Australian Government
Department of Industry, Innovation and Science

National Measurement Institute

Statistical Manual

NMI North Ryde – CRV

Issue No.:	3.10	**Issued Date:**	26 February 2019
Approved By:	CRV Manager	**Reviewed Date:**	26 February 2021
Prepared By:	Raluca Iavetz		
Amendments:	Refer to revision history		
Control:	The electronic copy on the WAN is the latest version of this document. Any paper copy is UNCONTROLLED and should be checked against the electronic copy before use.		

Date Printed: 5/03/2019

Contents

1 Introduction

The Chemical Proficiency Testing (CPT) Statistical Manual outlines the statistical methods used by CPT. These methods are based on the procedures described in ISO 13528:2005 (E) "Statistical methods for use in proficiency testing by interlaboratory comparisons"[1] and "The International Harmonized Protocol for the Proficiency Testing of Analytical Chemistry Laboratories"[2].

The role of the CPT Statistical Manual is to set out the procedures used in assessing the homogeneity of the test materials sent to the participants', the method of establishing the assigned value and the target standard deviation of a PT study as well as the tools used to assess and compare individual laboratory performance.

2 Sufficient Homogeneity Testing

2.1 Sample Selection and Measurement

Homogeneity testing of the prepared and packaged proficiency test samples should be conducted as soon as possible after packaging.

Select a minimum of 7 (but preferably 10) of the packaged units strictly at random from the entire batch, or by stratified random sampling throughout the fill sequence if fill trend effects are suspected. This must be done in a formal way, by assigning a sequential number to the units (either by label or by their position in a linear sequence). The selection is made by use of a random number table or computer random number generation software. It is not acceptable to select the units in any other way (eg by "shuffling" or "selection at random").

Homogenise each selected test unit within its container, then take two appropriately sized test portions from each. Label the test portions as "1a", "1b", "2a", "2b" etc. Test portions must be sufficiently large, particularly for solid samples, so as not to compromise the precision of the test results.

Sort the entire set of test portions into a random order, again using a random number table or computer random number generation software.

Analyse each test portion for each analyte of interest, maintaining this random order throughout. The testing should be performed under repeatability conditions (in as short a time as is practical, by a single analyst, preferably in a single sample batch). The analytical method selected must be sufficiently precise to allow a satisfactory estimation of between-sample variance and therefore should have a repeatability standard deviation (san) of less than half of the target standard deviation (σ) set for the study.

Include appropriate quality control samples (blanks, recoveries, control samples) with each batch of test samples.

2.2 Statistical Analysis of Homogeneity Data

The statistical procedure below follows the "The International Harmonized Protocol for the Proficiency Testing of Analytical Chemistry Laboratories"2.

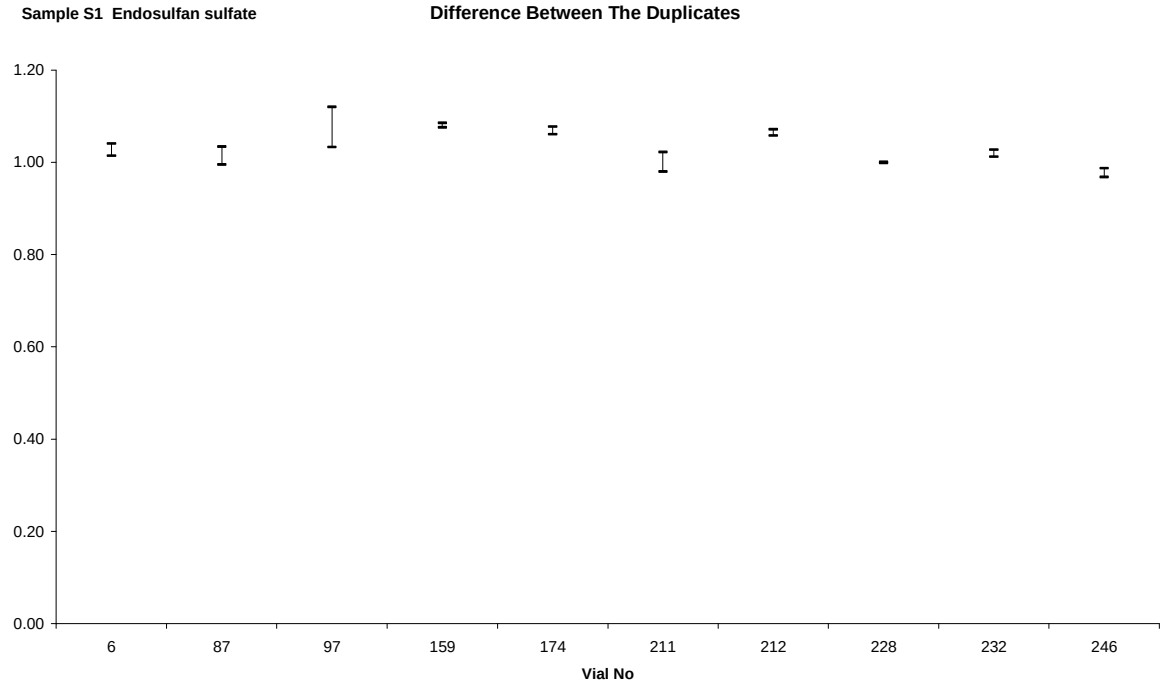
The data in the Table 1 are taken from AQA 06-02, Sample S1 Endosulfan Sulfate

Table 1 Duplicated results for ten distribution units and intermediate stages of calculation in Cochran's test

Sample	A (mg/kg)	B (mg/kg)	D = A-B	S = A+B	$D^2=(A-B)^2$
6	1.041	1.014	0.027	2.055	0.00070
87	1.034	0.995	0.039	2.029	0.00151
97	1.120	1.033	0.087	2.153	0.00756
159	1.076	1.086	-0.010	2.161	0.00010
174	1.078	1.061	0.017	2.139	0.00028
211	1.023	0.980	0.042	2.003	0.00178
212	1.058	1.072	-0.013	2.130	0.00018
228	1.001	0.998	0.002	1.999	0.00001
232	1.012	1.028	-0.015	2.040	0.00023
246	0.987	0.969	0.019	1.956	0.00035

2.2.1 Visual Appraisal for Data Pathologies

The data presented is inspected visually for suspect features such as discordant duplicated results, outlying samples, trends or discontinuities.



No obvious trends, outliers or discontinuities.

2.2.2 Cochran's Test

Analytical outliers should be deleted from the data before one-way analysis of variance (ANOVA) is carried out; Cochran's test is suitable.

Calculate the test statistic (C):

$$C = \frac{D_{\max}^2}{\sum D_i^2}$$
$$= \frac{0.00756}{0.0127}$$
$$= 0.595$$

where C = Cochran's statistic test

Dmax = the largest difference between duplicates

Di = difference of each pair of duplicates

Table 2 Critical values for the Cochran test statistic for duplicates

m[1]	95%
7	0.727
8	0.680
9	0.638
10	0.602
11	0.570
12	0.541
13	0.515
14	0.492
15	0.471
16	0.452
17	0.434
18	0.418
19	0.403
20	0.389

[1]m is the number of samples that have been measured in duplicate.

The 5% critical value for ten samples from Table 2 is 0.602.

No analytical outlier was identified.

2.2.3 Estimate of Analytical and Sampling Variances

One-way ANOVA is used to estimate the analytical and sampling variance and is performed in Excel.

The output from one-way Anova is presented in the table below:

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	0.0244	9	0.00271	4.27	0.0166	3.020
Within Groups	0.00635	10	0.000635			

$$\text{So} \quad s_{an}^2 = MS_{within}$$
$$= 0.0006351$$

where s_{an}^2 = the analytical variance

and

$$s_{sam}^2 = \frac{MS_{between} - MS_{within}}{2}$$
$$= \frac{0.00271 - 0.000635}{2}$$
$$= 0.00104$$

where s_{sam}^2 = the between-sample variance

2.2.4 Test for Sufficient Analytical Precision ($s_{an} < 0.5\sigma$)

The target standard deviation (σ) is the product of the mean of all duplicate results (χ) and the performance coefficient of variation (PCV) which is established by the study coordinator.

$$\sigma = \chi * PCV$$
$$= 1.03 * 0.15$$
$$= 0.155 \text{ mg/kg}$$

The analytical standard deviation (san) is the square root of the analytical variance estimated from ANOVA above.

$$s_{an} / \sigma = \frac{0.0252}{0.155}$$
$$= 0.163$$

This is less than the critical value of 0.5. The method is precise enough to detect significant in-homogeneity.

2.2.5 Test for Acceptable Between Sample Variance

Calculate the allowable sampling variance (σ_{all}^2) as

$$\sigma_{all}^2 = (0.3 * \sigma)^2$$
$$= (0.3 * 0.155)^2$$
$$= 0.00216$$

where σ = target standard deviation

The critical value is:

$$c = F_1\sigma_{all}^2 + F_2 s_{an}^2$$
$$c = 1.88 * 0.00216 + 1.01 * 0.000635$$
$$= 0.00471$$

The values for factors F1 and F2 are presented in Table 3.

Table 3 Factors F1 and F2 for use in testing for sufficient homogeneity

m[1]	20	19	18	17	16	15	14	13	12	11	10	9	8	7
F_1	1.59	1.60	1.62	1.64	1.67	1.69	1.72	1.75	1.79	1.83	1.88	1.94	2.01	2.10
F_2	0.57	0.59	0.62	0.64	0.68	0.71	0.75	0.80	0.86	0.93	1.01	1.11	1.25	1.43

[1]m is the number of samples that have been measured in duplicate.

Compare the sampling variance s_{sam}^2 with the critical value.

The sampling variance (s_{sam}^2 = 0.00104) is less than the critical value (0.00471). The samples are sufficiently homogeneous.

The results of the sufficient homogeneity testing is summarised in Table 4.

Table 4: Homogeneity test results

	Value	Critical	Result
Cochran	0.595	0.602	Pass
s_{an}/σ	0.16	0.5	Pass
s^2_{sam}	0.00104	0.00471	Pass

Note: even though statistically significant differences between the test samples have been detected using one-way Anova (P value < 0.02), the inhomogeneity is small enough to be of no practical consequence when compared to the expected between laboratory variability.

2.3 Uncertainty due to Inhomogeneity

The uncertainty associated with inhomogeneity (uhom) is incorporated into the uncertainty of the assigned value.

- If F > 1, then u_{hom} = the sampling standard deviation (ssam) estimated from ANOVA
- If F < 1, then u_{hom} = the standard deviation of all results (stotal) divided by root 6.

The logic is:

If F > 1, sampling variance has been observed, so this can be used to estimate the uncertainty due to inhomogeneity.

If F < 1, then the sampling variance is smaller than the analytical variance. This means that any inhomogeneity is so small that the homogeneity testing does not have the power to detect it. The observed variation is almost all due to analytical variance. However this is not proof that the samples are perfectly homogeneous. Inhomogeneity is somewhere between zero, and the analytical variance (estimated as the standard deviation of all results, stotal), and it is likely to be closer to 0 than to stotal. This approximates a triangular distribution, hence the choice of root 6 as the divisor.

2.4 Alternative Homogeneity Testing Procedure used in NMI CPT

Sometime the above approach for homogeneity testing is not practical. For the analysis of total petroleum hydrocarbons and PFOS/PFOA in water it is necessary to use the whole sample for each analysis and so it is not possible to analyse in duplicate. An alternative is to perform single analyses on a minimum of 5 packaged units (but preferably 7 to 10). The standard deviation of replicate analysis results is an indicator of sample homogeneity. When is not possible to conduct replicate measurements, the standard deviation of the results can be used as ssam 1

The proficiency testing samples may be considered to be adequately homogeneous if:

$$Ssam \leq 0.3\ \sigma$$

3 Establishing the Assigned Value (X)

The assigned value is the "best practicable estimate of the true value of the concentration (or amount) of analyte in the test material."[3] Methods for establishing assigned value are presented below.

3.1 Consensus of Participants' Results (Robust Average)

The consensus of participants results is used as the assigned value when this value is the only practical method available for the proficiency test. The consensus of participants results is not traceable to any external reference, so although expressed in SI units, metrological traceability is not established.

CPT will calculate an assigned value by this method only if there is a minimum of six results to ensure a reasonable estimate.

The assigned value for the test material used in a proficiency study is the robust average of the results reported by all the participants in the round. This is a modern approach to the outlier problems in a proficiency study in which the influence of the outliers and heavy tails is down-weighted and is calculated using the procedure described in "ISO13528:2015(E), Statistical methods for use in proficiency testing by interlaboratory comparisons – Annex C"1.

When the assigned value is derived from robust average the uncertainty is estimated as:

$$u_{rob\ mean} = 1.25 * S_{rob\ mean} / \sqrt{p}$$

where:

$u_{rob\ mean}$ = robust mean standard uncertainty

$s_{rob\ mean}$ = robust mean standard deviation

p = number of results

The expanded uncertainty $(U_{rob\ mean})$` is the standard uncertainty multiplied by a coverage factor k = 2 at approximately 95% confidence level.

A worked example is set out below in Table 5 and 6.

Table 5 Participant results AQA 08-13 methamphetamine

Lab Code	Concentration Sample S3
2	71.2
3	57.0
4	55.4
5	58.1
6	55.4
7	58.4
8	60.67
9	55.65
10	57.2
11	55.4
12	59.6
13	45.9
14	57.3
15	56.0

Lab Code	Concentration Sample S3
16	55.3
17	61
18	56.5
19	57.7
20	100
21	58.4
22	54.3

Table 6 Robust average and associated uncertainty

No. results (p)	21
Robust mean	57.4
Srob mean	2.6
urob mean	0.7
k	2
Urob mean	1.4

So the assigned value is 57.4 ± 1.4% methamphetamine base (m/m).

Participants results that are extreme outliers (outside the range of ±50% of the robust average) will be excluded from the assigned value calculation.

3.2 Measurement by a Reference Laboratory

An assigned value and uncertainty may be obtained by a suitably qualified measurement laboratory using a method with sufficiently small uncertainty. This is probably the closest approach to obtaining the true value for the test material but it may be very expensive. This approach is used when practical and when resources are available for certain analytes and matrices.

NMI uses primary methods such as Isotope Dilution Mass Spectrometry for which the result is traceable directly to SI and is of the smallest achievable uncertainty. When reference value is used as the assigned value, performance scores are **calculated for any number of participants.**

3.3 Use of a Certified Reference Material

When the material used in a proficiency testing scheme is a certified reference material (CRM) its certified reference value is used as the assigned value. The uncertainty of the assigned value is derived from the information on uncertainty provided on the certificate.

3.4 Formulation

Formulation is the addition of a known amount or concentration of analyte to a base material which is either free of the analyte or its concentration accurately known. The assigned value is then determined from the proportions of the materials used and the known concentrations added.

This method is advantageous if pure substances are available to spike the test samples, as the added amount can be measured extremely accurately by gravimetric or volumetric methods. Consequently, there is usually no difficulty in establishing the traceability of the assigned value.

The uncertainty is estimated from the uncertainties in analyte concentrations of the materials used and gravimetric and volumetric uncertainties, through moisture content or any other changes during

mixing if significant. For more details to estimate standard uncertainty follow the approach described in the "Guide to the expression of uncertainty in measurement"[5].

4 Setting the Target Standard Deviation (σ)

The target standard deviation (σ) is the product of the assigned value (X) and the performance coefficient of variation (PCV).

The performance coefficient of variation is a measure of the between laboratory variation that in the judgement of the study coordinator would be expected from participants given the analyte concentration. It is important to note that this is not the coefficient of variation of participants results.

4.1 By Perception

The target standard deviation could be fixed arbitrarily by the study coordinator based on a perception of how laboratory should perform. The perception is based on practical experience and published models4, 5, 6 and varies depending on the concentration in the matrix. The values of target standard deviation for various projects are presented in the CPT Study Protocol.

4.2 From a Predictive Model

Thompson6 suggested a contemporary model to calculate the reproducibility standard deviation (σ) based on the Horwitz function4. This model predicts a standard deviation from a given concentration (c) and requires c to be dimensionless mass ratio, eg.1ppm≡10-6 or %≡10-2.

$$\sigma = \begin{cases} 0.22 * c & if\ c < 1.2 * 10^{-7} \\ 0.02 * c^{0.8495} & if\ 1.20 * 10^{-7} \le c \le 0.138 \\ 0.01 * c^{0.5} & if\ c > 0.138 \end{cases}$$

where c = concentration, (eg. the assigned value X expressed as a dimensionless mass ratio 1ppm≡10-6 or %≡10-2)

5 Calculation of z-scores and E_n-scores

5.1 Introduction

Scoring is the method of converting a participant's raw result into a standard form that adds judgemental information about performance.

Laboratory performance is assessed by comparing reported test results to the assigned value using both z-scores and En-scores.

5.2 Invalid results

Results are identifiably invalid and/or gross error if they are

- expressed in the wrong units,
- transposed
- non-numerical (eg NR not reported, NT not tested, 'less than')

and excluded from statistical analysis.[1,2]

5.3 Calculation of z-scores

z-scores are an indication of how much the reported result differs from the assigned value. The assigned value (X) and the target standard deviation (σ) have a critical influence on the calculation of

z-scores and must be selected with care if they are to provide a realistic assessment of laboratory performance.

$$z = \frac{(\chi - X)}{\sigma}$$

where:

z = z-score
χ = individual laboratory result
X = assigned value
σ = target standard deviation.

z-scores are interpreted as follows:

- $|z| \leq 2$ satisfactory.
- $2 < |z| < 3$ questionable
- $|z|$ of ≥ 3 unsatisfactory

Z-scores will be rounded to two decimal places.

5.4 Calculation of E_n-scores

E_n-scores (more properly called E_n numbers) are an alternative to z-scores. They provide a measure of how closely a reported laboratory result agrees with the assigned value, taking account of uncertainties in both the result and assigned value. Where a laboratory does not report an uncertainty estimate, an uncertainty of zero (0) is used to calculate the E_n-score.

The E_n-score is an objective measure of whether or not an individual result is consistent with the assigned value. Unlike z-scores, E_n-scores do not require the setting of a target standard deviation.

$$E_n = \frac{(\chi - X)}{\sqrt{U_\chi^2 + U_X^2}}$$

where:

E_n = E_n-score
χ = individual laboratory result
U_χ = expanded uncertainty of the individual laboratory result
X = assigned value
U_X = expanded uncertainty of the assigned value

E_n-scores are interpreted as follows:

- $|E_n| \leq 1$ satisfactory
- $|E_n| > 1$ questionable

E_n-scores will be rounded to two decimal places.

5.5 z, E_n-score adjustments

To account for possible bias in the consensus values due to laboratories using inefficient analytical/extraction techniques, some z-scores greater than 2 are adjusted to 2.
The maximum acceptable concentration for which z-scores are adjusted is set to two target standard deviations more than the spiked level. For results higher than the maximum acceptable concentration z-scores are not adjusted. This adjustment ensures that laboratories reporting results close to the spiked concentration are not penalised. The corresponding E_n-scores are adjusted to 1 if needed.

6 Summary Statistics and Graphs

6.1 Summary Statistics

Summary statistics: mean, median, maximum, minimum, robust standard deviation and robust coefficient of variation are calculated from the participants' results and tabulated with the participant results.

A guide to the number of significant figures for the summary statistics is given by Hibbert and Gooding[7]. The recommendation is two significant figures for uncertainty and then the result to the same order of magnitude (eg. uncertainty 0.011 M then the concentration would be expressed as 0.115 ± 0.011 M – 95% confidence interval).

6.2 Bar Plots

Bar charts of results and performance scores are included in the final report. An example chart with interpretation guide is shown in Figure 1. Included with the participant results chart is a histogram.

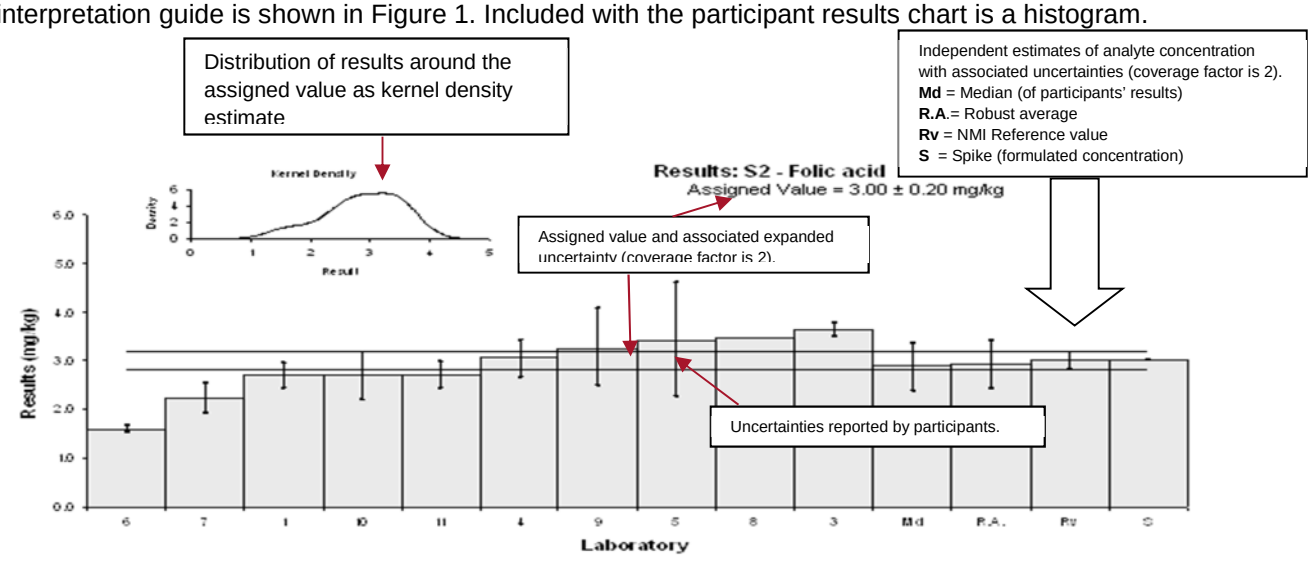


Figure 1 Guide to Presentation of Results

Z-scores and E_n-scores are plotted against the Lab Code number. Example z-score chart is presented in Figure 2.

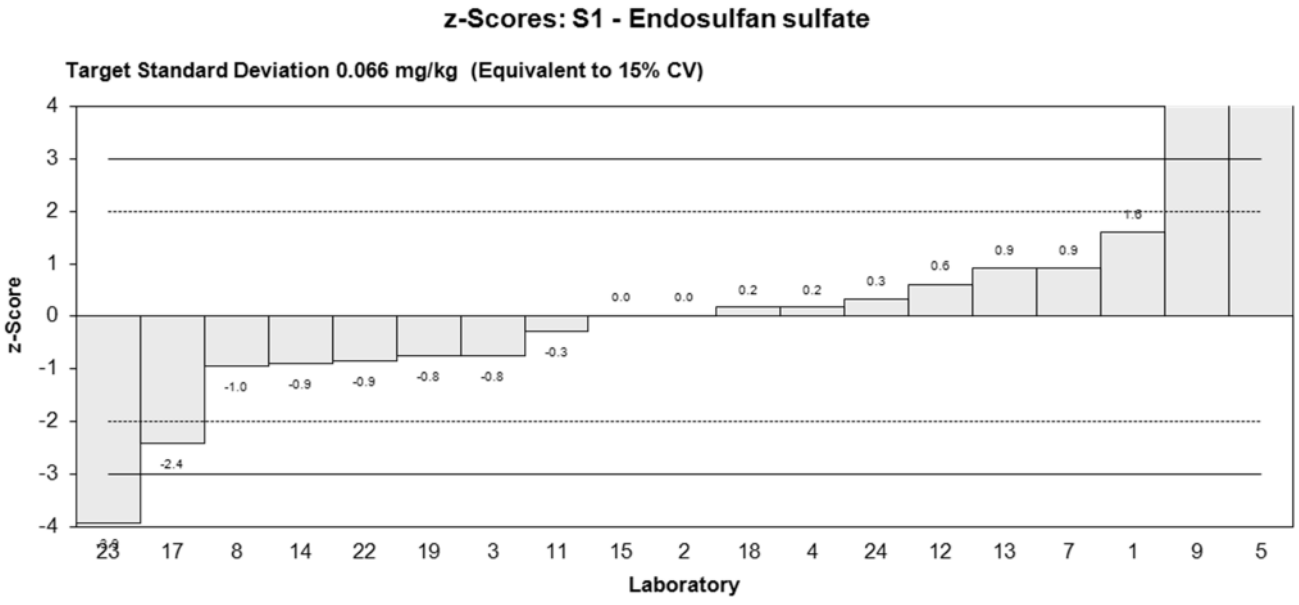


Figure 2. Bar chart z-scores

6.3 Scatter Plots of z-Scores

The z-score scatter plot is presented in Figure 3.

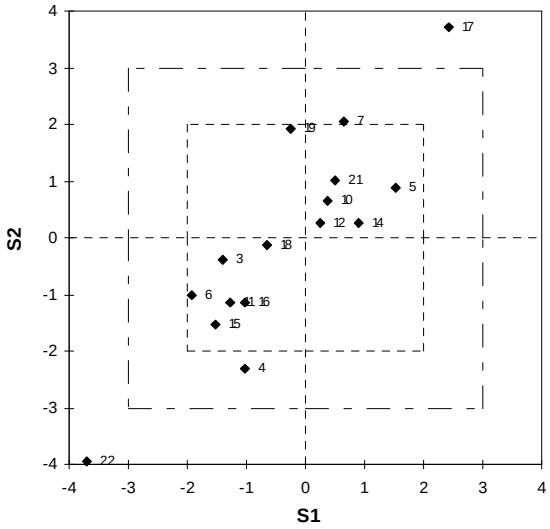


Figure 3 z-score scatter plot for sample S1 and S2

The plot has two squares, the inner square corresponding to a z-score of |2|, the outer square corresponding to a z-score of |3|. Laboratories falling within the centre square have z-scores with |z| < 2 for both samples. Laboratories falling between the inner and outer squares have z-sores with |z| between 2 and 3 for at least one sample. Laboratories falling outside the outer square have at least one z-score with |z| > 3.

Within laboratory and between laboratory variability is indicated in the same fashion as for a conventional Youden Plot. For laboratories plotted in the upper right and lower left quadrants, between laboratory variability predominates. For laboratories plotted in the upper left and lower right quadrants, within laboratory variation predominates.

6.4 Box-and-whisker plot

Box and whisker plots8 are helpful in interpreting the distribution of data. The diagram shows the quartiles of the data, using these as an indication of the spread. It is made up of a "box", which lies between the upper and lower quartiles. The median can also be indicated by dividing the box into two. The "whiskers" are straight line extending from the ends of the box to the maximum and minimum values. Example is presented in Figure 4.

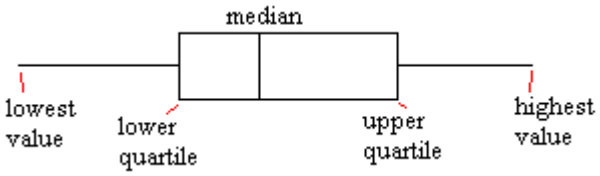


Figure 4 Box-and-whisker plot

6.5 Kernel density plot

An alternative to histograms for visualising the distribution of results is the kernel density estimate. Details about kernel density estimates are presented in AMC Technical Brief no 4. The technical brief

and the software required to produce kernel density plots are found at the Royal Society of Chemistry UK.9

The Kernel density plot is used to identify modes in the distribution of participants' results. It is also used to identify outlying results.

7 References

1. ISO13528:2015 (E), Statistical methods for use in proficiency testing by interlaboratory comparisons, ISO, Geneva, Switzerland.

2. Thompson, M., Ellison, S. L. and Wood, R., The International Harmonised Protocol for the Proficiency Testing of Analytical Chemistry Laboratories, Pure Appl. Chem., 78 (1), 145-196, 2006.

3. Lawn, R. E., Thompson, M. and Walker, R. F., Proficiency Testing in Analytical Chemistry, LGC, Teddington, UK, 1997.

4. Horwitz, W., Evaluation of analytical methods used for regulations of food and drugs, Anal. Chem., 54, 67A-76A.6, 1982.

5. Thompson, M., and Lowthian, P.J., A Horwitz-like function describes precision in a proficiency test, Analyst, 120, 271-272, 1995.

6. Thompson, M., Recent trends in inter-laboratory precision at ppb and sub-ppb concentrations in relation to fitness for purpose criteria in proficiency testing, Analyst, 125, 385-386, 2000.

7. Hibbert, D. B. and Gooding J. J., Data Analysis for Chemists – An introductory guide for students and laboratory scientists, first edition, Unversity Press, New York, 2006.

8. Stephen L. R. E., Barwick V. J. and Farrant T. J. D., Practical Statistics for the Analytical Scientist – A bench guide, 2nd edition, RSC Publishing, Cambridge, 2009.

9. Royal Society of Chemistry UK, http://www.rsc.org/, 2010.

8 Revision/Review History

Date	Issue Number	Reasons for revision
April 2006	1.0	First issue after move to NSW
August 2006	1.1	Issues raised at NATA audit addressed
November 2007	1.2	Issues raised at Internal audit addressed
February 2009	2.0	Issues raised at NATA audit addressed
December 2010	3.0	Complete revision
February 2012	3.1	Small amendments to Chapter 3, 5 and 6
August 2012	3.2	Changed from Pymble to North Ryde
September 2012	3.3	Issue raised at Internal audit addressed
July 2013	3.4	Review minor change to example chart.
February 2014	3.5	Histogram replaced with Kernel plot
May 2016	3.6	Invalid result definition expanded
October 2016	3.7	Amendments for homogeneity
September 2018	3.8	Renamed between laboratory CV to PCV
January 2019	3.9	Amended 3.1 and 5.2. Added 5.5
February 2019	3.10	Amended 5.3 and 5.4.