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

KSA-SM-15

Disinfectant test results: How to average across laboratories

[Key Words: estimator, LR, REML, SE, *TestLD*, unbalanced data, variance, weighted average]

A comprehensive assessment of a disinfectant product often entails multi-laboratory testing, in which case the individual test outcomes may be averaged across the laboratories. The article KSA-SM-13 (2013) showed how to average across laboratories for a study with balanced data. The term "balanced data" indicates that each laboratory conducted the same number of tests, whereas "unbalanced data" indicates that the number of replicate tests was not the same for all laboratories and may in fact be quite different (Searle et al. 1992). Most multi-laboratory disinfectant efficacy studies are purposely designed to produce balanced data, but even the best plan can go awry, such as when it belatedly is discovered that some tests were flawed, leaving the investigators with an unbalanced data set. Some studies are conducted with prior knowledge that the data will be unbalanced, e.g., a study based on aggregated test results from the data archives of several laboratories. For unbalanced data, several different averaging techniques have been proposed by statisticians (Cochran 1954; Iyer et al. 2004).

This article is a review of techniques for averaging across laboratories. We describe and compare popular averaging techniques, and recommend the one that seems most appropriate in the context of disinfectant testing. The recommendation is based on the statistical literature and our many years of experience in analyzing data from multi-laboratory studies. We assume that all disinfectant tests conducted by all laboratories were judged by the study director to follow the pre-specified test protocol closely enough to justify calculation of an overall average.

To illustrate the issues, we use two unbalanced data sets from multi-laboratory studies. The first is from a historical study of Use Dilution Method (UDM) test results, gathered from the archives of four laboratories (Tomasino et al. 2012). The UDM is a dried surface carrier test. The study was conducted to investigate the response quantity *TestLD*, the mean log density of viable cells for six untreated carriers (KSA SM-10 2012; Hamilton et al. 2013). The data pertain to tests conducted on *Pseudomonas aeruginosa* bacteria in the presence of an organic soil load. Table 1 shows the mean and the standard deviation (SD) of the *TestLD* values in each laboratory. In all, 185 UDM tests were conducted, with the number of tests per laboratory ranging from 36 to 62.

Table 1. Mean and SD of *TestLD* for each laboratory; UDM tests against *P. aeruginosa* in the presence of an organic soil load (Tomasino et al. 2012)

Lab index (i)	No. of Tests (n_i)	Mean *TestLD* (L_i)	SD of *TestLD* (S_i)
1	36	6.71293	0.29341
2	62	6.51515	0.27459
3	46	6.90142	0.22578
4	41	6.79364	0.24231

The second data set is from a designed collaborative study of the Quantitative Carrier Test (QCT-1) in 14 laboratories (Sattar et al. 2005). QCT-1 is a dried surface carrier test. The study was

conducted to evaluate the reproducibility of the log reduction (LR) disinfectant efficacy measure. We will use the data to calculate the overall average LR for a diluted quaternary ammonium compound when applied to *Bacillus subtilis* spores. Table 2 shows the mean and the standard deviation (SD) of the LR values in each laboratory. In all, 18 tests were conducted; 1 test in each of 10 laboratories, 2 tests in each of 4 laboratories.

Table 2. Mean and SD of LR for each laboratory; QCT-1 tests of a diluted quaternary ammonium compound applied to *B. subtilis* spores (Sattar et al. 2012)

Lab index (i)	No. of Tests (n_i)	Mean LR (L_i)	SD of LR (S_i)
1	2	4.495	0.00130
2	1	6.110	NA*
3	1	6.330	NA
4	1	3.550	NA
5	2	6.955	0.25200
6	2	7.350	1.84300
7	1	5.750	NA
8	1	6.430	NA
9	1	7.760	NA
10	1	5.340	NA
11	2	5.685	0.00005
12	1	7.760	NA
13	1	5.820	NA
14	1	4.910	NA

*NA is Not Available; $n_i = 1$

Two Averages: Mean of Laboratory Means (MLM) and Grand Mean (GM)

Let i denote the laboratory index. Let I denote the number of laboratories in the study. Note that I = 4 for Table 1 and I = 14 for Table 2. For laboratory i, let n_i denote the number of replicate tests. Let j denote the j^{th} replicate test in a laboratory. Let Y_{ij} denote the response quantity for replicate test j in laboratory i, where j = 1, 2, …, n_i and i = 1, 2, …, I. In the context of Table 1, Y is an observed *TestLD* value; in the context of Table 2, Y is an observed LR value. For laboratory i, L_i denotes the mean of Y_{ij} values across the n_i tests, and S_i denotes the SD of Y_{ij} values across the n_i tests. Let N denote the total number of tests across all laboratories, $N = \sum_{i=1}^{I} n_i$. Let **MLM** denote the mean of laboratory means, $\mathbf{MLM} = \sum_{i=1}^{I} L_i / I$. For Table 1, **MLM** = 6.73079, the arithmetic average of the four laboratory means. Let **GM** denote the grand mean of the N response quantities, $\mathbf{GM} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} Y_{ij} / N$. For Table 1, **GM** = 6.71140, the arithmetic average of the 185 *TestLD* values. For Table 2, **MLM** = 6.0175 and **GM** = 6.0406. Calculated averages for the examples are summarized in Table 3, along with a third average, **REMLM**, described below. The two obvious choices for calculating the average across laboratories (**MLM** and **GM**) yield different answers for unbalanced data. The difference depends on the extent to which the data are unbalanced; i.e., how much the n_i numbers differ. In neither of the examples are the data greatly unbalanced.

Table 3. Estimates & SEs for data in Tables 1 & 2; SE calculation formulas in Appendix equations (A3-A5)

Estimator	Estimate	SE(Estimate)
	Data in Table 1	
MLM	6.7308	0.08239
REMLM	6.7300	0.08238
GM	6.7114	0.08401
	Data in Table 2	
MLM	6.0175	0.32669
REMLM	6.0231	0.32560
GM	6.0406	0.33621

Which is better, MLM or GM?

The question that heads this section arises only when the data are unbalanced because **MLM** and **GM** are identical for balanced data. Let μ denote the true mean response across the entire population of laboratories. The goal is to find an accurate and precise estimator of μ. Both **MLM** and **GM** are accurate in the sense of being consistent, unbiased estimators of μ. However, the two estimators are not equally precise. The preferable estimator is the less imprecise one, i.e., the estimator that has a smaller variance (variance measures imprecision). To compare the variances of **MLM** and **GM**, we have developed a comparison tool, displayed as equation (1).

For the observed quantity *Y* (e.g., *TestLD* or LR), let σ_L^2 denote the true variance among laboratories and let σ_r^2 denote the true repeatability variance (**Appendix** equation (A1)). Let σ_{MLM}^2 and σ_{GM}^2 denote the true variances of the estimators **MLM** and **GM**, respectively. Recently, Levin and Leu (2013) published a formula for σ_{GM}^2 that makes the variance comparison relatively easy. From their work, we derive a quantity *Q*, calculated from the numbers of tests n_1, …, n_I (**Appendix** equation (A2)), with the following property:

$$\text{If } \sigma_r^2 < Q \cdot \sigma_L^2 \text{ then } \sigma_{MLM}^2 < \sigma_{GM}^2, \text{and conversely.} \qquad (1)$$

In words, **MLM** is preferable to **GM** if and only if *Q* times the variance among laboratories is larger than the repeatability variance.

For the n_i values in Table 1, $Q = 50.145$. For the data in Table 1, the estimated variance among laboratories is $S_L^2 = 0.025628$ and the estimated repeatability variance is $S_r^2 = 0.067695$ (Tomasino et al. 2012). Substituting these estimates for the corresponding parameters in equation (1), we find that $S_r^2 = 0.068 < 1.304 = Q \cdot S_L^2$, leading us to surmise that **MLM** is better than **GM** for the numbers of tests n_i for the Table 1 data.

For the n_i values in Table 2, $Q = 1.5556$. For the data in Table 2, the estimated variance among laboratories is $S_L^2 = 1.0494$ and the estimated repeatability variance is $S_r^2 = 0.51889$. Substituting these estimates for the corresponding parameters in equation (1), we find that $S_r^2 = 0.519 < 1.632 = Q \cdot S_L^2$, leading us to surmise that **MLM** is better than **GM** for the numbers of tests n_i for the Table 2 data.

The statement displayed in equation (1) is consistent with the long-known result that the **MLM** is a better estimator than **GM** except when the variance among the laboratories is nearly zero (Cochran 1954; Birkes et al. 1981). For the Table 1 example, if the variance among laboratories had turned out to be zero, that would indicate the *TestLD* values were not affected by what lab did the test. In that case, we could pool the data into one big sample and use the **GM** as the estimator; the effective sample size would be 185. However, most interlaboratory studies of disinfectant tests produce a significant interlaboratory variance, as for the Table 1 example, in which case **MLM** will be preferred over **GM**. Intuitively speaking, when the variance among laboratories is large, the information content of the data depends more on the number of laboratories than on the total number of tests. In the Table 1 example, the effective sample size is $I = 4$, the number of laboratories, not 195, the number of tests. Therefore, it makes sense to average the 4 laboratory means instead of averaging all 195 response quantities.

Body weight measurements provide a convenient analogy. When your scale produces an unexpected weight, you might reweigh yourself a second or even a third time. Then, realizing that further readings will little alter the outcome, you walk away (wearing a smile or a grimace). Probably, the next day your weight will be closer to what you expect. You know from experience

that day-to-day variability is much greater than the variability among repeated readings within a few minutes. To get a good idea of your weight, you will average over days, and not attach extra importance to the day when you weighed yourself multiple times.

An Alternative to MLM and GLM

There are many alternative averaging techniques, but we will discuss only the most commonly mentioned one, which is a specific weighted average. To calculate a weighted average, numerical weights are required; i.e., for each laboratory, a non-negative numerical weight must be specified by the analyst. Let w_i denote the weight for laboratory i. Then **WM**, the weighted average of the laboratory means, is

$$\mathbf{WM} = \frac{\sum_{i=1}^{I} w_i \cdot L_i}{\sum_{i=1}^{I} w_i}. \tag{2}$$

Note that **GM** and **MLM** can be expressed as weighted averages. If each weight equals 1 (or any other fixed positive number), then **WM** = **MLM**; if the weight for laboratory i is the corresponding number of tests, $w_i = n_i$, $i = 1, \ldots, I$, then **WM** = **GM** (Cochran 1954; Searle et al. 1992).

Statisticians have shown that there are optimal weights, "optimal" because they provide both the Best Linear Unbiased Estimate and the Maximum Likelihood Estimate (Cochran 1954; Piepho 1996). The variance of the optimal estimator is never larger than the variances of **GM** and **MLM**. The optimal weights, denoted by W_i, $i = 1, \ldots, I$, are displayed in equation (3) (Piepho 1996). The W_i weights differ only because the n_i values differ among laboratories.

$$W_i = \left[\sigma_L^2 + \frac{\sigma_r^2}{n_i}\right]^{-1}. \tag{3}$$

Unfortunately, the W_i weights are impractical because the numerical values of σ_L^2 and σ_r^2 are unknown. However, the variance parameters can be estimated from the observed data (KSA-SM-13 2013). Denote the observed variance estimates by S_L^2 and S_r^2. The variance estimates can be substituted for the corresponding parameters in equation (3), to produce $\widehat{W}_i$ weights of equation (4). If S_L^2 is near zero, then the **WM** based on $w_i = \widehat{W}_i$ is essentially the **GM**; if S_L^2 is large, then **WM** is essentially the **MLM**. In essence, the $\widehat{W}_i$-weighted **WM** blends **MLM** and **GM**, and it is the preferred estimator, especially when S_L^2 is neither zero nor "large" (Cochran 1954; Birkes et al. 1981).

$$\widehat{W}_i = \left[S_L^2 + \frac{S_r^2}{n_i}\right]^{-1}. \tag{4}$$

We recommend the restricted maximum likelihood (REML) method for calculating S_L^2 and S_r^2 (KSA-SM-13 2013; Hamilton et al. 2013). Let **REMLM** denote the REML mean; i.e., **REMLM** is the $\widehat{W}_i$-weighted **WM** in equation (2) with weights calculated by equation (4) using REML variance estimates (Rukhin et al. 2000; Kacker 2004)**.**

For the data of Table 1, **REMLM** is 6.72998 (Table 3). Note that, for the Table 1 data, the difference between **MLM** and **REMLM** is quite small, less than 0.001. Also the standard errors of the mean (**Appendix** equations (A3)-(A5)) are not different, SE(**MLM**) = 0.08239 and SE(**REMLM**) = 0.08238 (Table 3). For the Table 1 study, the large S_L^2 relative to S_r^2 and the degree of unbalance (as represented by Q) indicate that **MLM** is an acceptable estimator of μ.

For the data of Table 2, **REMLM** is 6.0231. Note that, for the Table 2 data, **REMLM** is between **MLM** and **GM**, but close to **MLM**. The standard errors of the estimates for **MLM** and **REMLM** are almost the same, SE(**MLM**) = 0.327and SE(**REMLM**) = 0.326 (Table 3). For the Table 2 data, **MLM** is easier to calculate and explain than **REMLM** and provides essentially the same result as **REMLM**. We believe that **MLM** is sufficiently reliable for use in estimating μ if one had insufficient computational resources for calculating **REMLM**.

Although, more computationally complex, the best analysis of unbalanced data is a REML method analysis of variance (Searle et al. 1992). Computer software can produce S_L^2, S_r^2, **REMLM**, and SE(**REMLM**). Some computational guidance is provided in the next section.

Calculating REMLM using the statistical programming language R

The section headed "Statistical calculations using the statistical programming language R" in KSA-SM-13 (2013) shows how to use the package **nlme** and the function **lme** to perform a REML method analysis. The computer code is the same for unbalanced data as for balanced data.

When the **lme** function is used to analyze the data set summarized in Table 1, using R-language computer code copied from KSA-SM-13 (2013), the `Fixed Effects` component of the output is:

```
Fixed effects: TestLD ~ 1
               Value  Std.Error  DF  t-value p-value
(Intercept) 6.729978 0.08238387 181 81.69049       0
```

When the **lme** function is used to analyze the data set summarized in Table 2, the `Fixed Effects` component of the output is:

```
Fixed effects: LR ~ 1
               Value Std.Error DF  t-value p-value
(Intercept) 6.023061 0.3255979 14 18.49846       0
```

In each case, the `Value` is the **REMLM** and the `Std.Error` is the SE(**REMLM**). It is a bonus that the KSA-SM-13 (2013) R code for calculating S_L^2 and S_r^2 also produces the most reliable estimate of the overall mean μ and the associated standard error of the estimate. Note that we have crossed out the `DF t-value p-value` results in the output. In our opinion, those values should be ignored. For t-statistic confidence interval calculations, it is appropriate to use I-1 for the degrees of freedom (`DF`); that is, we suggest `DF` = 3 for the Table 1 data and `DF` = 13 for the Table 2 data. This recommendation is somewhat conservative in that the confidence interval will be a little wider than would be produced by the generalized confidence interval procedure of Iyer et al. (2004) which is computationally intensive because of the necessary computer simulations.

Recommendations

We recommend estimating μ with the **REMLM** average which is automatically calculated by the REML analysis of variance calculations for unbalanced data (KSA-SM-13 2013). Among easily calculated estimates, **MLM**, the mean of laboratory means, is a reliable estimator of μ for practically any unbalanced, multi-laboratory, disinfectant test data set. The **MLM** is more precise than the **GM** except when there is a negligibly small variance among laboratories, a rare and unexpected circumstance because a multi-laboratory study is expensive and will be conducted only when prior information indicates that the variability among laboratories is large. Although this article is written for disinfectant test data, the issues, notation, and formulas apply to any data set that can be modeled with a one-factor random effects linear model.

Appendix

One-factor random effects linear model

The Levin and Leu (2013) formula (that we used to derive equation (1)), the **REMLM** estimate, and conventional analysis of variance calculations are based on a one-factor random effects linear model for multi-laboratory data. Let j denote the j^{th} replicate test in a laboratory. Let Y_{ij} denote the measured response for replicate test j in laboratory i, where j = 1, 2, …, n_i and i = 1, 2, …, I. In the context of our first example, Y is an observed *TestLD* value. The true, unknown mean response for laboratory i is denoted by λ_i and the true mean of λ across all laboratories in the population is denoted by μ. The population variance of the laboratory means (i.e., of the λ values) is denoted by σ_L^2 and it is called the "true variance among laboratories." Let ε_{ij} denote the deviation of the j^{th} replicate outcome from the true mean for the i^{th} laboratory, λ_i. The variance of these deviations is denoted by σ_r^2 and it is called the "true repeatability variance" (or alternatively, the "within-laboratory variance"). The model assumes that the true repeatability variance is homogeneous across laboratories. To evaluate this homogeneity assumption, compare the intra-laboratory SD values; e.g., see the final column of Table 1.

The variance of Y_{ij} is denoted by σ_R^2 and it is called the "true reproducibility variance." Under this model $\sigma_R^2 = \sigma_L^2 + \sigma_r^2$; i.e., the true reproducibility variance is the sum of the true variance among laboratories and the true repeatability variance. This notation is the same as used in Table 1 of KSA-SM-14. The one-factor random effects linear model can be expressed succinctly as follows:

$$Y_{ij} = \lambda_i + \varepsilon_{ij}, \text{ where} \qquad \text{(A1)}$$

all λ_i and ε_{ij} values are mutually statistically independent random variables,
λ_i has mean μ and variance σ_L^2, and
ε_{ij} has mean 0 and variance σ_r^2.

For statistical inference purposes, an additional assumption is that all λ_i and ε_{ij} values are normally distributed random variables. In our experience, disinfectant efficacy testing data are suited to the normality assumption.

The ***parameters*** of the model are λ_i for i = 1, …, I, μ, σ_L^2, and σ_r^2. Parameters are conceptual values that we will never know exactly. However, we can estimate the numerical value of each parameter by applying analysis of variance techniques to the observed data, as described in KSA-SM-13 (2013). That is, L_i estimates λ_i for i = 1, …, I; **REMLM** estimates μ, S_L^2 estimates σ_L^2, S_r^2 estimates σ_r^2, and $S_R^2 = S_L^2 + S_r^2$ estimates $\sigma_R^2 = \sigma_L^2 + \sigma_r^2$.

Q of equation (2)

The quantity Q is a function of the numbers of tests in the I individual laboratories. For these numbers, let $\bar{n}_a$ denote the arithmetic mean, $\bar{n}_a = \sum_{i=1}^{I} n_i / I$, $\bar{n}_h$ denote the harmonic mean, $\bar{n}_h = \left[\sum_{i=1}^{I} n_i^{-1} / I\right]^{-1}$, and $\bar{n}_q$ denote the quadratic mean (also called root-mean-square), $\bar{n}_q = \left[\sum_{i=1}^{I} n_i^2 / I\right]^{1/2}$. For balanced data, these three means are equal. For unbalanced data, the well-known inequality for general means shows that $\bar{n}_h < \bar{n}_a < \bar{n}_q$ (Beckenbach and Bellman 1965). The quantity Q is defined only for unbalanced data (equation A2).

$$Q = \left(\bar{n}_h \cdot \left(\bar{n}_q^2 - \bar{n}_a^2\right)\right) / \left(\bar{n}_a \cdot \left(\bar{n}_a - \bar{n}_h\right)\right) \qquad \text{(A2)}$$

Formulas for the variance and standard error of the estimate

For **MLM**, see Miller (1986): $\mathrm{Var}(\mathbf{MLM}) = \frac{\sigma_L^2}{I} + \frac{\sigma_r^2}{I \cdot \overline{n}_h}$;

$$\mathrm{SE}(\mathbf{MLM}) = \sqrt{\frac{S_L^2}{I} + \frac{S_r^2}{I \cdot \overline{n}_h}}. \tag{A3}$$

For **GM**, see Levin and Leu(2013): $\mathrm{Var}(\mathbf{GM}) = \left(\frac{\sigma_L^2}{I}\right) \cdot \frac{\overline{n}_q^2}{\overline{n}_a^2} + \frac{\sigma_r^2}{I \cdot \overline{n}_a}$;

$$\mathrm{SE}(\mathbf{GM}) = \sqrt{\left(\frac{S_L^2}{I}\right) \cdot \frac{\overline{n}_q^2}{\overline{n}_a^2} + \frac{S_r^2}{I \cdot \overline{n}_a}}. \tag{A4}$$

For the weighted average **WM** calculated with the weights W_i of equation (3), see Cochran (1954):

$$\mathrm{Var}(\mathbf{WM}) = \sum_{i=1}^{I} \left(\frac{1}{\sigma_L^2 + \frac{\sigma_r^2}{n_i}} \right).$$

The standard error for **REMLM** is $\sqrt{\mathrm{Var}(\mathbf{WM})}$ with the variance parameters replaced by the corresponding REML variance estimates; see Kacker (2004):

$$\mathrm{SE}(\mathbf{REMLM}) = \sqrt{\sum_{i=1}^{I} \left(\frac{1}{S_L^2 + \frac{S_r^2}{n_i}} \right)}. \tag{A5}$$

References

Beckenbach, E.F. and Bellman, R. (1965) *Inequalities*. Springer-Verlag, New York

Birkes, D., Seely, J., and Abdul-Mordy, A. (1981) An efficient estimator of the mean in a two-stage nested model. Technometrics 23:143-148.

Cochran, W.G. (1954) The combination of estimates from different experiments. *Biometrics* 10:101-129

Hamilton, MA, Hamilton, GC, Goeres, DM, and Parker, AE (2013) Guidelines for the statistical analysis of a collaborative study of a laboratory method for testing disinfectant product performance. *J AOAC Int* 96(5):1138-1151
[The appendix to this article can be downloaded as a MS Word docx file from:
`http://aoac.publisher.ingentaconnect.com/content/aoac/jaoac/2013/00000096/00000005/art00033/supp-data/content-12217_Hamilton_Appendix`]

Iyer, H.K., Wang, D.M.J., and Mathew, T. (2004) Models and confidence intervals for true values in interlaboratory trials. *J. American Statistical Association* 99:1060-1071.

KSA-SM-10 (2012) *Assessing resemblance, repeatability, and reproducibility for quantitative methods*, Center for Biofilm Engineering at Montana State University, Bozeman (CBE Knowledge Sharing Articles)

KSA-SM-13 (2013) *Using R to assess Resemblance, Repeatability, and Reproducibility for quantitative and semi-quantitative disinfectant test methods*, Center for Biofilm Engineering at Montana State University, Bozeman (CBE Knowledge Sharing Articles)

Kacker, R.N. (2004) Combining information from interlaboratory evaluations using a random effects model. *Metrologia* 41:132-136

Levin, B. and Leu, C-S. (2013) Note on an identity between two unbiased variance estimators for the grand mean in a simple random effects model. *The American Statistician* 67:42-43

Miller, R.G. Jr. (1986) *Beyond ANOVA, Basics of Applied Statistics.* John Wiley & Sons, New York

Piepho, H.P. (1996) Weighted estimates of interlaboratory consensus values. *Computational Statistics & Data Analysis* 22:471-479

Rukhin, A.L., Biggerstaff, B.J., and Vangel, M. G. (2000) Restricted maximum likelihood estimation of a common mean and the Mandel-Paule algorithm. *Journal of Statistical Planning and Inference* 83:319-330

Sattar, S.A., Springthorpe, V.S., and Hamilton, M.A. (2005) Appendix 9: Test data based on QCT-1 and QCT-2. in *Harmonization of hard surface disinfectant test methodology in OECD countries* – A report to the European Commission [DG ENV. F.2 (BU-5 00/122), Project No. ENV.C.4/ETU/2003/0063r] October, 2005

Searle, S.R., Casella, G., and McCulloch, C.E. (1992) *Variance Components*. John Wiley & Sons, New York

Tomasino, S.F., Pines, R.M., and Hamilton, G.C. (2012) Procedural Revision to the Use-dilution Method: Establishment of Maximum Log Density Value for Test Microbes on Inoculated Carriers. *J AOAC Int* 95(4):1059-1063

Version date: 19 March 2015

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

Acknowledgments: Darla Goeres and Gordon Hamilton provided helpful comments