range. The best performances were achieved by laboratories 1, 4, 5, and 9 ($|Z| < 2$). Taking into account both the notes as the classification accuracy, it appears that it was impossible to distinguish between the two groups in simulation 1 (Table 2).

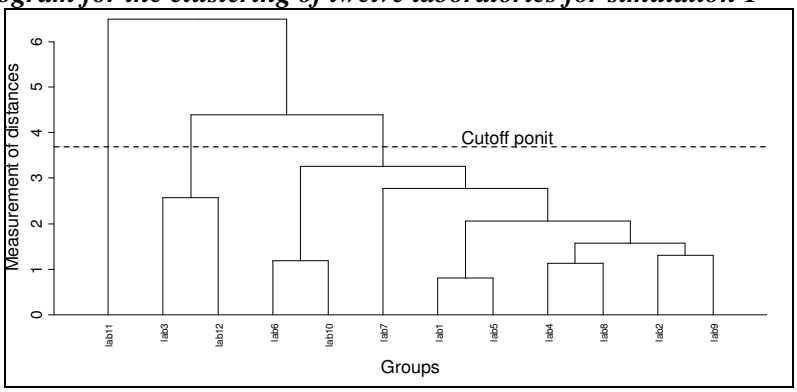
In this context, the risk that occurs by classifying a laboratory as, for example, BMP can be questioned; for example laboratory 3 should have a similar performance to other laboratories in its initial group. This oddity exists at least in theory, since the working was done with simulated data and it is expected, therefore, in relatively controlled conditions. So it seems that this technique may not be the best suited to the situation where the variance and covariance matrices are different.

Table 2 Results of the Z-score method for simulation 1

Group	Lab.	Classification	Grade	Bias	Precision	Accuracy	Accuracy classification
1	1	A	5	0.6	1.2	1.4	Acceptable
	2	A	5	1.1	1.4	1.7	Critical
	3	BMP	0	2.8	5.7	6.4	Not acceptable
	4	A	5	0.5	0.1	0.5	Acceptable
	5	A	5	0.6	1.2	1.3	Acceptable
	6	C	3	2.1	4.1	4.6	Not acceptable
	7	C	3	1.8	1.7	2.4	Not acceptable
	8	B	4	1.4	0.5	1.5	Acceptable
	9	A	5	0.9	0.7	1.1	Acceptable
	10	C	3	2.1	4.4	4.8	Not acceptable
2	11	C	3	1.8	1.8	2.6	Not acceptable
	12	BMP	0	2.6	5.5	6.1	Not acceptable

In the case of the cluster analysis results for simulation 1, initially using the criteria proposed by Mojena (1977), only a single group would be formed because the cutoff was equal to 4.63 $\left(\bar{\alpha} = 2.51; s_a = 1.70\right)$ and the value of the distance from the corresponding junction stage to a group was 6.49 according to the dendrogram (Figure 1).

Figure 1 Dendrogram for the clustering of twelve laboratories for simulation 1



However, this distance value was considered an outlier by the Hampel method. Therefore, in this study is proposed a modification to the original procedure of Mojena's criterion; that is, the possible presence of an outlier should be checked and then the mean, standard deviation, and therefore the cutoff point should be calculated. In this case, the new cutoff point was equal to 3.52 $\left(\bar{\alpha} = 2.10; s_a = 1.13\right)$ and the measure of distance corresponding to the junction of the formation of two groups was 4.39. Thus, laboratory 11 was separated from the other laboratories, but it was not possible to also separate laboratory 12 using the cluster analysis technique. The grouping could be considered of good quality because the value of the cophenetic correlation was equal to 86.52%.

In the case of simulation 2, although the Z-score method was unable to satisfactorily distinguish between the two groups, the scores for the three samples from laboratories 11 and 12 were highly negative. This pattern did not show the results of two laboratories for the conditions of simulation 1. Thus, these considerations seem to be an indicator that an intermediate classification between C and BMP could be proposed. In the case of clustering technique, the cutoff point established after the withdrawal of distance 54.01 was equal to 3.19 $\left(\bar{\alpha} = 1.94; s_a = 1.00\right)$ and the measure of the distance corresponding to the junction of the formation of two groups was 3.74. Thus, laboratories 11 and 12 fit into a different group from the others. The grouping was considered

to be good quality as the value of the cophenetic correlation was equal to 99.86%.

In the case of simulation 3, it was also not possible to differentiate the two groups by the Z-score method. For the clustering technique, the cutoff point established after the withdrawal of distance 52.56 (using the Hampel outlier method) was equal to 6.18 $\left(\bar{\alpha} = 2.99; s_a = 2.54\right)$ and the measure of the distance corresponding to the junction of the formation of two groups was 8.09. Thus, laboratories 11 and 12 fit into a different group from the others. The clustering can be considered to be of good quality, as the cophenetic correlation value was equal to 99.39%.

CONCLUSIONS

In a more general discussion about the use of the Z-score method for analysis of the results of a proficiency test, it should be noted that the exclusive use of this technique should be done with caution because it may jeopardize some laboratories' real quality conditions. Moreover, its use in conjunction with other procedures can also cause doubts about the validity of results, as is the case in this study. The low distinction between groups in the case of the Z-score method may be associated with sample size, consequently triggering an increase in the error rate for a pre-established degree of confidence. The cluster analysis technique distinguished the two groups well. It should be noted that by using this procedure it was possible to make assumptions about the performance of laboratories based only on dendrogram, and it was not necessary to obtain indexes of precision and accuracy to assess the laboratories.

In general, laboratories that presented distorted results were identified by at least one of the techniques. Thus, these results may serve to orientate the guidelines of the programme of quality control for a laboratory and the process of accreditation of seed laboratories along with the certifying agencies.

REFERENCES

Association of Analytical Communities (1993) International harmonized protocol for proficiency testing of (chemical) analytical laboratories. *Journal of AOAC International*, 76 (4).

Brasil (1992). Ministério da Agricultura e Reforma Agrária. *Regras para análise de sementes*. Brasília: SNDA/DNDV/CLAV. 365 p.

Ferreira, D.F. (2008) *Estatística multivariada*. Lavras: UFLA/DEX. 661p.

International Seed Testing Association (2002) ISTA Referee Test Programme: *ISTA Statistics Seminar*. Corvallis, Oregon.

International Seed Testing Association (2007) *The ISTA Proficiency Test Programme*. Switzerland.

Milligan, G.W.; Cooper, M.C. (1985) An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50, pp. 159–179.

Mojena, R. (1977) Hierarchical grouping methods and stopping rules: an evaluation. *Computer Journal, London*, 20, pp. 359–363.

Oliveira, P.R.P.; Cicero, S.M. (1996). Causas de variação dos resultados das análises de sementes de capim colonião (Panicum maximum Jacq.). *Revista Brasileira de Sementes*, 18 (1), pp. 122–128.

Rencher, A.C. (2002) *Methods of multivariate analysis*. 2nd ed. New York: John Wiley. 708 p.

Rennie, W.J.; Tomlin, M.N. (1984) Repeatability, reproducibility and interrelationship of results of tests on wheat seed samples infected with Septoria nodorum. *Seed Science & Technology*, 12, pp. 863–880.

Zorato, F. (2005) Evolução do laboratório de análise de sementes. *Revista Seed News*, 9 (6), pp. 1-6.