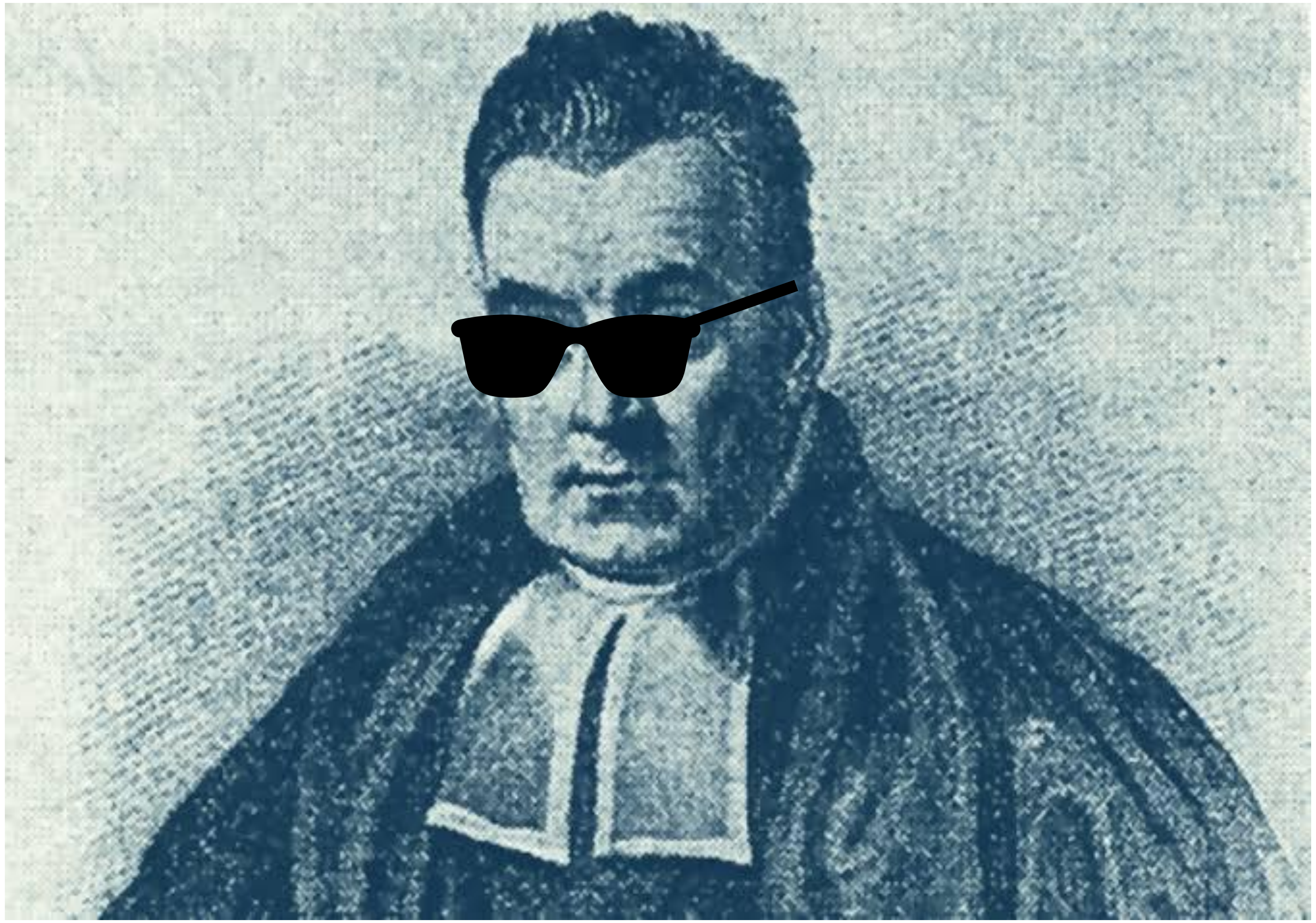
Causal Inference as a Missing Data Problem

(Bayes' Version)

Cindy J. Pang

Casual Inference Journal Club – November 12, 2025



Data Set-Up

Individual (i)	Potential Outcomes $Y_i(0)$	$Y_i(1)$	Actual Treatment $W_i = \begin{cases} 1 \text{ if treated} \\ 0 \text{ if control} \end{cases}$	Observed Outcome Y_i^{obs}
1	66	?	0	66
2	0	?	0	0
3	0	?	0	0
4	?	0	1	0
5	?	607	1	607
6	?	436	1	436

Objective: we are interested in estimating ? or Y^{mis}

Bayes 101 (Missing Data's Version)

$$Pr(A|B) = \frac{Pr(A \cap B)}{Pr(B)} = \frac{Pr(B|A)Pr(A)}{Pr(B)}$$

Drop P(B) since it's a normalizing constant

$$\underbrace{Pr(A|B)}_{\text{Posterior}} \propto \underbrace{Pr(B|A)}_{\text{Likelihood}} \times \underbrace{Pr(A)}_{\text{Prior}}$$

Posterior
the quantity
we want

Likelihood
Updates the
Prior

Prior
We *aSsUmE*
this

Let $\mathcal{U}$ denote the Unknowns and $\mathcal{K}$ denote the Knowns:

$$Pr(\mathcal{U}|\mathcal{K}) \propto Pr(\mathcal{K}|\mathcal{U}) \times Pr(\mathcal{U})$$

Bayes 101 (Missing Data Version)

We write $\mathcal{U} = Y^{mis}$ and $\mathcal{K} = Y^{obs}, W, X, \theta$ where X is a covariate matrix and θ is a model parameter that governs some said distribution. Then we rewrite our problem as:

$$Pr\left(\boldsymbol{Y}^{\text{mis}} \middle| \boldsymbol{Y}^{\text{obs}}, \boldsymbol{W}, \boldsymbol{X}, \theta\right) \propto Pr(\boldsymbol{Y}(0), \boldsymbol{Y}(1), \boldsymbol{W}, \boldsymbol{X}, \theta)$$

$$\propto \prod_{i=1}^{N} Pr(W_i | Y_i(0), Y_i(1), X_i, \theta) \cdot Pr(Y_i(0), Y_i(1) | \, X_i, \theta) \ \cdot Pr(X_i | \theta)$$

This is kinda complicated tho… so let's introduce….

Bayes 101 (Missing Data Version)

We write $\mathcal{U} = Y^{mis}$ and $\mathcal{K} = Y^{obs}, W, X, \theta$ where X is a covariate matrix and θ is a model parameter that governs some said distribution. Then we rewrite our problem as:

$$Pr\left(\boldsymbol{Y}^{\text{mis}} \middle| \boldsymbol{Y}^{\text{obs}}, \boldsymbol{W}, \boldsymbol{X}, \theta\right) \propto Pr(\boldsymbol{Y}(0), \boldsymbol{Y}(1), \boldsymbol{W}, \boldsymbol{X}, \theta)$$

$$\propto \prod_{i=1}^{N} Pr(W_i | Y_i(0), Y_i(1), X_i, \theta) \cdot Pr(Y_i(0), Y_i(1) | X_i, \theta) \cdot Pr(X_i | \theta)$$

Unconfoundedness $(W_i \perp \left(Y_i(0), Y_i(1)\right) | X_i)$ allows us to drop $Pr(W_i | Y_i(0), Y_i(1), X_i, \theta)$ *and* $Pr(X_i | \theta)$

$$\propto \prod_{i:W_i=0} Pr(Y_i(1) | Y_i(0), X_i, \theta) \prod_{i:W_i=1} Pr(Y_i(0) | Y_i(1), X_i, \theta)$$

Toy Example

SUPPOSE $(Y_i(0), Y_i(1))$ has the following joint distribution:

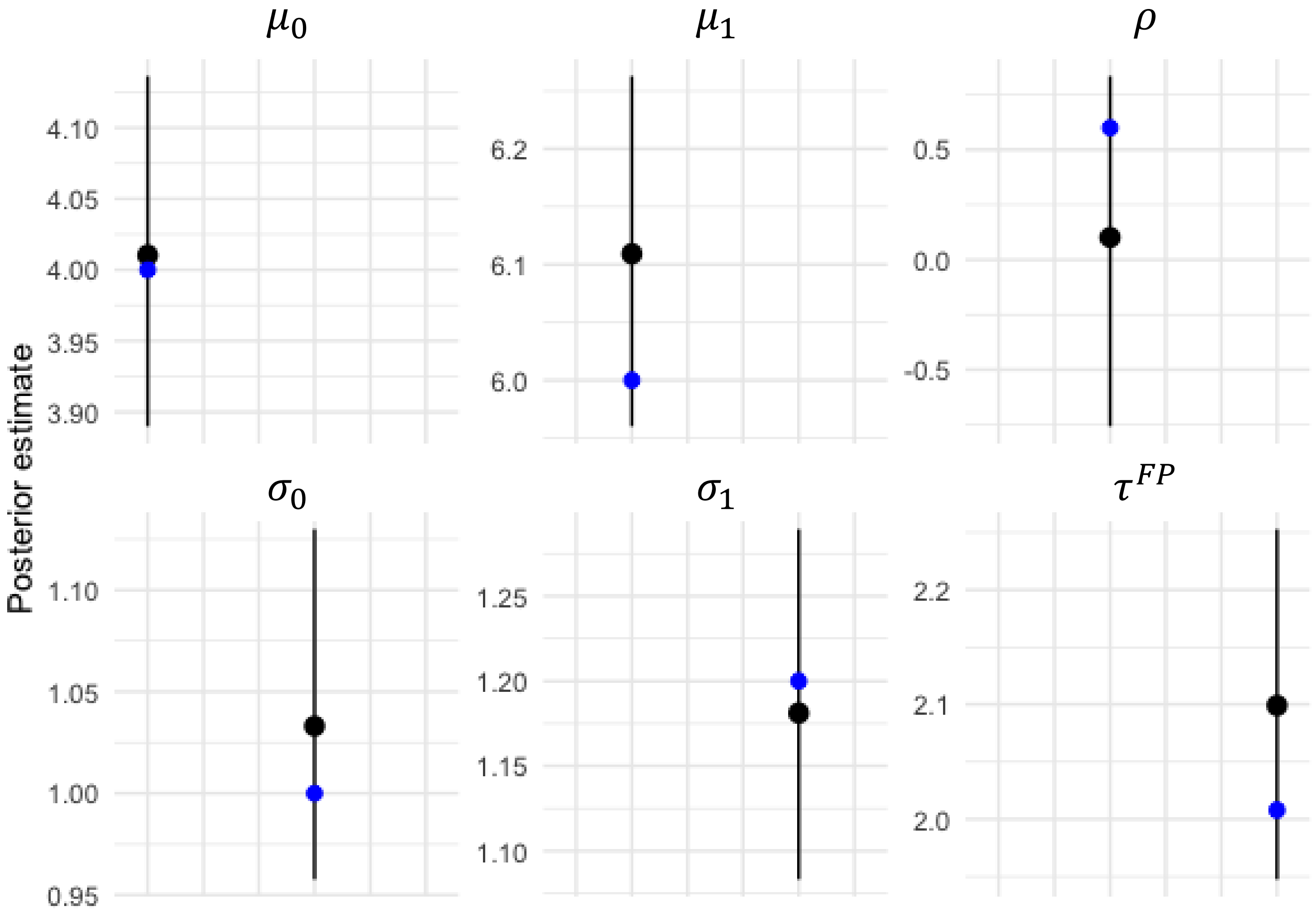
$$\binom{Y_i(0)}{Y_i(1)} \sim N(\mu, \Sigma) \text{ where } \mu = (\mu_0, \mu_1) \text{ and } \Sigma = \begin{pmatrix} \sigma_0^2 & \rho\sigma_0\sigma_1 \\ \rho\sigma_0\sigma_1 & \sigma_1^2 \end{pmatrix}$$

We observe $Y_i^{obs} = W_i Y_i(1) + (1 - W_i) Y_i(0)$ so some of the $Y_i(0)$'s and $Y_i(1)$'s are observed.

In this example, we want to find $\theta = (\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$ and subsequently capture the true Average Treatment Effect (ATE) for the finite population setting:

$$\tau^{FP} = \frac{1}{N} \sum_{i=1}^{N} Y_i(1) - Y_i(0)$$

Toy Example



μ_0
μ_1
ρ
σ_0
σ_1
τ^{FP}
Posterior estimate
4.10
4.05
4.00
3.95
3.90
6.2
6.1
6.0
0.5
0.0
-0.5
1.10
1.05
1.00
0.95
1.25
1.20
1.15
1.10
2.2
2.1
2.0

Toy Example – Some Issues

We never observe $(Y_i(0), Y_i(1))$ jointly, which raises concerns about whether ρ is **identifiable**

SUPPOSE $(Y_i(0), Y_i(1))$ has the following joint distribution:

$$\binom{Y_i(0)}{Y_i(1)} \sim N(\mu, \Sigma) \text{ where } \mu = (\mu_0, \mu_1) \text{ and } \Sigma = \begin{pmatrix} \sigma_0^2 & \rho\sigma_0\sigma_1 \\ \rho\sigma_0\sigma_1 & \sigma_1^2 \end{pmatrix}$$

We observe $Y_i^{obs} = W_i Y_i(1) + (1 - W_i) Y_i(0)$ so some of the $Y_i(0)$'s and $Y_i(1)$'s are observed.

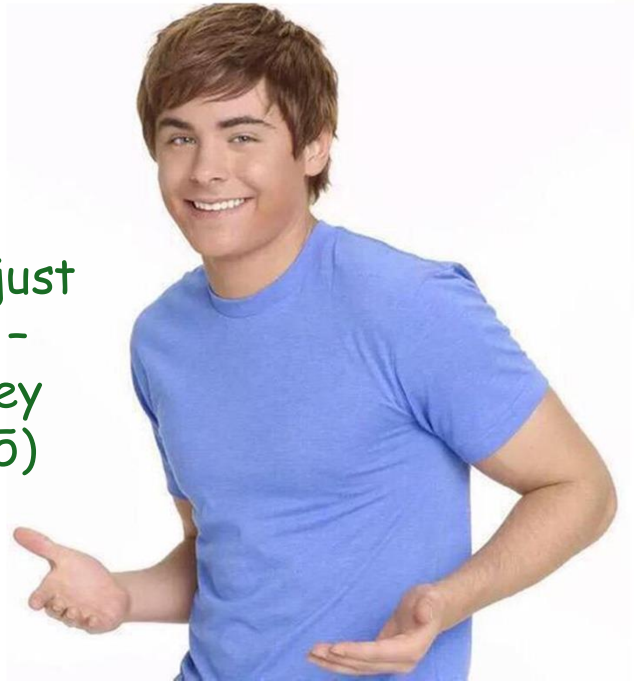
Toy Example – Identifiability

- ***Frequentists:*** A parameter θ is identifiable if

$$\theta \coloneqq f\left(Y_i^{obs}\right)$$

- ***Bayesians:***

"they just are" – Lindley (2015)



There is a thing called "weakly" or "partially identifiable" – Gustafson (2015)

No real consensus – Bayesians

Toy Example – Some Issues

SUPPOSE $(Y_i(0), Y_i(1))$ has the following joint distribution:

$$\binom{Y_i(0)}{Y_i(1)} \sim N(\mu, \Sigma) \text{ where } \mu = (\mu_0, \mu_1) \text{ and } \Sigma = \begin{pmatrix} \sigma_0^2 & \rho\sigma_0\sigma_1 \\ \rho\sigma_0\sigma_1 & \sigma_1^2 \end{pmatrix}$$

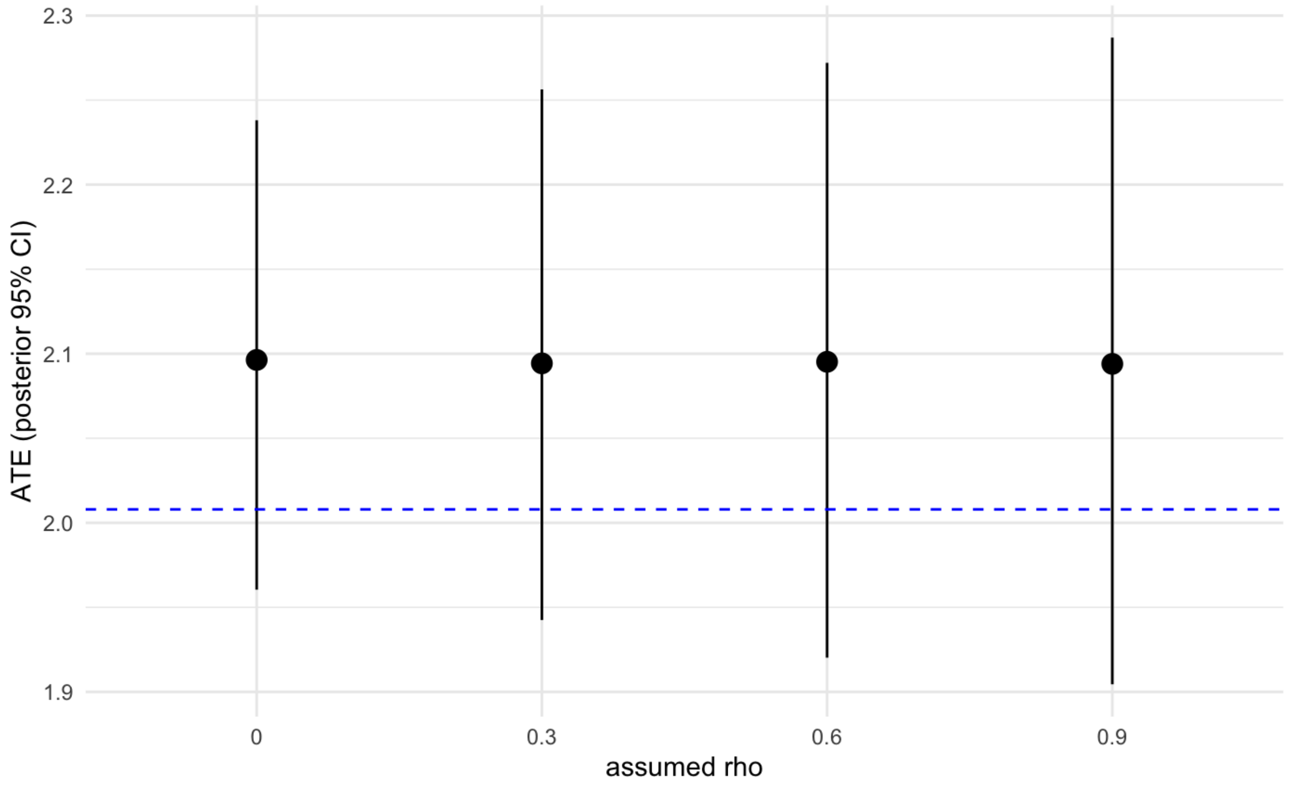
We observe $Y_i^{obs} = W_i Y_i(1) + (1 - W_i) Y_i(0)$ so some of the $Y_i(0)$'s and $Y_i(1)$'s are observed.

Strategy 2: Separate $\theta = (\theta^a, \theta^m)$ where $\theta^m = (\mu_1, \mu_2, \sigma_0, \sigma_1)$ and $\theta^a = \rho$. Now our posterior becomes:

$$Pr\left(\boldsymbol{Y}^{\text{mis}} \middle| \boldsymbol{Y}^{\text{obs}}, \boldsymbol{W}, \boldsymbol{X}, \theta\right) \propto Pr(\theta^a) Pr(\theta^m) \prod_{i:W_i=1} Pr(Y_i(1) | Y_i(0), X_i, \theta^m) \prod_{i:W_i=0} Pr(Y_i(0) | Y_i(1), X_i, \theta^m)$$

where $\theta^a \perp \theta^m$

Toy Example – Strategy 2



➔ Varying ρ doesn't impact the ATE that much

the association (which you don't like *actually* have) does not inform ρ

In other words, ρ doesn't "learn" from missing data

Takeaways

- According to Ding and Li (2018) we still need to find like a "Frequentist"-like prior
- Future Bayesian Research: find better priors duh
- Cindy's Hot Take: find something better to do with your life than this (don't tell Dr. Banerjee)
  - But if you were to do it, do a sensitivity analysis for associational parameters (ρ)