Stat 576
10-12-17

(1)

We had

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

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

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

↑ correction factor to make an unbiased estimator

(2)

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

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

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

2 wins if $S_e^2 < S_Y^2$

$$\parallel$$

$$S_Y^2 + B^2 S_X^2 - 2B S_{XY}$$

$$B^2 S_X^2 < 2B S_{XY}$$

Case 1: $B > 0$ $\quad B < 2 \frac{S_{XY}}{S_X^2}$

$$\frac{\bar{Y}}{\bar{X}} < 2 \frac{S_{XY}}{S_X S_Y} \quad \frac{S_Y}{S_X}$$

$$\frac{S_{XY}}{S_X S_Y} > \frac{1}{2} \frac{S_X}{\bar{X}} \Big/ \frac{S_Y}{\bar{Y}}$$

$$R > \frac{1}{2} \frac{CV_X}{CV_Y}$$

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 $CV_X \approx CV_Y$ & $|R| > .5$

(5)

(3) vs (1)

(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 $R^2 \neq 0$

(6)

(3) vs (2)

Compare $s_Y^2 + s_X^2 - 2s_{XY}$ with

$$s_Y^2 + B^2 s_X^2 - 2Bs_{XY}$$

(2) wins if $\cancel{s_Y^2} + B^2 s_X^2 - 2Bs_{XY} < \cancel{s_Y^2} + s_X^2 - 2s_{XY}$

$$(B^2-1)s_X^2 < 2(B-1)s_{XY}$$

$$\frac{(B-1)(B+1)}{2} < (B-1)\frac{s_{XY}}{s_X^2}$$

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

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

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

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.

9

① Is $E(\bar{y}_1) = \bar{Y}_1$?

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

$$\bar{y}_1 = \frac{\sum_{\text{group 1}} y_i}{n_1}$$ 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 $x_i = \begin{cases} 1 & \text{if the person is from group 1} \\ 0 & \text{otherwise} \end{cases}$

10

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

And $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{n_1}{n}$

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

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

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

but is asymptotically unbiased

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

$$s_e^2 = s_u^2 + \hat{B}^2 s_x^2 - 2\hat{B} s_{xy}$$

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

$$s_u^2 = \frac{\sum_{i=1}^{n} u_i^2 - \frac{(\sum u_i)^2}{n}}{n-1} = \frac{\sum_{sp1} y_i^2 - \frac{(\sum_{sp1} y_i)^2}{n}}{n-1}$$

12

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

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

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

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

$$= \frac{n_1\bar{y}_1 - \frac{n_1 n_1 \bar{y}_1}{n}}{n-1}$$ (13)

Now

$$s_e^2 = \frac{1}{n-1}\Big[\sum_1 y_i^2 - \cancel{\frac{(n_1\bar{y}_1)^2}{n}} + \bar{y}_1^2\Big(n_1 - \cancel{\frac{n_1^2}{n}}\Big)$$
$$- 2\bar{y}_1\Big(n_1\bar{y}_1 - \cancel{n_1^2\frac{\bar{y}_1}{n}}\Big)\Big]$$

$$= \frac{1}{n-1}\Big[\sum_1 y_i^2 + n_1\bar{y}_1^2 - 2n_1\bar{y}_1^2\Big]$$

$$= \frac{1}{n-1}\Big[\sum_1 y_i^2 - n_1\bar{y}_1^2\Big] = \frac{1}{n-1}(n_1-1)s_1^2$$

So $\hat{V}(\bar{y}_1) = \frac{1}{\bar{x}^2}\,\frac{1}{n-1}(n_1-1)s_1^2\,\frac{1}{n}\Big(1-\frac{n}{N}\Big)$ (14)

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

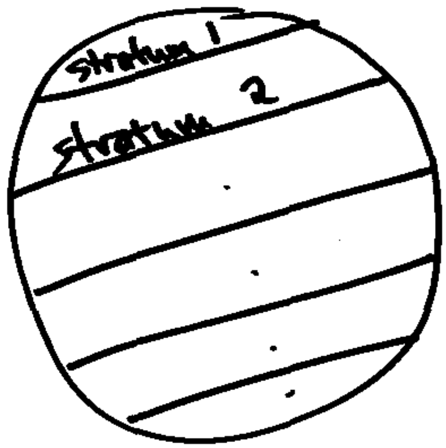
$$\hat{V}(\bar{y}_1) = \frac{n^2}{n_1^2}\,\frac{1}{n-1}(n_1-1)\underbrace{\frac{s_1^2}{n_1}\Big(1-\frac{n_1}{N_1}\Big)}\,\frac{n_1}{\Big(1-\frac{n_1}{N_1}\Big)}\,\frac{1}{n}\Big(1-\frac{n}{N}\Big)$$

$$= \underbrace{\frac{n}{n-1}\,\frac{(n_1-1)}{n_1}\,\frac{1-\frac{n}{N}}{1-\frac{n_1}{N_1}}}_{\neq 1,\text{ but } \approx 1}\,\frac{s_1^2}{n_1}\Big(1-\frac{n_1}{N_1}\Big)$$

15

Heterogeneity between,
Homogeneity within

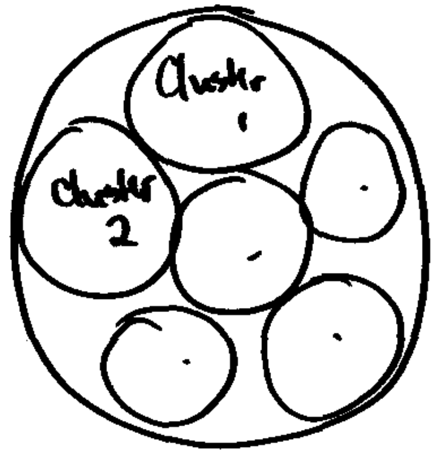
Stratified sampling



Collect a SRSWOR within each stratum, independently

Homogeneity between
Heterogeneity within

Cluster Sampling



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

HW #3

Use data on p.156 #3

Answer the questions, using:

1. $\bar{y}$
2. $\bar{y}_{rat}$
3. $\bar{y}_{diff}$
4. $\bar{y}_{reg}$

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

a Draw a scatterplot of y vs. x.

b Estimate the population mean age of trees in the stand using ratio estimation and give an approximate standard error for your estimate.

c Repeat (b) using regression estimation.

d Label your estimates on your graph. How do they compare?