LA-UR-03-6568

The Expected Sample Variance of Uncorrelated Random Variables With a Common Mean and an Application in Unbalanced Random Effects Models

by

Stephen B. Vardeman[*]
Departments of Statistics and Industrial and Manufacturing Systems Engineering
Iowa State University
vardeman@iastate.edu

Joanne Wendelberger
Statistical Sciences Group
Los Alamos National Laboratory
joanne@lanl.gov

October 9, 2003 (Corrected Version)

Abstract

There is a little-known but very simple generalization of the standard result that for uncorrelated variables with a common mean and variance, the expected sample variance is the marginal variance. The generalization justifies the use of the usual standard error of the sample mean in possibly heteroscedastic situations and motivates some simple estimators for unbalanced linear random effects models. The latter is illustrated for the simple one-way context.

The Mean of the Sample Variance

It is completely standard in first courses in statistical theory to prove that the mean of the sample variance of iid observations is the common marginal variance. It is little 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 μ) and possibly different variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$, and

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

[*] Financial support of the Deutsche Forschungsgemeinschaft (SFB 475, "Reduction of Complexity in Multivariate Data Structures") through the University of Dortmund and of Los Alamos National Laboratory Statistical Sciences Group is gratefully acknowledged.

is their sample variance, then

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

Proof: First note that

$$S^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} Y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} Y_i^2 + \sum_{i\neq j} Y_i Y_j\right)\right)$$
$$= \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \frac{1}{n(n-1)}\sum_{i\neq j} Y_i Y_j$$

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 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 $\mathrm{E}\,Y_i Y_j = 0\ \forall i, j$. Since with mean 0, $\mathrm{E}\,Y_i^2 = \sigma_i^2$, the lemma is proved.□

Note also that under the hypotheses of the lemma

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

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

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

functions as a standard error for $\overline{Y}$ 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 page 52 of Stapleton (1995) without explicit mention of Lemma 1. We proceed to illustrate that the lemma has uses beyond this most obvious one.

Application in Linear Random Effects Models for Unbalanced Data

It is possible to use Lemma 1 to produce simple estimators based on (even) unbalanced data under linear random effects models. This is because the lemma shows expected sample variances of appropriate sample average observations to be easily-identified linear combinations of variance components. Here we illustrate in the context of the one-way random effects model. (For more complicated examples, see Section 4.2 of the class notes at http://www.public.iastate.edu/~vardeman/stat531/stat531.html)

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

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

for μ some constant, $\alpha_1, \alpha_2, \ldots, \alpha_I$ with mean 0 and variance σ_α^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 σ^2, and all of the α_i and ε_{ij} uncorrelated. We may apply the foregoing to the sample means

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

The unweighted mean of sample means

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

is an unbiased estimator of μ with variance

$$\frac{1}{I}\left(\sigma_\alpha^2 + \left(\frac{1}{I} \sum_{i=1}^{i} \frac{1}{n_i} \right) \sigma^2 \right)$$

If we write

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

this sample variance has mean

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

and a standard error for $\hat{\mu}$ is

$$\mathrm{SE}_{\hat{\mu}} = \frac{S_{\overline{X}}}{\sqrt{I}}$$

regardless of whether or not the data are balanced.

The authors' original motivation for considering Lemma 1 was a calibration problem where σ_α^2 represented a day-to-day variance component in the measurement of a standard, σ^2 represented a within-day variance component, and constraints in the measurement process led to an error analysis based on the average values. This approach also is 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 produce sensible estimators of the variance components. Let

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

be the usual pooled sample variance (or mean squared error). This, of course, has mean σ^2. In light of equation (1),

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

which suggests the simple estimators of variance components

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

Final Comments

The results in this note are very simple and arguably "obvious." But they are not well known. The only reference that the authors know for something like Lemma 1 is the Stapleton problem. And they have been unable to locate references for the application of it to estimation in the one-way random effects model. For example, in this latter regard, the estimator of σ_{α}^{2} presented above does not seem to be among those listed in Searle (1971) for unbalanced data contexts.

References

Searle, S.R. (1971). *Linear Models*, John Wiley & Sons, New York.

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

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 WOLOG one may assume that $\mu = 0$. They also thank Joachim Kunert, Ron Christensen, Bobby Mee, Karen Kafadar, Glen Meeden, and H.A. David for comments on an earlier draft of the note.