Optimal Data Collection When Strata and Strata Variances Are Known*

Ken Cor† Gaurav Sood‡

July 7, 2021

What's the least amount of data you need to collect to estimate the population mean with a particular standard error? For the simplest case—estimating the mean of a binomial variable using simple random sampling, a conservative estimate of the variance ($p = .5$), and a $\pm 3\%$ confidence interval—the answer ($n \sim 1,000$) is well known. The simplest case, however, assumes little to no information. Often, we know more. In opinion polling, we generally know sociodemographic strata in the population. And we have historical data on the variability in strata. Take, for instance, measuring support for Mr. Obama. A polling company like YouGov will usually have a long time series, including information about respondent characteristics. Using this data, the company could derive how variable the support for Mr. Obama is among different sociodemographic groups. With information about strata and strata variances, we can often poll fewer people (vis-a-vis random sampling) to estimate the population mean with a particular s.e. In this note, we show how.

Let's say that we are interested in estimating the population mean (x) within a particular error bound. Let's assume that there are two strata in the population: a and b. Let's say that the proportion of a in the population is w_a and the proportion of b is $1 - w_a$. The corresponding

*All the code for the note is posted on https://github.com/soodoku/optimal_data_collection/

†Ken can be reached at mcor@ualberta.ca

‡Gaurav can be reached at: gsood07@gmai

standard deviation for a and b is σ_a and σ_b. And let n_a and n_b denote the sample size of groups a and b and let n denote the total sample size. Finally, let $\sigma_{\bar{x}}$ denote the s.e. of $\bar{x}$, our estimand.

The s.e. of the mean is given by equation 1:

$$\sigma_{\bar{x}} = \sqrt{\frac{w_a^2 * \sigma_a^2}{n_a} + \frac{(1-w_a)^2 * \sigma_b^2}{n_b}} \tag{1}$$

If the n, σ_a, σ_b, and w_a are known, the formula for optimal allocation across strata is well known. The intuition behind the formula is straightforward—allocation depends on how large the proportion of the population a strata is and how variable it is. The formulas for n_a and n_b are given by equations 2 and 3 respectively.

$$n_a = \frac{n w_a \sqrt{\sigma_a^2}}{w_a \sqrt{\sigma_a^2} + (1-w_a)\sqrt{\sigma_b^2}} \tag{2}$$

$$n_b = \frac{n w_b \sqrt{\sigma_b^2}}{w_a \sqrt{\sigma_a^2} + (1-w_a)\sqrt{\sigma_b^2}} \tag{3}$$

Doing a bit of algebra using equations 1, 2, and 3 allows us to express n as a function of w_a, σ_a, σ_b, $\sigma_{\bar{x}}$ (see equation):

$$n = \frac{w_a^2\sigma_a^2 + 2w_a(1-w_a)\sqrt{\sigma_a^2\sigma_b^2} + (1-w_a)^2\sigma_b^2}{\sigma_{\bar{x}}^2} \tag{4}$$

We provide a script that allows users to calculate n given $\sigma_{\bar{x}}, \sigma_a^2, \sigma_b^2, w_a$. The script also includes a function that provides a way to use constrained optimization to solve for the two

unknowns (n and proportion of n_a versus n_b) without using the optimal allocation formula.

The benefit of using optimal allocation

In a realistic example, we find the benefit of using optimal allocation over simple random sampling is 6.5% (see code block 1). We assume two groups with $w_a = .8$, $\sigma_a^2 = .25$, and $\sigma_b^2 = .16$. The simple random sampling formula that only uses the population variance would need us to sample 1095 people to get us a mean with a standard error of .015. With our allocation rule, you only need to sample 1024 people. The net benefit is $\frac{(1095-1024)}{1095} = .065$.[1]

```
## Benefit of Using Optimal Allocation Rules
## wa = .8
## vara = .25; pa = .5
## varb = .16; pb = .8
## SRS: pop_mean of .8*.5 + .2*.8 = .56

# sqrt(p(1 -p)/n) = .015
# n = p*(1- p)/.015^2 = 1095

# optimal_n_plus_allocation(.8, .25, .16, .015)
#    n    na    nb
#1024   853   171
```

Listing 1: Benefit of Using Optimal Sampling

[1] We are aware that with a binary variable, the variance and the mean are mechanically linked. And with a normal distribution, we will have a more compelling illustration.

What's the next best data point to collect when you know the strata and strata variances?

You can adapt the equations above to answer a more subtle question—what's the next best data point to collect when you know the strata and strata variances? Let's say that once again, we want to measure support for Mr. Obama. Let's assume that we have information about different strata in the population and know the variability of the response in each stratum. Let's say that our objective is to estimate the population mean with the smallest confidence interval. If I could collect only one additional data point, which strata would I sample from? Once again, the heuristic is that the greater reduction in error will come from collecting data from the stratum where the responses are the most unpredictable, pro-rated by how big the stratum is. As above, we provide a script that allows users to calculate the strata from which the next point should come.

Future Work

In this note, we solved a simple version of the problem. The simplicity comes at the cost of realism. A more realistic version of the problem may be as follows:

- We want to estimate the mean of x at time t.
- We know strata proportions and historic strata variances (before time t).

We design a sampling strategy for 1 given 2. But as new data comes in, we find that the new data differs 'substantially' from the old data. How do we dynamically adapt to a sampling scheme that takes less and less account of the historical strata variances and gets the sampling scheme closer to stratified random sampling?

You could also extend the work in other compelling directions. Rather than solve for the smallest n and optimal allocation when calculating the population mean, you could compute the

optimal data collection strategy for experiments when heterogeneity in treatment effects by strata is known.