Estimation in Domains

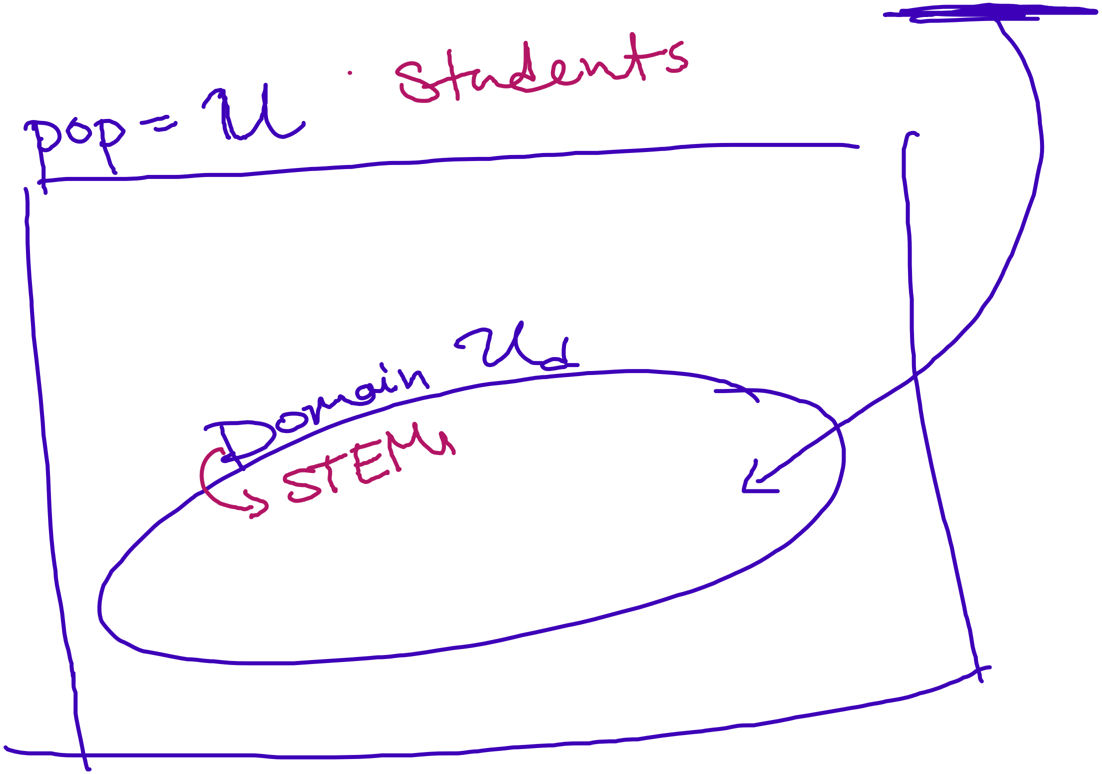
Week 5 (4.2)

Stat 260, St. Clair

Domains

Domain: subpopulations of interest $\mathcal{U}_d$

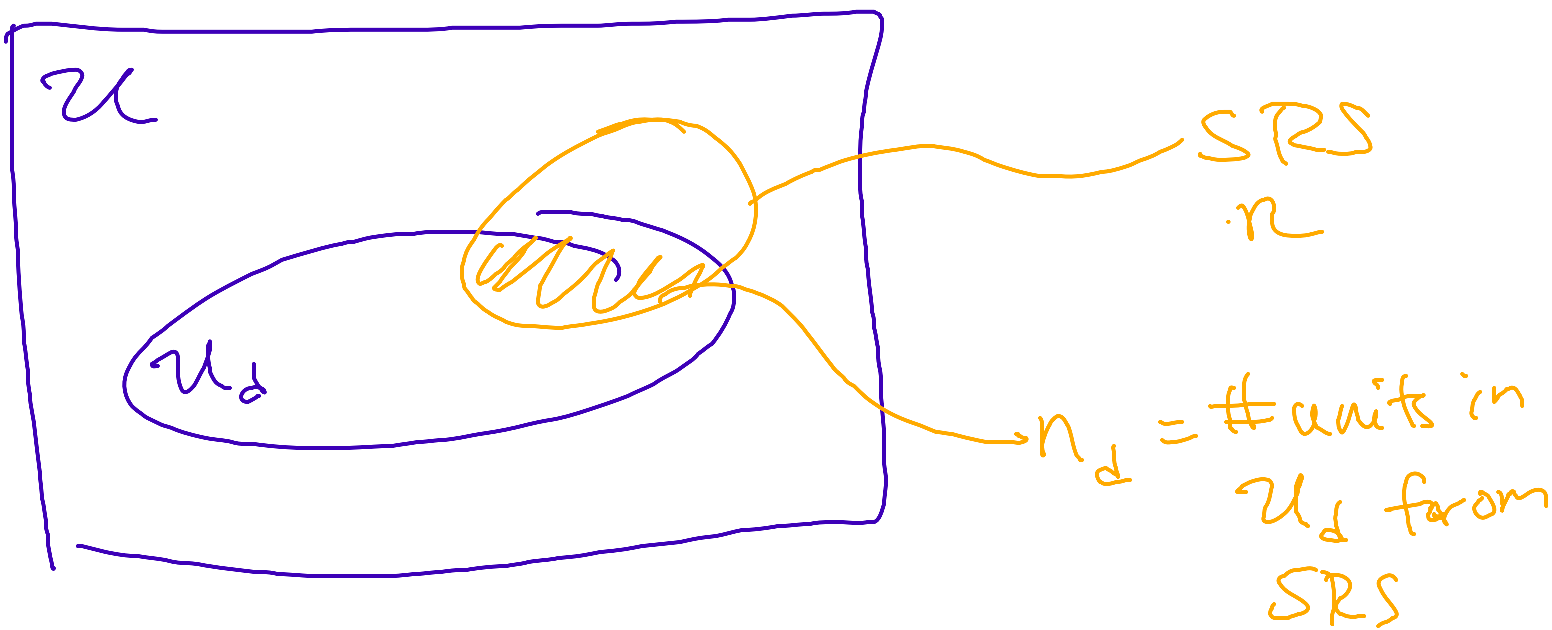
- We want to estimate a domain parameter: $t_d, \bar{y}_{\mathcal{U}_d}, p_d$



Design

We take a SRS of size n from a population $\mathcal{U}$

- **Problem:** n_d, the number of respondents in the domain, varies from sample to sample



Domain estimation: a ratio estimator

- Domain indicator:

$$x_i = \begin{cases} 1 & \text{if unit } i \text{ is in the domain} \\ 0 & \text{if unit } i \text{ is not in the domain} \end{cases}$$

STEM

not STEM

- Domain sample size:

$$\# \quad n_d = \sum_{i=1}^{n} x_i$$

= # STEM in sample

$$N_d = \sum_{i=1}^{N} x_i =$$

(domain) STEM in population

Domain estimation: a ratio estimator

- Domain responses: u_i for every sampled unit

$$u_i = x_i y_i = \begin{cases} y_i & \text{if unit } i \text{ is in the domain} \\ 0 & \text{if unit } i \text{ is not in the domain} \end{cases}$$

- Domain sample mean:

$$\bar{y}_d = \frac{\sum_{i \in \text{domain}} y_i}{n_d} = \frac{\sum_{i=1}^{n} u_i}{\sum_{i=1}^{n} x_i} \longrightarrow \text{SE calc. based on ratio est. SE}$$

ratio

$$\frac{\sum_{i=1}^{n} \cdot}{\sum_{i=1}^{n} \cdot}$$

Domain estimation for a mean

- **Domain Parameter** mean $\bar{y}_{\mathcal{U}_d}$
- **Estimator** with a SRS of units: a ratio estimator

$$\bar{y}_d = \frac{\sum_{i=1}^n u_i}{\sum_{i=1}^n x_i}$$

domain sample mean

- **SE**: when n is "large" (or bias about 0)

$$SE(\bar{y}_d) \approx \sqrt{\left(1 - \frac{n}{N}\right) \frac{s_e^2}{n\bar{x}^2}} = \sqrt{\left(1 - \frac{n}{N}\right) \left(\frac{n}{n-1}\right) \left(\frac{n_d - 1}{n_d}\right) \frac{s_d^2}{n_d}}$$

where s_d is the sample SD of the measurements y_i in the sampled domain.

Stats in domain: $\bar{y}_d$ s_d

if n & n_d large: $SE(\bar{y}_d) \approx \frac{s_d}{\sqrt{n_d}}$

$N >> n$

Domain estimation for a proportion

- **Domain Parameter** proportion p_d
- Response y_i is an indicator of "success"
- **Estimator** with a SRS of units: a ratio estimator

$$\hat{p}_d = \frac{\sum_{i=1}^n u_i}{\sum_{i=1}^n x_i} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i}$$

= sample prop. of success in domain

- **SE**: same as the mean except $s_d^2 = \dfrac{n_d}{n_d - 1}\hat{p}_d(1 - \hat{p}_d)$

Domain estimation for a total

- **Domain Parameter** total t_d

Two scenarios:

1. If N_d, population size of the domain, is **known**

$$\hat{t}_d = N_d \bar{y}_d, \quad SE(\hat{t}_d) = N_d SE(\bar{y}_d)$$

Domain estimation for a total

- **Domain Parameter** total t_d $= \sum_{i=1}^{N} y_i \cdot x_i$

$= \sum_{i=1}^{N} u_i$

Two scenarios:

2. If N_d, population size of the domain, is **unknown**

- **Estimator**: SRS total estimate with u_i as the response

Measurement u_i?

SRS
est total
u_i

$$\hat{t}_d = N\bar{u}$$

$$\hat{t}_d = N\bar{u} = \frac{N}{n}\sum_{i=1}^{n} u_i$$
$$= \frac{N}{n}\sum_{i=1}^{n} x_i y_i \quad \bar{y}_d$$
$$= \frac{N}{n} \sum_{\text{domain}} y_i \cdot \frac{n_d}{n_d}$$
$$= \underbrace{N\left(\frac{n_d}{n}\right)}_{\tilde{N}_d} \bar{y}_d$$

- **SE**:

$$SE(\hat{t}_d) = N SE(\bar{u}) = N\sqrt{\left(1 - \frac{n}{N}\right)\frac{s_u^2}{n}}$$

where s_u is the sample SD of the measurements u_i in the sample and

$$s_u^2 = \frac{1}{n-1}\left[(n_d - 1)s_d^2 + n_d \bar{y}_d^2 \left(1 - \frac{n_d}{n}\right)\right]$$

Example

An economist wants to estimate the average weekly amount spent on food by households containing children in a small town. A complete list of all 2500 households in the county is available, but identifying those households with children is impossible. So the economist selects a SRS of 500 households and observes 420 that contain children. Of the 420 households with children, he records an average of $120.35 spent on food during a week and a sample SD of $42.20.

Population

units = households (HH) $\quad N = 2500$

Domain : HH with children

y_i = cost of food

SRS : $n = 500 \quad n_d = 420 \rightarrow \bar{y}_d = \$120.35$

$s_d = \$42.20$

Example:

$\bar{y}_{\mathcal{U}_d}$

Estimate the average weekly amount of money spent on food by all households with children in the county and compute the standard error of your estimate.

<u>Est</u> : $\bar{y}_d = \$120.35$

$s_d = 42.20$

<u>SE</u> : $n = 500$ $\quad n_d = 420$

$N = 2500$

$$SE(\bar{y}_d) = \sqrt{\left(1 - \frac{500}{2500}\right)\left(\frac{500}{500-1}\right)\left(\frac{420-1}{420}\right)\frac{42.2^2}{420}} = \$1.84$$

Example:

t_d

Estimate the total weekly amount of money spent on food by households with children and compute the standard error of your estimate.

$N_d = ??$ # HH in pop. with kids

$\hat{t}_d = N\bar{u}$

$$\bar{u} = \frac{\sum_{i=1}^{\hat{}} u_i}{500} = \frac{\bar{y}_d \times n_d}{500} = \frac{(120.35)(420)}{500} = \frac{50,547}{500}$$

$$\hat{t}_d = 2500\left(\frac{50,547}{500}\right) = \boxed{\$252,735} = 2500\left(\frac{420}{500}\right)120.35$$

$$SE(\hat{t}_d) = 2500\sqrt{\left(1-\frac{500}{2500}\right)\frac{s_u^2}{500}} = \$5870.18$$

$$s_u^2 = \frac{1}{500-1}\left[(420-1)42.2^2 + 420(120.35)^2\left(1-\frac{420}{500}\right)\right]$$

$= 3445.902$ D

sample SD of u_i's

Bias of domain mean estimator

Small when n is large or n_d/n is close to 1

want small

Ratio estimator bias facts

- $$\left|\text{Bias}(\bar{y}_d)\right| \le \frac{SE(\bar{y}_d)\,SE(\bar{x})}{|\bar{x}_u|} = \frac{SE(\bar{y}_d)\sqrt{1-\frac{n}{N}}\,\frac{s_x}{\sqrt{n}}}{|\bar{x}_u|}$$

$$\bar{x}_u = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{N_d}{N} \approx \frac{n_d}{n}$$

x_i = ind.

$\hat{p}(1-\hat{p})$

$$\frac{s_x}{|\bar{x}_u|} \approx \frac{\frac{n_d}{n}\left(1-\frac{n_d}{n}\right)}{n_d/n} = 1-\frac{n_d}{n} \approx 0$$

$$\frac{n_d}{n} \approx 1$$

Proof

$\bar{y}_d = \frac{\sum u_i}{\sum x_i}$

$$SE(\bar{y}_d) \approx \sqrt{\left(1-\frac{n}{N}\right)\frac{s_e^2}{n\bar{x}^2}} \overset{\nearrow \text{proof}}{=} \sqrt{\left(1-\frac{n}{N}\right)\left(\frac{n}{n-1}\right)\left(\frac{n_d-1}{n_d}\right)\frac{s_d^2}{n_d}}$$

① $\bar{x}^2 = \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 = \left(\frac{n_d}{n}\right)^2$

② Ratio $e_i = u_i - \bar{y}_d x_i$

s_e^2 = sample SD of e_i

$= \frac{1}{n-1}\sum_{i=1}^{n}(e_i - \bar{e})^2$

$$\bar{e} = \frac{1}{n}\sum(u_i - \bar{y}_d x_i) = \frac{1}{n}\left[\sum u_i - \bar{y}_d \sum x_i\right]$$

$\bar{y}_d \sum x_i \downarrow \frac{\sum u_i}{\sum x_i}$

$$= \frac{1}{n}\left[\sum u_i - \sum u_i\right] = \underline{\underline{0}}$$

$$s_e^2 = \frac{1}{n-1}\sum(e_i - 0)^2 = \frac{1}{n-1}\sum_{i=1}^{n}(u_i - \bar{y}_d x_i)^2$$

Fact SD: $s^2 = \frac{1}{n-1}\sum(y_i-\bar{y})^2 = \cdots = \frac{1}{n-1}\left[\sum y_i^2 - n\bar{y}^2\right]$

$$\sum_{i=1}^{n} y_i^2 = (n-1)s^2 + n\bar{y}^2$$

Next! $\longrightarrow$

$$s_e^2 = \frac{1}{n-1}\sum_{i=1}^{n}(u_i - \bar{y}_d x_i)^2 \rightarrow \sum \text{ is } 0 \text{ for any } i \text{ not in domain}$$

d

$x_i y_i$

$$= \frac{1}{n-1} \underbrace{\sum_{\text{domain}}(y_i - \bar{y}_d)^2 \cdot \frac{n_d-1}{n_d-1}}_{s_d^2} = \frac{n_d-1}{n-1} s_d^2$$

$$SE(\bar{y}_d) \approx \sqrt{\left(1-\frac{n}{N}\right)\frac{\frac{n_d-1}{n-1}s_d^2}{n\,(n_d/n)^2}} = \sqrt{\left(1-\frac{n}{N}\right)\left(\frac{n_d-1}{n_d}\right)\frac{n}{n-1}\frac{s_d^2}{n_d}}$$

$\swarrow$

Proof

$$s_u^2 = \frac{1}{n-1}\sum_{i=1}^{n}(u_i-\bar{u})^2 \overset{\text{prove}}{=} \frac{1}{n-1}\left[(n_d-1)s_d^2 + n_d\bar{y}_d^2\left(1-\frac{n_d}{n}\right)\right].$$

$$S_u^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} u_i^2 - n\bar{u}^2\right] = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 y_i^2 - n\left(\frac{1}{n}\sum x_i y_i\right)^2\right]$$

$\hookrightarrow x_i = 0$ when i not in domain

$$= \frac{1}{n-1}\left[\underbrace{\sum_{\text{domain}} y_i^2}_{(n_d-1)S_d^2 + n_d\bar{y}_d^2} - \frac{n}{n^2}\Big(\underbrace{\sum_{\text{domain}} y_i \cdot \frac{n_d}{n_d}}_{\bar{y}_d}\Big)^2\right]$$

$$= \frac{1}{n-1}\left[(n_d-1)S_d^2 + n_d\bar{y}^2 - \frac{n_d^2}{n}\bar{y}_d^2\right]$$

$$= \frac{1}{n-1}\left[(n_d-1)S_d^2 + n_d\bar{y}_d^2\left(1-\frac{n_d}{n}\right)\right]$$