FNRI PROFICIENCY TESTING SUPPLEMENT ON STATISTICAL PROCEDURES

Feb 2016
2nd Edition

LEAH C. DAJAY
PTL Head

MARIO V. CAPANZANA, Ph.D.
Director

Republic of the Philippines
Department of Science and Technology
FOOD AND NUTRITION RESEARCH INSTITUTE
Gen. Santos Ave., Bicutan, Taguig City, Philippines
Website: www.fnri.dost.gov.ph
Tel. Nos.: (632)837-2934; 837-2071 to 82; Fax No.: (632) 837-3164

PREFACE

This document describes the statistical procedures that the Food and Nutrition Research Institute – Department of Science and Technology (FNRI-DOST) uses in (a) the analysis of the results of its proficiency testing (PT) programs, and (b) evaluation of test material homogeneity.

The procedures for the evaluation of PT results and test material homogeneity detailed here are based on ISO 13528:2005 and the International Harmonized Protocol for the Proficiency Testing of Analytical Chemical Laboratories (IUPAC Technical Report: 2006 IUPAC), with the exception of the criteria for the identification/deletion/exclusion of outlying laboratory result/s as well as the construction of the boxplot. For procedures, which are not specified in the aforementioned documents, the best judgment of the author was relied on.

Other statistical procedures and plots detailed in ISO 13528:2005 and IUPAC Protocol may be applied, whenever necessary and applicable, and with proper guidance from a statistician.

I. EVALUATION OF PROFICIENCY TESTING (PT) RESULTS

A. INTRODUCTION

Proficiency test results are assessed by comparison with assigned values derived from the consensus of results (consensus value) from participants, or values determined by a reference laboratory.

The consensus values are estimated using robust procedures. Robust procedures are used in the estimation of consensus values because the most commonly used measures of location and dispersion – **arithmetic mean and standard deviation** – are highly influenced by the presence of extreme outliers and their interpretation depends on an implicit assumption that they are a random sample from a normal distribution. The mean and standard deviation are the optimal estimators of location and dispersion, respectively, for a normal distribution but they can be substantially sub-optimal for distributions close to the normal.

It is very common in many fields to encounter data that have skewed distributions or contain outliers. Analytical data from testing laboratories often depart from the assumption that the data are a random sample from a normal distribution. It is often heavy tailed – contains a higher than expected proportion of results far from the mean – and sometimes contains outliers.

Outliers are values that are so far in value from the rest of the data that they may be viewed as coming from a different population, or the result of a measurement error. One way of coping with outliers is to exclude them from the calculation of the statistics. But when is it justifiable to exclude outliers in the calculation? The decision to exclude or retain an outlier depends on the understanding of the cause of the outlier and its impact on the results.

Employing tests such as Grubbs' test or the boxplot usually identifies suspect outliers. The use of the Grubbs' test presumes that the distribution of the variable is approximately normal. A boxplot, on the other hand, can be used in identifying outliers for both normal and non-normal distributions.

On the basis of some simple assumptions, outlier tests identify where it is likely to have a technical error but it does not assess or judge that the point is "wrong". In a data set, the value may be extreme but it could be the correct one. Only with experience or by identification of a certain cause can data be declared "wrong" and excluded from the computations. Generally, if more than 20% of the data are identified as outlying, the assumption about the data distribution and/or the quality of the data collected becomes questionable.

A convenient way of coping with outliers is to use **robust statistics**. Robust statistics includes methods that are largely unaffected by the presence of extreme values. "It provides an alternative way of summarizing results when they include a small proportion of outliers, without the requirement to identify specific observations as outliers or exclude them." [**1**].

Examples of robust statistics are the median and the mode for they are not highly influenced by the presence of outliers. "The ***median*** is the value in an ordered data set that has an equal number of data points on either side while the ***mode*** is the value of the peak of the distribution." [**2**].

Among the three statistics – mean, median and mode – the mode is least affected by the presence of outliers. However, because the calculation of the mode is more difficult than that of the mean or median, the mode has limited application.

B. SETTING THE STANDARD DEVIATION FOR PROFICIENCY ASSESSMENT

"The standard deviation for proficiency assessment (σ_p) is a parameter that is used to provide a scaling for the laboratory deviations from the assigned value and thereby define a z-score. The value is determined by "fitness-for-purpose" as it does not represent a general idea of how laboratories are performing, but how they ought to perform to fulfill their commitment to their clients." [**3**].

Fitness-for-purpose is the ability of a value to satisfy a set of conditions given by the application. "The **uncertainty of measurement** is a parameter associated with the results of a measurement that characterizes the dispersion of the value that could reasonably be attributed to the measurand." [**4**].

Most common approaches in setting the $\sigma_{pt,}$ of a measurand are the following:

(i) by collaborative trial data calculated using the formula:

$$\sigma_p = RSD_R \times x_{pt}$$

where:

RSD_R is the relative standard deviation of reproducibility from collaborative trials

x_{pt} is the assigned value from consensus of PT participants' results derived as a robust average using Alogarithm A of ISO 13528 expressed in appropriate units

(ii) by perception of how laboratories should perform, based on CV of previous PT results on appropriate (same or similar) matrix

$$\sigma_p = \frac{(CV \times x_{pt})}{100}$$

where:

CV is the coefficient of variation

x_{pt} is the assigned value from consensus of PT participants' results, derived as a robust average using Algorithm A of ISO 13528 expressed in appropriate units

(iii) by Horwitz equation in the absence of collaborative trial data for minerals using any of the formula:

$\sigma_p = 0.02(X_c^{0.8495})$

where:

X_c is the consensus value, σ_p and X_c are expressed as mass fraction

The standard deviation of reproducibility found in collaborative trials is generally considered an appropriate indicator of the best agreement that can be obtained between laboratories [**5**].

C. EVALUATION PROCESS FOR PT RESULTS

The evaluation of proficiency test results proceeds as follows:

- **Exclusion of invalid data**

  There may be instances where a participant's test result will be excluded from the calculation of a measurand's consensus value and its associated standard uncertainty. Reasons for exclusion are the following:

  a. method used for the measurand is not applicable to the food matrix (e.g. fat analysis using direct solvent extraction instead of acid hydrolysis);
  b. removal of extreme results or results that are identifiably invalid, (e.g., results caused by calculation errors or used wrong unit of measurements). Results which are out of range of the median ± 5 * σ_{pt} will be excluded based on the General Protocol of LGC standards Proficiency Testing **[4]**: and
  c. extremely low or high values identified by the boxplot rule.

- **Determination of the assigned value (x_{pt}) and its standard uncertainty (u_x)**

  - Calculation of the robust average (x*) or median (med (x)) for use as consensus value and the corresponding robust standard deviation (s*) or *MADe(x)*, whichever is applicable, of the test results:
  - Computation of the standard uncertainty of the consensus value (u) using the formula:

$$U_x = \frac{1.25 \times MADe(x) \text{ or } s^*}{\sqrt{n_2}}$$

where:

MADe(x) is the scaled median absolute deviation calculated using the formula:

$MADe(x) = 1.483 * med(d)$

s* is the robust standard deviation computed using Algorithm A of ISO 13528: 2005

n_2 is the number of data included in the computation

of consensus value

- **Calculation of performance statistics**

  z-scores are typically used in the evaluation of performance. The z-scores are calculated, using the consensus value and σ_{pt}, only when the consensus value is suitable for use as an assigned value.

  Section I.C.2 gives the details on the calculation of performance statistics.

- **Evaluation of performance**

  Section I.C.3 describes the steps in evaluating the performance of participating laboratories.

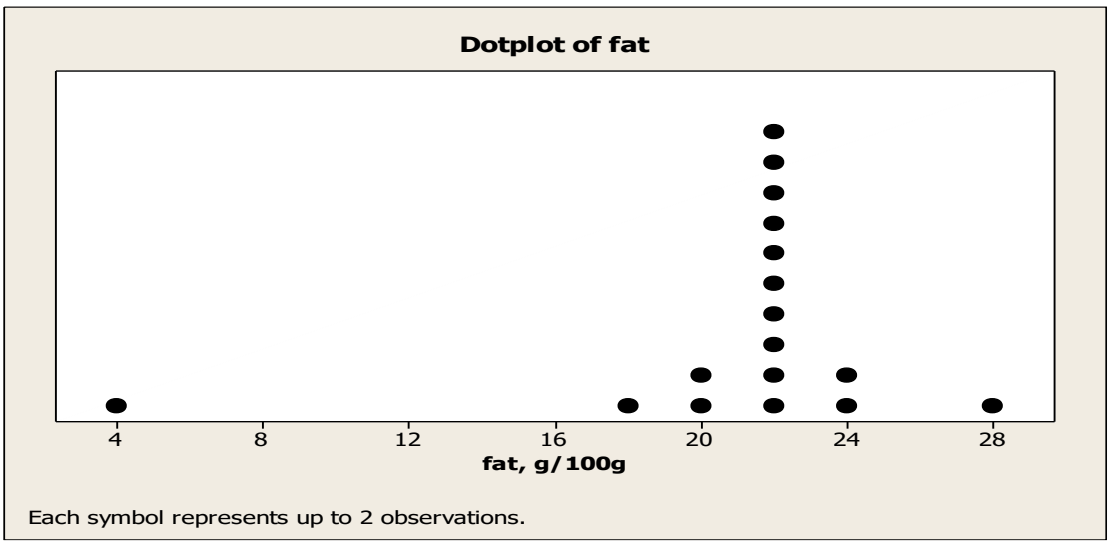
Graphical representations of some evaluation outputs are in Figures 1 to 5. The flowchart of the statistical evaluation process is presented in Figure 6 and the detailed steps and explanatory notes are given below.

1. Determination of the Assigned Value (X) and its Standard Uncertainty (u_X)

1.1. *Construction of a dotplot or histogram of the results*

When the number of data points in the plot is less than 50, a dotplot is constructed. When there are 50 or more data points, the histogram is used instead.

Figure 1



1.2. *Calculation of initial statistics (e.g. mean, median, minimum value, maximum value) of all data*

1.3. *Exclusion of data obtained using inapplicable method or expressed in wrong units of measure (e.g. fat analysis using direct solvent extraction instead of acid hydrolysis prior to solvent extraction)*

1.4. *Identification and exclusion of laboratories with extremely low or high reported results.*

Test results that satisfy the following boxplot criteria are considered as extreme outliers:

$$x > (Q3 + (3 \times IQR)) \text{ or } x < (Q1 - (3 \times IQR))$$

where

- x is the laboratory's average result
- Q1 is the first quartile of the laboratory averages
- Q3 is the third quartile of the laboratory averages and
- IQR is the interquartile range and computed as follows:

$$IQR = (Q3 - Q1)$$

1.5. *Removal of extreme results or results that are identifiably invalid, (e.g., results caused by calculation errors or used wrong unit of measurements). Results which are out of range of the median ± 5 * σ_{pt} will be excluded based on the General Protocol of LGC standards Proficiency Testing:*

1.6. *Calculation of the robust average (x*) for use as consensus value and the corresponding robust standard deviation (s*) of the test results with outlying data excluded using Algorithm A of ISO 13528:2005*

1.7. *Calculation of the standard uncertainty of the consensus value (*u*) using the following formula:*

$$u = \frac{1.25 \times s^*}{\sqrt{n}}$$

where:

- n is the number of data included in the computation of the robust average, and
- s* is the robust standard deviation computed using Algorithm A

1.8. *Determination of the suitability of the consensus value to be used as assigned value based on the ISO 13528:2015 criteria [3]:*

if $u \leq 0.3\sigma_{pt}$	-	u is negligible, z-scores can be issued;
if $u > 0.3\sigma_{pt}$	-	u is high, use the uncertainty of the assigned value

in the interpretation of performance, i.e. z'-scores can be issued,

if $\mu_x^2 + \sigma_{pt}^2 \leq \sigma_{rob}^2$

where:

σ_{pt} is the standard deviation for proficiency assessment
μ_x is the standard uncertainty of the assigned value
σ_{rob} is the robust standard deviation

1.9 *Abandonment of attempt to determine a consensus value*

The attempt to determine a consensus value is abandoned if the uncertainty of the consensus value is not negligible or is too high, i.e. $\mu_x^2 + \sigma_{pt}^2 \leq \sigma_{rob}^2$. There is no real consensus of results, thus no z-score is issued. However, the participants are provided with summary statistics (e.g. mean, median) of the data set as a whole.

2. Calculation of Performance Statistics

z-scores are the basis for evaluating the performance of participating laboratories. The z-scores are calculated using the consensus value and the standard deviation for proficiency assessment (σ_{pt}), only when the consensus value is suitable for use as the assigned value. The performance of individual PT participant laboratories was evaluated using the formula:

$$z = \frac{x - x_{pt}}{\sigma_{pt}}$$

where
- x is the participant's reported result
- x_{pt} is the assigned value from the consensus of the PT participants' results derived as a robust average or median
- σ_{pt} is the standard deviation for proficiency assessment

The laboratory z-scores are interpreted as follows:

|z-score| ≤ 2.0: "Satisfactory" **(S)** performance
2.0 < |z-score| < 3.0: "Warning" **(W)** signal
|z-score| ≥ 3.0: "Action" **(A)** signal

If the uncertainty of the assigned value is greater than $0.3\sigma_{pt}$, then the uncertainty can be taken into account by expanding the denominator for the calculation of performance score, such that:

$$z' = \frac{x - x_{pt}}{\sqrt{(\sigma_{pt}^2 + \mu_x^2)}}$$

where:
- x is the participant's reported result
- x_{pt} is the assigned value from the consensus of PT participants'
- σ_{pt} is the standard deviation for proficiency assessment
- μ_x is the standard uncertainty of the assigned value

z'-scores are interpreted in the same way as z-scores and using the same critical values of 2.0 and 3.0.

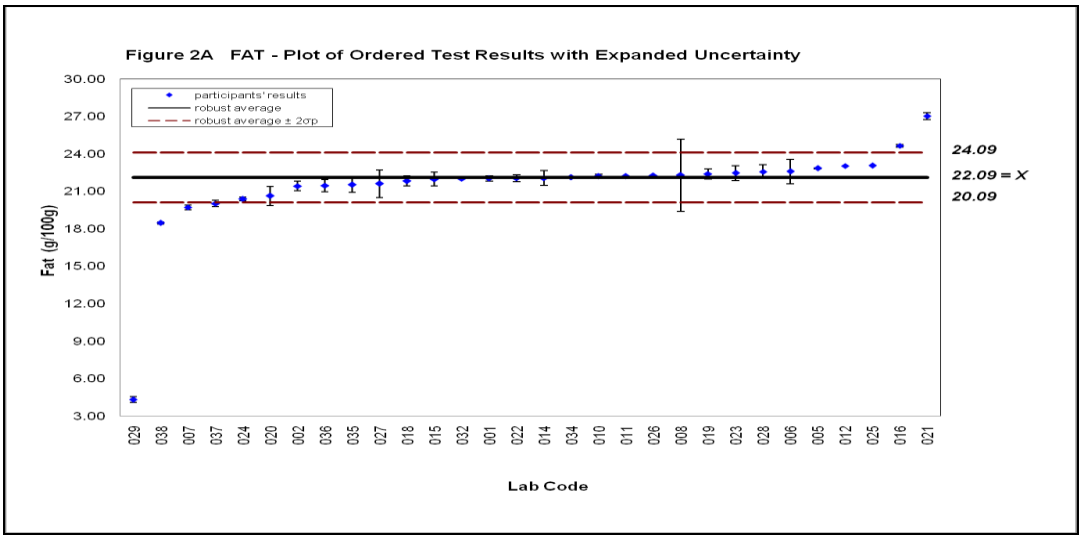
The plots of ordered test results with expanded uncertainty, ordered test results according to methods used and ordered z-scores are also used in evaluating performance. These, including the dotplot, are graphical means by which a participating laboratory can be readily compare its performance relative to the other laboratories.

3. Construction of Plots

3.1 *Construction of plot of ordered test results with expanded uncertainty*

The plot of ordered test results with expanded uncertainty (Figure 2) is a graphical display of each laboratory's test result with the reported expanded uncertainty. It shows the performance of each laboratory relative to the other laboratories. For example, in Figure 2, the test results starting from **Lab 024** to **Lab 025** are within the range of values of "Satisfactory" range: 20.37 to 23.06 g/100g. However, the test results of **Labs 029, 038, 007** and **037** are below the lower limit of the value for "Satisfactory" range, while **Labs 016** and **021** obtained test results above the upper limit of the value for "Satisfactory" range.

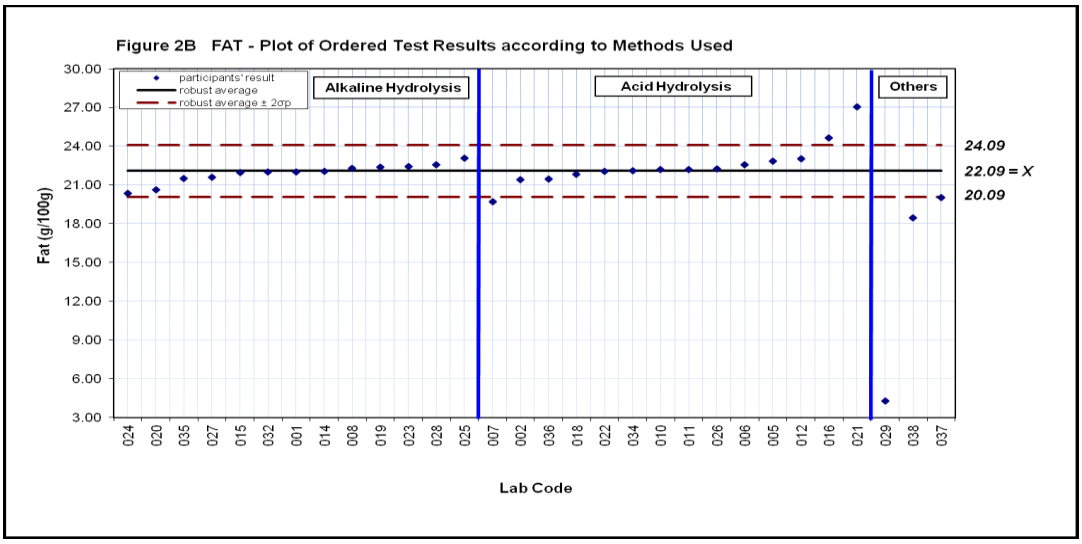
Figure 2



3.2 *Construction of plot of ordered test results according to methods used*

This plot is a graphical display of the participant's performance according to methods used. It shows if there are differences and clustering of results by method used. For example in Figure 3, comparable behavior of results was observed for both alkali and acid hydrolysis for fat (i.e. no clustering of data). When there is clustering of results, there is a need to conduct separate evaluation of participants' results by method used.

Figure 3



3.3 *Construction of plot of ordered z-scores*

The plot of ordered z-scores is a graphical display of the participants' performance. This plot shows each participant laboratory's performance relative to that of the other laboratories. From this plot, results outside the "Satisfactory" range (i.e. (| z-score | > 2.0) can be quickly identified. As illustrated in Figure 4, **Labs 029, 038, 007, 037, 016,** and **021** have results outside the "Satisfactory" range.

Figure 4

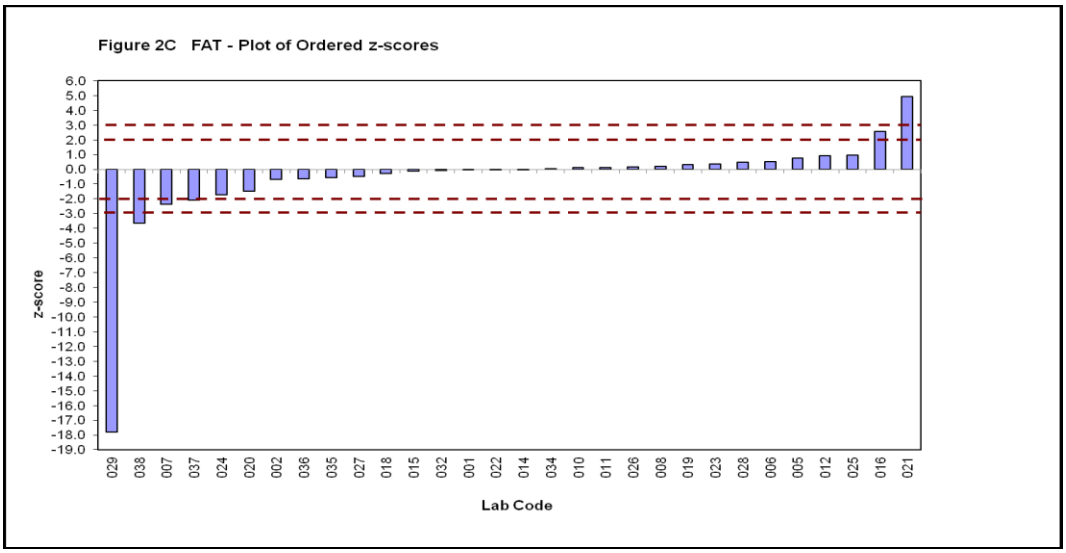
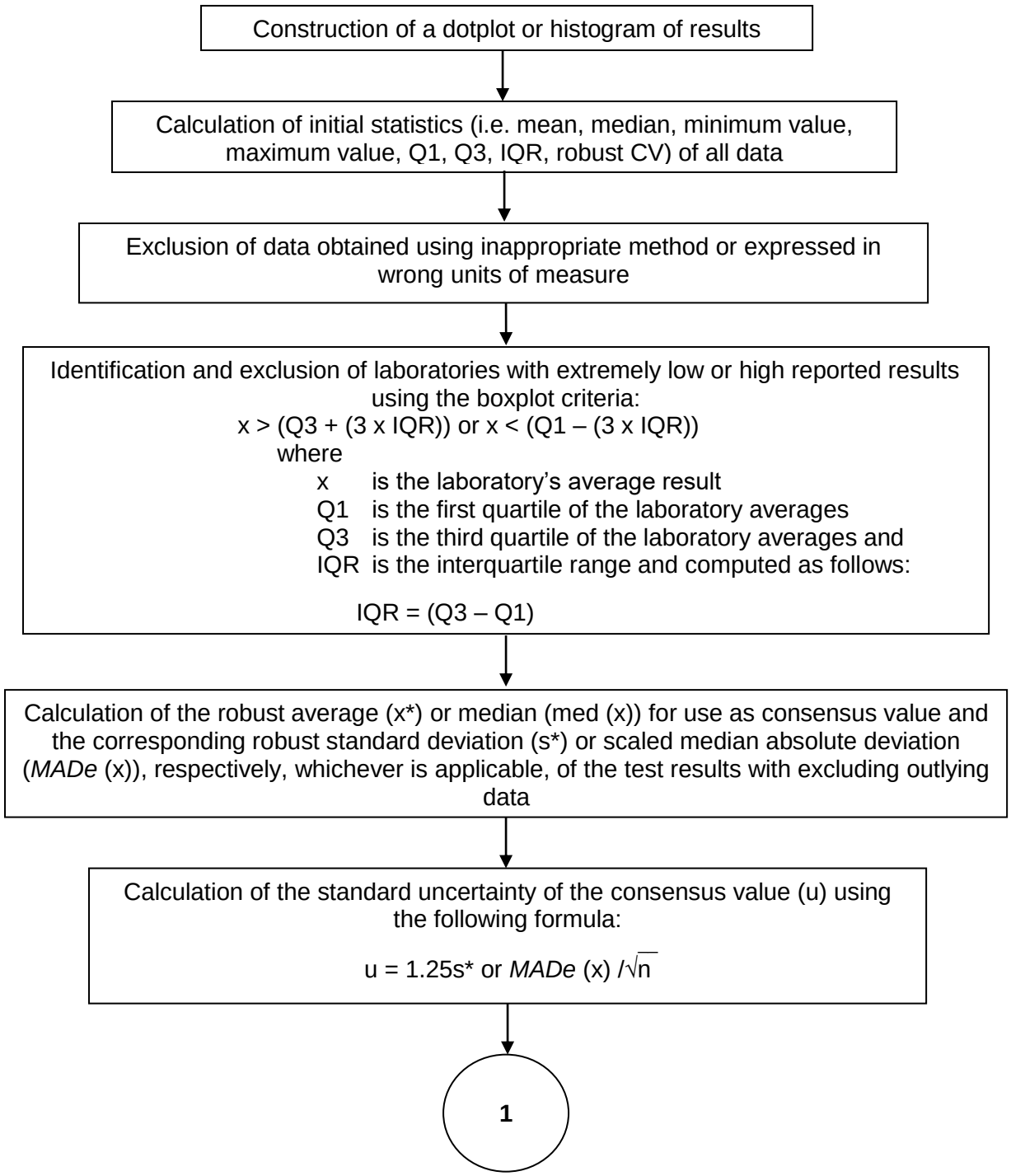
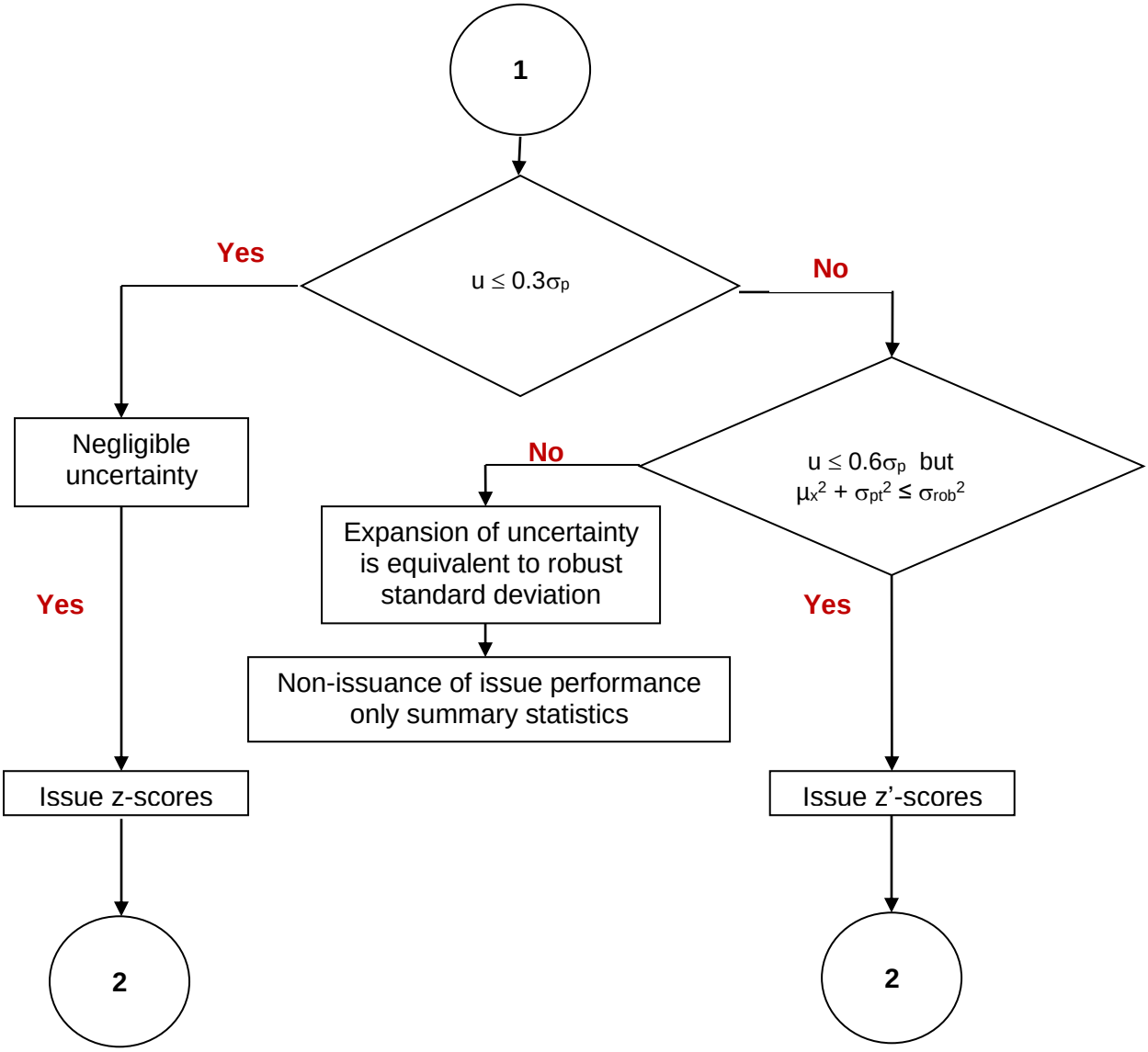
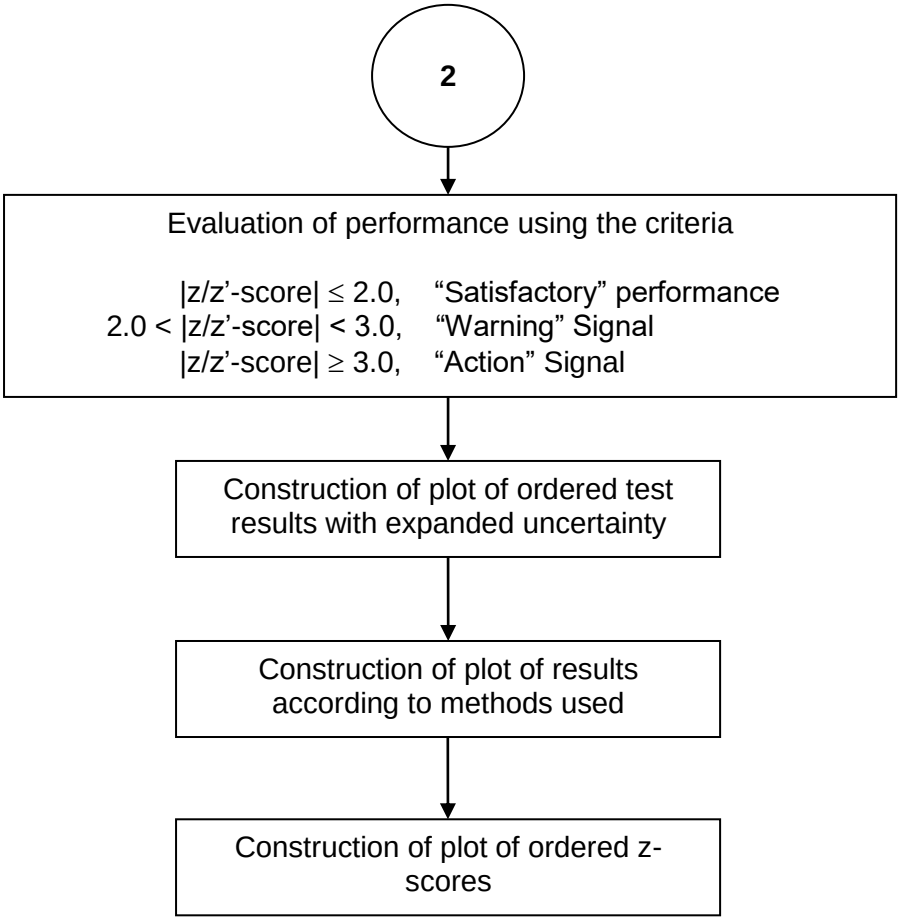


Figure 5
STATISTICAL EVALUATION PROCESS




1
$u \leq 0.3\sigma_p$
Yes
No
Negligible uncertainty
Yes
$u \leq 0.6\sigma_p$ but $\mu_x^2 + \sigma_{pt}^2 \leq \sigma_{rob}^2$
No
Expansion of uncertainty is equivalent to robust standard deviation
Non-issuance of issue performance only summary statistics
Yes
Issue z-scores
Issue z'-scores
2
2


2
Evaluation of performance using the criteria
|z/z'-score| ≤ 2.0, "Satisfactory" performance
2.0 < |z/z'-score| < 3.0, "Warning" Signal
|z/z'-score| ≥ 3.0, "Action" Signal
Construction of plot of ordered test results with expanded uncertainty
Construction of plot of results according to methods used
Construction of plot of ordered z-scores

II. EVALUATION OF TEST MATERIAL HOMOGENEITY

The statistical methods used in testing a material for homogeneity are:

- Cochran's test procedure for duplicate results
- Test for "adequate" homogeneity using ISO 13528 assessment criterion for homogeneity check
- Test for "sufficient" homogeneity by Fearn and Thompson (2001)

The following are the steps in conducting the homogeneity test:

1. **Selection** of 10 test samples in their final packaged form using systematic sampling using Microsoft Excel software,

2. **Separate homogenization** of the contents of each of the 10 selected packages by the appropriate techniques, to minimize within-package variability,

3. **Preparation** of two (2) sub-samples from each test sample using techniques appropriate to the test material, to minimize between-test-portion differences,

4. **Obtaining of a measurement result** on each of the twenty (20) sub-samples in a random order as in Step 1 of this Section, where applicable, and completing the whole series of measurements under repeatability conditions (i.e. same analyst, equipment, glassware, etc.).

5. **Examination of data** for pathologies,

    5.1. *Construction of a simple plot of duplicate results*

    Use any software that has the capability to construct a scatterplot of duplicate results, e.g., Excel.

    5.2. *Visual examination of a simple plot of the duplicate results and searching for diagnostic features such as:*

    - trends or discontinuities
    - nonrandom distribution of differences between first and second test results
    - excessive rounding; and
    - outlying results within samples

    5.3. *Testing for outlying results within samples using Cochran's test procedure for duplicate results*

    The Cochran's test procedure is as follows:

    5.3.1. Calculation of the sum, S_i, and difference, D_i, of each pair of duplicates, for i = 1, ..., m, where m = 10

    5.3.2. Calculation of the sum of squares, S_{DD}, of the 10 differences

$$S_{DD} = \sum D_i^2$$

5.3.3. Calculation of the ratio, C, and comparison of the result with the appropriate critical value

The Cochran's test statistic is the ratio of D^2_{max}, the largest squared difference, to this sum of squared differences

$$C = D^2_{max}/S_{DD}$$

For 10 test samples analyzed in duplicate, the critical values at 95% and 99% levels of confidence are 0.602 and 0.718, respectively. Refer to the IUPAC Technical Report for other values.

5.3.3.1. Close inspection of outlying pairs detected at the 95% or higher level of confidence for transcription or other errors. An outlying pair is rejected when there are irremediable analytical errors or if the difference between duplicate results is significant at the 99% level.

5.3.3.2. Deletion of duplicate results from a single test sample if they are significantly different from each other at the 99% level of significance.

5.3.3.3. Discarding of data if they contain discrepancies in two or more test samples. However, pairs of results with outlying mean (average) value but with no evidence of extreme variance (difference) are not discarded.

6. Testing for homogeneity

6.1. *Testing for "adequate" homogeneity*

6.1.1. Use of the same sum of squared differences in Step 5.3.2 to calculate the analytical variance, s^2_{an}

$$s^2_{an} = \sum D_i^2/2m$$

where:
D_i^2 is the difference of each pair of duplicates
m is the total number of samples, i.e. 10

6.1.2. Calculation of the variance V_S of the sums, S_i

$$V_S = \frac{\sum(S_i - \overline{S})^2}{(m-1)}$$

where:
S_i is the sum of each pair of duplicates
$\overline{S}$ is the mean of the S_i, $(1/m)\sum S_i$

6.1.3. Calculation of the sampling variance, s^2_{sam}

$$s^2_{sam} = \frac{(V_S/2 - s^2_{an})}{2}$$

or

$s_{sam}^2 = 0$, if the above estimate is negative.

Note: The quantities $V_S/2$ and s^2_{an} may be extracted from the analysis of variance table as the "between" and "within" mean squares, respectively.

6.1.4. Calculation of the sampling standard (between-samples) deviation, s_{sam}, and comparing it with $0.3\sigma_p$

6.1.4.1. If $s_{sam} > 0.3\sigma_p$, the test for adequate homogeneity has failed.

6.1.4.2. If $s_{sam} \leq 0.3\sigma_p$, the test for adequate homogeneity has been passed.

6.2. *Testing for "sufficient" homogeneity*

6.2.1. Calculation of the allowable sampling variance, σ^2_{all}, as

$$\sigma^2_{all} = (0.3\sigma_p)^2$$

where σ_p is the SD for PT assessment

6.2.2. Calculation of the critical value for the test as

$$c = F_1\sigma^2_{all} + F_2 s_{an}^2$$

where
$F_1 = 1.88$ and $F_2 = 1.01$ (for 10 test samples measured in duplicate; 95% level of confidence). Refer to the IUPAC 2006 Technical Report for other values.

6.2.3. Use of the calculated sampling variance, s^2_{sam}, in Step 6.1.3 and making decision based on the following criteria:

6.2.3.1. If $s_{sam}^2 > c$

there is evidence at the 95% level of confidence that the sampling standard deviation in the population of samples exceeds the allowable fraction of σ_p; therefore, the test for homogeneity has failed.

6.2.3.2. If $s_{sam}^2 \leq c$

there is no evidence at the 95% level of confidence that the sampling standard deviation in the population of samples exceeds the allowable fraction of σ_p; therefore, the test for homogeneity has been passed.

BIBLIOGRAPHY

1. Analytical Methods Committee. "Robust statistics: a method of coping with outliers", AMC Technical Brief No. 6. 2001. Available online at <http://www.rsc.org/images/ brief6_tcm18-25948.pdf> [April 30, 2007].

2. Hedges, S. and Shah, P. 2003. *Comparison of mode estimation methods and application in molecular clock analysis*Available online at <http://www.biomedcentral.com/ content/pdf/1471-2105-4-31.pdf> [April 30, 2007]

3. Thompson, M., Ellison, S.L.R. and Wood, R. 2006. The International Harmonized Protocol for the proficiency testing of analytical chemistry laboratories. *Pure Applied Chem.* **78** (1), 145-196.

4. International Organization of Standardization. *ISO/IEC Guide 43-1: Proficiency testing by interlaboratory comparisons – Part 1: Development and operation of proficiency testing schemes*, 2nd ed., ISO, Geneva, Switzerland.1997.

5. International Organization of Standardization. *ISO 13528:2005. Statistical methods for use in proficiency testing by Interlaboratory comparisons,* ISO, Geneva, Switzerland.

6. Food Analysis Performance Assessment Scheme (FAPAS). 2002. *Proficiency Testing Protocol: Organization and Analysis of Data*, 6th ed., FAPAS, York, UK.

7. LGC Standards Proficiency Testing. (September 2014). General Protocol: Proficiency Testing Schemes. p 14.

8. _____, *Random Error and Systematic Error*. Available online at <http://www.phys.selu.edu/rhett/plab193/labinfo/Error_Analysis/05_Random_vs_systematic.html> [April 30, 2007]

9. Burke, S. *Missing values, outliers, robust statistics & non-parametric methods.* Available online at <http://www.lgceurope.com/lgceurope/data/articlestandard/lgceurope/502001/4509/article.pdf> [April 30, 2007]

10. _____, *Box and Whisker Plots.* 2003. Available online at <http://www.mathcentre.ac.uk/resources/leaflets/mathcentre.business/box_and_whisker_plots.pdf > [April 30, 2007]

11. Thompson, M. and Fearn, T. 2001. A new test for 'sufficient homogeneity. *Analyst*, **126**, 1414-1417.

12. Hall, J.M., Bean, V.E. and Mattingly, G.E. 1999. *Preliminary Results from Interlaboratory Comparisons of Air Speed Measurements Between 0.3 m/s and 15 m/s.* 1999 NCSL Workshop & Symposium. Available online at <http://www.cstl.nist.gov/ div836/836.01/PDFs/1999/NCSL_056.pdf> [April 30, 2007].