# Meng on Big Data: The basic result, but in excruciating detail

Will Lowe

October 13, 2020

In a population of size $N$, Let $G_j(X_j) = G_j$ be some quantity we want to estimate and $R_j$ be an indicator that element $j$ is in a sample of size $n = \sum_j^J R_j$. The population quantity that wants estimating is

$$\overline{G}_N = \frac{1}{N} \sum_j^N G_j$$

whereas the sample we take is

$$\overline{G}_n = \frac{1}{N} \sum_j^N \frac{R_j G_j}{R_j}$$

Letting $R$, $X$, and $G$ be fixed then we can characterize the difference between sample and true in Meng's overly busy notation as

$$\begin{aligned}
\overline{G}_n - \overline{G}_N &= \frac{\mathrm{E}_J[R_J G_J]}{\mathrm{E}_J[R_J]} - \mathrm{E}_J[G_J] \\
&= \frac{\mathrm{E}_J[R_J G_J]}{\mathrm{E}_J[R_J]} - \frac{\mathrm{E}_J[R_J]\mathrm{E}_J[G_J]}{\mathrm{E}_J[R_J]} \\
&= \frac{\mathrm{E}_J[R_J G_J] - \mathrm{E}_J[R_J]\mathrm{E}_J[G_J]}{\mathrm{E}_J[R_J]} \\
&= \frac{\mathrm{Cov}[R_J, G_J]}{\mathrm{E}_J[R_J]}
\end{aligned}$$

Now, recalling that

$$\rho_{R,G} = \mathrm{Cor}(R_J, G_J) = \frac{\mathrm{Cov}[R_J, G_J]}{\sqrt{\mathrm{Var}[R_J]}\sqrt{\mathrm{Var}[G_J]}}$$

and denoting $f = \mathrm{E}_J[R_J] = \frac{n}{N}$, so that, since $R$ is binary, $\mathrm{Var}[R_J] = f(1-f)$.
Denote $\sigma_G^2 = \mathrm{Var}[G_J]$ and write

$$\begin{aligned}
\overline{G}_n - \overline{G}_N &= \frac{\rho_{R,G}\sqrt{\mathrm{Var}[R_J]}\sqrt{\mathrm{Var}[G_J]}}{\mathrm{E}_J[R_J]} \\
&= \rho_{R,G} \frac{\sqrt{f(1-f)}}{f} \sigma_G \\
&= \rho_{R,G} \frac{f(1-f)}{f^2} \sigma_G \\
&= \rho_{R,G} \sqrt{\frac{(1-f)}{f}} \sigma_G
\end{aligned}$$

and the expected value of this thing (the mean squared error) is

$$\begin{aligned}
\mathrm{MSE}_R[G_n] &= \mathrm{E}[(\overline{G}_n - \overline{G}_N)^2] \\
&= \mathrm{E}[(\rho_{R,G}\sqrt{\frac{(1-f)}{f}}\sigma_G)^2] \\
&= \mathrm{E}[\rho^2_{R,G}]\frac{(1-f)}{f}\sigma^2_G
\end{aligned}$$

Now assume simple random sampling. What's $\rho_{R,G}$ for that? Well there we know that the MSE is the same as the variance because it's unbiased. Reminder, the (finite sample) variance is

$$\frac{1-f}{n}S^2_G \quad \text{where} \qquad S^2_G = \frac{N}{N-1}\sigma^2_G$$

so putting this together and plugging it into the left hand side we can back out $\rho_{R,G}$:

$$\begin{aligned}
\frac{1-f}{n}\frac{N}{N-1}\sigma^2_G &= \rho^2_{R,G}\frac{(1-f)}{f}\sigma^2_G \\
&= \rho^2_{R,G}\left[\frac{(1-f)N}{n}\sigma^2_G\right] && \text{expand } f \text{ in the denominator and group} \\
\frac{\frac{1-f}{n}\frac{N}{N-1}\sigma^2_G}{\frac{(1-f)N}{n}\sigma^2_G} &= \rho^2_{R,G} && \text{divide both sides by the group} \\
\frac{1}{N-1} &= \rho^2_{R,G}
\end{aligned}$$

Note that $N$ is the population size so this terms is usually very small.

In general we can't estimate $\rho^2_{R,G}$ from data (because it only has $R_j = 1$ cases by construction), but if we happen to have an estimate of the actual error then we can nevertheless back it out.