STA304

Neil Montgomery
2016-02-29

Solving for n

The sample size required is (approximately):

$$n = \frac{\sum_{i=1}^{L} N_i^2 \sigma_i^2 / a_i}{N^2 B^2/4 + \sum_{i=1}^{L} N_i \sigma_i^2}$$

with σ_i^2 guessed from best available knowledge, like before.

But this is a really poor formula for hand calculation, which you'll have to practice. I had on the board divided through by N (for understanding). Even better is to divide through by N^2:

$$n = \frac{\sum_{i=1}^{L} W_i^2 \sigma_i^2 / a_i}{B^2/4 + \frac{1}{N} \sum_{i=1}^{L} W_i \sigma_i^2}$$

example—transformer age from term test

Suppose we want to estimate the average transformer age with error bound of 1 year. We're pretty sure the oldest transformers are around 60 years old. If we stick with a proportional allocation, what sample size is required?

Recall:

Size	N	W
50	9882	0.3797994
75	9405	0.3614666
100	6732	0.2587340

In the special case of proportional allocation, $W_i = a_i$ (noice!) and the formula is just:

$$n = \frac{\sum_{i=1}^{L} W_i \sigma_i^2}{B^2/4 + \frac{1}{N} \sum_{i=1}^{L} W_i \sigma_i^2}$$

example—transformer age

The desired bound is $B = 1$. We think the oldest is 60 years old, so within each subgroup we can use the guess $\sigma_i \approx \mathrm{range}/4 = 15$. The population total is $N = 26019$.

To get the required sample size I'll augment the table from before:

Size	N	W	sigma^2	W_i*sigma^2
50	9882	0.3797994	225	85.45486
75	9405	0.3614666	225	81.32999
100	6732	0.2587340	225	58.21515

And we get:

$$n = \frac{225}{0.25 + \frac{1}{26019} \cdot 225} = 869.91$$

Why did the formula get *even simpler*?

example—"fittings" age

Suppose we want to estimate the average fitting age to within 0.5 years.

We choose to sample *equally* from each stratum. In other words $a_i = 1/6$ for each stratum.

We need guesses for stratum variances. We'll use the sample variances from last week's stratified sample. (Perhaps not realistic.) Here is an augmented summary of what we need to do the calculation:

Municipality	N	W	a_i	sigma_i^2	W_i^2*sigma_i^2/a_i	W_i * sigma_i^2
BRAMPTON	18374	0.139	0.167	11.203	1.299	1.557
ETOBICOKE	13504	0.102	0.167	6.845	0.429	0.699
MISSISSAUGA	32584	0.247	0.167	11.047	4.028	2.723
NORTH YORK	19086	0.144	0.167	11.489	1.437	1.659
OTTAWA	11564	0.087	0.167	8.528	0.392	0.746
SCARBOROUGH	37071	0.280	0.167	12.217	5.765	3.426

example—"fittings" age

The sums of the last two columns from the previous slide are:

sum(W_i^2*sigma_i^2/a_i)	sum(W_i * sigma_i^2)
13.34932	10.81099

so the required sample size is:

$$n = \frac{13.3493241}{0.0625 + \frac{1}{132183}10.8109925} = 213.31$$

Is this surprising?

optimal allocation

The word "optimal" gets thrown around a bit too casually. To *optimize* anythng you need a criterion. And often optimization also requires *prediction*. But I digress…

The sample size calculation requires an allocation fraction to be determined. We've seen *proportional* and *equal* allocations.

If there is variation in *cost per unit sampled* c_i among strata, then it is possible to allocate the sample in a way that minimizes total cost.

It makes sense that the optimal allocation should be *larger* for strata that are:

1. Larger (bigger N_i)

2. More variable (bigger σ_i^2)

3. Cheaper (smaller c_i)

optimal allocation formulae

Textbook formula (bad for hand calculation):

$$a_i = \frac{N_i \sigma_i \Big/ \sqrt{c_i}}{\sum_{k=1}^{L} N_k \sigma_k \Big/ \sqrt{c_k}}$$

and note that textbook doesn't call this a_i directly…

Better for hand calculation is to divide by N to get:

$$a_i = \frac{W_i \sigma_i \Big/ \sqrt{c_i}}{\sum_{k=1}^{L} W_k \sigma_k \Big/ \sqrt{c_k}}$$

optimal allocation example

Let's use the fittings example, with prior variances used again. Suppose the cost per unit is as follows (along with some other calculations we'll need):

Municipality	N	W	sigma_i	c_i	W_i*sigma_i/sqrt(c_i)
BRAMPTON	18374	0.139	3.347	15.75	0.030
ETOBICOKE	13504	0.102	2.616	9.45	0.028
MISSISSAUGA	32584	0.247	3.324	15.75	0.052
NORTH YORK	19086	0.144	3.390	9.75	0.050
OTTAWA	11564	0.087	2.920	21.39	0.012
SCARBOROUGH	37071	0.280	3.495	9.75	0.101

optimal allocation example

The sum of the final column is 0.273. The optimal allocation now replaces the `a_i` column from the table on slide 5 which now becomes:

Municipality	N	W	a_i	sigma_i^2	W_i^2*sigma_i^2/a_i	W_i * sigma_i^2
BRAMPTON	18374	0.139	0.108	11.203	1.997	1.557
ETOBICOKE	13504	0.102	0.104	6.845	0.688	0.699
MISSISSAUGA	32584	0.247	0.191	11.047	3.517	2.723
NORTH YORK	19086	0.144	0.184	11.489	1.300	1.659
OTTAWA	11564	0.087	0.044	8.528	1.489	0.746
SCARBOROUGH	37071	0.280	0.369	12.217	2.605	3.426

The required sample size is now:

$$n = \frac{11.5964692}{0.0625 + \frac{1}{132183}10.8109925} = 185.3$$

efficiency of SRS versus various allocations

We've seen $\overline{y}_{SRS}$ and $\overline{y}_{st}$. Let's specify two versions of the latter: $\overline{y}_{prop}$ for the stratified population mean estimator with proportional allocation and $\overline{y}_{opt}$ for the stratified population mean estimator with optimal allocation.

Then when the overall sample size is the same the following is then usually true (as long as the N_i are all relatively large):

$$V(\overline{y}_{opt}) \leq V(\overline{y}_{prop}) \leq V(\overline{y}_{SRS})$$

Ponder when they might be close and when they might be far…