STAT 521: Homework Assignment 4 - Solution

Problem 1:

A city block is divided into 100 blocks from which 5 blocks are selected with replacement and with probability proportional to the number of households enumerated in a previous census. Within each sampled block, the average household income and the average household size (=number of people in the household) are obtained from the sampled blocks. The following table presents a summary of information obtained from the sample blocks.

Table 1: Summary of information obtained from the sampled households

Block	Block Size	Average Household income ($\times 10^{-3}$\$)	Average Household size
1	50	30	2
2	60	70	4
3	47	80	5
4	50	50	4
5	70	60	4

1. What is the estimated average household income and its estimated variance?

Solution: By the property of the PPS sampling,

$$\begin{aligned} \hat{\bar{Y}} &= \frac{1}{n}\sum_{k=1}^{n} \bar{y}_{a_k} \\ &= \frac{1}{5}(30+70+80+50+60) \\ &= 58\ (\times 10^3\$) \end{aligned}$$

and

$$\begin{aligned} \hat{V}(\hat{\bar{Y}}) &= \frac{1}{n}\frac{1}{n-1}\sum_{k=1}^{n}\left(\bar{y}_{a_k} - \hat{\bar{Y}}\right)^2 \\ &= 74 \end{aligned}$$

2. What is the estimated per capita income (= income per person) and its estimated variance? (You may need to use a Taylor linearization.)

Solution: Since $\theta = \bar{Y}/\bar{X}$ where $\bar{X}$ is the average household size. Thus, we have

$$\hat{\theta} = \frac{\hat{\bar{Y}}}{\hat{\bar{X}}} = \frac{58}{3.8} = 15.26 \; (\times 10^3 \$)$$

because

$$\hat{\bar{X}} = \frac{1}{5}(2+4+5+4+4) = 3.8.$$

Also, by a Taylor linearization,

$$\hat{\theta} \cong \theta + \frac{1}{\bar{X}}\left(\hat{\bar{Y}} - \theta \hat{\bar{X}}\right)$$

and

$$\hat{V}(\hat{\theta}) \cong \left(\frac{1}{\hat{\bar{X}}}\right)^2 \frac{1}{n}\frac{1}{n-1}\sum_{k=1}^{n}\left(y_{a_k} - \hat{\theta} x_{a_k}\right)^2 = \frac{1}{3.8^2}\cdot\frac{1}{5}\cdot 54.29 \doteq 0.752$$

Problem 2:

Suppose that we have a population of clusters with equal size M. Suppose that the population has the following ANOVA structure as summarized in the folloiwng table.

Table 2: ANOVA table

Source	d.f.	Mean Sum of Square
Between Clusters	49	6,218
Within Clusters	450	2,918

1. Find the cluster size M.

Solution: In the AONVA, the d.f. for "Between clusters" sum of squares is $N_I - 1 = 49$ and the d.f for "within cluster" sum of squares is $N_I(M-1) = 450$. Thus, $M = 10$.

2. Estimate the intracluster correlation coefficient.

Solution: We can use

$$S^2 = \frac{1}{M} S_b^2 + \left(1 - \frac{1}{M}\right) S_w^2 = 0.1 \times 6218 + 0.9 \times 2918 = 3248$$

and so

$$\hat{\rho} = 1 - \frac{S_w^2}{S^2} = 1 - 2918/3248 = 0.1016$$

3. What is the variance of the mean estimator under this cluster sampling?

Solution:

$$V(\bar{y}) = \frac{1}{n_I}\left(1 - \frac{n_I}{N_I}\right) S_b^2 \cong \frac{1}{50} \cdot 6218 = 124.36$$

4. Compute the design effect of this sampling design and give an interpretation.

Solution: The formula for design effect is $1 + (M - 1)\rho$. Thus, the estimated design effect is $1 + 9 \times 0.01016 = 1.9144$. The effective sample size is $n^* = n/\text{diff} = n/1.9144$. Thus, the above cluster sampling has the same efficiency of the SRS of size $n^* = (n_I M)/1.9144$ from the same finite population.

Problem 3: (30 pt) A statistician wishes to carry out a survey on the quality of health care in the cardiology service of hospitals. For that, he selects by simple random sampling of $n = 100$ hospitals among the $N = 1,000$ hospitals listed and then, in each of the selected hospitals, he collects the opinions of all the cardiology patients.

1. We consider that each cardiology unit is comprised of exactly $M = 50$ beds and that the 95% confidence interval on the true proportion P of dissatisfied patient is:

$$P \in [0.10 \pm 0.018],$$

(that signifies in particular that, in the sample, 10 % of patients are dissatisfied with the quality of care). How do you estimate the intracluster correlation coefficient ?

2. How would the accuracy of the statistician's survey on satisfaction evolve if, all at once, there are $M = 25$ beds and $n = 200$ hospitals are selected in the sample using the same sampling design ?

3. Compute the ratio of the two variances in (1) and (2) and explain it in terms of intracluster correlation.

Solution:

1. First note that $\hat{P} = 0.1$ and $1.96\sqrt{\hat{V}\left(\hat{P}\right)} = 0.018$. Now, since

$$Var\left(\hat{P}\right) = \frac{1}{n_I M}\left(1 - \frac{n_I}{N_I}\right) S_b^2 = \frac{1}{n_I M}\left(1 - \frac{n_I}{N_I}\right) S^2\left[1 + (M-1)\rho\right]$$

and S^2 is estimated by $P(1-P) = 0.1 * 0.9 = 0.09$, we can solve

$$\left(\frac{0.018}{1.96}\right)^2 = \frac{1}{100 * 50}\left(1 - \frac{100}{1000}\right) 0.09\left[1 + (50-1)\rho\right]$$

to get $\rho \doteq 0.0858$.

2. The variance under the new design will be

$$\frac{1}{200 * 25}\left(1 - \frac{200}{1000}\right) 0.09\left[1 + (25-1)\rho\right] = 4.405 \times 10^{-5}.$$

3. The variance under (1) is $8.434/4.405 = 1.915$ time bigger than the variance under (2). The ratio can be explained by

$$\frac{(1-0.1)[1 + (50-1)\rho]}{(1-0.2)[1 + (25-1)\rho]}.$$