STA304

Neil Montgomery

2016-02-22

recap of example from 2016-02-22

The SRS estimate is $\overline{y} = 27.7377764$ with error bound $B_{\text{SRS}} = 1.3819573$.

The stratified estimates are:

$$\overline{y}_{st} = 27.9942797$$

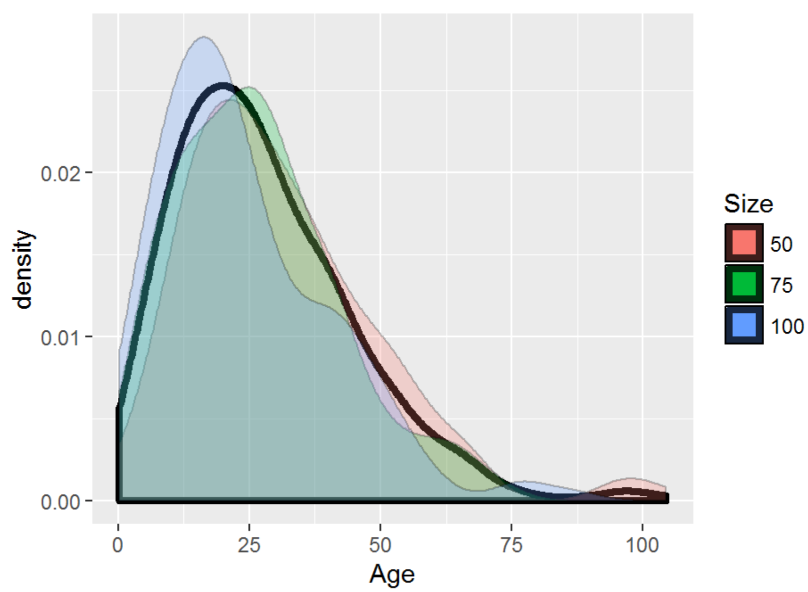
$$\hat{V}(\overline{y}_{st}) = 0.4877308$$

with bound $B_{\text{st}} = 2\sqrt{\hat{V}(\overline{y}_{st})} = 1.3967546$.

comparing the bounds I - homogeneity

The stratified bound was than the SRS bound. What happened? There were two things that we will examine in a little more detail.

First, let's look at density plots of the Age variable from the elements in the simple random sample. The thick line is for the whole sample and the coloured filled densities are for the sub-samples.



The strata all look very similar to each other and to the population. The strata

comparing the bounds II - random variation

Second, keep in mind that the bounds are based on of the population variance σ^2 and the stratum variances σ_i^2 based on randomly selected items. The estimates will of course be higher or lower than the true values just by random chance.

Let's look at a table of both the estimated and (in practice un-knowable) true variances:

Size	N	Mean	Variance	SD	n	Sample Mean	Sample Var	Sample SD
All (population)	26019	27.3	308.7	17.6				
50	9882	31.2	327.0	18.1	228	32.9	364.4	19.1
75	9405	26.7	290.2	17.0	217	27.5	320.7	17.9
100	6732	22.6	263.3	16.2	155	21.6	174.9	13.2

another example "fittings"

The data have been adapted from a study I did with a gas distribution company. They were concerned with properties of a certain old type of , whose age might be associated with failure, leak, and safety risk.

The company's "territory" covers at least the GTA, Ottawa, and other areas. Some areas might have an older population of this fitting than others.

The company wishes to estimate the overall average age of the fittings, and also the average ages within each area.

To determine the age of a fitting they may need to check a paper record, as the ages are not all in the database.

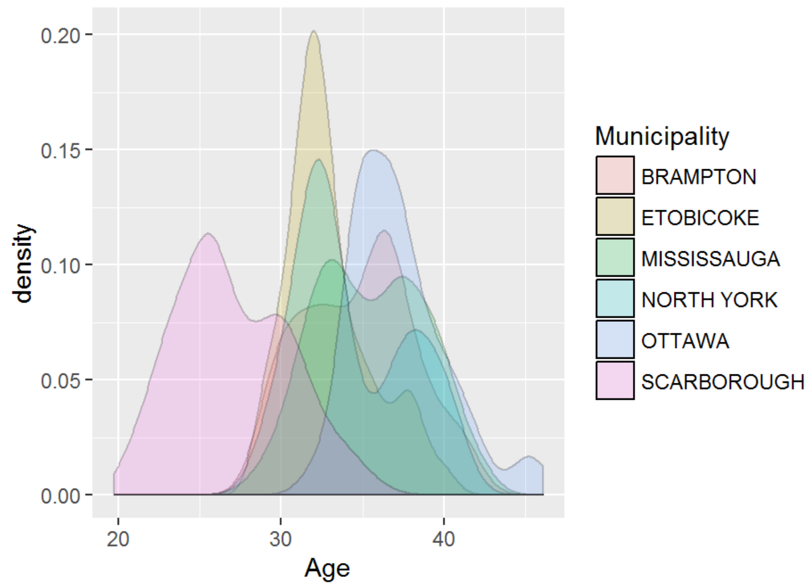
fittings population summary

Municipality	N	W
BRAMPTON	18374	0.1390043
ETOBICOKE	13504	0.1021614
MISSISSAUGA	32584	0.2465067
NORTH YORK	19086	0.1443907
OTTAWA	11564	0.0874848
SCARBOROUGH	37071	0.2804521

The population size is 132183.

We'll use an overall sample of size 1000, allocated proportionally to the strata.

summaries of the stratified sample



Municipality	n	mean	sd
BRAMPTON	139	34.71942	3.347135
ETOBICOKE	102	32.92157	2.616347
MISSISSAUGA	247	35.50550	3.323675
NORTH YORK	144	34.31250	3.389554
OTTAWA	87	37.21522	2.920263
SCARBOROUGH	280	27.03434	3.495274

stratified estimates of average age

Note: one goal has been acheived, which was to get estimates for each municipality. Use tables on previous two slides and the usual SRS theory to obtain CI and/or error bounds.

Estimate of population average age is:

$$\bar{y}_{st} = 32.7338166$$

with error bound:

$$B_{st} = 0.2072594$$

compared with SRS of same size

The overall sample size was (due to rounding) 999. A simple random sample of the same size gives:

$$\bar{y}_{SRS} = 32.7537864$$

and

$$B_{SRS} = 0.3075457$$

The stratified estimate had a lower bound on the error of estimation. There is the notion of "relative efficiency" of estimators, which is simply the ratio of their variances (see 6.8 of the text). In this case we might estimate the relative efficiency with

$$\frac{B_{SRS}^2}{B_{st}^2} = 2.2018669$$

.

sample size and allocation (for population mean and total)

In the two examples I arbitrarily chose a sample size, and allocated the sample size proportionally to each stratum according to its sub-population size.

We need to consider how large the overall sample size be and also how it should actually be allocated to the strata.

The sample size is determined based on a desired bound B. Just like in the case of SRS it comes down to solving this equation for n:

$$2\sqrt{\hat{V}(\overline{y}_{st})} = B$$

$$\hat{V}(\overline{y}_{st}) = \frac{B^2}{2}$$

the allocation fractions, and solving for n

The sample size n and the allocation are two peas in a pod. (Love and marriage, horse and carriage, etc.)

The allocation $n = n_1 + \cdots + n_L$ is described by the "allocation fractions" defined as in:

$$a_i = \frac{n_i}{n} \qquad n_i = n a_i \qquad 0 < a_i < 1 \qquad a_1 + \cdots + a_L = 1$$

Given these, the sample size required is (approximately):

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

Of course σ_i^2 have to be guessed from best available knowledge, like before.

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

example—fitting age

Suppose we decide to sample equally from each municipality, and we want to estimate the average fitting age to within 0.5 years. What sample size is required?