Lecture 3: Unequal Prob. Sampling

SRS: iid, sampling w/ equal probability

Unequal prob sample:

1. Stratified random sample
2. Survey non-response

} → survey weights

Probability of being in sample: π_i

→ SRS: $\pi_i = \frac{n}{N}$ ← sample / ← population

→ Stratified sample
→ Differential non-response

} variable π_i

Survey weights (sampling weights) ← depend on π_i

1. design weights
2. nonresponse weights or poststratification weights

Finite population Horvitz-Thompson Estimator for the Total

Total: $Y = \sum_{i=1}^{N} y_i$ ← population

population $i = 1, \cdots, N$

Survey: probability π_i → sample $i = 1, \cdots, n$

Estimator: $\hat{Y} = \sum_{i=1}^{n} w_i y_i$

e.g. $w_1 y_1 + w_2 y_2 + w_3 y_3 \cdots$

How to set w_i? Unbiased...

$$E[\hat{Y}] = Y$$

$$E\left[\sum_{i=1}^{n} w_i y_i\right] = Y$$

$$E\left[\sum_{i=1}^{N} 1[i \in S] w_i y_i\right] = Y$$

$$\sum_{i=1}^{N} E[1[i \in S] w_i y_i] = \sum_{i=1}^{N} y_i$$

$$\sum_{i=1}^{N} \pi_i w_i y_i = \sum_{i=1}^{N} y_i$$

→ need to set $w_i = \frac{1}{\pi_i}$ ← # of people represented by i

Weighted average

$$\hat{\mu} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} = \sum_{i=1}^{n} \left(\frac{w_i}{\underbrace{\sum_{i=1}^{n} w_i}} \right) y_i$$

$$\sum_{i=1}^{n} w_i = N$$

$$= \frac{\hat{Y}}{N} \rightarrow \text{unbiased estimator for } \mu$$

This estimator works for:

$\rightarrow w_i = \frac{1}{\pi_i}$

$\rightarrow w_i = \frac{n}{N} \frac{1}{\pi_i}$

$\rightarrow w_i = 1m * \frac{1}{\pi_i}$

Unbiased <u>and</u> consistent

Weighted average in a stratified sample:

→ $\bar{y}^s$ for stratum $s \in \{u, r\}$

→ $\hat{\mu} = \frac{N^u}{N} \bar{y}^u + \frac{N^r}{N} \bar{y}^r$

→ $\pi^u = \frac{n^u}{N^u}$, $\pi^r = \frac{n^r}{N^r}$ → $w^u = \frac{1}{\pi^u}$, $w^r = \frac{1}{\pi^r}$

→ $\hat{\mu} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} =$

Weighted least squares (WLS)

$$\min_{\hat{b}_0^{WLS},\, \hat{b}_1^{WLS}} \sum_{i=1}^{n} w_i \left(y_i - \hat{b}_0^{WLS} - \hat{b}_1^{WLS} x_i\right)^2$$

$$\hat{\beta}_1^{WLS} = \frac{\sum_{i=1}^{n} w_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} w_i (x_i - \bar{x})^2}$$

$\hat{\beta}_1^{WLS}$ is unbiased and consistent for β_1^{POP}, where β_1^{POP} comes from the regression in the population.

Good! In this sense $\hat{\beta}_1^{WLS}$ is <u>representative</u>.

Structural equation:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

constant / homogeneous

$$y_i = \beta_{0i} + \beta_{1i} x_i + u_i$$

heterogeneous $\rightarrow$ what is $\bar{\beta}_1 = \frac{1}{N}\sum_{i=1}^{N} \beta_{1i}$

In general, $\hat{\beta}_1^{WLS} \neq \bar{\beta}_1$

because $\beta_1^{POP} \neq \bar{\beta}_1$

$$\beta_1^{POP} = \frac{cov(x_i, y_i)}{V[x_i]} = \frac{cov(x_i, \beta_{0i} + \beta_{1i} x_i + u_i)}{V[x_i]}$$

$$= \frac{cov(x_i, \beta_{0i})}{V[x_i]} + \frac{cov(x_i, \beta_{1i} x_i)}{V[x_i]} = \bar{\beta}_1 \text{ if } x_i \text{ randomized}$$

not otherwise