STA304

Neil Montgomery
2016-03-21

ratio and regression estimators

rats on drugs - comparing the estimators

	Estimate	$2\sqrt{\hat{V}}$
SRS	16.85	1.15
Ratio	17.59	0.27
Regression	17.51	0.25

Ratio and regression bounds very close - both far better than the SRS bound.

some theoretical notes and conclusions

Ratio is best when there is a linear relationship through the origin (especially when the variance of y is proportional to x).

Regression is best when there is a linear relationship not through the origin.

Ratio estimator is *not* a special case of regression with intercept forced to be 0, since $r = \sum y_i \Big/ \sum x_i$ while $\hat{\beta}_1 = \sum x_i y_i \Big/ \sum x_i^2$ in a "regression through the origin" model.

The so-called "difference estimator" is just the special case of a regression estimator with b fixed in advance, typically to $b = 1$. We will not consider this special case.

systematic random sampling

a new sampling design

SRS and stratified sampling were examples of *sampling designs*.

Ratio and regression estimations were examples of calculation techniques in the special case where the population consisted of paired measurements (y_i, x_i). The calculation techniques would apply to stratified sampling as well (at the stratum level).

Systematic sampling is a new sampling design which can be useful in two seemingly "opposite" cases:

1. Where there are known trends of certain types in the population (in which case systematic sampling can give improved estimators or population parameters.)

2. Where there are no trends in a population and it is otherwise difficult to sample due to difficulties in making a frame.

systematic sampling

A *1 in k systematic sample with random start* works as follows:

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

Choose first element of sample y_1 out of first k elements, and then choose each k^{th} element after that.

Can be implemented from a frame, or can be implemented in cases where the population presents itself sequentially (manufacturing, quality control, customer interviews).

Each element has an equal chance of being selected. But this is not a simple random sample. Why? (This was a question on Test 1.)

minor technical issue

The sample size ends up being $n = N/k$. Alternatively, a sample size calculation to obtain n implies $k = N/n$.

These quotients are not likely to be integers. So a systematic sample might result in a final sample size of $n - 1$, n, or $n + 1$, depending on the nature of the rounding error.

But when N and/or n is not small, the "issue" is trivial, and negligible. And sample size calculations themselves are so dependent on the bound B (arbitrarily chosen) and the population variance (either guessed or estimated with not very much precision), so it's nothing worth worrying about it practice.

analysis of a systematic sample

The key question: is the order of the population related to the quantity of interest? Examples:

Population: shoppers at a store. Order: the order in which they leave. Possible quantities of interest: satisfaction with the store (yes/no); how much purchased ($).

Population: farms. Order: size. Possible quantities of interest: how much livestock on the farm? revenues? return on investment?

Population: items produced in a factory. Possible quantities of interest would include any matter related to *quality*.

properties of a systematic random sample

To estimate the population mean μ we use the sample mean of the elements in the systematic sample chosen:

$$\bar{y}_{sy} = \frac{\sum_{i=1}^{n} y_i}{n}$$

This is unbiased for μ (no matter what.)

The variance of $\bar{y}_{sy}$ is the crux of the matter. It can be expressed as:

$$V(\bar{y}_{sy}) = \frac{\sigma^2}{n}(1 + (n-1)\rho)$$

where ρ is the so-called "intra-cluster correlation coefficient".

intra-cluster correlation coefficient

Is bounded between $-1/(n-1)$ and 1 and measures the similarity of items within each possible systematic sample.

Interesting cases:

$\rho \approx 1$ (the worst case for systematic sampling)

$\rho \approx 0$ (in which case systematic sampling can be treated with SRS theory)

$\rho < 0$ (in which case systematic sampling gives improved population parameter estimators, but can be tricky to measure this improvement)

randomly ordered population

When the population is randomly ordered we can treat the systematic sample with SRS theory.

$$\bar{y}_{sy} = \frac{\sum_{i=1}^{n} y_i}{n}$$

$$\hat{V}(\bar{y}_{sy}) = \left(1 - \frac{n}{N}\right)\frac{s^2}{n}$$

Why is this a good estimator for the variance of $\bar{y}_{sy}$ when the population is randomly ordered?