Lecture 14
Inverse Probability weighting

Outline

- IPW
  - Using normalized weights
  - Connection with weighted least squares
  - IPW V.S. stratification
  - Bootstrap
- Suggested reading: Imbens and Rubin book 12.4.2, Peng's book Chapter 11.2

Motivation

- Matching methods can improve covariate balance
- Potential limitations of matching methods:
  - Inefficient: it may throw away many control units
  - Ineffective: it may not be able to balance covariates
  - Biased: not estimating the ATT if a lot of treated units are not matched

- Matching is a special case of weighting

$$\hat{\tau}_{\text{match}} = \frac{1}{N_t}\sum_{i=1}^{N} W_i \left(Y_i^{\text{obs}} - \frac{1}{|\mathcal{M}_i^c|}\sum_{i' \in \mathcal{M}_i^c} Y_{i'}^{\text{obs}} \right)$$

$$= \frac{1}{N_t}\sum_{i:W_i=1} Y_i^{\text{obs}} - \frac{1}{N_c}\sum_{i:W_i=1} \left(\frac{N_c}{N_t}\sum_{i':W_{i'}=1} \frac{1_{i \in \mathcal{M}_{i'}^c}}{|\mathcal{M}_{i'}^c|} \right) Y_i^{\text{obs}}$$

- Idea: weight each observation in the control group such that it looks like the treatment group

Inverse probability weighting (IPW)

- Weighting makes use the following properties to estimate $\mathbb{E}(Y_i(1))$ and $\mathbb{E}(Y_i(0))$

$$\mathbb{E}\left[\frac{Y_i^{\text{obs}} \cdot W_i}{e(X_i)}\right] = \mathbb{E}_{\text{sp}}\left[Y_i(1)\right], \quad \text{and} \quad \mathbb{E}\left[\frac{Y_i^{\text{obs}} \cdot (1 - W_i)}{1 - e(X_i)}\right] = \mathbb{E}_{\text{sp}}\left[Y_i(0)\right]$$

  - Intuitively, unit that has a smaller $e(\boldsymbol{X}_i)$ has less chance to appear in the treatment group, so we should give it a higher weight (the less likely a subject is sampled, then the larger population it should represent)

$$\hat{\tau}_{\text{IPW}} = \frac{1}{N}\sum_{i=1}^{N}\frac{W_i \cdot Y_i^{\text{obs}}}{e(X_i)} - \frac{1}{N}\sum_{i=1}^{N}\frac{(1 - W_i) \cdot Y_i^{\text{obs}}}{1 - e(X_i)}$$

$$= \frac{1}{N}\sum_{i:W_i=1}\lambda_i \cdot Y_i^{\text{obs}} - \frac{1}{N}\sum_{i:W_i=0}\lambda_i \cdot Y_i^{\text{obs}},$$

where

$$\lambda_i = \frac{1}{e(X_i)^{W_i} \cdot (1 - e(X_i))^{1-W_i}} = \begin{cases} 1/(1 - e(X_i)) & \text{if } W_i = 0, \\ 1/e(X_i) & \text{if } W_i = 1. \end{cases}$$

IPW for observational studies

- The propensity scores are estimated
- Estimate ATE and ATT
  - ATE

$$\widehat{\mathrm{ATE}} = \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{W_i Y_i^{\mathrm{obs}}}{\hat{e}(\boldsymbol{X}_i)} - \frac{(1-W_i)Y_i^{\mathrm{obs}}}{1-\hat{e}(\boldsymbol{X}_i)}\right\}$$

  - ATT

$$\widehat{\mathrm{ATT}} = \frac{1}{N_t}\sum_{i=1}^{N}\left\{W_i Y_i^{\mathrm{obs}} - \frac{\hat{e}(\boldsymbol{X}_i)(1-W_i)Y_i^{\mathrm{obs}}}{1-\hat{e}(\boldsymbol{X}_i)}\right\}$$

- For units that have identical propensity scores → difference-in-means estimator

Normalizing the weights

- When use any weighting method (e.g. IPW), good practice is to normalize weights – sum of the total of weights within one group should be 1

- Divide each unit's weight (ω_i) by the sum of all weights in that group $\omega_i / \sum_{i':W_{i'}=w} \omega_{i'}$ for $w = 0,1$, i.e. the Hajek estimator:
- The new ATE estimator:

$$\widehat{\mathrm{ATE}} = \frac{\sum_{i=1}^{N} W_i Y_i^{\mathrm{obs}} / \hat{e}(\boldsymbol{X}_i)}{\sum_{i=1}^{N} W_i / \hat{e}(\boldsymbol{X}_i)} - \frac{\sum_{i=1}^{N} (1 - W_i) Y_i^{\mathrm{obs}} / (1 - \hat{e}(\boldsymbol{X}_i))}{\sum_{i=1}^{N} (1 - W_i) / (1 - \hat{e}(\boldsymbol{X}_i))}$$

- The new ATT estimator:

$$\widehat{\mathrm{ATT}} = \frac{1}{N_t} \sum_{i=1}^{N} W_i Y_i^{\mathrm{obs}} - \frac{\sum_{i=1}^{N} (1 - W_i) Y_i^{\mathrm{obs}} \hat{e}(\boldsymbol{X}_i) / (1 - \hat{e}(\boldsymbol{X}_i))}{\sum_{i=1}^{N} (1 - W_i) \hat{e}(\boldsymbol{X}_i) / (1 - \hat{e}(\boldsymbol{X}_i))}$$

- Using normalized weights, we can reduce variance and lead to more stable estimate (Hirano, Imbens, Ridder, 2003)

Connection between IPW estimator and WLS

- Define inverse probability weights

$$\lambda_i = \frac{1}{e(X_i)^{W_i} \cdot (1 - e(X_i))^{1-W_i}} = \begin{cases} 1/(1 - e(X_i)) & \text{if } W_i = 0, \\ 1/e(X_i) & \text{if } W_i = 1. \end{cases}$$

- Weighted least square with no covariate adjustments

$$(\hat{\alpha}, \hat{\tau}) = \min_{\alpha,\tau} \sum_{i=1}^{N} \lambda_i \left(Y_i^{\text{obs}} - \alpha - \tau W_i\right)^2$$

  - Solution: $\hat{\alpha} = \frac{\sum_{i=1}^{N}(1-W_i)\lambda_i Y_i^{\text{obs}}}{\sum_{i=1}^{N}(1-W_i)\lambda_i}$ and $\hat{\alpha} + \hat{\tau} = \frac{\sum_{i=1}^{N} W_i \lambda_i Y_i^{\text{obs}}}{\sum_{i=1}^{N} W_i \lambda_i}$

- Solution is the same as IPW with normalizing weights
- If we ignore the uncertainty in estimating the propensity score, we can estimate the variance of $\hat{\tau}$ from Sandwich estimator for WLS

- We can also use WLS to adjust for other pre-treatment covariates

IPW advantages v.s. disadvantages

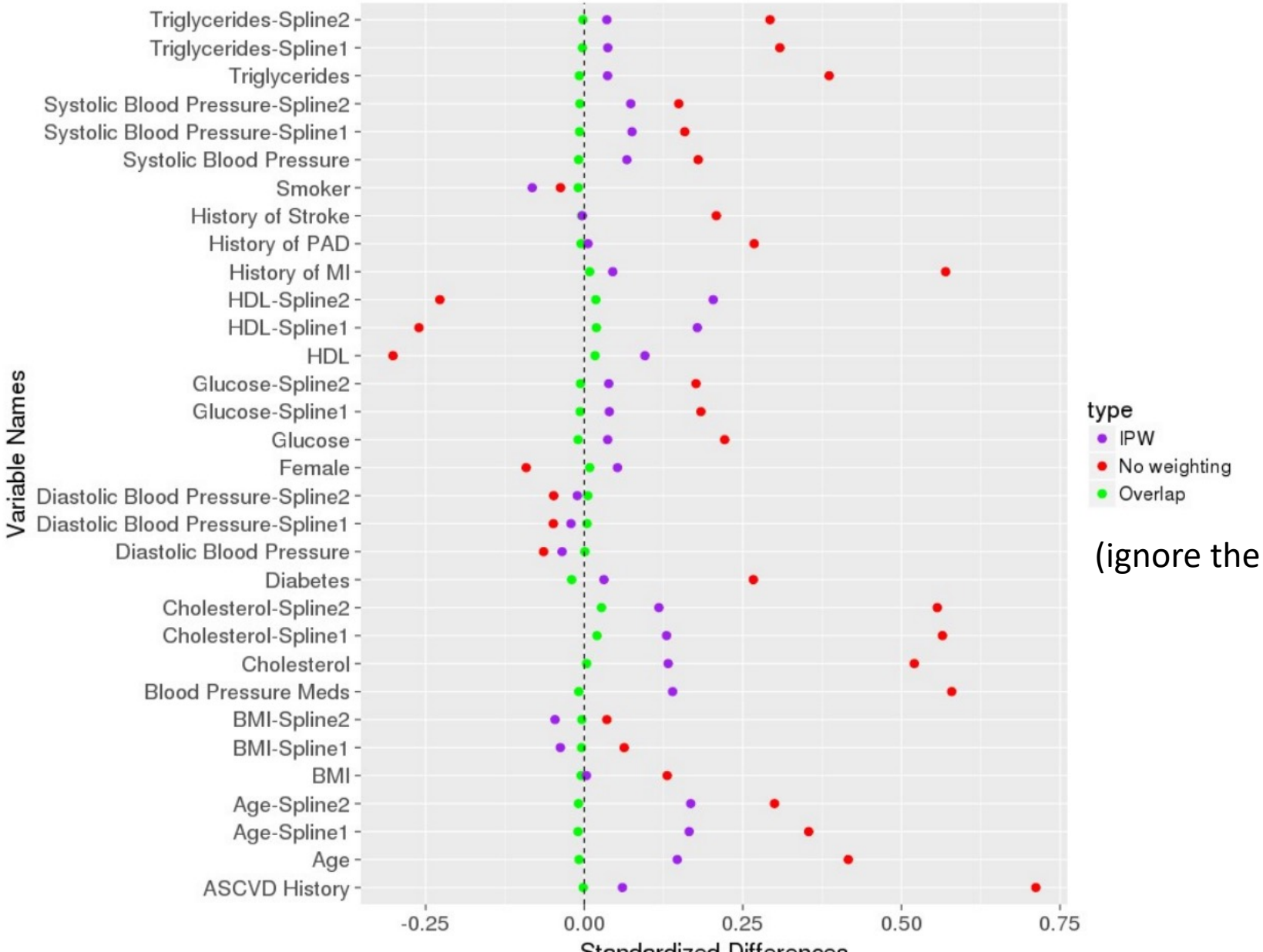
- Advantages
  - Simple, with theoretical foundation
  - Global balance
  - Use all data
  - Can be extended to more complex settings

- Disadvantages
  - More sensitive to misspecification of propensity scores then matching
  - Estimated propensity scores near 0 or 1 can yield extreme weights

Example: Framingham Heart Study

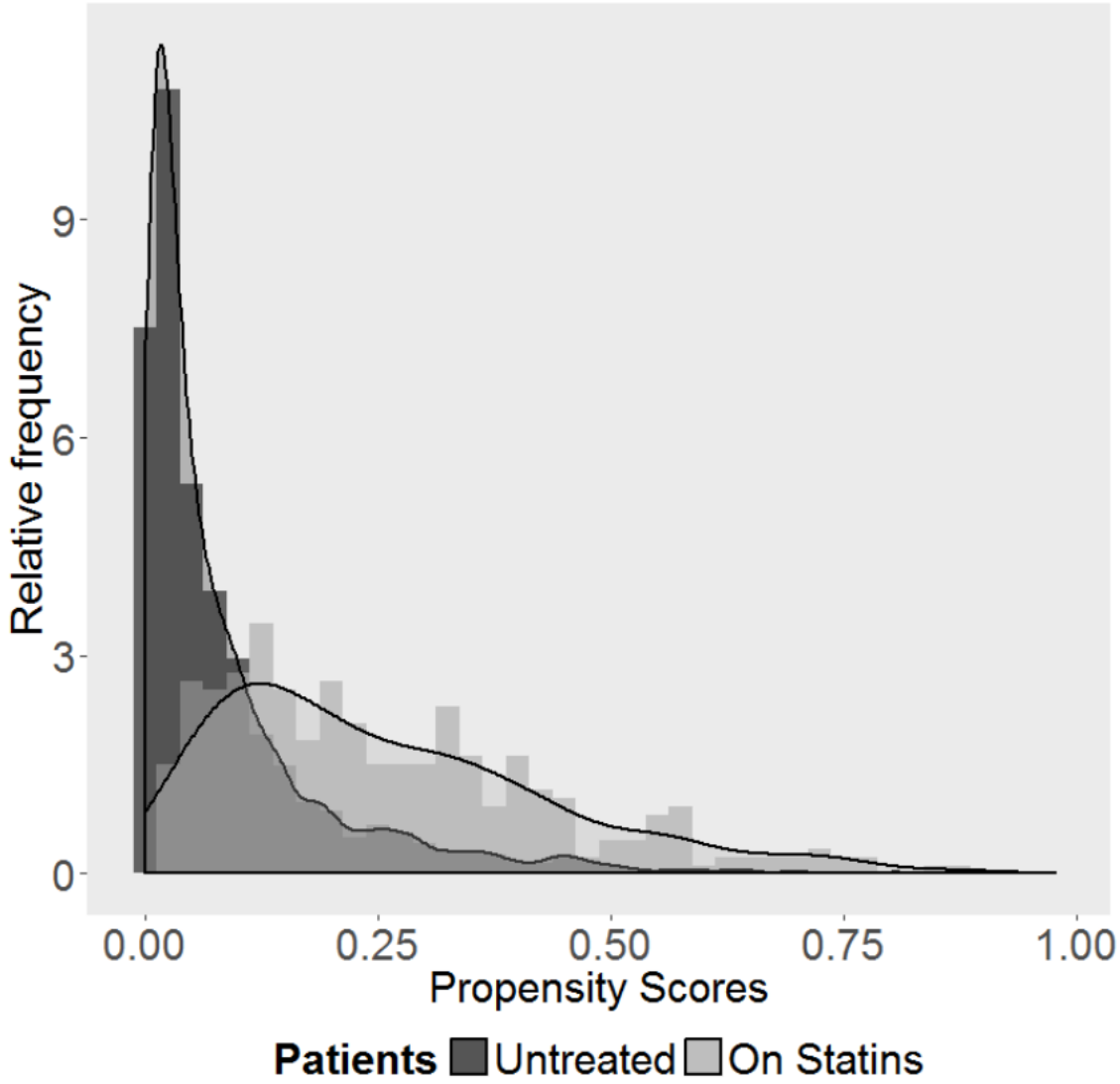
- Goal: evaluate the effect of statins on health outcomes
- Patients: cross-sectional population from the offspring cohort with a visit 6 (1995-1998)
- Treatment: statin use at visit 6 vs. no statin use
- Outcomes: CV(cardiovascular) death, myocardial infarction (MI), stroke
- Confounders: sex, age, body mass index, diabetes, history of MI, history of PAD, history of stroke...
- Significant imbalance between treatment and control groups in covariates motivates IPW (or some form of propensity score adjustment)

Love plot for covariate balancing



(ignore the green dots)

Distribution of estimated propensity scores



- For treated units with $\hat{e}(\boldsymbol{X}_i)$ close to 0, then can greatly influence the IPW estimator value
- Trimming removes individuals with extremely large weights

Stratification V.S. IPW estimators

- Stratification estimator can be treated as a weighting estimator

$$\hat{\tau}^{\text{strat}} = \frac{1}{N}\sum_{i=1}^{N} W_i \cdot Y_i^{\text{obs}} \cdot \lambda_i^{\text{strat}} - \frac{1}{N}\sum_{i=1}^{N} (1 - W_i) \cdot Y_i^{\text{obs}} \cdot \lambda_i^{\text{strat}},$$

where the weights λ_i^{strat} satisfy

$$\lambda_i^{\text{strat}} = \sum_{j=1}^{J} B_i(j) \cdot \left(\frac{1 - W_i}{N_{\text{c}}(j)/N(j)} + \frac{W_i}{N_{\text{t}}(j)/N(j)} \right)$$

$$= \begin{cases} \sum_{j=1}^{J} B_i(j) \cdot \frac{N(j)}{N_{\text{c}}(j)} & \text{if } W_i = 0, \\ \sum_{j=1}^{J} B_i(j) \cdot \frac{N(j)}{N_{\text{t}}(j)} & \text{if } W_i = 1. \end{cases}$$

- Instead of using the eps $\hat{e}(\boldsymbol{X}_i)$ to obtain weights, stratification estimator estimates the propensity scores as the block proportions (averaging $\hat{e}(\boldsymbol{X}_i)$ within subclasses)

$$\tilde{e}(X_i) = \sum_{j=1}^{J} B_i(j) \cdot \frac{N_{\text{t}}(j)}{N(j)}$$

Stratification V.S. IPW estimators

- If there are many blocks, then the dispersion within each stratum is limited, two estimators are similar
- The weights will be different only if, in at least some blocks, there is substantial variation in the propensity score, which is most likely to happen in blocks with propensity score values close to zero and one.
- Smoothing the weights by averaging them within blocks, as the stratification estimator does, may remove some of the biases introduced by the estimation of propensity scores (avoids extreme weights).
- Stratification is more robust to model mis-specification.
- Stratification as a coarsening method is more ad-hoc.

The Imbens-Rubin-Sacerdote lottery data

[Estimating the effect of unearned income on labor earnings, savings, and consumption: Evidence from a survey of lottery players. *American economic review*, 2001]

- Goal: Estimate magnitude of lottery prizes (unearned income) on economic behavior, including labor supply, consumption and savings

- Data collection:
  - "Winners": individuals who had played and won large sums of money in the Massachsetts lottery
  - "Losers": individuals who played the lottery and had won only small prizes
  - Constructing a comparison group of lottery players who did not win anything was not feasible as the Lottery Commission did not have contact information of such individuals

- Surveys are sent to these individuals with financial incentives

- We analyze a subset of $N_t = 259$ and $N_c = 237$ individuals with complete answers

- We use the model forward selection procedure to estimate the propensity scores

The Imbens-Rubin-Sacerdote lottery data

TABLE 1—RESPONSE RATES BY MAILING

		Sent		Responses		Response rates		
Mailing	Date	Winners	Nonwinners	Winners	Nonwinners	Winners	Nonwinners	Total
Pilot	July '95	50	50	17	25	0.34	0.50	0.42
Main	July '96	752	637	272	262	0.36	0.41	0.38
Follow-up ($50 check)	Sept. '96	248	248	39	40	0.16	0.16	0.16
Follow-up ($10 cash, $40 check)	Sept. '96	49	49	11	12	0.22	0.24	0.23
Total		802	687	339	339	0.42	0.49	0.46

Covariate balancing after IPW or stratification

Table 17.1. ***Normalized Differences in Covariates after Subclassification for the IRS Lottery Data***

Variable	Full Sample		Trimmed Sample			
	One Block	Horvitz-Thompson	One Block	Two Blocks	Five Blocks	Horvitz-Thompson
Year Won	−0.26	0.10	−0.06	−0.03	0.07	0.07
# Tickets	0.91	0.10	0.51	0.17	0.07	−0.04
Age	−0.50	−0.30	−0.09	−0.03	0.05	0.05
Male	−0.19	0.09	−0.11	−0.10	−0.14	−0.13
Education	−0.70	0.48	−0.51	−0.18	−0.10	−0.01
Work Then	0.09	0.05	0.03	0.03	0.01	0.00
Earn Year -6	−0.32	0.01	−0.18	−0.10	−0.03	0.06
Earn Year -5	−0.28	0.01	−0.19	−0.07	−0.00	0.09
Earn Year -4	−0.29	−0.01	−0.23	−0.09	−0.01	0.06
Earn Year -3	−0.26	0.05	−0.18	−0.03	0.03	0.10
Earn Year -2	−0.31	0.06	−0.19	−0.03	0.01	0.09
Earn Year -1	−0.23	0.11	−0.17	−0.01	0.00	0.06
Pos Earn Year -6	0.03	0.16	−0.00	−0.09	−0.09	−0.01
Pos Earn Year -5	0.14	−0.14	0.10	0.01	−0.01	0.06
Pos Earn Year -4	0.10	−0.19	0.06	−0.00	−0.01	0.03
Pos Earn Year -3	0.13	−0.17	0.03	−0.04	−0.05	−0.00
Pos Earn Year -2	0.14	−0.17	0.06	0.00	−0.04	0.01
Pos Earn Year -1	0.10	0.17	−0.01	−0.04	−0.07	−0.01

- Trimming:
  results from optimal trimming only keep individuals whose $\hat{e}(\boldsymbol{X}_i) \in [0.0891, 0.9109]$
- Horvitz-Thompson: IPW with normalized weights

Results on the lottery data: stratification

- Estimated ATE

Covariates	Full Sample 1 Block		Trimmed Sample 1 Block		Trimmed Sample 2 Blocks		Trimmed Sample 5 Blocks	
	Est	(s.e.)	Est	(s.e.)	Est	(s.e.)	Est	(s.e.)
None	−6.2	(1.4)	−6.6	(1.7)	−6.0	(1.9)	−5.7	(2.0)
# Tickets, Education, Work Then, Earn Year-1	−2.8	(0.9)	−4.0	(1.1)	−5.6	(1.2)	−5.1	(1.2)
All	−5.1	(1.0)	−5.3	(1.1)	−6.4	(1.1)	−5.7	(1.1)

Results on the lottery data: IPW using weighted linear regression

- Estimated ATE

Table 17.9. ***Least Squares Regression Estimates for the IRS Lottery Data***

Covariate	Full Sample		Trimmed Sample	
	Est	(s. e.)	Est	(s. e.)
Intercept	21.20	(4.80)	22.76	(6.49)
Treatment Indicator	−5.08	(0.95)	−5.34	(1.08)

- These standard errors from WLS tend to underestimate the actual uncertainty as they assume weights are fixed (estimated propensity scores are true). Will discuss bootstrap later

Stratification V.S. IPW estimators

- On the lottery data, summary statistics of the weights

	Full Sample		Trimmed Sample	
	Horvitz-Thompson	Subclass	Horvitz-Thompson	Subclass
Minimum	0.92	1.06	1.00	1.19
Maximum	79.79	17.71	18.18	6.15
Standard deviation	4.20	2.63	1.69	1.35

- Uncertainty on the lottery data

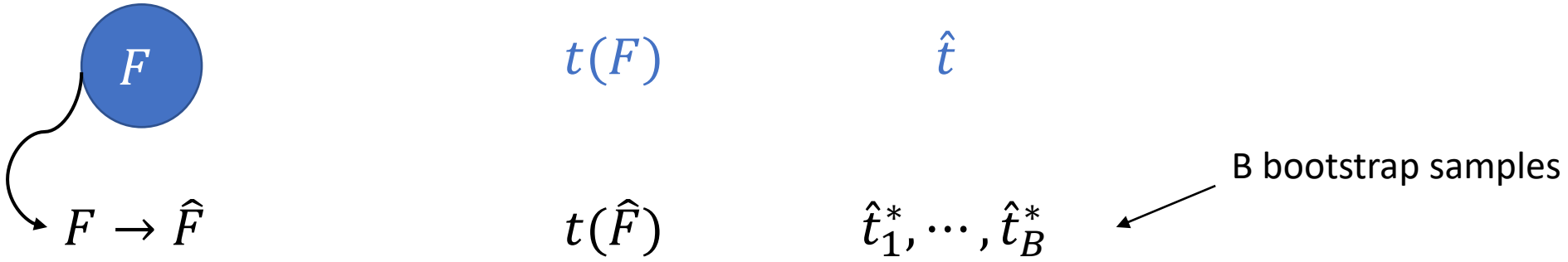
	Full Sample		Trimmed Sample	
	Horvitz-Thompson	Subclass	Horvitz-Thompson	Subclass
Bias	4.34	2.68	1.29	0.30
Variance	2.59^2	0.83^2	1.29^2	1.15^2
$\text{Bias}^2+\text{Variance}$	5.06^2	2.81^2	1.83^2	1.19^2

- Horvitz-Thompson: IPW with normalized weights
- Subclass: propensity score stratification

Variance of IPW estimator

- Researchers have shown that using the estimated propensity score asymptotically results in smaller variance of the IPW estimator (Hirano, Imbens and Ridder, 2003)
- Closed-form sandwich estimator (M-estimator) of variance that takes into account of the uncertainty in estimating the propensity score (Lunceford and Davidian, 2004)
- Bootstrap: Resample units and refit PS and estimate the causal effects every time – computationally intensive for large sample

Bootstrap



- Nonparametric bootstrap:
  - Repeat B times: for each time b
    - sample N units with replacement (or resample the treated and controls separately)
    - Follow the whole procedure (starting from propensity score estimation to estimate the ATE/ATT using IPW)
    - Obtain an IPW estimator $\hat{\tau}_{IPW}^{(b)}$
  - Use the histogram of $\{\hat{\tau}_{IPW}^{(1)}, \cdots, \hat{\tau}_{IPW}^{(B)}\}$ as the approximated distribution of $\hat{\tau}_{IPW}$
    - The standard deviation of these estimates approximates the standard error of $\hat{\tau}_{IPW}$