SRS: design based theory

Week 2 (2.8)

Stat 260, St. Clair

"Intro stats" vs. "design-based"

An Intro Stats story: Responses are randomly selected from a "normal population"

- the population is infinite
- the observed value Y_i is **random**: $Y_i \sim N(\mu, \sigma^2)$

"Intro stats" vs. "design-based"

Design-based story:

- Responses are **fixed** values (not random, just unknown)
- the population is finite
- the **sampling design** induces randomness, who is picked for the sample is random (not their response)

Design-based derivations

Use the design-based perspective to prove that $\bar{y}$ is an unbiased estimator of the population mean $\bar{y}_{\mathcal{U}}$ when the design is a SRS.

$\hat{\bar{y}}_{\mathcal{U}} = \bar{y}$

$E(\bar{y}) = \bar{y}_{\mathcal{U}}$ (pop. mean)

under SRS

Y_i's fixed : Construct a random variable Z_i that indicates inclusion in the sample

$$Z_i = \begin{cases} 1, & \text{unit } i \text{ is include} \\ 0, & \text{otherwise} \end{cases}$$

$\boxed{Z_i \sim \text{Bernoulli}(\pi_i)}$

SRS: $Z_i \sim \text{Bern}\left(\pi_i = \frac{n}{N}\right)$

For all units in pop.

$Z_1, Z_2, \ldots, Z_N$

aside: $\sum_{i=1}^{N} Z_i = n$

Facts: $E(Z_i) = 0\left(1-\frac{n}{N}\right) + 1\left(\frac{n}{N}\right) = \frac{n}{N} = \pi_i$

$Var(Z_i) = \left(0-\frac{n}{N}\right)^2\left(1-\frac{n}{N}\right) + \left(1-\frac{n}{N}\right)^2\left(\frac{n}{N}\right) = \frac{n}{N}\left(1-\frac{n}{N}\right)$

$= \pi_i(1-\pi_i)$

Back to estimator

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{N} Z_i y_i$$

$$E(\bar{y}) = E\left[\frac{1}{n}\sum_{i=1}^{N} Z_i y_i\right] = \frac{1}{n} E\left[\sum^{N} Z_i y_i\right] = \frac{1}{n}\sum_{i=1}^{N} \overbrace{E[Z_i]}^{\frac{n}{N}} y_i$$

SRS

$$= \frac{1}{n}\sum_{i=1}^{N}\left(\frac{n}{N}\right) y_i = \frac{1}{N}\sum_{i=1}^{N} y_i = \bar{y}_u = \text{pop. mean}$$

✓ $E(\bar{y}) = \bar{y}_u$ we know $\bar{y}$ is unbiased!

Design-based derivations

Use the design-based perspective to prove that $\hat{t}_{HT} = \sum y_i/\pi_i$ is an unbiased estimator of the population total.

Prove: under any design with inclusion probs. π_i

$$E(\hat{t}_{HT}) = t = \sum_{i=1}^{N} y_i$$

Rewrite:

$$\hat{t}_{HT} = \sum_{i=1}^{n} \frac{y_i}{\pi_i} = \sum_{i=1}^{N} Z_i \frac{y_i}{\pi_i}$$

$$\underline{E(\hat{t}_{HT})} = \sum_{i=1}^{N} \underbrace{E(Z_i)}_{\pi_i} \frac{y_i}{\pi_i} = \sum_{i=1}^{N} \pi_i \cdot \frac{y_i}{\pi_i} = \underline{\sum_{i=1}^{N} y_i = t} \checkmark$$

so $\hat{t}_{HT}$ is an unbiased estimator of t.

SRS variance

Prove that for a SRS: $\longrightarrow$ $\bar{y}_u = \bar{y} = \frac{1}{n}\sum_{i=1}^{N} Z_i y_i$

$$SE(\hat{\bar{y}}_{\mathcal{U}}) = \frac{SE(\hat{t}\,)}{N} = \sqrt{\left(1 - \frac{n}{N}\right)\frac{s^2}{n}}$$

$$\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{N} Z_i y_i\right) = *$$

If all Z_i's are independent, this is "easy" $\left(\frac{1}{n}\right)^2 \cancel{\sum \mathrm{Var}(Z_i) y_i^2}$

But when sampling ~~without~~ replacement from a finite population the Z_i's are dependent.

$$* = \left(\frac{1}{n}\right)^2 \sum_{i=1}^{N} \underbrace{\mathrm{Var}(Z_i)}_{\frac{n}{N}\left(1-\frac{n}{N}\right)} y_i^2 + 2\left(\frac{1}{n}\right)^2 \boxed{\sum_{\substack{\text{all pairs}\\(i,j)}} \mathrm{Cov}(Z_i, Z_j)} y_i y_j$$

What is Cov?

$$\mathrm{Cov}(Z_i, Z_j) = \underbrace{E(Z_i Z_j)}_{\text{only 1 when both} = 1} - \underbrace{E(Z_i)}_{\frac{n}{N}} \underbrace{E(Z_j)}_{\frac{n}{N}}$$

$$E(Z_i Z_j) = 1 \cdot \underbrace{P(Z_i = 1, Z_j = 1)}_{P(Z_i = 1) P(Z_j = 1 \mid Z_i = 1)} + 0 \cdot \underline{\qquad\qquad} = \frac{n}{N} \times \frac{n-1}{N-1}$$

$$\mathrm{Cov}(Z_i, Z_j) = \frac{n}{N} \cdot \frac{n-1}{N-1} - \left(\frac{n}{N}\right)^2 = \frac{n}{N}\left(\frac{n-1}{N-1} - \frac{n}{N}\right)$$

$$= \frac{n}{N}\left(\frac{N(n-1) - n(N-1)}{(N-1)N}\right) = \cdots = \frac{n}{N}\left(1 - \frac{n}{N}\right)\underbrace{\left(\frac{-1}{N-1}\right)}_{\text{negative!}}$$

SRS

$$V(\bar{y}) = \frac{1}{n^2} \sum^{N} y_i^2 \frac{n}{N}\left(1 - \frac{n}{N}\right) + 2 \frac{1}{n^2} \sum_{\text{all pairs}} y_i y_j \left(\frac{n}{N}\left(1 - \frac{n}{N}\right) \frac{-1}{N-1}\right)$$

$$= \text{lots of algebra} = \left(1 - \frac{n}{N}\right) \frac{s^2}{n}$$