Stratified sampling estimation

Week 3 (3.1-3.3)

Stat 260, St. Clair

Design: Stratified Sample

Definition: A population is **stratified** if it's sampling units are divided into H non-overlapping subpopulations.

- The subpopulations are called **strata** (plural)
- Notation: N_h is the population size of stratum h

Design: Stratified Sample

Defined: We take H separate SRS from *each* of the H strata.

- Assumption: sampling unit = observation unit
- Assumption: done *without replacement*
- Notation: n_h is the SRS sample size of stratum h

Design: Stratified Sample

Why?

- Can be more precise than a SRS
- May want to estimate within strata
- May want to oversample smaller strata to achieve a certain level of precision
- May want to use different contact methods in different stratum

Inclusion probabilities: Stratified

What is the probability that unit j from stratum h is selected?

Sampling weights: Stratified

What is the sampling weight for unit j from stratum h under a stratified design?

Estimation plan: Stratified

- y_{hj} is the response from unit j in stratum h
- Use a Horvitz-Thompson estimator to estimate the (overall) **population total**

$$\hat{t}_{HT} = \sum_{\text{sampled units}} w_{hj} y_{hj}$$

Population Total: Stratified

- Parameter: $t = \sum_{h=1}^{H} \sum_{j=1}^{N_h} y_{hj} = \sum_{h=1}^{H} t_h$

Population Total: Stratified

- Parameter: $t = \sum_{h=1}^{H} \sum_{j=1}^{N_h} y_{hj} = \sum_{h=1}^{H} t_h$
- Estimator (unbiased)

$$\hat{t}_{str} = \sum_{h=1}^{H} N_h \bar{y}_h = \sum_{h=1}^{H} \hat{t}_h$$

where $\bar{y}_h$ is the sample mean response in stratum h.

Population Total: Stratified

- Parameter: $t = \sum_{h=1}^{H} \sum_{j=1}^{N_h} y_{hj} = \sum_{h=1}^{H} t_h$
- Estimator (unbiased)

$$\hat{t}_{str} = \sum_{h=1}^{H} N_h \bar{y}_h = \sum_{h=1}^{H} \hat{t}_h$$

where $\bar{y}_h$ is the sample mean response in stratum h.

- Standard error (estimated variation in $\hat{t}_{str}$)

$$SE(\hat{t}_{str}) = \sqrt{\sum_{h=1}^{H} SE(\hat{t}_h)^2} = \sqrt{\sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_h^2}{n_h}}$$

where s_h is the sample standard deviation in stratum h.

Population Mean: Stratified

- Parameter: $\bar{y}_{\mathcal{U}} = \dfrac{t}{N} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_{\mathcal{U},h}$

Population Mean: Stratified

- Parameter: $\bar{y}_{\mathcal{U}} = \dfrac{t}{N} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_{\mathcal{U},h}$
- Estimator (unbiased)

$$\bar{y}_{str} = \frac{\hat{t}_{str}}{N} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h$$

Population Mean: Stratified

- Parameter: $\bar{y}_{\mathcal{U}} = \dfrac{t}{N} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_{\mathcal{U},h}$
- Estimator (unbiased)

$$\bar{y}_{str} = \frac{\hat{t}_{str}}{N} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h$$

- Standard error (estimated variation in $\bar{y}_{str}$)

$$SE(\bar{y}_{str}) = \frac{SE(\hat{t}_{str})}{N} = \sqrt{\sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_h^2}{n_h}}$$

Population Proportion of "successes": Stratified

- $y_{hj} = 1$ if unit j's response is a "success" and 0 otherwise
- Parameter: $p = \dfrac{\text{number of successes in pop.}}{\text{pop. size}} = \sum_{h=1}^{H} \frac{N_h}{N} p_h$
- Estimator (unbiased)

$$\hat{p}_{str} = \sum_{h=1}^{H} \frac{N_h}{N} \hat{p}_h$$

  where $\hat{p}_h$ is the sample proportion of successes in stratum h.
- Standard error (estimated variation in $\hat{p}_{str}$)

$$SE(\hat{p}_{str}) = \sqrt{\sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\hat{p}_h(1-\hat{p}_h)}{n_h - 1}}$$

Confidence intervals: Stratified

- Same idea as a SRS for an overall total/mean/proportion CI:

$$\text{estimate} \pm q \times SE$$

- For overall population estimate: use $df = n - H$ where $n = n_1 + \cdots + n_H$ is the total sample size

Estimation plan within a stratum: SRS

To estimate a **stratum total/mean/proportion**, use SRS estimation methods.