STA304

Neil Montgomery
2016-04-04

systematic random sampling

today's assumption

$$N = nk$$

the theoretical variance revisited

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 usefully 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". Back to this in a moment

basic $V(y_{sy})$ formula

It's easy to write down the correct (but not as useful) formula for the variance of $\overline{y}_{sy}$ from first principles.

In fact you were asked to do so on the first test!.

y_{sy} can take on one of k possibilities, specifically: $\overline{y}_i = \sum_{j=1}^{n} y_{ij}/n$ where $\{y_{i1}, \ldots, y_{in}\}$ is the i^{th} possible systematic sample.

Each possibility has the same probability $1/k$.

The average of all the possible $\overline{y}_{sy}$ is simply μ. (Why?) So:

$$V(\overline{y}_{sy}) = \frac{1}{k}\sum_{i=1}^{k}(\overline{y}_i - \mu)^2$$

the road to ρ

The actual intra-cluster correlation really does look like a correlation…

$$\rho = \frac{\sum_{j \neq u}(y_{ij} - \mu)(y_{iu} - \mu) \Big/ (kn(n-1)/2)}{\sigma^2}$$

…but is pretty nasty to calculate. In principle it means if the elements within each possible stratified sample are "far apart", ρ will be negative.

The book gives a perplexing (because they don't explain it) but manageable version that depends on an ANOVA-style sum of squares decomposition:

$$\rho = \frac{(k-1)nMSB - SST}{(n-1)SST}$$

It's actually worth explaining this version.

the ANOVA of systematic sampling

(Assume $N = nk$ perfectly.) Consider any bunch of numbers $\{y_1, y_2, \ldots, y_N\}$. Call their mean $\bar{y}$. (Book uses double-bar notation that I cannot nicely replicate in slides: $\overline{\bar{y}}$. And we already call this μ anyway!)

They can be divided into k groups of n elements, called $\{y_{i1}, \ldots, y_{in}\}$. Each group has mean $\bar{y}_i$

The following (sum of squares decomposition) is always true:

$$\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y})^2 = n\sum_{i=1}^{k}(\bar{y}_i - \bar{y})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_i)^2$$

Often called something like "*total SS* equals *between SS* plus with *within SS*" or $SST = SSB + SSW$. Often SSW is called the "error sum of squares" or SSE.

ANOVA continued

Alert! We've seen SSB already. In fact $V(\bar{y}_{sy}) = SSB/k$.

Alert! We've seen SST already. In fact $\sigma^2 = SST/N = SST/(nk)$

These three components usually have "degrees of freedom" associated with them. In this case it would be $nk-1$, $k-1$, and $n(k-1)$ respectively. The last two are used to construct the so-called "mean squares". From this we get, for example:

$$MSB = \frac{SSB}{k-1} = \frac{n\sum_{i=1}^{k}(\bar{y}_i - \bar{y})^2}{k-1}$$

(Remainder of work done on board…)

what's the use of this?

In practice, limited, since the calculations require the population values to be known. But it helps with the understanding of when systematic sampling can produce better estimates than SRS. Consider, for example:

1. randomly ordered populations
2. populations with increasing trend
3. periodic trend - good and bad

repeated systematic sampling

With one systematic sample, there is no data-driven way to know if you'll end up with a better-than-SRS estimate (knowledge of population required).

Another approach can be to select multiple systematic samples, n_s in number, resulting in n_s estimates of μ we can call $\bar{y}_i$.

We average these estimates to obtain:

$$\hat{\mu} = \frac{1}{n_s} \sum_{i=1}^{n_s} \bar{y}_i$$

which will have estimated variance:

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

where $s^2_{\bar{y}}$ is just the sample variances of the $\bar{y}_i$.