

We had

Stat 576
4-17-25

$$V(\bar{y}_{reg}) = \frac{1}{n} \left(1 - \frac{1}{N}\right) S_y^2 (1 - R^2) \quad (1)$$

when $B = \frac{S_{xy}}{S_x^2}$

$$\hat{V}(\bar{y}_{reg}) = \frac{1}{n} \left(1 - \frac{1}{N}\right) S_y^2 (1 - r^2) \frac{n-1}{n-2}$$

↑ correction factor
to make an
unbiased
estimator

Estimator	Biased?	Variance
① $\bar{y}$	no	$\frac{S_y^2}{n} \cdot fpc$
② $\bar{y}_{tot}$	yes	$\frac{S_e^2}{n} \cdot fpc$ *
③ $\bar{y}_{dft}$	no	$\frac{S_y^2 + S_x^2 - 2S_{xy}}{n} \cdot fpc$
④ $\bar{y}_{reg}$	no	$\frac{S_y^2 (1 - R^2)}{n} \cdot fpc$

$$* S_e^2 = S_y^2 + B^2 S_x^2 - 2BS_{xy}, B = \bar{y}/\bar{x}$$

(3)

② vs ① : Already did this, but let's look at another approach

② wins if $s_e^2 < s_y^2$

$$s_y^2 + B^2 s_x^2 - 2B s_{xy}$$

$$B^2 s_x^2 < 2B s_{xy}$$

Case 1: $B > 0$ $B < 2 \frac{s_{xy}}{s_x^2}$

$$\frac{\bar{y}}{\bar{x}} < 2 \frac{s_{xy}}{s_x s_y} \frac{s_y}{s_x}$$

(4)

$$\frac{s_{xy}}{s_x s_y} > \frac{1}{2} \frac{s_x}{\bar{x}} / \frac{s_y}{\bar{y}}$$

$$R > \frac{1}{2} \frac{CV_x}{CV_y}$$

So, if $CV_x \approx CV_y$, then ② beats ①

if $B > 0$ & $R > .5$

Case 2: $B < 0$ Similarly, ② beats ① if $R < -.5$

Summary of ② vs ① : ② wins if

$$CV_x \approx CV_y \text{ \& } |R| > .5$$

③ vs ①

⑤

③ wins if $S_y^2 + S_x^2 - 2S_{xy} < S_y^2$

$$S_x^2 < 2S_{xy}$$

$$\frac{S_{xy}}{S_x^2} > \frac{1}{2}$$

④ vs ①

④ wins if $R^2 \neq 0$

③ vs ②

⑥

Compare $S_y^2 + S_x^2 - 2S_{xy}$ with

$$S_y^2 + B^2 S_x^2 - 2BS_{xy}$$

② wins if $S_y^2 + B^2 S_x^2 - 2BS_{xy} < S_y^2 + S_x^2 - 2S_{xy}$

$$(B-1)S_x^2 < 2(B-1)S_{xy}$$

$$\frac{(B-1)(B+1)}{2} < (B-1) \frac{S_{xy}}{\frac{S_x^2}{2}}$$

Case 1: $B > 1$ $\frac{B+1}{2} < \frac{S_{xy}}{S_x^2}$ (7)

Case 2: $B < 1$ $\frac{B+1}{2} > \frac{S_{xy}}{S_x^2}$

(2) beats (3) if $1 < B < 2 \frac{S_{xy}}{S_x^2} - 1$

or if $2 \frac{S_{xy}}{S_x^2} - 1 < B < 1$

(4) vs (2): (4) wins, but is equivalent to (2)
if the L.S. line passes through
the origin

(4) vs (3): (4) wins

(8)

Post-stratification

We find out that we separate estimates
of the population mean of 2 subpopulations

All we have is a SRSWOR from the
combined population

Let $\bar{y}_1$ be the sample mean for the 1st group.

① Is $E(\bar{y}_1) = \bar{Y}_1$?

⑨

② Is $V(\bar{y}_1) = \frac{S_1^2}{n_1} (1 - \frac{n_1}{N_1})$?

$\bar{y}_1 = \frac{\sum_{\text{group 1}} y_i}{n_1}$ $\swarrow \searrow$ both are random variables

Let $u_i = \begin{cases} y_i & \text{if the person is from group 1} \\ 0 & \text{otherwise} \end{cases}$

Let $k_i = \begin{cases} 1 & \text{if the person is from group 1} \\ 0 & \text{otherwise} \end{cases}$

Then $\bar{u} = \frac{\sum_{i=1}^n u_i}{n} = \frac{\sum_{\text{gp 1}} y_i}{n}$

⑩

And $\bar{k} = \frac{\sum_{i=1}^n k_i}{n} = \frac{n_1}{n}$

So $\frac{\bar{u}}{\bar{k}} = \frac{\sum_{\text{gp 1}} y_i}{n_1} = \bar{y}_1$

We know that $\frac{\bar{u}}{\bar{k}}$ is a ratio estimator of $\frac{\bar{U}}{\bar{X}} = \bar{Y}_1$

(11)

So $\bar{y}_1$ is a biased estimator of $\bar{Y}_1$,

but is asymptotically unbiased

$$\text{Also } \hat{V}(\bar{y}_1) = \frac{1}{N^2} \frac{S_e^2}{n} \left(1 - \frac{n}{N}\right)$$

$$S_e^2 = S_u^2 + \hat{B}^2 S_x^2 - 2\hat{B}S_{xy}$$

$$\hat{B} = \frac{\bar{u}}{\bar{x}} = \bar{y}_1$$

$$S_u^2 = \frac{\sum_{i=1}^n u_i^2 - \frac{(\sum u_i)^2}{n}}{n-1} = \frac{\sum_{i=1}^n y_i^2 - \frac{(\sum y_i)^2}{n}}{n-1}$$

$$S_u^2 = \frac{\sum_1 y_i^2 - \frac{(n_1 \bar{y}_1)^2}{n}}{n-1}$$

(12)

$$S_x^2 = \frac{\sum_{i=1}^n x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1}$$

$$= \frac{n_1 - \frac{n_1^2}{n}}{n-1}$$

$$S_{xy} = \frac{\sum_{i=1}^n x_i u_i - \frac{(\sum x_i)(\sum u_i)}{n}}{n-1}$$

$$= \frac{n_1 \bar{y}_1 - \frac{n_1 n_1 \bar{y}_1}{n}}{n-1}$$

(13)

Now

$$S_e^2 = \frac{1}{n-1} \left[\sum_i y_i^2 - \cancel{\frac{(n_1 \bar{y}_1)^2}{n}} + \bar{y}_1^2 \left(n_1 - \cancel{\frac{n_1^2}{n}} \right) - 2 \bar{y}_1 \left(n_1 \bar{y}_1 - \cancel{n_1^2 \frac{\bar{y}_1}{n}} \right) \right]$$

$$= \frac{1}{n-1} \left[\sum_i y_i^2 + n_1 \bar{y}_1^2 - 2 n_1 \bar{y}_1^2 \right]$$

$$= \frac{1}{n-1} \left[\sum_i y_i^2 - n_1 \bar{y}_1^2 \right] = \frac{1}{n-1} (n_1 - 1) S_1^2$$

$$\text{So } \hat{V}(\bar{y}_1) = \frac{1}{\bar{x}^2} \frac{1}{n-1} (n_1 - 1) S_1^2 \frac{1}{n} \left(1 - \frac{n_1}{N} \right) \quad (14)$$

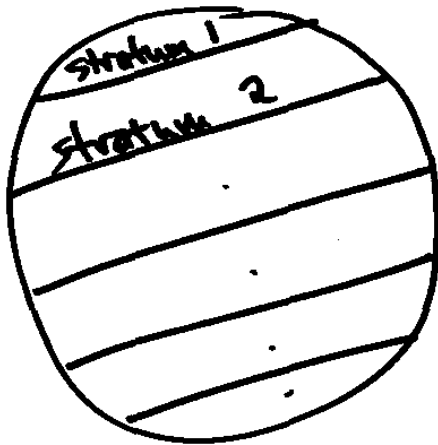
$$\text{Also } \bar{x} = \frac{n_1}{n}$$

$$\hat{V}(\bar{y}_1) = \frac{n^2}{n_1^2} \frac{1}{n-1} (n_1 - 1) \underbrace{\frac{S_1^2}{n_1} \left(1 - \frac{n_1}{N_1} \right)}_{\left(1 - \frac{n_1}{N_1} \right)} \frac{n_1}{\left(1 - \frac{n_1}{N_1} \right)} \frac{1}{n} \left(1 - \frac{n_1}{N} \right)$$

$$= \underbrace{\frac{n}{n-1} \frac{(n_1 - 1)}{n_1} \frac{1 - \frac{n_1}{N}}{1 - \frac{n_1}{N_1}}}_{\neq 1, \text{ but } \approx 1} \frac{S_1^2}{n_1} \left(1 - \frac{n_1}{N_1} \right)$$

$\neq 1, \text{ but } \approx 1$

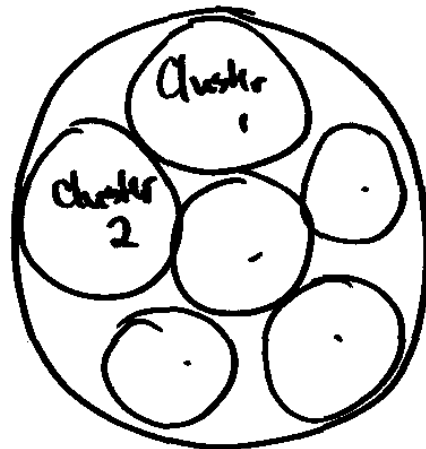
Heterogeneity between,
homogeneity within
Stratified sampling



Collect a SRSWOR
within each stratum,
independently

Homogeneity between,
heterogeneity within
Cluster Sampling

(15)



Collect a SRSWOR of
clusters, + keep every item
in those clusters

Stat 576 HW3

- 3 Foresters want to estimate the average age of trees in a stand. Determining age is cumbersome, because one needs to count the tree rings on a core taken from the tree. In general, though, the older the tree, the larger the diameter, and diameter is easy to measure. The foresters measure the diameter of all 1132 trees and find that the population mean equals 10.3. They then randomly select 20 trees for age measurement.

Tree No.	Diameter, x	Age, y	Tree No.	Diameter, x	Age, y
1	12.0	125	11	5.7	61
2	11.4	119	12	8.0	80
3	7.9	83	13	10.3	114
4	9.0	85	14	12.0	147
5	10.5	99	15	9.2	122
6	7.9	117	16	8.5	106
7	7.3	69	17	7.0	82
8	10.2	133	18	10.7	88
9	11.7	154	19	9.3	97
10	11.3	168	20	8.2	99

Estimate the population mean age of trees in the stand and give an approximate standard error for your estimate, using:

- a) the sample mean
- b) the ratio estimator
- c) the difference estimator
- d) the regression estimator