Published in final edited form as:

J Stat Plan Inference. 2008 November 1; 138(11): 3336–3349. doi:10.1016/j.jspi.2006.07.017.

Statistical Scoring Procedures Applicable to Laboratory Performance Evaluation¹

A. S. Hedayat,

Department of Mathematics, Statistics, and Computer Science (M/C 249), University of Illinois at Chicago, 851 S. Morgan Street, Chicago, IL 60607-6045

Guoqin Su, and

Novartis Pharmaceuticals Corporation, 59 Route 10, East Hanover, NJ 07936-1080

W. Elane Streets

Spallation Neutron Source, Oak Ridge National Laboratory, 1 Bethel Valley Road, Bldg. 8600, PO Box 2008, MS 6460, Oak Ridge, TN 37831-6460

Abstract

Two statistical scoring procedures based on p-values have been developed to evaluate the overall performance of analytical laboratories performing environmental measurements. The overall score of bias and standing are used to determine how consistently a laboratory is able to measure the true (unknown) value correctly over time. The overall score of precision and standing are used to determine how well a laboratory is able to reproduce its measurements in the long run. Criteria are established for qualitatively labeling measurements as Acceptable, Warning, and Not Acceptable, and for identifying areas where laboratories should re-evaluate their measurement procedures. These statistical scoring procedures are applied to two real environmental data sets.

Keywords

Bias; precision; interlaboratory study; proficiency testing; random effect model

1. Introduction

Environmental measurements are made daily by many federal, state, and private agencies. The data from these measurements are used for a wide variety of purposes, including assessment of health effects, the establishment of standards and guidelines, and enforcement of regulatory requirements. It is imperative that the precision and bias of the data be assured so that decisions concerning environmental safety and quality are based on data with known reliability.

Therefore, relevant laboratories should be evaluated for their competence and be accredited appropriately. For the evaluation of laboratory competence and laboratory accreditation, the reader could consult, for example, the American Society for Testing and Materials standards E548 (ASTM, 1995a), E994 (ASTM, 1995c) and E1187 (1995d). In the assessment of a laboratory's competence, important criteria include the education, training, technical

¹The research related to this work is supported by the National Science Foundation (NSF) Grant DMS-0603761. The contents are solely the responsibility of the authors and do not necessarily represent the official view of NSF

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

knowledge, and experience of its staff; its equipment and reference materials; its standard operating procedures; and its quality assurance system. Although a laboratory's facilities, equipment, and personnel may be assessed through questionnaires and on-site examination, the actual ability of a laboratory to produce reliable testing results can only be verified by examining its performance through proficiency testing (Kirkpatrick and Horlick, 1983).

There are two types of interlaboratory studies. The first type monitors and evaluates a laboratory's performance. Such studies are generally called proficiency testing studies, interlaboratory performance evaluation studies, or interlaboratory comparison studies (Twomey, 1983; Kirkpatrick and Horlick, 1983; Britton and Lewis, 1986; ASTM, 1995e; Dilbeck and Honsa, 1989; Lawn, *et al.*, 1997). The second type of interlaboratory study determines the precision and bias of a test method (Youden and Steiner, 1975; Scott, *et al.*, 1990; Kristiansen, 1991; Mandel, 1991, 1994, 1997; ASTM, 1995b). This type is called a study of a test method or method validation study (Lawn, *et al.*, 1997).

The purpose of proficiency testing studies is to measure a laboratory's ability to reproduce reference values provided by the study coordinator. A proficiency test checks laboratory-testing performance by means of interlaboratory studies. It usually consists of the following steps: decision on ground rules, which may include development of pass/fail criteria; preparation and distribution of specimens; receipt of specimens in satisfactory condition; testing of specimens; expeditious return of data to source; data analysis; and report preparation. According to ASTM standard E1301 (ASTM, 1995e), the following test designs have been used: the Youden two-sample design wherein two similar specimens are submitted for each test during a given period, one or more unrelated specimens are submitted once for a single test, one or more specimens are submitted for two to four or more replicate tests at one time or over a period of time. The main statistical methods utilized here include the estimates of mean and standard deviation and the construction of prediction/evaluation intervals.

A method validation study yields the precision and bias of the method under study. Organizations in charge of standardization, such as ASTM and the Association of Official Analytical Chemists (AOAC), require that all test methods include statements of precision and bias. The precision of a test method is a measure of the closeness of agreement between test results obtained using the prescribed conditions of the test method. In general, this approach has been used by standardization organizations to consider the precision under two conditions, i.e., repeatability and reproducibility. Bias is a consistent or systematic difference between a set of test results and an accepted reference value of the property being measured. In general, p laboratories test q samples and perform n replicate tests in each sample. The samples should be carefully chosen so as to cover the entire range of interest. The p laboratories and q samples form a two-way table in which the (i, j) -th cell contains the n replicate test results of the i -th sample produced by the j -th laboratory. There is a vast literature on the estimation of repeatability and reproducibility and analysis of two-way tables (Youden and Steiner, 1975; Kristiansen, 1991; Mandel, 1991, 1994, 1997; ASTM, 1995e).

The design and statistical analysis are generally quite different for these two types of studies. In design of a method validation study, only one method is specified to measure a property, and that method is usually a new method. Further, samples should be prepared so that their concentrations cover the range of interest, and laboratories are usually instructed to perform equal numbers of replicate tests on each sample. In design of proficiency testing studies, test methods may or may not be specified, and different laboratories may or may not be allowed to use different test methods. These test methods are usually standard/routine ones. Single or multiple samples with single or multiple analytes may be distributed to participating laboratories. The statistical analyses of a method validation study usually include the estimation of repeatability, reproducibility, and bias of the test method. Such studies emphasize the overall

view of understanding and execution of the test method by a suite of laboratories and do not focus on the performance of an individual laboratory. The statistical analyses of proficiency testing studies usually include the estimation of mean and standard deviation and the construction of evaluation intervals for assessing the test results of all participating laboratories. Such studies focus on an individual laboratory's test results compared with assigned values (Lawn, *et. al.*, 1997) or test results for other laboratories.

In this paper we have developed statistical scoring procedures to evaluate laboratory performance. The first example is for a national proficiency testing study. The second example evaluates laboratory performance in a round-robin test of a method, as a way to help us decide whether or not to include a given laboratory's data in the assessment of the method.

2. Setup of performance evaluation program

In this section, we introduce the setup of a performance evaluation program to apply our statistical scoring procedures for assessing overall laboratory performance. This approach uses data from all the analytes in a performance evaluation study that a laboratory has reported to give a single performance score to that laboratory. The performance of a group of laboratories can be fairly compared, even when the performance pattern of individual laboratories (i.e., analysis of individual matrix-analytes) is different. This technique can also be used to identify specific areas of analysis where a laboratory needs to improve its performance.

The organizer of the performance evaluation program prepares homogeneous environmental and/or spiked samples in matrices and distributes them to the participating laboratories. A homogeneous sample may contain more than one analyte, and different homogeneous samples may contain different analytes or a given analyte at different levels. The organizer pre-specifies the analytes and their units for each sample. A combination of an analyte and a homogeneous sample in a matrix is called a *cell*. Since different cells most likely contain different analytes that may require a laboratory to use different measurement procedures, our statistical analysis will consider the measurements for each cell as an independent sample. A cell score of bias will be assigned to a laboratory that provides measurements in a cell. If a laboratory provides replicate measurements, a cell score of precision will also be provided. Cell scores of bias will be combined into an overall score of bias for all cells that a laboratory has participated. Similarly, cell scores of precision will be combined into an overall score of precision for all cells that a laboratory has participated and made replicate measurements. Finally, our statistical purposes include:

- a. Setting up a criterion for labeling the reported concentrations as Acceptable (A), Warning (W), and Not Acceptable (NA).
- b. Setting up a criterion for scoring (standing) the overall performance of each laboratory.
- c. Identifying areas where laboratories should re-evaluate their measurement procedures.

3. Random Effect Model

For a given cell, let m denote the number of laboratories that determine the concentration of the analyte and n_i denote the number of independent measurements requested by the organizer and reported by Laboratory i . The replication number, n_i , usually does not depend on i and for almost every cell, $n_i = 1, 2, 3$, or 4 . Let y_{ij} denote the j -th measurement reported by Laboratory i , where $i = 1, 2, \dots, m$; $j = 1, 2, \dots, n_i$.

We postulate the following random effect model for the data collected for the given cell,

$$y_{ij} = \mu + \beta_i + \varepsilon_{ij}, \quad i=1, 2, \dots, m, \quad j=1, \dots, n_i, \quad (1)$$

where

- μ is the true concentration of the analyte in the cell.
- β_i is a random effect produced by Laboratory i (between laboratory variation) when measuring the concentration of the analyte in the cell with $E(\beta_i) = 0$ and $\text{Var}(\beta_i) = \sigma_1^2$.
- ε_{ij} is the error term produced by Laboratory i when it makes the j -th measurement of the concentration of the analyte in the cell (within laboratory variation) with $E(\varepsilon_{ij}) = 0$ and $\text{Var}(\varepsilon_{ij}) = \sigma_2^2$. If all laboratories make only one measurement, ε_{ij} is dropped from the model or σ_2^2 is assumed to be 0.
- β_i and ε_{ij} are assumed to be independent random variables. They are also assumed to be normal random variables. Other distributions with left- or right-skewed tails could be considered to fit the data for some cells. When necessary, we will transform the data in such cells.

Remark 1

This random effect model is postulated after analyzing several very large data sets, consulting with chemists, and reviewing the literature. In the analyses of several very large data sets, the between-laboratory variation (σ_1^2) is usually much larger than the within-laboratory variation (σ_2^2). The random effect model was used by ASTM standard E691-92 (ASTM, 1995b), in which $\sigma_1^2 + \sigma_2^2$ is the variance of a measurement made by any laboratory and is called *reproducibility variance*, and σ_2^2 is the within-laboratory variation and is called *repeatability variance*.

We introduce the following notations to simplify our presentation:

$$N = \sum_{i=1}^m n_i, \quad \bar{y}_i = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i}, \quad s_i^2 = \frac{\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2}{n_i - 1},$$

$$\bar{y} = \frac{\sum_{i=1}^m \sum_{j=1}^{n_i} y_{ij}}{N}, \quad Q_1 = \sum_{i=1}^m n_i (\bar{y}_i - \bar{y})^2, \quad Q_2 = \sum_{i=1}^m (n_i - 1) s_i^2,$$

Analysis of variance method, also called Henderson's method 1 (see Searle, 1971), will be used to estimate μ , σ_1^2 , σ_2^2 . More specifically, we have

$$E(\bar{y}) = \mu, \quad E(Q_1) = (N - \frac{\sum_{i=1}^m n_i^2}{N}) \sigma_1^2 + (m - 1) \sigma_2^2,$$

$$E(Q_2) = (N - m) \sigma_2^2,$$

and unbiased estimators of μ , σ_2^2 , and σ_1^2 are given by

$$\widehat{\mu} = \bar{y}, \quad (3)$$

$$\widehat{\sigma}_2^2 = \frac{Q_2}{N - m}, \quad (4)$$

$$\widehat{\sigma}_1^2 = \frac{Q_1 - (m - 1)\widehat{\sigma}_2^2}{N - \frac{\sum_{i=1}^m n_i^2}{N}}, \quad (5)$$

respectively. If all n_i 's are equal to 1 ($n_i \equiv 1$), then we assume that $\sigma_2^2 = 0$. If $\widehat{\sigma}_1^2$ is negative, it will be replaced by 0. When n_i does not depend on i , the above three estimators are minimum variance unbiased estimators [see Graybill (Theorem 15.2.4), 1976]. In general, $\bar{y}$, Q_1 , and Q_2 are independent and

$$\frac{Q_2}{\sigma_2^2} \sim \chi^2(N - m).$$

Further, if $n_i = n$ for all i , then

$$\begin{aligned} E(Q_1) &= n(m - 1)\sigma_1^2 + (m - 1)\sigma_2^2, \\ \frac{Q_1}{n(m - 1)} &= \widehat{\sigma}_1^2 + \frac{\sigma_2^2}{n}, \\ \frac{Q_1}{n\sigma_1^2 + \sigma_2^2} &\sim \chi^2(m - 1). \end{aligned}$$

4. Statistical Scoring Procedures

4.1 Cell Score of Bias and Cell Score of Precision

For a given cell as specified in Section 3 and a particular laboratory, say Laboratory k , that reports measurements for the concentration of the cell, we are interested in two important quantities of the laboratory's measurement procedure: $\bar{y}_k$ and s_k^2 . $\bar{y}_k$ and s_k^2 are unbiased estimates for μ and σ_2^2 , respectively. We have developed procedures to assess these two quantities.

First we observe that

$$z_k^c = \frac{\bar{y}_k - \mu}{\sqrt{\sigma_1^2 + \sigma_2^2/n_k}} \sim N(0, 1). \quad (6)$$

This z -score was used by Lawn, *et. al.* (1997) as performance score when μ and σ_1^2 are known and there is no replication for any laboratory (ie., $n_i \equiv 1$ and $\sigma_2^2 = 0$). The approach we will

recommend here is to estimate the unknown parameters μ , σ_1^2 , and σ_2^2 in (6) using estimators in (3), (4), and (5) using data excluding the observations produced by Laboratory k . That is, a laboratory's measurements will be evaluated by using the consensus values of other laboratories' measurements. More specifically, let

$$\bar{y}^{(k)} = \frac{\sum_{i \neq k} \sum_{j=1}^{n_i} y_{ij}}{N - n_k}, \quad Q_1^{(k)} = \sum_{i \neq k} n_i (\bar{y}_i - \bar{y}^{(k)})^2, \quad Q_2^{(k)} = \sum_{i \neq k} (n_i - 1) s_i^2 = Q_2 - (n_k - 1) s_k^2.$$

μ is estimated by $\bar{y}^{(k)}$, and σ_2^2 and σ_1^2 are estimated by:

$$\begin{aligned} \widehat{\sigma}_2^{(k)^2} &= \frac{Q_2^{(k)}}{N - n_k - (m - 1)}, \\ \widehat{\sigma}_1^{(k)^2} &= \frac{Q_1^{(k)} - (m - 2) \widehat{\sigma}_2^{(k)^2}}{\frac{\sum_{i \neq k} n_i^2}{N - n_k}}, \end{aligned}$$

respectively. After substituting these estimators into (6), we obtain the following result:

$$T_k^c = \frac{\bar{y}_k - \bar{y}^{(k)}}{\sqrt{(\widehat{\sigma}_1^{(k)^2} + \widehat{\sigma}_2^{(k)^2} / n_k)(1 + 1 / [m - 1])}} \sim t_{m-2} \quad (7)$$

where t_v denotes a central t distribution with v degrees of freedom. If n_i does not depend on i , then the distribution of T_k^c in (7) is exactly a central t distribution. The approximation in (7) is reasonable since n_i 's are not expected to differ greatly from laboratory to laboratory as they are usually decided to be equal by the organizer. Further simplification leads to

$$\begin{aligned} Q_1^{(k)} &= Q_1 - \frac{N n_k}{N - n_k} (\bar{y}_k - \bar{y})^2, \\ T_k^c &= \frac{N (\bar{y}_k - \bar{y})}{(N - n_k) \sqrt{(\widehat{\sigma}_1^{(k)^2} + \widehat{\sigma}_2^{(k)^2} / n_k)(1 + 1 / [m - 1])}}. \end{aligned}$$

These two formulas greatly simplify the computations of $\widehat{\sigma}_1^{(k)^2}$ and T_k^c since one does not need to compute $\bar{y}^{(k)}$. If μ is known (for example, an assigned value or a known value for a spiked sample), then T_k^c might be replaced by

$$T_k^c = \frac{\bar{y}_k - \mu}{\sqrt{(\widehat{\sigma}_1^{(k)^2} + \widehat{\sigma}_2^{(k)^2} / n_k)}}.$$

To reflect the degree of bias for Laboratory k 's measurements in the cell, we define the following score based on two-sided p -value (in percentage) of T_k^c , called the score of bias, be assigned to Laboratory k :

$$B_k^c = 100 \cdot P(|t_{m-2}| \geq |T_k^c|). \quad (8)$$

That is, B_k^c is the probability in percentage that the t distribution with $m - 2$ degrees of freedom could produce a result whose absolute value is at least as large as the observed $|T_k^c|$. B_k^c is a decreasing function of $|T_k^c|$. If μ , σ_1^2 , and σ_2^2 are known, we replace T_k^c by z_k^c and t distribution by normal distribution in (8). Hence, B_k^c is a uniform random variable on $[0, 100]$.

The use of p-value as performance score is similar to the use of z -score (Lawn, *et. al.* 1997), since both are interpreted based on the normal approximation. However, the interpretation of p-value is more straightforward.

If Laboratory k has a bias ($E(\bar{y}_k) \neq \mu$) and/or a lower precision ($\text{Var}(y_{kj}) > \sigma_1^2 + \sigma_2^2$), $|T_k^c|$ is more likely to be larger than the absolute value produced by a central t distribution. Hence, the cell score of bias, B_k^c , for Laboratory k is more likely to be smaller than expected. If Laboratory k does not have a bias ($E(\bar{y}_k) = \mu$) but has a higher precision ($\text{Var}(y_{kj}) < \sigma_1^2 + \sigma_2^2$), $|T_k^c|$ is more likely to be smaller than the absolute value produced by a central t distribution. Hence, the cell score of bias, B_k^c , for Laboratory k is more likely to be larger than expected. Hence, a good laboratory is expected to receive a high cell score and a poor laboratory is expected to receive a low cell score.

If $n_k > 1$ and $N - n_k - m + 1 > 1$, then

$$F_k^c = \frac{(N - n_k - m + 1)s_k^2}{Q_2^{(k)}} \sim F_{n_k-1, N-n_k-m+1} \quad (9)$$

where F_{v_1, v_2} denotes the F distribution (or random variable) with v_1 and v_2 degrees of freedom. To reflect the precision of the performance by laboratories participating in the study, we define the following score based on the p-value (in percentage) of F_k^c :

$$P_k^c = 100 \cdot P(F_{n_k-1, N-n_k-m+1} \geq F_k^c). \quad (10)$$

P_k^c is the probability in percentage that the random mechanism of the F distribution could produce a result no better than the observed F_k^c . Hence, P_k^c is a decreasing function of F_k^c and is a uniform random variable on $[0, 100]$.

If Laboratory k has a lower precision ($\text{Var}(y_{kj}) > \sigma_1^2 + \sigma_2^2$), then F_k^c is more likely to be larger than that produced by the F distribution, and P_k^c is more likely to be smaller than expected. On the other hand, if Laboratory k has a higher precision ($\text{Var}(y_{kj}) < \sigma_1^2 + \sigma_2^2$), then F_k^c is more likely to be smaller than that produced by the F distribution, and P_k^c is more likely to be larger than expected. Hence, a laboratory with high precision is expected to receive a high cell score and a laboratory with low precision is expected to receive a low cell score.

4.2 Overall Score of Bias

We recommend the following overall score of bias be given to Laboratory k for its participation in the study based on all measurements it reported:

$$B_k = \text{Mean of } B_k^c \text{ for those cells Lab } k \text{ participated.} \quad (11)$$

To compare the biases of two laboratories that have participated in a different collection of cells, we must do more than compare their overall scores of bias. Analogously, one cannot assert that a student with a grade point average of 5.0 based on two courses is better than another student with an average of 4.9 based on ten courses. To compare the performance of bias for laboratories more reasonably, we recommend an overall adjusted score of bias based on the standing for the overall score of bias for Laboratory k as follows

$$D_k^b = 100 \cdot P(\bar{U}_M \leq B_k), \quad (12)$$

where M is the number of cells Lab k participated and $\bar{U}_v$ denotes the average of v uniform random variables on $[0, 100]$. D_k^b is an increasing function of B_k . The interpretation of D_k^b is that, as far as bias is concerned, Laboratory k is better than D_k^b percent of laboratories participating in the performance evaluation study. Thus the combination of the overall score of bias and its standing can be used to determine how consistently a laboratory is able to measure the true (unknown) concentration correctly over time.

Lawn, *et. al.* (1997) proposed two combinations of scores based on z -scores, the rescaled sum of z -scores (RSZ) and the sum of squared z -scores (SSZ). The RSZ score is interpreted based on the normal approximation while the interpretation of the SSZ score is based on the χ^2 approximation. However, high scores of opposite sign tend to cancel out each other and allow the RSZ score to remain at a small absolute value. The SSZ score overcomes this drawback. But, there is no simple way of making the statistic commensurate for different degrees of freedom. On the other hand, the influence of one or more very large or very small (negative) z -scores on RSZ or SSZ scores is unlimited.

4.3 Overall Score of Precision

Finally, we recommend the following overall score of precision be given to Laboratory k for its participation in the study:

$$P_k = \text{Mean of } P_k^c \text{ for those cells with replicate measurements Lab } k \text{ participated.} \quad (13)$$

Comparing the performance of precision for two laboratories requires more than comparing their overall scores of precision. Similar to the proposed overall score of bias, in order to compare the performance of precision for laboratories more reasonably, we recommend that the following standing for the overall score of precision be given to Laboratory k :

$$D_k^p = 100 \cdot P(\bar{U}_{M'} \leq P_k), \quad (14)$$

where M' is the number of cells Lab k participated and made replicate measurements. Note that D_k^p is an increasing function of S_k^p . The interpretation of D_k^p is that the precision of Laboratory k is better than D_k^p percent of laboratories that participated in the study. In short, the overall score of precision and its standing can be used to determine how well a laboratory is able to reproduce its measurements over time.

4.4 Outliers

There is a practical issue concerning outliers that should be discussed. For real data sets, it is not unusual that some laboratories might report measurements that are in serious discord with the measurements reported by the great majority of laboratories. For environmental pollution data sets, dealing with such measurements (or outliers) is an unavoidable problem (Gilbert, 1987). Hence, data screening and data validation should be done before using the above scoring procedures, albeit they are not the main subjects of this paper. For the above scoring procedures, the most important aspects are the estimation of μ , σ_1^2 , and σ_2^2 , and the determination (approximate) of the distributions of T_k^c and F_k^c . We will use statistical methods that involve outlier detection and elimination. However, outlier detection is a very sensitive issue here. Discarding a fair proportion of the measurements could lead to underestimation of both the between-laboratory variation (σ_1^2) and the within-laboratory variation (σ_2^2). Therefore, a laboratory's results could appear to be more precise than they really are. This is clearly an undesirable state of affairs. Thus, we will use a rather conservative method to handle outliers.

For a given cell, each measurement y_{ij} has the same mean μ and the same variance $\sigma_1^2 + \sigma_2^2$. Under our approach in which observations between laboratories are independent, observations within a laboratory are dependent. This generates a problem if we use the kurtosis method (Barnett and Lewis, 1978) without any modification for outlier detection. The modification we recommend is this. First use the kurtosis method at the 1% level to detect any potential outliers. Then, draw a histogram of the data to see if these potential outliers are in serious discord with the great majority of measurements. If that proves to be the case, then treat the measurement as an outlier. Otherwise, keep it in the data set.

4.5 Labeling Measurements

We will now set up a criterion for labeling the reported concentrations as Acceptable (**A**), Warning (**W**), and Not Acceptable (**NA**). The emphasis is *not* on individual measurements made, but rather the collection of measurements made on a cell. With this philosophy, for a given cell, if a laboratory, say k , participates in measuring the concentration for this cell, its reported measurements $\{y_{kj}, j = 1, \dots, n_k\}$ will be labeled with the same symbol. More specifically, the measurements are labeled as follows:

NA: if one of them is determined to be an outlier, or either B_k^c or P_k^c is smaller than 1.

W: if they are not labeled as **NA**, but either B_k^c or P_k^c is smaller than 5.

A: if they are not labeled as **NA** or as **W**.

Although the cutpoints of 5% for **A** and 1% for **W** seem arbitrary, they are consistent with the uses of 95% evaluation limits for **A** and 99% evaluation limits for **W**. Similar labeling based on z -score was suggested by Lawn, *et. al.* (1997), who classified a z -score with absolute value >3 as unsatisfactory, >2 but ≤ 3 as questionable, and ≤ 2 as satisfactory.

4.6 Guideline for Re-evaluation of Procedures

We will set up a guideline to identify areas where laboratories should re-evaluate their measurement procedures. If at least 5% of a laboratory's measurements are determined to be

outliers, then its measurement procedures should be totally re-evaluated. If a laboratory receives an overall score of bias with a standing in the bottom 5%, the bias of its measurement procedures should be re-evaluated. If it receives an overall score of precision with a standing in the bottom 5%, the precision of its measurement procedures should be re-evaluated as well. This separation of scores of bias and precision allows a laboratory to direct its resources to the particular area that seems to be the weakest. This guideline may be used for any subclass of measurement procedures, such as those for metal analytes, radionuclides, air matrix, water matrix, and soil matrix.

5. Score Distribution and Score Standing

For the scoring procedures proposed in Section 4, we first evaluated a laboratory's performance cell by cell and gave a cell score of bias (and precision if there were replicate measurements for the cell) to it. Then, we assigned an overall score of bias (and an overall score of precision for those cells with replicate measurements) to each laboratory by averaging its corresponding cell scores. In this section we will investigate the statistical distribution of the overall score. The cell scores received by participating laboratories are identical and independent uniformly distributed random variables on $[0,100]$. Hence, the overall score for a laboratory is the average of these uniform random variables. Below we will investigate some simple statistical properties of the average of these uniform random variables.

Let U_1, L, U_v be v identical and independent uniform random variables on $[0,100]$ and

$$\bar{U}_v = \sum_{i=1}^v U_i / v,$$

then

$$E(\bar{U}_v) = 50 \text{ and } \text{Var}(\bar{U}_v) = 2500/3v.$$

Further, the distribution of $\bar{U}_v$ is symmetric with respect to its mean 50 and has the following cumulative distribution function:

$$\sum_{l=0}^{\lfloor vx/100 \rfloor} (-1)^l \binom{v}{l} \frac{(\frac{vx}{100} - l)^v}{v!}, 0 \leq x \leq 100, \quad (15)$$

where $\lfloor x \rfloor$ is the largest integer below x . When v is larger than 4, the distribution of $\bar{U}_v$ is well approximated by a normal distribution with a mean of 50 and a standard deviation of $28.87 / \sqrt{v}$.

Assume that U_1, L, U_v are the cell scores for a laboratory. The distribution of the overall score $\bar{U}_v$ for this laboratory has the mean of 50 and the standard deviation of $28.87 / \sqrt{v}$. However, this standard deviation depends on the number of cells in which this laboratory participated. If we define its score standing as the probability of observed $\bar{U}_v$ in percentage then D is the percentage of laboratories that could receive an overall score not better than $\bar{U}_v$. The interpretation of D is that this laboratory is better than D percent of laboratories that participated in the study. If

two laboratories receive the same score, it does not mean that they perform equally well unless they have participated in the same number of cells. But if they receive the same score standing, it means that they are both on the same level of laboratory performance.

Since the mean of $\bar{U}_v$ is 50, it is expected that most laboratories receive overall scores of about 50. Laboratories receiving such a score may feel disappointed, since traditionally we think a good score should be 80 or higher. The score standing can alleviate this disappointment. For example, assuming a laboratory participated in three cells and received an overall score of 72 (which does not seem to be a very high score), it could appear to be an average laboratory. However, when its overall score of 72 is converted into a score standing of 90.3, then the laboratory is considered to be better than 90.3 percent of all participating laboratories; and it should be very pleased with its performance.

6. Examples

In this section, we will demonstrate our statistical scoring procedures with two environmental data sets. We will use Quality Assessment Program (QAP) 39 data from the U.S. Department of Energy (DOE) Environmental Measurements Laboratory, received via the Analytical Chemistry Laboratory at Argonne National Laboratory, to evaluate lab performance. To assess the proficiency of methods, we used a strontium data set from the Strontium Study Group at Argonne National Laboratory.

6.1 QAP 39 Data Set

The QAP 39 data set consists of 2209 observations on 46 cells (matrix/analyte combination) provided by 99 participating laboratories. These laboratories are coded as 1, 2, ..., 99. (The complete data set is very large and cannot be reproduced here. The reader can obtain the entire data set from DOE or from us). Each laboratory participating in a given cell provided an observation. The 46 cells and their numbers of participating laboratories are listed in Table 1. For each participating laboratory, Table 2 gives the number of cells it participated in (# Cells), the number of outliers (# Cells Out), the number of *not acceptable* cells (# Cells NA), the number of *warning* cells (# Cells W), the number of *acceptable cells* (# Cells A), the number of cells (# Cells) used in scoring, the score of bias, and the laboratory standing. If a laboratory had at least 5% of its observations determined to be outliers, the column # Cells Out is marked by an asterisk. It is surprising that 39 (39%) laboratories had at least 5% of its observations determined to be outliers. Further, it seems that most laboratories' performances were either very good or very poor.

6.2 Strontium Data Set

The strontium data set has two radioactive analytes (Sr-89 and Sr-90). The organizer provided seven spiked samples in water with different levels of Sr-89 and -90 for six laboratories to measure the concentrations. The protocol also required that each laboratory make four measurements of Sr-89 and Sr-90 for each sample using each of three preconditioning methods: calcium carbonate (CaCO_3) precipitation, calcium phosphate (CaPO_4) precipitation, and application of a cation exchange resin. The whole data set is provided in the appendix. For this data set, a cell is defined to be a combination of an analyte, a sample, and a preconditioning method.

Table 3 gives the statistical scoring results for each participating laboratory. The results include the number of cells it participated in (# Cells), the number of total observations (# Obs), the number of outliers (# Out), the number of *not acceptable* cells (# NA), the number of *warning* cells (# W), the number of *acceptable cells* (# A), the number of cells used in scoring (Cells) of bias, score of bias and its standing, the number of cells used in scoring (Cells) of

precision, score of precision and its standing are listed in Table 3. If a laboratory had at least 5% of its observations determined to be outliers, the column # Out is marked by a '*'. For example, for Laboratory 2, its # Out is marked by a '*' and therefore as we recommended in Section 4, this laboratory should be alerted about this problem. Note that for Laboratory 5, although only three of its measurements were determined to be outliers, it received very low scores of bias and precision. The scoring procedures were also carried out to the sub data set containing only Sr-89 or -90 in Tables 4 and 5 respectively. From Tables 3-5, the following conclusions may be drawn:

- Laboratories 1 and 6 performed very well.
- Laboratory 3's performance for bias in its measurement procedure of Sr-89 was marginally acceptable and needed future monitoring for this bias.
- Laboratory 4 needed investigation of precision for its measurement procedure of Sr-89, and Laboratory 2 needed investigation (especially for bias) for its measurement procedure of Sr-89.
- Laboratory 5 needed a complete investigation for its measurement procedures of both analytes. Because of its poor performance, data from Laboratory 5 were not used in the subsequent analysis of the comparison of the three preconditioning methods.

7. Summary

Two statistical scoring procedures based on p-values are proposed to evaluate the overall performance of analytical laboratories making environmental measurements. The overall score of bias and its standing evaluate how well a laboratory can measure the true (unknown) concentration correctly in the long run. The overall score of precision and its standing evaluate how well a laboratory can reproduce its measurements in the long run. Criteria have been developed for classifying the reported concentrations as Acceptable (A), Warning (W), and Not Acceptable (NA) and for identifying areas where laboratories should re-evaluate their measurement procedures.

Scores of bias and scores of precision do not penalize outliers. Once measurements are identified as outliers, they are excluded from the computation of score of bias or score of precision. Therefore, scores of bias and scores of precision should be used in conjunction with the number of outliers identified. (See Sections 4.5 and 4.6.)

The meaning of the two combination scores proposed by Lawn, et. al. (1997) is not straightforward. To interpret the RSZ score properly, one must know the standard normal distribution. Unless someone knows the degree of freedom and has a χ^2 distribution function table, it is almost impossible to interpret the SSZ score. The interpretation of the two scoring procedures proposed in this paper is obvious since the standing is the probability that the given laboratory is better than other laboratories.

Although our methods are designed for use with laboratories that measure and collect environmental data, the methods could be used for laboratories in other fields as well. As an example, they could be used to evaluate and monitor medical laboratories that measure blood levels of drugs.

We are investigating the use of robust estimation of location (Andrews, et al., 1972; Stigler, 1977; Huber, 1981; Hampel, et al., 1986) and robust estimation of variance components (Rocke, 1983, 1991, and 1992; Fellner, 1986; Stahel and Welsh, 1997) for the estimation of μ , σ_1^2 , and σ_2^2 . From these results the approximate distributions are determined for the corresponding $T_k^{C_1}$ s and $F_k^{C_1}$ s.

Acknowledgments

Many colleagues helped while we were preparing this article. In particular, we would like to acknowledge the following individuals: A.E. Scandora, Argonne National Laboratory (ANL), for his assistance with data transfers and manipulations and for his many insightful comments; Catherine Klusek and Colin Sanderson (US DOE Environmental Measurements Laboratory) for facilitating access to the QAP data sets; and finally, Mitch Erickson, Mary Ann Edgell, Lesa Smith, and Peter Lindahl (ANL) for allowing us to test our methods on the strontium data set.

Appendix: The Strontium Data Set

The entire data set for the strontium study is listed in Tables 6-11, where blanks represent missing data. Tables 6-8 list the measurements for Sr-89, while Tables 9-11 list the measurements for Sr-90. Six laboratories (Lab) made four measurements (Seq) with seven samples and three pretreatments.

Table 6
Sr-89: Pretreatment Method: Cation Exchange

		Sample						
LabSeq		1	2	3	4	5	6	7
1	1	1.4700	0.3400	0.8700	1.2000	1.1700	0.7600	1.1100
1	2	1.2900	0.2920	1.0300	1.2100	1.0300	0.6900	1.1400
1	3	1.5000	0.2400	0.8200	1.2000	0.9500	0.6800	1.2100
1	4	1.4200	0.2150	0.7200	1.0300	1.0600	0.6600	1.0600
2	1	1.9700	0.5200	22.2000	1.3100	2.7500	2.7700	2.4100
2	2	1.8800	2.7100	13.0800	1.3300	2.9100	2.4100	2.2900
2	3	1.8000	0.6200	2.6900	1.4200	2.2100	1.2600	2.4400
2	4	1.8400	1.9800	5.9800	1.2400	3.4500	6.2700	4.5100
3	1	1.3300	0.3300	1.0200	0.6700	0.9500	0.3100	0.5500
3	2	0.9400	0.4800	1.0900	0.6900	0.1600	0.5600	1.2700
3	3	1.8200	0.1600	0.9700	0.5400	0.5000	0.6300	0.7500
3	4	0.5200	0.1800	0.5300	0.8100		0.4700	1.3300
4	1	1.7290	0.6162	1.3460	1.9810	4.4000	1.3770	
4	2	1.6100	0.6572	1.6770	1.3970	4.5290	1.3240	1.2130
4	3	1.8500	0.3789	0.5931	1.1512		1.0740	1.0030
4	4	1.2400	1.3280	0.9356	1.4720	1.1120	2.8830	0.4161
5	1	0.9100			1.4900			
5	2	1.8200		0.2400	2.1400			0.5500
5	3	2.0000			1.8900			0.3900
5	4	1.7000		0.3700	1.8900			
6	1	1.4590	0.3343	0.7818	1.3110	1.1640	0.7421	1.4470
6	2	0.8874	0.5411	0.9050	1.3070	0.8100	0.3717	1.0560
6	3	1.3140	0.3212	0.9185	1.3810	0.8764	0.7062	1.1870
6	4	1.3860	0.2883	0.7816	1.2910	1.1310	0.6807	1.2510

Table 7
Sr-89: Pretreatment Method: Calcium Carbonate Precipitation

		Sample						
LabSeq		1	2	3	4	5	6	7
1	1	1.4100	0.1100	0.8700	1.2100	0.8800	0.6200	1.1100
1	2	1.4100	0.0190	0.5800	1.1300	1.0200	1.1800	0.9500
1	3	1.9100	0.8200	1.1000	1.9900	1.4600	1.2900	1.3200
1	4	1.3500	0.2700	0.9400	1.3300	1.1400	0.8200	1.0400
2	1	1.8000	0.8000	2.7800	1.0800	1.9200	0.6400	3.2900
2	2	2.0400	0.8800	2.4300	1.2300	1.8600	1.8300	1.5900
2	3	1.7200	0.6100	4.5100	1.3300	2.0200	2.6900	2.0500
2	4	2.6000	0.8900	2.8200	1.1900	3.4400	5.3300	1.5300
3	1	0.4500			0.6700		0.5500	
3	2	0.6200	0.0600		0.7400			0.7100
3	3	0.7400	0.3100	1.4500	0.6900	0.8400	0.9400	1.1500
3	4	0.7800	0.4400	0.7600	0.6600	0.9300	0.5100	0.7200
4	1	1.5070	0.3752	0.6842	1.6130	4.1510	0.0567	0.9858
4	2	1.4410	0.0117	1.8020	1.6100	1.0620		0.8955
4	3	1.3170	0.8207	0.1877	1.4780	0.8816		1.5400
4	4	1.3890		1.2760	1.4550	1.1080	4.8810	0.5575
5	1			0.6800	0.7400			
5	2	1.4200	0.2100	0.7900	1.3400	0.7500		0.4100
5	3				0.4900			
5	4	0.3100			0.9200			0.1900

LabSeq	Sample						
	1	2	3	4	5	6	7
6 1	1.6240	0.4007	0.7770	1.3180	1.1920	1.0050	1.3110
6 2	1.4780	0.3955	0.8079	1.2790	1.1490	0.8818	1.1410
6 3	1.3190	0.3129	0.8305	1.4110	1.1580	0.7516	1.3240
6 4	1.4580	0.3050	0.8314	1.3510	1.3010	0.6830	1.4150

Table 8
Sr-89: Pretreatment Method: Calcium Phosphate Precipitation

LabSeq	Sample						
	1	2	3	4	5	6	7
1 1	1.3400	0.3400	0.6700	0.9900	0.9200	0.6500	0.9700
1 2	1.2800	0.2600	0.9100	1.1200	1.0200	0.4900	1.3300
1 3	0.9600	0.4310	1.0000	1.0000	1.0100	0.8900	2.0000
1 4	1.5200	0.3440	0.7800	0.9700	0.8900	0.6500	1.1800
2 1	3.7600		2.4800	1.1700	1.9100	1.4500	1.2300
2 2	1.3600	0.9400	2.1100	1.2000	0.4000	1.2800	1.2200
2 3	1.9800	0.5800	1.7700	1.1500	1.5200	1.6300	2.0300
2 4	2.1000	0.7700	2.3400	1.3500	4.8400	1.2200	2.0200
3 1	1.1100	0.1400		0.7300	1.3400	1.2700	1.0000
3 2	0.8800	0.2200	0.3000	0.8500	0.3800	0.3100	0.8600
3 3	0.6400	0.1600	0.7000	0.7100	0.3200	0.7500	
3 4	0.9400	0.1800	0.2200	0.6800	0.3300	0.9300	0.7500
4 1		0.2819	0.9679	1.5690	0.2489	0.4485	0.2868
4 2	1.3170	0.0845	0.2178	1.6670	0.2647		
4 3	1.5010	0.4524	0.5789	1.6250		0.9957	
4 4	1.7980	0.3968		1.3480	1.4820	1.7020	1.0700
5 1	2.4600	0.5300	2.5700	1.6100	1.6800	0.6800	2.1300
5 2	3.4200	0.6200	1.3800	1.6700	1.6100	1.4600	2.8100
5 3	2.7600	0.6100	1.9900	1.1900		0.1700	1.3500
5 4	2.7400	0.7200	1.8800	1.8600	0.5400	0.2800	1.4500
6 1	1.3730	0.4521	0.8393	1.2110	1.2930	0.5303	1.2930
6 2	1.3790	0.3593	0.8700	1.2760	1.0470	0.7085	1.3670
6 3	0.9563	0.3365	0.7746	1.3230	1.1110	0.6914	1.1770
6 4	1.1990	0.3131	0.8194	1.3050	1.1306	0.8237	1.1100

Table 9
Sr-90: Pretreatment Method: Cation Exchange

LabSeq	Sample						
	1	2	3	4	5	6	7
1 1	2.0800	1.5300	6.0000	0.2400	5.1000	4.1600	3.9100
1 2	2.0200	1.6740	6.2300	0.2290	5.6500	4.6500	3.8000
1 3	2.1800	1.5700	6.1200	0.2720	5.3700	4.6800	3.9300
1 4	2.1600	1.6100	6.0600	0.2220	5.5400	4.5700	3.8600
2 1	2.0900	1.5800	5.9700	0.2200	5.3700	9.2500	4.0100
2 2	2.1700	0.7600	2.3400	0.2200	5.1600	4.8200	3.9500
2 3	2.1400	1.5600	5.8800	0.2100	5.8400	5.1400	4.2100
2 4	2.2100	1.1500	5.2000	0.2000	4.8600	3.2900	2.9900
3 1	1.4900	1.3100	4.8700	0.5000	3.6800	3.7300	2.6400
3 2	1.6100	1.1400	4.4800	0.1200	4.2800	3.6700	2.6200
3 3	1.1700	1.7000	4.9500	0.1900	4.5000	3.9300	3.3700
3 4	1.7400	1.5100	4.9600	0.1000	4.8800	3.9700	2.9200
4 1	2.2100	1.6370	6.1210	0.1897	5.0580	4.5310	
4 2	2.1770	1.5880	6.2540	0.2249	5.3660	4.7890	4.2110
4 3	2.3230	1.7050	6.4950	0.2164	6.1360	5.0310	4.1050
4 4	2.3840	1.2640	6.8250	0.2224	5.4880	4.2000	4.4770
5 1	5.8800	4.1500	14.4000	0.9000	12.8000	11.2000	9.5300
5 2	5.2200	3.7500	13.4000	0.4900	11.8000	10.5000	8.3900
5 3	5.2400	3.8700	13.4000	0.6700	12.2000	10.2000	8.9400
5 4	5.9800	4.1100	13.5000	0.6900	11.9000	10.7000	9.1200
6 1	1.5490	1.2380	5.1980	0.1996	3.6220	4.3640	3.4980
6 2	1.4980	1.2520	5.1720	0.1723	4.4900	4.0500	3.2830
6 3	1.6420	1.2270	5.0430	0.1734	4.3090	3.6640	3.0800
6 4	1.6210	1.2810	5.1750	0.1918	4.5060	4.0130	3.2030

Table 10
Sr-90: Pretreatment Method: Calcium Carbonate Precipitation

LabSeq	Sample						
	1	2	3	4	5	6	7
1 1	2.2400	1.6300	6.1000	0.2250	5.4400	4.6100	3.8300
1 2	2.1500	1.5600	6.0400	0.2800	5.3700	4.5500	3.7500
1 3	2.2600	1.5300	6.0100	0.2250	5.2200	4.7300	3.8500
1 4	2.1600	1.5200	5.9800	0.2310	5.2700	4.7400	3.9000
2 1	2.2100	1.5500	6.4600	0.2400	5.6100	5.3400	3.5100
2 2	1.9800	1.5500	5.9400	0.2100	5.5000	4.9500	4.1400
2 3	2.2300	1.5100	6.4800	0.2200	4.5200	4.7300	3.6600
2 4	2.1200	1.5200	5.8800	0.2400	5.2300	3.9000	4.1900
3 1	1.9400	1.5000	6.2500	0.0900	5.3200	3.9000	3.7500
3 2	1.8300	1.4600	5.5100	0.1300	6.3200	5.2800	3.2000
3 3	1.7100	1.2200	3.6700	0.1100	3.9600	3.7700	2.9900
3 4	1.7500	1.1200	4.8000	0.0900	4.2100	4.1000	3.0800
4 1	2.4180	1.6550	6.4080	0.2359	5.6470	5.0810	4.0180
4 2	2.3970	1.7650	6.4200	0.2111	5.5650	5.5440	4.3470
4 3	2.2710	1.4590	6.1810	0.1776	5.4480	4.8140	3.5820
4 4	2.2390	1.6900	6.5070	0.2415	5.4110	3.6190	3.9640
5 1	4.1400	3.1100	11.4000	0.6400	10.9000	9.8000	7.8600
5 2	5.2300	3.5400	13.5000	0.5000	12.1000	10.8000	8.7800
5 3	4.4000	3.4400	11.8000	0.5700	10.1000	8.9000	7.0200
5 4	4.4900	3.3700	11.9000	0.4100	10.3000	8.9100	6.9700
6 1	1.6340	1.2450	5.0120	0.1873	4.3450	4.0860	3.3480
6 2	1.8940	1.2730	5.0270	0.1701	4.5650	4.1060	3.2640
6 3	1.5880	1.2590	5.3390	0.1393	4.5230	4.1890	3.3310
6 4	1.7390	1.2210	5.0750	0.1875	4.9400	4.3270	3.2270

Table 11
Sr-90: Pretreatment Method: Calcium Phosphate Precipitation

LabSeq	Sample						
	1	2	3	4	5	6	7
1 1	2.0800	1.6100	5.8100	0.1940	5.1500	4.6900	3.8100
1 2	2.2900	1.5200	5.9800	0.2040	5.3900	4.5200	3.8800
1 3	2.2500	1.4800	6.0900	0.2000	5.3700	4.1200	1.4100
1 4	1.9700	1.6600	6.1300	0.1590	5.3300	4.5900	3.5500
2 1	1.6000	2.0200	5.2500	0.2100	5.3300	4.7600	3.9100
2 2	2.1100	1.1800	5.4600	0.2100	5.6700	4.6400	3.9300
2 3	2.3300	1.5300	6.1100	0.2300	5.6900	5.0000	3.8900
2 4	2.1900	1.5500	6.1800	0.2400	4.7300	4.8800	3.8600
3 1	1.5300	1.4100	5.4100	0.1500	4.0900	3.1900	3.0500
3 2	1.8300	1.2300	5.0000	0.1500	4.5900	4.2500	3.1300
3 3	1.7500	1.2600	5.1500	0.1100	4.7400	3.5500	3.9300
3 4	1.6900	1.2400	5.0900	0.1300	4.7300	3.8400	3.2800
4 1	3.7320	1.7840	6.3540	0.2841	6.2050	5.1610	4.0600
4 2	2.3480	1.8250	7.0270	0.2317	5.9640	5.2180	4.4230
4 3	2.5350	1.6000	6.4060	0.2248	6.2610	5.1040	5.0900
4 4	2.3080	1.8040	7.0610	0.2424	5.9970	4.5060	4.4250
5 1	4.7700	2.9900	12.2100	0.4800	10.7000	9.3200	7.3200
5 2	4.0600	3.1100	11.7700	0.4600	10.3000	8.7700	6.9800
5 3	3.9500	2.7400	10.9000	0.7200	11.2000	10.0000	7.4200
5 4	4.3000	2.9000	11.4000	0.2200	10.3000	9.5000	7.1200
6 1	1.5730	1.3130	4.9130	0.1693	4.3750	2.2030	3.2710
6 2	1.6500	1.2930	5.3080	0.1666	4.4460	3.4960	3.1360
6 3	1.4160	1.2610	5.0530	0.1922	4.4950	4.2910	3.2110
6 4	1.4990	1.2890	5.1560	0.1486	4.5860	4.7560	3.1530

References

1. American Society for Testing and Materials. Annual Book of ASTM Standards. Vol. 14.02. Philadelphia: 1995a. E548-94: Standard Guide for General Criteria Used for Evaluating Laboratory Competence; p. 311-317.
2. American Society for Testing and Materials. Annual Book of ASTM Standards. Vol. 14.02. Philadelphia.: 1995b. E691-92: Standard Practice for Conducting an Interlaboratory Study to Determine the Precision of a Test Method; p. 426-445.

3. American Society for Testing and Materials. Annual Book of ASTM Standards. Vol. 14.02. Philadelphia: 1995c. E994-95: Standard Practice for Calibration and Testing Laboratory Accreditation Systems: General Requirements for Operation and Recognition; p. 656-660.
4. American Society for Testing and Materials. Annual Book of ASTM Standards. Vol. 14.02. Philadelphia: 1995d. E1187-90: Standard Terminology Related to Laboratory Accreditation; p. 738-739.
5. American Society for Testing and Materials. Annual Book of ASTM Standards. Vol. 14.02. Philadelphia: 1995e. E1301-89: Standard Guide for Development and Operation of Laboratory Proficiency Testing Programs; p. 819-824.
6. Andrews, DF.; Bickel, PJ.; Hampel, FR.; Huber, PJ.; Rogers, WH.; Tukey, JW. Robust Estimates of Location: Survey and Advances. Princeton University Press; Princeton, New Jersey: 1972.
7. Barnett, V.; Lewis, T. Outliers in Statistical Data. John Wiley & Sons; New York: 1978.
8. Britton, PW.; Lewis, DF. Statistical Basis for Laboratory Performance Evaluation Limits. U.S. EPA Quality Assurance Branch, Environmental Monitoring and Support Laboratory; Cincinnati: 1986.
9. Dilbeck, G.; Honsa, P. Environmental Radioactivity Laboratory Intercomparison Studies Program and Radioactive Standard Distribution Program. Nuclear Radiation Assessment Division, Environmental Monitoring Systems Laboratory; Las Vegas: 1989.
10. Fellner WH. Robust Estimation of Variance Components. *Technometrics* 1986;28:51–60.
11. Gilbert, RO. Statistical Methods for Environmental Pollution Monitoring. Van Nostrand Reinhold; New York: 1987.
12. Graybill, FA. Theory and Application of the Linear Model. Duxbury Press; North Scituate: 1976.
13. Hampel, FR.; Ronchetti, EM.; Rousseeuw, PJ.; Stahel, WA. Robust Statistics. John Wiley & Sons; New York: 1986.
14. Huber, P. Robust Statistics. John Wiley & Sons; New York: 1981.
15. Kafadar K. A Biweight Approach to the One-Sample Problem. *JASA* 1982;77:416–424.
16. Kirkpatrick, D.; Horlick, J. Proficiency Testing: An Essential Element of Laboratory Performance Evaluation and Accreditation. In: Schock, Harvey, editor. Evaluation and Accreditation of Inspection and Test Activities, ASTM STP 814. American Society for Testing and Materials; Philadelphia: 1983. p. 128-140.
17. Kristiansen S. A Statistical Procedure for the Estimation of Bias Parameters in Interlaboratory Studies. *Statistics in Medicine* 1991;10:843–854. [PubMed: 1876777]
18. Lawn, R.; Thompson, M.; Walker, RF. Proficiency Testing in Analytical Chemistry. Royal Society of Chemistry; Cambridge, UK: 1997.
19. Mandel, J. Evaluation and Control of Measurements. Marcel Dekker; New York: 1991.
20. Mandel, J. Analysis of Two-way Layouts. Chapman & Hall; New York: 1994.
21. Mandel J. Repeatability and Reproducibility for Pass/Fail Data. *Journal of Testing and Evaluation* 1997;25:151–153.
22. Quality Assurance Branch. Environmental Radioactivity Laboratory Intercomparison Studies Program 1978-1979. U.S. Environmental Protection Agency, Office of Research and Development, Environmental Monitoring and Support Laboratory; 1978. EPA-60/4-78-032
23. Rocke DM. Robust Statistical Analysis of Interlaboratory Studies. *Biometrika* 1983;70:421–431.
24. Rocke DM. Robustness and Balance in the Mixed Model. *Biometrics* 1991;47:303–309. [PubMed: 2049504]
25. Rocke DM. Estimation of Variation after Outlier Rejection. *Computational Statistics and Data Analysis* 1992;13:9–20.
26. Scott EM, Baxter MS, Aitchison TC, Harkness DD, Cook GT. An Overview of Some Interlaboratory Studies. *Radiocarbon* 1990;32:259–265.
27. Stahel WA, Welsh A. Approaches to Robust Estimation in the Simplest Variance Components Model. *Journal of Statistical Planning and Inference* 1997;57:295–319.
28. Stigler SM. Do Robust Estimators Work with Real Data? *Annals of Statistics* 1977;5:1055–1078.
29. Twomey, A. Interlaboratory Comparisons as Used in the Accreditation of Laboratories in New Zealand's Dairy Industry. In: Schock, Harvey, editor. Evaluation and Accreditation of Inspection and

- Test Activities, ASTM STP 814. American Society for Testing and Materials; Philadelphia: 1983. p. 153-161.
30. Youden, WJ.; Steiner, EH. Statistical Manual of the Association of Official Analytical Chemists. Association of Official Analytical Chemists; Washington, D.C.: 1975.

Table 1
Forty-six Cells and Their Number of Participating Laboratories in QAP39 Data Set

Matrix	Analyte	# of Labs
Air	Am - 241	34
Air	Ce - 144	65
Air	Co - 57	66
Air	Co - 60	69
Air	Cs - 134	68
Air	Cs - 137	71
Air	Mn - 54	66
Air	Pu - 238	48
Air	Pu - 239	49
Air	Sb - 125	55
Air	Sr - 90	40
Air	U (ug)	26
Air	U (Bq)	11
Air	U - 234	26
Air	U - 238	26
Soil	Am - 241	34
Soil	Cs - 137	78
Soil	K - 40	64
Soil	Pu - 239	54
Soil	Sr - 90	38
Soil	U (ug)	25
Soil	U (Bq)	9
Soil	U - 234	32
Soil	U - 238	36
Vegetation	Am - 241	27
Vegetation	Co - 60	56
Vegetation	Cs - 137	68
Vegetation	K - 40	62
Vegetation	Pu - 238	35
Vegetation	Pu - 239	38
Vegetation	Sr - 90	36
Water	Am - 241	50
Water	Ce - 144	78
Water	Co - 60	81
Water	Cs - 134	80
Water	Cs - 137	84
Water	Fe - 55	9
Water	H - 3	61
Water	Mn - 54	79
Water	Pu - 238	58
Water	Pu - 239	60
Water	Sr - 90	49
Water	U (ug)	33
Water	U (Bq)	10
Water	U - 234	32
Water	U - 238	33

Table 2

Statistical Scoring Results for the QAP 39 Data set

Lab#	Cells A										# Cells	Bias		Standing
	Cells#	Out#	Cells NA#	Cells W#	Cells A							Score	Score	
1	46	0	0	0	46						46	63.5	99.9	
2	34	1	2	2	30						33	47.9	33.9	
3	19	0	0	0	19						19	62.8	97.4	
4	36	2*	3	2	31						34	44.1	11.7	
5	17	7*	7	3	7						10	29.3	1.2	
6	39	2*	2	0	37						37	69.1	100	
7	14	0	0	0	14						14	52.0	60.3	
8	42	1	1	0	41						41	56.4	92.2	
9	42	1	1	1	40						41	53.1	75.2	
10	23	8*	10	2	11						15	24.4	0.0	
11	36	1	3	3	30						35	39.7	1.8	
12	13	3*	4	0	9						10	65.0	95.0	
13	19	0	0	3	16						19	57.6	87.4	
14	11	0	0	0	11						11	74.7	99.8	
15	37	2*	4	0	33						35	45.3	16.8	
16	42	5*	8	0	34						37	48.0	33.6	
17	31	5*	5	2	24						26	35.3	0.5	
18	18	1*	1	0	17						17	59.3	90.7	
19	16	0	0	0	16						16	62.9	96.3	
20	20	1*	2	0	18						19	58.8	90.9	
21	15	0	1	0	14						15	34.0	1.6	
22	39	1	1	1	37						38	70.6	100	
23	14	0	0	0	13						14	37.9	5.8	
24	19	0	0	0	19						19	61.8	96.2	
25	17	9*	10	2	5						8	24.3	0.6	
26	23	0	1	2	20						23	41.4	7.7	
27	22	4*	6	1	15						18	50.2	51.4	
28	1	1*	1	0	0						.	.	.	
29	21	2*	2	0	19						19	65.0	98.8	
30	38	4*	6	1	31						34	54.3	80.7	
31	36	1	1	0	35						35	65.2	99.9	
32	17	0	0	0	17						17	66.5	99.1	
33	38	1	2	2	34						37	62.3	99.5	
34	24	0	0	0	24						24	63.8	99.1	
35	31	1	2	0	29						30	63.9	99.6	
36	36	0	2	3	31						36	44.2	11.2	
37	23	2*	4	0	19						21	52.4	65.0	
38	19	1*	1	0	18						18	56.9	84.6	
39	31	0	0	0	31						31	68.2	100	
40	38	0	1	1	36						38	50.0	49.7	
41	16	0	1	1	14						16	45.3	25.6	
42	1	0	0	0	1						1	38.7	38.7	
43	21	7*	7	1	13						14	49.6	48.1	
44	14	3*	3	0	11						11	52.3	60.6	
45	11	4*	4	0	7						7	38.5	14.6	
46	37	0	0	0	37						37	64.0	99.8	
47	32	0	0	0	32						32	66.5	99.9	
48	20	0	0	0	19						20	61.1	95.8	
49	2	0	0	0	2						2	42.7	36.4	
50	35	5*	5	1	29						30	66.3	99.9	

Bias										Score	Standing	
# Cells												
Lab#	Cells#	Cells Out#	Cells NA#	Cells W#	Cells A							
51	41	0	0	1	40	41				62.3	99.7	
52	14	1*	1	0	13	13				61.8	93.1	
53	28	11*	14	1	13	17				43.5	17.8	
54	42	0	0	0	42	42				67.0	100	
55	27	0	0	0	27	27				68.0	99.9	
56	12	0	1	1	10	12				36.8	5.6	
57	11	0	0	1	10	11				50.9	54.3	
58	6	0	0	0	6	6				43.9	30.3	
59	10	1*	1	1	8	9				32.3	3.3	
60	14	0	0	0	14	14				70.1	99.5	
61	8	0	0	0	8	8				56.7	74.5	
62	13	1*	1	0	12	12				66.8	97.8	
63	7	0	0	0	7	7				53.6	62.8	
64	5	0	0	0	5	5				57.7	72.5	
65	34	2*	2	0	32	32				62.6	99.3	
66	17	12*	12	3	2	5				10.7	0.1	
67	16	0	0	0	16	16				65.7	98.5	
68	35	0	0	1	34	35				64.4	99.8	
69	14	2*	5	2	7	12				9.1	0.0	
70	11	1*	1	0	10	10				41.4	17.3	
71	8	2*	3	0	5	6				49.6	48.5	
72	8	0	0	2	6	8				48.9	45.6	
73	1	0	0	0	1	1				86.6	86.6	
74	23	2*	4	0	19	21				44.5	19.2	
75	32	2*	5	4	23	30				47.7	33.0	
76	19	0	0	1	18	19				63.5	97.9	
77	5	0	0	0	5	5				32.7	9.0	
78	25	0	5	3	17	25				37.9	1.8	
79	12	0	0	0	12	12				67.1	98.0	
80	20	0	0	1	19	20				60.9	95.4	
81	21	0	1	2	18	21				33.3	0.4	
82	17	0	0	0	17	17				60.5	93.2	
83	29	1	1	0	28	28				73.9	100	
84	4	0	1	0	3	4				52.2	55.9	
85	36	0	0	1	35	36				62.6	99.6	
86	34	1	3	2	29	33				47.4	30.1	
87	40	1	4	5	31	39				41.6	3.5	
88	34	0	1	2	31	34				62.2	99.3	
89	34	3*	4	2	28	31				42.0	6.3	
90	14	4*	4	0	10	10				49.2	46.4	
91	18	1*	1	0	17	17				64.2	97.9	
92	15	1*	1	3	11	14				22.7	0.0	
93	42	0	3	0	39	42				60.0	98.8	
94	11	0*	0	0	11	11				43.6	23.2	
95	30	2*	2	1	27	28				62.9	99.1	
96	30	1*	1	0	29	29				62.4	99.0	
97	3	1*	1	0	2	2				90.4	98.2	
98	30	8*	12	5	13	22				27.1	0.0	
99	3	0	0	0	3	3				49.7	49.4	

* At least 5% of observations were determined to be outliers.

Table 3

Statistical Scoring Results for the Strontium Data Set

#	Lab#	Cells#	Obs#	Out#	NA#	W#	A	Bias			Precision		
								Cells	Score	Standing	Cells	Score	Standing
1	42	168	0	1	0	41		42	74.5	100	42	84.2	100
2	42	167	10*	12	2	28		41	53.6	78.5	41	51.4	61.8
3	42	158	0	2	0	40		42	46.9	24.5	42	58.5	97.2
4	42	156	1	7	0	35		42	68.0	100	42	45.3	14.6
5	38	135	3	22	1	15		38	16.2	0.0	35	29.4	0.0
6	42	168	0	1	0	41		42	62.6	99.8	42	89.4	100

* At least 5% of observations were determined to be outliers.

Table 4
Statistical Scoring Results of Analyte Sr-89 for the Strontium Data Set

										Bias			Precision		
Lab#	Cells#	Obs#	Out#	NA#	W#	A	Cells	Score	Standing	Cells	Score	Standing	Score	Standing	Standing
1	21	84	0	0	0	21	21	69.8	99.9	21	80.4	100	80.4	100	100
2	21	83	10*	8	2	11	20	29.3	0.1	20	49.0	43.8	49.0	43.8	43.8
3	21	74	0	0	0	21	21	40.2	5.9	21	65.0	99.2	65.0	99.2	99.2
4	21	73	1	6	0	15	21	52.8	67.3	21	24.3	0.0	24.3	0.0	0.0
5	17	51	0	2	0	15	17	32.4	0.6	15	31.8	0.7	31.8	0.7	0.7
6	21	84	0	0	0	21	21	70.1	99.9	21	92.0	100	92.0	100	100

* At least 5% of observations were determined to be outliers.

Table 5
Statistical Scoring Results of Analyte Sr-90 for the Strontium Data Set

										Bias			Precision		
Lab#	Cells#	Obs#	Out#	NA#	W#	A	Cells	Score	Standing	Cells	Score	Standing	Cells	Score	Standing
1	21	84	0	1	0	20	21	79.2	100	21	88.0	100	21	88.0	100
2	21	84	0	4	0	17	21	76.7	100	21	53.6	71.7	21	53.6	71.7
3	21	84	0	2	0	19	21	53.7	72.2	21	52.0	62.6	21	52.0	62.6
4	21	83	0	1	0	20	21	83.2	100	21	66.3	99.5	21	66.3	99.5
5	21	84	3	20	1	0	21	3.1	0.0	20	27.6	0.0	20	27.6	0.0
6	21	84	0	1	0	20	21	55.1	79.2	21	86.7	100	21	86.7	100