Sample size estimates
Preliminary results - please do not disseminate

Dennis M. Feehan
feehan@berkeley.edu

28 June, 2017

Overview

In order to get sample size estimates, we have to understand the sampling variance of the estimator. This understanding can be a mathematical expression or a simulation. Once we have an understanding of the sampling variance, we have to pick the criterion that will be used to pick a sample size. There are many possibilities; here, we'll focus on the coefficient of variation and the margin of error.

In this note, we first develop an understanding of the sampling variance of the estimator; then we look at the coefficient of variation and the margin of error implied by the variance. Finally, we look specifically at what the results imply for possible sample sizes.

Sampling variance

The basic scale-up estimator is

$$\widehat{N}_H = \frac{\widehat{y}_{F,H}}{\widehat{\bar{d}}_{F,F}} = N_F \frac{\widehat{y}_{F,H}}{\widehat{d}_{F,F}} = N \frac{\widehat{\bar{y}}_i}{\widehat{\bar{d}}_i}.$$

To estimate a prevalence, we divide the estimator by the population size, N

$$\widehat{p} = \frac{\widehat{\bar{y}}_i}{\widehat{\bar{d}}_i}. \tag{1}$$

As a ratio estimator, we can approximate the sampling variance of $\widehat{p}$ with

$$\mathrm{V}(\widehat{p}) \approx \left(\frac{1}{\bar{d}}\right)^2 \left[\mathrm{V}(\widehat{\bar{y}}_i) + \left(\frac{\bar{y}}{\bar{d}}\right)^2 \mathrm{V}(\widehat{\bar{d}}_i) - 2\left(\frac{\bar{y}}{\bar{d}}\right) \mathrm{C}(y_i, d_i)\right]. \tag{2}$$

In the case of simple random sampling, we can simplify this expression by using the fact that $\mathrm{V}(\widehat{\bar{y}}) = (1 - \mathrm{fpc})S_y^2/n$, $\mathrm{V}(\widehat{\bar{d}}) = (1 - \mathrm{fpc})S_d^2/n$, and $\mathrm{C}(\widehat{\bar{y}}, \widehat{\bar{d}}) = (1 - \mathrm{fpc})S_{y,d}/n$, where fpc is the finite population correction factor $(1 - \frac{n}{N})$, and S_y^2, S_d^2, and $S_{y,d}$ are the unit variances and covariances. This leads to

$$\mathrm{V}(\widehat{p}) \approx (1 - \mathrm{fpc}) \left(\frac{1}{n}\right) \left(\frac{1}{\bar{d}_i}\right)^2 \left[S_y^2 + \left(\frac{\bar{y}}{\bar{d}}\right)^2 S_d^2 - 2\left(\frac{\bar{y}}{\bar{d}}\right) S_{yd}\right]. \tag{3}$$

Equation 3 shows that the sampling variance of the prevalence estimate will depend on

- n - the sample size
- $\text{fpc} = 1 - \frac{n}{N}$ - the finite population correction factor
- $\bar{d}$ - the average size of the personal network
- $\bar{y}$ - the average number of reported women who had an abortion in the past 3 years
- $S_y^2 = \text{V}(y_i)$ - the population variance in the reported connections to the hidden population
- $S_d^2 = \text{V}(d_i)$ - the population variance in personal network sizes
- $S_{y,d} = \text{C}(y_i, d_i)$ - the population covariance between the reported connections and the personal network sizes

In order to use Equation 3, we must be able to plug in values for each of these terms. One approach to obtaining sample sizes is to conduct simulation studies based on Equation 3. Based on the study design, we can make the following assumptions about the remaining variables:

- the average number of confidants is $\bar{d} = 1.5$
- prevalence of women with abortions in the past 3 years is $p_{\text{true}} = 0.025$
- abortions are under-reported by $\tau = 0.66$
- the estimand is thus $\frac{\bar{y}}{\bar{d}} = \tau p_{\text{true}}$; we'll call this quantity p

We also have to make assumptions about the distributions of d_i and y_i. We make simple assumptions here:

- the number of confidants is Poisson distributed: $d_i \sim \text{Poisson}(\bar{d})$
- the number of reported abortions is binomially distributed: $y_i \sim \text{Binomial}(\bar{d}, p)$

These are both likely not perfectly realistic; in particular, if there is a lot of clustering in the network, the actual distributions may have higher variance than the Binomial and Poisson we use here. In the absence of any empirical information, these seem like a reasonable starting place.

In addition to simulations based on Equation 5, a different approach is to make more approximations and assumptions in order to keep simplifying Equation 5 until we have a tractable analytical expression. (Equation 5 is hard to use analytically because it depends on S_y^2, S_d^2, and $S_{y,d}$ which only have simple forms in special cases.)

To proceed, we start with two assumptions:

1. We assume that the finite population correction is 1; this is tantamount to saying that our sample size n is negligible compared to the population size N, i.e. that $1 - \frac{n}{N} \approx 1$. This is a conservative assumption in the sense that it can be expected to lead us to conclude that we need a slightly larger sample than we actually do (because incorporating the fpc would reduce the variance; the fpc is less than 1)

2. We assume that the covariance between the reported connections and the personal network sizes is 0. This is very likely conservative, because this covariance is probably positive (ie, people who have bigger networks are also expected to report that they know more women who had abortions). If the covariance is actually positive, then the $-2\left(\frac{\bar{y}}{\bar{d}}\right) S_{y,d}$ term would reduce the variance; by treating this term as 0, we again believe that we would err on the side of calculating that we need a sample size larger than we actually do.

After these two simplifications, we are left with

$$\mathrm{V}(\widehat{p}) \approx \left(\frac{1}{n}\right)\left(\frac{1}{\bar{d}}\right)^2 \left[S_y^2 + \left(\frac{\bar{y}}{\bar{d}}\right)^2 S_d^2\right]. \tag{4}$$

Note also that

- $\bar{y} = \bar{d}p_{\text{true}}\tau = \bar{d}p$.
- the distributional assumption about the d_i implies that $S_d^2 = \bar{d}$.
- there is no simple exact expression for S_y^2, but a loose approximation is given by $S_y^2 = \frac{1}{n}\sum_i (y_i - \bar{y})(d_i - \bar{d}_i) \approx \bar{d}p(1-p)$.

With these further simplifications, our expression can be reduced to

$$\begin{aligned}
\mathrm{V}(\widehat{p}) &\approx \left(\frac{1}{n}\right)\left(\frac{1}{\bar{d}}\right)^2 \left[\bar{d}p(1-p) + \left(\frac{\bar{y}}{\bar{d}}\right)^2 \bar{d}\right] \\
&= \left(\frac{1}{n}\right)\left(\frac{1}{\bar{d}}\right)^2 \left[\bar{d}p(1-p) + \bar{d}p^2\right] \\
&= \left(\frac{1}{n\bar{d}}\right) p\left[1 - p + p\right] \\
&= \frac{p}{n\bar{d}}
\end{aligned} \tag{5}$$

Thus, to evaluate possible sample sizes, we can take two approaches: simulations based on Equation 4 and analytic results based on Equation 5.

Criteria for choosing sample size

Coefficient of variation

The coefficient of variation is defined as

$$\mathrm{CV} = \frac{\sqrt{\mathrm{V}(\widehat{p})}}{\bar{p}} \tag{6}$$

The coefficient of variation is typically useful, because it describes the accuracy of the estimate as a fraction of the estimate itself. For example, we might specify that we want the standard errors for our estimate to be about five percent of the estimate itself; this corresponds to a coefficient of variation of $CV = 0.05$. For very rare populations, though, this can result in very large sample sizes: a standard error that is within five percent of $p = 0.025$ is very small because 0.025 is small to begin with (and so five percent of 0.025 is extremely small).

Margin of error

An alternative is to specify how wide we want the (sampling-based) confidence intervals to be, regardless of how big the estimate is; for example, we might specify that we want to be able to estimate the fraction of women who had an abortion with a confidence interval of plus or minus 1 percent, regardless of how big we estimate the group to be; this is equivalent to specifying a margin of error of $e = .01$.

If the true size of the group is very small (ie, $p = 0.025$), then this would mean that we are willing to accept a confidence interval like $(0.0125, 0.0325)$.

More specifically, the margin of error, e, is defined as the half-width of a confidence interval. If the sample size is large enough for the estimator $\widehat{p}$ to be approximately Normally distribution, and for a conventional 95% confidence interval,

$$e = 1.96\sqrt{\mathrm{V}(\widehat{p})}.$$

Sample size results

Analytic - coefficient of variation

If we pick a target coefficient of variation CV_0, we can use the analytical approximation to the sampling variance (Equation 4) to solve for the implied sample size:

$$\begin{aligned} \mathrm{CV}_0^2 &= \frac{\mathrm{V}(\widehat{p})}{\bar{p}^2} = \frac{1}{pn\bar{d}} \\ \leftrightarrow n &= \frac{1}{p\ \bar{d}\ \mathrm{CV}_0^2} \end{aligned}$$

For example, with $p_{\text{true}} = 0.025$, $\tau = 0.66$, $\bar{d} = 1.5$, and a target $\mathrm{CV}_0 = 0.05$, this formula suggests a sample size of $n = 16,161$. This is very large because $p = p_{\text{true}}\tau$ is very small: we need a large sample to get a small relative standard error.

Analytic - margin of error

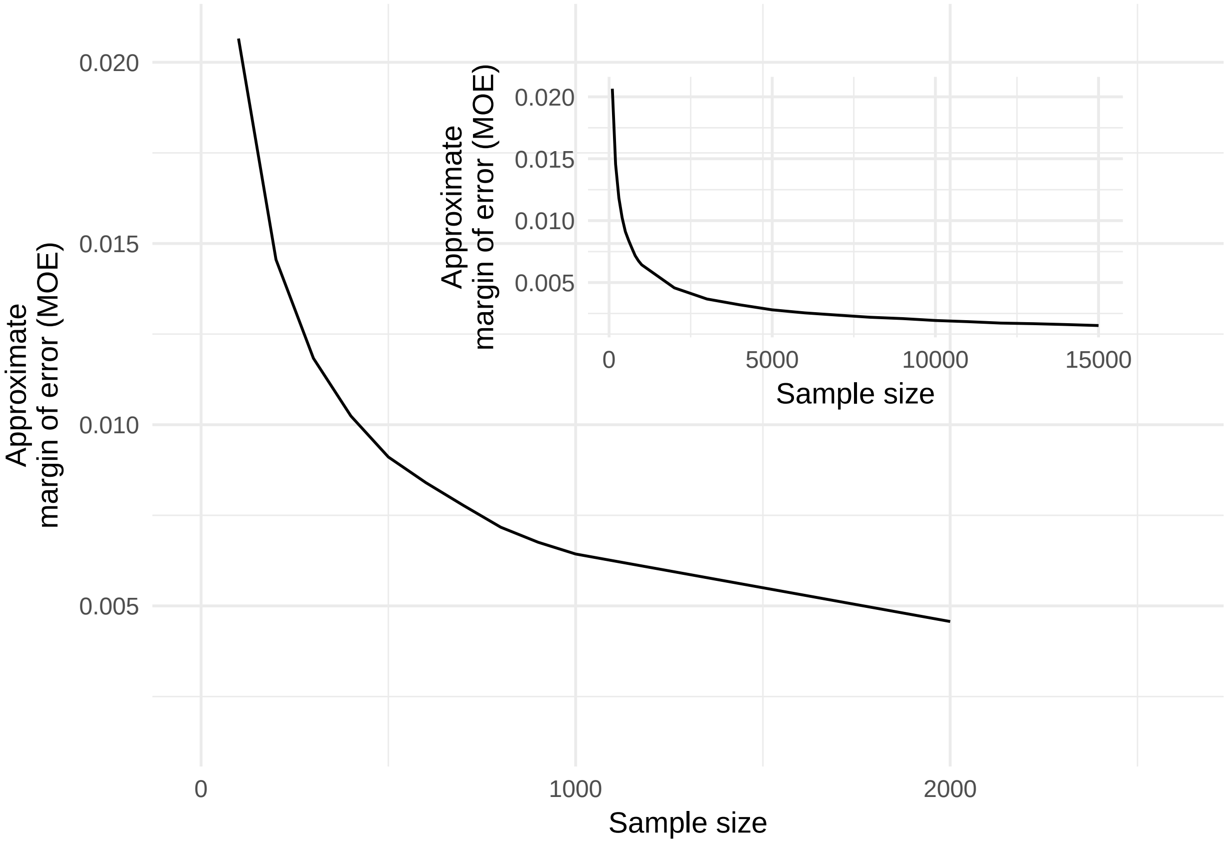
If we pick a target margin of error e_0, we can use the analytical approximation to the sampling variance (Equation 4) to solve for the implied sample size:

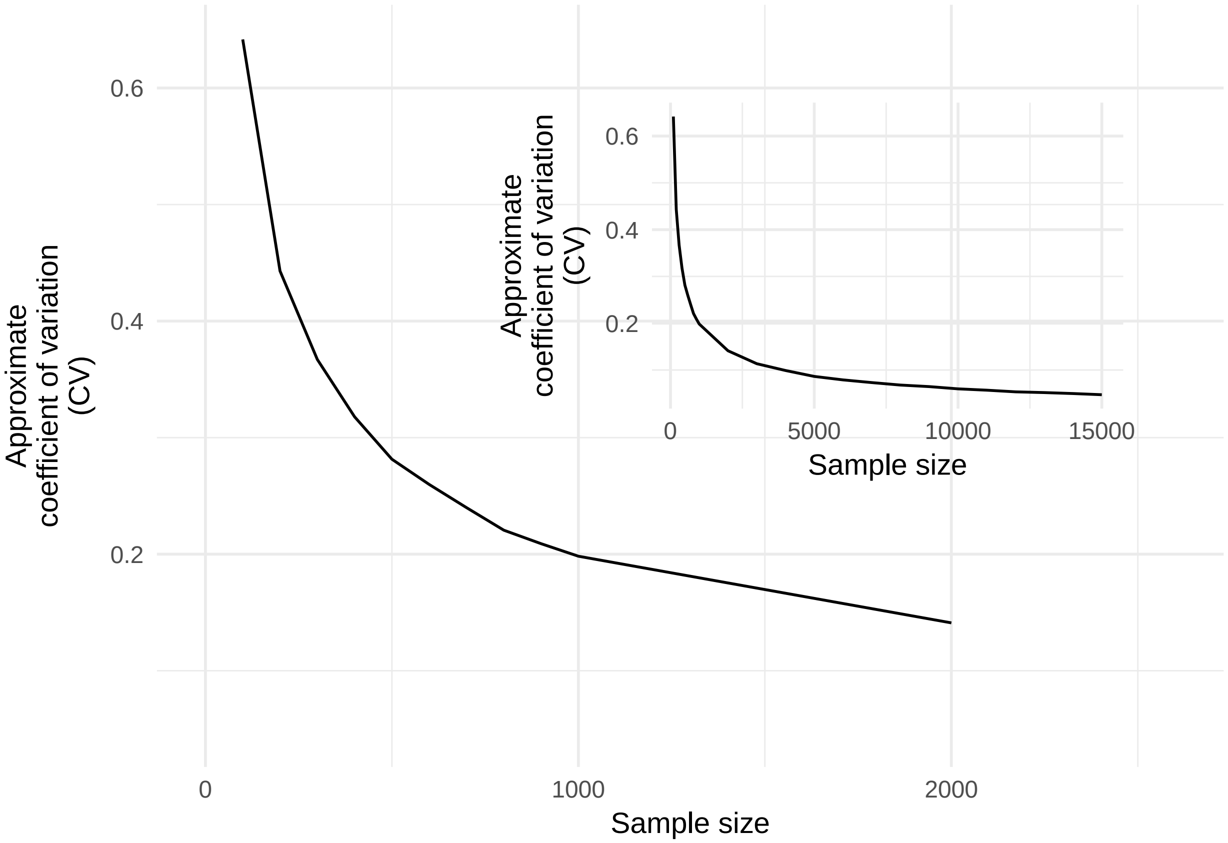
$$\begin{aligned} e_0^2 &= (1.96^2)\mathrm{V}(\widehat{p}) = (1.96^2)\frac{p}{n\bar{d}} \\ \leftrightarrow n &= (1.96^2)\frac{p}{e_0^2\ \bar{d}} \end{aligned}$$

For example, with $p_{\text{true}} = 0.025$, $\tau = 0.66$, $\bar{d} = 1.5$, and a target $e_0 = .01$, this formula suggests a sample size of $n = 423$. This number is much smaller than the one we obtained based on a coefficient of variation of 0.05; this is because for $p = (0.025)(.66)$, the coefficient of variation that corresponds to a margin of error of 0.01 is $\frac{1}{\sqrt{p\ n\ \bar{d}}} = 0.3$. This looks huge, but that is because p is so small.

Simulation results

The plots below show the results of the simulation described above. The curves show the approximate relationship between sample size and (i) the margin of error; and (ii) the coefficient of variation.




Approximate
coefficient of variation
(CV)
0.6
0.4
0.2
0
1000
2000
Sample size
Approximate
coefficient of variation
(CV)
0.6
0.4
0.2
0
5000
10000
15000
Sample size