STAT 521: Homework Assignment 3

Due at 10:00am on May 8, 2025

Problem 1: (20 pt)

A researcher wants to estimate the total number of patients discharged from hospitals in Iowa in January 2019. It is known that there are $N = 145$ hospitals in Iowa and the researcher obtains a list of all $N = 145$ hospitals in Iowa from administrative data. The list contains the number of inpatient beds in each hospital in the population. She (= the researcher) decides to select a sample using probability proportional to size sampling with replacement. The size variable (x_i) is the number of inpatient beds in the hospital. The total of x_i for all 145 hospitals in Iowa is $T_x = \sum_{i=1}^{N} x_i = 13,785$ inpatient beds. She selects a probability proportional to size sample with replacement with three independent draws and draw probability proportional to x_i. She collects the number of patients discharged in January 2019 for the sampled hospitals. The table below contains the data for the hospitals obtained in the three draws.

Draw	Hospital ID	Number of beds (x_i)	Number of patients (y_i)
1	46	250	754
2	88	100	321
3	113	450	1362

1. What is the estimate of the total number of patients discharged in January 2019 from the population of hospitals in this region?

2. Provide a 95% confidence interval for the total number of patients discharged in January 2019. Show intermediate steps.

3. Estimate the average number of patients discharged per hospital in January 2019 and provide a corresponding standard error. Show intermediate steps.

Problem 2: (15 pt)

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?

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

Problem 3: (25 pt)

A researcher wants to estimate the average household income in a city using two-phase sampling.

Phase 1: Basic Survey

200 households are selected using simple random sampling (SRS) from 5,000 households. Collected info: the household size x_i which is the total number of adults and children in household i.

Phase 2: Detailed Income Survey

From the 200 households, 80 households are selected to Collected info: Household income y_i (\$1,000).

1. If the second phase sample were selected using probability proportional to household size (PPS). Calculate the second-phase conditional inclusion probabilities $\pi^{(2)}_{i|A_1}$ for a household i with 2 adults and 1 child. Can you compute the overall inclusion probability for this household?

The researcher decide to use a simple random sample in the second phase to select the 80 households, and the summary statistics from both phases are as follows:

Phase 1 Summary Statistics

$$\bar{x}_1 = 3.2, \quad s^2_{x1} = 2.0$$

Phase 2 Summary Statistics

$$\bar{x}_2 = 3.5, \quad s^2_{x2} = 2.2, \quad \bar{y}_2 = 58, \quad s^2_{y2} = 100, \quad r_{xy} = 0.6$$

2. Estimate the mean household income using $\pi^\star$-estimator.
3. Calculate the approximate variance of the $\pi^\star$-estimate.
4. Calculate the regression estimator of the mean household income using household size as the covariate.
5. Calculate the approximate variance of the regression estimator.
6. What advantage does the regression estimator have over the $\pi^\star$-estimate?

Problem 4: (20 pt)

A health researcher is studying the effect of a new drug treatment ($T = 1$) versus a control ($T = 0$) on patient blood pressure reduction (Y). Because treatment was not randomly assigned, the researcher uses observational data and applies causal inference methods.

The data below summarize 10 patients:

ID	Treatment (T)	Blood Pressure Change (Y)	Age (X)	Propensity Score $\hat{\pi}(X)$	$\hat{Q}(X,1), \hat{Q}(X,0)$
1	1	-12	55	0.7	-11, -6
2	1	-10	60	0.6	-12, -7
3	1	-13	50	0.8	-10, -5
4	1	-15	65	0.5	-14, -8
5	0	-5	55	0.7	-11, -6
6	0	-6	60	0.6	-12, -7
7	0	-7	50	0.8	-10, -5
8	0	-9	65	0.5	-14, -8
9	1	-11	58	0.65	-11, -6
10	0	-8	62	0.55	-12, -7

1. Calculate the IPW estimate of the average blood pressure change for the treated and control groups.
2. Compute the DIME estimate of average treatment effect (ATE) without considering the propensity scores. Is this in general a good estimate for ATE? Briefly explain your reasoning.
3. Calculate the **IPW estimate** of the ATE.
4. Calculate the optimal AIPW estimate of the ATE.
5. What is the advantage of using AIPW over IPW?
6. Explain why AIPW is called **doubly robust**.

Notes to Students

- Round estimates to two decimal places.
- Show all intermediate steps.
- You may use R, Excel, or a calculator.
- Write 2–3 sentences for each conceptual answer.