A Brief Overview of Survey Sampling (and the Connection to Offline Policy Evaluation)

Vincent Liu

Outline

Introduction

- A finite population $\mathcal{U}$ of size N
- For each unit $i \in \mathcal{U}$, y_i is the study variable and $\mathbf{x}_i$ is the auxiliary variable
- A subset of the population, called a probability sample, is selected according to some *sampling selection scheme*, then we observe y_i for each unit in this subset
- The goal is to estimate some population parameters, for example, the population total $t = \sum_{i \in \mathcal{U}} y_i$ or mean $\bar{Y} = \frac{1}{N} \sum_{i \in \mathcal{U}} y_i$

Sampling Design

- The *design* vector $\mathbf{I} = (I_1, \ldots, I_N)$ is a random vector describing how the samples are drawn from the population, for example, $I_1 > 0$ means that the unit 1 is selected and $I_2 = 0$ means the unit 2 is not selected
- The distribution of the design vector is called the sampling design $\mathbf{I} \sim P$

WOR and WR Design

- Example of WOR design: simple random design
  Any sample of n distinct units has the same probability, that is

$$P(\mathbf{I}_1 = i_1, \ldots, \mathbf{I}_N = i_N) = \begin{cases} \frac{1}{\binom{N}{n}}, \text{ if } \sum_i |i_i| = n, i_i \in \{0,1\} \\ 0 \end{cases}$$

- Example of WR design: multinomial design
  Each unit is drawn according to $p_1, \ldots, p_N$ (with $\sum p_i = 1$) and we repeat the process n times (independently), that is

$$P(\mathbf{I}_1 = i_i, \ldots, \mathbf{I}_N = i_N) = \begin{cases} \frac{n!}{\Pi_{i=1}^N i_i!} p_1^{i_1} \ldots p_N^{i_N}, \text{ if } \sum_i |i_i| = n \\ 0 \end{cases}$$

Example 1: percentage of population who are vaccinated

- $y_i = 1$ if the person i is vaccinated and 0 otherwise
- We want to know $\bar{Y} = \frac{1}{N} \sum_{i \in \mathcal{U}} y_i$
- We choose n people randomly to conduct survey/interview

Example 2: OPE

- Consider the contextual bandits problem: $\mathcal{S}$ be the set of context, $\mathcal{A}$ be the set of actions, and $r_a(s) \in \mathbb{R}$ be the associated reward
- Given a policy π, let $\mathcal{U} = \mathcal{S} \times \mathcal{A}$ and $y_{s,a} = P(s)\pi(a|s)r_a(s)$. We want to estimate the population total

$$t = \sum_{(s,a)\in\mathcal{U}} y_{s,a} = \sum_{s\in\mathcal{S}, a\in\mathcal{A}} P(s)\pi(a|s)r_a(s) = V^{\pi}$$

- Samples are drawn according to a multinomial design with $p_{s,a} = P(s)\pi_b(a|s)$
- Some limitations:
  - finite population
  - deterministic rewards (the study variable)
  - known behavior policy (the sampling design)

Example 3: multilabel classification with partial feedback

- $\mathcal{S}$ be the set of inputs, $\mathcal{A}$ be the set of classes, and $r_a(s)$ be the associated (binary) relevance
- Given a top-k classifier $\pi : \mathcal{S} \to \{0,1\}^{|\mathcal{A}|}$, we want to estimate Precision@$k = \sum_{s \in \mathcal{S}, a \in \mathcal{A}} \frac{1}{|\mathcal{S}|k} \pi(a|s) r_a(s)$
- Samples are drawn according to some sampling design and we only observe the relevance for these samples

General Estimator

- Example: $\mathcal{U} = \{1, 2, \ldots, 10\}$ and $S = \{1, 1, 2, 5, 10\}$
- An estimator of the total $t = \sum_{i \in \mathcal{U}} y_i$ is

$$\hat{t} = \sum_{i \in S} \frac{y_i}{\mathbb{E}[I_i]} = \sum_{i \in \mathcal{U}} \frac{I_i y_i}{\mathbb{E}[I_i]}$$

- The only randomness is $I_i \sim P(\mathbf{I})$
- $\hat{t}$ is unbiased:

$$\mathbb{E}\left[\hat{t}\right] = \mathbb{E}\left[\sum_{i \in U} \frac{I_i y_i}{\mathbb{E}[I_i]}\right] = \sum_{i \in U} \frac{\mathbb{E}[I_i] y_i}{\mathbb{E}[I_i]} = \sum_{i \in \mathcal{U}} y_i = t$$

Variance

- The variance of $\hat{t}$ is

$$V(\hat{t}) = \sum_{i \in \mathcal{U}} \sum_{j \in \mathcal{U}} \frac{y_i}{\mathbb{E}[I_i]} \frac{y_j}{\mathbb{E}[I_j]} Cov(I_i, I_j)$$

- An unbiased variance estimator is

$$\hat{V}(\hat{t}) = \sum_{i \in S} \sum_{j \in S} \frac{y_i}{\mathbb{E}[I_i]} \frac{y_j}{\mathbb{E}[I_j]} \frac{Cov(I_i, I_j)}{\mathbb{E}[I_i I_j]}$$

- When the sample size is fixed, the Sen-Yates-Grundy variance estimator is

$$\hat{V}(\hat{t}) = -\frac{1}{2} \sum_{i \in S} \sum_{j \in S} Cov(I_i, I_j) \left(\frac{y_i}{\mathbb{E}[I_i]} - \frac{y_j}{\mathbb{E}[I_j]} \right)^2$$

Hansen-Hurwitz estimator

- For multinomial design, the Hansen-Hurwitz estimator is

$$\hat{t} = \sum_{i \in S} \frac{y_i}{\mathbb{E}[I_i]} = \sum_{i \in S} \frac{y_i}{np_i}$$

$$V(\hat{t}) = \frac{1}{n} \left(\sum_{i \in U} \frac{y_i^2}{p_i} - t^2 \right)$$

$$\hat{V}(\hat{t}) = \frac{1}{n(n-1)} \left(\sum_{i \in S} \frac{y_i^2}{p_i^2} - n\hat{t}^2 \right)$$

- This is the design-based estimator

Model-assisted approach: ratio estimator

- **Model-assisted**: use auxiliary information to improve estimator
- Recall $\mathbf{x}_i$ is the auxiliary variable (which is assumed to be known for all units), the ratio estimator is

$$\hat{t}_r = t_x \frac{\hat{t}}{\hat{t}_x}$$

where $\hat{t}$ is the HH estimator for the total of y, $\hat{t}_x$ is the HH estimator for the total of x, and $t_x = \sum_{i \in \mathcal{U}} x_i$

Model-assisted approach: ratio estimator

- Bias:

$$|\mathbb{E}[\hat{R}] - R| \le \frac{\sqrt{V(\hat{R})}\sqrt{V(\hat{t}_x)}}{t_x}$$

  where $R = t/t_x$ and $\hat{R} = \hat{t}/\hat{t}_x$

- Variance: use the Taylor linearization technique

$$\hat{R} = R + \frac{1}{t_x}(\hat{t}_y - R\hat{t}_x) + O(n^{-\frac{1}{2}}),$$

  the approximate variance is

$$V(\hat{t}_r) = V(t_x\hat{R}) \approx V(\hat{t}_y - R\hat{t}_x) = V\left(\sum_{i\in S}\frac{y_i - Rx_i}{np_i}\right)$$

  which is the variance of the HH estimator for the total of $y_i - Rx_i$

Model-assisted approach: difference estimator

- Let $\hat{y}_i = m(\mathbf{x}_i; \beta)$ be the proxy value for y_i (assume β is given to us for now), the difference estimator is

$$\hat{t}_d = \sum_{i \in \mathcal{U}} \hat{y}_i + \sum_{i \in S} \frac{y_i - \hat{y}_i}{np_i}$$

- $\hat{t}_d$ is unbiased
- The variance is

$$V(\hat{t}_d) = V\left(\sum_{i \in S} \frac{y_i - \hat{y}_i}{np_i}\right)$$

which is the variance of the HH estimator for the total of residual $y_i - \hat{y}_i$

Model-assisted approach: optimal coefficients for difference estimator

- We find β by minimizing $V(\hat{t}_d)$

$$\beta^* = \arg\min V(\hat{t}_d)$$

which involves some population quantities which needs to be estimated from samples

Model-assisted approach: regression estimator

- We find β by fitting a regression

$$\beta^* = \arg\min \sum_{i \in \mathcal{U}} (y_i - \beta \mathbf{x}_i)^2$$

  which involves some population quantities which needs to be estimated from samples
- If we use the same sample to estimate β^* and $\hat{t}$ then we introduce bias (similar to the ratio estimator), but we can still obtain approximate variance estimator
- The ratio estimator is a special case of regression estimator where $m(x_i; \beta) = \beta x_i$

Model-based approach

- The values y_i are assumed to be generated by a stochastic model, e.g., $\mathbb{E}[y_i|\mathbf{x}_i] = m(\mathbf{x}_i;\beta)$
- The selected sample S is treated as a constant and the sample values of y_i are random variables
- The total can be decomposed as

$$t = \sum_{i \in S} y_i + \underbrace{\sum_{i \notin S} y_i}_{t_{yr}}$$

and the model-based estimator is

$$\hat{t}_{MB} = \sum_{i \in S} y_i + \hat{t}_{yr}$$

for example, $\hat{t}_{yr} = \sum_{i \notin S} m(\mathbf{x}_i;\hat{\beta})$

Stratified estimator

- Population can be divided into strata and sampling takes place in each strata separately
- Assume there are H strata, then the stratified estimator is

$$\hat{t}_s = \sum_{h=1}^{H} \hat{t}_h$$

where $\hat{t}_h$ is the estimator for the h-th stratum

- We can use multinomial sampling for each strata with sample size n_h ($\sum n_h = n$) and $n_h \propto S_{y_h} = \frac{1}{N_h - 1} \sum_{i \in \mathcal{U}_h} (y_i - \bar{Y}_h)^2$

Combing multiple surveys

- Combining probability samples from multiple independent surveys
- Combining probability samples from multiple surveys with missingness: observe the design weight and auxiliary variable only in one survey and study variable in another survey
- Combining a probability sample with a non-probability sample (non-probability samples have unknown selection mechanisms and are typically biased, and they do not represent the population)

OPE

- OPE from survey sampling perspective:
  - $y_{s,a}$ is fixed but $I_{s,a}$ is random
  - Use auxiliary information to reduce MSE
- OPE from Monte Carlo perspective:
  - $r_{s,a}$ is a random variable drawn from P, estimate the mean under Q
  - Techniques to reduce MSE (e.g., clipping, blending)

Survey sampling and OPE estimators

- HH estimator = importance sampling estimator/IPS estimator
- Ratio estimator = weighted IS estimator/self-normalized IPS estimator
- Difference estimator = doubly robust estimator
- Optimal coefficient for difference estimator = more robust doubly robust estimator
- Model based estimator = direct method

Summary

- We provide a very brief overview of survey sampling and discuss its connection to OPE
- What can we borrow from survey sampling to improve OPE?
    - ideas of using auxiliary information
    - variance estimators
    - beyond multinomial design