This series of knowledge sharing articles is a project of the Standardized Biofilm Methods Laboratory in the CBE

KSA-SM-10

(*this version, 2012-09-21, further clarifies and provides references for the similarity of REML and MOM results on p. 5*)

Assessing Resemblance, Repeatability, and Reproducibility for quantitative methods

[*Key Words*: analysis of variance, ANOVA, random effect, variance component]

Of the eight desirable attributes of a standardized method given in KSA-SM-3, three are the focus of this article: *resemblance* of the untreated control data, *repeatability* of the response of interest across multiple tests, and *reproducibility* of the response across multiple laboratories. These attributes can be assessed by considering three *random effects* on the response: among-carrier differences within each test; among-test differences in each lab (e.g. tests are performed on different days with different reactors and inoculums); and among-lab differences (e.g. labs are in different geographic locations and use different equipment). The magnitude of each of these effects on the response is quantified by a *variance component*, or a standard deviation (SD), which is the square root of a variance component. By estimating the SD associated with each random effect, a multiple-laboratory study can provide an assessment of resemblance, repeatability, and reproducibility. A study in a single lab can provide a limited assessment of only resemblance and repeatability.

In this article, we describe how to calculate the SDs necessary to assess the qualities of resemblance, repeatability and reproducibility when the responses of interest are quantitative. Assessments of semi-quantitative methods (see KSA-SM-2 and KSA-SM-8) will be addressed in a separate article. We will focus on disinfectant tests for which the response of interest is the log reduction (LR, see KSA-SM-7). KSA-SM-3 established notation for each of the SDs that we will focus on, and gave historically acceptable values for each (see Table 1 below). When validating a disinfectant test method, repeatability and reproducibility are among the most important considerations (Bloomfield and Looney, 1992).

Table 1. Three desirable attributes of a standardized method; the statistical measures used to assess the attribute; and the historical upper bound for acceptability (KSA-SM-3).

Desirable Attribute	Statistical measure	Symbol	HistoricallyAcceptable Upper Bound
Resemblance of the untreated controls	Within-test SD	CS	--
	Among-test SD	CS_{test}	--
	Repeatability SD	CS_r	0.5
	Among-lab SD	CS_{lab}	--
	Reproducibility SD	CS_R	0.7
Repeatability of the LR	Within-test SD	S	--
	Among-test SD	S_{test}	--
	Repeatability SD	S_r	1.0
	Among-lab SD	S_{lab}	--
Reproducibility of the LR	Reproducibility SD	S_R	1.3

It is important to consider "labs" and "tests" as random effects since a researcher is not really interested in how a method performs during one test at a specific lab, but rather how the method will perform in a randomly chosen test at a randomly chosen lab. The random effects are *nested* [1]: among-carrier effects are nested within each test or experiment; among-test effects are nested within each laboratory; the top-most level is the among-lab effect. Thus, the statistical models which quantify the random effects (by estimating the SDs in Table 1) are called *nested random effects analysis of variance* (ANOVA) *models* (Neter et. al., 1996). The mathematical equations for these models are given in the Appendix for the interested reader. We focus on the application and interpretation of the output from these models in the rest of this article.

Resemblance

An assessment of resemblance provides information about the "typical" bio-challenge posed by a test method. In many cases, the response of interest when measuring the bio-challenge for a test is the log density of organisms on each untreated control carrier. Two statistics used to assess resemblance are the mean and the standard deviation (SD) for the untreated control log densities. These statistics are calculated differently for each of the following scenarios: multiple carriers from a single test; multiple tests in a single laboratory; and multiple tests conducted in multiple laboratories.

For a single test, we will use *TestLD* to denote the mean of the *J* untreated carrier log densities, calculated by $TestLD = \frac{1}{J}\sum_{j=1}^{J} C_j$. The response of interest is C_j, the LD for the j^{th} untreated control carrier, j = 1, 2, …, *J*. The resemblance of the carriers in the single test is quantified by the *within-test SD* for the untreated control carriers, calculated by

$$CS = \sqrt{\frac{1}{J-1}\sum_{j=1}^{J}(C_j - TestLD)^2}. \quad (1)$$

The value for *CS* is interpreted as the typical distance of the LD for a randomly chosen untreated control carrier from the true *Test LD* for a given test. Measuring more carriers in the same test (called *pseudo-replication*) only increases one's knowledge of the value of the within-test SD *CS*, and does not provide a better estimate of the true variability of the untreated controls across multiple tests and labs.

Across multiple tests or experiments in a single laboratory, the mean bio-challenge is calculated for each test. If we use $TestLD_k$ to denote the mean of the *J* untreated control log densities for the k^{th} test out of a total of *K* tests, then the mean bio-challenge across all *K* tests is $LabLD = \frac{1}{K}\sum_{k=1}^{K} TestLD_k$. The resemblance of the tests is quantified in this case by the *resemblance repeatability SD*, CS_r, which is the standard deviation of all of the $TestLD_k$*'s* from that laboratory. Alternatively, CS_r can be calculated by fitting the one-factor random effects ANOVA model described in the Appendix. The ANOVA provides estimates of two variance components: the within-test variance CS^2 (given in equation (1) when calculated from a single test), and the variance among tests, CS^2_{test}. Using this ANOVA output, CS_r can be found by

$$CS_r = \sqrt{\frac{CS^2}{J} + CS^2_{test}}. \quad (2)$$

The value for CS_r is interpreted as the typical distance of the *TestLD* for a randomly chosen test from the true mean *TestLD* across all tests in that lab (which could also be called the true *LabLD* for the lab).

[1] A formal statistical definition can be found at http://www.itl.nist.gov/div898/handbook/pri/section7/pri7.htm

Performing more tests in the same lab only increases one's knowledge of the value of the within-lab SD CS_r, and does not provide a better estimate of the true variability of the untreated controls across multiple labs. Small CS_r values indicate good resemblance of the untreated control carriers within the lab (see Table 1).

For studies involving multiple laboratories (L), the mean bio-challenge is calculated for each laboratory. If we use $LabLD_l$ to denote the mean of the K *TestLDs* from the l^{th} lab, with each *TestLD* calculated from J untreated control log densities, then the mean bio-challenge across all L tests is $OverallMeanLD = \frac{1}{L}\sum_{l=1}^{L} LabLD_l$. The reproducibility of the untreated controls across the multiple labs is assessed by the *resemblance reproducibility SD*, CS_R, which can be calculated by fitting the two-factor random effects ANOVA model described in the Appendix. The ANOVA provides estimates of three variance components, the variance within-tests (CS^2, given in equation (1) when calculated from a single test), the variance among tests (CS^2_{test}), and the variance among laboratories (CS^2_{lab}). Based on this ANOVA output, when each test is conducted with J untreated control carriers, then CS_R is calculated by

$$CS_R = \sqrt{\frac{CS^2}{J} + CS^2_{test} + CS^2_{lab}} \ . \tag{3}$$

The value of CS_R is interpreted as the typical distance of the *TestLD* for a randomly chosen test in a randomly chosen lab from the true mean *TestLD* <u>for all</u> labs (which could also be called the true *OverallMeanLD*). Note the extra variance component due to labs, when comparing equation (3) to (2). Small CS_R values indicate good resemblance of the untreated controls across multiple labs (see Table 1).

Repeatability

The response of interest when measuring disinfectant efficacy is the LR. An assessment of the repeatability of the LRs across multiple tests performed in the same lab requires two statistics, the mean and standard deviation (SD) of the LRs.

The LR for a single test is found by calculating $\mathrm{LR} = TestLD - \overline{T}$ where $\overline{T}$ is the mean of the LDs for I treated carriers, and *TestLD* is the mean for the J untreated carriers. Let *TS* denote the *treated within-test SD*, defined by

$$TS = \sqrt{\frac{1}{I-1}\sum_{i=1}^{I}(T_i - \overline{T})^2},$$

where T_i denotes the LD for the i^{th} treated carrier. Now the within-test SD of the LR (Zelver et al., 2001) can be given as

$$S = \sqrt{\frac{CS^2}{J} + \frac{TS^2}{I}} \tag{4}$$

where *CS* is given in equation (1). For a single test or experiment, the primary responses are LR and *S*.

Across K multiple tests in a given laboratory, let $\mathrm{LR_k}$ be the LR from the k^{th} test. The mean response of interest is $\overline{LR}$, the mean LR over all K tests. When I treated carriers and J untreated control carriers are used in each test, then the repeatability of the LR across the K tests is quantified by the *repeatability SD*, which is the SD of the K LR values,

$$S_r = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}(LR_k - \overline{LR})^2} \ . \tag{5}$$

For a given lab, the value for S_r is interpreted as the typical distance of the LR for a randomly chosen test from the true mean LR. Performing more tests in the same lab only increases one's knowledge of the value of the within-lab SD S_r, and does not provide a better estimate of the true variability of the LRs across multiple labs. Small S_r values indicate good repeatability of the LR (see Table 1).

Reproducibility

Across multiple laboratories (L), let $\overline{LR}_l$ be the mean LR over all K tests performed at the l^{th} lab. The response of interest is the mean LR over all L laboratories, calculated by $OverallMeanLR = \frac{1}{L}\sum_{l=1}^{L}\overline{LR}_l$.

The reproducibility of the LR across the L labs can be calculated by fitting the one-factor random effects ANOVA model described in the Appendix. The ANOVA provides estimates of two variance components: the variance among labs, S^2_{lab}, and the within-lab variance of the LR, S_r^2. The value of S_r is called the repeatability SD pooled across all L labs; equation (5) shows how S_r^2 is calculated for a single lab. Based on these variance components from an ANOVA, S_R is calculated by (Mandel 1998)

$$S_R = \sqrt{S_r^2 + S_{lab}^2}\ . \tag{6}$$

The value of S_R is interpreted as the typical distance of the LR for a randomly chosen test from the true mean LR <u>for all</u> labs. Small values of S_R indicate good reproducibility (see Table 1).

Examples

Let us assess the resemblance, repeatability and reproducibility of tests conducted using the quantitative three step method (AOAC official method 2008.05 (2008)) using spores of *Bacillus subtilis* on glass carriers as described by Tomasino et al. (2008, data is in Appendix 3). In this collaborative study, each of L = 8 labs performed K = 9 tests, with each test using J = 3 untreated control carriers and I = 3 treated carriers.

Resemblance

To assess the resemblance of the untreated control LDs, the variance components presented in Table 2, calculated by a 2-way random effects ANOVA, are required.

Table 2. The ANOVA results for the untreated control carrier LDs for the tests described by Tomasino et. al. (2008).

Source	Estimated Variance	Symbol
lab	0.04899	CS^2_{lab}
test in lab	0.01607	CS^2_{test}
within-test	0.02097	CS^2

Substituting the estimated variances from Table 2 into equation (2), the repeatability SD for the untreated control mean LD for J = 3 untreated carriers per test is

$$CS_r = \sqrt{\frac{0.02097}{3} + 0.01607} = 0.152.$$

Thus, in a given lab, the mean of the 3 untreated control carriers in a randomly chosen test is typically about 0.152 from the true mean control LD for <u>that lab</u>. Using equation (3), the reproducibility SD is

$$CS_R = \sqrt{\frac{0.02097}{3} + 0.01607 + 0.04899} = 0.268.$$

Hence, the untreated control mean LD for a randomly chosen test in a randomly chosen lab is typically about 0.268 from the true overall mean LD for all labs. Using the acceptance criteria given in Table 1, the quantitative three step method exhibited acceptable resemblance across multiple tests in each lab, and acceptable resemblance across multiple labs.

It is common to report the variance components CS^2/J, CS^2_{test} , and CS^2_{lab} as proportions of the total variance CS^2_R. For example, since

$$CS^2_{lab}/CS^2_R = 0.04899/0.07205 = 0.6799,$$

$CS^2_{test}/CS^2_R = 0.2230$ and $(CS^2/3)/CS^2_R = 0.0970$, then one would report that 68% of the variability of the *TestLDs* is due to among-lab sources, 22% is due to among-test sources in each lab, and the final 10% of variance is due to among-carrier sources. Similar calculations, $(CS^2/J)/CS^2_r$ and CS^2_{test}/CS^2_r, can be calculated for a single lab study.

Repeatability and Reproducibility

Of the many tests performed by each lab in the collaborative study described by Tomasino et. al. (2008), multiple disinfectants and efficacy levels were considered. This example uses the results from $K = 3$ tests of glutaraldehyde at a "low efficacy level." To assess the repeatability and reproducibility of the LR in this case, we will use the variance components presented in Table 3, which were calculated using one-way random effects ANOVA.

Table 3. The ANOVA results for the LRs for glutaraldehyde at a low efficacy level in tests described by Tomasino et. al. (2008).

Source	Estimated Variance	Symbol	Percentage
lab	0.0894	S^2_{lab}	75%
within-lab	0.0293	S^2_r	25%

The repeatability SD for the LR is $S_r = \sqrt{0.0293} = 0.17$. Thus, in a given lab, the LR for a randomly chosen test is typically about 0.17 from the true mean LR for that lab. Using equation (6), the reproducibility SD is $S_R = \sqrt{0.0293 + 0.0894} = 0.34$. Hence, the LR for a randomly chosen test in a randomly chosen lab is typically about 0.34 from the true overall mean for all labs. Using the acceptance criteria given in Table 1, the quantitative three step method exhibited acceptable repeatability of the LR for a low efficacy treatment of glutaraldehyde across multiple tests in each lab, and acceptable reproducibility of the LR across multiple labs.

Although this example focused on the quantitative three step method (AOAC official method 2008.05), the above steps can be applied to assess the attributes of resemblance, repeatability (with-in a lab), and reproducibility (across multiple labs) for any quantitative method.

Appendix: ANOVA models

This appendix presents mathematical equations for the ANOVA models described in this article.

Two alternative statistical procedures are commonly used for fitting ANOVA models, thereby estimating the variances for the random effects: the method of moments (MOM) and the restricted maximum likelihood (REML) method. The statistical software Minitab implements MOM; package *nlme* in R (Pinheiro et al., 2009; R Development Core Team, 2010) implements REML. Both MOM and REML give the same results for balanced data when the MOM estimates are positive. Even for balanced data with very small variance components, it is not uncommon for MOM variance component estimates to be

negative. REML is recommended for unbalanced data (Pinheiro & Bates, 2000; Searle et al., 1992, in which the MOM is called the ANOVA method). Examples showing how to use Minitab and R, including how to check relevant assumptions, will be provided in a future KSA.

Resemblance model of the untreated control LDs in one lab

Let LD_{jk} denote the log density for the j^{th} untreated control carrier in the k^{th} replicate test in a single laboratory. The 1-factor random effects ANOVA is

$$LD_{jk} = \mu + \beta_k + \varepsilon_{jk},$$

where μ is the true mean of the control log densities at the single lab, β_k is the random effect due to the k^{th} test, and ε_{jk} is the random effect due to the j^{th} replicate control carrier in the k^{th} test. The analysis requires that β_k and ε_{jk}, for all j and k, are independent normal random variables having means of zero. The estimated variance of β_k is CS^2_{test}, the variance among tests, and the estimated variance of ε_{jk} is CS^2, the within-test variance. The estimate of μ is the overall mean LD for untreated control carriers over all tests in the one lab.

Resemblance model of the untreated control LDs in multiple labs

Let LD_{jkl} denote the log density for the j^{th} untreated control carrier in the k^{th} test performed at the l^{th} laboratory. The 2-factor, nested, random effects ANOVA is

$$LD_{ijk} = \mu + \gamma_l + \beta_{k(l)} + \varepsilon_{jkl},$$

where μ is the true mean LD across <u>all labs</u>, γ_l is the random effect due to the l^{th} laboratory, $\beta_{k(l)}$ is the nested random effect due to the k^{th} test in the l^{th} laboratory, and ε_{jkl} is the nested random effect due to the j^{th} carrier in the k^{th} test in the l^{th} laboratory. The analysis requires that γ_l, $\beta_{k(l)}$, and ε_{jkl}, for all j, k, and l, are independent normal random variables having means of zero. The estimated variance of γ_l is CS^2_{lab}, the variance among laboratories; the estimated variance of $\beta_{k(l)}$ is CS^2_{test}, the variance among tests within a laboratory; and the estimated variance of ε_{jkl} is CS^2, the variance among untreated control carriers within a test. The estimate of μ is the overall mean log density for untreated carriers over all tests and labs.

Repeatability model of the LR in one lab

Let LR_k denote the LR for the k^{th} test in a single laboratory. The ANOVA model is

$$LR_k = \mu + \varepsilon_k,$$

where μ is the true mean LR at the single lab, and ε_k is the random effect due to the k^{th} replicate test. The analysis requires that ε_k is a normal random variable with a mean of zero. The estimated variance of ε_k is S_r^2, the repeatability variance within a laboratory. The estimate of μ is the mean LR over all tests in the one lab.

Reproducibility model of the LR in multiple labs

Let LR_{kl} denote the LR for the k^{th} test in the l^{th} laboratory. The one-factor, random effects ANOVA model is

$$LR_{kl} = \mu + \gamma_l + \varepsilon_{kl},$$

where μ is the true mean LR over <u>all labs</u>, γ_l is the random effect due to the l^{th} laboratory, and ε_{kl} is the random effect due to the k^{th} test in the l^{th} laboratory. The analysis requires that γ_l and ε_{kl}, for all k and l, are

normal independent random variables having means of zero. The estimated variance of γ_l is S^2_{lab}, the variance among laboratories, and the estimated variance of ε_{kl} is S_r^2, the repeatability variance within a lab. The estimate of μ is the overall mean LR over all tests and labs

References

AOAC International (2008) Official Method 2008.05: *Efficacy of Liquid Sporicides against Spores of Bacillus subtilis on a Hard Nonporous Surface Quantitative Three Step Method (First Action)*. AOAC International, Gaithersburg, MD.

Bloomfield, S. F. and Looney, E. (1992) Evaluation of the repeatability and reproducibility of European suspension test methods for antimicrobial activity of disinfectants and antiseptics. *J. Applied Bacteriology* 73:87-93.

Bloomfield, S. F., Arthur, M., Van Klingeren, B., Pullen, W., Holah, J. T., and Elton, R. (1994) An evaluation of the repeatability and reproducibility of a surface test for the activity of disinfectants. *J. Applied Microbiology* 76:86-94.

Mandel, J. (1998) Interlaboratory study, pp 2073- 2076 in *Encyclopedia of Biostatistics*, Armitage, P. and Colton, T. (Eds), Wiley, New York.

Neter, J., Kutner, M.H., Wasserman, W., and Nachtsheim, C.J. (1996) *Applied Linear Statistical Models, 4th Ed.*, McGraw-Hill, Boston.

Pinheiro, J. C. and Bates, D. M. (2000) *Mixed Effects Models in S and S-PLUS*. *Statistics and Computing Series*, Springer-Verlag, New York.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. and the R Core team (2009) nlme: Linear and Nonlinear Mixed Effects Models, R package version 3.1-96.

R Development Core Team (2010) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.\

Searle, S. R., Casella, G., and McColloch, C. E. (1992) *Variance Components*. Wiley, New York, NY.

Tilt, N. and Hamilton, M.A. (1999) Repeatability and reproducibility of germicide tests: a literature review. *J. AOAC Int.*, **82**, 384 – 389.

Tomasino, S.F., Pines, R.M., Cottrill, M.P., and Hamilton, M.A. (2008) Determining the efficacy of liquid sporicides against spores of *Bacillus subtilis* on a hard nonporous surface using the quantitative three step method: collaborative study. *J. AOAC Int*. 91(4), 833-852.

Zelver, N., Hamilton, M., Goeres, D., and Heersink, J. (2001) Development of a standardized antibiofilm test, in *Methods in Enzymology – Biofilms II*, Vol. 337, R.J. Doyle (Ed), Academic Press, New York, NY, pp 363-376.

Version date: 21 September 2012
Orig. publication date: 23 June 2011
Updated: 23 January 2012

Author: Albert E. Parker, parker@math.montana.edu
Martin A. Hamilton, mhamilton@biofilm.montana.edu