Three Stage Sampling

Doug

2017-10-04

One of IDinsight's project teams is in the process of designing the sampling strategy for a large scale household survey and is considering using a three stage sampling design in which they would first select districts, then villages (or urban wards), and then households. In addition, someone was asking about three stage clustering for an RCT somewhere on Slack (I can't seem to find the slack post now) so I thought it might be useful to write a short post on three stage designs.

In this post, I'll try to answer four questions: 1. When do you need to take into account both stages of clustering in a survey or evaluation? 2. How do you properly account for a three stage design when performing sample size / power calculations? 3. How should you estimate the inputs required for these calculations? 4. How do you properly account for a three stage design when analyzing data?

When you do you need to take into both stages?

With an RCT, it's pretty rare that you really need to take into account two stages of clustering. Remember that just because units exhibit some sort of clustering doesn't mean that you need to adjust for clustering in your analysis. For example, if you randomize at the student level it doesn't matter that student learning outcomes exhibit clustering at the classroom level. An example of when you might want to take into account two stages of clustering is when you randomize large clusters (e.g. schools) and then only collect data from units in a randomly sampled set of smaller clusters (e.g. kids in classrooms). With surveys, anytime you have a three stage design you should theoretically take into account the clustering at both levels.

Even when it makes sense in theory to take into account both stages of clustering, you can usually get by with just considering the most aggregate (highest) level of clustering. We'll see below why that makes sense. In some cases, e.g. when you are trying to find the optimal survey design for a given budget, you do really need to take into account both stages of clustering.

How do you account for a three stage design?

(The advice given below is tailored to someone performing sampling size calcs for a survey with a three stage design. All of the advice holds true for power calcs as well. You just need to multiply the final variance by 2 (since you have 2 groups – treatment and control) and then use the standard adjustment to the standard error for power calcs – i.e. instead of multiplying the standard error by +/-1.96 to create a 95% confidence interval you multiply by ~2.8 to calculate an MDE for alpha .05 and power .8.)

Let's first recap how one stage of clustering affects the variance of your estimator. Let's say that you will use a two stage sampling strategy in which you will first randomly sample J clusters and then randomly sample K units from each cluster to estimate the mean of some variable y. Further assume that the total number of units per cluster does not vary and is pretty large. If values of y are correlated within each cluster, we can think of the values for y as being made up of a cluster component and an independent within-cluster component, i.e.

$$y_{j,k} = \eta_j + \phi_{j,k}$$

This allows us to calculate the variance the of y as:

$$\sigma_y^2 = \sigma_\eta^2 + \sigma_\phi^2$$

And the variance of the mean as:

$$Var(\bar{y}) = \frac{\sigma_\eta^2}{J} + \frac{\sigma_\phi^2}{JK} = \sigma_y^2 \left(\frac{\rho}{J} + \frac{(1-\rho)}{JK} \right)$$

Where $\rho = \frac{\sigma_\eta^2}{\sigma_y^2}$. It's also useful to calculate the design effect, or the ratio of the variance of this estimator to the ratio of the estimator if the sample had been collected using simple random sampling (SRS). Since the variance under SRS would be $\frac{\sigma_y^2}{JK}$ the design effect=$1 + (K-1)\rho$.

Let's now suppose that we have a higher level sampling stage. We first pick Q mega-clusters, then J clusters from each mega-cluster, and then K households from each cluster. Similarly, we can think of the values y as made of three components:

$$y_{q,j,k} = \gamma_q + \eta_{q,j} + \phi_{q,j,k}$$

The variance of y is then:

$$\sigma_y^2 = \sigma_\gamma^2 + \sigma_\eta^2 + \sigma_\phi^2$$

And the variance of the mean is:

$$Var(\bar{y}) = \frac{\sigma_\gamma^2}{Q} + \frac{\sigma_\eta^2}{QJ} + \frac{\sigma_\phi^2}{QJK} = \sigma_y^2 \left(\frac{\rho_\gamma}{Q} + \frac{\rho_\eta}{QJ} + \frac{(1-\rho_\gamma-\rho_\eta)}{QJK} \right)$$

Where $\rho_\eta = \frac{\sigma_\eta^2}{\sigma_y^2}$ and $\rho_\gamma = \frac{\sigma_\gamma^2}{\sigma_y^2}$. For our three stage sampling design, the design effect is:

$$DEFF = 1 + (JK-1)\rho_\gamma + (K-1)\rho_\eta$$

This also shows why just looking at the most aggregate level of clustering is usually pretty reasonable – assuming the two ICCs are relatively similar in size, the adjustment to the variance will be driven primarily by the most aggregate level of clustering.

The formula above ignores the finite population correction. With multi-stage sampling, we often want to take into account the finite population correction for at least one stage. We can do this using the

$$DEFF = 1 + (f_\gamma JK - f_\phi)\,\rho_\gamma + (f_\eta K - f_\phi)\,\rho_\eta$$

Where f_γ, f_η, and f_ϕ are the finite population corrections at each stage. For example $f_\gamma = 1 - \frac{Q}{\sum Q}$ – i.e. 1 minus the proportion of total mega-clusters sampled. If we wish to ignore any of the FPCs, you can replace them with 1.

How should you estimate the inputs required for these calculations?

Estimating ICCs at each level of clustering is a bit tricky. You need to find a dataset that has both levels of clustering that you are interested in as well as the variable you are interested in (or a similar variable).

If you have such a dataset, you can estimate the ICC components using Stata's xtmixed command followed by the user written iccvar command and specifying a random effect for each level of clustering you are interested in as well as each level of clustering used in the sampling design for the survey. For example, suppose you have data from a survey which has info on the mega cluster ID and cluster ID, which used a two stage sampling design in which PSUs were selected and households randomly selected within PSUs, and PSUs are nested within clusters. Then you could estimate these components using the following command:

```
xtmixed y_var || megacluster_id: || cluster_id: || psu_id:

iccvar
```

Note that the standard errors on the estimates of the components are probably going to be pretty large so it may be useful to run some sensitivity analyses on your estimates.

How do you properly account for a three stage design when analyzing data?

Unfortunately, most Stata commands only allow for a single stage of clustering. To account for two or more stages of clustering, you need to first "svyset" your data and then use the "svy" prefix before running any command.