DATA 606: Statistical Methods in Data Science

—— Ratio estimation

Wenjun Jiang

Department of Mathematics & Statistics
The University of Calgary

Lecture 4

An example

Goal: *France had no population census in 1802, and Laplace wanted to estimate the number of persons living there.*

- He obtained a sample of 30 communes spread throughout the country. These communes had a total of 2,037,615 inhabitants on September 23, 1802.
- In the 3 years preceding September 23, 1802, a total of 215,599 births were registered in the 30 communes.
- Laplace determined the annual number of registered births in the 30 communes to be $215{,}599/3 = 71{,}866.33$.
- Dividing 2,037,615 by 71,866.33, Laplace estimated that each year there was one registered birth for every 28.352845 persons.
- One could estimate the total population of France by multiplying the total number of annual births in all of France by 28.352845.

Ratio estimation in SRS

- For ratio estimation to apply, two quantities y_i and x_i must be measured on each sample unit.
- x_i is called an auxiliary variable.
- In the population of size N,

$$t_y = \sum_{i=1}^{N} y_i, \quad t_x = \sum_{i=1}^{N} x_i.$$

- Their ratio is

$$B = \frac{t_y}{t_x} = \frac{\bar{y}_U}{\bar{x}_U}.$$

Ratio estimation in SRS

- Ratio estimation takes advantage of the correlation of y and x. The more correlated they are, the better prediction x can make.
- Measure the correlation (Pearson correlation coefficient)

$$R = \frac{\sum_{i=1}^{N}(x_i - \bar{x}_U)(y_i - \bar{y}_U)}{(N-1)S_x S_y}.$$

- If an SRS is taken, then

$$\hat{B} = \frac{\bar{y}_S}{\bar{x}_S} = \frac{N \cdot \bar{y}_S}{N \cdot \bar{x}_S} = \frac{\hat{t}_y}{\hat{t}_x},$$

$$\hat{t}_{yr} = \hat{B} \cdot t_x,$$

$$\hat{\bar{y}}_r = \hat{B} \cdot \bar{x}_U.$$

Why we use ratio estimation

- *We simply want to estimate a ratio*
  A tip: If you took a different sample, the denominator would be a different number.

Example 1

Suppose you are interested in the percentage of pages in Good Housekeeping magazine that contain at least one advertisement.You might take an SRS of 10 issues from the most recent 60 issues of the magazine and for each issue measure the following:

$$x_i = \text{total number of pages in issue } i,$$

$$y_i = \text{total number of pages in issue } i \text{ that contain at least one advertisement.}$$

The percentage could be estimated as $\hat{B} = \frac{\hat{t}_y}{\hat{t}_x}$.

Why we use ratio estimation

- *We want to estimate a population total, but the population size N is unknown*

$$N = \frac{t_x}{\bar{x}_U} \implies \hat{N} = \frac{t_x}{\bar{x}_S}.$$

Example 2

To estimate the total number of fish in a haul that are longer than 12 cm, you could take a random sample of fish, estimate the proportion that are larger than 12 cm, and multiply that proportion by the total number of fish, N. Such a procedure cannot be used if N is unknown.You can, however, weigh the total haul of fish, and use the fact that having a length of more than 12 cm y is related with weight x:

$$\hat{t}_{yr} = \bar{y}_S \cdot \frac{t_x}{\bar{x}_S}.$$

Why we use ratio estimation

- Ratio estimation is used to increase the precision of estimated means and totals.

Example 3

Other option of estimating the total population of France by multiplying the average number of persons in the 30 communes by the total number of communes in France.

He reasoned that the ratio estimator would attain more precision: on average, the larger the population of a commune, the higher the number of registered births.

Why we use ratio estimation

- Ratio estimation may be used to adjust for nonresponse.

Example 4

Suppose a sample of businesses is taken: let y_i be the amount spent on health insurance by business i and x_i be the number of employees in business i. Some businesses would not respond to the survey. Suppose the total number of employees t_x is known. To estimate the total amount spent on health insurance, we use

$$\hat{t}_y = t_x \cdot \frac{\bar{y}_S}{\bar{x}_S}.$$

Bias and MSE of ratio estimators

- For SRS, the sample mean is unbiased in the sense that

$$\mathbf{E}[\bar{y}_S] = \bar{y}_U.$$

Incorrect: should be reciprocal

- The ratio estimator is usually biased

$$\hat{\bar{y}}_r = \frac{\bar{y}_S}{\bar{x}_S} \cdot \bar{x}_U = \frac{\bar{x}_S}{\bar{x}_U} \cdot \bar{y}_S.$$

$$\mathbf{E}[\hat{\bar{y}}_r] = \mathbf{E}[\frac{\bar{x}_S}{\bar{x}_U} \cdot \bar{y}_S] \neq \bar{y}_U.$$

Incorrect: should be reciprocal

Bias and MSE of ratio estimators

- *The bias*:

$$\hat{\bar{y}}_r - \bar{y}_U = \frac{\bar{y}_S}{\bar{x}_S} \cdot \bar{x}_U - \bar{y}_U = \bar{y}_S \left(1 - \frac{\bar{x}_S - \bar{x}_U}{\bar{x}_S}\right) - \bar{y}_U.$$

$$\begin{aligned} \text{Bias}(\hat{\bar{y}}_r) &= \mathbf{E}[\hat{\bar{y}}_r - \bar{y}_U], \\ &= \mathbf{E}[\bar{y}_S] - \bar{y}_U - \mathbf{E}[\frac{\bar{y}_S}{\bar{x}_S}(\bar{x}_S - \bar{x}_U)] \\ &= -\mathbf{E}[\hat{B}(\bar{x}_S - \bar{x}_U)] \\ &= -\mathrm{Cov}(\hat{B}, \bar{x}_S). \end{aligned}$$

Bias and MSE of ratio estimators

- An approximation:

$$\text{Bias}(\hat{\bar{y}}_r) \approx \left(1 - \frac{n}{N}\right) \frac{1}{n \cdot \bar{x}_U} (B \cdot V_x - R\sqrt{V_x \cdot V_y})$$

To make it small:

- sample size n is large
- fraction $\frac{n}{N}$ is large
- $\bar{x}_U$ is large
- V_x is small
- R is close to one and $y_i \approx a \cdot x_i$.

Bias and MSE of ratio estimators

- *The MSE*:

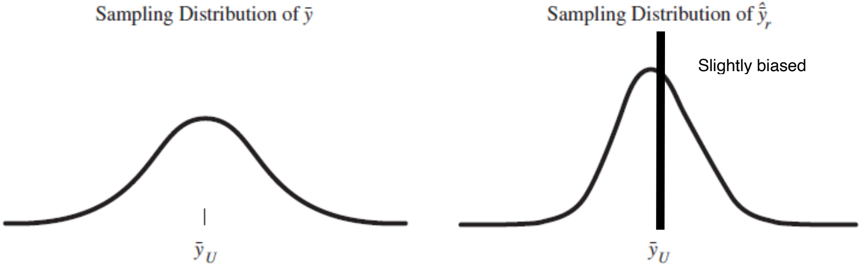
$$\mathrm{MSE}(\hat{\bar{y}}_r) = \mathbf{E}[(\hat{\bar{y}}_r - \bar{y}_U)^2]$$

$$\approx \mathbf{E}[(\bar{y}_S - B \cdot \bar{x}_S)^2] = \left(1 - \frac{n}{N}\right) \cdot \frac{V_y - 2B \cdot R \cdot \sqrt{V_x \cdot V_y} + B^2 V_x}{n}$$

To make it small:

- sample size n is large
- fraction $\frac{n}{N}$ is large
- R is close to 1 and $y_i \approx a \cdot x_i$.

Illustrative graph



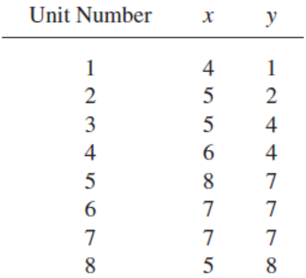
Sampling Distribution of $\bar{y}$

Sampling Distribution of $\hat{\bar{y}}_r$

Figure 1: Left: simple random sample estimate; Right: ratio estimator.

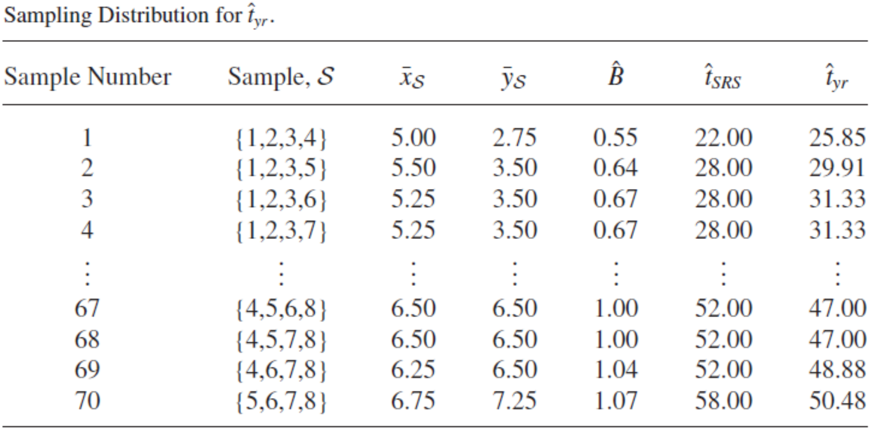
An example

Example 4.4 from the textbook:



Unit Number	x	y
1	4	1
2	5	2
3	5	4
4	6	4
5	8	7
6	7	7
7	7	7
8	5	8

An example (cont.)

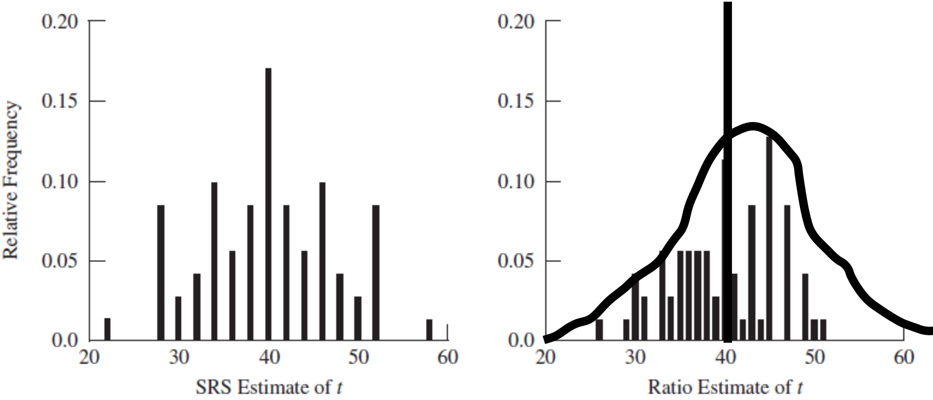


Sampling Distribution for $\hat{t}_{yr}$.

Sample Number	Sample, $\mathcal{S}$	$\bar{x}_{\mathcal{S}}$	$\bar{y}_{\mathcal{S}}$	$\hat{B}$	$\hat{t}_{SRS}$	$\hat{t}_{yr}$
1	{1,2,3,4}	5.00	2.75	0.55	22.00	25.85
2	{1,2,3,5}	5.50	3.50	0.64	28.00	29.91
3	{1,2,3,6}	5.25	3.50	0.67	28.00	31.33
4	{1,2,3,7}	5.25	3.50	0.67	28.00	31.33
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
67	{4,5,6,8}	6.50	6.50	1.00	52.00	47.00
68	{4,5,7,8}	6.50	6.50	1.00	52.00	47.00
69	{4,6,7,8}	6.25	6.50	1.04	52.00	48.88
70	{5,6,7,8}	6.75	7.25	1.07	58.00	50.48

An example (cont.)

$$t_x = 47 \qquad t_y = 40$$
$$S_x = 1.3562027 \qquad S_y = 2.618615$$
$$R = 0.6838403 \qquad B = 0.8510638$$


Relative Frequency
SRS Estimate of *t*
Ratio Estimate of *t*