# HW3

Sam Olson

## Problem 1:

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?

#### Answer

For this problem, we use the PPS estimator, which is of the form:

$$\hat{T}_y = \frac{1}{n} \sum_{j=1}^{n} \frac{y_{i_j}}{\pi_{i_j}}$$

With $n = 3$.

Calculating using the information provided, our estimate is:

$$\hat{T}_y = \frac{1}{3} \sum_{j=1}^{3} \frac{y_{i_j}}{\pi_{i_j}} = \frac{1}{3} \left( \frac{754}{250/13785} + \frac{321}{100/13785} + \frac{1362}{450/13785} \right) \approx 42,516$$

```r
denom <- 13785
pi_ij <- c((250/denom), (100/denom), (450/denom))
y_ij <- c(754, 321, 1362)
n <- 3

tHat <- (1/n) * sum(y_ij / pi_ij)
tHat
```

```
## [1] 42516
```

Or, we estimate the total number of patients discharged in January 2019 from the population of hospitals in this region is 42,516.

Note: When I manually calculated this (greater precision) the estimate came out to 42,517.87, which is close but nonetheless different than the above. This also has implications for the resulting confidence interval, though only slightly.

## 2.

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

### Answer

We center our confidence interval on the point estimate calculated in the prior step.

We then need to find the critical value, which for a 95% confidence interval is roughly 1.96 (assuming normally distributed mean/average via CLT).

Then, we have to calculate our Standard Error. To begin with, we calculate the variance using the estimated variance formula for PPS, of the form:

$$\hat{V}(\hat{T}_y) = \frac{1}{n(n-1)} \sum_{j=1}^{3} \left( \frac{y_{i_j}}{\pi_{i_j}} - \hat{T}_y \right)^2 \approx 759{,}188.4$$

```r
variance_est <- (1/(n*(n-1))) * sum(((y_ij/pi_ij) - tHat)^2)
variance_est
```

```
## [1] 753357.8
```

With the variance estimate, we then get the standard error and combine with our typical confidence interval formula. Giving us:

$$\hat{SE}(\hat{T}_y) = \sqrt{753533.3} \approx 868.063$$

Taken together, our 95% confidence interval is of the form:

$$\text{CI}_{95\%} = \hat{T}_y \pm 1.96 \cdot \hat{SE} \approx (40{,}833.34,\ 44{,}236.14)$$

So, our 95% confidence interval for the total number of patients discharged in January 2019 is (40,833.34, 44,236.14).

```r
SE <- sqrt(variance_est)
SE
```

```
## [1] 867.9619
```

```r
CI_lower <- tHat - 1.96 * SE
CI_upper <- tHat + 1.96 * SE
c(CI_lower, CI_upper)
```

```
## [1] 40814.80 44217.21
```

Note: Again, we get a slightly different answer when explicitly calculating this without any rounding in intermediate steps. The resulting confidence interval would be: (40,826.78, 44,242.70)

## 3.

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

**Answer**

$$\hat{\bar{Y}} = \frac{\hat{T}_y}{N} = \frac{42{,}516}{145} \approx 293.34$$

$$\hat{SE}(\hat{\bar{Y}}) = \frac{\hat{SE}(\hat{T}_y)}{N} = \frac{868.063}{145} \approx 5.99$$

Via R with less precision, we have: Explicitly, without rounding, we have: The average number of patients discharged per hospital is 293.34 with Standard Error 5.99.

Explicitly, without rounding, we have: The average number of patients discharged per hospital is 293.34 with Standard Error 6.01.

```r
N <- 145
tHat/N
```

```
## [1] 293.2138
```

```r
SE/N
```

```
## [1] 5.985944
```

# Problem 2:

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.

| 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?

**Answer**

For this problem, we again use the PPS estimator, noting we are doing PPS with replacement.

For our point estimate, we have:

$$\hat{\bar{Y}} = \frac{1}{n}\sum_{k=1}^{n} \bar{y}_k = \frac{30+70+80+50+60}{5} = 58$$

And for the variance formula, we have:

$$\hat{V}(\hat{\bar{Y}}) = \frac{1}{n(n-1)}\sum_{k=1}^{n}(\bar{y}_k - \hat{\bar{Y}})^2 = \frac{(30-58)^2+\cdots+(60-58)^2}{20} = 74$$

Taken together, the estimated average household income is $\$58 \times 10^3$ and its estimated variance is $\$74 \times 10^3$.

Note: I believe there is a typo in the table provided, and that $\times 10^3$ should be used instead of $\times 10^{-3}$. This assumption will be carried into the following problems when giving units of the estimates.

## 2.

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

**Answer**

First compute average household size:

$$\hat{\bar{X}} = \frac{2+4+5+4+4}{5} = 3.8$$

Giving us a point estimate of on average 3.8 people per household. We need to convert this into a ratio though, so we then have:

$$\hat{\theta} = \frac{\hat{Y}}{\hat{\hat{X}}} = \frac{58}{3.8} \approx 15.26$$

Giving us the ratio $15.26 \times 10^3$

Via linearizaton, we define the variable z (linarized variable) as:

$$z_k = y_k - \hat{\theta} x_k$$

Compute $z_k$ values for each of the five blocks by:

$$\begin{aligned} z_1 &= 30 - 15.26 \cdot 2 = -0.52 \\ z_2 &= 70 - 15.26 \cdot 4 = 8.96 \\ z_3 &= 80 - 15.26 \cdot 5 = 3.70 \\ z_4 &= 50 - 15.26 \cdot 4 = -11.04 \\ z_5 &= 60 - 15.26 \cdot 4 = -1.04 \end{aligned}$$

We can then compute the variance:

$$s_z^2 = \frac{1}{n-1} \sum_{k=1}^{n} (z_k - \bar{z})^2 = \frac{(-0.52 - 0.012)^2 + \cdots + (-1.04 - 0.012)^2}{4} \approx 54.29$$

Converting back to $\theta$ notation, we have:

$$\hat{V}(\hat{\theta}) = \left( \frac{1}{\hat{\hat{X}}} \right)^2 \frac{s_z^2}{n} = \frac{1}{3.8^2} \cdot \frac{54.29}{5} \approx 0.752$$

Giving us an estimated per capita income of $15.26(\times 10^3)$ with estimated variance $0.752(\times 10^3)$.

Note: The answer here is in units of "income per person", which I believe would be something like "$\$\times 10^3$/person".

# Problem 3:

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?

**Answer** We cannot explicitly compute the overall inclusion probability $\pi_i$ unless we know the total household size $T_x^{(1)} = \sum_{j=1}^{200} x_j$ from Phase 1. And this typically is something we don't know, as it would mean we already have data from all Phase 1 households. I believe this is a fundamental limitation of two-phase designs and why estimators such as $\pi^*$ estimator are used for two-phase sampling designs.

Explicitly, the above comes from:

Under the Two-Phase Sampling design, the second-phase inclusion probability for a household with $x_i = 3$ (2 adults + 1 child) is given by:

$$\pi^{(2)}_{i|A_1} = \frac{n_2 \cdot x_i}{T_x^{(1)}}$$

Using all known quantities, this simplifies to:

$$\pi^{(2)}_{i|A_1} = \frac{80 \cdot 3}{\sum_{j \in A_1} x_j} = \frac{240}{T_x^{(1)}}$$

Where $T_x^{(1)}$ is the total household size in the Phase 1 sample.

We then have the overall inclusion probability given by:

$$\pi_i = \pi^{(1)} \cdot \pi^{(2)}_{i|A_1} = \frac{200}{5000} \cdot \left(80 \cdot \frac{x_i}{T_x^{(1)}}\right)$$

At most, we may simplify to:

$$\pi_i = \frac{16000 x_i}{5000 T_x^{(1)}}$$

However, as noted, we still require knowing $T_x^{(1)}$ explicitly to calculate the overall inclusion probability.

**Back to the Question** 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**

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

**Phase 2 Summary Statistics**

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

## 2.

Estimate the mean household income using $\pi^\star$-estimator.

**Answer**

The formula for the $\pi^\star$-estimator is of the form:

$$\bar{y}_{\pi^\star} = \bar{y}_2 + (\bar{x}_1 - \bar{x}_2) \cdot \frac{s_{xy}}{s_{x2}^2}$$

Where the term $s_{xy}$ is the covariance between X and Y, given by:

$$s_{xy} = r_{xy} \cdot s_{x2} \cdot s_{y2} = 0.6 \cdot \sqrt{2.2} \cdot \sqrt{100} \approx 8.9$$

Giving us the term $\frac{s_{xy}}{s_{x2}^2}$ given by:

$$\frac{8.9}{2.2} \approx 4.045$$

Taken together, combining all known quantities in the above formula, we have:

$$58 + (3.2 - 3.5) \cdot 4.045 \approx 56.79$$

Giving an estimate of the mean household income using $\pi^\star$-estimator of \$56,790.

## 3.

Calculate the approximate variance of the $\pi^\star$-estimate.

**Answer**

We use the following formula for the approximate variance:

$$\text{Var}(\hat{Y}^*) \approx \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2 + \left(\frac{1}{r} - \frac{1}{n}\right) S_e^2$$

Where the term $\left(\frac{1}{n} - \frac{1}{N}\right) S_y^2$ is the approximate Phase 1 variance and $\left(\frac{1}{r} - \frac{1}{n}\right) S_e^2$ is the approximate Phase 2 variance.

Using known values and calculating:

First Term (Phase):

We treat the term $\frac{1}{N} \approx 0$, giving us:

$$\frac{S_y^2}{n} = \frac{100}{200} = 0.5$$

Second Term: (Phase)

$$S_e^2 = S_y^2(1 - \rho^2)$$

Where:

$$\rho^2 = (0.6)^2 = 0.36 \to 1 - \rho^2 = 1 - 0.36 = 0.64 \to S_e^2 = 100 \cdot 0.64 = 64$$

And:

$$\frac{1}{r} - \frac{1}{n} = \frac{1}{80} - \frac{1}{200} = 0.0125 - 0.005 = 0.0075 \to \left(\frac{1}{r} - \frac{1}{n}\right) S_e^2 = 0.0075 \cdot 64 = 0.48$$

Combining:

$$\text{Var}(\bar{y}_{\pi^*}) \approx 0.5 + 0.48 = 0.98$$

## 4.

Calculate the regression estimator of the mean household income using household size as the covariate.

**Answer**

The regression estimator is formula is of the form:

$$\bar{y}_{\text{reg}} = \bar{y}_2 + (\bar{x}_1 - \bar{x}_2) \cdot b$$

Where:

$$b = \frac{s_{xy}}{s_{x2}^2} \approx 4.045$$

All other quantities are known values, so we may calculate:

$$\bar{y}_{\text{reg}} = 58 - 1.2135 = 56.79$$

So, the regression estimator of the mean household income using household size as the covariate provides an estimate of \$56,790, which happens to be the same as the estimate using $\pi^\star$-estimator previously.

## 5.

Calculate the approximate variance of the regression estimator.

**Answer**

The approximate variance is of the form:

$$\text{Var}(\bar{y}_{reg,tp}) \approx \left(\frac{1}{n} - \frac{1}{N}\right) B' S_{xx} B + \left(\frac{1}{r} - \frac{1}{n}\right) S_{ee}$$

Where the term $\left(\frac{1}{n} - \frac{1}{N}\right) B' S_{xx} B$ is the approximate Phase 1 variance and $\left(\frac{1}{r} - \frac{1}{n}\right) S_{ee}$ is the approximate Phase 2 variance.

Using known quantities, we calculate:

First Term (Phase):

$$B' S_{xx} B = \rho^2 S_y^2 = 0.36 \cdot 100 = 36 \rightarrow \left(\frac{1}{200} - 0\right) \times 36 = 0.18$$

Second Term (Phase):

$$\left(\frac{1}{80} - \frac{1}{200}\right) \cdot 64 = 0.48$$

Combining:

$$0.18 + 0.48 = 0.66$$

## 6.

What advantage does the regression estimator have over the $\pi^\star$-estimate?

**Answer**

The regression estimator is more flexible, efficient, and robust. In that order: The regression estimator can incorporate multiple covariates, which can improve the generalizability of results. Also, because the regression estimator takes advantage of covariates, it "exploits" the correlation structure between the response and covariate(s) more effectively, and as a result can be more efficient (have smaller variance) than the $\pi^\star$-estimator, especially so when the covariate(s) are strongly correlated with the response. Lastly, the regression estimator is more robust in the sense that it is less sensitive to violations of the model assumptions, i.e., it can still be consistent (in the traditional sense, i.e., converges to the true parameter) when the model is mis-specified.

For the purposes of this problem, it just so happens that the estimates and approximate variances in the above are equal, which is a result of setting the intercept term to 0 in the regression model (in this scenario the regression reduces to the ratio estimator, which in the context of this problem is equivalent to the $\pi^\star$-estimator).

# Problem 4:

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.

**Answer**

**Treatment Group, $T = 1$:** Formula:

$$\bar{Y}^{(1)}_{IPW} = \frac{\sum_{i=1}^{n} \frac{T_i Y_i}{\hat{\pi}(X_i)}}{\sum_{i=1}^{n} \frac{T_i}{\hat{\pi}(X_i)}}$$

Calculating:

$$\text{Numerator} = \frac{-12}{0.7} + \frac{-10}{0.6} + \frac{-13}{0.8} + \frac{-15}{0.5} + \frac{-11}{0.65} = -17.14 - 16.67 - 16.25 - 30.00 - 16.92 = -96.98$$

$$\text{Denominator} = \frac{1}{0.7} + \frac{1}{0.6} + \frac{1}{0.8} + \frac{1}{0.5} + \frac{1}{0.65} = 1.429 + 1.667 + 1.25 + 2 + 1.538 = 7.884$$

Combining

$$\bar{Y}^{(1)}_{IPW} = \frac{-96.98}{7.884} \approx -12.31$$

**Control Group, $T = 0$:** Formula:

$$\bar{Y}_{IPW}^{(0)} = \frac{\sum_{i=1}^{n} \frac{(1-T_i)Y_i}{1-\hat{\pi}(X_i)}}{\sum_{i=1}^{n} \frac{(1-T_i)}{1-\hat{\pi}(X_i)}}$$

Calculating:

$$\text{Numerator} = \frac{-5}{0.3} + \frac{-6}{0.4} + \frac{-7}{0.2} + \frac{-9}{0.5} + \frac{-8}{0.45} = -16.67 - 15.00 - 35.00 - 18.00 - 17.78 = -102.45$$

$$\text{Denominator} = \frac{1}{0.3} + \frac{1}{0.4} + \frac{1}{0.2} + \frac{1}{0.5} + \frac{1}{0.45} = 3.333 + 2.5 + 5 + 2 + 2.222 = 15.055$$

Combining terms:

$$\bar{Y}_{IPW}^{(0)} = \frac{-102.45}{15.055} \approx -6.81$$

Explicitly, in R:

```
data <- data.frame(
  ID = 1:10,
  T = c(1, 1, 1, 1, 0, 0, 0, 0, 1, 0),
  Y = c(-12, -10, -13, -15, -5, -6, -7, -9, -11, -8),
  pi_X = c(0.7, 0.6, 0.8, 0.5, 0.7, 0.6, 0.8, 0.5, 0.65, 0.55)
)

numer_treated <- sum((data$T * data$Y) / data$pi_X)
denom_treated <- sum(data$T / data$pi_X)
ipw_treated <- numer_treated / denom_treated

numer_control <- sum(((1 - data$T) * data$Y) / (1 - data$pi_X))
denom_control <- sum((1 - data$T) / (1 - data$pi_X))
ipw_control <- numer_control / denom_control

ipw_treated
```

```
## [1] -12.30166
```

```
ipw_control
```

```
## [1] -6.804428
```

Again, I think the difference in the manual calculations and my R validation is a result of R's precision/rounding when making these types calculation that involve a fair amount of precision. Though for this problem, the differences are very small (in the hundreths of decimal place).

## 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.

**Answer**

The DIME (Difference in Means Estimator) is:

$$\widehat{ATE}_{DIME} = \bar{Y}_T - \bar{Y}_C$$

Where:

Treated group average: $\bar{Y}_T = \frac{-12-10-13-15-11}{5} = -12.2$

And:

Control group average: $\bar{Y}_C = \frac{-5-6-7-9-8}{5} = -7.0$

$$\widehat{ATE}_{DIME} = \frac{-12-10-13-15-11}{5} - \frac{-5-6-7-9-8}{5} = -5.2$$

```
treated_outcomes <- c(-12, -10, -13, -15, -11)
control_outcomes <- c(-5, -6, -7, -9, -8)

mean_treated <- mean(treated_outcomes)
mean_control <- mean(control_outcomes)
ate_dime <- mean_treated - mean_control

ate_dime
```

```
## [1] -5.2
```

No, this is not a generally good estimate! This estimate is problematic because it assumes treatments were randomly assigned, which is not the case! This introduces bias and the estimate is possibly being confounded by other factors that may be relevant to the analysis such as age, gender, or other factors.

But at least that one doesn't have any discrepancy with the R validation!

## 3.

Calculate the IPW estimate of the ATE.

**Answer**

$$\widehat{ATE}_{IPW} = -12.31 - (-6.81) = -5.50$$

```
ipw_treated - ipw_control
```

```
## [1] -5.497233
```

The IPW estimate of the ATE is $\approx -5.50$ (Treatment - Control).

## 4.

Calculate the optimal AIPW estimate of the ATE.

**Answer**

The AIPW estimator formula is given by:

$$\widehat{ATE}_{AIPW} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{T_i(Y_i - \hat{Q}(X_i,1))}{\hat{\pi}(X_i)} + \hat{Q}(X_i,1) - \frac{(1-T_i)(Y_i - \hat{Q}(X_i,0))}{1-\hat{\pi}(X_i)} - \hat{Q}(X_i,0)\right]$$

I am not going to attempt this manually. I defer strictly to the R calculation below:

```
df <- data.frame(
  Trt = c(1,1,1,1,0,0,0,0,1,0),
  Y = c(-12,-10,-13,-15,-5,-6,-7,-9,-11,-8),
  pi = c(0.7,0.6,0.8,0.5,0.7,0.6,0.8,0.5,0.65,0.55),
  Q1 = c(-11,-12,-10,-14,-11,-12,-10,-14,-11,-12),
  Q0 = c(-6,-7,-5,-8,-6,-7,-5,-8,-6,-7)
)

df$aipw <- df$Trt * (df$Y - df$Q1) / df$pi + df$Q1 -
           (1 - df$Trt) * (df$Y - df$Q0) / (1 - df$pi) - df$Q0

round(mean(df$aipw), 2)
```

```
## [1] -4.75
```

Giving us an optimal AIPW estimate of the ATE of -4.75.

## 5.

What is the advantage of using AIPW over IPW?

**Answer**

The advantages the AIPW estimator has compared to IPW is the fact it is "doubly robust" and its improved efficiency. The AIPW estimator combines both the propensity model (PM) and the outcome model (OM), making it doubly robust, covered in more detail in the next part. On the point of efficiency: The AIPW estimator reduces variance compared to IPW by combining weighting with regression adjustment, making it at least as efficient as the IPW estimator. The AIPW is also more stable with extreme propensity scores, i.e., scores near the tails, 0 or 1.

## 6.

Explain why AIPW is called doubly robust.

**Answer**

The AIPW estimator is called "doubly robust" because the AIPW estimator is consistent if either the OM or the PM is correct. Specifically: The AIPW model can still be unbiased if either the OM or PM are mis-specified. This is not XOR though, at least one needs to be correctly specified!