Published in final edited form as:

Am Stat. 2013 January 1; 67(1): 42–43. doi:10.1080/00031305.2012.752105.

Note on an Identity Between Two Unbiased Variance Estimators for the Grand Mean in a Simple Random Effects Model

Bruce Levin and Cheng-Shiun Leu

Department of Biostatistics, Columbia University, 722 West 168th Street 6th Floor, New York, NY 10032, USA

Abstract

We demonstrate the algebraic equivalence of two unbiased variance estimators for the sample grand mean in a random sample of subjects from an infinite population where subjects provide repeated observations following a homoscedastic random effects model.

Keywords

Intraclass correlation coefficient; Random effects; Ratio-unbiased estimators; Unbiased estimators; Variance estimators; Variance inflation factor

Let X be a random variable that can be observed repeatedly in subjects randomly sampled from an infinite population. We suppose the subjects have true subject-specific means μ such that given μ we have $E(X|\mu) = \mu$ and $Var(X|\mu) = \sigma^2$, while across subjects in the population we have $E(\mu) = \nu$ and $Var(\mu) = \tau^2$. We assume that observations from different subjects are statistically independent, while for two observations from the same subject, say X and X' , they are conditionally independent given μ , so that marginally, $Cov(X, X') = \tau^2$. In an alternate notation our model is, for the j^{th} observation from subject i ,

$$X_{ij} = \nu + x_i + e_{ij}$$

where $E(x_i) = E(e_{ij}) = 0$, $Var(x_i) = \tau^2$, $Var(e_{ij}) = \sigma^2$, $Cov(x_i, e_{ij}) = 0$, and $\mu_i = \nu + x_i$. Note that the population variance of a single observation is $Var(X_{ij}) = \sigma^2 + \tau^2$. We refer to this common situation as the simple random effects model. Examples include the simple replication study for estimating the reliability of measurements (see, e.g., Fleiss, 1986); each arm of a classical comparative repeated measures study with constant means across time; or an observational study of subjects who provide data clusters of varying size.

Suppose our goal is to estimate ν , the population average of true subject-specific means. To do that we complete the simple random effects model specification by assuming that the within-subject variance σ^2 and the between-subject variance τ^2 of the true subject-specific means are constant for all subject-specific sample sizes m , and by assuming for the moment that the true subject-specific means do not vary systematically with their observed sample sizes, i.e., $E(\mu|m) = \nu$ for all m . Now let $i = 1, \dots, K$ index the simple random sample of subjects with observed sample sizes $m_1, \dots, m_K$ and let X_{ij} denote the unit observations in the total sample where $j = 1, \dots, m_i$ for $i = 1, \dots, K$. The assumption that μ and m do not covary will allow us to condition on the observed sample sizes $m_1, \dots, m_K$ and treat them as constants. In particular, we will suppress them from the notation in all conditional means and variances. We briefly comment on the case of covarying μ and m at the end of this note.

Let $\bar{X}_{i.} = m_i^{-1} \sum_{j=1}^{m_i} X_{ij}$ denote the sample subject-specific means for $i = 1, \dots, K$ and let the sample grand mean of the observations be denoted by $\bar{X}_{..} = \sum_{i=1}^K \sum_{j=1}^{m_i} X_{ij} / \sum_{i=1}^K m_i = \sum_{i=1}^K (m_i/N) \bar{X}_{i.}$, where $N = m_1 + \dots + m_K$ is the total number of observations in the sample. The following facts are well-known:

$$E(\bar{X}_{i.} | \mu_i) = \mu_i, \text{Var}(\bar{X}_{i.} | \mu_i) = \sigma^2 / m_i, E(\bar{X}_{i.}) = \nu, \text{Var}(\bar{X}_{i.}) = \tau^2 + \sigma^2 / m_i,$$

from which it follows that

$$E(\bar{X}_{..} | \mu_1, \dots, \mu_K) = \sum_{i=1}^K (m_i/N) \mu_i, \text{Var}(\bar{X}_{..} | \mu_1, \dots, \mu_K) = \sum_{i=1}^K (m_i/N)^2 \sigma^2 / m_i = \sigma^2 / N,$$

whence unconditionally $E(\bar{X}_{..}) = \nu$. Thus $\bar{X}_{..}$ is an unbiased estimator of ν under our assumptions and, writing $\underline{\mu} = (\mu_1, \dots, \mu_K)$, the variance of the grand mean is

$$\text{Var}(\bar{X}_{..}) = E\{\text{Var}(\bar{X}_{..} | \underline{\mu})\} + \text{Var}\{E(\bar{X}_{..} | \underline{\mu})\} = (\sigma^2 / N) + \sum_{i=1}^K (m_i/N)^2 \tau^2 = (\sigma^2 + \bar{m} \tau^2) / N = \{(\sigma^2 + \tau^2) / N\} \{1 + (\bar{m} - 1) \rho\} \quad (1)$$

where $\bar{m} = \sum_{i=1}^K m_i^2 / \sum_{i=1}^K m_i = \sum_{i=1}^K (m_i/N) m_i$ is the size-weighted average sample size, and ρ is the intraclass correlation coefficient (ICC), $\rho = \text{Cov}(X, X') / \text{Var}(X) = \tau^2 / (\sigma^2 + \tau^2)$. Note that the final expression in (1) expresses $\text{Var}(\bar{X}_{..})$ in the familiar form of a population variance $\sigma^2 + \tau^2$ times the variance inflation factor (VIF) of $1 + (\bar{m} - 1) \rho$.

The purpose of this note is to recognize an algebraic identity between two versions of an unbiased estimator for $\text{Var}(\bar{X}_{..})$ in (1). To that end, consider the familiar ANOVA table below. The constant k in $E(MSB)$ is given by $k = (N - \bar{m}) / (K - 1)$ (see, e.g., Fleiss, 1986). Because the expected value of the total mean square is often omitted from the ANOVA table, it may be less familiar that

$$E(s^2) = (\sigma^2 + \tau^2) [1 - \{(\bar{m} - 1) \rho / (N - 1)\}]. \quad (2)$$

ANOVA table for the simple random effects model

Source	Sum of squares	Degrees of Freedom	Mean square	Expected mean square
Between subjects	$SSB = \sum_{i=1}^K m_i (\bar{X}_{i.} - \bar{X}_{..})^2$	$K - 1$	$MSB = SSB / (K - 1)$	$\sigma^2 + k \tau^2$
Within subjects	$SSW = \sum_{i=1}^K \sum_{j=1}^{m_i} (\bar{X}_{ij} - \bar{X}_{i.})^2$	$N - K$	$MSW = SSW / (N - K)$	σ^2

Source	Sum of squares	Degrees of Freedom	Mean square	Expected mean square
Total	$SST = \sum_{i=1}^K \sum_{j=1}^{m_i} (\bar{X}_{ij} - \bar{X}_{..})^2$	$N - 1$	$s^2 = MST = SST/(N - 1)$	(2)

Note, in particular, that when the *ICC* is non-zero, s^2 is a biased estimator of the population variance $\sigma^2 + \tau^2$ of a single observation. The bias arises from the fact that the N observations are not marginally independent, whereas the usual mean squared estimator s^2 of the population variance relies on independence for its unbiasedness. Cochran (1977, p. 238) makes a similar comment in the finite population approach to cluster sampling.

Substituting the well known ANOVA estimators for σ^2 and τ^2 in (1) we obtain an estimator of $N \text{Var}(\bar{X}_{..})$ of the form $\hat{\sigma}^2 + \bar{m}\hat{\tau}^2$ where $\hat{\sigma}^2 = MSW$, $\hat{\tau}^2 = (MSB - MSW)/k$. An interesting algebraically equivalent form of this estimator is given by

$$\frac{s^2\{1 + (\bar{m} - 1)\hat{\rho}\}}{1 - (\bar{m} - 1)\hat{\rho}/(N - 1)} \quad (3)$$

where s^2 is the ordinary sample variance of the X_{ij} 's and $\hat{\rho} = \hat{\tau}^2/(\hat{\sigma}^2 + \hat{\tau}^2)$ is a common estimator for the ICC ρ . This equivalence can be seen easily if

$$s^2 = (\hat{\sigma}^2 + \hat{\tau}^2)\{1 - (\bar{m} - 1)\hat{\rho}/(N - 1)\} \quad (4)$$

because then it follows that (3) becomes

$$(\hat{\sigma}^2 + \hat{\tau}^2)(1 + (\bar{m} - 1)\hat{\rho}) = (\hat{\sigma}^2 + \hat{\tau}^2) + (\bar{m} - 1)\hat{\tau}^2 = \hat{\sigma}^2 + \bar{m}\hat{\tau}^2$$

(cf. equation (1)). Indeed (4) does hold because the right side of the equality equals

$$(\hat{\sigma}^2 + \hat{\tau}^2)\{1 - (\bar{m} - 1)(\frac{\hat{\tau}^2}{\hat{\sigma}^2 + \hat{\tau}^2})/(N - 1)\} = (MSW + \frac{MSB - MSW}{k}) - \frac{\bar{m} - 1}{N - 1}(\frac{MSB - MSW}{k}) = c_1 MSB + c_2 MSW = s^2$$

where, after some simplification, $c_1 = (1/k) - (\bar{m} - 1)/\{(N - 1)k\} = (K - 1)/(N - 1)$ and $c_2 = 1 - (1/k) + (\bar{m} - 1)/\{(N - 1)k\} = 1 - c_1 = (N - K)/(N - 1)$ and the last equality comes from the familiar analysis of variance decomposition $SST = SSB + SSW$.

Expression (3) appears to be new, exhibiting the standard unbiased estimator in the familiar form of an “independent-sample variance estimate times VIF” with bias-correction, and is note-worthy because ordinarily, substituting an unbiased (or ratio-unbiased) estimator of an unknown population parameter into a second unbiased estimator of a different population parameter would not be expected to maintain the unbiasedness of the second estimator.

COMMENT

When μ and m covary, the sample mean $\bar{X}_{..}$ is no longer an unbiased estimate of ν . This situation could arise in observational studies where each subject provides a cluster of data whose sizes m_i are not under the control of the statistician and vary from subject to subject. Conditional on the cluster sizes the expectation of $\bar{X}_{..}$ is

$$\sum_{i=1}^K (m_i/N) E(\mu_i|m_i) = \frac{K^{-1} \sum_{i=1}^K E(m_i \mu_i | m_i)}{K^{-1} \sum_{i=1}^K m_i}$$

so unconditionally as K becomes large the sample mean is a consistent and ratio-unbiased estimator of

$$\frac{E(m\mu)}{E(m)} = \nu + \frac{\text{Cov}(m, \mu)}{E(m)}.$$

Post-hoc stratification of the sample into subsets of subjects homogeneous in m would be required to remove the bias.

Finally we note how easy it is to miss a simple and clear explanation for an algebraic identity: our original proof was much longer and less revealing than the above. We thank the referee and editors for their insightful suggestions.

REFERENCES

- Cochran, WG. Sampling Techniques. 3rd Edition. Hoboken: John Wiley & Sons; 1977.
 Fleiss, JL. The Design and Analysis of Clinical Experiments. Hoboken: John Wiley & Sons; 1986.