STA304

Neil Montgomery

2016-02-22

stratified random sampling

recap

- Stratified sampling can be useful for:
  - getting better population parameter estimates, if the strata are (more) homogeneous
  - further investigation of strata
- SRS done within each strata to get the usual estimates with the usual properties
- Population paramaters are estimated using suitably weighted combinations of stratum estimates.

(from last time)

Going backwards: $\hat{\tau} = N\overline{y}_{st}$.

True and estimated variance (because of independence):

$$V(\overline{y}_{st}) = \sum_{i=1}^{L} W_i^2 V(\overline{y}_i)$$

$$\hat{V}(\overline{y}_{st}) = \sum_{i=1}^{L} W_i^2 \hat{V}(\overline{y}_i)$$

Confidence interval $\overline{y}_{st} \pm 2\sqrt{\hat{V}(y_{st})}$, bound on error of estimation, etc.

example(s)

We will look at the transformer "data" from the term test, by which I mean the actually simulated population's worth of data that was used for the test, contained in the file `tx.csv`. The story here has to be a little different. On the test knew only the locations of the transformer. For this example we'll pretend we know, say, the location and the `Size` (50KVA, 75KVA, or 100KVA) but not the Manufacturer or the Age.

plan: compare SRS versus stratified by `Size`.

We will try to estimate the average age of the population of transformers.

We'll do this in two ways. One is using a SRS of size $n = 600$. For the other we'll stratify by transformer "Size", which is one of 50KVA, 75KVA, or 100KVA. A summary of the population by Size is as follows, including weights to be used in the stratified formulae later on. Note that here "Size" is a property of a transormer and not anything to do with how many of them there are.

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

the simple random sample

We select a simple random sample of 600 transformers and get the following sample mean, standard deviation, and "bound on the error of estimation"

mean	sd	B
27.73778	17.12404	1.381957

stratify by `Size`

To make a fair comparison we'll keep an overall sample size of 600. We will choose to allocate the sample proportionally to the size of the strata (more on this later). Here is a summary of the results by stratum:

Size	n	means	variances	sds
50	228	32.85426	364.4449	19.09044
75	217	27.48206	320.7445	17.90934
100	155	21.57585	174.9298	13.22610

stratified estimates

The stratified estimator for the population mean is:

$$\bar{y}_{st} = \sum_{i=1}^{L} W_i \bar{y}_i$$

with estimated variance:

$$\hat{V}(\bar{y}_{st}) = \sum_{i=1}^{L} W_i^2 \hat{V}(\bar{y}_i)$$

where $\hat{V}(\bar{y}_i) = \frac{s^2}{n_i} \frac{N_i - n_i}{N_i}$.

stratified estimates

Plug in numbers and weights from previous slides ago to get:

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

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

The "usual bound on the error of estimation" is $\sqrt{\hat{V}(\bar{y}_{st})} = 1.3967546$.