Why the standard deviation of the school means is too big.

Miratrix

August 26, 2019

Say we want to estimate the variability of mean math achievement across schools. I.e., each school has some average math achievement of its students, and we want to know how different schools are.

The naïve way of doing this is to estimate the mean math achievement for a sample of schools, and take the standard deviation (square root of variance) of this sample as a reasonable estimate. In math terms, we would calculate $\bar{Y}_j$ for each school j and then use as our estimate

$$\widehat{\tau^2} = var(\bar{Y}_j) = \frac{1}{n-1} \sum_{j=1}^{J} \bar{Y}_j.$$

This will give a number that is too big. The following is a math derivation on a simple scenario that illustrates why.

First, pretend our Data Generation Process (DGP) is Mother Nature making a bunch of schools, and then for each school making a bunch of kids. Our model is that the schools are represented by school-level true mean math achievement, and the kids are made by adding an individual kid effect to the mean math achievement of their schools.

So we have

$$\alpha_j \sim N(\mu, \tau^2)$$

meaning each school is a random draw from a normal distribution with a mean μ and a standard deviation τ. These are the *true* means of the schools. We wish we knew them, but we do not. Instead we see a sample of kids from the school and we hope the mean of the kids is close to this true mean α_j.

For any kid i we have

$$Y_i = \alpha_{j[i]} + \epsilon_i$$

with

$$\epsilon_i \sim N(0, \sigma^2).$$

These ϵ_i are the classic residuals we are used to.

For the moment, assume each school j has n kids. Then the average observed math achievement is

$$\bar{Y}_j = \frac{1}{n} \sum_{i:j[i]=j} Y_i,$$

the average of all kids in the school. Note the "$i : j[i] = j$" term, which reads as "i for those i where $j[i] = j$" meaning "sum over all students which go to school j."

Ok, so now we have math achievement for school j. We then have

$$\bar{Y}_j = \frac{1}{n} \sum_{i:j[i]=j} Y_i = \frac{1}{n} \sum_{i:j[i]=j} \alpha_{j[i]} + \epsilon_i = \frac{1}{n} \sum_{i:j[i]=j} \alpha_j + \frac{1}{n} \sum_{i:j[i]=j} \epsilon_i = \alpha_j + \frac{1}{n} \sum_{i:j[i]=j} \epsilon_i = \alpha_j + \bar{\epsilon}_j.$$

Here we have $\bar{\epsilon}_j = \bar{Y}_j - \alpha_j$, i.e., we have a school-level residual, the error in our estimate of α_j using $\bar{Y}_j$. This residual is the sum of a bunch of student residuals, which we assume are all independent of each other. When you average a bunch of independent, identically distributed (i.i.d.) residuals, each with variance σ^2, you get something which still has the same mean (of 0) but a smaller variance by a factor of n:

$$\text{Var}[\bar{\epsilon_j}] = \text{Var}\left[\frac{1}{n} \sum_{i:j[i]=j} \epsilon_i\right] = \frac{1}{n^2} \sum_{i:j[i]=j} \text{Var}[\epsilon_i] = \frac{1}{n}\sigma^2$$

This is the familiar result that the mean of a bunch of variables has a standard deviation $1/\sqrt{n}$ of the original standard deviation (part of the Central Limit Theorem).

We can think of $\bar{Y}_j$ as a random quantity, random for two reasons: school j is a randomly made school, and the students in school j are randomly made students. Under the assumption that the students' error terms are independent of the school's mean math achievement we can easily calculate the variance of our estimator:

$$\text{Var}\left[\bar{Y}_j\right] = \text{Var}[\alpha_j] + \text{Var}[\bar{\epsilon}_j] = \tau^2 + \frac{1}{n}\sigma^2$$

This is bigger than our target of τ^2, the true variability in mean math achievement across schools. The uncertainty in estimating the α_j has entered into the variability.

Our estimate $\widehat{\tau^2}$ will be an unbiased estimate of $\tau^2 + \frac{1}{n}\sigma^2$. One way to fix is to estimate σ^2 and then adjust our estimate of the variance of τ^2 by subtracting $\frac{1}{n}\hat{sigma}^2$. Another is to use multilevel modeling, which does this for us, in effect.