Statistical Methods for Analysis with Missing Data

Lecture 12: inverse-probability weighting

Mauricio Sadinle

Department of Biostatistics

UNIVERSITY of WASHINGTON

Previous Lectures

Approaches to handling missing data covered so far

- Ad-hoc approaches (imputation, complete cases)
  - Not likelihood-based but we want to avoid them if possible
- Frequentist likelihood-based inference
  - Estimation via the EM algorithm
- Bayesian inference
  - Estimation via Gibbs sampling and data augmentation
- Multiple imputation
  - Versions: proper, MICE (others not covered here)
  - Congeniality requires being able to see overall procedure as approximation to Bayesian model (prior + likelihood)

Generally speaking, the last three approaches require a parametric model (likelihood function), either explicitly or implicitly

Today's Lecture[1]

- Inverse-probability weighting
  - Origins in survey sampling
  - Augmented inverse-probability weighting
  - Double robustness

[1] Acknowledgment: today's slides are partially based on materials developed by Gary Chan

Outline

Sampling from Finite Populations

Consider a *finite population* composed of N units

- We know a vector of *design variables* x_i for each unit $i = 1, \ldots, N$
- We want to learn the mean of an unknown quantity in the population $(y_1, \ldots, y_N)$,

$$\bar{y} = \sum_{i=1}^{N} y_i / N$$

- N is large, so measuring y_i on every unit is not feasible
- Idea: take a sample of units and measure y_i on them
- Remark: all x_i and y_i values are considered fixed quantities
  - Example: every household i has a number y_i that represents their income last year; that number is fixed, regardless of whether the value came to exist as the result of a random process

Sampling from Finite Populations

Consider a *finite population* composed of N units

- We know a vector of *design variables* x_i for each unit $i = 1, \ldots, N$
- We want to learn the mean of an unknown quantity in the population $(y_1, \ldots, y_N)$,

$$\bar{y} = \sum_{i=1}^{N} y_i / N$$

- N is large, so measuring y_i on every unit is not feasible
- Idea: take a sample of units and measure y_i on them
- Remark: all x_i and y_i values are considered fixed quantities
  - Example: every household i has a number y_i that represents their income last year; that number is fixed, regardless of whether the value came to exist as the result of a random process

Sampling from Finite Populations

Consider a *finite population* composed of N units

- We know a vector of *design variables* x_i for each unit $i = 1, \ldots, N$
- We want to learn the mean of an unknown quantity in the population $(y_1, \ldots, y_N)$,

$$\bar{y} = \sum_{i=1}^{N} y_i / N$$

- N is large, so measuring y_i on every unit is not feasible
- Idea: take a sample of units and measure y_i on them
- Remark: all x_i and y_i values are considered fixed quantities
  - Example: every household i has a number y_i that represents their income last year; that number is fixed, regardless of whether the value came to exist as the result of a random process

Sampling from Finite Populations

Consider a *finite population* composed of N units

- We know a vector of *design variables* x_i for each unit $i = 1, \ldots, N$
- We want to learn the mean of an unknown quantity in the population $(y_1, \ldots, y_N)$,

$$\bar{y} = \sum_{i=1}^{N} y_i / N$$

- N is large, so measuring y_i on every unit is not feasible
- Idea: take a sample of units and measure y_i on them
- Remark: all x_i and y_i values are considered fixed quantities
  - Example: every household i has a number y_i that represents their income last year; that number is fixed, regardless of whether the value came to exist as the result of a random process

Sampling from Finite Populations

Consider a *finite population* composed of N units

- We know a vector of *design variables* x_i for each unit $i = 1, \ldots, N$
- We want to learn the mean of an unknown quantity in the population $(y_1, \ldots, y_N)$,

$$\bar{y} = \sum_{i=1}^{N} y_i / N$$

- N is large, so measuring y_i on every unit is not feasible
- Idea: take a sample of units and measure y_i on them
- Remark: all x_i and y_i values are considered fixed quantities
  - Example: every household i has a number y_i that represents their income last year; that number is fixed, regardless of whether the value came to exist as the result of a random process

Sampling from Finite Populations

- Denote $R_i = 1$ if unit i is in the sample, 0 otherwise
- The *random vector* $R = (R_1, \ldots, R_N) \in \{0,1\}^N$ indicates the units included in the sample
- A *sample design* is a joint probability distribution

$$p(R_1, \ldots, R_N \mid x_1, \ldots, x_N),$$

giving the probability of selecting each possible sample

- The following two conditions need to hold:
  - The probabilities of inclusion depend on the design variables x_i only

$$p(R_1, \ldots, R_N \mid x_1, \ldots, x_N, y_1, \ldots, y_N) = p(R_1, \ldots, R_N \mid x_1, \ldots, x_N)$$

  - Every unit has a positive probability of being selected

$$\pi_i \equiv p(R_i = 1 \mid x_1, \ldots, x_N) > 0$$

- Note that the *sample design* and therefore the π_i's are *known*!

Sampling from Finite Populations

- Denote $R_i = 1$ if unit i is in the sample, 0 otherwise
- The *random vector* $R = (R_1, \ldots, R_N) \in \{0,1\}^N$ indicates the units included in the sample
- A *sample design* is a joint probability distribution

$$p(R_1, \ldots, R_N \mid x_1, \ldots, x_N),$$

giving the probability of selecting each possible sample

- The following two conditions need to hold:
  - The probabilities of inclusion depend on the design variables x_i only

$$p(R_1, \ldots, R_N \mid x_1, \ldots, x_N, y_1, \ldots, y_N) = p(R_1, \ldots, R_N \mid x_1, \ldots, x_N)$$

  - Every unit has a positive probability of being selected

$$\pi_i \equiv p(R_i = 1 \mid x_1, \ldots, x_N) > 0$$

- Note that the *sample design* and therefore the π_i's are *known*!

Sampling from Finite Populations

- Denote $R_i = 1$ if unit i is in the sample, 0 otherwise
- The *random vector* $R = (R_1, \ldots, R_N) \in \{0,1\}^N$ indicates the units included in the sample
- A *sample design* is a joint probability distribution

$$p(R_1, \ldots, R_N \mid x_1, \ldots, x_N),$$

giving the probability of selecting each possible sample

- The following two conditions need to hold:
  - The probabilities of inclusion depend on the design variables x_i only

$$p(R_1, \ldots, R_N \mid x_1, \ldots, x_N, y_1, \ldots, y_N) = p(R_1, \ldots, R_N \mid x_1, \ldots, x_N)$$

  - Every unit has a positive probability of being selected

$$\pi_i \equiv p(R_i = 1 \mid x_1, \ldots, x_N) > 0$$

- Note that the *sample design* and therefore the π_i's are *known*!

Sampling from Finite Populations

- Denote $R_i = 1$ if unit i is in the sample, 0 otherwise
- The *random vector* $R = (R_1, \ldots, R_N) \in \{0,1\}^N$ indicates the units included in the sample
- A *sample design* is a joint probability distribution

$$p(R_1, \ldots, R_N \mid x_1, \ldots, x_N),$$

giving the probability of selecting each possible sample

- The following two conditions need to hold:
  - The probabilities of inclusion depend on the design variables x_i only

$$p(R_1, \ldots, R_N \mid x_1, \ldots, x_N, y_1, \ldots, y_N) = p(R_1, \ldots, R_N \mid x_1, \ldots, x_N)$$

  - Every unit has a positive probability of being selected

$$\pi_i \equiv p(R_i = 1 \mid x_1, \ldots, x_N) > 0$$

- Note that the *sample design* and therefore the π_i's are *known*!

Sampling from Finite Populations

- Denote $R_i = 1$ if unit i is in the sample, 0 otherwise
- The *random vector* $R = (R_1, \ldots, R_N) \in \{0,1\}^N$ indicates the units included in the sample
- A *sample design* is a joint probability distribution

$$p(R_1, \ldots, R_N \mid x_1, \ldots, x_N),$$

giving the probability of selecting each possible sample

- The following two conditions need to hold:
  - The probabilities of inclusion depend on the design variables x_i only

$$p(R_1, \ldots, R_N \mid x_1, \ldots, x_N, y_1, \ldots, y_N) = p(R_1, \ldots, R_N \mid x_1, \ldots, x_N)$$

  - Every unit has a positive probability of being selected

$$\pi_i \equiv p(R_i = 1 \mid x_1, \ldots, x_N) > 0$$

- Note that the *sample design* and therefore the π_i's are *known*!

Sampling from Finite Populations

- Denote $R_i = 1$ if unit i is in the sample, 0 otherwise
- The *random vector* $R = (R_1, \ldots, R_N) \in \{0,1\}^N$ indicates the units included in the sample
- A *sample design* is a joint probability distribution

$$p(R_1, \ldots, R_N \mid x_1, \ldots, x_N),$$

  giving the probability of selecting each possible sample
- The following two conditions need to hold:
  - The probabilities of inclusion depend on the design variables x_i only

$$p(R_1, \ldots, R_N \mid x_1, \ldots, x_N, y_1, \ldots, y_N) = p(R_1, \ldots, R_N \mid x_1, \ldots, x_N)$$

  - Every unit has a positive probability of being selected

$$\pi_i \equiv p(R_i = 1 \mid x_1, \ldots, x_N) > 0$$

- Note that the *sample design* and therefore the π_i's are *known*!

Sampling from Finite Populations

Examples of sample designs

- Simple random sample: every sample of size n has the same probability of being selected, and therefore each unit has the same probability of being selected

$$\pi_i = n/N$$

- Stratified sample: say the x_i design variables define J strata $S_1, \ldots, S_J$ (e.g., different combinations of categorical variables)
  - Randomly sample n_j units from the N_j units in stratum j
  - For a unit $i \in S_j$, inclusion probability is:

$$\pi_i = \frac{n_j}{N_j}$$

  - Stratum proportion:

$$p_j = \frac{N_j}{N}$$

Sampling from Finite Populations

Examples of sample designs

- Simple random sample: every sample of size n has the same probability of being selected, and therefore each unit has the same probability of being selected

$$\pi_i = n/N$$

- Stratified sample: say the x_i design variables define J strata $S_1, \ldots, S_J$ (e.g., different combinations of categorical variables)
  - Randomly sample n_j units from the N_j units in stratum j
  - For a unit $i \in S_j$, inclusion probability is:

$$\pi_i = \frac{n_j}{N_j}$$

  - Stratum proportion:

$$p_j = \frac{N_j}{N}$$

Sampling from Finite Populations

Examples of sample designs

- Simple random sample: every sample of size n has the same probability of being selected, and therefore each unit has the same probability of being selected

$$\pi_i = n/N$$

- Stratified sample: say the x_i design variables define J strata $S_1, \ldots, S_J$ (e.g., different combinations of categorical variables)
  - Randomly sample n_j units from the N_j units in stratum j
  - For a unit $i \in S_j$, inclusion probability is:

$$\pi_i = \frac{n_j}{N_j}$$

  - Stratum proportion:

$$p_j = \frac{N_j}{N}$$

Sampling from Finite Populations

Examples of sample designs

- Simple random sample: every sample of size n has the same probability of being selected, and therefore each unit has the same probability of being selected

$$\pi_i = n/N$$

- Stratified sample: say the x_i design variables define J strata $S_1, \ldots, S_J$ (e.g., different combinations of categorical variables)
  - Randomly sample n_j units from the N_j units in stratum j
  - For a unit $i \in S_j$, inclusion probability is:

$$\pi_i = \frac{n_j}{N_j}$$

  - Stratum proportion:

$$p_j = \frac{N_j}{N}$$

Sampling from Finite Populations

Examples of sample designs

- Simple random sample: every sample of size n has the same probability of being selected, and therefore each unit has the same probability of being selected

$$\pi_i = n/N$$

- Stratified sample: say the x_i design variables define J strata $S_1, \ldots, S_J$ (e.g., different combinations of categorical variables)
  - Randomly sample n_j units from the N_j units in stratum j
  - For a unit $i \in S_j$, inclusion probability is:

$$\pi_i = \frac{n_j}{N_j}$$

  - Stratum proportion:

$$p_j = \frac{N_j}{N}$$

Estimation of Mean in Finite Population

Want to estimate $\bar{y}$

- Simple random sample:

$$\hat{\bar{y}} = \frac{1}{n}\sum_{i=1}^{N} R_i y_i = \frac{1}{N}\sum_{i=1}^{N}\frac{R_i y_i}{n/N} = \frac{1}{N}\sum_{i=1}^{N}\frac{R_i y_i}{\pi_i}$$

- Stratified sample:
  - First, compute sample means $\hat{\bar{y}}_j$ in each stratum, $j = 1, \ldots, J$.
  - Estimate $\bar{y}$ by taking a weighted average, weighting by strata proportions

$$\begin{aligned}
\hat{\bar{y}} &= \sum_{j=1}^{J}\frac{N_j}{N}\hat{\bar{y}}_j \\
&= \sum_{j=1}^{J}\frac{N_j}{N}\frac{1}{n_j}\sum_{i\in S_j} R_i y_i \\
&= \frac{1}{N}\sum_{i=1}^{N}\frac{R_i y_i}{\pi_i}
\end{aligned}$$

where $\pi_i = n_j/N_j$ if unit i is in stratum j

Estimation of Mean in Finite Population

Want to estimate $\bar{y}$

- Simple random sample:

$$\hat{\bar{y}} = \frac{1}{n}\sum_{i=1}^{N} R_i y_i = \frac{1}{N}\sum_{i=1}^{N} \frac{R_i y_i}{n/N} = \frac{1}{N}\sum_{i=1}^{N} \frac{R_i y_i}{\pi_i}$$

- Stratified sample:
  - First, compute sample means $\hat{\bar{y}}_j$ in each stratum, $j = 1, \ldots, J$.
  - Estimate $\bar{y}$ by taking a weighted average, weighting by strata proportions

$$\begin{aligned}
\hat{\bar{y}} &= \sum_{j=1}^{J} \frac{N_j}{N} \hat{\bar{y}}_j \\
&= \sum_{j=1}^{J} \frac{N_j}{N} \frac{1}{n_j} \sum_{i \in S_j} R_i y_i \\
&= \frac{1}{N} \sum_{i=1}^{N} \frac{R_i y_i}{\pi_i}
\end{aligned}$$

where $\pi_i = n_j/N_j$ if unit i is in stratum j

Estimation of Mean in Finite Population

Want to estimate $\bar{y}$

- Simple random sample:

$$\hat{\bar{y}} = \frac{1}{n}\sum_{i=1}^{N} R_i y_i = \frac{1}{N}\sum_{i=1}^{N} \frac{R_i y_i}{n/N} = \frac{1}{N}\sum_{i=1}^{N} \frac{R_i y_i}{\pi_i}$$

- Stratified sample:
  - First, compute sample means $\hat{\bar{y}}_j$ in each stratum, $j = 1, \ldots, J$.
  - Estimate $\bar{y}$ by taking a weighted average, weighting by strata proportions

$$\begin{aligned}
\hat{\bar{y}} &= \sum_{j=1}^{J} \frac{N_j}{N} \hat{\bar{y}}_j \\
&= \sum_{j=1}^{J} \frac{N_j}{N} \frac{1}{n_j} \sum_{i \in S_j} R_i y_i \\
&= \frac{1}{N} \sum_{i=1}^{N} \frac{R_i y_i}{\pi_i}
\end{aligned}$$

where $\pi_i = n_j/N_j$ if unit i is in stratum j

Estimation of Mean in Finite Population

Want to estimate $\bar{y}$

- Simple random sample:

$$\hat{\bar{y}} = \frac{1}{n}\sum_{i=1}^{N} R_i y_i = \frac{1}{N}\sum_{i=1}^{N} \frac{R_i y_i}{n/N} = \frac{1}{N}\sum_{i=1}^{N} \frac{R_i y_i}{\pi_i}$$

- Stratified sample:
  - First, compute sample means $\hat{\bar{y}}_j$ in each stratum, $j = 1, \ldots, J$.
  - Estimate $\bar{y}$ by taking a weighted average, weighting by strata proportions

$$\begin{aligned}
\hat{\bar{y}} &= \sum_{j=1}^{J} \frac{N_j}{N} \hat{\bar{y}}_j \\
&= \sum_{j=1}^{J} \frac{N_j}{N} \frac{1}{n_j} \sum_{i \in S_j} R_i y_i \\
&= \frac{1}{N}\sum_{i=1}^{N} \frac{R_i y_i}{\pi_i}
\end{aligned}$$

where $\pi_i = n_j/N_j$ if unit i is in stratum j

Estimation of Mean in Finite Population

Want to estimate $\bar{y}$

- Simple random sample:

$$\hat{\bar{y}} = \frac{1}{n}\sum_{i=1}^{N} R_i y_i = \frac{1}{N}\sum_{i=1}^{N} \frac{R_i y_i}{n/N} = \frac{1}{N}\sum_{i=1}^{N} \frac{R_i y_i}{\pi_i}$$

- Stratified sample:
  - First, compute sample means $\hat{\bar{y}}_j$ in each stratum, $j = 1, \ldots, J$.
  - Estimate $\bar{y}$ by taking a weighted average, weighting by strata proportions

$$\begin{aligned}
\hat{\bar{y}} &= \sum_{j=1}^{J} \frac{N_j}{N} \hat{\bar{y}}_j \\
&= \sum_{j=1}^{J} \frac{N_j}{N} \frac{1}{n_j} \sum_{i \in S_j} R_i y_i \\
&= \frac{1}{N}\sum_{i=1}^{N} \frac{R_i y_i}{\pi_i}
\end{aligned}$$

where $\pi_i = n_j/N_j$ if unit i is in stratum j

Estimation of Mean in Finite Population

Want to estimate $\bar{y}$

- Simple random sample:

$$\hat{\bar{y}} = \frac{1}{n}\sum_{i=1}^{N} R_i y_i = \frac{1}{N}\sum_{i=1}^{N} \frac{R_i y_i}{n/N} = \frac{1}{N}\sum_{i=1}^{N} \frac{R_i y_i}{\pi_i}$$

- Stratified sample:
  - First, compute sample means $\hat{\bar{y}}_j$ in each stratum, $j = 1, \ldots, J$.
  - Estimate $\bar{y}$ by taking a weighted average, weighting by strata proportions

$$\begin{aligned}
\hat{\bar{y}} &= \sum_{j=1}^{J} \frac{N_j}{N} \hat{\bar{y}}_j \\
&= \sum_{j=1}^{J} \frac{N_j}{N} \frac{1}{n_j} \sum_{i \in S_j} R_i y_i \\
&= \frac{1}{N}\sum_{i=1}^{N} \frac{R_i y_i}{\pi_i}
\end{aligned}$$

where $\pi_i = n_j/N_j$ if unit i is in stratum j

The Horvitz-Thompson Estimator[2]

The above ideas can be generalized

- Suppose each unit is included in the sample with probability $\pi_i > 0$, π_i being an arbitrary but known function of the design variables
- The *Horvitz-Thompson estimator* of the mean is

$$\bar{y}_{HT} = \frac{1}{N} \sum_{i=1}^{N} \frac{R_i y_i}{\pi_i}$$

- Nowadays also called *Inverse-Probability Weighted* (IPW) estimator, where R_i/π_i is seen as the weight of unit i in the sample

[2] Second author was Donovan J. Thompson, former chair of UW Biostat!
`https://www.jstor.org/stable/2280784`

The Horvitz-Thompson Estimator[2]

The above ideas can be generalized

- Suppose each unit is included in the sample with probability $\pi_i > 0$, π_i being an arbitrary but known function of the design variables
- The *Horvitz-Thompson estimator* of the mean is

$$\bar{y}_{HT} = \frac{1}{N} \sum_{i=1}^{N} \frac{R_i y_i}{\pi_i}$$

- Nowadays also called *Inverse-Probability Weighted* (IPW) estimator, where R_i/π_i is seen as the weight of unit i in the sample

[2] Second author was Donovan J. Thompson, former chair of UW Biostat!
https://www.jstor.org/stable/2280784

The Horvitz-Thompson Estimator[2]

The above ideas can be generalized

- Suppose each unit is included in the sample with probability $\pi_i > 0$, π_i being an arbitrary but known function of the design variables
- The *Horvitz-Thompson estimator* of the mean is

$$\bar{y}_{HT} = \frac{1}{N} \sum_{i=1}^{N} \frac{R_i y_i}{\pi_i}$$

- Nowadays also called *Inverse-Probability Weighted* (IPW) estimator, where R_i/π_i is seen as the weight of unit i in the sample

[2]Second author was Donovan J. Thompson, former chair of UW Biostat!
https://www.jstor.org/stable/2280784

The Horvitz-Thompson Estimator

- The Horvitz-Thompson estimator is appealing because it is unbiased

$$E_R(\bar{y}_{HT}) = E_R\left(\frac{1}{N}\sum_{i=1}^{N}\frac{R_i y_i}{\pi_i}\right)$$

$$= \frac{1}{N}\sum_{i=1}^{N}\frac{E_R(R_i)y_i}{\pi_i}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\frac{p(R_i = 1 \mid x_1, \ldots, x_N)y_i}{\pi_i}$$

$$= \frac{1}{N}\sum_{i=1}^{N} y_i$$

- In *survey sampling* the *randomization*-based approach to inference is mainstream, under which the only thing that is random is the sample selection (the R_i's)
- Take a course on survey sampling to learn more about this!

The Horvitz-Thompson Estimator

- The Horvitz-Thompson estimator is appealing because it is unbiased

$$
\begin{aligned}
E_R(\bar{y}_{HT}) &= E_R\left(\frac{1}{N}\sum_{i=1}^{N}\frac{R_i y_i}{\pi_i}\right) \\
&= \frac{1}{N}\sum_{i=1}^{N}\frac{E_R(R_i)y_i}{\pi_i} \\
&= \frac{1}{N}\sum_{i=1}^{N}\frac{p(R_i = 1 \mid x_1, \ldots, x_N)y_i}{\pi_i} \\
&= \frac{1}{N}\sum_{i=1}^{N} y_i
\end{aligned}
$$

- In *survey sampling* the *randomization*-based approach to inference is mainstream, under which the only thing that is random is the sample selection (the R_i's)
- Take a course on survey sampling to learn more about this!

The Horvitz-Thompson Estimator

- The Horvitz-Thompson estimator is appealing because it is unbiased

$$
\begin{aligned}
E_R(\bar{y}_{HT}) &= E_R\left(\frac{1}{N}\sum_{i=1}^{N}\frac{R_i y_i}{\pi_i}\right) \\
&= \frac{1}{N}\sum_{i=1}^{N}\frac{E_R(R_i)y_i}{\pi_i} \\
&= \frac{1}{N}\sum_{i=1}^{N}\frac{p(R_i = 1 \mid x_1, \ldots, x_N)y_i}{\pi_i} \\
&= \frac{1}{N}\sum_{i=1}^{N} y_i
\end{aligned}
$$

- In *survey sampling* the *randomization*-based approach to inference is mainstream, under which the only thing that is random is the sample selection (the R_i's)
- Take a course on survey sampling to learn more about this!

The Horvitz-Thompson Estimator

- The Horvitz-Thompson estimator is appealing because it is unbiased

$$
\begin{aligned}
E_R(\bar{y}_{HT}) &= E_R\left(\frac{1}{N}\sum_{i=1}^{N}\frac{R_i y_i}{\pi_i}\right) \\
&= \frac{1}{N}\sum_{i=1}^{N}\frac{E_R(R_i)y_i}{\pi_i} \\
&= \frac{1}{N}\sum_{i=1}^{N}\frac{p(R_i = 1 \mid x_1, \ldots, x_N)y_i}{\pi_i} \\
&= \frac{1}{N}\sum_{i=1}^{N} y_i
\end{aligned}
$$

- In *survey sampling* the *randomization*-based approach to inference is mainstream, under which the only thing that is random is the sample selection (the R_i's)
- Take a course on survey sampling to learn more about this!

The Horvitz-Thompson Estimator

- The Horvitz-Thompson estimator is appealing because it is unbiased

$$
\begin{aligned}
E_R(\bar{y}_{HT}) &= E_R\left(\frac{1}{N}\sum_{i=1}^{N}\frac{R_i y_i}{\pi_i}\right) \\
&= \frac{1}{N}\sum_{i=1}^{N}\frac{E_R(R_i)y_i}{\pi_i} \\
&= \frac{1}{N}\sum_{i=1}^{N}\frac{p(R_i = 1 \mid x_1, \ldots, x_N)y_i}{\pi_i} \\
&= \frac{1}{N}\sum_{i=1}^{N} y_i
\end{aligned}
$$

- In *survey sampling* the *randomization*-based approach to inference is mainstream, under which the only thing that is random is the sample selection (the R_i's)
- Take a course on survey sampling to learn more about this!

The Horvitz-Thompson Estimator

- The Horvitz-Thompson estimator is appealing because it is unbiased

$$
\begin{aligned}
E_R(\bar{y}_{HT}) &= E_R\left(\frac{1}{N}\sum_{i=1}^{N}\frac{R_i y_i}{\pi_i}\right) \\
&= \frac{1}{N}\sum_{i=1}^{N}\frac{E_R(R_i)y_i}{\pi_i} \\
&= \frac{1}{N}\sum_{i=1}^{N}\frac{p(R_i = 1 \mid x_1, \ldots, x_N)y_i}{\pi_i} \\
&= \frac{1}{N}\sum_{i=1}^{N} y_i
\end{aligned}
$$

- In *survey sampling* the *randomization*-based approach to inference is mainstream, under which the only thing that is random is the sample selection (the R_i's)
- Take a course on survey sampling to learn more about this!

Basu's Elephant

Many people criticize the Horvitz-Thompson estimator, in particular Debabrata Basu (1971): *An essay on the logical foundations of survey sampling, Part I.*

- Circus owner planning to ship 50 elephants and needs an estimate of the total weight
- She plans to weight just one elephant: Sambo, the middle-sized elephant, and take $50y_{Sambo}$ (y_{Sambo} is the weight of Sambo) to be an estimate of the total weight
- Circus' statistician is horrified because the circus owner gives 0 probability for sampling other elephants
- Statistician developed the following plan: 99% prob. of selecting Sambo; and equal probability to each of other 49 elephants
- As expected, Sambo is selected so the owner should be happy

Basu's Elephant

Many people criticize the Horvitz-Thompson estimator, in particular Debabrata Basu (1971): *An essay on the logical foundations of survey sampling, Part I.*

- Circus owner planning to ship 50 elephants and needs an estimate of the total weight
- She plans to weight just one elephant: Sambo, the middle-sized elephant, and take $50y_{Sambo}$ (y_{Sambo} is the weight of Sambo) to be an estimate of the total weight
- Circus' statistician is horrified because the circus owner gives 0 probability for sampling other elephants
- Statistician developed the following plan: 99% prob. of selecting Sambo; and equal probability to each of other 49 elephants
- As expected, Sambo is selected so the owner should be happy

Basu's Elephant

Many people criticize the Horvitz-Thompson estimator, in particular Debabrata Basu (1971): *An essay on the logical foundations of survey sampling, Part I.*

- Circus owner planning to ship 50 elephants and needs an estimate of the total weight
- She plans to weight just one elephant: Sambo, the middle-sized elephant, and take $50y_{Sambo}$ (y_{Sambo} is the weight of Sambo) to be an estimate of the total weight
- Circus' statistician is horrified because the circus owner gives 0 probability for sampling other elephants
- Statistician developed the following plan: 99% prob. of selecting Sambo; and equal probability to each of other 49 elephants
- As expected, Sambo is selected so the owner should be happy

Basu's Elephant

Many people criticize the Horvitz-Thompson estimator, in particular Debabrata Basu (1971): *An essay on the logical foundations of survey sampling, Part I.*

- Circus owner planning to ship 50 elephants and needs an estimate of the total weight
- She plans to weight just one elephant: Sambo, the middle-sized elephant, and take $50y_{Sambo}$ (y_{Sambo} is the weight of Sambo) to be an estimate of the total weight
- Circus' statistician is horrified because the circus owner gives 0 probability for sampling other elephants
- Statistician developed the following plan: 99% prob. of selecting Sambo; and equal probability to each of other 49 elephants
- As expected, Sambo is selected so the owner should be happy

Basu's Elephant

Many people criticize the Horvitz-Thompson estimator, in particular Debabrata Basu (1971): *An essay on the logical foundations of survey sampling, Part I.*

- Circus owner planning to ship 50 elephants and needs an estimate of the total weight
- She plans to weight just one elephant: Sambo, the middle-sized elephant, and take $50y_{Sambo}$ (y_{Sambo} is the weight of Sambo) to be an estimate of the total weight
- Circus' statistician is horrified because the circus owner gives 0 probability for sampling other elephants
- Statistician developed the following plan: 99% prob. of selecting Sambo; and equal probability to each of other 49 elephants
- As expected, Sambo is selected so the owner should be happy

Basu's Elephant

Many people criticize the Horvitz-Thompson estimator, in particular Debabrata Basu (1971): *An essay on the logical foundations of survey sampling, Part I.*

- Circus owner planning to ship 50 elephants and needs an estimate of the total weight
- She plans to weight just one elephant: Sambo, the middle-sized elephant, and take $50y_{Sambo}$ (y_{Sambo} is the weight of Sambo) to be an estimate of the total weight
- Circus' statistician is horrified because the circus owner gives 0 probability for sampling other elephants
- Statistician developed the following plan: 99% prob. of selecting Sambo; and equal probability to each of other 49 elephants
- As expected, Sambo is selected so the owner should be happy

Basu's Elephant

- Circus owner asked if the estimated total weight is $50y_{Sambo}$ since Sambo was sampled
- Statistician said no; IPW estimate is

$$50 \times \frac{1}{50} \times \frac{1}{0.99} y_{Sambo} = \frac{100}{99} y_{Sambo}$$

- Owner asked what if the largest elephant, Jumbo, had been selected
- Statistician answered: IPW estimate would be

$$50 \times \frac{1}{50} \times \frac{1}{0.01/49} y_{Jumbo} = 4900 y_{Jumbo}$$

- The statistician was immediately fired!

Basu's Elephant

- Circus owner asked if the estimated total weight is $50 y_{Sambo}$ since Sambo was sampled
- Statistician said no; IPW estimate is

$$50 \times \frac{1}{50} \times \frac{1}{0.99} y_{Sambo} = \frac{100}{99} y_{Sambo}$$

- Owner asked what if the largest elephant, Jumbo, had been selected
- Statistician answered: IPW estimate would be

$$50 \times \frac{1}{50} \times \frac{1}{0.01/49} y_{Jumbo} = 4900 y_{Jumbo}$$

- The statistician was immediately fired!

Basu's Elephant

- Circus owner asked if the estimated total weight is $50 y_{Sambo}$ since Sambo was sampled
- Statistician said no; IPW estimate is

$$50 \times \frac{1}{50} \times \frac{1}{0.99} y_{Sambo} = \frac{100}{99} y_{Sambo}$$

- Owner asked what if the largest elephant, Jumbo, had been selected
- Statistician answered: IPW estimate would be

$$50 \times \frac{1}{50} \times \frac{1}{0.01/49} y_{Jumbo} = 4900 y_{Jumbo}$$

- The statistician was immediately fired!

Basu's Elephant

- Circus owner asked if the estimated total weight is $50 y_{Sambo}$ since Sambo was sampled
- Statistician said no; IPW estimate is

$$50 \times \frac{1}{50} \times \frac{1}{0.99} y_{Sambo} = \frac{100}{99} y_{Sambo}$$

- Owner asked what if the largest elephant, Jumbo, had been selected
- Statistician answered: IPW estimate would be

$$50 \times \frac{1}{50} \times \frac{1}{0.01/49} y_{Jumbo} = 4900 y_{Jumbo}$$

- The statistician was immediately fired!

Basu's Elephant

- Circus owner asked if the estimated total weight is $50y_{Sambo}$ since Sambo was sampled
- Statistician said no; IPW estimate is

$$50 \times \frac{1}{50} \times \frac{1}{0.99} y_{Sambo} = \frac{100}{99} y_{Sambo}$$

- Owner asked what if the largest elephant, Jumbo, had been selected
- Statistician answered: IPW estimate would be

$$50 \times \frac{1}{50} \times \frac{1}{0.01/49} y_{Jumbo} = 4900 y_{Jumbo}$$

- The statistician was immediately fired!

Basu's Elephant

- What's going on with Basu's elephant example?
- Sample size is too small
- Selection probabilities are too extreme
- Huge standard error by having such extreme selection probabilities
- Over repeated samples, the average estimate is close to the truth, but each estimate can be far off

Basu's Elephant

- What's going on with Basu's elephant example?
- Sample size is too small
- Selection probabilities are too extreme
- Huge standard error by having such extreme selection probabilities
- Over repeated samples, the average estimate is close to the truth, but each estimate can be far off

Basu's Elephant

- What's going on with Basu's elephant example?
- Sample size is too small
- Selection probabilities are too extreme
- Huge standard error by having such extreme selection probabilities
- Over repeated samples, the average estimate is close to the truth, but each estimate can be far off

Basu's Elephant

- What's going on with Basu's elephant example?
- Sample size is too small
- Selection probabilities are too extreme
- Huge standard error by having such extreme selection probabilities
- Over repeated samples, the average estimate is close to the truth, but each estimate can be far off

Basu's Elephant

- What's going on with Basu's elephant example?
- Sample size is too small
- Selection probabilities are too extreme
- Huge standard error by having such extreme selection probabilities
- Over repeated samples, the average estimate is close to the truth, but each estimate can be far off

Outline

Infinite Population Set-Up

The usual set-up in this class consists of an *infinite population* represented by the joint distribution of a vector of random variables. In particular today we will consider:

- X: vector of always observed random variables
- Y: random variable subject to nonresponse
- R: indicator of whether Y is observed
- Note that the infinite population is the full-data distribution with density

$$p(x, y, r) = p(x, y)p(r \mid x, y)$$

- The data are n i.i.d. copies of (X, Y_R, R)

Infinite Population Set-Up

The usual set-up in this class consists of an *infinite population* represented by the joint distribution of a vector of random variables. In particular today we will consider:

- X: vector of always observed random variables
- Y: random variable subject to nonresponse
- R: indicator of whether Y is observed
- Note that the infinite population is the full-data distribution with density

$$p(x, y, r) = p(x, y)p(r \mid x, y)$$

- The data are n i.i.d. copies of (X, Y_R, R)

Infinite Population Set-Up

The usual set-up in this class consists of an *infinite population* represented by the joint distribution of a vector of random variables. In particular today we will consider:

- X: vector of always observed random variables
- Y: random variable subject to nonresponse
- R: indicator of whether Y is observed
- Note that the infinite population is the full-data distribution with density

$$p(x, y, r) = p(x, y)p(r \mid x, y)$$

- The data are n i.i.d. copies of (X, Y_R, R)

Infinite Population Set-Up

- Say we want to estimate the mean of Y

$$\mu = E(Y) = \int yp(y)dy$$

- Assume MAR, which in this case is $R \perp\!\!\!\perp Y \mid X$
- Define the *propensity score* to be

$$\pi(x) = p(R = 1 \mid x)$$

Inverse-Probability Weighting

- The inverse-probability weighted (IPW) estimator of the mean is

$$\hat{\mu}^{ipw0} = \frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\pi(X_i)}Y_i$$

- Which, again, is unbiased

$$\begin{aligned}
E(\hat{\mu}^{ipw0}) &= \frac{1}{n}\sum_{i=1}^{n} E\left(\frac{R_i}{\pi(X_i)}Y_i\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} E_{X_i}\left(E_{R_i,Y_i}\left(\frac{R_i}{\pi(X_i)}Y_i \mid X_i\right)\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} E_{X_i}\left(\frac{E(R_i \mid X_i)E(Y_i \mid X_i)}{\pi(X_i)}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} E_{X_i}\left(E(Y_i \mid X_i)\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu
\end{aligned}$$

Inverse-Probability Weighting

- The inverse-probability weighted (IPW) estimator of the mean is

$$\hat{\mu}^{ipw0} = \frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\pi(X_i)}Y_i$$

- Which, again, is unbiased

$$E(\hat{\mu}^{ipw0}) = \frac{1}{n}\sum_{i=1}^{n}E\left(\frac{R_i}{\pi(X_i)}Y_i\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}E_{X_i}\left(E_{R_i,Y_i}\left(\frac{R_i}{\pi(X_i)}Y_i \mid X_i\right)\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}E_{X_i}\left(\frac{E(R_i \mid X_i)E(Y_i \mid X_i)}{\pi(X_i)}\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}E_{X_i}\left(E(Y_i \mid X_i)\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}E(Y_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu$$

Inverse-Probability Weighting

- The inverse-probability weighted (IPW) estimator of the mean is

$$\hat{\mu}^{ipw0} = \frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\pi(X_i)}Y_i$$

- Which, again, is unbiased

$$\begin{aligned}
E(\hat{\mu}^{ipw0}) &= \frac{1}{n}\sum_{i=1}^{n} E\left(\frac{R_i}{\pi(X_i)}Y_i\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} E_{X_i}\left(E_{R_i,Y_i}\left(\frac{R_i}{\pi(X_i)}Y_i \mid X_i\right)\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} E_{X_i}\left(\frac{E(R_i \mid X_i)E(Y_i \mid X_i)}{\pi(X_i)}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} E_{X_i}\left(E(Y_i \mid X_i)\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu
\end{aligned}$$

Inverse-Probability Weighting

- The inverse-probability weighted (IPW) estimator of the mean is

$$\hat{\mu}^{ipw0} = \frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\pi(X_i)}Y_i$$

- Which, again, is unbiased

$$\begin{aligned}
E(\hat{\mu}^{ipw0}) &= \frac{1}{n}\sum_{i=1}^{n} E\left(\frac{R_i}{\pi(X_i)}Y_i\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} E_{X_i}\left(E_{R_i,Y_i}\left(\frac{R_i}{\pi(X_i)}Y_i \mid X_i\right)\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} E_{X_i}\left(\frac{E(R_i \mid X_i)E(Y_i \mid X_i)}{\pi(X_i)}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} E_{X_i}\left(E(Y_i \mid X_i)\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu
\end{aligned}$$

Inverse-Probability Weighting

- The inverse-probability weighted (IPW) estimator of the mean is

$$\hat{\mu}^{ipw0} = \frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\pi(X_i)}Y_i$$

- Which, again, is unbiased

$$\begin{aligned}
E(\hat{\mu}^{ipw0}) &= \frac{1}{n}\sum_{i=1}^{n} E\left(\frac{R_i}{\pi(X_i)}Y_i\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} E_{X_i}\left(E_{R_i,Y_i}\left(\frac{R_i}{\pi(X_i)}Y_i \mid X_i\right)\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} E_{X_i}\left(\frac{E(R_i \mid X_i)E(Y_i \mid X_i)}{\pi(X_i)}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} E_{X_i}\left(E(Y_i \mid X_i)\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu
\end{aligned}$$

Inverse-Probability Weighting

- The inverse-probability weighted (IPW) estimator of the mean is

$$\hat{\mu}^{ipw0} = \frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\pi(X_i)}Y_i$$

- Which, again, is unbiased

$$\begin{aligned}
E(\hat{\mu}^{ipw0}) &= \frac{1}{n}\sum_{i=1}^{n} E\left(\frac{R_i}{\pi(X_i)}Y_i\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} E_{X_i}\left(E_{R_i,Y_i}\left(\frac{R_i}{\pi(X_i)}Y_i \mid X_i\right)\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} E_{X_i}\left(\frac{E(R_i \mid X_i)E(Y_i \mid X_i)}{\pi(X_i)}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} E_{X_i}\left(E(Y_i \mid X_i)\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu
\end{aligned}$$

Inverse-Probability Weighting

- Unlike in the context of a well-designed survey, $\pi(x)$ is unknown and needs to be estimated
- Estimate the propensity scores as $\pi(x; \hat{\psi})$, e.g. using a logistic regression, and use

$$\hat{\mu}^{ipw} = \frac{1}{n} \sum_{i=1}^{n} \frac{R_i}{\pi(X_i; \hat{\psi})} Y_i$$

- This estimator is consistent if $\pi(x; \hat{\psi})$ is correctly specified HW4
- IPW was re-introduced by James Robins, Andrea Rotnitzky and Lue Ping Zhao (JASA, 1994)[3] for parameter estimation in semiparametric models

[3] https://www.jstor.org/stable/2290910

Inverse-Probability Weighting

- Unlike in the context of a well-designed survey, $\pi(x)$ is unknown and needs to be estimated
- Estimate the propensity scores as $\pi(x; \hat{\psi})$, e.g. using a logistic regression, and use

$$\hat{\mu}^{ipw} = \frac{1}{n} \sum_{i=1}^{n} \frac{R_i}{\pi(X_i; \hat{\psi})} Y_i$$

- This estimator is consistent if $\pi(x; \hat{\psi})$ is correctly specified HW4
- IPW was re-introduced by James Robins, Andrea Rotnitzky and Lue Ping Zhao (JASA, 1994)[3] for parameter estimation in semiparametric models

[3] https://www.jstor.org/stable/2290910

Inverse-Probability Weighting

- Unlike in the context of a well-designed survey, $\pi(x)$ is unknown and needs to be estimated
- Estimate the propensity scores as $\pi(x; \hat{\psi})$, e.g. using a logistic regression, and use

$$\hat{\mu}^{ipw} = \frac{1}{n} \sum_{i=1}^{n} \frac{R_i}{\pi(X_i; \hat{\psi})} Y_i$$

- This estimator is consistent if $\pi(x; \hat{\psi})$ is correctly specified HW4
- IPW was re-introduced by James Robins, Andrea Rotnitzky and Lue Ping Zhao (JASA, 1994)[3] for parameter estimation in semiparametric models

[3] https://www.jstor.org/stable/2290910

Inverse-Probability Weighting

- Unlike in the context of a well-designed survey, $\pi(x)$ is unknown and needs to be estimated
- Estimate the propensity scores as $\pi(x; \hat{\psi})$, e.g. using a logistic regression, and use

$$\hat{\mu}^{ipw} = \frac{1}{n} \sum_{i=1}^{n} \frac{R_i}{\pi(X_i; \hat{\psi})} Y_i$$

- This estimator is consistent if $\pi(x; \hat{\psi})$ is correctly specified HW4
- IPW was re-introduced by James Robins, Andrea Rotnitzky and Lue Ping Zhao (JASA, 1994)[3] for parameter estimation in semiparametric models

[3] https://www.jstor.org/stable/2290910

Outline

Augmented IPW

- IPW is straightforward to implement but can be quite inefficient (uses only complete cases)
- Consistency of IPW relies on the correctness of model assumptions for missing data mechanism (propensity score)
- The *augmented IPW* (AIPW) estimator of the population mean is

$$\hat{\mu}^{aipw} = \frac{1}{n}\sum_{i=1}^{n}\frac{R_i Y_i}{\pi(X_i;\hat{\psi})} - \frac{1}{n}\sum_{i=1}^{n}\frac{(R_i - \pi(X_i;\hat{\psi}))}{\pi(X_i;\hat{\psi})} m(X_i;\hat{\xi})$$

  where $m(x;\hat{\xi})$ is an estimate of $E(Y \mid x)$ among the complete cases, since under MAR $E(Y \mid x) = E(Y \mid x, R = 1)$
- AIPW estimators were introduced by Robins, Rotnitzky and Zhao (1994)
- In the survey sampling world these are called *model-assisted survey estimators* (Särndal, Swensson and Wretman, 1992, Springer)

Augmented IPW

- IPW is straightforward to implement but can be quite inefficient (uses only complete cases)
- Consistency of IPW relies on the correctness of model assumptions for missing data mechanism (propensity score)
- The *augmented IPW* (AIPW) estimator of the population mean is

$$\hat{\mu}^{aipw} = \frac{1}{n}\sum_{i=1}^{n}\frac{R_i Y_i}{\pi(X_i;\hat{\psi})} - \frac{1}{n}\sum_{i=1}^{n}\frac{(R_i - \pi(X_i;\hat{\psi}))}{\pi(X_i;\hat{\psi})} m(X_i;\hat{\xi})$$

  where $m(x;\hat{\xi})$ is an estimate of $E(Y \mid x)$ among the complete cases, since under MAR $E(Y \mid x) = E(Y \mid x, R = 1)$
- AIPW estimators were introduced by Robins, Rotnitzky and Zhao (1994)
- In the survey sampling world these are called *model-assisted survey estimators* (Särndal, Swensson and Wretman, 1992, Springer)

Augmented IPW

- IPW is straightforward to implement but can be quite inefficient (uses only complete cases)
- Consistency of IPW relies on the correctness of model assumptions for missing data mechanism (propensity score)
- The *augmented IPW* (AIPW) estimator of the population mean is

$$\hat{\mu}^{aipw} = \frac{1}{n}\sum_{i=1}^{n}\frac{R_i Y_i}{\pi(X_i;\hat{\psi})} - \frac{1}{n}\sum_{i=1}^{n}\frac{(R_i - \pi(X_i;\hat{\psi}))}{\pi(X_i;\hat{\psi})} m(X_i;\hat{\xi})$$

  where $m(x;\hat{\xi})$ is an estimate of $E(Y \mid x)$ among the complete cases, since under MAR $E(Y \mid x) = E(Y \mid x, R = 1)$

- AIPW estimators were introduced by Robins, Rotnitzky and Zhao (1994)
- In the survey sampling world these are called *model-assisted survey estimators* (Särndal, Swensson and Wretman, 1992, Springer)

Augmented IPW

- IPW is straightforward to implement but can be quite inefficient (uses only complete cases)
- Consistency of IPW relies on the correctness of model assumptions for missing data mechanism (propensity score)
- The *augmented IPW* (AIPW) estimator of the population mean is

$$\hat{\mu}^{aipw} = \frac{1}{n}\sum_{i=1}^{n}\frac{R_i Y_i}{\pi(X_i;\hat{\psi})} - \frac{1}{n}\sum_{i=1}^{n}\frac{(R_i - \pi(X_i;\hat{\psi}))}{\pi(X_i;\hat{\psi})} m(X_i;\hat{\xi})$$

  where $m(x;\hat{\xi})$ is an estimate of $E(Y \mid x)$ among the complete cases, since under MAR $E(Y \mid x) = E(Y \mid x, R = 1)$

- AIPW estimators were introduced by Robins, Rotnitzky and Zhao (1994)
- In the survey sampling world these are called *model-assisted survey estimators* (Särndal, Swensson and Wretman, 1992, Springer)

Augmented IPW

- IPW is straightforward to implement but can be quite inefficient (uses only complete cases)
- Consistency of IPW relies on the correctness of model assumptions for missing data mechanism (propensity score)
- The *augmented IPW* (AIPW) estimator of the population mean is

$$\hat{\mu}^{aipw} = \frac{1}{n}\sum_{i=1}^{n}\frac{R_i Y_i}{\pi(X_i;\hat{\psi})} - \frac{1}{n}\sum_{i=1}^{n}\frac{(R_i - \pi(X_i;\hat{\psi}))}{\pi(X_i;\hat{\psi})} m(X_i;\hat{\xi})$$

  where $m(x;\hat{\xi})$ is an estimate of $E(Y \mid x)$ among the complete cases, since under MAR $E(Y \mid x) = E(Y \mid x, R = 1)$
- AIPW estimators were introduced by Robins, Rotnitzky and Zhao (1994)
- In the survey sampling world these are called *model-assisted survey estimators* (Särndal, Swensson and Wretman, 1992, Springer)

Augmented IPW

- AIPW usually has a smaller standard error than IPW
- AIPW enjoys a *double robustness* property: it is consistent for μ if either
  - The propensity score model $\pi(x;\psi)$ is correctly specified
  - The conditional mean model $m(x;\xi)$ is correctly specified
- HW4: show that if $\hat{\psi} \xrightarrow{p} \psi^*$ and $\hat{\xi} \xrightarrow{p} \xi^*$ then

$$\hat{\mu}^{aipw} \xrightarrow{p} E\left(\frac{RY}{\pi(X;\psi^*)} - \frac{(R-\pi(X;\psi^*))}{\pi(X;\psi^*)} m(X;\xi^*)\right)$$

- HW4: show that the above expression can be written as

$$\mu + E_X\left[E_R\left(\frac{(R-\pi(X;\psi^*))}{\pi(X;\psi^*)} \mid X\right) E_Y(Y - m(X;\xi^*) \mid X)\right]$$

- We conclude that $\hat{\mu}^{aipw} \xrightarrow{p} \mu$ when either
  - $\pi(X;\psi^*) = p(R=1 \mid X)$
  - $m(X;\xi^*) = E(Y \mid X)$

Augmented IPW

- AIPW usually has a smaller standard error than IPW
- AIPW enjoys a *double robustness* property: it is consistent for μ if either
  - The propensity score model $\pi(x;\psi)$ is correctly specified
  - The conditional mean model $m(x;\xi)$ is correctly specified
- HW4: show that if $\hat{\psi} \xrightarrow{p} \psi^*$ and $\hat{\xi} \xrightarrow{p} \xi^*$ then

$$\hat{\mu}^{aipw} \xrightarrow{p} E\left(\frac{RY}{\pi(X;\psi^*)} - \frac{(R-\pi(X;\psi^*))}{\pi(X;\psi^*)} m(X;\xi^*)\right)$$

- HW4: show that the above expression can be written as

$$\mu + E_X\left[E_R\left(\frac{(R-\pi(X;\psi^*))}{\pi(X;\psi^*)} \mid X\right) E_Y(Y - m(X;\xi^*) \mid X)\right]$$

- We conclude that $\hat{\mu}^{aipw} \xrightarrow{p} \mu$ when either
  - $\pi(X;\psi^*) = p(R=1 \mid X)$
  - $m(X;\xi^*) = E(Y \mid X)$

Augmented IPW

- AIPW usually has a smaller standard error than IPW
- AIPW enjoys a *double robustness* property: it is consistent for μ if either
  - The propensity score model $\pi(x;\psi)$ is correctly specified
  - The conditional mean model $m(x;\xi)$ is correctly specified
- HW4: show that if $\hat{\psi} \xrightarrow{p} \psi^*$ and $\hat{\xi} \xrightarrow{p} \xi^*$ then

$$\hat{\mu}^{aipw} \xrightarrow{p} E\left(\frac{RY}{\pi(X;\psi^*)} - \frac{(R-\pi(X;\psi^*))}{\pi(X;\psi^*)} m(X;\xi^*)\right)$$

- HW4: show that the above expression can be written as

$$\mu + E_X\left[E_R\left(\frac{(R-\pi(X;\psi^*))}{\pi(X;\psi^*)} \mid X\right) E_Y(Y - m(X;\xi^*) \mid X)\right]$$

- We conclude that $\hat{\mu}^{aipw} \xrightarrow{p} \mu$ when either
  - $\pi(X;\psi^*) = p(R=1 \mid X)$
  - $m(X;\xi^*) = E(Y \mid X)$

Augmented IPW

- AIPW usually has a smaller standard error than IPW
- AIPW enjoys a *double robustness* property: it is consistent for μ if either
  - The propensity score model $\pi(x;\psi)$ is correctly specified
  - The conditional mean model $m(x;\xi)$ is correctly specified
- HW4: show that if $\hat{\psi} \xrightarrow{p} \psi^*$ and $\hat{\xi} \xrightarrow{p} \xi^*$ then

$$\hat{\mu}^{aipw} \xrightarrow{p} E\left(\frac{RY}{\pi(X;\psi^*)} - \frac{(R-\pi(X;\psi^*))}{\pi(X;\psi^*)} m(X;\xi^*)\right)$$

- HW4: show that the above expression can be written as

$$\mu + E_X\left[E_R\left(\frac{(R-\pi(X;\psi^*))}{\pi(X;\psi^*)} \mid X\right) E_Y(Y - m(X;\xi^*) \mid X)\right]$$

- We conclude that $\hat{\mu}^{aipw} \xrightarrow{p} \mu$ when either
  - $\pi(X;\psi^*) = p(R=1 \mid X)$
  - $m(X;\xi^*) = E(Y \mid X)$

Augmented IPW

- AIPW usually has a smaller standard error than IPW
- AIPW enjoys a *double robustness* property: it is consistent for μ if either
  - The propensity score model $\pi(x;\psi)$ is correctly specified
  - The conditional mean model $m(x;\xi)$ is correctly specified
- HW4: show that if $\hat{\psi} \xrightarrow{p} \psi^*$ and $\hat{\xi} \xrightarrow{p} \xi^*$ then

$$\hat{\mu}^{aipw} \xrightarrow{p} E\left(\frac{RY}{\pi(X;\psi^*)} - \frac{(R-\pi(X;\psi^*))}{\pi(X;\psi^*)} m(X;\xi^*)\right)$$

- HW4: show that the above expression can be written as

$$\mu + E_X\left[E_R\left(\frac{(R-\pi(X;\psi^*))}{\pi(X;\psi^*)} \mid X\right) E_Y(Y - m(X;\xi^*) \mid X)\right]$$

- We conclude that $\hat{\mu}^{aipw} \xrightarrow{p} \mu$ when either
  - $\pi(X;\psi^*) = p(R=1 \mid X)$
  - $m(X;\xi^*) = E(Y \mid X)$

Augmented IPW

- AIPW usually has a smaller standard error than IPW
- AIPW enjoys a *double robustness* property: it is consistent for μ if either
  - The propensity score model $\pi(x;\psi)$ is correctly specified
  - The conditional mean model $m(x;\xi)$ is correctly specified
- HW4: show that if $\hat{\psi} \xrightarrow{p} \psi^*$ and $\hat{\xi} \xrightarrow{p} \xi^*$ then

$$\hat{\mu}^{aipw} \xrightarrow{p} E\left(\frac{RY}{\pi(X;\psi^*)} - \frac{(R-\pi(X;\psi^*))}{\pi(X;\psi^*)} m(X;\xi^*)\right)$$

- HW4: show that the above expression can be written as

$$\mu + E_X\left[E_R\left(\frac{(R-\pi(X;\psi^*))}{\pi(X;\psi^*)} \mid X\right) E_Y(Y - m(X;\xi^*) \mid X)\right]$$

- We conclude that $\hat{\mu}^{aipw} \xrightarrow{p} \mu$ when either
  - $\pi(X;\psi^*) = p(R=1 \mid X)$
  - $m(X;\xi^*) = E(Y \mid X)$

Augmented IPW

- AIPW usually has a smaller standard error than IPW
- AIPW enjoys a *double robustness* property: it is consistent for μ if either
  - The propensity score model $\pi(x;\psi)$ is correctly specified
  - The conditional mean model $m(x;\xi)$ is correctly specified
- HW4: show that if $\hat{\psi} \xrightarrow{p} \psi^*$ and $\hat{\xi} \xrightarrow{p} \xi^*$ then

$$\hat{\mu}^{aipw} \xrightarrow{p} E\left(\frac{RY}{\pi(X;\psi^*)} - \frac{(R-\pi(X;\psi^*))}{\pi(X;\psi^*)} m(X;\xi^*)\right)$$

- HW4: show that the above expression can be written as

$$\mu + E_X\left[E_R\left(\frac{(R-\pi(X;\psi^*))}{\pi(X;\psi^*)} \mid X\right) E_Y(Y - m(X;\xi^*) \mid X)\right]$$

- We conclude that $\hat{\mu}^{aipw} \xrightarrow{p} \mu$ when either
  - $\pi(X;\psi^*) = p(R=1 \mid X)$
  - $m(X;\xi^*) = E(Y \mid X)$

Augmented IPW

- AIPW usually has a smaller standard error than IPW
- AIPW enjoys a *double robustness* property: it is consistent for μ if either
  - The propensity score model $\pi(x;\psi)$ is correctly specified
  - The conditional mean model $m(x;\xi)$ is correctly specified
- HW4: show that if $\hat{\psi} \xrightarrow{p} \psi^*$ and $\hat{\xi} \xrightarrow{p} \xi^*$ then

$$\hat{\mu}^{aipw} \xrightarrow{p} E\left(\frac{RY}{\pi(X;\psi^*)} - \frac{(R-\pi(X;\psi^*))}{\pi(X;\psi^*)} m(X;\xi^*)\right)$$

- HW4: show that the above expression can be written as

$$\mu + E_X\left[E_R\left(\frac{(R-\pi(X;\psi^*))}{\pi(X;\psi^*)} \mid X\right) E_Y(Y - m(X;\xi^*) \mid X)\right]$$

- We conclude that $\hat{\mu}^{aipw} \xrightarrow{p} \mu$ when either
  - $\pi(X;\psi^*) = p(R=1 \mid X)$
  - $m(X;\xi^*) = E(Y \mid X)$

Augmented IPW

- AIPW usually has a smaller standard error than IPW
- AIPW enjoys a *double robustness* property: it is consistent for μ if either
  - The propensity score model $\pi(x;\psi)$ is correctly specified
  - The conditional mean model $m(x;\xi)$ is correctly specified
- HW4: show that if $\hat{\psi} \xrightarrow{p} \psi^*$ and $\hat{\xi} \xrightarrow{p} \xi^*$ then

$$\hat{\mu}^{aipw} \xrightarrow{p} E\left(\frac{RY}{\pi(X;\psi^*)} - \frac{(R-\pi(X;\psi^*))}{\pi(X;\psi^*)} m(X;\xi^*)\right)$$

- HW4: show that the above expression can be written as

$$\mu + E_X\left[E_R\left(\frac{(R-\pi(X;\psi^*))}{\pi(X;\psi^*)} \mid X\right) E_Y(Y - m(X;\xi^*) \mid X)\right]$$

- We conclude that $\hat{\mu}^{aipw} \xrightarrow{p} \mu$ when either
  - $\pi(X;\psi^*) = p(R=1 \mid X)$
  - $m(X;\xi^*) = E(Y \mid X)$

Are Two Models Better Than One?[4]

Statistical Science
2007, Vol. 22, No. 4, 523–539
DOI: 10.1214/07-STS227

Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data[1]

Joseph D. Y. Kang and Joseph L. Schafer

Abstract. When outcomes are missing for reasons beyond an investigator's control, there are two different ways to adjust a parameter estimate for covariates that may be related both to the outcome and to missingness. One approach is to model the relationships between the covariates and the outcome and use those relationships to predict the missing values. Another is to model the probabilities of missingness given the covariates and incorporate them into a weighted or stratified estimate. Doubly robust (DR) procedures apply both types of model simultaneously and produce a consistent estimate of the parameter if either of the two models has been correctly specified. In this article, we show that DR estimates can be constructed in many ways. We compare the performance of various DR and non-DR estimates of a population mean in a simulated example where both models are incorrect

[4] https://projecteuclid.org/euclid.ss/1207580167

Simulation Studies of Kang and Schafer

- The authors used extensive simulation scenarios to evaluate different estimators
- Simulation scenarios resemble a quasi-experiment to measure the effect of dieting on body mass index (BMI) in a large sample of high-school students
- At baseline, covariates measured include BMI, self-perceived physical fitness, social acceptance and personality measures
- Outcome is BMI in 1 year, which may be missing
- Response bias is moderate, good overlap between the missing and non-missing
- Good predictors of the outcomes are available, $R^2 = 0.81$
- Both models are approximately but not exactly true

Simulation Studies of Kang and Schafer

- The authors used extensive simulation scenarios to evaluate different estimators
- Simulation scenarios resemble a quasi-experiment to measure the effect of dieting on body mass index (BMI) in a large sample of high-school students
- At baseline, covariates measured include BMI, self-perceived physical fitness, social acceptance and personality measures
- Outcome is BMI in 1 year, which may be missing
- Response bias is moderate, good overlap between the missing and non-missing
- Good predictors of the outcomes are available, $R^2 = 0.81$
- Both models are approximately but not exactly true

Simulation Studies of Kang and Schafer

- The authors used extensive simulation scenarios to evaluate different estimators
- Simulation scenarios resemble a quasi-experiment to measure the effect of dieting on body mass index (BMI) in a large sample of high-school students
- At baseline, covariates measured include BMI, self-perceived physical fitness, social acceptance and personality measures
- Outcome is BMI in 1 year, which may be missing
- Response bias is moderate, good overlap between the missing and non-missing
- Good predictors of the outcomes are available, $R^2 = 0.81$
- Both models are approximately but not exactly true

Simulation Studies of Kang and Schafer

- The authors used extensive simulation scenarios to evaluate different estimators
- Simulation scenarios resemble a quasi-experiment to measure the effect of dieting on body mass index (BMI) in a large sample of high-school students
- At baseline, covariates measured include BMI, self-perceived physical fitness, social acceptance and personality measures
- Outcome is BMI in 1 year, which may be missing
- Response bias is moderate, good overlap between the missing and non-missing
- Good predictors of the outcomes are available, $R^2 = 0.81$
- Both models are approximately but not exactly true

Simulation Studies of Kang and Schafer

- The authors used extensive simulation scenarios to evaluate different estimators
- Simulation scenarios resemble a quasi-experiment to measure the effect of dieting on body mass index (BMI) in a large sample of high-school students
- At baseline, covariates measured include BMI, self-perceived physical fitness, social acceptance and personality measures
- Outcome is BMI in 1 year, which may be missing
- Response bias is moderate, good overlap between the missing and non-missing
- Good predictors of the outcomes are available, $R^2 = 0.81$
- Both models are approximately but not exactly true

Simulation Studies of Kang and Schafer

- The authors used extensive simulation scenarios to evaluate different estimators
- Simulation scenarios resemble a quasi-experiment to measure the effect of dieting on body mass index (BMI) in a large sample of high-school students
- At baseline, covariates measured include BMI, self-perceived physical fitness, social acceptance and personality measures
- Outcome is BMI in 1 year, which may be missing
- Response bias is moderate, good overlap between the missing and non-missing
- Good predictors of the outcomes are available, $R^2 = 0.81$
- Both models are approximately but not exactly true

Simulation Studies of Kang and Schafer

- The authors used extensive simulation scenarios to evaluate different estimators
- Simulation scenarios resemble a quasi-experiment to measure the effect of dieting on body mass index (BMI) in a large sample of high-school students
- At baseline, covariates measured include BMI, self-perceived physical fitness, social acceptance and personality measures
- Outcome is BMI in 1 year, which may be missing
- Response bias is moderate, good overlap between the missing and non-missing
- Good predictors of the outcomes are available, $R^2 = 0.81$
- Both models are approximately but not exactly true

Some Conclusions of Kang and Schafer

- "*Methods that use inverse-probabilities as weights, whether they are DR [double robust] or not, are sensitive to misspecification of the propensity model when some estimated propensities are small*"
- "*Many DR methods perform better than simple inverse-probability weighting*"
- "*None of the DR methods we tried, however, improved upon the performance of simple regression-based prediction of the missing values*"
- "*This study does not represent every missing-data problem that will arise in practice. But it does demonstrate that, in at least some settings, two wrong models are not better than one*"

Some Conclusions of Kang and Schafer

- "*Methods that use inverse-probabilities as weights, whether they are DR [double robust] or not, are sensitive to misspecification of the propensity model when some estimated propensities are small*"
- "*Many DR methods perform better than simple inverse-probability weighting*"
- "*None of the DR methods we tried, however, improved upon the performance of simple regression-based prediction of the missing values*"
- "*This study does not represent every missing-data problem that will arise in practice. But it does demonstrate that, in at least some settings, two wrong models are not better than one*"

Some Conclusions of Kang and Schafer

- "*Methods that use inverse-probabilities as weights, whether they are DR [double robust] or not, are sensitive to misspecification of the propensity model when some estimated propensities are small*"
- "*Many DR methods perform better than simple inverse-probability weighting*"
- "*None of the DR methods we tried, however, improved upon the performance of simple regression-based prediction of the missing values*"
- "*This study does not represent every missing-data problem that will arise in practice. But it does demonstrate that, in at least some settings, two wrong models are not better than one*"

Some Conclusions of Kang and Schafer

- "*Methods that use inverse-probabilities as weights, whether they are DR [double robust] or not, are sensitive to misspecification of the propensity model when some estimated propensities are small*"
- "*Many DR methods perform better than simple inverse-probability weighting*"
- "*None of the DR methods we tried, however, improved upon the performance of simple regression-based prediction of the missing values*"
- "*This study does not represent every missing-data problem that will arise in practice. But it does demonstrate that, in at least some settings, two wrong models are not better than one*"

Stabilizing Weights

- What is happening? Similar to Basu's elephant: weights $R_i/\pi(X_i;\hat{\psi})$ are too unstable
- Under the true π-model, the weights are expected to be 1:

$$E\left(\frac{R}{\pi(X)}\right) = E\left(\frac{p(R=1 \mid X)}{\pi(X)}\right) = 1$$

- Therefore, when the model is correctly specified,

$$C = \frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\pi(X_i;\hat{\psi})} \approx 1$$

- However, when the model is misspecified, often $C >> 1$
- Define the inverse of $\tilde{\pi}_i = C\pi(X_i;\hat{\psi})$ as the stabilizing weight, so that

$$\frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\tilde{\pi}_i} = \frac{1}{C}\frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\pi(X_i;\hat{\psi})} = 1$$

Stabilizing Weights

- What is happening? Similar to Basu's elephant: weights $R_i/\pi(X_i;\hat{\psi})$ are too unstable
- Under the true π-model, the weights are expected to be 1:

$$E\left(\frac{R}{\pi(X)}\right) = E\left(\frac{p(R=1 \mid X)}{\pi(X)}\right) = 1$$

- Therefore, when the model is correctly specified,

$$C = \frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\pi(X_i;\hat{\psi})} \approx 1$$

- However, when the model is misspecified, often $C >> 1$
- Define the inverse of $\tilde{\pi}_i = C\pi(X_i;\hat{\psi})$ as the stabilizing weight, so that

$$\frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\tilde{\pi}_i} = \frac{1}{C}\frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\pi(X_i;\hat{\psi})} = 1$$

Stabilizing Weights

- What is happening? Similar to Basu's elephant: weights $R_i/\pi(X_i;\hat{\psi})$ are too unstable
- Under the true π-model, the weights are expected to be 1:

$$E\left(\frac{R}{\pi(X)}\right) = E\left(\frac{p(R=1 \mid X)}{\pi(X)}\right) = 1$$

- Therefore, when the model is correctly specified,

$$C = \frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\pi(X_i;\hat{\psi})} \approx 1$$

- However, when the model is misspecified, often $C >> 1$
- Define the inverse of $\tilde{\pi}_i = C\pi(X_i;\hat{\psi})$ as the stabilizing weight, so that

$$\frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\tilde{\pi}_i} = \frac{1}{C}\frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\pi(X_i;\hat{\psi})} = 1$$

Stabilizing Weights

- What is happening? Similar to Basu's elephant: weights $R_i/\pi(X_i;\hat{\psi})$ are too unstable
- Under the true π-model, the weights are expected to be 1:

$$E\left(\frac{R}{\pi(X)}\right) = E\left(\frac{p(R=1 \mid X)}{\pi(X)}\right) = 1$$

- Therefore, when the model is correctly specified,

$$C = \frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\pi(X_i;\hat{\psi})} \approx 1$$

- However, when the model is misspecified, often $C >> 1$
- Define the inverse of $\tilde{\pi}_i = C\pi(X_i;\hat{\psi})$ as the stabilizing weight, so that

$$\frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\tilde{\pi}_i} = \frac{1}{C}\frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\pi(X_i;\hat{\psi})} = 1$$

Stabilizing Weights

- What is happening? Similar to Basu's elephant: weights $R_i/\pi(X_i;\hat{\psi})$ are too unstable
- Under the true π-model, the weights are expected to be 1:

$$E\left(\frac{R}{\pi(X)}\right) = E\left(\frac{p(R=1 \mid X)}{\pi(X)}\right) = 1$$

- Therefore, when the model is correctly specified,

$$C = \frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\pi(X_i;\hat{\psi})} \approx 1$$

- However, when the model is misspecified, often $C >> 1$
- Define the inverse of $\tilde{\pi}_i = C\pi(X_i;\hat{\psi})$ as the stabilizing weight, so that

$$\frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\tilde{\pi}_i} = \frac{1}{C}\frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\pi(X_i;\hat{\psi})} = 1$$

Outline

Summary

Main take-aways from today's lecture:

- Inverse-probability weighting
  - Origins in survey sampling (Horvitz-Thompson estimator)
  - Does not require modeling of the full-data distribution
  - Sensitive to misspecification of the propensity score model and to extreme weights
- Augmented IPW
  - Enjoys *double-robustness* property
  - However "*in at least some settings, two wrong models are not better than one*" (Kang and Schafer, 2007)

Next lecture:

- Weighted Generalized Estimating Equations