Ch. 5: Optimal sample size allocation for cluster sampling

Math 255, St. Clair

Determining sample sizes for a cluster sample

Problem: You have a quantitative variable y and you want to estimate its population mean/total.

Question 1: How many SSU (elements) to sample?

Question 2: How many PSU (clusters) to sample?

Optional: How to do this in an optimal way to (a) achieve a desired margin of error or (b) not exceed by fixed survey budget?

- An optimal solution is "easily" computable assuming equal cluster sizes!
  - M's are the equal and m's are equal

Optimal Allocation

This allocation is **optimal** because it both

- **minimizes costs** for a fixed SE/margin of error, *or*
- **minimizes SE/margin of error** for a fixed survey cost.

Mathematical Problem:

- Let c_1 be the cost per PSU (cluster) and c_2 be the cost per SSU (element). With c_0 fixed costs, the total survey costs are

$$C(m, n) = c_0 + c_1 n + c_2(mn)$$

- Variance is also a function of m and n and ANOVA MS.

$$\begin{aligned} V(\hat{\bar{y}}_{unb}; m, n) &= \frac{N^2}{(NM)^2}\left(1 - \frac{n}{N}\right)\frac{S_t^2}{n} + \frac{N}{n(NM)^2}\sum_i M^2\left(1 - \frac{m}{M}\right)\frac{S_i^2}{m} \\ &= \left(1 - \frac{n}{N}\right)\frac{MSB}{nM} + \left(1 - \frac{m}{M}\right)\frac{MSW}{nm} \end{aligned}$$

Optimal Allocation: 1. SSU sample size

Solution: Use Lagrange Multiplier method to minimize one function (C or V) subject to the contraints of the other function.

- The optimal SSU (element) sample size is

$$m_{opt} = \sqrt{\frac{c_1 M(N-1)(1-R_a^2)}{c_2(NM-1)R_a^2}} \approx \sqrt{\frac{c_1(1-R_a^2)}{c_2 R_a^2}} \text{ when } N >> M$$

where $R_a^2 = 1 - \frac{MSW}{S^2}$ is (roughly) the proportion of variability in y explained by the clusters

- sample lots of SSU when
  - PSU are more expensive, $c_1 > c_2$
  - clusters are heterogeneous, $R_a^2 < 0.5$

Optimal Allocation: 2. PSU sample size:

(a) achieving a margin of error

Problem: How many PSU to sample to estimate $\bar{y}_{\mathcal{U}}$ with $(1-\alpha)100\%$ confidence and a margin of error $e = z_{\alpha/2} SE(\hat{\bar{y}}_{unb})$?

Solution: Get optimal SSE size m_{opt}, if you ignore the FPC then

$$n_{opt} = \frac{\nu z^2_{\alpha/2}}{e^2} \quad \text{where} \quad \nu = \frac{MSB}{M} + \left(1 - \frac{m_{opt}}{M}\right)\frac{MSW}{m_{opt}}$$

- If N is smaller, don't ignore FPC and use:

$$n_{opt} = \frac{\nu z^2_{\alpha/2}}{e^2 + \frac{z^2 MSB}{NM}}$$

- To estimate t with e_t margin of error, just set $e = e_t/(NM)$.

Optimal Allocation: 2. PSU sample size:

(b) Do not go over budget

Problem: How many PSU to sample if your budget is C dollars (or man hours, etc...)?

Solution: Get optimal SSE size m_{opt}, then

$$n_{opt} = \frac{C - c_0}{c_1 + c_2 m_{opt}}$$

Optimal Allocation

- The previous solutions for n are optimal when m_{opt} is used
  - but you can use any m to obtain a given ME or cost, but it will not minimize the value of the other function
- You need a guess at MSB and MSW
  - guess at variability of cluster totals: $MSB = S_t^2/M$
  - guess at variability of within clusters: $MSW = \sum_i^N S_i^2/N$

Example: Dorms

- New GPA study: want to estimate average dorm GPA with a 95% ME of 0.2
  - $N = 100$ rooms with $M = 4$ students per room
- Previous study: One-stage example 1(b)
  - $msw = 0.18504$, $msb = 0.56392$ and $\hat{S}^2 = 0.279$
  - $\hat{R}_a^2 = 1 - 0.18504/0.279 \approx 0.337$
- Costs? $c_1 = 2$ minutes to travel between rooms and $c_2 = 1$ minute to talk to each student.
- We want to minimize cost and get a ME of $e = 0.2$

Example: Dorms

- SSU sample size:

$$m_{opt} = \sqrt{\frac{2(4)(100-1)(1-0.337)}{1(400-1)(0.337)}} \approx 1.98 \approx 2$$

- Sample 2 students per room

```
> (m_opt <- sqrt(2*4*(100-1)*(1-0.337)/(1*(400-1)*0.337)) )
[1] 1.976141
```


Example: Dorms

- PSU sample size:

$$\nu = \frac{0.56392}{4} + \left(1 - \frac{2}{4}\right)\frac{0.18504}{4} \approx 0.18724$$

- Using FPC:

$$n_{opt} = \frac{(0.18724)1.96^2}{0.2^2 + \frac{1.96^0.56392}{400}} \approx 15.8 \approx 16$$

- Optimal solution: Sample 16 rooms, 2 students per room.

```
> (nu <- 0.56392/4 + (1-2/4)*0.18504/2 )
[1] 0.18724
> (n_opt <- 1.96^2*nu/(.2^2 + 1.96^2*0.56392/400) )
[1] 15.8381
```

Example: Dorms - check answer

- We used $z = 1.96$ for 95% confidence, but we should be using a t-distribution with $n-1$ degrees of freedom for CI when n is "small"

- Recheck solution with our larger multiplier, suggests a larger n

```
> qt(.975, df= 16-1)
[1] 2.13145
> (n_opt <- qt(.975,15)^2*nu/(.2^2 + qt(.975,15)^2*0.56392/400) )
[1] 18.33097
```

- Using a more accurate multiplier says we need n above 18!
- Try using $n = 19$

```
> qt(.975, df= 19-1)
[1] 2.100922
> (n_opt <- qt(.975,18)^2*nu/(.2^2 + qt(.975,18)^2*0.56392/400) )
[1] 17.87983
```

- Final answer: $n = 19$ will give a ME of at most 0.2

Optimal Allocation: Unequal cluster sizes

- What if your clusters are different sizes?!

  - if clusters are **not too variable**, the (almost) optimal solution could use $\bar{M}$ to get m_{opt}

- If clusters sizes are variable, don't use the optimal solution for equal sizes!