STA304

Neil Montgomery

2016-02-04

an ongoing discussion—sources of errors in surveys

a comedy of errors

In general we're seeking to estimate population parameters using samples. Suppose the parameter is θ, estimated with $\hat{\theta}$.

The *error* is simply $\left|\hat{\theta} - \theta\right|$.

Errors can be caused because the actual sample doesn't reflect the target population. In other words, we do not get information from the units we should have. These are *errors of nonobservation*.

Errors can be also be caused because the observations we do get are wrong. These are *errors of observation*

errors of nonobservation

sampling error

Sampling error is due to the randomness inherent in sampling and can be described using the laws of probability, and takes up the bulk of our time in this course.

Sampling error is a form of *nonobservation error* because the actual sample is not representitive of the target population. But sampling error is just bad luck.

Sampling error can be reduced with larger samples (always) or better sampling design (sometimes), both of which will reduce the chance of bad luck.

We are now looking into one way of luck augmentation: stratified random sampling.

coverage errors

Coverage errors occur when there are discrepancies between the target population, the sampling popuation, and the frame.

Examples?

nonresponse

Nonresponse occurs when no information can be obtained about a unit in a sample.

The book uses human-centred language to describe this, but it can happen with non-human sampling units!

The information might be unavailable because the sampling unit:

1. doesn't respond/can't be contacted/can't be located

2. is unable to answer

3. is unwilling to answer ("refusal")

Examples? Possible remedies (a few appear later in the chapter)?

The challenge with nonresponse is in determining its impact on the results, as this is inevitably a subject-matter problem.

stratified random sampling

when is it a good idea?

- Stratified sampling can be more effective (in terms of cost and/or estimate variation) than SRS when one or more of the following are true:
    - auxiliary information suggests that values within strata have less variation, i.e. strata are *homogeneous*
    - convenience of accessing stratum elements results in lower cost
- In addition, stratification allows further estimation within strata.
- It can be a bad idea when the auxiliary information is misused.
- Questions: population parameters versus stratum equivlents? estimators and their properties? overall sample size? allocation of sample size? number and composition of strata?

practical example—gas distribution company

The gas distribution company operates over a large part of south, central, and eastern Ontario, and in a small part of western Quebec.

The company grew over decades as the result of mergers and acquisitions.

Some of the work in some regions is done mostly by contractors.

Its territory ranges from urban to suburban to semi-rural.

All these practical factors could be reasons to do stratification when sampling *depending on what exactly is being studied.*

notation

Population: $\{y_1, y_2, \ldots, y_N\}$

N_i: units in stratum i (the stratum itself is: $\{y_{i1}, y_{i2}, \ldots, y_{iN_i}\}$)

L: number of strata

$N = \sum_{i=1}^{L} N_i$: population size

n_i: sample size in stratum i (the sample itself is $\{y_{i1}, y_{i2}, \ldots, y_{in_i}\}$)

$n = \sum_{i=1}^{L} n_i$ overall sample size

$W_i = \frac{N_i}{N}$: "weight" of stratum i (not in book but possibly should be.)

Note that $\sum_{i=1}^{L} W_i = 1$.

parameters - population and stratum

We still have the population mean, total, and variance: μ, τ, σ^2 .Strata have their analogues, imagining a stratum as a (sub-)population of its own:

$$\begin{aligned}
\tau_i &= \sum_{j=1}^{N_i} y_{ij} &&= N_i\mu_i \\
\mu_i &= \frac{\tau_i}{N_i} \\
\sigma_i^2 &= \frac{1}{N_i} \sum_{j=1}^{N_i} \left(y_{ij} - \mu_i\right)^2
\end{aligned}$$

Obviously $\tau = \sum_{i=1}^{L} \tau_i$

parameters - population and stratum means

$$\mu = \frac{\tau}{N} = \frac{1}{N}\sum_{i=1}^{L}\tau_i = \frac{1}{N}\sum_{i=1}^{L}N_i\mu_i = \sum_{i=1}^{L}W_i\mu_i$$

The population variance is (not obviously or very easily) related to the stratum variances.

You may or may not have learned of a technique "anaylsis of variance" (ANOVA), which is based on dividing "overall" variation into two parts: "within-group" variation plus "between-group" variation.

In stratified sampling it is similar:

$$\sigma^2 = \sum_{i=1}^{L}W_i\sigma_i^2 + \sum_{i=1}^{L}W_i(\mu_i - \mu)^2$$
$$\text{Total} = \text{Within Group} + \text{Between Group}$$

a few illustrations

$$\sigma^2 = \sum_{i=1}^{L} W_i \sigma_i^2 + \sum_{i=1}^{L} W_i (\mu_i - \mu)^2$$

If all the strata have the *same* mean $\mu_i = \mu_s$ and variance $\sigma_i^2 = \sigma_s^2$, then the population variance is

$$\sigma^2 = \sum_{i=1}^{L} W_i \sigma_s^2 = \sigma_s^2$$

If all the strata are *perfectly homogeneous* then $\sigma_i^2 = 0$ and:

$$\sigma^2 = \sum_{i=1}^{L} W_i (\mu_i - \mu)^2$$

properties of a stratified sample

The sample $\{y_{i1}, y_{i2}, \ldots, y_{in}\}$ *from within stratum* i is a simple random sample and shares all familiar properties from before. They have the same distribution with:

$$E\left(y_{ij}\right) = \mu_i \qquad V(y_{ij}) = \sigma_i^2$$

but are *not* independent (they are negatively correlated, as before)

The estimators of stratum parameters and so on are just as before:

$$\hat{\mu}_i = \overline{y_i} \qquad V(\overline{y_i}) = \frac{\sigma_i^2}{n_i}\left(\frac{N_i - n_i}{N_i - 1}\right) \qquad \hat{V}(\overline{y_i}) = \frac{s_i^2}{n_i}\left(\frac{N_i - n_i}{N_i}\right)$$

(multiply as appropriate by N_i and N_i^2 for stratum total τ_i)

But anything related to samples from *different strata* are ***independent***.

estimation of population parameters

The best way to see how "obvious" it all is, start with population total τ. I claim it is obvious that:

$$\hat{\tau} = \sum_{i}^{L} \hat{\tau}_i$$

Next move to the population mean, by dividing through by N:

$$\hat{\mu} = \frac{\hat{\tau}}{N} = \frac{\sum_i^L \hat{\tau}_i}{N}$$

Then recall that $\hat{\tau}_i = N_i \overline{y_i}$ to get:

$$\hat{\mu} = \frac{\hat{\tau}}{N} = \frac{\sum_i^L \hat{\tau}_i = N_i \bar{y}_i}{N} = \sum_{i=1}^{L} W_i \bar{y}_i = \bar{y}_{st} \quad \text{(definition)}$$

properties of the estimators

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 $\overline{y}_{st} \pm 2\sqrt{\hat{V}(y_{st})}$.