STAT 521: Homework Assignment 2

Due on March 25, 2025

Problem 1: (20 pt)

A city has a total of 100,000 dwelling units, of which 35,000 are houses, 45,000 are apartments, and 20,000 are condominiums. A stratified sample of size $n = 1000$ is selected using proportional allocation (and rounding the sample sizes to the nearest integer). The three strata are houses (h = 1), apartments (h = 2), and condominiums (h = 3). The table below gives the estimates of the mean energy consumption per dwelling unit for the three strata and the corresponding standard errors.

Stratum	Estimated mean energy consumption	Estimated standard error of the sample mean
(h)	kWh per dwelling unit $(\bar{y}_h)$	$\widehat{SE}(\bar{y}_h)$
House $(h = 1)$	915	4.84
Apartments $(h = 2)$	641	2.98
Condominium $(h = 3)$	712	7.00

1. Estimate the total energy consumption for the full population of 100,000 dwelling units.

2. Estimate the standard error of the estimator used in (1).

3. What would be the sample size if the optimal allocation is to be used (under $n = 1000$) for this population? Assume that the survey costs are the same for each stratum.

   Hint: Use the following steps:

   (a) What is the sample size n_h for each stratum under proportional allocation?

   (b) Note that

   $$\widehat{SE}(\bar{y}_h) = \sqrt{\frac{1}{n_h}\left(1 - \frac{n_h}{N_h}\right) s_h^2}.$$

   Thus, you can obtain s_h^2.

   (c) Apply Neyman allocation (optimal allocation) using s_h in place of S_h.

4. What would be the estimated standard error of the total estimator under the optimal allocation in (3)? Compare it with the answer in (2). Which one is smaller?

Problem 2: (10 pt)

Consider a simple random sample of size $n = 200$ from a finite population with size $N = 10,000$, measuring (X, Y), taking values on $\{(0,0), (0,1), (1,0), (1,1)\}$. The finite population has the following distribution.

	$X = 1$	$X = 0$	
$Y = 1$	N_{11}	N_{10}	N_{1+}
$Y = 0$	N_{01}	N_{00}	N_{0+}
	N_{+1}	N_{+0}	N

The population count N_{ij} are unknown.

Suppose that the realized sample has the following sample counts:

	$X = 1$	$X = 0$	
$Y = 1$	70	30	100
$Y = 0$	50	50	100
	120	80	200

1. If it is known that $N_{+1} = N_{+0} = 5,000$, how can you make use of this information to obtain a post-stratified estimator of $\theta = E(Y)$, using X as the post-stratification variable?

2. If we are interested in estimating $\theta = P(Y = 1 \mid X = 1)$, discuss how to estimate θ from the above sample and how to estimate its variance (Hint: Use Taylor expansion of ratio estimator to obtain the sampling variance.)

Problem 3: (10 pt)

Suppose that we have a finite population of $(Y_{hi}(1), Y_{hi}(0))$ generated from the following superpopulation model

$$\begin{pmatrix} Y_{hi}(0) \\ Y_{hi}(1) \end{pmatrix} \sim \left[\begin{pmatrix} \mu_{h0} \\ \mu_{h1} \end{pmatrix}, \begin{pmatrix} \sigma_{h0}^2 & \sigma_{h01} \\ \sigma_{h01} & \sigma_{h1}^2 \end{pmatrix} \right] \tag{1}$$

for $i = 1, \ldots, N_h$ and $h = 1, \ldots, H$. Instead of observing $(Y_{hi}(0), Y_{hi}(1))$, we observe $T_{hi} \in \{0, 1\}$ and $Y_{hi} = T_{hi} Y_{hi}(1) + (1 - T_{hi}) Y_{hi}(0)$ for $i = 1, \ldots, N_h$ and $h = 1, \ldots, H$. The parameter of interest is the average treatment effect, which can be expressed as

$$\tau = \sum_{h=1}^{H} W_h \left(\mu_{h1} - \mu_{h0} \right)$$

where $W_h = N_h / N$. To estimate the average treatment effect, we can use

$$\hat{\tau}_{\text{sre}} = \sum_{h=1}^{H} W_h \hat{\tau}_h$$

where

$$\hat{\tau}_h = \frac{1}{N_{h1}} \sum_{i=1}^{N_h} T_{hi} Y_{hi} - \frac{1}{N_{h0}} \sum_{i=1}^{N_h} \left(1 - T_{hi} \right) Y_{hi}$$

and $N_{ht} = \sum_{i=1}^{N_h} \mathbb{I}\left(T_{hi} = t \right)$ for $t = 0, 1$. We assume that the treatment assigment mechanism is the stratified randomized experiment covered in the class (Chapter 4 lecture). In this case, we can easily show that $\hat{\tau}_{\text{sre}}$ is unbiased for τ.

1. Compute the variance of $\hat{\tau}_{\text{sre}}$ using the model parameters in (1). (Do not just give the formula. Show your work.)

2. Assuming the model parameters in (1) are known, what is the optimal sample allocation such that $Var\left(\hat{\tau}_{\text{sre}} \right)$ is minimized subject to $N_h = N_{h1} + N_{h0}$ for $h = 1, \ldots, H$ are fixed? That is, how to choose N_{h1} and N_{h0} for a given N_h?

Problem 4: (10 pt) Assume that a simple random sample of size n is selected from a population of size N and (x_i, y_i) are observed in the sample. In addition, we assume that the population mean of x, denoted by $\bar{X}$, is known.

1. Use a Taylor linearization method to find the variance of the product estimator $\bar{x}\bar{y}/\bar{X}$, where $(\bar{x}, \bar{y})$ is the sample mean of (x_i, y_i).
2. Find the condition that this product estimator has a smaller variance than the sample mean $\bar{y}$.
3. Prove that if the population covariance of x and y is zero, then the product estimator is less efficient than $\bar{y}$.

Problem 5: (10 pt) In a population of 10,000 businesses, we want to estimate the average sales $\bar{Y}$. For that, we sample $n = 100$ businesses using simple random sampling. Furthermore, we have at our disposal the auxiliary information "number of employees", denoted by x, for each business. It is known that $\bar{X} = 50$ in the population. From the sample, we computed the following statistics:

- $\bar{y}_n = 5.2 \times 10^6$ \$ (average sales in the sample)
- $\bar{x}_n = 45$ employees (sample mean)
- $s_y^2 = 25 \times 10^{10}$ (sample variance of y_k)
- $s_x^2 = 15$ (sample variance of x_k)
- $r = 0.8$ (sample correlation coefficient between x and y)

Answer the following questions.

1. Compute a 95% confidence interval for $\bar{Y}$ using the ratio estimator.
2. Compute a 95% confidence interval for $\bar{Y}$ using the regression estimator based on the simple linear regression of y on x (with intercept).

Problem 6: (10 pt)

Under the setup of the Chapter 6, Part 1 lecture, prove the last two equalities on page 23. That is, show that

$$\begin{aligned} Cov\left(\frac{1}{N_1}\sum_{i=1}^{N} T_i e_i(1), \frac{1}{N_0}\sum_{i=1}^{N}(1-T_i)\mathbf{x}_i'\mathbf{B}_0 \mid \mathcal{F}_N\right) &= 0 \\ Cov\left(\frac{1}{N_0}\sum_{i=1}^{N}(1-T_i) e_i(0), \frac{1}{N_0}\sum_{i=1}^{N}(1-T_i)\mathbf{x}_i'\mathbf{B}_0 \mid \mathcal{F}_N\right) &= 0 \end{aligned}$$

Problem 7: (10 pt)

Under the setup of the Chapter 6, Part 2 lecture,

1. Prove Lemma 3.
2. Show that the final weight in (13) satisfies a hard calibration for $\mathbf{x}_1$ in the sense that

$$\sum_{i \in A} \hat{\omega}_i \mathbf{x}_{1i} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_{1i}.$$