Ch. 3: Optimal sample size allocation

Math 255, St. Clair

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.

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

(Optional) Question 2: How many units should be selected to (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 $a_1, a_2, \ldots, a_H$ 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:** special case of 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 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_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 arethe 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 $e = e_t/N$.
- ⋆ 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

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

where p_h is an educated guess at the population proportion within stratum h.