Ratio estimation - the details

Week 4 (4.1)

Stat 260, St. Clair

Ratio estimation: for a ratio parameter

$\ast$ SRS

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

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

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

$$\hat{B} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \frac{\bar{y}}{\bar{x}} = \frac{N\bar{y} \to \hat{t}_y}{N\bar{x} \to \hat{t}_x}$$

- SE and bias: now!

Ratio estimation: for a ratio parameter

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

$$SE(\hat{B}) \approx \sqrt{\left(1 - \frac{n}{N}\right) \frac{s_e^2}{n\bar{x}^2}}$$

where s_e is the sample SD of the "errors" $e_i = y_i - \hat{B}x_i$

$\sqrt{\frac{1}{n-1}\sum(e_i-\bar{e})^2}$

- Computational "by hand" short cut:

$$s_e^2 = s_y^2 + \hat{B}^2 s_x^2 - 2\hat{B}\hat{R}s_y s_x$$

  - s_x, s_y: sample SD's of x/y measurements
  - $\hat{R}$: correlation coefficient of the sampled x/y measurements

intro stats

$-1 \le r = \hat{R} \le 1$

$|\hat{R} \approx \pm 1|$ $\boxed{y \approx c \cdot x}$

$\frac{y}{x} \approx c = B.$

$e_i \approx 0$ (small)

Example 1: revisit petition

$N = 676$ $\quad y_i = \#$ voter sig.

$n = 50$ $\quad x_i = \#$ sig.

Signatures to a petition were collected on 676 sheets of paper.

- You take a SRS of 50 sheets and find 1515 signatures, 771 of which are from registered voters.
- The estimated proportion signatures belonging to registered voters is

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

Compute the SE for this estimate.

$$p = B = \frac{\sum_{i=1}^{676} y_i}{\sum_{i=1}^{676} x_i}$$

$$\bar{x} = \frac{\# \text{ of sig}}{\# \text{ sheets}} = \# \text{ sig/sheet} = \frac{1515}{50} \checkmark$$

Example 1: revisit petition

$\hat{B} = .5089$

S_x

- The standard deviation for the number of signatures per sheet is 2.929 and the standard deviation for the number of registered voter signatures per sheet is 3.909. $\to S_y$
- The sample correlation between the total number of signatures and number of registered voter signatures per sheet is 0.4497. $\to \hat{R}$

$$s_e^2 = 3.909^2 + (.5089)^2\, 2.929^2 - 2(.5089)(.4497)(2.929)(3.909)$$

$$\approx 11.51$$

$$SE(\hat{B}) = \sqrt{\left(1 - \frac{50}{676}\right)\frac{11.51}{(50)(1515/50)^2}} \approx .015$$

est. $\Rightarrow$ 50.9% of sig. were from reg. voters (SE of 1.5%)

Ratio estimation for a mean or total

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

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

$$t_y = \frac{t_y}{t_x} \cdot t_x$$
$$= B \cdot t_x$$

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

$$\hat{t}_{y,r} = \hat{B} t_x$$

- **Population mean estimate** for response y can be written

$$\hat{\bar{y}}_r = \hat{B} \bar{x}_{\mathcal{U}}$$

- SE and bias: Now!

Ratio estimation for a mean or total

- SE: when n is "large" (or bias of $\hat{B}$ about 0)

$$SE(\hat{t}_{y,r}) \approx \sqrt{\left(1 - \frac{n}{N}\right)\left(\frac{t_x}{\bar{x}}\right)^2 \frac{s_e^2}{n}}$$

$$SE(\hat{\bar{y}}_r) \approx \sqrt{\left(1 - \frac{n}{N}\right)\left(\frac{\bar{x}_{\mathcal{U}}}{\bar{x}}\right)^2 \frac{s_e^2}{n}}$$

where s_e is the sample SD of the "errors" $e_i = y_i - \hat{B}x_i$

- Computational "by hand" short cut:

$$s_e^2 = s_y^2 + \hat{B}^2 s_x^2 - 2\hat{B}\hat{R}s_y s_x$$

$SE(\hat{t}_{y,r}) = SE(\hat{B}t_x) = t_x \underline{SE(\hat{B})}$

$N=196 \quad n=49$ SRS of cities $\Rightarrow$ pop. size

Example 4: a classic

$t_x = \sum_{i=1}^{196} x_i = 22,919$

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

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

what?	1920 (x)	1925 (y)
population total	$t_x = 22,919$	$t_y = ??$
sample total	$\sum_{i=1}^{49} x_i = 5054$	$\sum_{i=1}^{49} y_i = 6262$
sample variance	$s_x^2 = 10,900.4$	$s_y^2 = 15,158.8$
sample correlation	$\hat{R} = 0.9817$	

$\hat{t}_{y,ratio} = 28,397 \quad SE(\hat{t}_{y,ratio}) = ??$ $\rightarrow$ 4b

$$\hat{t}_{y,r} = \left(\frac{\hat{t}_{y,SRS}}{\hat{t}_{x,SRS}}\right) t_x = 25,048\left(\frac{22919}{20,216}\right) = 25,048\underbrace{(1.134)}_{\text{calibration weights}}$$

$\hat{B}$

$s_e^2 = 621.85$ (Alg. here)

$$SE(\hat{t}_{y,r}) = \sqrt{\left(1-\frac{49}{196}\right)\left(\frac{22919}{5054/49}\right)^2 \frac{621.85}{49}} = \boxed{685.5}$$

smaller than $SE(\hat{t}_{y,SRS})$

Ratio estimation for a mean or total vs. SRS estimation

When is

$$V(\hat{\bar{y}}_r) \overset{?}{<} V(\bar{y}_{srs})$$

$$\left(1-\frac{n}{N}\right)\left(\frac{\bar{x}_u}{\bar{x}}\right)^2 \frac{s_e^2}{n} \overset{?}{<} \left(1-\frac{n}{n}\right)\frac{s_y^2}{n}$$

* Assume $\frac{\bar{x}_u}{\bar{x}} \approx 1 \implies \boxed{s_e^2 \overset{?}{<} s_y^2}$

$e_i = y_i - \hat{B}x_i \approx 0$

$y_i \approx c x_i \Rightarrow$ positive correlation!

(c is B)

general idea:
"strong" corr. between x & y
means $s_e^2 < s_y^2$

Theory: Bias approximation

Cov = how do two variables "covary" together

corr = $\frac{\text{cov}}{SD \cdot SD}$

$$B = \frac{t_y}{t_x} \quad \hat{B} = \frac{\hat{t}_y}{\hat{t}_x}$$

$$Bias(\hat{B}) = \underline{E(\hat{B})} - B = -\frac{Cov(\hat{B}, \bar{x})}{\bar{x}_{\mathcal{U}}}$$

Proof

A.2

$$Cov(\hat{B}, \bar{x}) = E[\hat{B}\bar{x}] - E(\hat{B}) E(\bar{x})$$

$E(\bar{x}) \Rightarrow \bar{x}_u$

$$= E\left[\frac{\bar{y}}{\bar{x}} \cdot \bar{x}\right] - E(\hat{B})\bar{x}_u$$

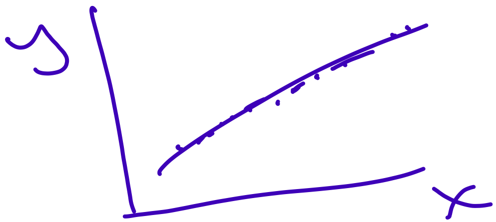
$$= \bar{y}_u - E[\hat{B}]\bar{x}_u$$

$\Rightarrow$ solve for $E(\hat{B})$

$$E[\hat{B}] = \frac{\bar{y}_u - Cov(\hat{B}, \bar{x})}{\bar{x}_u} = \underline{B} - \frac{Cov(\hat{B}, \bar{x})}{\bar{x}_u}$$

Theory: Bias approximation

What is bias if $y_i \approx cx_i$ for all i where c is some fixed value?



$$\hat{B} = \frac{\bar{y}}{\bar{x}} \approx \frac{c\bar{x}}{\bar{x}} = c$$

$$\mathrm{Cov}(\hat{B}, \bar{x}) \approx \mathrm{Cov}(c, \bar{x}) = 0$$

constant

$$\mathrm{Bias}(\hat{B}) \approx 0$$

Theory: Bias approximation

What is the maximum value of bias (approximately)?

$$|\mathrm{Bias}(\hat{B})| = \left| \frac{-\mathrm{Cov}(\hat{B}, \bar{x})}{\bar{x}_u} \right|$$

$$= \left| \frac{\mathrm{Corr}(\hat{B}, \bar{x})\, SE(\hat{B})\, SE(\bar{x})}{\bar{x}_u} \right|$$

$$\leq \frac{SE(\hat{B})\, SE(\bar{x})}{|\bar{x}_u|} \to SRS$$

large n!

x & y positive

corr.

$*$ probl (c)