

Ilya Kuselman

Nonparametric assessment of comparability of analytical results obtained in proficiency testing based on a metrological approach

Received: 10 October 2005
Accepted: 5 January 2006
Published online: 18 February 2006
© Springer-Verlag 2006

I. Kuselman (✉)
The National Physical Laboratory of
Israel (INPL),
Danciger “A” Bldg., Givat Ram,
Jerusalem 91904, Israel
e-mail: ilya.kuselman@moital.gov.il

Abstract A nonparametric sign test is implemented for assessment of comparability of proficiency testing (PT) results when their distribution differs from the normal or other known distribution. It allows testing the null hypothesis about insignificance of the bias of median of results obtained in PT from the traceable certified value of the reference material used in PT as test items, i.e., the hypothesis stating that comparability of the PT results is successful. Probability of type I error of rejecting the null hypothesis when it is true, and probability of type II

error of it not rejecting when it is false (the alternative hypothesis about unsuccessful comparability is true) are considered. The test can be helpful for PT providers and laboratory accreditation bodies in analysis of PT results when the comparability criterion developed for a normal results distribution (Accred. Qual. Assur. 10:466–470) is not applicable.

Keywords Proficiency testing · Distribution of analytical results · Comparability · Test of hypotheses · Reliability

Introduction

Comparability of analytical results obtained in proficiency testing (PT) based on a metrological approach is discussed in Ref. [1] for cases when hypothesis about normal distribution of the results is accepted. Since the approach implies the use of a reference material or a spike with traceable property value as test items/samples sent to laboratories participating in PT, comparability was assessed by the bias of the mean of PT result from the assigned/certified value C_{cert} of the analyte concentration in the material, taking into account standard uncertainty σ_{cert} of this value and standard deviation σ_{PT} of the PT results.

The hypothesis on normal distribution of analytical results is widely used because of the central limit theorem. According to the theorem, a sum of independent random values/components, when their number is large enough, has normal distribution, even if the components are distributed arbitrarily. The condition is only that the components are of the same order, i.e., none of them dominates [2, 3]. The reason is that analytical results obtained by a validated method have uncertainty budget with a number of studied components, which are under control [4]. Moreover, as follows from the central limit theorem, PT or other analytical

results calculated as means of some replicates should have a distribution tending towards the normal one, even if the original populations of the replicates are not normal [5].

However, the conditions of applicability of this theorem are not always fulfilled. Such a situation is typical for analyses made close to the range limits of the analyte content in the material/substance under analysis [6, 7]. In particular, it occurs at the limit of determination where zero and even negative results (obtained by subtraction of a blank value, due to interference, because of calibration inappropriate extrapolation or by other reasons) are a part of the population. Excluding unrealistic negative values leads to asymmetric distributions in these cases [8]. Asymmetric distributions of results in calibration curves are also described for analyte concentrations that are far from the limit of determination [9]. It is shown that deviation of a distribution of analytical results from the normal one can be caused by the dominating uncertainty source arising at any stage of analysis, from aliquot measuring to result rounding [10, 11]. Analyte heterogeneity in the material under analysis, use of different lots of reagents (with different impurities) by different laboratories, as well as drift of measuring instruments lead to asymmetric and/or polymodal distributions of analytical results [12, 13].

Table 1 Probability P of the events according to the alternative hypothesis H_{11}

γ	Δ/σ_{PT}	P
0.4	0.50	0.69
0.7	0.75	0.77
1.0	1.04	0.85

For example, during validation of analytical methods used in secondary nonferrous metallurgy (recycling scrap and waste products) in the former USSR, performed under my supervision in 1978–1990, more than 1,500 statistical samples were tested. About 10% of them had distributions that were significantly different from normal distributions. Eight percent of the analytical data was obtained by spectral methods and 21% obtained by chemical (wet) methods [14].

In the book [15], about 1,900 statistical samples of results of interlaboratory trials for certification of reference materials of minerals are classified by chemical elements/analytes as the class of symmetrical distributions, the class with the left distribution asymmetry (at the limits of determination) and the class with the right asymmetry (at the upper limit of the concentration range).

It is a good case when sources of a deviation from normal distribution are identified and eliminated, or there are theoretical reasons to use another distribution, such as log-normal [16], uniform [12], Poisson's [17], etc. However, in many cases, statistical tests for goodness-of-fit of empirical and theoretical distributions are not selective and do not allow to prefer one of the known distributions [18]. In these cases, nonparametric methods for statistical analysis of analytical/chemical results can be useful [2].

The purpose of the present paper is to develop in continuation of Ref. [1] a nonparametric test allowing a PT provider and/or an accreditation body to assess comparability of analytical results obtained in PT, when distribution of the results differs from the normal or other known distribution, and it is impossible to determine and to remove the causes of this difference.

The sign test of the hypotheses about comparability of PT results

For unknown distributions differing from the normal one, the median is more robust than the average, i.e., is better reproduced on experiment repetition, being less sensitive to extreme results/outliers. Besides, a confidence interval to the median can be formulated not depending on the results distribution [2, 5]. Therefore, when PT results are not distributed normally, their median can be used just as the mean was used in Ref. [1] for comparability assessment of the results. In this case, the null hypothesis H_0 , assuming that the PT results bias exceeds σ_{cert} by a value which is insignificant in comparison with random interlaboratory errors of the analysis, has the following form:

$$H_0: |M_{PT} - C_{cert}| = [(0.3\sigma_{PT})^2 + \sigma_{cert}^2]^{1/2} = \Delta, \quad (1)$$

Table 2 A values (the bias norms)

$\alpha/2$	N						
	5	10	15	20	30	40	50
0.025	–	1	3	5	9	13	17
0.05	0	1	3	5	10	14	18

where 0.3 is used according to known metrological rule defining one standard deviation insignificant in comparison with other when the first one is not increasing one-third of the second one (it means that the first variance is smaller than the second one for an order); and M_{PT} is the median of PT results of hypothetically infinite number N of participants, i.e., the population median.

If $M_{PT} \geq C_{cert}$, the null hypothesis H_0 means that probability P of an event when a result C_i of i -th laboratory participating in the PT exceeds the value $C_{cert} + \Delta$, is $P\{C_i > C_{cert} + \Delta\} \leq 1/2$ according to the median definition. If $M_{PT} < C_{cert}$, the probability that C_i yields to the value $C_{cert} - \Delta$ is also $P\{C_i < C_{cert} - \Delta\} \leq 1/2$. The alternative hypothesis H_1 assumes that the bias exceeds σ_{cert} significantly and probabilities of the events described above are $P > 1/2$:

$$H_1: |M_{PT} - C_{cert}| > \Delta, \quad (2)$$

where Δ is the same as in Eq. (1). For example, by analogy with [1]

$$H_{11}: |M_{PT} - C_{cert}| = 2.0\Delta, \quad (3)$$

Probabilities P of the events according to the alternative hypothesis H_{11} at normal distribution depending on the permissible bias Δ (in σ_{PT} units) at different ratio $\sigma_{cert}/\sigma_{PT} = \gamma$ values are shown in Table 1.

Since in practice the population median is unknown, and N laboratories participating in PT form a statistical sample (of size N) from the population results, the hypothesis H_0 is not rejected when the upper limit of the median confidence interval does not exceed $C_{cert} + \Delta$, or the lower limit does not yield to $C_{cert} - \Delta$. The limits can be evaluated based on the simplest nonparametric sign test [5]. According to this test, the number N_+ of results $C_i > C_{cert} + \Delta$ or the number N_- of results $C_i < C_{cert} - \Delta$ should not exceed the critical value A (the bias norm) in order to not reject H_0 . The A values are available in known statistical handbooks, for example, in Ref. [19]. These values for $N = 5$ –50 PT participants and two levels of confidence 0.975 ($\alpha/2 = 1 - 0.975 = 0.025$) and 0.95 ($\alpha/2 = 0.05$), are shown in Table 2. The A value for less than six participants at $\alpha/2 = 0.025$ cannot be determined, and therefore, is absent from Table 2 for $N = 5$.

Reliability of the test

Reliability of the test is determined by the probabilities of not rejecting the null hypothesis H_0 when it is true, and

Table 3 PT results of aluminum determination in SRM 2690 (simulated in percent by weight) ranked according to their increasing values

No. <i>j</i>	Result C_i (%)	$C_j - C_{\text{cert}}$ (%)	Sign	No. <i>j</i>	Result C_i (%)	$C_j - C_{\text{cert}}$ (%)	Sign	No. <i>j</i>	Result C_i (%)	$C_j - C_{\text{cert}}$ (%)	Sign
1	11.86	-0.49	-	18	12.44	0.09	0	35	12.53	0.18	0
2	11.88	-0.47	-	19	12.44	0.09	0	36	12.55	0.20	+
3	11.90	-0.45	-	20	12.45	0.10	0	37	12.56	0.21	+
4	11.91	-0.44	-	21	12.46	0.11	0	38	12.57	0.22	+
5	11.93	-0.42	-	22	12.46	0.11	0	39	12.60	0.25	+
6	11.96	-0.39	-	23	12.47	0.12	0	40	12.61	0.26	+
7	11.96	-0.39	-	24	12.48	0.13	0	41	12.64	0.29	+
8	11.97	-0.38	-	25	12.49	0.14	0	42	12.66	0.31	+
9	11.98	-0.37	-	26	12.49	0.14	0	43	12.67	0.32	+
10	11.99	-0.36	-	27	12.50	0.15	0	44	12.68	0.33	+
11	12.03	-0.32	-	28	12.50	0.15	0	45	12.69	0.34	+
12	12.07	-0.28	-	29	12.51	0.16	0	46	12.76	0.41	+
13	12.17	-0.18	0	30	12.51	0.16	0	47	12.81	0.46	+
14	12.19	-0.16	0	31	12.52	0.17	0	48	12.84	0.49	+
15	12.20	-0.15	0	32	12.52	0.17	0	49	12.90	0.55	+
16	12.34	-0.01	0	33	12.53	0.18	0	50	12.96	0.61	+
17	12.43	0.08	0	34	12.53	0.18	0	$N_- = 12,$ $N_+ = 15$			

rejecting it when it is false (the alternative hypothesis H_1 is true). The test does not allow rejecting hypothesis H_0 with probability $1 - \alpha/2$, when it is true. Probability of an error of type I by this test (to reject the H_0 hypothesis when it is true) is $\alpha/2$. Probability P_t of rejecting H_0 , when it is false, i.e., when the alternative hypothesis H_1 is actually true (the test power), is tabulated in Ref. [19]. The probability of an error of type II (not rejecting the H_0 when it is false) equals $\beta = 1 - P_t$. The operational characteristics of the test P_t and β are shown in Fig. 1 at $\alpha = 0.05$ for the alternative hypothesis H_{11} by formula (3) at different γ values and different numbers N of the PT participants.

Example

The hypothesis about normal distribution of PT results in the example discussed in Ref. [1] (regarding standard reference material SRM 2690 applicable for PT of aluminum determination in coal fly ashes) was not tested because of the small size of the statistical samples. Therefore, the sample size is increased here to $N = 50$: the simulated data are presented in Table 3 (the simulation is performed by the

known method of successive approximations). Such sample size allows testing the hypothesis about the data normal distribution applying the Cramer–von Mises ω^2 -criterion [18]:

$$\omega_N^2 = -N - 2 \sum_{j=1}^N \{[(2j-1)/2N] \ln F(x_j) + [1 - (2j-1)/2N] \ln [1 - F(x_j)]\}, \quad (4)$$

where $j = 1, 2, \dots, N$ is the number of the PT result C_j in the statistical sample ranked by an increasing C value ($C_1 \leq C_2 \leq \dots \leq C_N$); $x_j = (C_j - C_{\text{PT/av}})/S_{\text{PT}}$ is the normalized value of the j -th result which is distributed with the mean of 0 and the standard deviation of 1; $C_{\text{PT/av}}$ and S_{PT} are the natural average and the standard deviation of the PT results (percent by weight); $F(x_j)$ is the function of the normalized normal distribution.

The advantage of this criterion in comparison to the classical Kolmogorov or Kolmogorov–Smirnov criteria for goodness-of-fit distributions is the possibility to test a statistical sample of $N \geq 50$ without dividing the data into ranges according to their values. Therefore, ω^2 -criterion

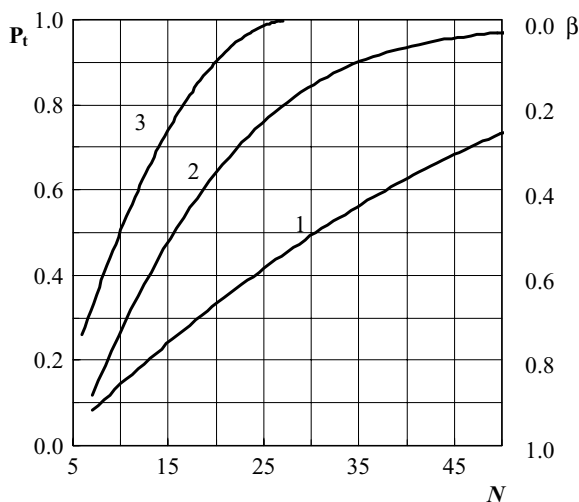


Fig. 1 Power P_t of the test and probability β of an error of type II in dependence on the number N of laboratories participating in PT, when probability of an error of type I is $\alpha/2 = 0.025$, and the null hypothesis H_0 is tested against the alternative hypotheses H_{11} at $\gamma = 0.4, 0.7$ and 1.0 —curves 1, 2 and 3, correspondingly

uses the maximum information contained in the statistical sample [18].

The probability that $\omega_N^2 = 1.95$ calculated by formula (4) is randomly exceeding the critical value 1.94 (for $N = 50$) equals 0.10 [19]. Therefore, the hypothesis about normal distribution of the data in Table 3 should be rejected at the level of confidence of 0.90. The corresponding empirical histogram and the theoretical (normal) distribution are shown in Fig. 2. It is clear that the empirical distribution is a bimodal one, therefore no normal distribution can fit it. Since other known distributions are also not suitable here, let us apply the proposed nonparametric test for the comparability assessment of the results.

Taking into account $C_{\text{cert}} = 12.35\%$, $\sigma_{\text{cert}} = 0.14\%$, $\sigma_{\text{PT}} = 0.38\%$ by weight, and $\gamma = 0.14/0.38 \approx 0.4$, one

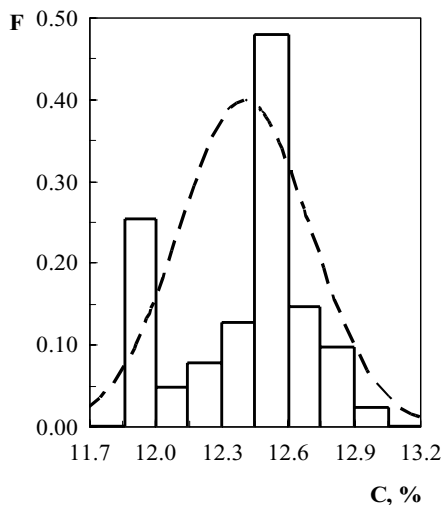


Fig. 2 Histogram of PT results (frequency F of a result value C) (solid line), and the corresponding normal approximation (dotted line)

can calculate $\Delta = 0.50 \cdot 0.38 = 0.19\%$ (see Table 1 for $\gamma = 0.4$), $C_{\text{cert}} + \Delta = 12.54\%$ and $C_{\text{cert}} - \Delta = 12.16\%$. The values $C_j > 12.54\%$ are marked in Table 3 by symbol “+”, and the values $C_j < 12.16\%$ by symbol “-”. There are $N_+ = 15$ and $N_- = 12$ of such values, correspondingly. Other $N - N_+ - N_- = 23$ values in the range $C_{\text{cert}} \pm \Delta$ are marked by symbol “0”. The sample median is $C_{25} = C_{26} = 12.49 > C_{\text{cert}} = 12.35\%$ and $N_+ > N_-$, however, N_+ is less than the critical value $A = 17$ at $\alpha/2 = 0.025$ and $N = 50$ (see Table 2). Therefore, the null hypothesis H_0 about successful comparability of the results is not rejected.

Reliability of the assessment using the hypotheses H_0 against H_{11} for this PT scheme ($\gamma = 0.4$) can be characterized by (1) probability $1 - \alpha/2 = 0.975$ of correctly assessing the comparability as successful (not rejecting the null hypothesis when it is true) for any number $N \geq 6$ of the PT participants, and (2) probability $P_t = 0.73$ of correctly assessing the comparability of $N = 50$ PT results as unsuccessful (rejecting H_0 when the alternative hypothesis H_{11} is true: see Fig. 1). Probability $\alpha/2$ of an error of type I is 0.025 for any $N \geq 6$, while probability β of type II error is 0.27 for $N = 50$.

Discussion

Since the sign test critical A values are determined for $N \geq 4/8$ at different probabilities α , and the test power is calculated also only for $N \geq 6/8$, the proposed comparability assessment cannot be performed for a smaller sample size. As already mentioned in paper [1], the power efficiency of the sign test in relation to the t -test (ratio of the sizes N of statistical samples from normal populations allowing the same power) is from 0.96 for $N = 5$ to 0.64 for infinite N . For example, practically the same power (0.73 and 0.75) was achieved in case of the sign test of comparability of results for aluminum determination in coal fly ashes at $N = 50$ discussed above, and using t -test for the same purpose at $N = 30$ in Ref. [1]. The power efficiency here is approximately $30/50 = 0.6$. On the other hand, when information about the distribution of PT results is limited by small sizes of statistical samples of experimental data ($N < 50$), it is a problem to evaluate the goodness-of-fit empirical and theoretical/normal distributions, a decrease of the t -test power and the corresponding decrease of reliability of the comparability assessment caused by deviation of the empirical distribution from the normal one.

Of course, reliability of the assessment and probabilities of erroneous decisions based on the sign test, as any statistical criterion or test, depends on how the hypotheses have been formulated. In particular, if $\gamma < 0.3$ and the null hypothesis is: $|M_{\text{PT}} - C_{\text{cert}}| \leq 0.3\sigma_{\text{PT}}$, the same power requires a bigger sample size, i.e., $N = 50/100$ [20]. Another null hypothesis: $|M_{\text{PT}} - C_{\text{cert}}| \leq 0.3(\sigma_{\text{PT}}^2 + \sigma_{\text{cert}}^2)^{1/2}$ also leads to a test that is stricter than the one described above. Such hypothesis is suitable for comparability assessment of analytical results obtained in a crucial PT, or may be in a key comparison when distribution of the results differs from normal or other known distribution.

The role of the adequacy of the reference material used, of the ratio γ and of an alternative hypothesis (its “distance” from the null hypothesis) is similar to the one shown in Ref. [1] for the t -test.

Conclusions

1. A nonparametric sign test is formulated for assessment of comparability of analytical results obtained in PT based on a metrological approach, when distribution of the results differs from a normal one. The test consists of estimation of a null hypothesis about insignificance of the bias of median of the PT results from the traceable certified value of the reference material used in the PT,
2. In spite of smaller sign test power, in comparison with power of the corresponding t -test, the former can be more effective for cases when number of laboratories participating in a PT is less than 50, and indication of a deviation of the results’ empirical distribution from a theoretical/normal distribution is problematic.
3. The sign test can be helpful for PT providers and laboratory accreditation bodies in analysis of PT results without any special statistical software tools available.

References

1. Kuselman I (2006) Comparability of analytical results obtained for proficiency testing based on a metrological approach. *Accred Qual Assur* 10:466–470
2. Mittag HJ, Rinne H (1993) *Statistical methods of quality assurance*. Chapman and Hall, London
3. ISO (1993) *Guide to expression of uncertainty in measurement*. ISO, Geneva
4. EURACHEM/CITAC Guide (2000) *Quantifying uncertainty in analytical measurement*, 2nd edn. EURACHEM/CITAC, www.eurachem.ul.pt
5. Miller JC, Miller JN (1993) *Statistics for analytical chemistry*, 3rd edn. Ellis Horwood Ltd., Bodmin, England
6. Parshin AK, Fadeev AG (1983) *J Anal Chem* 38/7:1174–1180 (in Russian)
7. Thompson M, Howarth RJ (1980) *Analyst* 105:1188–1195
8. Hill ARC, Von Holst C (2001) *Analyst* 126:2044–2052
9. Renman L, Jagner D (1997) *Anal Chim Acta* 357:157–166
10. Harris WE (1978) *Talanta* 25/26:325–329
11. Harris WE (1978) *Am Lab* 10/11:31–39
12. Shaevich AB (1981) *Analytical service as a system*. Khimia, Moscow (in Russian)
13. Sansoni B, Iyer RK, Kurth R (1981) *Fresenius’ Z Anal Chem* 306:212–232
14. Skryabina LG (1986) *Metrological validation of methods for analysis of samples of non-ferrous alloys from secondary raw materials*, Ph.D. thesis. Kharkov State University, Ukraine (in Russian)
15. Lontsih SV, Petrov LL (1988) *Reference materials of natural objects*. Nauka, Novosibirsk (in Russian)
16. Shaevich AB (1971) *Measurement and standardization of chemical composition of substances*. Physmathgis, Moscow (in Russian)
17. Doerffel K (1990) *Statistik in der Analytischen Chemie*, 5th edn. Deutscher Verlag fur Grundstoffindustrie GmbH, Leipzig (in German)
18. D’Agostino RB, Stephens MA (eds) (1986) *Goodness-of-fit techniques*. Marcel Dekker, New York
19. Owen DB (1962) *Handbook of statistical tables*. Addison-Wesley, London
20. Kuselman II, Skryabina LG (1983) *Metrologia* 3:59–64 (in Russian)