Friday, Sep 6

Optimum Allocation

Recall that *allocation* concerns specifying the sample sizes in a stratified sampling design with L strata — i.e., $n_1, n_2, \dots, n_L$. A couple of things we can take into consideration are (a) the *bound* of the error of estimation and (b) the *cost* of the survey.

Recall that under stratified random sampling the variances of $\hat{\mu}$ and $\hat{\tau}$ are

$$V(\hat{\mu}) = \frac{1}{N^2}\sum_{j=1}^{L} N_j^2 \left(1 - \frac{n_j}{N_j}\right)\frac{\sigma_j^2}{n_j} \quad \text{and} \quad V(\hat{\tau}) = \sum_{j=1}^{L} N_j^2 \left(1 - \frac{n_j}{N_j}\right)\frac{\sigma_j^2}{n_j},$$

respectively, where σ_j^2 is the variance of the observations of the elements in the i-th stratum. The bounds on the error of estimation for $\hat{\mu}$ and $\hat{\tau}$ are then

$$B = 2\sqrt{V(\hat{\mu})} \quad \text{and} \quad B = 2\sqrt{V(\hat{\tau})},$$

respectively.

Assume that the *cost* of the survey can be computed using

$$C = c_0 + \sum_{j=1}^{L} n_j c_j,$$

where c_0 is the *overhead cost* and c_j is the *cost-per-element* in the j-th stratum.

We will consider two different approaches to optimum allocation which depend on our objective.

1. For a fixed *bound*, how do we allocate to minimize the *cost*?
2. For a fixed *cost*, how do we allocate to minimize the *bound*?

These are *constrained optimization* problems, but the solutions are (relatively) simple as these kinds of problems go.

Step 1

First we determine how to divide n into $n_1, n_2, \ldots, n_L$. Regardless of which goal we have, it can be shown that

$$n_j = n\left(\frac{N_j\sigma_j/\sqrt{c_j}}{N_1\sigma_1/\sqrt{c_1} + N_2\sigma_2/\sqrt{c_2} + \cdots + N_L\sigma_L/\sqrt{c_L}}\right) = n\left(\frac{N_j\sigma_j/\sqrt{c_j}}{\sum_{k=1}^{L} N_k\sigma_k/\sqrt{c_k}}\right).$$

Note that in practice we need a good guess of $\sigma_1, \sigma_2, \ldots, \sigma_L$. Also note that this does not yet give us n or $n_1, n_2, \ldots, n_L$. It only tells us the *proportion* of the total sample size that should be allocated to each stratum because

$$\frac{n_j}{n} = \frac{N_j\sigma_j/\sqrt{c_j}}{N_1\sigma_1/\sqrt{c_1} + N_2\sigma_2/\sqrt{c_2} + \cdots + N_L\sigma_L/\sqrt{c_L}} = \frac{N_j\sigma_j/\sqrt{c_j}}{\sum_{k=1}^{L} N_k\sigma_k/\sqrt{c_k}}.$$

Example: Recall the sword fern survey.

Stratum	Region	N_j	n_j	$\bar{y}_j$	s_j
1	Forest	30	8	287	149.1
2	Prairie	87	5	11.3	16.8
		117	13		

If we were doing this survey again at the same location, we might use s_1 and s_2 as guesses of σ_1 and σ_2, respectively. Assume that $c_1 = 4$ and $c_2 = 1$. What would be n_1/n and n_2/n?

Observe that n_j is *proportional to* $N_j\sigma_j/\sqrt{c_j}$. What does this tell us about the relationship between n_j and N_j, n_j and σ_j, and n_j and c_j? To which strata do we allocate larger sample sizes?

Step 2

Second we compute n. How we do this depends on our goal.

1. If our goal is to *minimize cost* for a *fixed bound* on the error of estimation, then we compute

$$n = \frac{\left(\sum_{j=1}^{L} N_j\sigma_j/\sqrt{c_j}\right)\left(\sum_{j=1}^{L} N_j\sigma_j\sqrt{c_j}\right)}{N^2V + \sum_{j=1}^{L} N_j\sigma_j^2},$$

   where $V = B^2/4$ if we are estimating μ, and $V = B^2/(4N^2)$ if we are estimating τ.

   Example: Suppose we are estimating μ and we want a bound on the error of estimation of $B = 20$ g/m^2. What is the n that will give us the least expensive survey with that bound on the error of estimation? Similarly what would we use for n if we wanted to estimate τ with a bound on the error of estimation of $B = 2000$ g/m^2?

2. If our goal is to *minimize the bound* of estimation for a *fixed cost*, then we compute

$$n = \frac{(C - c_0)\sum_{j=1}^{L} N_j\sigma_j/\sqrt{c_j}}{\sum_{j=1}^{L} N_j\sigma_j\sqrt{c_j}}.$$

   Comment: A related goal is to *minimize the bound* for a *fixed total sample size* n. This can be viewed as a special case where we set $C = n$, $c_0 = 0$, and all $c_j = 1$. In that case n will necessarily equal C which equals n. So we do not need to do the above calculation and we can just proceed to the third step!

Example: Suppose we want to minimize the bound on the error of estimation subject to a total cost of $C =$ 100 and an overhead cost of $c_0 = 20$. What is n?

Step 3

Finally we combine our results from the first two steps to compute for each stratum

$$n_j = n\left(\frac{N_j\sigma_j/\sqrt{c_j}}{\sum_{k=1}^{L} N_k\sigma_k/\sqrt{c_k}}\right).$$

Example: Given the results from the earlier examples, if we are estimating μ what are n_1 and n_2 if we want to minimize cost for a bound on the error of estimation of $B = 20\ g/m^2$. What if we want to minimize the bound for a fixed cost with $C = 100$ and $c_0 = 20$ when estimating μ?

Summary of Optimum Allocation

1. Compute the allocation fraction

$$\frac{N_j\sigma_j/\sqrt{c_j}}{\sum_{k=1}^{L} N_k\sigma_k/\sqrt{c_k}}$$

   for *each* stratum.

2. Decide if you want to minimize cost for a fixed bound, or minimize the bound for a fixed cost, and then use the appropriate formula to compute n.

3. Compute $n_1, n_2, \ldots, n_L$ using the allocation fractions and n you computed in the previous two steps as

$$n_j = n\left(\frac{N_j\sigma_j/\sqrt{c_j}}{\sum_{k=1}^{L} N_k\sigma_k/\sqrt{c_k}}\right).$$

Special Cases

1. **Neyman allocation** is a special case where the *cost-per-element is the same for all strata* (i.e., all c_j are *equal*). In this case we have that in the first step

$$\frac{n_j}{n} = \frac{N_j \sigma_j}{\sum_{k=1}^{L} N_k \sigma_k},$$

and if we are want to minimize the cost for a fixed bound then the calculation of the n simplifies to

$$n = \frac{\left(\sum_{j=1}^{L} N_j \sigma_j\right)^2}{N^2 V + \sum_{j=1}^{L} N_j \sigma_j^2}.$$

Example: Assume that the cost-per-square is the same regardless of whether a square is forest or prairie. What are n, n_1, and n_2 if we want to estimate μ with a bound on the error of estimation of B = 20 g/m^2?

2. **Proportional allocation** is a special case where the fraction of sampled elements in each stratum equals the fraction of population elements in that stratum. That is

$$\frac{n_j}{n} = \frac{N_j}{N},$$

which implies that $n_j = nN_j/N$. Proportional allocation is an optimum allocation if the cost-per-element is the same for all elements *and* all σ_j^2 are equal. In practice we might have approximate proportional allocation where $n_j/n \approx N_j/n$.

Example: What would n_1/n and n_2/n be for the sword fern survey using proportional allocation?

Restrictions on Optimum Allocation

There are some practical restrictions on an optimum allocation.

1. Optimum n and n_j must be non-negative integers, so usually the optimum allocation is approximate.
2. An optimum allocation may produce $n_j = 0$ or $n_j = 1$. But to estimate σ_j^2 we need all $n_j \geq 2$.
3. It is possible to have an optimum allocation of $n_j > N_j$, which is an impossible design.

For the latter two cases, we can find an optimum allocation subject to the constraint that all $2 \leq n_j \leq N_j$ if we find that some $n_j < 2$ or $n_j > N_j$ using the method above, but how this would be done is beyond the scope of this lecture (although see below if you are curious).

The formulas given above are an *analytical* solution to the optimum allocation problem. These are *derived* using the necessary mathematics (namely calculus and what are called Lagrange multipliers). But the optimum allocation problem can also be solved *numerically* by using computing power instead. I have created a short demonstration of how to do this in R. You do not need to know how to do this for this course, but I have included it for any students that might be interested.