17.2.16

STK4600

Solutions to Ex# 2

1. Let $u_i = x_i + y_i$, $i=1,\dots N$

$\Rightarrow \bar{u}_s = \bar{x}_s + \bar{y}_s$, $\bar{u} = \mu_x + \mu_y$

$$(*)\quad Var(\bar{u}_s) = Var(\bar{x}_s) + Var(\bar{y}_s) + 2cov(\bar{x}_s, \bar{y}_s)$$

and also:

$$Var(\bar{u}_s) = (1-f)\frac{1}{n}\cdot\frac{1}{N-1}\sum_{i=1}^{N}(u_i-\bar{u})^2$$

where $\frac{1}{N-1}\sum_{i=1}^{N}(u_i-\bar{u})^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\mu_x+y_i-\mu_y)^2$

$$= \frac{1}{N-1}\left\{\sum_1^N (x_i-\mu_x)^2 + \sum_1^N (y_i-\mu_y)^2 + 2\sum_{i=1}^{N}(x_i-\mu_x)(y_i-\mu_y)\right\}$$

$$= \sigma_x^2 + \sigma_y^2 + 2\sigma_{xy}$$

From $(*)$:

$$(1-f)\frac{1}{n}\sigma_x^2 + (1-f)\frac{1}{n}\sigma_y^2 + 2cov(\bar{x}_s, \bar{y}_s)$$

$$= (1-f)\frac{1}{n}\sigma_x^2 + (1-f)\frac{1}{n}\sigma_y^2 + \frac{1-f}{n}2\sigma_{xy}$$

$$\Rightarrow cov(\bar{x}_s, \bar{y}_s) = \frac{1-f}{n}\sigma_{xy}$$

2. $\hat{R} \approx R + \frac{\bar{y}_s - R\bar{x}_s}{\mu_x}$, $\hat{t}_R = X_0 \cdot \hat{R}$

$$\Rightarrow E\hat{t}_R = X_0 E(\hat{R}) \approx X_0 R + \frac{1}{\mu_x}\{E(\bar{y}_s) - R\,E(\bar{x}_s)\}$$

$$= t + \frac{1}{\mu_x}(\mu_y - \underbrace{R\cdot\mu_x}_{\mu_y}) = t$$

$Var(\hat{t}_R)$

$$\approx X_0^2 Var\left(\frac{\bar{y}_s - R\bar{x}_s}{\mu_x}\right) = \frac{X_0^2}{(X_0/N)^2} Var(\bar{y}_s - R\bar{x}_s)$$

$$= N^2\left(Var(\bar{y}_s) + R^2 Var(\bar{x}_s) - 2R\,cov(\bar{y}_s, \bar{x}_s)\right)$$

$$= N^2 \cdot \frac{1-f}{n}(\sigma_y^2 + R^2\sigma_x^2 - 2R\sigma_{xy})$$

Second equality:

Show that:

$$\sigma_y^2 + R^2\sigma_x^2 - 2R\sigma_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - Rx_i)^2$$

We see that:

$$\frac{1}{N-1}\sum_{i=1}^{N}(y_i - Rx_i)^2$$

$$= \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu_y + \overbrace{R\mu_x}^{\mu_y} - Rx_i)^2$$

$$= \frac{1}{N-1}\sum_{i=1}^{N}((y_i - \mu_y) - R(x_i - \mu_x))^2$$

$$= \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu_y)^2 + R^2\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)^2$$

$$- 2R\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu_y)(x_i - \mu_x)^2$$

$$= \sigma_y^2 + R^2\sigma_x^2 - 2R\sigma_{xy}.$$

3. y = household income
x = number of adults in household

a) Sampling design: $p(\{i,j\}) = \frac{1}{6}, \ \forall (i<j)$

Sampling distribution of $\hat{t}_e$ and $\hat{t}_e$:

s	(1,2)	(1,3)	(1,4)	(2,3)	(2,4)	(3,4)
$p(s)$	1/6	1/6	1/6	1/6	1/6	1/6
$\hat{t}_e$	1120	1820	2940	2100	3220	3920
$\hat{t}_R$	1960	2123.3	2572.5	2450	2817.5	2744

$\Rightarrow E(\hat{t}_e) = \sum \hat{t}_e(s) \cdot \frac{1}{6} = \frac{1}{6} \sum_s \hat{t}_e(s) = \frac{15120}{6} = \underline{2520 = t}$ (we know this, of course)

$Var(\hat{t}_e) = E(\hat{t}_e - t)^2 = \sum_s (\hat{t}_e(s) - 2520)^2 \frac{1}{6} = \underline{875466.67}$

From general formula:

$Var(\hat{t}_e) = N^2 Var(\bar{y}_s) = N^2 \cdot \frac{\sigma_y^2}{n}(1 - \frac{n}{N}) = 4\sigma_y^2$

and $\sigma_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu_y)^2 = \frac{1}{3}\sum_{i=1}^{4}(y_i - 630)^2 = 218\,866.67$

$\Rightarrow Var(\hat{t}_e) = 4 \times 218866.67 = 875466.67$

Ratio-estimator:

$E(\hat{t}_R) = \sum_s \hat{t}_R(s) \cdot \frac{1}{6} = \frac{1}{6}\sum_s \hat{t}_R(s) = \frac{14667.3}{6} = \underline{2444.5}$

$\hat{t}_R$ has a small bias: slightly underestimates t (on the average)

$Var(\hat{t}_R) = \sum_s (\hat{t}_R(s) - 2444.5)^2 \frac{1}{6} = \underline{97192.20}$

Comparison: "True" $SE(\hat{t}_e) = \sqrt{875466.67} = 935.67$

"True" $SE(\hat{t}_R) = \sqrt{97192.20} = 311.76$

$\Rightarrow \hat{t}_R$ is chosen, even if it is biased

Note: $MSE(\hat{t}_R) = 75.5^2 + 97192.20 = 102892.45 \Rightarrow \sqrt{MSE(\hat{t}_R)} = \underline{320.77}$!

b) $\mathrm{Var}(\hat{t}_R) \approx N^2 \frac{1-f}{n} \frac{1}{3} \sum_{i=1}^{4} (y_i - \hat{R}x_i)^2$

$= 4 \cdot \frac{1}{3} \sum_{i=1}^{4} (y_i - 360x_i)^2$

$= 4 \cdot 18466{,}67 = \underline{73866.7}$

compared to exact value $= 9719.20$; not accurate

but $SE(\hat{t}_R)$ is "better" approximated

$\sqrt{97192{,}20} = 311.76$ and $\sqrt{73866.7} = 271.8$

Typical for small n: Approximate $\mathrm{Var}(\hat{t}_R)$ is too small

c) CI's:

Based on $\hat{t}_e$: $\hat{t}_e \pm 1.96 \cdot SE(\hat{t}_e)$

where $SE(\hat{t}_e) = \sqrt{\hat{V}(\hat{t}_e)} = N \cdot \sqrt{\frac{s^2}{n}(1-f)}$

$s^2 = \frac{1}{n-1} \sum_{i \in s} (y_i - \bar{y}_s)^2 = \sum_{i \in s} (y_i - \bar{y}_s)^2$

Based on $\hat{t}_R$: $\hat{t}_R \pm 1.96 \cdot SE(\hat{t}_R)$

where $SE(\hat{t}_R) = \frac{\mu_x}{\bar{x}_s} \cdot N \cdot \sqrt{\frac{1-f}{n} \sum_{i \in s} (y_i - \hat{R}x_i)^2}$; $\hat{R} = \sum_s y_i / \sum_s x_i$

s	$\hat{t}_e \pm 1.96 SE(\hat{t}_e)$	Cover t?	$\hat{t}_R \pm 1.96 \cdot SE(\hat{t}_R)$	Cover t?
(1,2)	1120 ± 388	no	1960 ± 679	yes
(1,3)	1820 ± 1358	yes	2123.3 ± 603	yes
(1,4)	2940 ± 2910	yes	2577.5 ± 764	yes
(2,3)	2100 ± 970	yes	2450 ± 0 (!)	no
(2,4)	3220 ± 2522	yes	2817.5 ± ~~265~~ 255	~~yes~~ no
(3,4)	~~3~~3920 ± 15~~5~~52	yes	2744 ± 326	yes

$\hat{t}_e$: $P(CI \ni t) = 5/6 = .83$, $\hat{t}_R$: $P(CI \ni t) = 4/6 = .67$

but still: CI based on $\hat{t}_R$ is much more accurate and reliable

d) Based on $\hat{t}_e$: E(length) $= \sum$ lengths $\frac{1}{6} = \frac{1}{6} \times \cancel{5254}\ \frac{19400}{6} = \underline{3233}$!

The CI is of no value!

Based on $\hat{t}_R$: E(length) $= \frac{1}{6} \times 5254 = \underline{875.7}$