10/2/25

Stat 440 Lecture: Domain Attributes, Problem Solutions, & Ratio Estimators

(A). Recap definitions of Domains, associated Domain attributes and their population and sample means & variances

(B). Discussion of Problem Solutions for HW2:
Ch. 4 #1 (used on quiz) and last problem (3).

(C). Math explaining why domain ratio estimators can be (much) better than simpler estimators using known denominators.

Domains & Attribute Means and Variances

Domains are subpopulations whose size or characteristics are generally not known in advance of a survey

Consider SRS sample S in population U with domain D

with sizes $|S| = n,\ |U| = N,\ |D| = N_D$

and attributes of interest $y_i,\ i = 1, \ldots, N$

To study means and variances of y_i within D

define **domain attributes** $z_i = y_i \, I_{[i \in D]}$

we relate the mean and variance of z_i over S, U to those of y_i

Domains & Means and Variances, cont'd

$$\bar{z}_U = \frac{1}{N} \sum_{i=1}^{N} y_i \, I_{[i \in D]} = \frac{N_D}{N} \cdot \frac{1}{N_D} \sum_{i \in D} y_i = \frac{N_D}{N} \, \bar{y}_D$$

Using the notation $n_D = \sum_{i \in S} I_{[i \in D]} = |D \cap S|$, similarly

$$\bar{z}_S = \frac{1}{n} \sum_{i \in S} y_i \, I_{[i \in D]} = \frac{1}{n} \sum_{[i \in D \cap S]} y_i = \frac{n_D}{n} \, \bar{y}_{D \cap S}$$

$$s_{z,U}^2 = \frac{1}{N-1} \sum_{i=1}^{N} (z_i - \bar{z}_U)^2 = \frac{1}{N-1} \Big[\sum_{i=1}^{N} y_i^2 \, I_{[i \in D]} - N \, (\bar{z}_U)^2 \Big]$$

$$= \frac{1}{N-1} \Big[\sum_{i \in D} (y_i - \bar{y}_D)^2 + N_D \, \bar{y}_D^2 - \frac{N_D^2}{N} \, \bar{y}_D^2 \Big]$$

Therefore $\quad s_{z,U}^2 = \frac{N_D - 1}{N-1} \, s_{y,D}^2 + \frac{N_D \, (N - N_D)}{N \, (N-1)} \, \bar{y}_D^2$

and similarly $\quad s_{z,S}^2 = \frac{n_D - 1}{n-1} \, s_{y,\, D \cap S}^2 + \frac{n_D \, (n - n_D)}{n \, (n-1)} \, \bar{y}_{D \cap S}^2$

Domain Means & Variances, Binary y_i

There is an **important special case** of the domain mean and variance formulas, when the **attribute** $y_i = I_{[i \in A]}$ is binary

Use similar notation as before, with $z_i = y_i \, I_{[i \in D]} = I_{[i \in D \cap A]}$:

$$N \bar{y}_U = N_A = |A|, \qquad n \, \bar{y}_S = \sum_{i \in S} I_{[i \in A]} = |A \cap S| = n_A$$

Then $\bar{z}_U = N_{A \cap D}/N$, $\bar{z}_S = n_{A \cap D}/n$ and variances are based on $s^2_{y,U} = \frac{N_A \, (N - N_A)}{N \, (N-1)}$, $s^2_{y,S} = \frac{n_A \, (n - n_A)}{n \, (n-1)}$

In this case, formulas derived for variances of z's become:

$$s^2_{z,U} = \frac{N_D - 1}{N-1} \, s^2_{y,D} + \frac{N_D \, (N - N_D)}{N \, (N-1)} \, \bar{y}^2_D = \frac{N_{A \cap D} \, (N - N_{A \cap D})}{N \, (N-1)}$$

and similarly with $s^2_{z,S}$ *and little* n*'s*

HW Problem Solutions

On this slide **discuss Ch. 4 #1 (the Quiz problem)**

Target of estimation: **total number of trees** in study area

Sampling units: **plots** (assume accurate count of trunks, i.e. trees, in sampled plots).

Attribute: $y_i =$ number of trees in plot i, for $i = 1, \ldots, N = 900$

Goal: 95% CI of specified half-width $\leq \delta = 1000$ **from SRS:**

$$N\left[\bar{y}_S \pm z_{.025}\left(\frac{1}{n} - \frac{1}{N}\right)^{1/2} s_{y,U}\right] \quad , \qquad s^2_{y,U} = \sigma^2 \approx 45$$

Method: $\frac{1}{n} - \frac{1}{900} \leq (1000/900)^2/(45 \cdot 1.96^2)$

Formula: $n \geq \left[\frac{1}{900} + (1000/900)^2/(45 \cdot 1.96^2)\right]^{-1} = 121.27$

As in HW Ch. 4 #2, note **without fpc** answer is $n \geq 141$.

HW2 problem (3)

2-part problem, done separately: sample size needed for the $D = M$ domain with analogous formula for $D = F$. The larger of the 2 required sample sizes is the overall answer.

For **D=M**: $N = 20,000$, $N_D = 10,500$, $s_{y,D} \approx 2.7$, $\bar{y}_D \approx 5$.

Survey **SRS n out of N**, target total $\sum_{i \in D} y_i = N_D \bar{y}_D = N \bar{z}_U$, for **domain attribute** $z_i = y_i I_{[i \in D]}$

Estimator is $N \bar{z}_S$ (later compare with ratio estimator $N \bar{y}_D$), with theoretical variance (from previous domain-attribute discussion)

$$N^2 \left(\frac{1}{n} - \frac{1}{N}\right) \underbrace{\left[\frac{N_D - 1}{N - 1} s_{y,D}^2 + \frac{N_D (N - N_D)}{N(N-1)} \bar{y}_D^2\right]}_{s_{z,U}^2}$$

HW2 problem (3), continued

So we want, for desired precision δ $(= 800$ in this problem)

$$z^2_{\alpha/2}\, N^2 \left(\frac{1}{n} - \frac{1}{N}\right) \left[\frac{N_D - 1}{N-1}\, s^2_{y,D} + \frac{N_D\,(N - N_D)}{N\,(N-1)}\, \bar{y}^2_D\right] \le \delta^2$$

Substituting, the inequality to solve is:

$$\frac{1}{n} \le \frac{1}{2e4} + \frac{800^2}{(1.96 \cdot 2e4)^2} \Big/ \left(\frac{10499}{19999}\, 2.7^2 + \frac{10500 \cdot 9500}{2e4 \cdot 19999}\, 25\right) \;=\; 9.139e{-}5$$

Corresponding inequality for $D = F$ domain has right-hand side

$$\frac{1}{n} \le \frac{1}{2e4} + \frac{800^2}{(1.96 \cdot 2e4)^2} \Big/ \left(\frac{9499}{19999}\, 3.0^2 + \frac{10500 \cdot 9500}{2e4 \cdot 19999}\, 25\right) \;=\; 8.963e{-}5$$

So require $n \ge 1/\min(8.963e - 5, 9.139e - 5)$ or $n \ge 11,157$.
(**Without fpc, $> 20,000$.**) Next compare ratio estimator.

Intro to Ratio Estimator

SRS n out of N, attribute y_i, domain D with size N_D known

Target of estimation $r = \bar{y}_D = \frac{\sum_{i=1}^N y_i \, I_{[i \in D]}}{\sum_{i=1}^N I_{[i \in D]}} = \frac{\sum_{i=1}^N z_i}{\sum_{i=1}^N x_i}$,

for **domain attributes** $z_i = y_i \, I_{[i \in D]}, \;\; x_i = I_{[i \in D]}$

Ratio Estimator $\hat{r} = \frac{\bar{z}_S}{\bar{w}_S} = \frac{\sum_{i \in S} y_i \, I_{[i \in D]}}{\sum_{i \in S} I_{[i \in D]}}$

Idea here is that random excess or too few random i's from D included in S balance each other in numerator and denominator.

Re-express $\hat{r} - r = \frac{1}{n_D} \sum_{i \in S} (y_i - r) \, I_{[i \in D]}$ as

$$\sqrt{n} \, (\hat{r} - r) = \frac{n}{n_D} \cdot \frac{1}{\sqrt{n}} \sum_{i \in S} (y_i - r) \, I_{[i \in D]}$$

Large-Sample Limiting Behavior of Ratio

The terms of the last expression satisfy large-sample limit theorems (Law of Large Numbers and Central Limit Theorem, respectively) that allow us to estimate the variance. First,

$$\frac{n_D}{n} - \frac{N_D}{N} = \frac{1}{n} \sum_{i \in D} (I_{[i \in S]} - \frac{n}{N})$$

is a sum of expectation 0 terms with variance $(1/n - 1/N)\, s^2_{x,U} \to 0$ as n, N get large, implying $n_D/n - N_D/N \to 0$ in probability.

Similarly, again as an average of mean 0 terms with variance $\to 0$,

$$\frac{1}{n} \sum_{i \in S} (z_i - r x_i) \longrightarrow 0 \quad \text{in probability}$$

Also, for large n, N (when n/N tends to a limit $\lambda < 1$ and $s^2_{y,D}$, $\bar{y}_D$ have limiting values),

$$\frac{1}{n}\sum_{i\in S}(z_i - rx_i) \Big/ \Big[\Big(\frac{1}{n} - \frac{1}{N}\Big)s^2_{z-rx,U}\Big]^{1/2} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$$

From this, we conclude that the ratio estimator $\hat{r}$ has a limiting normal distribution with mean $r = \bar{y}_D$ and variance

$$(N/N_D)^2\Big(\frac{1}{n} - \frac{1}{N}\Big)\, s^2_{z-rx,U}$$

The other estimator we previously used for $r = \bar{y}_D$ was $(N/N_D)\,\bar{z}_S$, and its variance is $(N/N_D)^2\left(\frac{1}{n} - \frac{1}{N}\right) s^2_{z,U}$, which is larger.

We compare these variances in some detail in our next class.