Ratio estimation - why and when

Week 4 (4.1 up to 4.1.2)

Stat 260, St. Clair

Example 1: revisit petition

Signatures to a petition were collected on 676 sheets of paper. Along with each signature is a record of whether they are registered to vote.

How do you measure the proportion of signatures that belong to registered voters on all sheets?

$$p = \frac{\#\text{ sig from registered voters}}{\#\text{ signatures}} = \frac{\sum_{i=1}^{676} y_i}{\sum_{i=1}^{676} x_i} = \frac{t_y}{t_x}$$

unit = sheet . For sheet i

y_i = # sig. from registered voters on sheet i

X_i = # sig. on sheet i

Example 1: revisit petition

x

You take a SRS of 50 sheets and find 1515 signatures, 771 of which are from registered voters.

y

units

Use this information to estimate the proportion of all signatures that are from registered voters.

$n = 50$ $\quad \sum_{i=1}^{50} x_i = 1515 \quad \sum_{i=1}^{50} y_i = 771$

$$\hat{p} = \frac{771}{1515} = \frac{\hat{t}_y}{\hat{t}_x} = \frac{676\left(\frac{771}{50}\right)}{676\left(\frac{1515}{50}\right)} \approx \underline{.51}$$

$\hat{t}_y \longrightarrow$ SRS

$\hat{t}_x \downarrow$ SRS

Example 1: revisit petition

The estimated proportion signatures belongin to registered voters is

$$\hat{p} = \frac{771}{1515} = 0.5089$$

Why can we **not use a SRS SE formula** to measure the variability of this estimate?

$SE(\hat{p}) = ??$

SRS prop = denominater = sampling unit count (n, N)

→ fixed quantity

Here: denom of $\hat{p}$ = count of sig. → vary from SRS to SRS

Need different SE calc. than SRS

Ratio estimation: for a ratio parameter

- For each sampling unit, we measure two variables: x_i and y_i
- **Ratio Parameter** Suppose our parameter of interest looks like

Pop.

$$B = \frac{\sum_{i=1}^N y_i}{\sum_{i=1}^N x_i} = \underbrace{\frac{t_y}{t_x}}_{\frac{t_y/N}{t_x/N}} = \frac{\bar{y}_{\mathcal{U}}}{\bar{x}_{\mathcal{U}}}$$

pop. mean ratio

Ratio

- **Estimator** with a SRS of units:

$$\hat{B} = \underbrace{\frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i}}_{\text{ratio of sample totals}} = \frac{\bar{y}}{\bar{x}} = \frac{N\bar{y}}{N\bar{x}} = \frac{\hat{t}_y}{\hat{t}_x}$$

SRS est.

- SE and bias: next video

Example 2: food costs

If there are N households in the city, how would you form the parameter that measures the average weekly food costs per resident in Northfield if you plan to take a SRS of n households in the city?

unit = household

N = known

n = # HH sampled

y_i = weekly food costs in HH i

x_i = HH size (# residents)

$\bar{y}$ = avg. food cost per household

$$B = \frac{\text{total food costs}}{\text{total \# residents}} = \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i} \quad *$$

$$\hat{B} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}$$

Example 3: food costs take 2

Suppose you don't know N, the number of households (sampling units) in the population.

t_x

You do know t_{income} from IRS records. → auxiliary info

If you have a SRS of n households with income recorded, how can you estimate N?

$\boxed{\text{New } x}$ x_i = income of HH i

$$\frac{t_{income}}{N} = \bar{x}_u = \text{pop. mean income per HH}$$

$$N = \frac{t_{income}}{\bar{x}_u}$$

$$\boxed{\hat{N} = \frac{t_{income}}{\bar{x}_{SRS}} = \frac{\text{known pop total}}{\text{SRS mean est.}}}$$

Example 3: food spending take 2

(Ex.2)

N = unknown

y_i

How can you estimate the total weekly food costs for all residents of Northfield if you have an SRS of n households and know the total income for all residents.

Goal: $t_y = N\bar{y}_u = \left(\frac{t_x}{\bar{x}_u}\right)\bar{y}_u = t_x\left(\frac{\bar{y}_u}{\bar{x}_u}\right)$

B

$$\boxed{t_y = t_x B}$$

Estimate: $\underbrace{\hat{t}_{y,r}}_{\text{ratio}} = t_x\hat{B} = \underbrace{t_x}_{\text{known}}\underbrace{\left(\frac{\bar{y}}{\bar{x}}\right)}_{\text{ratio of sample means (per HH)}}$

Ratio estimation for a mean or total

using $\hat{B}$

- Suppose we know a population mean or total for an **auxiliary** variable x

$$t_x \text{ and/or } \bar{x}_U \text{ are known}$$

- **Population total** for response y can be written

$$t_y = \frac{t_y}{t_x} t_x = B t_x \rightarrow t_x \text{ known}$$

- **Population ~~total~~** mean for response y can be written

$$\bar{y}_{\mathcal{U}} = \frac{\bar{y}_{\mathcal{U}}}{\bar{x}_{\mathcal{U}}} \bar{x}_{\mathcal{U}} = B \bar{x}_{\mathcal{U}} \rightarrow \bar{x}_u \text{ known}$$

- **Estimators** use an estimated ratio $\hat{B} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \frac{\bar{y}}{\bar{x}} = \frac{\hat{t}_y}{\hat{t}_x}$

Example 4: a classic

The 1920 census found that the total number of residents living in 196 selected major U.S. cities was 22,919 thousand people.

We would like to determine the total number of residents living in these cities in 1925 (a year with no census).

who?	**1920** (x)	**1925** (y)
population	$t_x = 22,919$	$t_y = ??$

Example 4: a classic

unit = city

N = 196

n = 49

A random sample of 49 of these cities was taken in 1925 and for each city in the sample the 1920 and 1925 population sizes (in the thousands) were recorded:

who?	1920 (x)	1925 (y)
population	$t_x = 22,919$	$t_y = ??$
sample	$\sum_{i=1}^{49} x_i = 5054$	$\sum_{i=1}^{49} y_i = 6262$

What is a SRS estimate of the total number of people in these 196 cities in 1920? in 1925?

$$1920: \quad \hat{t}_{x,SRS} = N\bar{x} = 196\left(\frac{5054}{49}\right) = 20,216$$

$$1925: \quad \hat{t}_{y,SRS} = N\bar{y} = 196\left(\frac{6262}{49}\right) = 25,048$$

Example 4: a classic

who?	1920 (x)	1925 (y)
population	$t_x = 22,919$	$t_y =??$
sample	$\sum_{i=1}^{49} x_i = 5054$	$\sum_{i=1}^{49} y_i = 6262$

SRS estimates:

$$\hat{t}_{1920} = 20,216 \quad \hat{t}_{1925} = 25,048$$

- Does this SRS over or under estimate the 1920 total? Is it likely that the SRS does the same for the 1925 total?

underestimating $\hat{t}_{1920} < 22,919$

1925: If 1920 + 25 sizes are positively correlated, then we might be underestimating t_y in 1925 with this SRS.

Example 4: a classic

who?	1920 (x)	1925 (y)
population	$t_x = 22,919$	$t_y = ??$
sample	$\sum_{i=1}^{49} x_i = 5054$	$\sum_{i=1}^{49} y_i = 6262$

What is a **ratio estimate** of the total number of people in these 196 cities in 1925? Is this greater or smaller than the SRS estimate?

$$t_y = B t_x = B(22,919)$$

$$\hat{B} = \frac{\sum_{i=1}^{49} y_i}{\sum_{i=1}^{49} x_i} = \frac{6262}{5054} = 1.239$$

Ratio est of 1925 total

$$\hat{t}_{y,\text{ratio}} = (1.239)(22,919) = \boxed{28,397}$$

1925

Calibration weights

- Usual SRS sampling weights are N/n
- Ratio estimates of total/mean have an added **calibration weight** of

$$g_i = \frac{t_x}{\hat{t}_x} = \frac{\bar{x}_{\mathcal{U}}}{\bar{x}}$$

- These weights adjust the SRS estimate up or down, depending on whether the auxilary mean/total is over- or underestimated in the SRS.

$$\hat{t}_{y,\text{ratio}} = \hat{B}\, t_x = \frac{\hat{t}_y}{\hat{t}_x} t_x = \left(\frac{t_x}{\hat{t}_x}\right)\hat{t}_y$$

SRS est. $t_y = N\bar{y}$

$$= \left(\frac{t_x}{\hat{t}_x}\right)\sum_{i=1}^{n}\left(\frac{N}{n}\right) y_i$$

$$\hat{t}_{y,\text{ratio}} = \sum_{i=1}^{n}\left(\frac{N}{n}\right)\left(\frac{t_x}{\hat{t}_x}\right) y_i = \sum_{i=1}^{n} w_i g_i y_i$$

Example 4: a classic

who?	**1920** (x)	**1925** (y)
population	$t_x = 22,919$	$t_y = ??$
sample	$\sum_{i=1}^{49} x_i = 5054$	$\sum_{i=1}^{49} y_i = 6262$

$$\hat{t}_{y,srs} = 25,048 \quad \hat{t}_{y,ratio} = 28,397$$

What is the calibration weight for this ratio estimator?

$$g_i = \frac{t_x}{\hat{t}_{x,SRS}} = \frac{22,919}{20,216} = 1.134$$

$$\hat{t}_{y,ratio} = \left(\frac{t_x}{\hat{t}_{x,SRS}}\right)\hat{t}_{y,SRS} = (1.134)(25,048)$$

$$= 28,397$$

(same ratio est.)