DESCRIPTION OF AOAC STATISTICS COMMITTEE GENERATED DOCUMENTS

ALTERNATIVE APPROACHES FOR MULTI-LAB STUDY DOCUMENTS.

Alternative approaches approved by Committee on Statistics and by OMB.

1. *tr322-SAIS-XXXV-Reproducibility-from-PT-data.pdf: Discussion on how to obtain reproducibility from proficiency test data, and the issues involved. ................................................................. 2

2. tr333-SAIS-XLII-guidelines-for-use-of-PT-data.pdf: Recommended use of proficiency test data for estimating repeatability and reproducibility. ................................................................. 8

3. *tr324-SAIS-XXXVII-Incremental-collaborative-studies.pdf: Proposed incremental collaborative studies to find repeatability and reproducibility via a sequential series of experiments. ................................................................. 11

4. tr326-SAIS-XXXIX-Min-degr-freed-for-random-factor-estimation.pdf: The relationship of number of replicates or number of collaborators to precision of standard deviation for repeatability or reproducibility. ................................................................. 19

5. tr323-SAIS-XXXVI-When-robust-statistics-make-sense.pdf: Caveats and recommendations on the use of so-called ‘robust’ statistics in accreditation studies. ................................................................. 21

TRADITIONAL PATHWAY MULTI-LAB STUDY DOCUMENTS AND INSTRUMENTS.

Traditional study approach spreadsheet remaining as part of acceptable approaches.

6. JAOAC 2006 89.3.797_Foster_Lee1.pdf: Journal article by Foster and Lee on the number of collaborators needed to estimate accurately a relative standard deviation for reproducibility. ................................................................. 29

7. LCFMPCalculator.exe: This program analyzes serial dilution assay data to obtain a most probable number estimate of concentration with confidence interval. Separate Program

8. AOAC_BindDup_v2-1.xls: This workbook analyzes multi-collaborative study data with up to 4 replicates and reports all necessary statistics concerning repeatability and reproducibility for quantitative studies. Separate Program

9. AOAC-binary-v2-3.xls: This workbook analyzes multi-collaborative study data with arbitrary number of replicates and reports all necessary statistics, including POD, repeatability and reproducibility with confidence intervals, for binary (presence/absence) qualitative studies. Separate Program
TECHNICAL REPORT

NUMBER: TR322    DATE: 2012 July 8

TITLE: Statistical analysis of interlaboratory studies. XXXV. Reproducibility estimates from proficiency test data.

AUTHOR: R. A. LaBudde

ABSTRACT: The issue of estimating reproducibility effects from proficiency test (‘PT’) data is discussed. It is recommended that such data may be used to estimate reproducibility when: 1) There are sufficient collaborators (8 or more remain in the final set of data); 2) Collaborator results are removed only for known cause based on subject-matter expertise and facts. It is recommended that the method used to estimate reproducibility effects is by the standard deviation of mean (across replicates, if any) results for the entire dataset net of crude errors with known causes, corrected for replication. If useful, a lower bound on reproducibility may be obtained conveniently via the interquartile range of the original dataset.

KEYWORDS: 1) PT  2) REPRODUCIBILITY  3) VARIANCE
           4) ROBUST  5) OUTLIER  6) IQR

REL.DOC.: REVISED:

Copyright 2012 by Least Cost Formulations, Ltd.
All Rights Reserved
In the absence of a properly designed randomized and controlled collaborative study, it is tempting to use the less expensive and more commonly available data from proficiency testing (‘PT’) to estimate variance components, such as intercollaborator, repeatability and reproducibility effects. PT data is compromised of independent and only loosely controlled testing of sample replicates by a number of collaborators purporting to use the method under question for the analyte of interest. The collaborators do not follow a study protocol, so may deviate in minor respects from the orthodox method proposed. PT has a primarily goal of measuring performance of a collaborator vs. a group of others, not that of validating the accuracy or precision of the method in question. ‘Check-sample’ testing is a common form of proficiency testing.

Repeatability is an intralaboratory component of variance, and is therefore less subject to controversy. Generally there is no obvious objection to using proficiency test data done in replicate to measure repeatability variance.

Interlaboratory and reproducibility variance components are where most objections arise. The source of the objections is principally due to the self-selection of the collaborators involved, the lack of method control, and the means by which the data are cleaned before estimating the effects.

This paper is concerned primarily with the last of these (data cleaning and estimation).

It will be assumed that PT data is available based on m collaborator results, and all collaborators at least purport to use a specific method for a specific analyte in question.

The purpose of estimating reproducibility effects (intercollaborator and reproducibility) is assumed to be in use as criteria by which the quality of the method in question might be assessed. For this purpose, any compromise in methodology should be biased against the method.

**CHOICE OF ALGORITHM TO ESTIMATE REPRODUCIBILITY VARIANCE**

There is a hierarchy of algorithms possible to estimate reproducibility effects from PT data, listed here in order of decreasing skepticism and increasing assumptions:

1. Do not use PT data for this purpose at all, due to lack of control of methodology. This option considers PT data too poor to be relied upon in decision-making about the test method.

The remaining choices assume the PT data are from a properly randomized experiment (allocation of test portions) and therefore are subject to allowable inference. Typically the collaborators, if sufficiently numerous (say, 8 or more in the cleaned data) to allow a claim of some sort of membership in a ‘volunteer’ hypothetical population which the reproducibility effects might characterize.
2. Do not clean the data. Use the entire set to estimate reproducibility effects. If true outliers or gross errors are present, they will bias variance components high, which will indicate the test method is less precise than it might actually be. This is a conservative approach, but is liberal in that it allows PT data to be used for the purpose.

3. Clean the data, but remove only outliers or gross errors which can be validated by subject matter expertise (rather than purely statistical identification) and external evidence. The reasons for the removal should be documented and non-controversial, and results both with all data and with cleaned data should be reported. This is still a conservative approach, as external and objective evidence of error is required for data removal.

For methods 2) and 3), the presence of unvalidated outliers brings into question assumptions about the type of statistical distribution, *not* the outliers themselves.

The now remaining choices assert the PT data come from a normal distribution (or at least a unimodal symmetric distribution), but may be contaminated by a mixture with other distributions, and this contamination is known a priori as not being related to the test method in question and so should be removed. Generally these assertions will be completely unsupported and therefore highly subject to criticism, unless a substantial quantity of pre-existing data justifies the claims involved. *Although these assertions have been made freely in the past, modern statistical thinking depreciates these assumptions in the absence of clear evidence.*

4. Identify by statistical means any outliers that are improbable given the sample size \( m \). (Generally a conservative 1% significance level is used for this purpose, if \( m < 30 \).) Remove the identified outliers and make estimates from the reduced set of data. This liberal procedure will *always* bias reproducibility effects low.

5. Use so-called ‘robust’ estimators to estimate reproducibility effects via statistics that are unaffected by the outer quantiles of the empirical data distribution. Typically a normal distribution is assumed for the inner quantiles (a unimodal distribution will almost always appear normal near the mode, except for cubic terms). Several such estimators the apparent intercollaborator effect \( s \) (equal to reproducibility if a single replicate is done by each collaborator) are:

5.1. The interquantile range (‘IQR’), normalized by a factor of 0.7413. I.e., \( s = 0.7413 \text{ IQR} \).

5.2. 25\% trimmed data, with

\[
s = s_w \sqrt{\frac{(m-1)/(k-1)}}
\]

where \( s_w \) is the Winsorized standard deviation of the data, \( m \) is the original number of data and \( k \) is the number of data kept after trimming.
5.3. Huber’s method, which dynamically adjusts trimming to the empirical data distribution and then follows a procedure similar to 5.2. This method is an M-estimator.

5.4. Use the median absolute deviation (‘MAD’) from the median and a normalizing factor of 1.4826, i.e., $s = 1.4826 \text{ MAD}.$

5.5. Plot a Q-Q normal graph, select the range of data which is linear in the center, and compute $s$ as the slope of the line fit.

Note that all of methods 5.1)-5.5) will result in a lower bound for $s$, and are therefore maximally liberal (in favor of the test method in question). These methods will generate comparable estimates of $s$ for typical datasets. These methods are heavily dependent upon the normal distribution assumption, and are really only appropriate if it is known a priori that the data do, in fact, follow a normal distribution, and any deviation from this must, in fact, be error.

*It is the author’s opinion that methods 5.1) are due to an error in thinking.* What starts as a valid ‘robust’ theory for estimates of location is improperly twisted into a heavily biased estimate of scale. In the author’s opinion, method 3) is best compromise for the use of PT data to develop estimates of reproducibility effects.
EXAMPLE

Consider the March subset of ‘AOAC-C01-Fat-Data.csv’. There are two replicates for each collaborator, and the average of these is the results analyzed for reproducibility effects. Repeatability is measured by the difference in replicate pairs. It will be assumed for purposes of illustration that all collaborators all used the same test method for fat.

There are 42 collaborators, with average results ranging from 13.08 to 36.08 percent fat, with median and mean results 14.42 and 15.10.

Using method 2), the repeatability standard deviation $s_r$ is 0.4739, the apparent intercollaborator standard deviation $s$ is 3.503, and the repeatability adjusted value for reproducibility standard deviation is 3.519.

The boxplot and normal Q-Q plot for these data are:
The boxplot shows outliers on both tails, with a noticeable skewness to the right. The normal Q-Q plot shows non-normality on both tails, but much more pronounced on the right. The block of data from −0.75 to 0.75 normal quantiles is well fit by a straight line. Note that the outliers on the right follow what appears to be a continuous curve of increasing deviation, even the extreme at 36% fat. There is little evidence this distribution is truly normal and contaminated with a few outliers.

Suppose that we have evidence that the extreme outlier at 36.08% fat is due to a crude error in the laboratory (e.g., mix-up in transcription or calculation). Removing this point for cause, and calculating reproducibility effects via method 3) gives $s = 1.143$ and $s_R = 1.192$, which are much more believable (based on prior experience in fat measurement) values, given the repeatability $s_r = 0.4739$. (These are the values recommended to report in the author’s opinion.)

Using the IQR = 0.62, the estimates are $s = 0.4596$ and $s_R = 0.5688$ using method 5.1). Note that this value is not much different from $s_r$, and is clearly too small given repeatability. This value of $s_R$ is clearly a lower bound on reproducibility, and should be reported as such. One could equally easily just report $s_r = 0.4739$ as such a lower bound on reproducibility as an effectively equivalent estimate.

Using MAD = 0.2875, the estimates from method 5.4) are $s = 0.4262$ and $s_R = 0.5422$. These are very close to those obtained by the IQR in method 5.1).

Using 25% trimming from both tails, 22 data remain, and method 5.2) gives $s = 0.3766$ and $s_R = 0.5041$. These are slightly lower, but comparable to the results of methods 5.1) and 5.4).

Finally, Huber’s method 5.3) gives $s = 0.4918$ and $s_R = 0.5951$, both similar to that of 5.1), 5.2) and 5.4).

The PT provides a conclusion such as:

“Based on the data, the best estimate of repeatability $s_r$ is 0.47% fat and the best estimate of reproducibility $s_R$ is 1.19% fat, with one collaborator removed for cause. The lower bound on reproducibility $s_R$ is no less than 0.57% fat, based on the interquantile range.”
TECHNICAL REPORT

NUMBER: TR333 

DATE: 2013 May 29

TITLE: Statistical analysis of interlaboratory studies. XLII. Guidelines for the use of proficiency test data to replace or supplement collaborative studies.

AUTHOR: R. A. LaBudde

ABSTRACT: Guidelines are given for the acceptable use of proficiency test data to replace or supplement collaborative studies in the estimation of repeatability and reproducibility.

KEYWORDS: 1) PT 2) COLLABORATIVE 3) REPRODUCIBILITY
4) REPEATABILITY

REL.DOC.: TR322, TR323, TR324, TR325

REVISED:

Copyright 2013 by Least Cost Formulations, Ltd.
All Rights Reserved
INTRODUCTION

Proficiency testing (‘PT’) is an economical approach to a collaborative study which has the specific principal goal of measuring a participating collaborator result with respect to the mass of the other collaborator results. This differs in aim from a randomized, controlled collaborative study that is designed specifically to measure repeatability, reproducibility and bias. PT studies are generally performed for a nominal (middle) concentration of analyte in a particular matrix. Designed collaborative studies typically span the gamut of practical concentration levels and use challenging matrices. Participants in PT studies may use nominally the same method, but typically there is no direct control over the exact protocol used. In designed collaborative studies, the precise protocol is specified. In PT studies, replication may or may not be present, and may vary among participants, sometimes without disclosure.

Traditionally, ‘robust’ statistical methodology has been used to analyze PT data. In TR322 and TR323, the use of such statistics for estimating reproducibility was deprecated.

Here guidelines are given for valid use of data and ‘robust’ statistical estimates derived from PT studies for repeatability and reproducibility. (See TR323 for more discussion.)

The choice of performing or not performing a designed collaborative study is that of the method developer. The principal premise assumed here is that of ‘caveat developer’: Statistical estimates are to be designed to be conservative with respect to method approval.

GENERAL GUIDELINES

1. Results must be reported as pertaining only to the specific matrix and concentration involved.
2. The combined set of estimates across all studies will be considered adequate only if the gamut of low to high concentrations for each matrix are studied.
3. All statistical estimates must be reported with 95% confidence intervals. These intervals are important to making the quality of the data visible to reviewers.
GUIDELINES FOR REPEATABILITY ESTIMATION

1. No collaborators should be removed, except for known cause. Such causes may include: 1) does not meet inclusion criteria for protocol, if protocols used are known; or 2) provable contamination. Statistical identification of outliers or influential data is not grounds for removal, only for investigation.
2. Replication may range from 2 to 4 replicates per collaborator. Repeatability should be estimated in the usual way as the pooled standard deviation of the combined set of data.
3. Alternatively, replication may exceed 4 for some collaborators, but each estimate of repeatability standard deviation should be assigned the minimum degrees of freedom across all collaborators, and this number should be used in pooling and reporting.
4. The final number of degrees of freedom assigned to the pooled estimate must be 8 or more.
5. There must be at least 3 collaborators with replication.
6. Robust statistics associated with repeatability may be estimated and reported (such as interquartile range), but not estimates which attempt to convert such statistics to standard deviations by, e. g., constant factors under a normality assumption. Reporting such robust estimates for designed collaborative studies should be encouraged so that comparative results may be accumulated over time.
7. Boxplots and half-normal plots are encouraged.

GUIDELINES FOR REPRODUCIBILITY ESTIMATION

1. No collaborators should be removed, except for known cause. Such causes may include: 1) does not meet inclusion criteria for protocol, if protocols used are known; or 2) provable contamination. Statistical identification of outliers or influential data is not grounds for removal, only for investigation.
2. The number of collaborators providing included data must be 8 or more.
3. If replication is present for most or all collaborators, repeatability, among-collaborator variability and reproducibility should be estimated as standard deviations estimated in the usual way from 1-way analysis of variance (additive model). No more than 4 replicates should be used for any collaborator.
4. If replication is not present, reproducibility only may be estimated (as the standard deviation of collaborator results).
5. Robust statistics associated with reproducibility may be estimated and reported (such as interquartile range), but not estimates which attempt to convert such statistics to standard deviations by, e. g., constant factors under a normality assumption. Reporting such robust estimates for designed collaborative studies should be encouraged so that comparative results may be accumulated over time.
6. Boxplots and half-normal plots are encouraged.
TECHNICAL REPORT

NUMBER: TR324 DATE: 2012 September 5

TITLE: Statistical analysis of interlaboratory studies. XXXVII. Incremental collaborative studies.

AUTHOR: R. A. LaBudde

ABSTRACT: Various experimental design are presented which break a traditional collaborative study into incremental modules that can be performed in sequence over time at individually lower cost. Such incremental collaborative studies would solve the enrollment problem often encountered, and would supply more reliable information than proficiency test studies.

KEYWORDS: 1) PT 2) REPRODUCIBILITY 3) COLLABORATIVE 4) INCREMENTAL

REL.DOC.: TR322 TR323

REVISED:

Copyright 2012 by Least Cost Formulations, Ltd.
All Rights Reserved
**INTRODUCTION**

A validation study for an analytical method strives to characterize the performance of the method on the specified analyte across a gamut of concentration levels and for the matrices of interest claimed. Such a validation study must consist of the following elements:

1. **Inclusivity study**: Validate performance on all commonly encountered variants of the analyte.
2. **Exclusivity study**: Validate performance (rejection or non-recovery) on near-neighbor analytes.
3. **Environmental study**: Validate resistance to interferents and situ modifiers expected to be present.
4. **Under worst-case conditions**.
5. **At end-of-life for reagents and equipment**.
6. **Across the range of analyte concentration from lowest of importance in practice (typically zero) to highest of importance in practice**.
7. **For each matrix for which the method claims adequate performance**.
8. **Characterization of variance source due to repeatability (same technician, same equipment, same reagents, same point in time)**.
9. **Characterization of variance source due to between-collaborator (same point in time)**.
10. **Characterization of variance due to reproducibility (collaborator + single replicate)**.
11. **Characterization of bias in recovery**.
12. **Equivalency or better to a current accepted reference method, if required**.
13. **Performance within required requirements, if specified**.

Achievement of all of these elements in a single planned experiment executed at a single point in time is very difficult in practice, so multiple experiments are typically required.

Traditionally, AOAC International has carried out such a validation study in three steps:

1. **Investigation within the method developer’s laboratory**.
2. **Verification in a single independent AOAC-selected laboratory**.
3. **Investigation in a large-scale collaborative study done in cross-section at a single point in time**.

The method developer performs testing adequate to elements 1), 2), 3), 6), 7), 8), 11) and 12). It also investigates 4) and 5) under a ‘ruggedness’ experiment reported separately.

The independent laboratory repeats a subset of the testing done by the method developer (except for ruggedness) to verify objective performance.

The collaborative study tests elements 6), 7), 8), 9), 10), 11), 12) and 13).

Despite the division of labor into separate parts, the collaborative study remains an expensive and difficult experiment to execute, due to difficulty of enlistment of a sufficient number of collaborators willing to invest the substantial effort involved and the preparation and dispersal of a large number of homogeneous test specimens over a short period of time. These difficulties, plus the availability of a lesser status designation based solely on single laboratory information
(i.e., ‘PTM’ vs. ‘Official’ designation), have led to a great reduction in validation studies which involve collaborative studies.

As of 2012, a new ‘alternative’ methodology to ‘official first action’ has been implemented at AOAC. This new policy allows an ‘official’ status to new methods based on presented single laboratory evidence plus other anecdotal data. The method would be transitioned to ‘final action’ after a period of a year or more in which reproducibility, recovery and repeatability information is collected. The type of evidence which will be considered acceptable for final action has not yet been defined.

Proficiency testing (‘PT’) is an economical approach to a multicollaborator study which has the specific principal goal of measuring a participating collaborator result with respect to the mass of the other collaborator results. PT studies are generally performed for a nominal (middle) concentration of analyte in a particular matrix. Participants may use nominally the same method, but typically there is no direct control over the exact protocol used. Replication may or may not be present, and may vary among participants, sometimes without disclosure.

The use of PT data has been proposed as a possible surrogate for the traditional collaborative study. PT experiments require less intensive involvement for collaborators, so recruitment is easier, and involve typically a single concentration of a single matrix, so deployment is easier. The difficulty is the lack of control and design in PT studies that results in lack of repeatability conditions and lack of interpretability of the reported results. Table 1 shows a comparison of the properties of a collaborative vs. a PT study:
Here we propose that the optimal solution to this issue is to divide a traditional collaborative into separate incremental experiments (‘modules’) that preserve the randomization and control of the planned collaborative study, but reduce the involvement and deployment load to that of a PT study. Such an incremental collaborative study (as opposed to a cross-sectional collaborative study) would have all of the advantages of the traditional collaborative study and of the PT study, with none of the disadvantages of either.

Table 1. Comparison of Collaborative and PT Studies

<table>
<thead>
<tr>
<th>Property</th>
<th>Collaborative</th>
<th>PT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Purpose</strong></td>
<td>Measure method variance components and recovery bias, and to show equivalency to a reference method or meet performance requirements</td>
<td>Measure collaborator result compared to others</td>
</tr>
<tr>
<td><strong>Method procedure</strong></td>
<td>Controlled</td>
<td>Variants possible</td>
</tr>
<tr>
<td><strong>Test portions</strong></td>
<td>Randomized</td>
<td>Randomize</td>
</tr>
<tr>
<td><strong>Levels of concentration of analyte</strong></td>
<td>Full range of interest</td>
<td>Single level, nominal</td>
</tr>
<tr>
<td><strong>Matrices</strong></td>
<td>Multiple</td>
<td>Single</td>
</tr>
<tr>
<td><strong>Disclosure</strong></td>
<td>Full</td>
<td>Simple result</td>
</tr>
<tr>
<td><strong>Collaborator reporting</strong></td>
<td>Controlled</td>
<td>Ad hoc</td>
</tr>
<tr>
<td><strong>Experimental design</strong></td>
<td>Controlled</td>
<td>Ad hoc</td>
</tr>
<tr>
<td><strong>Reproducibility conditions</strong></td>
<td>Controlled</td>
<td>May vary</td>
</tr>
<tr>
<td><strong>Repeatability conditions</strong></td>
<td>Controlled</td>
<td>May vary</td>
</tr>
<tr>
<td><strong>Time element</strong></td>
<td>Cross-sectional</td>
<td>Learning curve</td>
</tr>
<tr>
<td><strong>Cost</strong></td>
<td>High</td>
<td>Low to moderate</td>
</tr>
<tr>
<td><strong>Suspicious data</strong></td>
<td>Infrequent</td>
<td>Common</td>
</tr>
<tr>
<td><strong>Interpretability</strong></td>
<td>Usually clear</td>
<td>Quizzical</td>
</tr>
</tbody>
</table>
INCREMENTAL COLLABORATIVE STUDY

The results of a traditional collaborative study are typically reported separately for each concentration level measured for each matrix. Repeatability, reproducibility, recovery and comparative results frequently are different for different matrices; and repeatability, reproducibility and recovery are typically concentration dependent (cf. ‘HORRAT’ index).

DESIGN ELEMENTS COMMON TO ALL SCHEMES FOLLOWING

All of the proposed versions of incremental collaborative studies will have the following design elements:

1. Fixed number of replicates. (2 are suggested)
2. Repeatability conditions for replicates (same equipment and reagents, same technician, same point of time).
3. Specified and constant method protocol across all measurements and all collaborators (reproducibility conditions).
4. Controls to maintain study integrity.
5. Specified reporting formal for results.
6. Randomization and masking wherever possible and desirable (replications, order of testing concentrations).

INCREMENTAL BY MATRIX

The first major line of demarcation for splitting a collaborative study into modules is at the matrix level. For example, if the plan is to validate a test method for three different matrices, then three different increments of the collaborative study might be performed, one for each matrix involved. Generally, this will involve studies that are still fairly expensive, given the multiple concentration levels and replication involved. The order of the matrices studied may be arranged in declining order of importance so that early termination of the study yields maximum value at minimum cost. If the confounding of time sequence with matrix is unacceptable, the order of the matrices may be randomized. Different collaborators may be used for each increment, which will greatly improve ease of enrollment.

Current thinking proposes study of various matrices at the single laboratory level, with a subsequent single worst-case matrix chosen for the collaborative study. Note, however, that this does not allow measurement of reproducibility, and should only be considered when the number of replicates used provides a statistical power to test method equivalency or performance requirements at the necessary level (and no less than that provided from a collaborative study). If reproducibility varies with matrix, as it frequently does, this should be taken into account in selecting the worst-case matrix. Also note that testing only a single worst-case matrix in a collaborative study will characterize the candidate method by its worst-case reproducibility.
An alternative to the single worst-case matrix collaborative study is incremental collaborative studies for each matrix, but with a reduced (e.g., 3) number of collaborators for all but the worst-case matrix (see Fractional by Collaborators below). These (reduced and less expensive) collaborative studies will provide partial, suggestive indications of performance. If performance is poor, the collaborative study may be upgraded to a full collaborative study, or the matrix dropped from claims. These ‘pilot’ studies would provide information by which the single worst-case matrix full collaborative study could be designed.

**INCREMENTAL BY MATRIX AND BY CONCENTRATION LEVEL**

The next level of subdivision that is convenient for modularization is by concentration level. A typical collaborative study uses at least 3 levels of concentration (zero, low, high), and frequently 4 or more. Each of these, for a particular matrix, can be considered a separate increment of the collaborative study. The range of concentrations studied should span the range of concentration expected in use for which an adequate performance is claimed. The relevant study questions to be answered are:

1. Does the candidate method have a sufficiently low false positive fraction or response at the zero concentration (‘blank’) level?
2. Does the candidate method have adequate recovery and reproducibility at low to intermediate concentration levels?
3. Does the candidate method have adequate recover and reproducibility across the gamut of high concentration levels?
4. Is the candidate method better or equal to the specified reference method across all concentrations?

Each concentration level studied will require an adequate set of collaborators to determine reproducibility (*but different collaborators may be used for each matrix and level*, which will greatly improve ease of enrollment).

The concentration levels should be randomized across time, so that a systematic confounding of concentration with time (e.g., learning curve) does not occur. If ‘M’ denotes ‘matrix’ and ‘C’ denotes concentration level, then a possible sequence of study increments for two matrices, each with 4 concentration levels, might be, e.g.:


The time factor (learning curve) would be confounded with matrix here. If this is not acceptable, and a commitment to testing all matrices is made, the order of the M:C combinations may be completely randomized.
Note that the ‘Incremental by Matrix and by Concentration Level’ study is a randomized controlled versus of the PT study.

**FRACTIONAL BY COLLABORATORS**

A study with a dozen or more collaborators is still difficult and expensive to execute, due to problems with enrollment. One way around such large studies is to divide the collaborative study module into ‘fractions’ by groups of collaborators. These groups might be as small as 3 or as large as 6 or more. The collaborators involved in each fraction are different, but the same collaborators may be reused for different matrix-level combinations.

The expectation is that the results of these ‘fractional by collaborator’ studies would be composited to estimate reproducibility and equivalency or the meeting of performance requirements. In order for this to be feasible (without confounding with sample preparation or concentration level), the concentration level in the matrix must be reasonably accurately controllable, or sufficient time-stable test portions capable of being prepared *ab initio*.

As before, the matrix-level-collaborator combinations M:L:C should be randomized at least over level and collaborator, and also over matrices, if a commitment to the full course of testing can be made. The size of the ‘fraction’ effect can be estimated in the analysis of the composited data, and examined to see if it is sufficiently negligible, justifying the composition of data.

The time element will be confounded with matrix, if matrices are not randomized, otherwise with a higher order interaction term.

**MINIMUM SIZE REQUIREMENTS**

1. Repeatability standard deviation requires a minimum of 8 degrees of freedom for estimation will any accuracy. Most reasonable designs will provide many more than this.

2. Reproducibility standard deviation requires a minimum of 8 degrees of freedom for estimation with any accuracy. The number of collaborators must be several more than this in order to allow for disqualification for cause or drop-outs.

3. Recovery bias will require a sample size sufficient to provide a 95% confidence interval of acceptable width.

4. Performance requirements may require total sample sizes (across all replicates and collaborators) of 60 or more.
RECOMMENDED SEQUENTIAL VALIDATION PROCEDURE

1. Method developer provides test results for all required sub-studies with the exception of measurement of reproducibility. All matrices and all concentrations are studied, with verification that all performance requirements are met. Repeatability and recovery bias estimates are obtained.

2. A random selection or expertise-based selection of the developer studies are repeated in an independent laboratory chosen by AOAC. The goal is to objectively verify the results obtained by the developer.

3. Based upon favorable results from these studies, a ‘first action’ status is granted.

4. Subsequent incremental, sequential, fractional collaborative studies are carried out over the course of one or two years.

5. Based upon the composition results, ‘final action’ status is granted.

SHOWING EQUIVALENCY TO A REFERENCE METHOD

Suppose in lieu of performance requirements that the candidate method must be shown in the validation study to be equal or better in performance than a specified reference method of known quality.

To statistically test such 1-sided equivalency, several steps must occur:

1. A subject-matter expertise based estimate of a ‘material difference’ $\Delta$ must be specified. This is the amount by which the candidate method performance can differ on the average from the reference method performance and still be considered ‘equivalent’. The value of $\Delta$ depends upon the application, and cannot be estimated by statistics.

2. The validation study is carried out, and the mean difference between the candidate and reference method results estimated, along with a 1-sided 95% confidence lower limit.

3. If the 1-sided 95% confidence lower limit found is greater than $-\Delta$, then there is sufficient evidence to claim that the candidate method is equal or better in performance to the reference method.

4. If the 1-sided 95% confidence lower limit found is greater than $+\Delta$, then there is sufficient evidence to claim that the candidate method is better in performance to the reference method.
TECHNICAL REPORT

NUMBER: TR326    DATE: 2012 October 2

TITLE: Statistical analysis of interlaboratory studies. XXXIX. Minimum degrees of freedom for random factor (standard deviation) estimation.

AUTHOR: R. A. LaBudde

ABSTRACT: In collaborative studies question of the minimum number of collaborators required to estimate reproducibility or collaborator effect arises as a contentious issue, as collaborators are expensive. In single laboratory studies, the question of the minimum number of replicates needed to estimate repeatability is a similar, but less contentious, issue, as replicates are cheap to perform. Using as a paradigm the 95% confidence interval on the standard deviation σ, the recommendation is made that minimum number of degrees of freedom needed is no less than five and should be at least seven for reasonable results. Note that the normal distribution paradigm used is a ‘best case’ scenario. For distributions deviating from normal, even larger number of degrees of freedom should be used. These recommendations correspond to no less than 6 collaborator results in the final dataset, and preferably 8 or more. Guidelines should require 8 or more collaborators, with as low as 6 used in extenuating circumstances. It is also strongly recommended that all reported standard deviations also report a 95% confidence interval (based on a normal distribution, if necessary) so that the degree of imprecision can be assessed by the reader.

KEYWORDS: 1) REPEATABILITY  2) REPRODUCIBILITY
3) COLLABORATIVE  4) CHI-SQUARE
4) INCREMENTAL

REL.DOC.: TR298

REVISED:

Copyright 2012 by Least Cost Formulations, Ltd.
All Rights Reserved
95% confidence interval for Sigma, given normal distribution

<table>
<thead>
<tr>
<th>Degrees of Freedom</th>
<th>LCL Multiplier</th>
<th>UCL Multiplier</th>
<th>Multiplier Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.446</td>
<td>31.910</td>
<td>71.52</td>
</tr>
<tr>
<td>2</td>
<td>0.521</td>
<td>6.285</td>
<td>12.07</td>
</tr>
<tr>
<td>3</td>
<td>0.566</td>
<td>3.729</td>
<td>6.58</td>
</tr>
<tr>
<td>4</td>
<td>0.599</td>
<td>2.874</td>
<td>4.80</td>
</tr>
<tr>
<td>5</td>
<td>0.624</td>
<td>2.453</td>
<td>3.93</td>
</tr>
<tr>
<td>6</td>
<td>0.644</td>
<td>2.202</td>
<td>3.42</td>
</tr>
<tr>
<td>7</td>
<td>0.661</td>
<td>2.035</td>
<td>3.08</td>
</tr>
<tr>
<td>8</td>
<td>0.675</td>
<td>1.916</td>
<td>2.84</td>
</tr>
<tr>
<td>9</td>
<td>0.688</td>
<td>1.826</td>
<td>2.65</td>
</tr>
<tr>
<td>10</td>
<td>0.699</td>
<td>1.755</td>
<td>2.51</td>
</tr>
<tr>
<td>12</td>
<td>0.717</td>
<td>1.651</td>
<td>2.30</td>
</tr>
<tr>
<td>15</td>
<td>0.739</td>
<td>1.548</td>
<td>2.10</td>
</tr>
<tr>
<td>18</td>
<td>0.756</td>
<td>1.479</td>
<td>1.96</td>
</tr>
<tr>
<td>20</td>
<td>0.765</td>
<td>1.444</td>
<td>1.89</td>
</tr>
<tr>
<td>25</td>
<td>0.784</td>
<td>1.380</td>
<td>1.76</td>
</tr>
<tr>
<td>30</td>
<td>0.799</td>
<td>1.337</td>
<td>1.67</td>
</tr>
<tr>
<td>35</td>
<td>0.811</td>
<td>1.304</td>
<td>1.61</td>
</tr>
</tbody>
</table>

---

Recommended by Committee on Statistics: 07-17-2013
Reviewed and approved by OMB: 07-18-2013
TECHNICAL REPORT

NUMBER: TR323

DATE: 2012 August 13

TITLE: Statistical analysis of interlaboratory studies. XXXVI. When robust statistics make sense in proficiency test studies.

AUTHOR: R. A. LaBudde

ABSTRACT: The use of ‘robust’ statistical estimators for measures of central location and variation are discussed. Robust estimators for measures of central location are non-controversial and acceptable. Robust estimators for measures of variation, such as reproducibility standard deviation, are heavily biased and therefore deprecated. Examples of performance of robust estimators are given for four example distributions (normal, lognormal, student-t and gamma).

KEYWORDS: 1) PT 2) REPRODUCIBILITY 3) VARIANCE 4) ROBUST 5) OUTLIER 6) IQR

REL.DOC.: TR322

REVISED:

Copyright 2012 by Least Cost Formulations, Ltd.
All Rights Reserved
INTRODUCTION

Proficiency testing (‘PT’) is an economical approach to a multicollaborator study which has the specific principal goal of measuring a participating collaborator result with respect to the mass of the other collaborator results. PT studies are generally performed for a nominal (middle) concentration of analyte in a particular matrix. Participants may use nominally the same method, but typically there is no direct control over the exact protocol used. Replication may or may not be present, and may vary among participants, sometimes without disclosure.

Traditionally, ‘robust’ statistical methodology has been used to analyze PT data. In TR322, the use of such statistics for estimating reproducibility was deprecated.

Here the issues related to robust statistics is discussed, and indications are made as to when such methodology might actually make sense.

MEASURE OF CENTER (LOCATION)

The original use of robust statistics was with respect to measures of centrality, i.e., the center point of the distribution. The arithmetic mean (first moment) has many good theoretical properties, particularly when a normal distribution is present, but is subject to influence by outliers (with a coefficient of 1/n, where n is the number of data in the sample).

When far or multiple outliers are suspected to be present, there are two general policies in use:

1. Remove the outlier for cause, if investigation and subject-matter expertise renders the data point involved subject to crude error, contamination or other gross failure of methodology. (Statistical identification of outliers may be helpful, but removal solely upon such identification is deprecated.) After removal of any outliers, the usual statistics (e.g., arithmetic mean and standard deviation) are estimated from the remaining data.

2. Do not remove outliers, but remove their influence. This is done by using ‘robust’ statistics that give less weight to data in the far tails. Examples of such robust statistics as measures of center are:

   2.1. Median.
   2.2. $\alpha$-trimmed mean (where a fraction $\alpha$ of the data are removed from each tail).

The median may be interpreted as a 50%-trimmed mean, in which case both of the above examples are of the same class. Trimming eliminates the influence of far outliers and concentrates estimation using only the center points of the distribution. The immunity to outliers increases with $\alpha$, which typically is 10%, 25% or 50%.

Using data exclusively from the center of the empirical distribution to find a good measure of the location of the center of the distribution is non-controversial. Immunizing this measure against
skewness, kurtosis and suspect outliers makes good sense. Use of robust statistics for measures of central location is well-established and in common use in a variety of subject areas.

**MEASURE OF VARIATION (SPREAD)**

As reviewed in TR322, robust statistics have been extended to provide measures of variation that are less influenced by outliers than the standard deviation, which is based on the second central moment and amplifies the effect of far outliers. The standard deviation is much more sensitive to far outliers than is the arithmetic mean.

However, as mentioned in TR322, variation is intrinsically a property of the entire width of the data distribution, not just the center cluster. So use of robust statistics for this purpose results in heavily biased (downward) estimates, and is deprecated. Such robust statistics also commonly scale results to an assumed underlying normal distribution, which is a strong and frequently unwarranted assumption.

In studies that provide quantitative measurement of analytes (both microbiological counts and chemical components), the most common distribution encountered is the lognormal, which is heavily skewed. Data from the lognormal distribution appears to contain sporadic outliers due to this skewness, and consequently use robust estimates of variation are unacceptably low.

**RESULTS FOR EXAMPLE DISTRIBUTIONS**

It is instructive to see how robust measures of variation perform for several example distributions. In each case, the results are given for a sample set of data of size 24.

**NORMAL DISTRIBUTION**

Consider first the unit (standard) normal distribution, with mean 0 and standard deviation 1.

Based on 100,000 realizations of samples of size 24, the estimated mean standard deviation (‘s’) is 0.9999, the equivalent estimate based on the mean absolute deviation from the median (‘MAD’) is 0.9766, and the equivalent estimate based on the interquartile range (‘IQR’) is 0.9538. Note that there are residual biases in the MAD and IQR based estimates, due to use of asymptotic scale factors that are slightly in error for a finite sample of size 24.

The standard errors of the statistics (i.e., standard deviations of the sampling distributions) are 0.1466 for s, 0.2311 for the MAD-based estimate and 0.2219 for the IQR-based estimate. These correspond to efficiencies relative to s of 0.4024 for MAD and 0.4363 for IQR. This means is would take 2.5 times the sample size to get equivalent precision for the MAD-based estimate and 2.3 times the sample size for the IQR-based estimate.
Even for the normal distribution, the robust estimates of variation are biased by several percent for reasonable sample sizes and are of very low efficiency compared to the sample standard deviation. This is a step price to pay for protection from outliers. The mean results do reasonably reproduce the originating distribution (with the small bias obvious at the mode):
LOGNORMAL DISTRIBUTION

Now consider the standard lognormal distribution with log mean 0 and log standard deviation 1. The unlogged mean is 1.6487 and the unlogged median is 1.0, with an unlogged standard deviation of 2.1612.

Based on 100,000 realizations of samples of size 24, the estimated mean standard deviation (’s’) is 1.899, the equivalent estimate based on MAD is 0.8904, and the equivalent estimate based on the IQR is 1.062. The biases in the MAD- and IQR-based estimates are substantial.

The standard errors of the statistics (i.e., standard deviations of the sampling distributions) are 1.054 for s, 0.8904 for the MAD-based estimate and 0.3756 for the IQR-based estimate. The sample standard deviation s is imprecise, but unbiased. The MAD- and IQR-based estimates are precise, but heavily biased.

Use of ‘robust’ estimators for the standard deviation when the underlying distribution is lognormal (i.e., heavily skewed) results in estimates which are only $\frac{1}{2}$ of the correct value.
The standard student-t distribution with 4 degrees of freedom has mean 0 and standard deviation of 1.4142. It is an example of a symmetric distribution with long tails.

Based on 100,000 realizations of samples of size 24, the estimated mean standard deviation (‘s’) is 1.376, the equivalent estimate based on MAD is 1.090, and the equivalent estimate based on the IQR is 1.061. The biases in the MAD- and IQR-based estimates are substantial.

The standard errors of the statistics (i.e., standard deviations of the sampling distributions) are 0.3895 for s, 0.2752 for the MAD-based estimate and 0.2638 for the IQR-based estimate. The sample standard deviation s is less precise, but unbiased. The MAD- and IQR-based estimates are precise, but biased.

Use of ‘robust’ estimators for the standard deviation when the underlying distribution is platykurtic results in estimates which are too small by 30+%. 

Student-t (4 d.f.) Distribution

Recommended to OMB by Committee on Statistics: 07-17-2013
Reviewed and approved by OMB: 07-18-2013
GAMMA DISTRIBUTION

The gamma distribution with shape = 2 and scale = 1 (rate = 1) has mean 2 and standard deviation of 1.4142. It is an example of an asymmetric distribution skewed to the right, but less so than the lognormal distribution.

Based on 100,000 realizations of samples of size 24, the estimated mean standard deviation (‘s’) is 1.396, the equivalent estimate based on MAD is 1.187, and the equivalent estimate based on the IQR is 1.229. The biases in the MAD- and IQR-based estimates again are substantial.

0.3075745 0.3059277 0.3226098

The standard errors of the statistics (i.e., standard deviations of the sampling distributions) are 0.3076 for s, 0.3059 for the MAD-based estimate and 0.3226 for the IQR-based estimate. All estimates are comparable in precision, but the MAD- and IQR-based estimates are biased.

Use of ‘robust’ estimators for the standard deviation when the underlying distribution is skewed results in estimates which are too small by 20%.
RECOMMENDATIONS

1. Use of ‘robust’ estimators for measures of center location is non-controversial, as the measure of centrality is based on central data.

2. Use of ‘robust’ estimators for measures of variation or spread is deprecated, as they will be substantially biased low.

3. A circumstance in which ‘robust’ estimators of variation might be recommended is when:
   
a. The underlying distribution is known a priori to be normally distributed or substantial additional evidence (other than the actual data in question) supports this assertion.

b. The observed data are known a priori to be contaminated with data a foreign distribution, and this contamination is exclusively found in the tails of the empirical distribution.

c. The outliers present are known a priori to not be identifiable for assigned cause.

This circumstance might arise, for example, in PT data where it can be supposed that substantially different variants of the method in question may be in use, and these variants cannot be identified from the information collected in the study. Inclusion of all data in such a study may result in an estimate of reproducibility standard deviation that is known a priori to be much too large.

4. In all other circumstance, reproducibility standard deviation should be estimated in the usual way after removal of outliers for assignable cause.
STATISTICAL ANALYSIS

Determining a One-Tailed Upper Limit for Future Sample Relative Reproducibility Standard Deviations

FOSTER D. MCCLURE and JUNG K. LEE
U.S. Food and Drug Administration, Center for Food Safety and Applied Nutrition, Office of Scientific Analysis and Support, Division of Mathematics, Department of Health and Human Services, 5100 Paint Branch Pkwy, College Park, MD 20740-3835

A formula was developed to determine a one-tailed 100p% upper limit for future sample percent relative reproducibility standard deviations

\[ RSD_R, \% = \frac{100s_R}{\bar{y}} \]

where \( s_R \) is the sample reproducibility standard deviation, which is the square root of a linear combination of the sample repeatability variance \( (s_L^2) \) plus the sample laboratory-to-laboratory variance \( (s_{L1}^2) \), i.e., \( s_R = \sqrt{s_L^2 + s_{L1}^2} \), and \( \bar{y} \) is the sample mean. The future RSDR,% is expected to arise from a population of potential RSDR,% values whose true mean is

\[ \xi_R, \% = \frac{100\sigma_R}{\mu} \]

where \( \sigma_R \) and \( \mu \) are the population reproducibility standard deviation and mean, respectively.

The sample relative reproducibility standard deviation \( (RSD_R) \), usually expressed as a percent \( (RSD_R,\%) \) is obtained using a completely randomized model (CRM: 1) and is defined as \( RSD_R, \% = \frac{100s_R}{\bar{y}} \), where \( s_R \) is the sample reproducibility standard deviation, which is the square root of a linear combination of the sample repeatability variance \( (s_L^2) \) plus the sample laboratory-to-laboratory variance \( (s_{L1}^2) \), i.e., \( s_R = \sqrt{s_L^2 + s_{L1}^2} \), and \( \bar{y} \) is the sample mean. The sample RSDR,% is an important method performance indicator for validation organizations such as AOAC INTERNATIONAL. Therefore, we reasoned that it might be of great value to have a statistical procedure to determine a one-tailed 100p% upper limit \( (\gamma_p) \) for future sample RSDR,% values. A thorough literature search suggested that until now no such procedure, based on a CRM, has existed. However, we did note that Hald (2) had investigated the distribution of the coefficient of variation for the single sample model, i.e., \( y_i = \mu + e_i \), where \( \mu \) is an unknown constant and \( e_i \) is the random error associated with \( y_i \).

After considerable study of the problem, we came to the conclusion that an exact limit for an RSDR was unachievable, primarily because the exact distributions of the sample \( s_L^2 \) and \( s_R \) are very complicated, and possibly impossible to obtain. Therefore, we sought to develop a formula to determine an approximate one-tailed 100p% upper limit \( (\gamma_p) \) for future sample RSDR values, obtained under a CRM model, by extending Hald’s single sample approximation for \( y_p \). In doing so, we used a normal approximation and the delta-method (δ-method; 1, 3, 4).

Collaborative Study Model

Here, we will review the CRM used by AOAC to establish background notations. The model represents 2 sources of variation: the first is often referred to as “among-laboratories” and the other as “within-laboratory” variation. For the CRM, an analytical result \( (y_{ij}) \) obtained by laboratory \( i \) on test sample \( j \) is expressed as \( y_{ij} = \mu + \tau_i + \varepsilon_{ij} \), \( i = 1, 2, \ldots, L \) and \( j = 1, 2, \ldots, n \), where \( \mu \) is the grand mean of all potential analyses for the material, \( \tau_i \) a constant associated with laboratory \( i \), and \( \varepsilon_{ij} \) the random error associated with analysis \( y_{ij} \). It is also assumed that \( \tau_i \) and \( \varepsilon_{ij} \) are independent random variables, such that \( \tau_i \) is normally distributed \( (\sim N(0, \sigma^2_\tau)) \). Similarly, \( \varepsilon_{ij} \) is normally distributed with a mean of 0 and variance of \( \sigma^2_\varepsilon \), i.e., \( \varepsilon_{ij} \sim N(0, \sigma^2_\varepsilon) \).

Given the above model, we note that the expected value of \( y_{ij} \) equals the grand mean \( (\mu) \) \( E(y_{ij}) = \mu \), the variance of \( y_{ij} \) equals the reproducibility variance \( \text{var}(y_{ij}) = \sigma^2_\tau + \sigma^2_\varepsilon \), the covariance of \( y_{ij} \) and \( y_{ik} \) equals the “among-laboratories” component of variation \( \text{cov}(y_{ij}, y_{ik}) = \sigma^2_\tau \) for \( j \neq k \), and the correlation between \( y_{ij} \) and \( y_{ik} \) is \( \frac{\sigma^2_\tau}{\sigma^2_\tau + \sigma^2_\varepsilon} \) for \( j \neq k \), i.e., within a given laboratory the \( y_{ij} \) are correlated under the CRM (5, 6).
Data Analysis

To obtain the sample estimate of the repeatability and reproducibility variances \( s_r^2 \) and \( s_R^2 \), respectively, the data from the CRM are analyzed to obtain the mean squares reflecting the "among-laboratories" and "within-laboratory" variations. Using an analysis of variance (ANOVA) technique for analyzing the data, the sample mean for the \( i \)th laboratory \( \bar{y}_i = \frac{1}{n} \sum_{j=1}^{n} y_{ij} \) and the sample grand mean \( \bar{y} = \frac{1}{L} \sum_{i=1}^{L} \frac{1}{n} \sum_{j=1}^{n} y_{ij} \) are used in computing the "among-laboratories" mean square

\[
MS_L = \frac{n}{L-1} \left( \bar{y} - \bar{y}_i \right)^2
\]

and the "within-laboratory" mean square

\[
MS_s = \frac{1}{L(n-1)} \left( \sum_{i=1}^{L} \sum_{j=1}^{n} (y_{ij} - \bar{y}_i)^2 \right) = s_s^2
\]

The sample reproducibility variance

\[
\left( s_R^2 =\frac{n}{L(n-1)}(MS_L - MS_s) + MS_s = s_s^2 + s_r^2 \right)
\]

is an estimate of the population reproducibility variance \( \sigma_R^2 = s_R^2 + \sigma_s^2 \). The sample reproducibility standard deviation \( s_R \) is the square root of \( s_R^2 \) and is an estimate of the population reproducibility standard deviation \( \sigma_R \). The sample RSD\(_R \) is an estimate of the population relative reproducibility standard deviation \( \xi_R = \frac{\sigma_R}{\mu} \), where \( \mu \) is the population mean.

Statistical Distribution and Independence of \( s_R \) and \( \bar{y} \)

In developing a formula for \( \gamma_p \), it is important to establish that the distribution and independence of \( s_R \) and \( \bar{y} \) exist. In an earlier paper, McClure and Lee (1) detailed the derivation of the asymptotic distribution of \( s_R \), assuming that the reproducibility variance \( s_R^2 \) was approximately normally distributed (·) with mean \( \left( \sigma_R^2 \right) \) and variance \( V(s_R^2) \), i.e.,

\[
s_R^2 \sim N(\sigma_R^2, V(s_R^2)), \text{ by finding } V(s_R^2) \text{ and applying the } \delta\text{-method (3, 4). Thus, the distribution of } s_R^2 \text{ is asymptotically normal with mean } \left( \sigma_R^2 \right) \text{ and variance } V(s_R^2), \text{ i.e.,}
\]

\[
s_R^2 \sim N(\sigma_R^2, V(s_R^2)), \text{ where}
\]

\[
V(s_R^2) = \frac{1}{2\sigma_R^2} \left( \frac{n-1}{n} \sigma_R^2 + \frac{(\sigma_R^2 + n\sigma_s^2)^2}{n(L-1)} \right)^2.
\]

Also, based on the CRM, the sample mean \( \bar{y} \) is normally distributed with a mean \( (\mu) \) and variance \( V(\bar{y}) = \frac{\sigma_R^2 + n\sigma_s^2}{nL} \), i.e.,

\[
\bar{y} \sim N(\mu, V(\bar{y})).
\]

In establishing the independence of \( s_R \) and \( \bar{y} \), we direct attention to the work of Stuart et al. (5), who have shown the mean, "among-groups" and "within-groups" sums of squares, which are analogous to our mean \( \bar{y} \), "among-laboratories" sum of squares \( (SS_L) \) and "within-laboratory" sum of squares \( (SS_s) \), are statistically independent under the CRM, and, hence, the mean \( \bar{y} \) and reproducibility standard deviation

\[
s_R = \sqrt{s_R^2} = \sqrt{\frac{SS_L}{L} + \frac{SS_s}{n(L-1)}}
\]

are independent.

100p% One-Tailed Upper Limits for Future Sample RSD\(_R \) Values

In approximating the distribution of the sample RSD\(_R \), we want the probability that the sample RSD\(_R \) is less than the \( p \)th percentile value \( \gamma_p \) to equal \( p \), i.e.,

\[
Pr(\ RSD_R < \gamma_p) = p \Rightarrow Pr(\ s_R - \gamma p \bar{y} < 0) = p.
\]

Here we note that the variable \( z = s_R - \gamma p \bar{y} \) is in the probability statement

\[
Pr(\ s_R - \gamma p \bar{y} < 0) = p
\]

is approximately normally distributed with mean \( (E(z) = \sigma_R - \gamma p \mu) \) and variance \( V(z) = V(s_R) + \gamma^2 p V(\bar{y}) \).

We chose the variable \( z = s_R - \gamma p \bar{y} \) because it is known that a linear function of a normal and an approximately normal variable will usually deviate less from the normal distribution than the distribution of the ratio of the 2 variables (2). Substituting the variances \( V(s_R) \) and \( V(\bar{y}) \) into \( V(z) \), we obtained the following:

\[
V(z) = \frac{1}{2\sigma_R^2} \left( \frac{n-1}{n} \sigma_R^2 + \frac{(\sigma_R^2 + n\sigma_s^2)^2}{n(L-1)} \right)^2 + \gamma^2 p \frac{\sigma_R^2 + n\sigma_s^2}{nL}.
\]

Hence, we obtained

\[
Pr(s_R - \gamma p \bar{y} < 0) \equiv \Phi\left[\frac{\gamma p \mu - s_R}{V(z)^{1/2}}\right] = p,
\]

where \( \Phi(R) \) represents the cumulative standard normal distribution. Therefore, \( \gamma p \mu - s_R \approx z_p \), where \( z_p \) is the abscessa on the standard normal curve that cuts off an area \( p \) in the upper tail. Substituting the expression for \( V(z) \) in the above formula, we have

\[
z_p \equiv \frac{\gamma p \mu - s_R}{V(z)^{1/2}} = \frac{1}{2\sigma_R^2} \left( \frac{n-1}{n} \sigma_R^2 + \frac{(\sigma_R^2 + n\sigma_s^2)^2}{n(L-1)} \right)^2 + \gamma^2 p \frac{\sigma_R^2 + n\sigma_s^2}{nL}.
\]

Performing some algebra on the right-most expression above, we obtained the following:
z_p \approx \frac{\gamma_p \mu - \sigma_R}{\sigma_R} \left\{ \frac{(n-1)\sigma_R^2}{2nL} \left( \frac{\sigma_R^2}{\sigma_R^2} \right)^2 + \frac{\sigma_R}{2n(L-1)} \left( n - (n-1) \frac{\sigma_R^2}{\sigma_R^2} \right)^2 \right\}^{1/2} + \frac{\gamma_R^2 \sigma_R^2}{nL} \left( n - (n-1) \frac{\sigma_R^2}{\sigma_R^2} \right) \left\{ \frac{(n-1)\sigma_R^2}{2nL} \left( \frac{\sigma_R^2}{\sigma_R^2} \right)^2 + \frac{\sigma_R}{2n(L-1)} \left( n - (n-1) \frac{\sigma_R^2}{\sigma_R^2} \right)^2 \right\}^{1/2}.

Letting $\Theta = \frac{\sigma_L}{\sigma_R}$ (the ratio of the population repeatability and reproducibility standard deviations), we obtained the following:

$$z_p \approx \frac{\gamma_p \mu - \sigma_R}{\sigma_R} \left\{ \frac{(n-1)\Theta^4}{2nL} \left( \frac{\sigma_R^2}{\sigma_R^2} \right)^2 + \frac{\Theta}{2n(L-1)} \left( n - (n-1) \frac{\sigma_R^2}{\sigma_R^2} \right)^2 \right\}^{1/2} + \frac{\gamma_R^2 \Theta^2}{nL} \left( n - (n-1) \frac{\sigma_R^2}{\sigma_R^2} \right) \left\{ \frac{(n-1)\Theta^4}{2nL} \left( \frac{\sigma_R^2}{\sigma_R^2} \right)^2 + \frac{\Theta}{2n(L-1)} \left( n - (n-1) \frac{\sigma_R^2}{\sigma_R^2} \right)^2 \right\}^{1/2}.

Solving this equation for $\gamma_p$, we obtained:

$$\gamma_p = \frac{\frac{1}{\Theta} \frac{1}{2} \frac{z_p}{\sqrt{\frac{2}{\sigma_R^2}} \left( n - (n-1) \frac{\sigma_R^2}{\sigma_R^2} \right)}}{1 + \frac{z_p^2}{\sqrt{\frac{2}{\sigma_R^2}} \left( n - (n-1) \frac{\sigma_R^2}{\sigma_R^2} \right)}} \left\{ \frac{1}{\Theta} \frac{1}{2} \frac{z_p}{\sqrt{\frac{2}{\sigma_R^2}} \left( n - (n-1) \frac{\sigma_R^2}{\sigma_R^2} \right)} \right\}^{1/2}.

To reiterate, $\gamma_p$ is a one-tailed 100$p\%$ upper limit for future sample $RSD_R$ values, $\Theta = \frac{\sigma_L}{\sigma_R}$ (the ratio of the population repeatability and reproducibility standard deviations), $\frac{\sigma_R}{\mu}$ (the population relative reproducibility standard deviation), $z_p$ (the abscissa on the standard normal curve that cuts off an area $p$ in the upper tail), and $L$ and $n$ are the number of laboratories and replicates/laboratory, respectively.

**Accuracy of $\gamma_p$**

To assess the accuracy of $\gamma_p$ with respect to the intended probability level, a Monte Carlo (MC) simulation study was conducted (see Appendix for details). The MC simulation was developed for use with Statistical Analysis System (SAS) software to model a CRM ANOVA assuming $L$ laboratories and $n$ replicates/laboratory to draw a set of simulated data, assuming known laboratory-to-laboratory and within-laboratory standard deviations ($\sigma_L$ and $\sigma_R$), respectively, and population mean ($\mu$) or concentration of analyte. The simulated data were then used to obtain an estimate of the sample relative reproducibility standard deviation ($RSD_R$). For each set of $\sigma_L$, $\sigma_R$, and $\mu$, the cumulative distribution of a total of 10 000 simulated sample relative reproducibility standard deviations was examined to obtain the 95th and 99th percentile values to represent simulated one-tailed 95 and 99% upper limits for future sample relative reproducibility standard deviations.

The results of the simulation are presented in Table 1 for values of $\xi_R$, $\% = 2$, 16, and 64; $\Phi = 1/2$ and 2/3; number of laboratories = 8 and 20; number of replicates = 2, 5, and 20; and probability levels of 95 and 99%. In general, Table 1 presents one-tailed 95 and 99% upper limits in percent ($\gamma_{0.95,\%}$) and ($\gamma_{0.99,\%}$) for future sample $RSD_R$ obtained in a collaborative study employing $L = 8$ and $L = 20$ laboratories, each performing 2, 5, or 20 replicates. Also presented in Table 1 are the MC simulated one-tailed 95 and 99% upper limit values ($MC_{0.95,\%}$ and $MC_{0.99,\%}$). The probability levels ($p$) are simulated probability levels that are equivalent to percentiles for the simulated MC values that equal the ($\gamma_{0.95,\%}$) and ($\gamma_{0.99,\%}$) values.

Based on the results in Table 1, it can be seen that there is excellent agreement between the $MC_{0.95,\%}$ and $MC_{0.99,\%}$ values and corresponding $p$-values. Hence, the computational formula ($\gamma_p$) provides a satisfactory approximation for obtaining a 100$p\%$ one-tailed upper limit for future sample $RSD_R$ values.

**Determining $\gamma_p$**

**Consensus Values Assumed for Population Values for $\xi_R$, $\%$ and $\Theta$**

Usually, the population values for $\xi_R$, $\%$ and $\Theta$ will not be known. However, in some cases, consensus values, i.e., values obtained on the basis of long-time experience, may be satisfactory approximations. For some analytical methods and materials, consensus values for $\xi_R$, $\%$ and $\Theta$ may be obtained from the results of research by Horwitz and Albert (7, 8).

For example, one might use the “Horwitz equation” to predict a consensus value ($\xi_{R,C}$) for the population percent relative reproducibility standard deviation ($\xi_R$, $\%$). The predicted relative reproducibility standard deviation expressed as a percent ($PRSD_R$, $\%$) is computed as $\xi_{R,C} \% = PRSD_R \% = 2^{0.150}$ using for a known spike or a consensus level of analyte to provide a consensus value for ($\xi_{R,C}$).

To obtain a consensus value for $\Theta = \frac{\sigma_L}{\sigma_R}$, one might appeal to Horwitz’s conclusion based on his observation of several thousand historic collaborative studies (7, 8). That is, Horwitz
Table 1. Comparison of simulated one-tailed 95 and 99% upper limits (MC_{95%} and MC_{99%}) and calculated one-tailed 95 and 99% upper limits (γ_{95%} and γ_{99%}) for future sample percent relative reproducibility standard deviations

<table>
<thead>
<tr>
<th>No. reps^d</th>
<th>Probability level, %</th>
<th>(MC_{95%}^e)</th>
<th>(γ_{95%}^f)</th>
<th>(p*)^g</th>
<th>(MC_{99%}^e)</th>
<th>(γ_{99%}^f)</th>
<th>(p*)^g</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2/3</td>
<td>8</td>
<td>2</td>
<td>2.76</td>
<td>2.78</td>
<td>0.955</td>
<td>3.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>2.69</td>
<td>2.71</td>
<td>0.955</td>
<td>3.05</td>
<td>3.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20</td>
<td>2.67</td>
<td>2.67</td>
<td>0.950</td>
<td>3.04</td>
<td>2.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20</td>
<td>2.47</td>
<td>2.47</td>
<td>0.950</td>
<td>2.68</td>
<td>2.67</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>2.43</td>
<td>2.43</td>
<td>0.950</td>
<td>2.63</td>
<td>2.61</td>
<td>0.990</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>2.41</td>
<td>2.41</td>
<td>0.950</td>
<td>2.60</td>
<td>2.58</td>
<td>0.988</td>
</tr>
<tr>
<td></td>
<td>2/3</td>
<td>8</td>
<td>2</td>
<td>2.68</td>
<td>2.71</td>
<td>0.951</td>
<td>3.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>2.59</td>
<td>2.59</td>
<td>0.950</td>
<td>2.92</td>
<td>2.83</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20</td>
<td>2.52</td>
<td>2.51</td>
<td>0.949</td>
<td>2.80</td>
<td>2.73</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>2</td>
<td>2.44</td>
<td>2.43</td>
<td>0.949</td>
<td>2.62</td>
<td>2.61</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>2.36</td>
<td>2.36</td>
<td>0.950</td>
<td>2.54</td>
<td>2.50</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>2</td>
<td>2.31</td>
<td>2.31</td>
<td>0.950</td>
<td>2.47</td>
<td>2.44</td>
</tr>
<tr>
<td>16</td>
<td>2/3</td>
<td>8</td>
<td>2</td>
<td>22.32</td>
<td>22.51</td>
<td>0.955</td>
<td>25.44</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>21.83</td>
<td>21.95</td>
<td>0.952</td>
<td>24.99</td>
<td>24.51</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20</td>
<td>21.44</td>
<td>21.65</td>
<td>0.955</td>
<td>24.36</td>
<td>24.09</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>2</td>
<td>19.82</td>
<td>19.92</td>
<td>0.954</td>
<td>21.83</td>
<td>21.59</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>19.60</td>
<td>19.59</td>
<td>0.950</td>
<td>21.25</td>
<td>21.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20</td>
<td>19.39</td>
<td>19.41</td>
<td>0.951</td>
<td>21.08</td>
<td>20.85</td>
</tr>
<tr>
<td></td>
<td>2/3</td>
<td>8</td>
<td>2</td>
<td>21.81</td>
<td>21.94</td>
<td>0.953</td>
<td>24.77</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>21.01</td>
<td>20.93</td>
<td>0.948</td>
<td>23.25</td>
<td>23.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20</td>
<td>20.47</td>
<td>20.35</td>
<td>0.946</td>
<td>22.79</td>
<td>22.22</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>2</td>
<td>19.59</td>
<td>19.59</td>
<td>0.950</td>
<td>21.35</td>
<td>21.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>18.97</td>
<td>18.89</td>
<td>0.950</td>
<td>20.42</td>
<td>20.25</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20</td>
<td>18.66</td>
<td>18.63</td>
<td>0.949</td>
<td>19.90</td>
<td>19.75</td>
</tr>
<tr>
<td>64</td>
<td>2/3</td>
<td>8</td>
<td>2</td>
<td>108.31</td>
<td>109.45</td>
<td>0.954</td>
<td>142.35</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>103.31</td>
<td>105.86</td>
<td>0.957</td>
<td>131.18</td>
<td>135.24</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20</td>
<td>100.96</td>
<td>104.03</td>
<td>0.955</td>
<td>126.23</td>
<td>131.70</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>2</td>
<td>88.26</td>
<td>88.64</td>
<td>0.951</td>
<td>102.69</td>
<td>102.31</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>86.08</td>
<td>86.86</td>
<td>0.956</td>
<td>99.45</td>
<td>99.39</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20</td>
<td>85.10</td>
<td>85.93</td>
<td>0.954</td>
<td>96.75</td>
<td>97.89</td>
</tr>
<tr>
<td></td>
<td>2/3</td>
<td>8</td>
<td>2</td>
<td>102.82</td>
<td>105.28</td>
<td>0.955</td>
<td>135.37</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>97.41</td>
<td>98.97</td>
<td>0.956</td>
<td>119.42</td>
<td>121.87</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20</td>
<td>93.48</td>
<td>95.63</td>
<td>0.958</td>
<td>116.67</td>
<td>115.81</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>2</td>
<td>86.27</td>
<td>86.63</td>
<td>0.952</td>
<td>99.26</td>
<td>98.96</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>83.73</td>
<td>83.42</td>
<td>0.948</td>
<td>94.12</td>
<td>93.76</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20</td>
<td>81.41</td>
<td>81.68</td>
<td>0.952</td>
<td>90.96</td>
<td>90.99</td>
</tr>
</tbody>
</table>

---

a: Population percent relative reproducibility standard deviation.
b: \(\theta = \frac{\sigma_r}{\sigma_p}\) = Ratio of the population repeatability standard deviation to the population reproducibility standard deviation.
c: Number of laboratories.
d: Number of replicates/laboratory.
e: Monte Carlo simulated one-tailed 95 and 99% upper limits for future sample percent relative reproducibility standard deviations.
f: Calculated one-tailed 95 and 99% upper limits for future sample percent relative reproducibility standard deviations.
g: Simulated percentile corresponding to a simulated MC value that equals MC_{95%} or MC_{99%}.

Recommended to OMB by Committee on Statistics: 07-17-2013
Reviewed and approved by OMB: 07-18-2013
observed from his research that the estimate of $\Theta = \frac{\sigma_c}{\sigma_R}$, i.e., the ratio of the sample repeatability standard deviation to the sample reproducibility standard deviation ($\frac{s_r}{s_R}$) for most accepted methods ranged from 1/2 to 2/3 (i.e., 0.500 to 0.667).

Because for any $\xi_R,\%$ of $\gamma_p$ is at a maximum when $\Theta = 0.5$, relative to the $\gamma_p$ obtained when $\Theta = 0.667$, we recommend using Horwitz’s lowest observation limit of $\frac{s_r}{s_R} = 0.5$ as a consensus value for $\Theta$.

**Example 1**

In this example, we assume that a Study Director has no knowledge of $\xi_R,\%$ and $\Theta$ but would like to know the largest $RSD_{R,\%}$ that might be confidently obtained in a collaborative study on a given material having a specified concentration ($C$). Given the above, we will start by using the “Horwitz equation,” if analytically applicable, to predict a consensus value for the population percent relative reproducibility standard deviation as follows: $\xi_{R,C,\%} = PRSD_\% = 2C^{-0.1505}$ (using for $C$ a known spike or a consensus level of analyte) to provide a consensus value for $(\xi_R,\%)$. Assume that the spike level or consensus value for the concentration is $C = 5.1147 \times 10^{-5}$. Substituting the value for $C$ in $\xi_{R,C,\%} = PRSD_\% = 2C^{-0.1505}$, we obtained $\xi_{R,C,\%} = 8.84$. For use in calculations later, $\xi_{R,C,\%}$ will be converted to a decimal, i.e., $\xi_R = \frac{\xi_{R,C,\%}}{100} = \frac{8.84}{100} = 0.0884$.

Next, we assume that we want a 95% upper limit for future sample $RSD_{R,\%}$ values ($\gamma_{0.95}$) obtained from a collaborative study employing $L = 8$ laboratories each analyzing duplicates ($n = 2$). We assume further a consensus value of $\Theta = 0.5$. Upon substituting the special case values $L = 8$, $n = 2$, $\Theta = 0.5$, and $\xi_{0.95} = 1.645$ (the standard normal deviate for $p = 0.95$) into

$$\gamma_p = \left[ \frac{\left(n-1\right)^2 + \left(n-\left(n-1\right)\frac{\xi_R^2}{\xi_{R,C}^2}\right)^2}{2nL} \right]^{1/2}$$

we obtained an easier-to-use formula for computing $\gamma_{0.95}$, given the above special case values as follows:

$$\gamma_{0.95} = \frac{\xi_R \left(1 + 1.645 \sqrt{0.05566 + 0.02993 \xi_R^2} \right)}{1 - 0.29597 \xi_R^2}$$

Substituting $\xi_{R,C} = 0.88398$ for $\xi_R$ in the previous general formula and performing the indicated mathematical operations, we obtained $\gamma_{0.95} = 0.12321$ or $\gamma_{0.95,\%} = 12.321$. This is the 95% upper limit for sample $RSD_{R,\%}$ arising from a population whose true mean percent relative reproducibility standard deviation is $\xi_{R,C,\%} = 8.84$.

Provided in the following is an easier-to-use formula for computing a 99% upper limit ($\gamma_{0.99}$) for future sample $RSD_{R,\%}$ values obtained from collaborative studies employing $L = 8$ laboratories each analyzing duplicates ($n = 2$). Here, we substituted the special case values $L = 8$, $n = 2$, $\Theta = 0.5$, and $\xi_{0.99} = 2.326$ (the standard normal deviate for $p = 0.99$) into $\gamma_p$ above, and obtained the following:

$$\gamma_{0.99} = \frac{\xi_R \left(1 + 2.326 \sqrt{0.05566 + 0.07644 \xi_R^2} \right)}{1 - 0.59175 \xi_R^2}$$

**Example 2**

Those familiar with the results from the “Horwitz equation” or predicted relative reproducibility standard deviation, $PRSD_{R,\%}$, may recognize that the $\xi_{R,\%} = 2, 16$, and 64 in Table 1 coincide with $PRSD_{R,\%} = 2, 16$, and 64 when the concentrations $C = 10^0, 10^{-6}$, and $10^{-10}$, respectively, are used in $PRSD_{R,\%} = 2C^{-0.1505}$. This implies that $\gamma_p$ may also be used to obtain one-tailed 100$p%$ upper limits for future sample $RSD_R$ obtained from a population with known $RSD_R = PRSD_R$ using the “Horwitz equation.”

Figure 1 presents plots of $PRSD_{R,\%}$ and one-tailed 95 and 99% upper limits, assuming $L = 8$, $n = 2$, and $\Theta = 0.5$, for future sample $RSD_{R,\%}$ on predefined concentrations transformed to $log_{10}(C)$. In Figure 1, the lower curve represents a plot of the $PRSD_{R,\%}$ values on $log_{10}(C)$ of analyte. This curve is called the “Horwitz curve.” The 2 upper curves reflect, respectively, one-tailed 95 and 99% upper limits for future sample $RSD_{R,\%}$ values.
Figure 1 appears to suggest that if one were to use the 95% _U_Lim or 99% _U_Lim values to define method acceptability, when the variability is higher, usually for low concentrations, the limits are wider, as they should be, allowing a greater degree of leniency for a method to be classified as acceptable than when the variability is lower for the higher concentrations.

**Summary**

A formula was developed for use in computing an upper limit for future sample relative reproducibility standard deviations obtained using a given method to analyze a given material in a collaborative study. This formula, and to a degree the results in Table 1, will prove useful to Study Directors in the design of collaborative studies because they can use the formula calculations or the results in Table 1 as a barometer for the worst that can be expected, given a specified level of confidence, with respect to reproducibility precision prior to conducting a study. The one drawback in using the formula is that it assumes that the relative reproducibility standard deviation and the ratio of the repeatability standard deviation to the reproducibility standard deviation are known population parameters. However, in practice this assumption may be relaxed by accepting and using the research results by Horwitz and Albert (7, 8) with respect to reproducibility precision. The results of that research, particularly that relating to the "Horwitz equation," appear useful for obtaining reproducibility precision consensus values for the above mentioned parameters that are generally accepted as standards.

**Acknowledgements**

The authors are grateful to Robert Blodgett (FDA/CFSAN, College Park, MD) for assistance in developing the SAS simulation procedure. In addition, we thank the referee for comments, which have assisted in improving the paper.

**References**


**Appendix**

The following Statistical Analysis System (SAS) program was written and executed to obtain a simulated distribution of sample RSD _R_ values. It is an unabridged version of the program used to generate the simulation results presented earlier.

*SAS Program to Determine a One-Tailed 100p% Upper Limit for Future Sample Relative Reproducibility Standard Deviations*

```sas
DATA FSIM (KEEP=X LAB I RHO N_LABS REPS); /* NEEDED FOR GLM*/
  ARRAY XG{N_LABS} XG1 - XG&N_LABS;
  ARRAY SLGP{N_LABS} SLGP1 - SLGP&N_LABS;
  %LET TEST = 10000; /*INPUT NUMBER OF SAMPLE RSD_R SIMULATIONS*/
  %LET N_LABS = 8; /*INPUT NUMBER OF LABORATORIES*/
  %LET REPS= 2; /*INPUT NUMBER OF REPLICATES*/
  %LET C = 1; /*INPUT VALUE FOR CONCENTRATION LEVEL*/
  %LET XI_R = .02; /*INPUT CONSENSUS VALUE FOR POP. */
  %LET XI_R = .02; /*INPUT CONSENSUS VALUE FOR POP. */
  %LET THETA = 0.5; /*INPUT Theta = \sigma_{R}/\sigma_{L} */
  %LET RHO = 1 - &THETA**2; /**ICC CALC.**/;
  SIG_L = SQRT((&C.*&XI_R.)**2 - (&THETA.*&XI_R.*&C.)**2); /*LAB STD*/
  SIG_R = &THETA*RHO; /*REPEATABILITY STANDARD DEVIATION*/
  N_LABS = &N_LABS;
  REPS = &REPS;
  DO I = 1 TO &TEST;
    %LET TEST = 10000; /*INPUT NUMBER OF SAMPLE RSD_R SIMULATIONS*/
    %LET N_LABS = 8; /*INPUT NUMBER OF LABORATORIES*/
    %LET REPS= 2; /*INPUT NUMBER OF REPLICATES*/
    %LET C = 1; /*INPUT VALUE FOR CONCENTRATION LEVEL*/
    X = &C + SLGP{I} + SIG_R*RANNOR(0); /*REPLICATE SELECTION*/
    OUTPUT FSIM;
  END;
RUN;
PROC GLM DATA=FSIM NOPRINT OUTSTAT=STATS;
   CLASSES LAB;
   MODEL X= LAB;
   BY I;
   PROC GLM DATA=FSIM NOPRINT OUTSTAT=STATS;
   CLASSES LAB;
   MODEL X= LAB;
   BY I;
END;
END;
RUN;
PROC GLM DATA=FSIM NOPRINT OUTSTAT=STATS;
   CLASSES LAB;
   MODEL X= LAB;
   BY I;
RUN; QUIT;
```

33

Recommended to OMB by Committee on Statistics: 07-17-2013
Reviewed and approved by OMB: 07-18-2013
DATA ALL; SET STATS;
RETAIN MS_ERR DF_ERR MS_LAB DF_LAB S_R S_RR ;
IF _TYPE_ = 'SS1' THEN DELETE;
IF _SOURCE_ = 'ERROR' THEN DO;
    MS_ERR = SS/DF;
    DF_ERR = DF;
END;
IF _SOURCE_ = 'LAB' THEN DO;
    MS_LAB = SS/DF;
    DF_LAB = DF;
    IF MS_LAB < = MS_ERR THEN SIGMA2_L = 0;
    ELSE;
        SIGMA2_L = (MS_LAB - MS_ERR)/&REPS;
        S_R = SQRT(MS_ERR);
        S_RR = SQRT(MS_ERR + SIGMA2_L);
    OUTPUT;
END;
RUN;
PROC MEANS NOPRINT DATA=FSIM;
   BY I;
   VAR X ;
   OUTPUT OUT=A N=N MEAN= XBAR;
RUN;
DATA AB; SET A;
   N_LABS = &N_LABS;
   REPS = &REPS;
   DROP _TYPE_ _FREQ_; 
RUN;
DATA VV; MERGE AB ALL;
   RSD_R = ROUND(100*(S_R/XBAR),.01);
   RSD_RR = ROUND(100*(S_RR/XBAR),.01);
   THETA1 = S_R/S_RR;
RUN;
PROC SORT DATA=VV;
   BY RSD_RR;
RUN;
PROC FREQ DATA = VV;
TABLES RSD_RR;
RUN;
DATA D; SET VV;
   LOG10_MU = LOG10(&C);
   POP_THETA = &THETA;
   POP_RSD = &XI_R.; 
KEEP N_LABS REPS PCTILE POP_RSD LOG10_MU RSD_RR POP_THETA;
DO PCTILE = .99, .95, .90, .80, .70, .60, .50, .40, .30, .20, .10, .05, .01; 
    J=CEIL(PCTILE*&TEST);
    SET VV POINT=J;
    OUTOUT D;
END;STOP;
RUN;
PROC PRINT DATA=D NOOBS;
   VAR POP_RSD LOG10_MU POP_THETA N_LABS REPS RSD_RR PCTILE;
RUN;