Lecture Note 3: Unequal probability sampling

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

Unequal prob. samples:

1. stratified random sample
2. survey non-response

} π_i → survey weight **or** sampling weight

Types of weights:

1. design weights
2. nonresponse weights or post-stratification weights

Finite population: $i = 1, \dots, N$

Sample (S): $i = 1, \dots, n$

<u>Horvitz - Thompson</u>

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

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

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

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

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

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

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

need to set $w_i = \frac{1}{\pi_i}$ ← # people in the pop represented by i

Weighted average

$$\hat{\mu} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} = \sum_{i=1}^{n} \left(\frac{w_i}{\sum_{i=1}^{n} w_i} \right) y_i = \frac{\hat{Y}}{N} \rightarrow \text{unbiased and consistent estimator for } \mu = \frac{Y}{N}$$

$\nwarrow N$

Works for different variants of w_i:

- $\rightarrow w_i = \frac{1}{\pi_i}$
- $\rightarrow w_i = \frac{n}{N} \frac{1}{\pi_i}$
- $\rightarrow w_i = 1\text{m} \times \frac{1}{\pi_i}$

Suppose we $\bar{y}^s$ in stratum $s \in \{u, r\}$

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

Same as

$$\hat{\mu} = \frac{\sum_i w_i y_i}{\sum_i w_i}$$

$$\pi^u = \frac{n^u}{N^u}, \quad \pi^r = \frac{n^r}{N^r}$$

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

Weighted least squares

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

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

as an estimator for

$$\beta_1^{POP} = \frac{\sum_{i=1}^{N} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{N} (x_i - \bar{x})^2}$$

$\hat{\beta}_1^{WLS}$ is an unbiased and consistent estimator for β_1^{POP}

In this sense, $\hat{\beta}_1^{WLS}$ is representative

But efficiency cost

Structural equation:

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

constant, homogeneous coefficients

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

heterogeneous coefficients

$\hat{\beta}_1^{OLS}$ not unbiased or consistent estimator of $\bar{\beta}_1 = \frac{\sum_{i=1}^{N} \beta_{1i}}{N}$

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

$\beta_1^{POP} = \frac{cov(x_i, y_i)}{V[x]}$ only $= \bar{\beta}_1^{POP}$ if x_i randomly assigned