STA 304H1F-1003H Fall 2019
Assignment 2-Question 4-Solution

Question 4. (5 marks)

Consider a stratified design composed of H strata of size N_h, $h = 1, \ldots, H$. We want to estimate the population mean μ_y of the characteristic y. Let $\mu_{x,h}$, $h = 1, \ldots, H$ be the means in the strata (in the population) of an auxiliary characteristic x. The $\mu_{x,h}$ are supposedly known and we propose to estimate μ_y using the following estimator:

$$\widehat{\mu}_D = \overline{y}_{st} + \mu_x - \overline{x}_{st}$$

where $\overline{y}_{st}$ and $\overline{x}_{st}$ are the basic estimate of the population means μ_y and μ_x for y and x, respectively.

(a) (1 mark) Give an expression of μ_x in terms of $\mu_{x,h}$, $h = 1, \ldots, H$.

$$\mu_x = \sum_{h=1}^{H} W_h \times \mu_{x,h} = \sum_{h=1}^{H} \frac{N_h}{N} \times \mu_{x,h}, \quad \text{where} \quad N = \sum_{h=1}^{H} N_h$$

(1mk)

(b) (1 mark) Show that $\widehat{\mu}_D$ is unbiased estimator for μ_y.

We have that:

$\overline{y}_{st} = \sum_{h=1}^{H} \frac{N_h}{N} \times \overline{y}_h$ is unbiased estimator for μ_y, e.g.

$$\mathbf{E}(\overline{y}_{st}) = \mathbf{E}(\sum_{h=1}^{H} \frac{N_h}{N} \times \overline{y}_h) = \sum_{h=1}^{H} \frac{N_h}{N} \times \mathbf{E}(\overline{y}_h) = \sum_{h=1}^{H} \frac{N_h}{N} \times \mu_{y,h} = \mu_y,$$

and

$\overline{x}_{st} = \sum_{h=1}^{H} \frac{N_h}{N} \times \overline{x}_h$ is unbiased estimator for μ_x, e.g.

$$\mathbf{E}(\overline{x}_{st}) = \mathbf{E}(\sum_{h=1}^{H} \frac{N_h}{N} \times \overline{x}_h) = \sum_{h=1}^{H} \frac{N_h}{N} \times \mathbf{E}(\overline{x}_h) = \sum_{h=1}^{H} \frac{N_h}{N} \times \mu_{x,h} = \mu_x.$$

Therefore, the expected value of $\widehat{\mu}_D$ is

$$\mathbf{E}(\widehat{\mu}_D) = \mathbf{E}(\overline{y}_{st} + \mu_x - \overline{x}_{st}) = \mathbf{E}(\overline{y}_{st}) + \mu_x - \mathbf{E}(\overline{x}_{st}) = \mu_y + \mu_x - \mu_x = \boxed{\mu_y}$$

which means that $\widehat{\mu}_D$ is unbiased estimator for μ_y

(1mk)

(c) (1 mark) Give the variance of $\widehat{\mu}_D$

Let $d_{h,i} = y_{h,i} - x_{h,i}$ be the diffence between y and x for each observation i in strata h.

We have that the population mean for d is the different between the population mean for y and for x in the stata h, e.g.

$$\mu_{d,h} = \frac{1}{N_h}\sum_{i=1}^{N_h} d_{h,i} = \frac{1}{N_h}\sum_{i=1}^{N_h} y_{h,i} - \frac{1}{N_h}\sum_{i=1}^{n_h} x_{h,i} = \mu_{y,h} - \mu_{x,h}$$

Similary, we can show that

$$\mu_d = \mu_y - \mu_x$$

The same result holds for the sample mean

$$\overline{d}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} d_{h,i} = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{h,i} - \frac{1}{n_h}\sum_{i=1}^{n_h} x_{h,i} = \overline{y}_h - \overline{x}_h$$

We can show that

$$\overline{d}_{st} = \sum_{h=1}^{H} W_h \overline{d}_h = \sum_{h=1}^{H} W_h(\overline{y}_h - \overline{x}_h) = \sum_{h=1}^{H} W_h \overline{y}_h - \sum_{h=1}^{H} W_h \overline{x}_h = \boxed{\overline{y}_{st} - \overline{x}_{st}}$$

Therefore the variance of $\widehat{\mu}_D$ can be writen has

$$\mathbf{V}(\widehat{\mu}_D) = \mathbf{V}(\overline{y}_{st} + \mu_x - \overline{x}_{st}) = \mathbf{V}(\overline{y}_{st} - \overline{x}_{st} + \mu_x) = \mathbf{V}(\overline{d}_{st} + \mu_x) = \mathbf{V}(\overline{d}_{st})$$

which is equal to

$$\mathbf{V}(\widehat{\mu}_D) = \mathbf{V}(\overline{d}_{st}) = \boxed{\sum_{h=1}^{H} W_h^2 \times \frac{N_h - n_h}{N_h - 1} \times \frac{\sigma^2_{d,h}}{n_h} = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \times \frac{N_h - n_h}{N_h - 1} \times \frac{\sigma^2_{d,h}}{n_h}}$$

where

$$\sigma^2_{d,h} = \frac{\sum_{i=1}^{N_h}(d_{h,i} - \mu_{d,h})^2}{N_h}$$

denote the population variance for d in strata h.

Note that $\sigma^2_{d,h}$ can be written as

$$\sigma^2_{d,h} = \frac{\sum_{i=1}^{N_h}(d_{h,i} - \mu_{d,h})^2}{N_h} = \frac{\sum_{i=1}^{N_h}\left[(y_{h,i} - \mu_{y,h})^2 - 2(y_{h,i} - \mu_{y,h})(x_{h,i} - \mu_{x,h}) + (x_{h,i} - \mu_{x,h})^2\right]}{N_h} = \sigma^2_{y,h} - 2\sigma^2_{yx,h} + \sigma^2_{x,h}$$

where

$$\sigma^2_{y,h} = \frac{\sum_{i=1}^{N_h}(y_{h,i} - \mu_{y,h})^2}{N_h}, \quad \sigma^2_{yx,h} = \frac{\sum_{i=1}^{N_h}(y_{h,i} - \mu_{y,h})(x_{h,i} - \mu_{x,h})}{N_h}, \quad \sigma^2_{x,h} = \frac{\sum_{i=1}^{N_h}(x_{h,i} - \mu_{x,h})^2}{N_h}$$

(1mk)

The estimated variance of $\widehat{\mu}_D$(**expression not accepted as response in this part**) is

$$\widehat{\mathbf{V}}(\widehat{\mu}_D) = \boxed{\sum_{h=1}^{H} W_h^2 \times (1 - \frac{n_h}{N_h}) \times \frac{s^2_{d,h}}{n_h} = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \times (1 - \frac{n_h}{N_h}) \times \frac{s^2_{d,h}}{n_h}}$$

where $s^2_{d,h} = \dfrac{\sum_{i=1}^{n_h}(d_{h,i} - \overline{d}_{d,h})^2}{n_h - 1} = s^2_{y,h} - 2s^2_{yx,h} + s^2_{x,h}$, denote the sample variance for d in the h^{th} strata.

(d) (1 mark) Let $n = \sum_{h=1}^{H} n_h$ be the sample size. What is the optimal allocation of the n_h in order to minimise the variance of $\widehat{\mu}_D$? We consider that the init cost of the survey does not depend on the stratum.

Since the unit cost is the same in all of the strata, the optimal allocation is the Neyman allocation which minimizes the the variance of $\widehat{\mu}_D$. It is given by:

$$n_h = a_h \times n = \frac{N_h \times \sigma_{d,h}}{\sum_{h=1}^{H} N_h \times \sigma_{d,h}} \times n,$$

where $\sigma_{d,h} = \sqrt{\sigma_{d,h}^2}$ is the population standard error for d in the ith stratum.

(1mk)

(e) (1 mark) In which favourable case is $\widehat{\mu}_D$ preferable to $\overline{y}_{st}$?

From part a), we have that both $\widehat{\mu}_D$ and $\overline{y}_{st}$ are unbiased estimators for μ_y.

As the variance of $\widehat{\mu}_D$ is

$$\mathbf{V}(\widehat{\mu}_D) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \times \frac{N_h - n_h}{N_h - 1} \times \frac{\sigma_{d,h}^2}{n_h}$$

and the variance of $\overline{y}_{st}$ is

$$\mathbf{V}(\overline{y}_{st}) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \times \frac{N_h - n_h}{N_h - 1} \times \frac{\sigma_{y,h}^2}{n_h}$$

then the estimator $\widehat{\mu}_D$ is preferable to $\overline{y}_{st}$ when for all $h = 1, \ldots, H$ $\sigma_{y,h}^2 > \sigma_{d,h}^2$.

$$\sigma_{y,h}^2 > \sigma_{d,h}^2 \implies \sigma_{y,h}^2 > \sigma_{y,h}^2 - 2\sigma_{yx,h}^2 + \sigma_{x,h}^2 \implies 2\sigma_{yx,h}^2 < \sigma_{x,h}^2 \implies \boxed{\frac{\sigma_{yx,h}^2}{\sigma_{x,h}^2} > \frac{1}{2}}$$

(1mk)

This condition means that $\widehat{\mu}_D$ is preferable to $\overline{y}_{st}$ if the slope $\beta_1 = \dfrac{\sigma_{yx,h}^2}{\sigma_{x,h}^2}$ of the regression of y on x is greater than 0.5.