Optimal sample size allocation for cluster sampling

Week 7 (5.4)

Stat 260, 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.

obs. units

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?

Optimal Allocation

This allocation is **optimal** because it either

- **minimizes costs** for a fixed SE/margin of error, *or*
- **minimizes SE/margin of error** for a fixed survey cost.
- An optimal solution is "easy" to derive assuming equal cluster sizes:
  - $M_i = M$: cluster sizes are equal
  - $m_i = m$: cluster sample sizes are equal

Optimal Allocation

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

$$C(m,n) = c_0 + c_1 n + c_2(\underbrace{mn}_{\text{SSU}})$$

PSU

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

$$V(\hat{\bar{y}}_{unb}; m, n) = \left(1 - \frac{n}{N}\right)\frac{MSB}{nM} + \left(1 - \frac{m}{M}\right)\frac{MSW}{nm}$$

algebra

between

within

S_t

S_i

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 $R_a^2 = 1 - \dfrac{MSW}{S^2}$ $\Rightarrow$ homogeneity

- if $c_1 > c_2$, then m_{opt} is bigger $R_a^2 < \frac{1}{2}$
- if clusters are heterogenous (R_a^2 smaller) then m_{opt} is bigger
- if clusters homogeneous, $R_a^2 \approx 1$ + $m_{opt} \approx 0$
  $\rightarrow$ very few SSU/cluster need to be sample b/c response in each cluster are very similar.

Optimal Allocation

- We know m_{opt}
- final sample size is then determined by n:

$$n \times m_{opt} = \text{number of observation units sampled}$$

Question 2: Determine n subject to a constraint:

- fixed SE/margin of error, *or*
- fixed survey cost.

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 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:

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

- To estimate t with e_t margin of error, 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 m_{opt}, then

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

Optimal Allocation

- Note: The **cost** and **ME** solutions for n work for *any* values of m given a desired cost or ME.

- You need a guess at MSB and MSW
  - $MSB = S_t^2/M$
    - S_t: how to cluster totals vary?
  - $MSW = \sum_i^N S_i^2/N$
    - S_i: within cluster variation?

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
    - $msw = 0.18504, msb = 0.56392$ and $\hat{S}^2 = 0.279$
    - $\hat{R}_a^2 \approx 0.337$
- Costs?
    - $c_1 = 20$ minutes to travel between rooms and
    - $c_2 = 10$ minute to talk to each student.

Example: Dorms

What is the optimal number of student to sample per room?

SSU??

$$m_{opt} = \sqrt{\frac{c_1 M(N-1)(1-R_a^2)}{c_2 (NM-1) R_a^2}} = \sqrt{\frac{20(4)(100-1)(1-.337)}{10(4 \cdot 100-1)(.337)}}$$

$$\approx 1.98 \rightarrow m_{opt} = 2 \text{ students/room}$$

Example: Dorms

How many rooms to sample to get a ME of $e = 0.2$ for estimating mean GPA?

$m_{opt} = 2$, $n = ??$

$$v = \frac{.56392}{4} + \left(1 - \frac{2}{4}\right)\frac{.18504}{2} \approx .18724$$

$$n_{opt} = \frac{(.18724)(1.96)^2}{\underbrace{.2^2}_{e^2} + \frac{1.96^2(.56392)}{4 \cdot 100}} \approx 15.8 \longrightarrow$$

$n_{opt} = 16$

?? what if use t-dist to get ME

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"
- Check margin of error with our larger multiplier, suggests a larger n

```
> n <- 16
> qt(.975, df= n-1)
[1] 2.13145
> se_squared <- (1-n/100)*0.56392/(n*4) + (1-2/4)*0.18504/(n*2)
> qt(.975, df= n-1)*sqrt(se_squared) # 0.2 or less??
[1] 0.2162418
```

→ use instead $z = 1.96$

→ Var

→ ME = .22

n = 16

Example: Dorms - check answer

- Try n of 17, 18 and 19!

ME

```
> n <- c(17,18,19)
> se_squared <- (1-n/100)*0.56392/(n*4) + (1-2/4)*0.18504/(n*2)
> qt(.975, df= n-1)*sqrt(se_squared) # 0.2 or less??
[1] 0.2077542 0.2000704 0.1930670
```

17 18 19

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

Example: Dorms

(forget ME n_{opt})

How many rooms to sample if we have a fixed cost of 300 minutes?

$c_1 = 20$ $c_2 = 10$ $m_{opt} = 2$ $n = ??$

$$n_{opt} = \frac{300}{20 + 10(2)} = 7.5$$

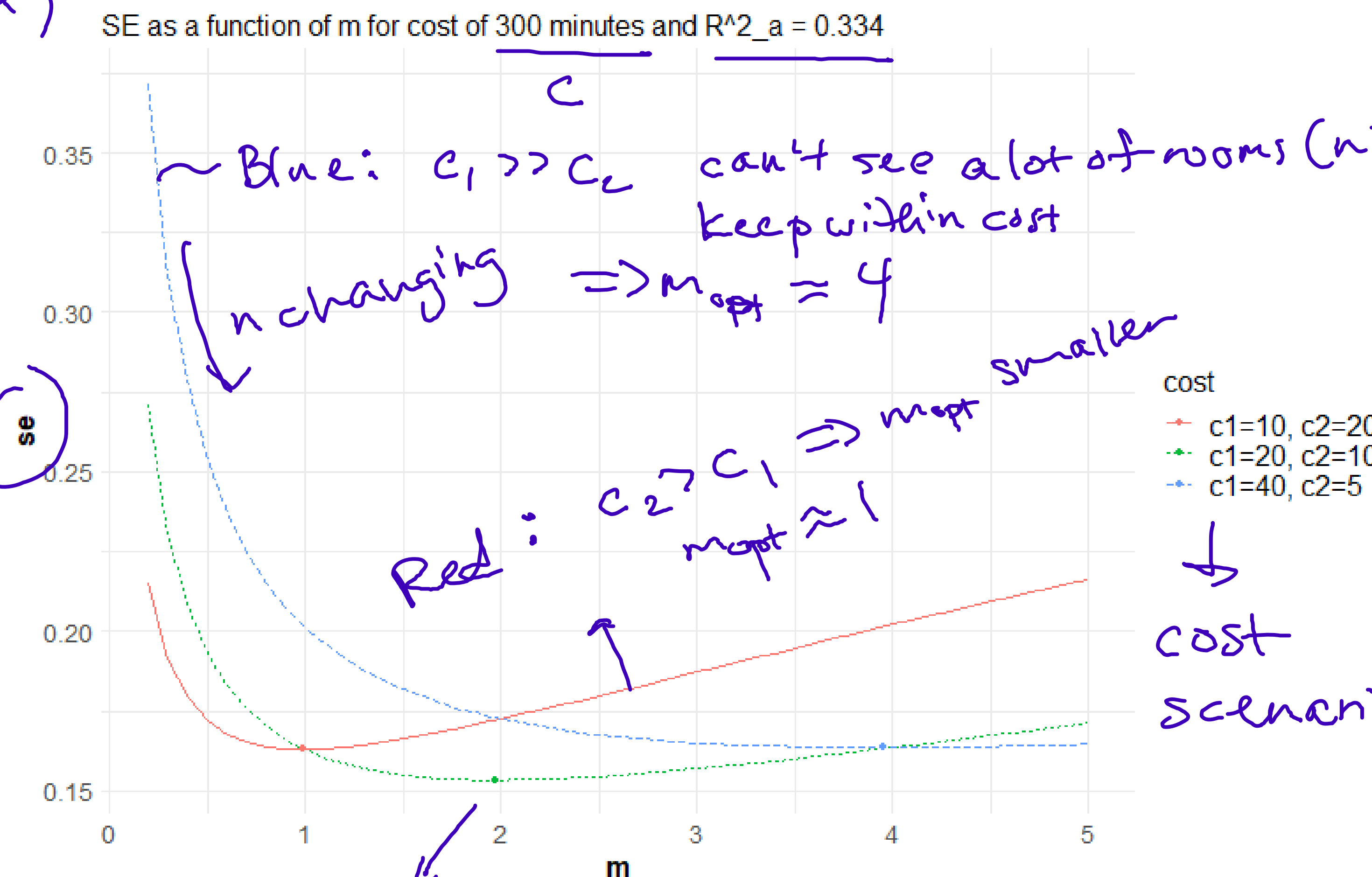
→ 7?
→ 8?

$\boxed{\text{cost for } n = 7 \text{ rooms : 280 min.}}$

$n = 8$ rooms : 320 min

Example: Dorms

(5.4)



C

Blue: $c_1 >> c_2$ can't see a lot of rooms (n) + keep within cost

n changing $\Rightarrow m_{opt} \approx 4$

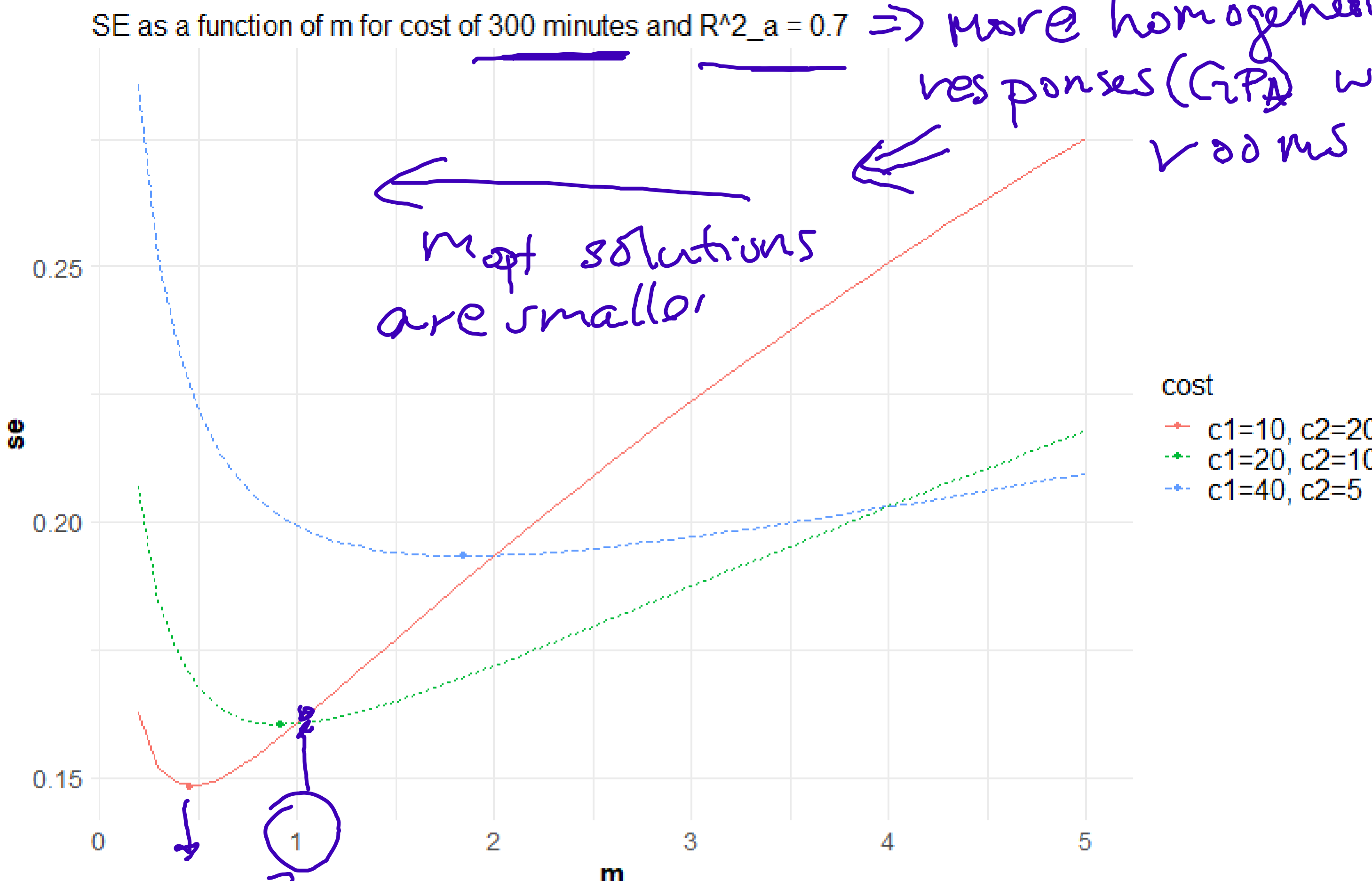
Red: $c_2 > c_1 \Rightarrow m_{opt}$ smaller

$m_{opt} \approx 1$

cost scenarios

$m_{opt} \approx 1.98$

Example: Dorms with $R_a^2 = 0.7$



⇒ more homogeneous responses (GPA) within rooms.

m_{opt} solutions are smaller

$\frac{1}{2}$

Optimal Allocation: Unequal cluster sizes

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

wrt size

avg. cluster size

  - use m_{opt} for all clusters or
  - use an average of m_{opt} → $\frac{m_i}{M_i} \approx$ constant $\Rightarrow \boxed{m_i = M_i \left(\frac{m_{opt}}{\bar{M}}\right)}$

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

e.g. California Schools

$1 \le M_i \le 72$

→ variable sizes!