Lecture 6: Introduction to estimation; From the G-computation formula to a simple substitution estimator

A roadmap for causal inference

1. Specify **Causal Model** representing real background knowledge
2. Specify **Causal Question**
3. Specify **Observed Data** and link to causal model
4. **Identify** : Knowledge + data sufficient?
5. Commit to an **estimand** as close to question as possible, and a **statistical model** representing real knowledge.
6. **Estimate**
7. **Interpret** Results

Outline

1. Definitions:
   - Parameters
   - Estimators
   - Substitution estimators
2. From the point treatment G-computation formula to a simple substitution estimator
   - Example and intuition
   - Comparison to standard MV regression
3. Motivation for new non-parametric approaches
   - The importance of respecting your statistical model
   - Evaluating estimator performance

Parameters

- Parameter Ψ: A mapping from the statistical model to the parameter space
  - Ψ: $\mathcal{M}$---> Real Numbers
- A function that
  - Takes as input any distribution in the statistical model $\mathcal{M}$
  - Gives as output a value in the parameter space (eg the real numbers)

Parameter of the observed data distribution

- $\Psi(P_0)=\psi_0$ is the true parameter value
  - It is a function of the (unknown) true observed data distribution P_0
  - It is an element of the parameter space
- Also referred to as the **estimand**

Parameter of the observed data distribution, or estimand

- Example: $\Psi(P_0)=E_W(E_0(Y|A=1,W)-E_0(Y|A=0,W))$
- If we knew P_0 (w,a,y) for all (w,a,y), we could plug this into Ψ and get a real number
- This number would be equivalent to the ATE under specific causal assumptions
  - Eg W satisfies the back door criteria

Empirical Distribution: P_n

- We sample n i.i.d. copies of the random variable O
- The empirical distribution P_n corresponds to putting a weight of 1/n on each copy $O_{i,}$ i=1,…n

Estimators

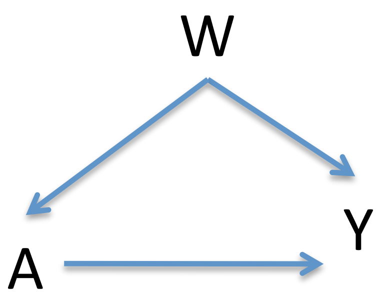
- Estimator: $\hat{\Psi}$: A mapping from the set of possible empirical distributions P_n to the parameter space
  - $\hat{\Psi}:\mathcal{M}_{\mathcal{NP}}$ ---> Real Numbers
- A function that
  - Takes as input our observed data
    - A realization of P_n
  - Gives as output a value in the parameter space
    - Ex. the real numbers

Estimators

- $\hat{\Psi}(P_n) = \psi_n$ is the estimate
  - It is a function of the empirical distribution of the data
  - It is an element of the parameter space
- If we plug in a realization of P_n (based on a sample of size n of the random variable O), we get back an estimate ψ_n of the true parameter value ψ_0

Our Classic Example

- $\Psi^F(P_{UX})=E_{U,X}(Y_1-Y_0)$



- Observe n i.i.d. copies of $O=(W,A,Y)\sim P_0$

- $\Psi(P_0)$
  $=E_{W,0}[E_0(Y|A=1,W)-E_0(Y|A=0,W)]$
- If we knew P_0, we could plug it into the function Ψ and get the true parameter value
  - In fact, we just need $E_0(Y|A,W)$ and $P_0(w)$
  - But we don't know P_0
- How might we define an estimator of $\Psi(P_0)$?

Substitution Estimators

- Also referred to as “plug in” estimators
- As in this example, often the target parameter is only a function of <u>part</u> of P_0
- Let Q_0 be defined as the part of P_0 that the target parameter Ψ is a function of
  - i.e. $\Psi(P_0)=\Psi(Q_0)$

Definition: Substitution Estimator

- A substitution estimator is an estimator based on

1. Defining an estimator Q_n of Q_0
   - Where Q_n respects the statistical model
2. Plugging the resulting estimate into the parameter mapping Ψ in order to generate an estimate of the true parameter value
   - $\hat{\Psi}(P_n) = \Psi(Q_n)$

Ex. Simple substitution estimator based on the G-computation formula

- $O=(W,A,Y)\sim P_0$
- $\Psi(P_0)=E_W(E_0(Y|A=1,W)-E_0(Y|A=0,W)$
- We use Q_0 to refer to the parts of the observed data distribution that our target parameter is a function of
  - i.e. $\Psi(P_0)=\Psi(Q_0)$
- Ex: $\Psi(P_0)=E_W(E_0(Y|A=1,W)-E_0(Y|A=0,W)$
  - $\Psi(P_0)$ only a function of $\bar{Q}_0(A,W) \equiv E_0(Y|A,W)$ and $Q_{0,W}$ (distribution of W)
  - $Q_0 = (\bar{Q}_0, Q_{0,W})$

Simple substitution estimator based on the G-computation formula

- We define

1. An algorithm that takes the observed data as input and gives us an estimate of $E_0(Y|A,W)$
2. An algorithm that takes the observed data as input and gives us an estimate of $P_0(W=w)$

- We can now substitute these estimates in place of the unknown observed data parameters

$$\Psi(P_0) = \sum_w \left(E_0(Y|A=1,W=w) - E_0(Y|A=0,W=w)\right) P_0(W=w)$$

$$\hat{\Psi}(P_n) = \sum_w \left(\hat{E}(Y|A=1,W=w) - \hat{E}(Y|A=0,W=w)\right) \hat{P}(W=w)$$

How might we estimate $P_0(W=w)$?

- Our estimator should respect our statistical model
  - Here, our statistical model is non-parametric
- A simple non-parametric estimator of $P_0(W=w)$: sample proportion $\frac{1}{n}\sum_{i=1}^{n} I(W_i = w)$
  - W_i is observed covariate value for subject i
- This doesn't assume anything about the distribution of W

A simple substitution estimator

- Target parameter value of observed data distribution:

$$\Psi(Q_0) = E_W[E_0(Y|A=1,W) - E_0(Y|A=0,W)]$$

- To take the expectation over W, we take the empirical mean over W_i, i=1,…,n
  - Same as estimating P(W=w) as the sample proportion
- An estimator of E_0 (Y|A,W) thus gives us a substitution estimator:

$$\hat{\Psi}(P_n) = \Psi(Q_n) = \frac{1}{n}\sum_{i=1}^{n}[\bar{Q}_n(1,W_i) - \bar{Q}_n(0,W_i)],$$

where $\bar{Q}_n(A,W)$ is an estimator of $E_0(Y|A,W)$.

General implementation of substitution estimator based on G-computation formula

1. Estimate $\bar{Q}_0(A,W) = E_0(Y|A,W)$
2. Use this estimate to generate a predicted outcome for each subject setting A=1 and setting A=0
   – Intuition: Mimics study where each individual received and did not receive the treatment
3. Estimate $\Psi(P_0)$ as the difference in the mean of these predicted outcomes

How might we estimate $E_0(Y|A,W)$?

- A simple non-parametric estimator of $E_0(Y|A,W)$: Take empirical mean of Y within strata defined by each possible value for (A,W)
  - Also referred to as non-parametric maximum likelihood estimator (NPMLE)
  - Same as fitting a saturated regression model

Empirical Mean of Y within strata defined by (A,W)

	W=1	W=0
A=1	35 (n=110)	5 (n=230)
A=0	10 (n=123)	27 (n=78)

HIV Example: Effect of switch to second line therapy on

- Intervention: a weekly pill organizer
- Designed to help patients remember to take their prescribed medications

Research Question:
Does use of a pill box improve adherence to antiretroviral drugs?

Example: Effect of Pill Box Use on Adherence to Antiretrovirals

- A= Pill Box “Mediset” Use
- Y= adherence to antiretroviral drugs
  - % of prescribed doses taken
- W= age, sex, recreational drug use, past adherence, type of regimen, CD4 count….

Research Question:
Does use of Pill Box improve adherence to antiretroviral drugs?

Simple Example: G-computation

Original Data

ID	Pill Box (A)	Crack Use (W)	Adherence (Y)
1	1	1	0.7
2	0	0	0.8
3	1	1	0.4
4	1	0	1
5	0	1	0.4
6	0	0	0.7

$$\hat{E}(Y|A=1,W=1)=0.55$$
$$\hat{E}(Y|A=0,W=1)=0.4$$
$$\hat{E}(Y|A=1,W=0)=1.0$$
$$\hat{E}(Y|A=0,W=0)=0.75$$

Expanded Data with Predicted Outcomes

ID	Pill Box (a)	Predicted Adherence ($\hat{Y}_a$)
1	0	0.4
2	0	0.75
3	0	0.4
4	0	0.75
5	0	0.4
6	0	0.75
1	1	0.55
2	1	1.0
3	1	0.55
4	1	1.0
5	1	0.55
6	1	1.0

Simple Example: G-computation

Expanded Data with Predicted Outcomes

ID	Pill Box (a)	Predicted Adherence ($\hat{Y}_a$)
1	0	0.4
2	0	0.75
3	0	0.4
4	0	0.75
5	0	0.4
6	0	0.75
1	1	0.55
2	1	1.0
3	1	0.55
4	1	1.0
5	1	0.55
6	1	1.0

Estimate of $E_W(E(Y|A=0,W)$= 0.575 (equal to $E(Y_0)$) if W satisfies the back door criterion)

$$\frac{1}{n}\sum_{i=1}^{n}\hat{E}(Y|A=0,W_i)=0.575$$

Estimate of $E_W(E(Y|A=1,W)$ (equal to $E(Y_1)$) if W satisfies the back door criterion)

$$\frac{1}{n}\sum_{i=1}^{n}\hat{E}(Y|A=1,W_i)=0.775$$

Simple Example: G-computation

- Estimate of $E_0[Y|A=1]-E_0[Y|A=0]$ (confounded association between pill box use and adherence):
  - 0.7-0.63=0.07

- Estimate of $E_W[E_0(Y|A=1,W)-E_0(Y|A=0,W)]$
  - 0.775-0.575=0.20
  - An estimate of $E[Y_1-Y_0]$ (effect of pill box use on adherence) if W satisfies the backdoor criteria

Note on Intuition

- Not really estimating what each subject's counterfactual outcome would have been…
  - In that case, we would not simulate the outcomes corresponding to the treatments we observed
  - This is just a hueristic to give some intuition
- Really, we are just implementing a substitution estimator
  - Plugging estimate of Q_0 into the parameter mapping Ψ

$$\hat{\Psi}(P_n) = \Psi(Q_n) = \frac{1}{n}\sum_{i=1}^{n}[\bar{Q}_n(1, W_i) - \bar{Q}_n(0, W_i)]$$

How to estimate $E_0(Y|A,W)$?

- NPMLE breaks down quickly if A and/or W are continuous or have multiple levels
  - As occurs when W has multiple components
  - End up with sparse or empty cells

Empirical Mean of Y within strata defined by (A,W)

	W=0	W=1	…	W=100	…
A=1	310 (n=1)	66 (n=12)		40 (n=30)	
A=0	10 (n=60)	5 (n=4)		?? (n=0)	

- We need alternative approaches to non-parametric estimation in this (very common) setting
  - Coming up next lecture.....

How else might we estimate $E_0(Y|A,W)$?

- Say we knew that this conditional expectation could be described by a lower dimensional parametric model
- We have real knowledge about the functional form of the relationship between the expectation of Y and (A,W)
  - i.e. Our statistical model is not Non parametric

How else might we estimate $E_0(Y|A,W)$?

- Ex. We know that

  $E(Y|A,W)=\beta_0+\beta_1 A+\beta_2 W+\beta_3 A*W$ for some β

- We can estimate β and thereby $E(Y|A,W)$ by fitting a simple linear regression

G-computation vs. MV Regression

- If $E_0(Y|A,W)$ is estimated using <u>a linear model without interactions</u> between A and W,
  - Estimated coefficient on treatment is equivalent to the G-computation estimate of the ATE
- Ex: Estimate of E[Y|A,W] :

$$\bar{Q}_n(A,W) = \hat{E}(Y|A,W) = \hat{\beta}_0 + \hat{\beta}_1 A + \hat{\beta}_2 W$$

- Estimate of ATE:

$$\hat{\Psi}(Q_n) = \frac{1}{n}\sum_{i=1}^{n}(\hat{E}(Y|A=1,W_i) - \hat{E}(Y|A=0,W_i))$$

$$= \frac{1}{n}\sum_{i=1}^{n}\hat{\beta}_1 = \hat{\beta}_1$$

G-computation vs. MV Regression

- If $E_0(Y|A,W)$ is estimated using a linear model <u>with interactions</u> between A and W
- Then the coefficients in the regression model provide a conditional effect estimate
  - Average treatment effect for a a given value of W
  - Average with respect to distribution of W to estimate the ATE

$$\bar{Q}_n(A,W) = \hat{E}(Y|A,W) = \hat{\beta}_0 + \hat{\beta}_1 A + \hat{\beta}_2 W + \hat{\beta}_3 WA$$

$$\begin{aligned}\hat{\Psi}(Q_n) &= \frac{1}{n}\sum_{i=1}^{n} \hat{E}(Y|A=1,W_i) - \hat{E}(Y|A=0,W_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \hat{\beta}_1 + \hat{\beta}_3 W_i \\ &= \hat{\beta}_1 + \hat{\beta}_3 \hat{E}(W)\end{aligned}$$

G-computation vs. MV Regression

- If $E_0(Y|A,W)$ is estimated using a nonlinear model
  - Ex. Logistic regression

$$\bar{Q}_n(A,W) = \hat{E}(Y|A,W) = \frac{1}{1 + exp^{-(\hat{\beta}_0 + \hat{\beta}_1 A + \hat{\beta}_2 W)}}$$

- Then the coefficient on A in the regression model provides a conditional effect estimate
  - Ex: Conditional casual odds ratio

$$exp(\hat{\beta}_1) = \frac{\hat{E}(Y|A=1,W)/(1-\hat{E}(Y|A=1,W))}{\hat{E}(Y|A=0|W)/(1-\hat{E}(Y|A=0,W))}$$

$$= \frac{\hat{E}(Y_1|W)/(1-\hat{E}(Y_1|W))}{\hat{E}(Y_0|W)/(1-\hat{E}(Y_0|W))}$$

G-computation vs. MV Regression

- Regardless of how $E_0(Y|A,W)$ is estimated, can use the G-comp formula to get an estimate of the ATE
  - Or other target causal quantity that is a function of $E(Y_a)$
- Example: From Logistic regression to ATE

$$\bar{Q}_n(A,W) = \hat{E}(Y|A,W) = \frac{1}{1+exp^{-(\hat{\beta}_0+\hat{\beta}_1 A+\hat{\beta}_2 W)}}$$

$$\hat{\Psi}(Q_n) = \frac{1}{n}\sum_{i=1}^{n}\hat{E}(Y|A=1,W_i) - \frac{1}{n}\sum_{i=1}^{n}\hat{E}(Y|A=0,W_i)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{1+exp^{-(\hat{\beta}_0+\hat{\beta}_1+\hat{\beta}_2 W_i)}} - \frac{1}{n}\sum_{i=1}^{n}\frac{1}{1+exp^{-(\hat{\beta}_0+\hat{\beta}_2 W_i)}}$$

General Implementation of G-Computation for point treatment

1. Estimate $\bar{Q}_0(A,W) = E_0(Y|A,W)$
2. Use this estimate to generate a predicted outcome for each subject setting A=1 and setting A=0
   - Intuition: Mimics study where each individual received and did not receive the treatment
3. Estimate $\Psi(P_0)$ as the difference in the mean of these predicted outcomes

Take home points

- Under specific conditions, the coefficient on exposure in a regression model equals the average treatment effect
- However, in many cases it does not
- It may still have a casual interpretation- eg it may be estimating a different casual parameter

Take home points

- Parametric multivariable regression is just one way to estimate $E(Y|A,W)$
- The resulting estimator can be plugged into the G-comp formula to get an estimate of the average treatment effect
- Whether or not this is a good idea depends on whether the regression is misspecified

Why do we need new tools?

- Even for a simple estimand like the Gcomp formula

1. NP MLE often breaks down in practical data settings: Sparse/empty cells
2. We often do <u>not</u> know that E(Y|A,W) can be described by a <u>lower dimensional parametric model</u>
   - Our true statistical model is non parametric

- We might still decide to estimate the conditional expectation by fitting the parameters of such a parametric model…

Why do we need new tools?

- Ex. We we do not know that

  $E(Y|A,W)=\beta_0+\beta_1 A+\beta_2 W+\beta_3 A*W$ for some β
- However, we can still decide to estimate β and thereby E(Y|A,W) by fitting a simple linear regression
- However, if our model is wrong it may result in a bad estimate, and thus a poorly performing (biased) estimator

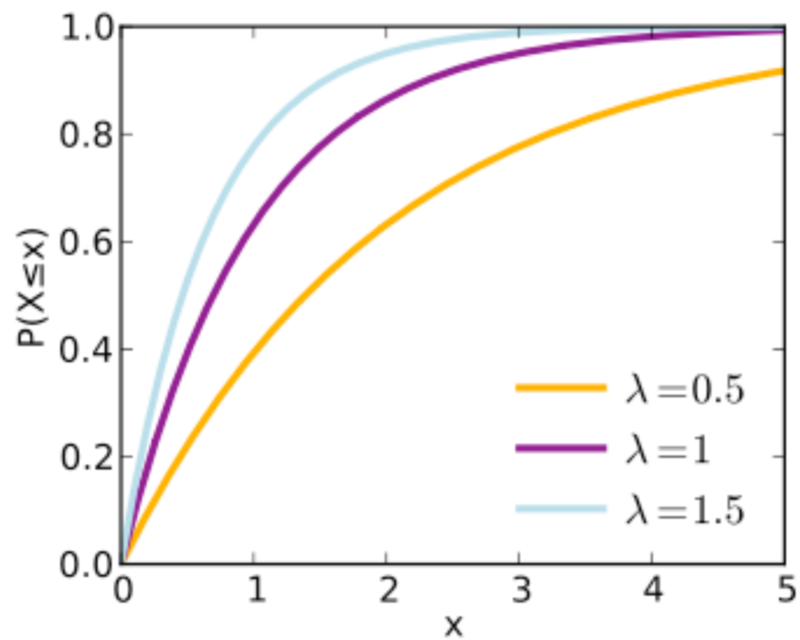
Motivation for Data adaptive approaches

- Often a statistical model that accurately represents our knowledge is non-parametric
  - Distribution of the observed data can take any form…
- If our statistical model does not represent our knowledge, it may not contain the truth
  - This can lead to biased estimators
- If we use an estimator that does not respect our true statistical model, it can lead to bias

Example: Why should we respect our model?

- Simple Example: X= Survival Time
- Estimand: P_0 (X ≤ 2 years)
- Say we know X is exponentially distributed
  - Model: the set of exponential distributions

$$F(x;\lambda) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$



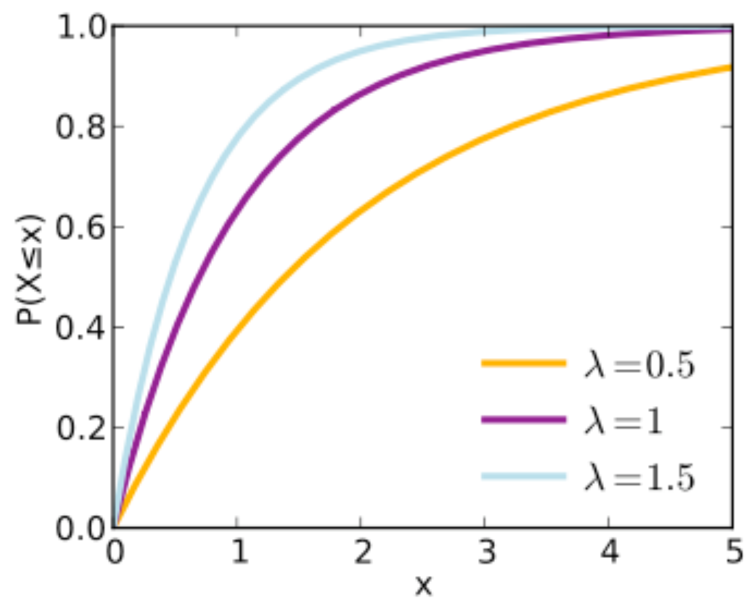
Example (1)

- Model: The set of exponential distributions
- To estimate $P_0(X \leq 2 \text{ years})$, we can just estimate λ
  - Gives us an estimate of the whole distribution of X (and thus an estimate of our target parameter)

$$F(x;\lambda) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

MLE estimate:

$$\hat{\lambda} = \frac{1}{1/n \sum_{i=1}^{n} x_i}$$



Example (2)

- We know nothing about the distribution of X
- Model: Non-parametric
  - Puts no restrictions on the allowed distributions for X
  - This doesn't mean we assume that X is <u>not</u> exponentially distributed, it just means we consider more possibilities

Example (2): Option 1

- We don't know anything about the distribution of X
- We could assume it is exponential (ie assume an exponential model)
  - This model does not respect the limits of our knowledge!!
- This route suggests one possible estimator:
  - MLE : $\hat{\lambda} = \frac{1}{1/n \sum_{i=1}^{n} x_i}$ $\hat{P}(X \leq 2) = 1 - \exp^{-\hat{\lambda}2}$

Example (2): Option 2

- We don't know anything about the distribution of X
- We thus assume a non-parametric model
- This suggests a different estimator
  - A natural non-parametric estimator: the sample proportion

$$\hat{P}(X \leq 2) = \frac{\sum_{i=1}^{n} I(X_i \leq 2)}{n}$$

  - Doesn't assume anything about the distribution of X

- Lets compare these two estimators….

Estimator performance

- Because an estimator is a function of random variables, it is itself a random variable
  - It has a distribution
- We can talk about its performance across many samples of size n (realizations P_n) drawn from the same underlying distribution P_0
- A few common measures of performance
  - Bias
  - Variance
  - Mean Squared Error

Some benchmarks for estimators

- Bias: How does the expectation of the estimator differ from the true parameter value?

$$Bias\left(\hat{\Psi}(P_n)\right) = E_0\left(\hat{\Psi}(P_n) - \Psi(P_0)\right)$$

- Variance: How much does the estimator vary across samples?

$$Variance\left(\hat{\Psi}(P_n)\right) = E_0\left[\left(\hat{\Psi}(P_n) - E_0(\hat{\Psi}(P_n))\right)^2\right]$$

- Mean Squared Error: On average, how far is the estimator from the truth?

$$MSE\left(\hat{\Psi}(P_n)\right) = E_0\left[\left(\hat{\Psi}(P_n) - \Psi(P_0)\right)^2\right]$$

Simple simulations

- Observed data: 200 i.i.d. copies of X drawn from an unknown distribution
- Target Parameter: $P_0(X \leq 2 \text{ years})$,
- Simulation 1
  - X~Exponential (rate λ=0.36)

$$F(x;\lambda) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

- Simulation 2
  - X ~ Weibull (shape k=5; scale λ=3)

$$F(x;k,\lambda) = \begin{cases} 1 - e^{-(x/\lambda)^k}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

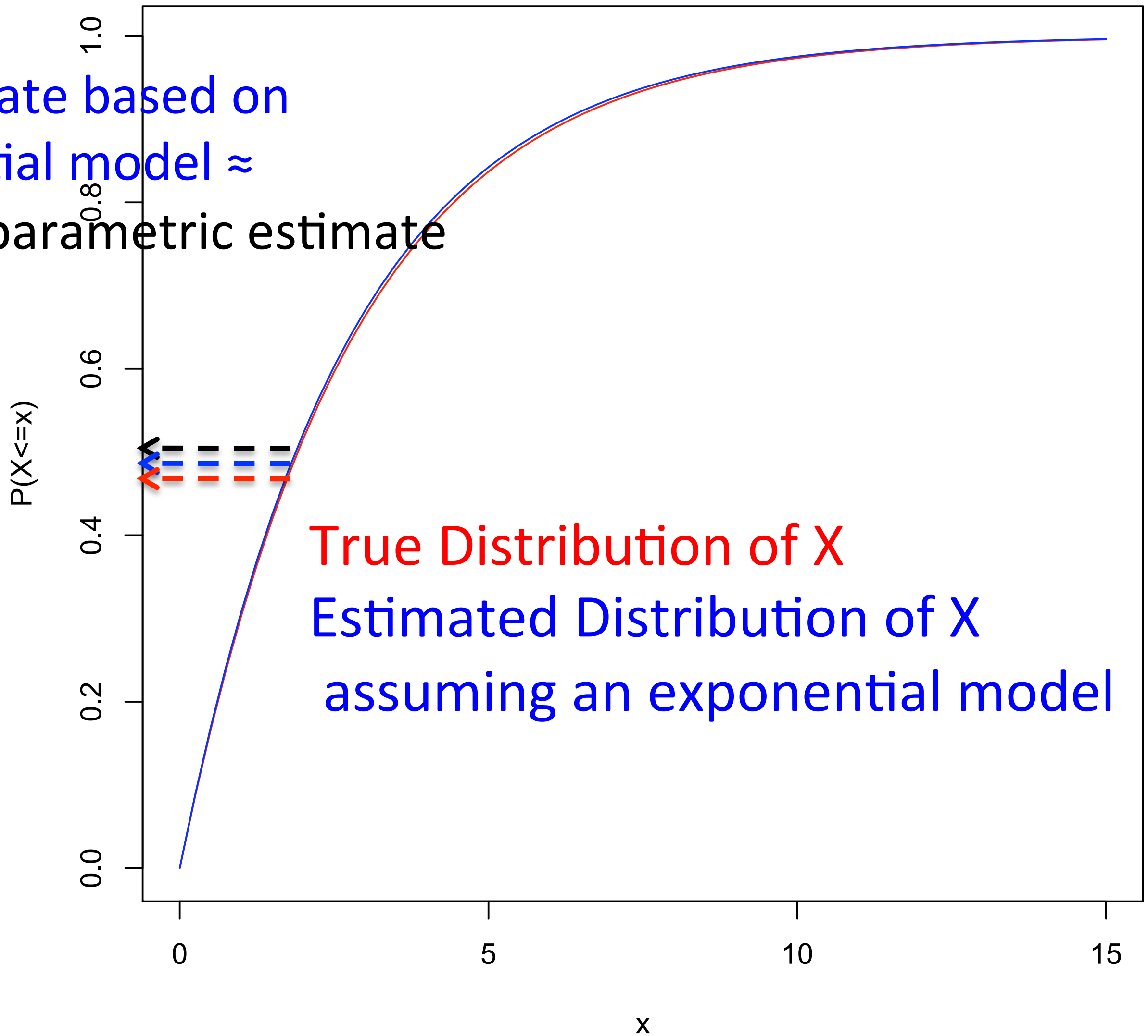
Results: Simulation 1 (X~Exponential)

- Bias/variance estimated based on 2000 samples each of size 200

Estimator	Truth	Mean estimate	Bias	Variance
Parametric (exponential model)	0.52	0.52	9e-4	5e-4
Non-parametric (sample proportion)	0.52	0.52	5e-4	1e-3

Results: Simulation 1 (X~Exponential)

Truth ≈
Avg Estimate based on
exponential model ≈
Avg Non-parametric estimate

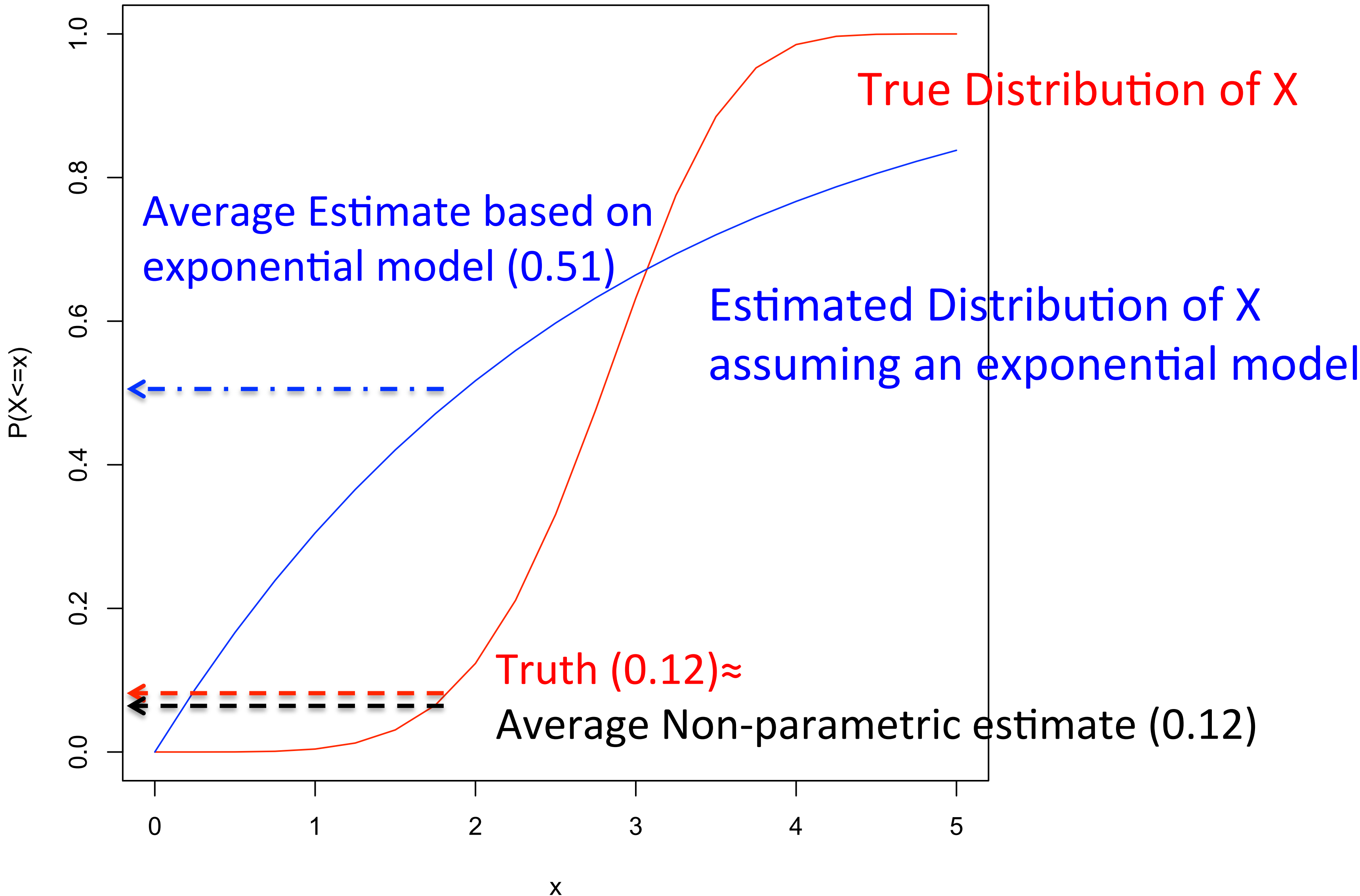


Results: Simulation 2 (X~Weibull)

- Bias/variance estimated based on 2000 samples each of size 200

Estimator	Truth	Mean estimate	Bias	Variance
Parametric (exponential model)	0.12	0.51	0.39	3e-5
Non-parametric (sample proportion)	0.12	0.12	3e-4	5e-4

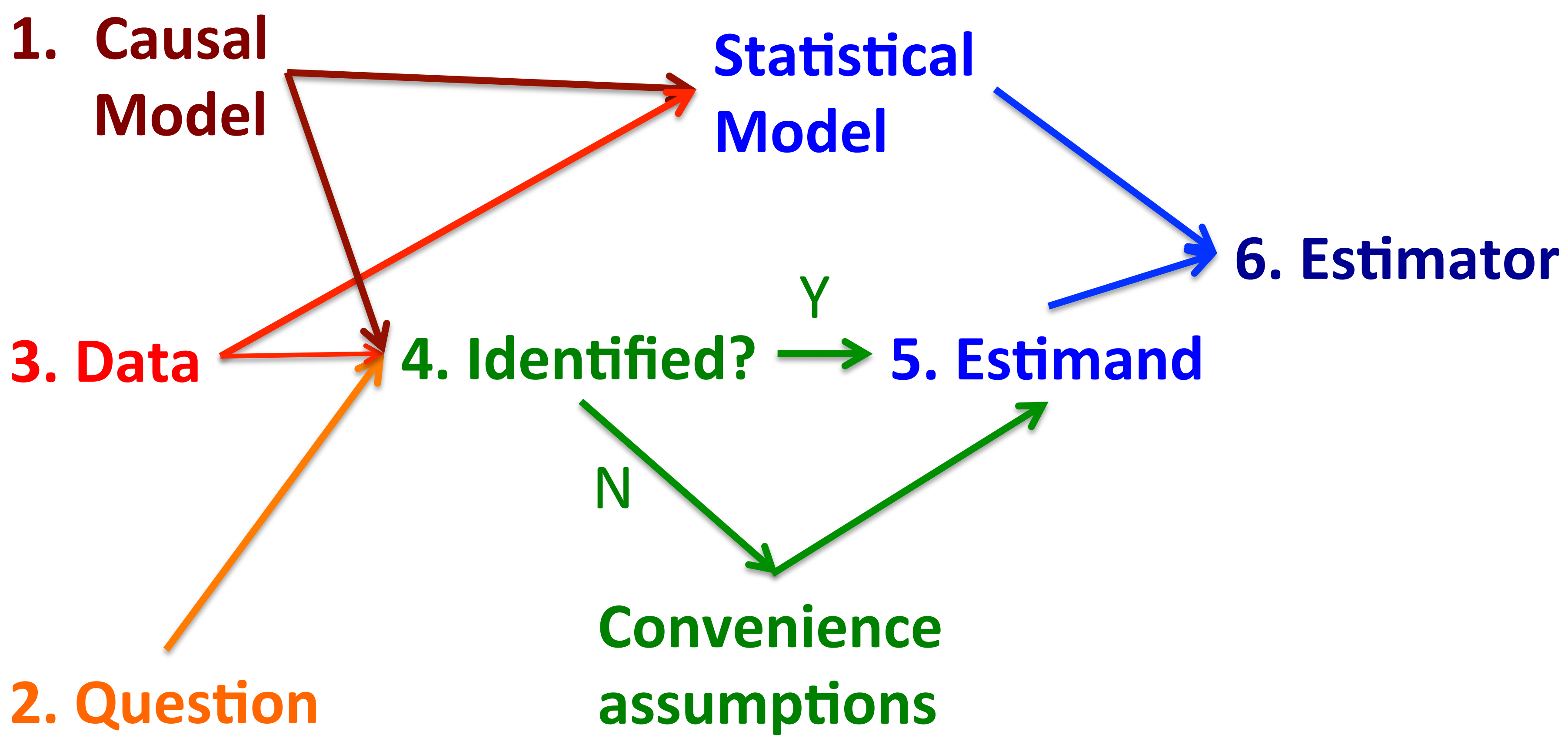
Results: Simulation 1 (X~Weibull)



This is a simple example

- There was an easy alternative here: the sample proportion provides a natural non-parametric estimator
- Real life is harder
  – More variables; More complex target parameters
- Coming up next…Estimation using high dimensional data in non-parametric statistical models

A Roadmap....



1. Causal Model
Statistical Model
6. Estimator
Y
3. Data
4. Identified?
5. Estimand
N
Convenience assumptions
2. Question

Key Points

- Parameter: a function with input a distribution in the statistical model and output a value in the parameter space
- Estimator: a function with input the observed data and output a value in the parameter space

- Simple substitution estimator for MSM parameter
  - Generate predicted values for each subject under each exposure of interest and regress on the MSM
- An estimator that does not respect statistical model can lead to poor estimates
  - Some measures of estimator performance: Bias, Variance, MSE