Statistics for Interlaboratory Comparisons

IAAC PT Workshop
São Paulo, Brazil
20 September, 2004

Dan Tholen, MS
Documents for Lecture

- ISO FDIS 13528 “Statistical Methods for use in Proficiency Testing by Interlaboratory Comparisons
- APLAC Statistical procedures
- ISO 5725 1-6, by reference in ISO 13528
Statistical methods for use in proficiency testing by interlaboratory comparisons

Méthodes statistiques utilisées dans les essais d'aptitude par comparaisons interlaboratoires

ICS 03.120.30
ISO FDIS 13528

- Originally approved as work item in 1997
- Cancelled as work item in 2001
- Re-approved in 2002
- Final comments resolved, 2002 & 2003
- Approved as FDIS, June 2003 after resolving 149 comments, to be published late 2004
ISO 13528

- Companion to ISO Guide 43-1
- Authored by TC-69, SC6, WG1
- Written as a Standard
- High interest / many comments
- Goal is to describe optimal procedures, but to allow other procedures as long as they are:
  - Statistically valid
  - Fully described to users
ISO 13528

Follows Template of Guide 43-1 Annex A

- Determine Assigned Value
  - Robust Statistics
  - Reference Values

- Calculate Performance Statistic
  - Z scores
  - Other performance scores

- Evaluate Performance
  - Relative
  - Fixed
SD for proficiency testing

- Discussed in detail in section 6
- SD as used in z scores
- Can also be thought of as 1/3 of evaluation interval
  (when \( z > 3 \) is action signal)

For example if fixed interval is \( E = \pm 10\% \)...
Then \( E = 3\sigma \)
\( \sigma = E/3 = 10\%/3 = 3.3\% \)
Limiting the uncertainty of the Assigned Value (X): Section 4.2

- Establish limits for uncertainty of AV
  \[ u(X) < 0.3\sigma \]

When using fixed limits…

\[ u(X) < 0.3(E/3) \]
\[ u(X) < E/10 \]
Limiting the uncertainty of the Assigned Value (X): Section 4.2

- If this cannot be met then
  - Look for a better way to determine AV
  - Incorporate uncertainty in score
  - Warn participants
Limiting the uncertainty of the Assigned Value (X): Example

When consensus mean and SD are used to determine performance:
Then \( u(X) = \frac{SD}{\sqrt{n}} \)
So one can have very small uncertainty with large number of labs.
What \( n \) is needed to assure criterion is met?
Need \( \frac{1}{\sqrt{n}} < 0.3 \), or \( n > (1/0.3)^2 \) or \( n > 11 \)
and \( n \leq 11 \), then cannot meet criterion.
Determine the appropriate number of Replicates: Section 4.3

- When a method’s repeatability is large, it can confuse interpretation of scores
- Determine $n$ replicates so that:

$$\frac{\sigma_r}{\sqrt{n}} < 0.3\sigma$$

- All labs must perform the same number of replicates, even if $\sigma_r$ is different.
Limiting the effect of repeatability

Example

Say $E=10\%$ and $\sigma_r = 2\%$:

Then $\sigma_r/\sqrt{n} < 0.1E$

So $2/\sqrt{n} < 0.1(10\%)$ or

$2/1.0 < \sqrt{n}$

Or $n > 4$

This criterion can lead to large $n$ replicates.
Reporting considerations: 4.6

- Reporting:
  - determine typical repeatability $\sigma_r$
  - Do not round digits by more than $\sigma_r/2$

- NO TRUNCATED RESULTS
  - “Less than” values not allowed
  - Results to be reported as determined
  - Conflict with need to report same as to clients
  - Can have check-off box to note “<MDL”.
ISO 13528 Procedures

Calculate Summary Statistics:
- Outlier detection and removal are allowed if done in a statistically valid way
- Robust measures are preferred
  - Mean
  - SD
- Preferred robust method is given, others are allowed if
  - Statistically valid
  - Fully described to participants
ISO 13528 Procedures

Determine Assigned Value (Target)*

- Determined before PT shipment
  - Result from formulation
  - Certified reference value
  - Other reference values

- Determined from PT data
  - Consensus of experts
  - Consensus of participants

* Control the uncertainty of the assigned value
Determine the Standard Uncertainty of the Assigned Value

**Determined** before PT shipment

- Result from formulation
  - Uncertainty per manufacture process, usually very small relative to measurement uncertainty

- Certified reference value
  - Uncertainty provided with certificate

- Other reference values
  - Uncertainty calculated per GUM or other procedure
Determine the Standard Uncertainty of the Assigned Value

**Determined** from data in PT shipment

- Consensus of expert laboratories (p of them)
  - Each lab should know their MU, and report it
  - \( u_x = 1.23(\sqrt{\Sigma u_i^2})/p \) for robust mean (median)
  - Caution about bias in experts

- Consensus of participants (p of them)
  - Calculate robust mean and SD (s*)
  - \( u_x = 1.23(s^*)/\sqrt{p} \)
  - Caution about bias due to method mix
  - Caution about lack of consensus
Robust Analysis

- Algorithm A for mean and SD
  Starts with $x^* = \text{median}$
  \[ s^* = 1.483 \times \text{median} |x_i - x^*| \]

  Limit data at $x^* + 1.5s^*$ and $x^* - 1.5s^*$
  Extreme values trimmed to $1.5s^*$

- New option added to use initial $x^*$ and $s^*$
Robust Analysis

- Algorithm A for mean and SD
  Calculate new: $x^* = (\Sigma x_i)/p$
  $s^* = 1.134 \sqrt{\Sigma (x_i^*-x^*)^2/(p-1)}$

  Trim data again, at 1.5$s^*$
  Recalculate new $x^*$ and $s^*$
  Repeat until convergence
Quality check: Section 5.7

- When AV is determined prior to PT:
  - Compare AV with robust mean or results
  - Determine uncertainty of comparison $u_d$
  - If difference exceeds $2u_d$ then investigate

- When AV is determined from consensus:
  - Compare AV with a reference value from a competent laboratory (could come from homogeneity data)

- Compare robust SD with experience
Determine Performance Interval

5 ways to get “SD for Proficiency” (for z scores)

- By prescription (set by AA)
- By experience (perception) of experts
- From a general model (Horwitz)
- By a precision experiment (ISO 5725-2)
- From participant data (robust SD)
Determine Performance Interval

- Fixed Limits (or “Fitness for Purpose”)
- Can come from methods for SD
- Not widely used
- Preferred for interpretation
  - Fixed percentage across range
  - Fixed value across range
  - Mixed or segmented.
Calculate Performance Statistic

Estimates of bias:

Difference: \[ D = (x - X) \]

Percentage Difference: \[ D\% = 100 \frac{(x - X)}{X} \]

\( D \) and \( D\% \) can be evaluated with Fixed Limits

Estimates of Relative Performance

– rank or percentage rank (not recommended)
– \( z \) score (recommended) \[ z = \frac{(x - X)}{\sigma} \]
$E_n$, $z'$, zeta, $E_z$ scores

- $E_n$ and zeta consider uncertainty of participant result and assigned value
  - Requires consistent determination of uncertainty by all laboratories
- $z'$ and $E_z$ use uncertainty of assigned value only
- $E_n$ in use in calibration, starting to be implemented
Calculate Performance Statistic

Scores that consider uncertainty:
- $E_n$ score ("Error, normalized")
  $$E_n = \frac{x-X}{\sqrt{U_{\text{lab}}^2 + U_{\text{ref}}^2}}$$

- $z'$ scores (like $z$, includes $u_x$)
  $$z' = \frac{x-X}{\sqrt{\sigma^2 + u_x^2}}$$

- zeta scores (like $E_n$, but with std. Uncert)
  $$\text{Zeta} = \frac{x-X}{\sqrt{u_{\text{lab}}^2 + u_{\text{ref}}^2}}$$

- $E_z$ scores (puts $U_x$ in numerator and denominator)
  $$E_z = \frac{x-(X\pm U_x)}{U_x}$$
Evaluate performance

Compare performance statistic against criteria, determine acceptability; i.e.,

- For fixed limits:
  - $\text{Bias} < \text{Limit}$ ➞ "acceptable"
  - $\text{Bias} \geq \text{Limit}$ ➞ "unacceptable"

- For $\zeta$:
  - $-2 < \zeta <+2$ ➞ "acceptable"
  - $-3 < \zeta \leq -2$ or $2 \leq \zeta < 3$ ➞ "warning signal"
  - $\zeta \leq -3$ or $\zeta \geq 3$ ➞ "unacceptable"
Evaluate performance

\[ E_n < 1 \quad \Rightarrow \quad \text{"acceptable"} \]
\[ E_n \geq 1 \quad \Rightarrow \quad \text{"unacceptable"} \]
Combined performance scores

- Analyze data for each item independently
- Special process for Youden pairs (&reps)
- Can be other reasons to combine results
  - Precision
  - Linearity
- Can count number of satisfactory scores
- Not recommended to combine performance scores (such as average z)
Graphic Reports for PT round

- Rank vs. Result (with or without MU)
Figure 6 — Normal probability plot of results of determinations of the lead content of water by 162 laboratories (the results for 19 laboratories are not included).
Figure 7 — Normal probability plot of expanded uncertainties for determinations of the lead content of water by 156 laboratories (the results for 25 laboratories are not included).
Graphic Reports for PT round

- Histograms – of results or scores
NOTE 1  The data are numbers of units (U) in thousands (k) per litre (L) of sample, where a unit is defined by the concentration of an international reference material.

NOTE 2  The numerical values given in the table are those that will be obtained when the calculation is carried out manually working to two decimal places.

Figure 2 — Concentrations of three allergen specific IgE antibodies: histograms of data as reported (data from Table 2).
EXAMPLE: Antibody concentrations. The z-scores for d1 are shown in Figure 8 in the form of a histogram.

Figure 8 — Histogram of z-scores for one round of a proficiency test (data for allergen d1 from Table 7).
Graphic Reports for PT round

- Bar plot of standardized performance statistics (z, h, k)
  - z, or other standardized scores (% error)

h and k plots from 5725
- h same as z, except always from sample SD
- k for repeatability (n≥2 replicates)
Figure 9 — Bar-chart of z-scores for one round of a proficiency test in which the participants determined the concentrations of three allergen specific IgE antibodies (data from Table 7).
Graphic Reports for PT round

- Youden plot (usually w/median lines)
  In this document uses only z scores.
  Should use sample results, for clarity
    Provides evidence of related results, which can suggest consistent bias
    Consistent bias can suggest lack of clearly defined method.
    Confirm with rank correlation test.
Rank Correlation Test

- Common statistical procedure
- Used when Youden Plot suggests relationship
- Use ranks of results from two samples
- Document shows critical values for correlation
Graphic Reports for PT round

- Repeatability SD
Figure 11 — Plot of standard deviations against averages for 25 laboratories (data from Table 13).
Graphic Reports for PT round

- Split sample
  - Check agreement in 2+labs
  - Check difference between results
Figure 14 — Differences between laboratory averages Y - X (with the differences calculated from the ln concentrations).
Figure 12 — Ranges of replicate determinations for Laboratory X (with the ranges calculated from the ln concentrations).

Figure 13 — Ranges of replicate determinations for Laboratory Y (with the ranges calculated from the ln concentrations).
Graphic Reports for More than One PT Round

- Line plot (Shewhart plot) for scores on previous rounds
  - Use any standardized score
  - Show evaluation intervals
  - Show test dates
Figure 15 — Shewhart control charts for the data from one laboratory for 20 rounds of a proficiency test (data from Table 16).
Graphic Reports for More than One PT Round

- CUSUM control chart
  - Can show trends affecting bias
  - Choose some number to use (rolling sum)
  - Sums should trend to zero
  - Not sensitive to current problems
Figure 16 — Cusum control charts for the data from one laboratory for 20 rounds of a proficiency test (data from Table 16).
Graphic Reports for More than One PT Round

- Plot of Standardized Laboratory Biases against assigned value
  - Shows relationship between score and level
  - Can mask time effect...do both
Figure 17 — Relation between z-score and level of concentration for determinations of allergen d1 in one laboratory.
Graphic Reports for more than One PT Round

- Dot Plot
  - Show all samples on same chart
  - Show evaluation intervals
  - Show dates or scheme codes
Figure 18 — Glucose z-Scores in Four PT Events
Annexes in 13528

A. Symbols

B. Homogeneity and stability procedures
   - No statistical test
   - Test relative to evaluation interval

C. Robust procedures
   - Procedure A for mean and SD
   - Procedure S for SD
Homogeneity

- 10 or more samples, 2 replicates
- $SD_S$ for samples (ANOVA calculation)
- $SD_S < 0.3\sigma$
- No F test

Can use experience to reduce testing
When evidence and theory prove homogeneous
Homogeneity

- Homogeneity F test (not recommended)

\[ F = \frac{(SD_S)^2}{s_r^2} \]

\[ F_{crit} = F(.05,k-1, s(n-1)) \]

- High \( S_r \) ➞ insensitive test (large \( SD_S \) passes)
- Low \( S_r \) ➞ too sensitive test (small \( SD_S \) fails)
Stability

- Stability
  - 3 or more samples, 2 replicates
  - Calculate overall mean
  - Mean(H) – Mean(S) < 0.3σ
  - No t test
    - High $S_r$ $\Rightarrow$ insensitive test (big difference passes)
    - Low $S_r$ $\Rightarrow$ too sensitive test (small difference fails)
Robust Analysis

- Algorithm S for pooled SD
- Applied when there are SD estimates from all participants (replicates) (called “w”)
- When 2 replicates, use range (difference)
  \[ w^* = \text{median } w_i. \]
  update with \( \eta \) from Table 19
  \[ w_i^* = \text{trimmed } w_i \text{ at } \eta \]
  \[ w^* = \xi \sqrt{\frac{\Sigma (w_i^*)^2}{p}} \]
- Iterate if necessary
APLAC (NATA) Robust procedure

- Calculate Quartiles Q1, median, Q3

\[ \text{IQR} = Q3 - Q1 \]

*Median* is an *estimate of mean*

*Normalized IQR* is an *estimate of SD*

\[ \text{IQR}_N = 0.7413 \times \text{IQR} \]
APLAC performance statistics

- Calculate relative performance measures
- Between lab agreement
  \[ S_i = (A_i + B_i)/\sqrt{2} \]
- Within lab agreement
  \[ D_i = (A_i - B_i)/\sqrt{2} \text{ if median } (A_i) > \text{median}(B_i) \]
  \[ (B_i - A_i)/\sqrt{2} \text{ if } \text{median}(A_i) < \text{median}(B_i) \]

- Calculate z-scores for these measures
Flour - Ash (%)
MedianY = 0.51, MedianX = 0.51, NIQR Sum = 0.05, NIQR Diff = 0.01
Possibilities for APLAC

- **Estimate Bias**
  
  Percentage Bias(\(Lab_i\)) =
  
  \[
  \frac{((A_i + B_i) - (M_A + M_B))}{(M_A + M_B)} \times 100%
  \]

- **Estimate repeatability**:
  
  \(D_A = (A_i - M_A)\)
  
  \(D_B = (B_i - M_B)\)
  
  \(D_{avg} = \frac{(D_A + D_B)}{2}\)
  
  \(r_{i1} = \sqrt{\frac{((D_A - D_{avg})^2 + (D_B - D_{avg})^2)}{2}}\) \hspace{1cm} \text{(no reps)}
  
  \(r_{i2} = \sqrt{\frac{((A_{i1} - A_{i2})^2 + (B_{i1} - B_{i2})^2)}{2}}\) \hspace{1cm} \text{(2 replicates)}
The End

Thank you!