Optimal sample size allocation

Week 4 (3.4)

Stat 260, St. Clair

Review: Tradeoff: Cost vs. Precision

As n (sample size) increases:

- SE's get decrease (more precise) but
- sampling costs increase

Stratified problem:

Issue: **Both** costs and precision can depend on how we **allocate** our overall sample size to each stratum

- Strata may be more/less costly to sample
- Measurements within stratum may have different SDs S_h

- The **allocation** fraction for stratum h is

$$a_h = \frac{n_h}{n} \quad \Rightarrow \quad n_h = a_h n$$

- Must have $\sum_{h=1}^{H} a_h = 1$

Determining sample sizes for a stratified sample

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

Question 1: If I sample n units total, what fraction of these units should be taken from stratum h?

Solution 1: Determine the **allocation fraction** a_h for each stratum.

$$a_h = \frac{n_h}{n}$$

Determining sample sizes for a stratified sample

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

(Optional) Question 2: How many units should be selected to **either**

(a) achieve a desired margin of error or

(b) not exceed by fixed survey budget?

Solution 2: Determine the total sample size n.

Q1. Sample size allocation

Goal: Determine the allocation fractions $a_1, a_2, \dots, a_H$ for all strata to get sample sizes:

$$n_h = na_h$$

- **Optimal allocation:**

  (a) minimize cost (sample size) for a fixed margin of error **OR**

  (b) minimize the margin of error for a fixed cost (sample size).

Q1. Sample size allocation

Goal: Determine the allocation fractions $a_1, a_2, \dots, a_H$ for all strata to get sample sizes:

$$n_h = na_h$$

- **Optimal allocation:** (a) minimize cost (sample size) for a fixed margin of error **OR** (b) minimize the margin of error for a fixed cost (sample size).
  - **Neyman allocation:** special case of optimal when all stratum **costs** are the same.
- **Proportional allocation:** $a_h = \dfrac{n_h}{n} = \dfrac{N_h}{N}$
  - This is optimal when stratum **costs** and **variances** are the same.
  - Use if the stratum SDs S_h are not known.
- Any other allocation that satisfies $\sum_{h=1}^{H} a_h = 1$.

Q1. Optimal Allocation

This allocation is **optimal** because it

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

Mathematical Problem:

- Let c_h be the cost of sampling one unit from stratum h and c_0 are your fixed costs. Total survey costs are

$$C(\{a_h\}, n) = c_0 + \sum_{h=1}^{H} c_h(na_h)$$

- Variance is also a function of $\{a_h\}$ and n, e.g. variance for estimated mean:

$$V(\{a_h\}, n) = \sum_{h=1}^{H} \left(1 - \frac{na_h}{N_h}\right) \left(\frac{N_h}{N}\right)^2 \frac{S_h^2}{na_h}$$

Q1. Optimal Allocation

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

- The optimal allocation fraction is

$$a_h = \frac{N_h S_h / \sqrt{c_h}}{\sum_{k=1}^{H} N_k S_k / \sqrt{c_k}} \quad \text{where } S_h = \text{ pop. SD in stratum } h$$

- Highest allocation for strata with
    - high variability S_h,
    - large size N_h, or
    - low costs c_h.

Q1. Neyman Allocation

Neyman allocation is an **optimal allocation** if you assume the cost per observation are the same for all strata $c_1 = c_2 = \cdots = c_H$.

- The Neyman allocation fraction is

$$a_h = \frac{N_h S_h}{\sum_{k=1}^{H} N_k S_k}$$

- Use this allocation if if costs c_h are unknown.

Q1. Proportional Allocation

Proportional allocation is an **optimal allocation** if the cost per observation and SDs are the same for all strata:

- $c_1 = c_2 = \cdots = c_H$ and
- $S_1 = S_2 = \cdots = S_H$.
- The proportional allocation fraction is

$$a_h = \frac{N_h}{N}$$

- Use this allocation if you don't have good guesses of the within stratum SD's S_h and costs are unknown or equal.
    - May not be optimal, but it is usually better than SRS.

2. Determining total sample size: (a) achieving a margin of error

Problem: what is n to estimate $\bar{y}_{\mathcal{U}}$ with $(1-\alpha)100\%$ confidence and a margin of error $e = z_{\alpha/2} SE(\bar{y}_{str})$?

Solution: Get allocations a_h's, if you ignore the FPC then

$$n_0 = \frac{\nu z^2_{\alpha/2}}{e^2} \quad \text{where} \quad \nu = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{S_h^2}{a_h}$$

- If your stratum population sizes are smaller, don't ignore FPC and use:

$$n = \frac{n_0}{1+D} \text{ where } D = \frac{z^2_{\alpha/2} \sum_{h=1}^{H} N_h S_h^2}{N^2 e^2}$$

- To estimate t with e_t margin of error, just set $e = e_t / N$.
- $\star$ If **optimal allocation** is used to determine a_h's, then you will **minimize the cost** of achieving this margin of error.

2. Determining total sample size: (b) Do not go over budget

Problem: what is n if your budget is C dollars (or man hours, etc...)?

Solution: Get allocations a_h's, then

$$n = \frac{C - c_0}{\sum_{h=1}^{H} c_h a_h}$$

- $\star$ If **optimal allocation** is used to determine a_h's, then you will **minimize the SE** of your estimate (and M.E.) while not exceeding your fixed budget C.

What about a Population Proportion?

- What if your variable of interest is categorical?
- All previous formulas apply but let

$$S_h \approx \sqrt{p_h(1 - p_h)}$$

Example

Suppose we know this about our population's heights:

strata	Nh	pop.mean	pop.var
Female	68	65.49	8.35
Male	60	70.64	11.73

What is the optimal allocation of if you know that sampling from the male stratum will cost twice as much as sampling from the female stratum?

Example

What are the optimal sample sizes if you want to sample a total of 40 people?

Example

Suppose it costs $1 to sample from the female stratum. Compute the cost and SE for estimating the population mean using the optimal sample sizes when $n = 40$.

Example

Compute the cost and SE for estimating the population mean using *proportional allocation* when $n = 40$.

Example

If you want to fix costs at $59, what is the SE of the optimal allocation? (and compare to proportional allocation)

Example

Suppose you want to estimate the average height in the population to within 0.5 inches with 95% confidence. Assuming that stratum costs are equal, what stratum sample sizes should you use?

When is optimal not actually optimal

- You may have an optimal solution for one variable but not others
  - E.g. An optimal solution for student height may not be optimal for student GPA.

- If your goal is to estimate with a fixed precision (MOE) **within** strata
  - Use SRS sample size calculations from ch. 2