STAT 521: Homework Assignment 2 - Solution

Problem 1:

A city has a total of 90,000 dwelling units, of which 35,000 are houses, 45,000 are apartments, and 10,000 are condominiums. A stratified sample of size $n = 750$ 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 (h)	Estimated mean energy consumption kWh per dwelling unit $(\bar{y}_h)$	Estimated standard error of the sample mean $\widehat{SE}(\bar{y}_h)$
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 90,000 dwelling units.

Solution: Use

$$\hat{T}_{str} = \sum_{h=1}^{H} N_h \bar{y}_h = 35,000 * 915 + 45,000 * 614 + 10,000 * 712 = 66,775,000(\text{Kwh})$$

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

Solution:

$$SE(\hat{T}_{str}) = \sqrt{\sum_{h=1}^{H} N_h^2 V(\bar{y}_h)} = \sqrt{35000^2(4.84)^2 + 45000^2(2.98)^2 + 10000^2(7.00)^2} = 227,100(\text{Kwh})$$

3. What would the sample size if the optimal allocation is to be used (under $n = 750$) for this population? Assume that the survey costs are the same for each stratum.

Solution: The overall sampling rate is $n/N = 750/90000 = 0.00833$. Under proportional allocation, the sample sizes are $n_h = N_h \times 0.00833$. Thus, we have $n_1 = 292$, $n_2 = 375$, and $n_3 = 83$.

Now, using

$$SE(\bar{y}_h) = \sqrt{\left(\frac{1}{n_h} - \frac{1}{N_h}\right)} S_h$$

we can obtain S_h. That is, we may solve

$$\begin{aligned}
\sqrt{\frac{1}{292} - \frac{1}{35000}} S_1 &= 4.84 \\
\sqrt{\frac{1}{375} - \frac{1}{45000}} S_2 &= 2.98 \\
\sqrt{\frac{1}{83} - \frac{1}{10000}} S_3 &= 7.00
\end{aligned}$$

to obtain $S_1 = 83.053$, $S_2 = 57.949$, and $S_3 = 64.039$.

Finally, we can apply Neyman allocation

$$n_h = \frac{N_h S_h}{\sum_{h=1}^{H} N_h S_h} n$$

with $n = 750$. Thus,

$$\begin{aligned}
n_1 &= \frac{35 \times 83.053}{35 \times 83.053 + 45 \times 57.949 + 10 \times 64.039} \times 750 \cong 354 \\
n_2 &= \frac{45 \times 57.949}{35 \times 83.053 + 45 \times 57.949 + 10 \times 64.039} \times 750 \cong 318 \\
n_3 &= \frac{10 \times 64.039}{35 \times 83.053 + 45 \times 57.949 + 10 \times 64.039} \times 750 \cong 78
\end{aligned}$$

are the final sample sizes from Neyman allocation.

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?

Solution:

$$SE(\hat{T}_{str}) = \sqrt{\sum_{h=1}^{H} \frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) S_h^2} = 223,781(\text{Kwh})$$

Note that it is smaller than the SE under proportional allocation.

Problem 2:

Consider a simple random sample of size $n = 100$ 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:

	$X=1$	$X=0$	
$Y=1$	30	20	50
$Y=0$	25	25	50
	55	45	100

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 to obtain the sampling variance.)

Solution: [Answer to (1)]: The postratification estimator is

$$\hat{\theta} = W_1\bar{y}_1 + W_2\bar{y}_2 = 0.5 * (30/55) + 0.5 * (20/45).$$

[Answer to (2)]: For $\theta = P(Y = 1 \mid X = 1)$, we can use

$$\hat{\theta} = \frac{\hat{P}(X = 1, Y = 1)}{\hat{P}(X = 1)} = \frac{n_{11}}{n_{1+}} = 30/55,$$

where n_{ij} is the number of sample elements with $(X = i, Y = j)$ and $n_{i+} = n_{i1} + n_{i0}$.

Now, to obtain variance estimation, Taylor method can be use to get

$$\begin{aligned}
\hat{\theta} &\cong \theta + \frac{1}{E(n_{1+})}(n_{11} - \theta n_{1+}) \\
&= \theta + \frac{1}{E(n_{1+})/n}\left(\frac{n_{11}}{n} - \theta\frac{n_{1+}}{n}\right) \\
&= \theta + \frac{1}{E(\hat{P}_{1+})}\left(\hat{P}_{11} - \theta\hat{P}_{1+}\right).
\end{aligned}$$

Thus,

$$V(\hat{\theta}) \cong \frac{1}{P_{1+}^2}\left\{V(\hat{P}_{11}) + \theta^2 V(\hat{P}_{1+}) - 2\theta Cov\left(\hat{P}_{11}, \hat{P}_{1+}\right)\right\}. \tag{1}$$

Now, under simple random sampling, we have

$$\begin{aligned}
V(\hat{P}_{11}) &= \frac{1}{n}(1 - f)\, P_{11}(1 - P_{11}) \\
V(\hat{P}_{1+}) &= \frac{1}{n}(1 - f)\, P_{1+}(1 - P_{1+}) \\
Cov(\hat{P}_{11}, \hat{P}_{1+}) &= \frac{1}{n}(1 - f)\, P_{11}(1 - P_{1+}).
\end{aligned}$$

Also, using $\theta = P_{11}/P_{1+}$, we can simplify (1) to get

$$\begin{aligned}
V(\hat{\theta}) &= \frac{1}{n}(1 - f)\frac{1}{P_{1+}^2}\left\{P_{11}(1 - P_{11}) + \frac{P_{11}^2}{P_{1+}^2}P_{1+}(1 - P_{1+}) - 2\frac{P_{11}}{P_{1+}}P_{11}(1 - P_{1+})\right\} \\
&= \frac{1}{n}(1 - f)\frac{1}{P_{1+}^2}\left\{P_{11} - \frac{P_{11}^2}{P_{1+}}\right\} \\
&= \frac{1}{n}(1 - f)\frac{1}{P_{1+}}\theta(1 - \theta).
\end{aligned}$$

Thus, we can estimate the variance of $\hat{\theta}$ by

$$\hat{V}(\hat{\theta}) = (1 - f)\frac{1}{n_{1+}}\hat{\theta}\left(1 - \hat{\theta}\right).$$

Using $f = 0.01$, $n_{1+} = 55$, and $\hat{\theta} = 30/55$, we can obtain $\hat{V}(\hat{\theta}) = 0.004463$.

Problem 3:

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

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

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

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

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

1. Express the variance of $\hat{\tau}_{\text{sre}}$ using the model parameters in (2).

Solution: Recall that

$$E\left(\hat{\tau}_{\text{sre}} \mid \mathcal{F}_N \right) = \sum_{h=1}^{H} W_h \bar{\tau}_h$$

where $\bar{\tau}_h = N_h^{-1} \sum_{i=1}^{N_h} \{ Y_{hi}(1) - Y_{hi}(0) \}$. Also, in the class, we have learned that

$$V\left(\hat{\tau}_{\text{sre}} \mid \mathcal{F}_N \right) = \sum_{h=1}^{H} W_h^2 \frac{1}{N_h} \left(\frac{N_{h0}}{N_{h1}} S_{h1}^2 + \frac{N_{h1}}{N_{h0}} S_{h0}^2 + 2 S_{h01} \right).$$

Hence, the total variance is

$$\begin{aligned} V\left(\hat{\tau}_{\text{sre}} \right) &= V\left\{ E\left(\hat{\tau}_{\text{sre}} \mid \mathcal{F}_N \right) \right\} + E\left\{ V\left(\hat{\tau}_{\text{sre}} \mid \mathcal{F}_N \right) \right\} \\ &= V\left\{ \sum_{h=1}^{H} W_h \bar{\tau}_h \right\} + E\left\{ \sum_{h=1}^{H} W_h^2 \frac{1}{N_h} \left(\frac{N_{h0}}{N_{h1}} S_{h1}^2 + \frac{N_{h1}}{N_{h0}} S_{h0}^2 + 2 S_{h01} \right) \right\}. \end{aligned}$$

Solution: Now, under model (2), we can obtain

$$V\left\{\sum_{h=1}^{H} W_h \bar{\tau}_h\right\} = \sum_{h=1}^{H} W_h^2 \frac{1}{N_h}\left(\sigma_{h1}^2 + \sigma_{h0}^2 - 2\sigma_{h01}\right)$$

and

$$E\left\{\sum_{h=1}^{H} W_h^2 \frac{1}{N_h}\left(\frac{N_{h0}}{N_{h1}} S_{h1}^2 + \frac{N_{h1}}{N_{h0}} S_{h0}^2 + 2S_{h01}\right)\right\} = \sum_{h=1}^{H} W_h^2 \frac{1}{N_h}\left(\frac{N_{h0}}{N_{h1}}\sigma_{h1}^2 + \frac{N_{h1}}{N_{h0}}\sigma_{h0}^2 + 2\sigma_{h01}\right)$$

Therefore, combining the two, we obtain

$$V\left(\hat{\tau}_{\text{sre}}\right) = \sum_{h=1}^{H} W_h^2 \left(\frac{\sigma_{h1}^2}{N_{h1}} + \frac{\sigma_{h0}^2}{N_{h0}}\right).$$

2. Assuming the model parameters in (2) 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?

Solution: For each h, we wish to minimize

$$Q(N_{h1}, N_{h0}) = \frac{\sigma_{h1}^2}{N_{h1}} + \frac{\sigma_{h0}^2}{N_{h0}}$$

subject to $N_h = N_{h1} + N_{h0}$ is constant. Thus, by Schwartz inequality, we can obtain

$$\left(\frac{\sigma_{h1}^2}{N_{h1}} + \frac{\sigma_{h0}^2}{N_{h0}}\right)(N_{h1} + N_{h0}) \geq \left(\sigma_{h1} + \sigma_{h0}\right)^2$$

which is equality to

$$\left(\frac{\sigma_{h1}^2}{N_{h1}} + \frac{\sigma_{h0}^2}{N_{h0}}\right)(N_{h1} + N_{h0}) \geq \frac{\left(\sigma_{h1} + \sigma_{h0}\right)^2}{N_h}$$

with the equality if and only if

$$\frac{\sigma_{ht}}{N_{ht}^{1/2}} \propto N_{ht}^{1/2}, \quad t = 0, 1.$$

That is, the minimum of $Q(N_{h1}, N_{h0})$ is achieved at

$$N_{h1}^* = N_h \frac{\sigma_{h1}}{\sigma_{h1} + \sigma_{h0}}$$

and $N_{h0}^* = N_h - N_{h1}^*$.