Journal of Statistics Education

# The Expected Sample Variance of Uncorrelated Random Variables with a Common Mean and Some Applications in Unbalanced Random Effects Models

Stephen B. Vardeman
Iowa State University

Joanne R. Wendelberger
Los Alamos National Laboratory

*Journal of Statistics Education* Volume 13, Number 1 (2005),
www.amstat.org/publications/jse/v13n1/vardeman.html

**Key Words:** Heteroscedastic; Method of moments; One-way model; Two-factor hierarchical model; Standard error of the mean; Variance component

## Abstract

There is a little-known but very simple generalization of the standard result that for uncorrelated random variables with common mean $\mu$ and variance $\sigma^2$, the expected value of the sample variance is $\sigma^2$. The generalization justifies the use of the usual standard error of the sample mean in possibly heteroscedastic situations, and motivates elementary estimators in even unbalanced linear random effects models. The latter both provides nontrivial examples and exercises concerning method-of-moments estimation, and also helps "demystify" the whole matter of variance component estimation. This is illustrated in general for the simple one-way context and for a specific unbalanced two-factor hierarchical data structure.

## 1. The Expected Value of the Sample Variance

It is completely standard in first courses in statistical theory at a variety of levels to prove that the expected value of the sample variance of independent identically distributed observations is the common variance. (See for example Wackerly, Mendenhall and Scheaffer (2002, page 372), Miller and Miller (2004, page 321), Wasserman (2004, page 52) and Casella and Berger (2002, page 213).) It is no harder to show something more general. Namely, there is the simple result below.

**Lemma 1** If $Y_1, Y_2, \ldots, Y_n$ are uncorrelated random variables with a common mean (say $\mu$) and possibly different variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$, and

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2$$

is their sample variance, then

$$\mathrm{E}\,S^2 = \frac{1}{n}\sum_{i=1}^{n}\sigma_i^2 .$$

**Proof:** First note that

$$\begin{aligned} S^2 &= \frac{1}{n-1}\left(\sum_{i=1}^{n} Y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} Y_i\right)^2\right) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n} Y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} Y_i^2 + \sum_{i \neq j} Y_i Y_j\right)\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \frac{1}{n(n-1)}\sum_{i \neq j} Y_i Y_j \quad . \end{aligned}$$

Then observe that one may with no loss of generality assume that $\mu = 0$. (The $Y_i$ and the $Y_i^* = Y_i - \mu$ have the same sample variance, and if necessary one could replace the $Y_i$ with $Y_i^*$ above.) The assumption that the $Y_i$ are uncorrelated then implies that $\mathrm{E}Y_iY_j$ for all $i \neq j$. Since with mean 0, $\mathrm{E}\,Y_i^2 = \sigma_i^2$, the lemma is proved.

A referee has suggested that in many classroom proofs of Lemma 1, it will be best to write $\sum_{i \neq j} Y_i Y_j$ in the form $2\sum_{i=1}^{n}\sum_{j>i}^{n} Y_i Y_j$ and further suggests that a good exercise will often be to ask students to redo the proof without simplifying to the $\mu = 0$ case. Notice that under the $\mu = 0$ case assumption, the type of summation notation used may not be so important, in that in either notation it is immediate from the fact that $\mathrm{E}Y_iY_j = 0$ for all $i \neq j$ that $\mathrm{E}\,S^2 = \mathrm{E}\frac{1}{n}\sum_{i=1}^{n} Y_i^2$. Not making use of the observation that one may reduce to the $\mu = 0$ case requires using the facts that $\mathrm{E}\,Y_i = \sigma^2 + \mu^2$ and $\mathrm{E}\,Y_iY_j = \mu^2$ for all $i \neq j$, and being able to count that there are $n^2 - n$ terms in $\sum_{i \neq j} Y_i Y_j$ in order to get the necessary cancellation of squared means. How it is easiest for students to see the counting fact from the type of summation notation used depends upon what has gone before in a course. In any case, we think that it is important to use the device of reducing to $\mu = 0$ in classroom proofs, not simply because it is "elegant," but more importantly because it foreshadows how the lemma can be applied in variance component estimation. (See the use of the fact that sample variances are unchanged by the addition of a common value to each element of a "data set" in our later discussion of estimation in an unbalanced two-factor nested design.)

Lemma 1 is very simple and arguably "obvious." But it is not well known and provides a mathematically satisfying extension of the standard result. Further, it can be applied to good effect in important teaching and data analysis contexts.

Note, for example, that under the hypotheses of the lemma

$$\mathrm{E}\,\bar{Y} = \mu \text{ and } \mathrm{Var}\,\bar{Y} = \frac{1}{n^2}\sum_{i=1}^{n}\sigma_i^2 = \frac{\mathrm{E}\,S^2}{n}.$$

So $\bar{Y}$ is potentially a sensible estimator of $\mu$ (at least where the relative precisions of the $Y_i$ are unknown) and

$$\mathrm{SE}_{\bar{Y}} = \frac{S}{\sqrt{n}}.$$

functions as a standard error for $\bar{Y}$ in the potentially heteroscedastic case of the lemma as well as the more familiar iid situation. This is a kind of "robustness" result for the usual standard error of the sample mean and appears as Problem 2.2.3 on page 52 of Stapleton (1995) without explicit mention of Lemma 1. (This is the only reference known to the authors that even hints at Lemma 1.)

We proceed to illustrate that the lemma has important additional uses beyond this most obvious one.

## 2. Applications in the One-Way Random Effects Model With Unbalanced Data

Typical introductions to random effects models and analyses are made in terms of ANOVA mean squares and mysterious "EMS algorithms" for balanced data that are of largely unexplained origin, and really provide little insight into the basic structure of the estimation problems and methods. (See for example Chapter 6, page 172 of Hicks and Turner (1999) or Appendix D, page 1377 of Neter, Kutner, Wasserman, and Nachtsheim (1996) for examples of EMS algorithms.) The possibility of facing the analysis of unbalanced data is either not admitted, or mentioned as an advanced topic requiring application of unspecified specialized advanced techniques.

But it is possible to use Lemma 1 to produce simple/from-first-principles estimators based on (even) unbalanced data under linear random effects models (and in the process demystify the problem of estimation in these models). This is because the lemma shows expected sample variances of appropriate sample average observations to be easily-identified linear combinations of variance components. We first illustrate in the general context of the one-way random effects model.

That is, suppose that for $i = 1, 2, ..., I,$ and $j = 1, 2, ..., n_i$

$$X_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

for $\mu$ some constant, $\alpha_1, \alpha_2, \ldots, \alpha_I$ with mean 0 and variance $\sigma_\alpha^2$, $\varepsilon_{11}, \ldots, \varepsilon_{1n_1}, \varepsilon_{21}, \ldots, \varepsilon_{2n_2}, \ldots, \varepsilon_{I1}, \ldots, \varepsilon_{In_I}$ with mean 0 and variance $\sigma^2$, and all of the $\alpha_i$ and $\varepsilon_{ij}$ uncorrelated. We may apply the foregoing to the uncorrelated sample means

$$Y_i = \bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}$$

that have

$$\mathrm{E}\,\bar{X}_i = \mu \text{ and } \mathrm{Var}\,\bar{X}_i = \sigma_\alpha^2 + \frac{1}{n_i}\sigma^2$$

The unweighted mean of sample means

$$\hat{\mu} = \bar{Y} = \frac{1}{I} \sum_{i=1}^{I} \bar{X}_i$$

is an unbiased estimator of $\mu$ with

$$\mathrm{Var}\,\hat{\mu} = \frac{1}{I}\left(\sigma_\alpha^2 + \left(\frac{1}{I}\sum_{i=1}^{I}\frac{1}{n_i}\right)\sigma^2\right). \tag{1}$$

If we write

$$S_{\bar{X}}^2 = \frac{1}{I-1}\sum_{i=1}^{I}\left(\bar{X}_i - \hat{\mu}\right)^2$$

by Lemma 1, this sample variance (of sample means) has expected value

$$\mathrm{E}\,S_{\bar{X}}^2 = \frac{1}{I}\sum_{i=1}^{I}\mathrm{Var}\,\bar{X}_i = \sigma_\alpha^2 + \left(\frac{1}{I}\sum_{i=1}^{I}\frac{1}{n_i}\right)\sigma^2. \tag{2}$$

So in light of (1) and (2), a standard error for the unbiased estimator of $\mu$ is

$$\mathrm{SE}_{\hat{\mu}} = \frac{S_{\bar{X}}}{\sqrt{I}}. \tag{3}$$

regardless of whether or not the data are balanced.

The authors' original motivation for considering applications of Lemma 1 (and in particular, standard error (3)) in the one-way context was a calibration problem where $\sigma_\alpha^2$ represented a day-to-day variance component in the measurement of a standard, $\sigma^2$ represented a within-day variance component, and constraints in the measurement process led to an error analysis based on the average values. The approach was also applicable in another situation, where analysis of summary data was required, and the sample sizes (and individual observations $X_{ij}$) were not available.

What is more, where the sample sizes and within-group sample variances are available, it is easy to use Lemma 1 to motivate simple estimators of the variance components. Let

$$S_{\text{pooled}}^2 = \frac{\sum_{i,j}\left(X_{ij} - \bar{X}_i\right)^2}{\sum_{j=1}^{I} n_{ij} - I}$$

be the usual pooled sample variance (or mean squared error). This has mean $\sigma^2$. In light of equation (2),

$$\mathrm{E}\left(S_{\bar{X}}^2 - \left(\frac{1}{I}\sum_{i=1}^{I}\frac{1}{n_i}\right)S_{\text{pooled}}^2\right) = \sigma_\alpha^2$$

which suggests the simple estimators of variance components

$$\hat{\sigma}^2 = S_{\text{pooled}}^2 \text{ and } \hat{\sigma}_\alpha^2 = \max\left(0, \left(S_{\bar{X}}^2 - \left(\frac{1}{I}\sum_{i=1}^{I}\frac{1}{n_i}\right)S_{\text{pooled}}^2\right)\right). \tag{4}$$

which appear, for example, in Rao (1997, page 20) and Cox and Solomon (2003, pages 74-76).

# 3. An Application to an Unbalanced Two-Factor Nested Design

The basic pattern used to motivate the estimator of $\sigma_\alpha^2$ in display (4) can be generalized and Lemma 1 applied to produce elementary unbalanced-data estimators of variance components in more complicated linear random effects models. We illustrate this for a particular small unbalanced two-factor nested design consisting of 13 observations $X_{ijk}$ represented in Figure 1. (General formulas for unbalanced two-factor nested designs are possible, but our intention here is to illustrate that Lemma 1 has wide utility, not to do an exhaustive treatment of these designs.)

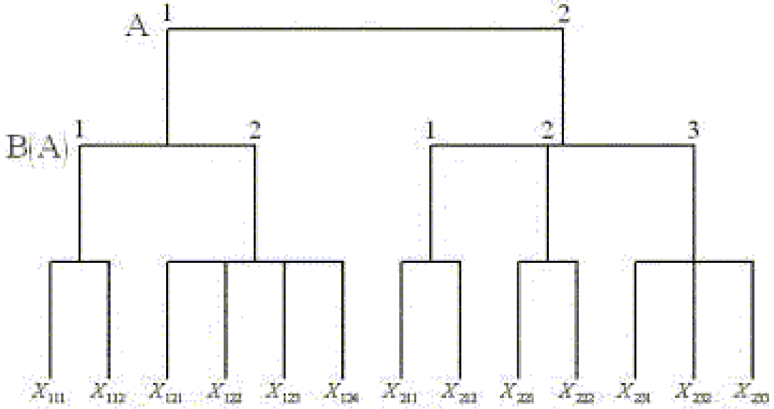

Figure 1

Figure 1: Schematic of a particular unbalanced two-factor hierarchical data structure

That is, with

$X_{ijk}$ = the $k^{\text{th}}$ observation at the $j^{\text{th}}$ level of B within the $i^{\text{th}}$ level of A

suppose that

$$X_{ijk} = \mu + \alpha_i + \beta_{ij} + \varepsilon_{ijk}$$

for $\mu$ some constant, the $\alpha_i$ with mean 0 and variance $\sigma_\alpha^2$, the $\beta_{ij}$ with mean 0 and variance $\sigma_\beta^2$, the $\varepsilon_{ijk}$ with mean 0 and variance $\sigma^2$, and all of the $\alpha_i$, $\beta_{ij}$, and $\varepsilon_{ijk}$ uncorrelated. Let

$n_{ij}$ = the number of observations at level $j$ of B within level $i$ of A

and define sample means

$$\bar{X}_{ij} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} X_{ijk}$$

and unweighted means of these

$$\bar{X}_1 = \frac{1}{2}\left(\bar{X}_{11} + \bar{X}_{12}\right) \text{ and } \bar{X}_2 = \frac{1}{3}\left(\bar{X}_{21} + \bar{X}_{22} + + \bar{X}_{23}\right)$$

and the unweighted mean of these

$$\bar{X} = \frac{1}{2}\left(\bar{X}_1 + \bar{X}_2\right)$$

We consider estimators of the variance components $\sigma_\alpha^2$, $\sigma_\beta^2$, and $\sigma^2$ based on the sample variances (of unweighted sample means)

$$S_1^2 = \frac{1}{2-1}\sum_{j=1}^{2}\left(\bar{X}_{1j} - \bar{X}_1\right)^2 \text{ and } S_2^2 = \frac{1}{3-1}\sum_{j=1}^{3}\left(\bar{X}_{2j} - \bar{X}_2\right)^2$$

and

$$S^2 = \frac{1}{2-1}\sum_{i=1}^{2}\left(\bar{X}_i - \bar{X}\right)^2 .$$

To begin, as always, the usual pooled sample variance

$$S_{\text{pooled}}^2 = \frac{\sum_{i,j,k}\left(X_{ijk} - \bar{X}_{ij}\right)^2}{13-5}$$

serves as an unbiased estimator of $\sigma^2$ . Note then that using the usual notation for averages of $\varepsilon_{ijk}$'s

$$\bar{X}_{1j} = \mu + \alpha_1 + \beta_{1j} + \bar{\varepsilon}_{1j}$$

and that $S_1^{\ 2}$ is not only the sample variance of $\bar{X}_{11}$ and $\bar{X}_{12}$, *but also of* $\bar{X}_{11} - (\mu + \alpha_1) = \beta_{11} + \bar{\varepsilon}_{11}$ and $\bar{X}_{12} - (\mu + \alpha_1) = \beta_{12} + \bar{\varepsilon}_{12}$ (using the same reasoning applied in the proof of Lemma 1 to reduce to the $\mu$ = 0 case). Since $\beta_{11} + \bar{\varepsilon}_{11}$ and $\beta_{12} + \bar{\varepsilon}_{12}$ are uncorrelated with the same mean and $\mathrm{Var}\left(\beta_{11} + \bar{\varepsilon}_{11}\right) = \sigma_\beta^2 + \frac{1}{2}\sigma^2$ while $\mathrm{Var}\left(\beta_{12} + \bar{\varepsilon}_{12}\right) = \sigma_\beta^2 + \frac{1}{4}\sigma^2$, Lemma 1 promises that

$$\mathrm{E}\, S_1^2 = \frac{1}{2}\left(\left(\sigma_\beta^2 + \frac{1}{2}\sigma^2\right) + \left(\sigma_\beta^2 + \frac{1}{4}\sigma^2\right)\right) = \sigma_\beta^2 + \frac{3}{8}\sigma^2 .$$

Similarly,

$$\mathrm{E}\, S_2^2 = \sigma_\beta^2 + \frac{4}{9}\sigma^2 .$$

So for any *c* between 0 and 1,

$$\mathrm{E}\left(cS_1^2 + (1-c)\, S_2^2\right) = \sigma_\beta^2 + \left(c\frac{3}{8} + (1-c)\frac{4}{9}\right)\sigma^2 ,$$

which then suggests that for such *c*, $\sigma_\beta^2$ be estimated as

$$\hat{\sigma}_{\beta,c}^2 = \max\left(0, \left(cS_1^2 + (1-c)\, S_2^2\right) - \left(\left(c\frac{3}{8} + (1-c)\frac{4}{9}\right) S_{\text{pooled}}^2\right)\right).$$

Finally, consider estimating $\sigma_\alpha^2$. With the usual notation for averages of $\beta_{ij}$'s and $\varepsilon_{ijk}$,

$$\bar{X}_1 = \mu + \alpha_1 + \bar{\beta}_1 + \frac{1}{2}\left(\bar{\varepsilon}_{11} + \bar{\varepsilon}_{12}\right) \text{ and } \bar{X}_2 = \mu + \alpha_2 + \bar{\beta}_2 + \frac{1}{3}\left(\bar{\varepsilon}_{21} + \bar{\varepsilon}_{22} + \bar{\varepsilon}_{23}\right).$$

So once more applying Lemma 1 (to the sample variance of uncorrelated variables with a common mean $\bar{X}_1$ and $\bar{X}_2$),

$$\begin{aligned}\mathrm{E}\, S^2 &= \frac{1}{2}\left(\left(\sigma_\alpha^2 + \frac{1}{2}\sigma_\beta^2 + \frac{1}{4}\left(\frac{3}{4}\right)\sigma^2\right) + \left(\sigma_\alpha^2 + \frac{1}{3}\sigma_\beta^2 + \frac{1}{9}\left(\frac{4}{3}\right)\sigma^2\right)\right)\\ &= \sigma_\alpha^2 + \frac{5}{12}\sigma_\beta^2 + \frac{145}{864}\sigma^2 \quad ,\end{aligned}$$

which in turn suggests the estimator

$$\hat{\sigma}^2_{\alpha,c} = \max\left(0, S^2 - \frac{5}{12}\hat{\sigma}^2_{\beta,c} - \frac{145}{864}S^2_{\text{pooled}}\right)$$

# 4. Final Comments

Lemma 1 is simple and interesting in its own right, and on that basis alone probably deserves to replace the standard independent and identically distributed (iid) result in introductions to mathematical statistics. But beyond this motivation, the examples offered here illustrate that it can be used to find elementary estimators in all kinds of unbalanced random effects models. Such applications are potentially useful both in “rough and ready” practical data analysis, and in important teaching contexts. The first author has found it useful when providing students some exposure to random effects analyses where little familiarity with ANOVA can be assumed. As we’ve argued above, it can be used to demystify otherwise obscure EMS values and provide simple methods for unbalanced data in experimental design courses. And even in mathematical statistics courses, it can be used to provide nontrivial examples and exercises concerning method-of-moments estimation.

While our discussion has focused exclusively on moment results (and is thus not restricted to Gaussian models), there is much traditional interest and a huge literature concerned with distributional (and inference) results when one adds normality to the kind of assumptions we’ve made. Our reviewers have made several interesting points regarding connections to that literature. If one adds (joint) normality to the assumptions of Lemma 1, the resulting distribution for $S^2$ is not chi-squared, but rather that of a weighted average of independent chi-square variables. On the other hand, under the normal one-way random effects model, our $S^2_{\bar{x}}$ is sometimes referred to as the unweighted mean square, and pages 68-73 of Burdick and Graybill (1992) argue that suitably scaled, it is approximately chi-square. Further, this result has been used by El-Bassiouni and Abelhafez (2000) to produce valid confidence intervals for $\mu$ in this context. Finally, pages 98-106 of Burdick and Graybill argue that in the normal version of the two-factor nested design, provided $c$ is suitably chosen, the quantity $cS_1^2 + (1-c)S_2^2$ is approximately chi-square.

# Acknowledgments

The authors thank Jim Stapleton for pointing out the reference to his text, and the fact that the proof of Lemma 1 can be simplified by noting that without loss of generality one may assume that $\mu = 0$. They also thank Joachim Kunert, Ron Christensen, Bobby Mee, Karen Kafadar, Glen Meeden, H.A. David, Dennis Gilliland, Götz Trenkler, Bob Stephenson, and three anonymous reviewers for comments on earlier drafts of the note that have worked to make it more complete and readable, and hopefully more useful.

Financial support of the Deutsche Forschungsgemeinschaft (SFB 475, “Reduction of Complexity in Multivariate Data Structures”) through the University of Dortmund and of the Los Alamos National Laboratory Statistical Sciences Group is gratefully acknowledged by the first author.

# References

Burdick, R.K. and Graybill, F.A. (1992), *Confidence Intervals on Variance Components*, New York: Marcel Dekker.

Casella, G. and Berger, R.L. (2002), *Statistical Inference*, Pacific Grove, California: Duxbury.

Cox, D.R. and Solomon, P.J. (2003), *Components of Variance*, New York: Chapman & Hall.

El-Bassiouni, M.Y. and Abdelhafez, M.E.M. (2000), "Interval estimation of the mean in a two-stage nested model," *Journal of Statistical Computation and Simulation*, 67 (4), pp. 333-350.

Hicks, C.R. and Turner, K.V. (1999), *Fundamental Concepts in the Design of Experiments, 5th Ed.*, Oxford: Oxford University Press.

Miller, I. and Miller, M. (2004), *John E. Freund's Mathematical Statistics, 7th Edition*, Upper Saddle River, New Jersey: Prentice Hall.

Neter, J., Kutner, M.H., Wasserman, W., and Nachtsheim, C.J. (1996), *Applied Linear Statistical Models, 4th Edition*, Chicago: McGraw-Hill/Irwin.

Rao, P.S.R.S. (1997), *Variance Components Estimation*, New York: Chapman & Hall.

Stapleton, J.H. (1995), *Linear Statistical Models*, New York: John Wiley & Sons.

Wackerly, D.D., Mendenhall W., and Scheaffer, R.L. (2002), *Mathematical Statistics with Applications, 6th Edition*, Pacific Grove, California: Duxbury.

Wasserman, L. (2004), *All of Statistics: A Concise Course in Statistical Inference*, New York: Springer-Verlag.

---

Stephen B. Vardeman
Departments of Statistics and Industrial and Manufacturing Systems Engineering
Iowa State University
Ames, IA 50011-1210
U.S.A.
vardeman@iastate.edu

Joanne R. Wendelberger
Statistical Sciences Group
Los Alamos National Laboratory
Los Alamos, NM
U.S.A.
joanne@lanl.gov

---