Selection bias in 2020

Qingyuan Zhao

Statistical Laboratory, University of Cambridge

December 8, 2020 @ Online Causal Inference Seminar

Selection bias: An umbrella term

- The Cambridge Dictionary of Statistics: "The bias that may be introduced into all types of scientific investigations whenever **a treatment is chosen by the individual involved or is subject to constraints that go unobserved by the researcher**".
- Wikipedia: "the bias introduced by the **selection of individuals, groups or data for analysis in such a way that proper randomization is not achieved**, thereby ensuring that the sample obtained is not representative of the population intended to be analyzed."
- Wikipedia collects many types of selection bias: (non-random) sampling bias; time interval (censoring/truncation); susceptibility bias; indication bias; data dredging; attrition/survivorship bias; observer selection bias; volunteer bias; Berkson's paradox (collider bias).

My experiences: Before 2020

- Don't remember taking a course that involved > 1 lectures about selection bias.
- In consulting sessions, clients often want to know how to interpret their results after an array of data processing and model selection. What should I say?
- Have seen many anecdotes about selection bias.
- Always thought they are far away from me. The current practice cannot be that bad, right?

My experiences: 2020

A statistician's existential crisis

Rest of the talk: Two topical examples

1. Initial estimates of COVID-19's infectiousness and incubation period.
   - Reference: Z, Ju, Bacallado, Shah. (2020). BETS: The dangers of selection bias in early analyses of the coronavirus disease (COVID-19) pandemic. *The Annals of Applied Statistics* (in press). arXiv: 2004.07743.
2. Racial bias in policing.
   - Reference: Z, Keele, Small, Joffe. (2020). A note on post-treatment selection in studying racial discrimination in policing. arXiv: 2009.04832.

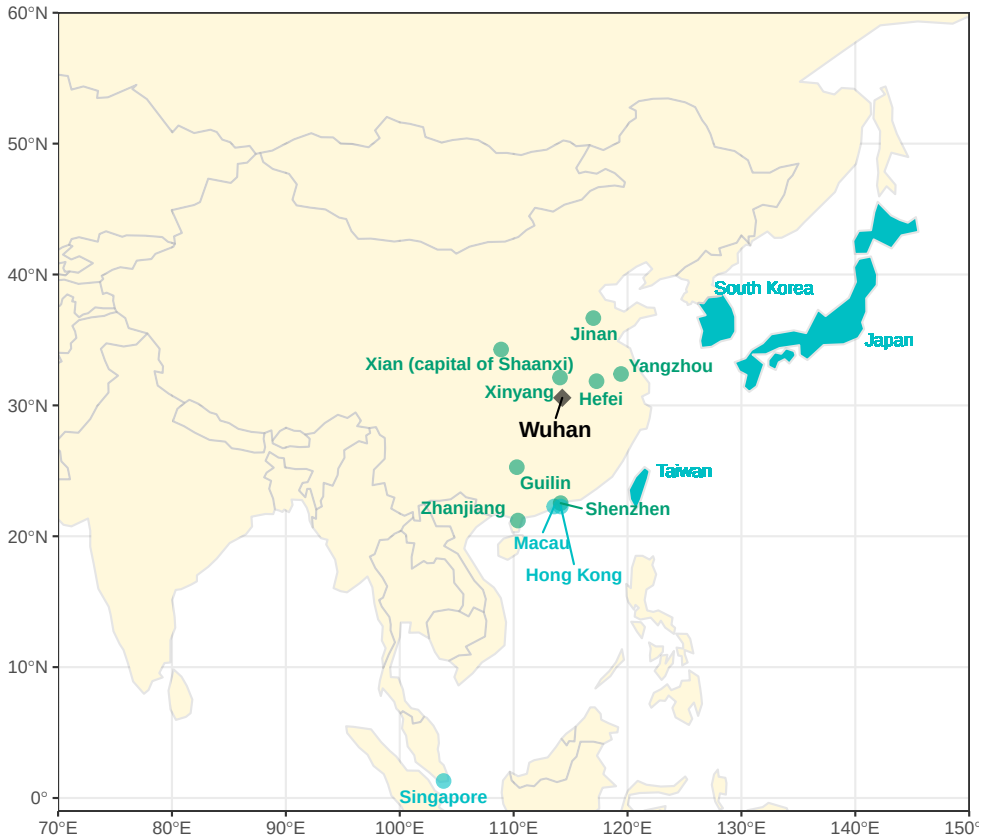
Acknowledgement

Many people who have offered helpful suggestions: Cindy Chen, Yang Chen, Yunjin Choi, Hera He, Michael Levy, Marc Lipsitch, James Robins, Andrew Rosenfeld, Dylan Small, Yachong Yang, Zilu Zhou; Dean Knox, Joshua Loftus, and Jonathan Mummolo.

Initial COVID-19 studies

- Many were based on "exported" cases from Wuhan.
- Extremely influential.
- Many types of selection bias incurred: Under-ascertainment; Non-random sample selection; right-truncation; ignoring travel restrictions and fast epidemic growth on unobserved data.
- Common mistake: New data + Existing model = New results.

Data collection



- 14 locations where the local health agencies published full case reports.
- 1,460 COVID-19 cases that were confirmed by February 29 for locations in mainland China (February 15 for international locations).
- 378 exported cases from Wuhan.

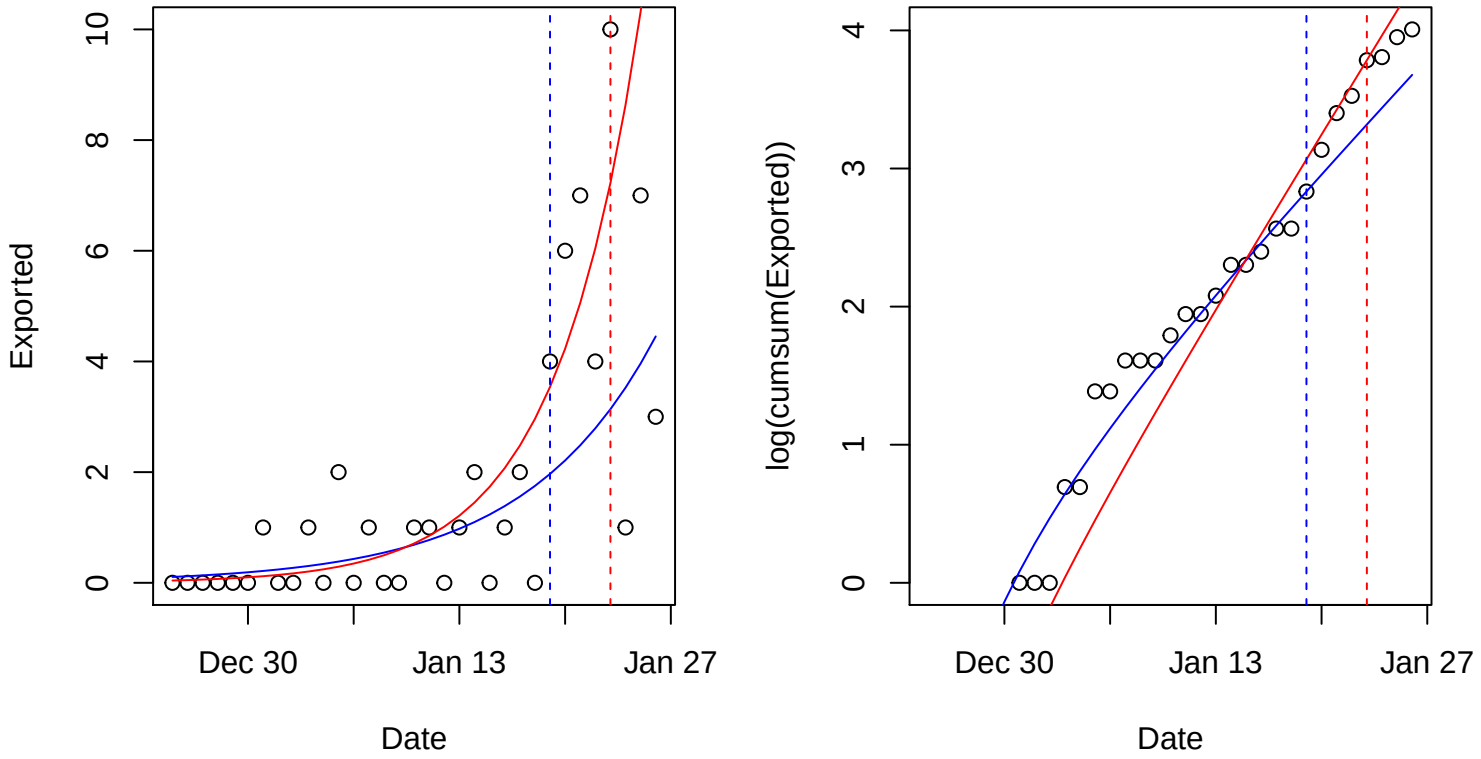
Overview of the dataset

Column name	Description	Example	Summary statistics
`Case`	Unique identifier for each case	HongKong-05	1460 in total
`Residence`	Nationality or residence of the case	Wuhan	21.5% reside in Wuhan
`Gender`	Gender	Male/Female	52.1%/47.7% (0.2% NA)
`Age`	Age	63	Mean=45.6, IQR=[34, 57]
`Known Contact`	Known epidemiological contact?	Yes/No	84.7%/15.3%
`Cluster`	Relationship with other cases	Husband of HongKong-04	32.1% known
`Outside`	Transmitted outside Wuhan?	Yes/Likely/No	58.5%/7.7%/33.8%
`Begin Wuhan`	**Begin of stay in Wuhan (B)**	30-Nov[4]	
`End Wuhan`	**End of stay in Wuhan (E)**	22-Jan	
`Exposure`	Period of exposure	1-Dec to 22-Jan	58.9% known period/date 8.2% known date
`Arrived`	Final arrival date at the location where confirmed a COVID-19 case	22-Jan	40.6% did not travel
`Symptom`	**Date of symptom onset (S)**	23-Jan	9.0% NA
`Initial`	Date of first medical visit	23-Jan	6.5% NA
`Confirmed`	Date confirmed	24-Jan	

Naive method

- Wu, J. T. et al. (2020). Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: A modelling study. *The Lancet*, 395(10225).
- They used a SEIR (Susceptible-Exposed-Infectious-Recovered) model for the epidemic in Wuhan and a Poisson process to model case exportation.
- They fitted the model using 17 (!) international cases who showed symptoms before January 20, 2020.
- To replicate their analysis, I fitted some simple Poisson log-linear models.

Initial doubling time



- **Blue** (using symptom onsets before January 20): 5.9 days (95% CI 3.4–15.7).
- **Red** (before January 24): 3.9 days (2.9–5.5).
- Original study: 6.4 days (5.8–7.1).

Problems

These models

- Do **NOT** take into account Wuhan's travel ban on January 23.
- Ignore the rich information available for the individual cases.

Let's start from the first principles

Four crucial epidemiological events

- B: Beginning of stay in Wuhan;
- E: End of stay in Wuhan;
- T: Time of transmission (unobserved);
- S: Time of symptom onset.

Below we will:

- Define the support $\mathcal{P}$ of (B, E, T, S) for the **Wuhan-exposed** population;
- Construct a generative model for (B, E, T, S);
- Define the sample selection set $\mathcal{D}$ corresponds to **Wuhan-exported** cases;
- Derive likelihood functions to adjust for sample selection.

Wuhan-exposed population $\mathcal{P}$

Intuitively, $\mathcal{P}$ = All people who stayed in Wuhan between 12am December 1, 2019 (time 0) and 12am January 24, 2020 (time L, the lockdown).

Conventions

- $B = 0$: **Started their stay in Wuhan before time 0.**
- $E = \infty$: **Did not arrive in the 14 locations we are considering before time L.** (We do not differentiate between people who stayed in Wuhan or went to a different location).
- $T = \infty$: **Were not infected during their stay in Wuhan.** (We do not differentiate between infection outside Wuhan and never infected.)
- $S = \infty$: **Did not show symptoms of COVID-19** (never infected or asymptomatic).

Under these conventions.

$$\mathcal{P} = \Big\{(b, e, t, s) \mid b \in [0, L], e \in [b, L] \cup \{\infty\}, t \in [b, e] \cup \{\infty\}, s \in [t, \infty]\Big\}.$$

A generative BETS model

$$f(b, e, t, s) = \underbrace{f_B(b) \cdot f_E(e \mid b)}_{\text{travel}} \cdot \underbrace{f_T(t \mid b, e)}_{\text{disease transmission}} \cdot \underbrace{f_S(s \mid b, e, t)}_{\text{disease progression}}.$$

To allow extrapolation from Wuhan-exported sample to Wuhan-exposed population, the BETS model makes two basic assumptions

Assumption 1: Disease transmission independent of travel

$$f_T(t \mid b, e) = \begin{cases} g(t), & \text{if } b < t < e, \\ 1 - \int_b^e g(x)\, dx, & \text{if } t = \infty. \end{cases}$$

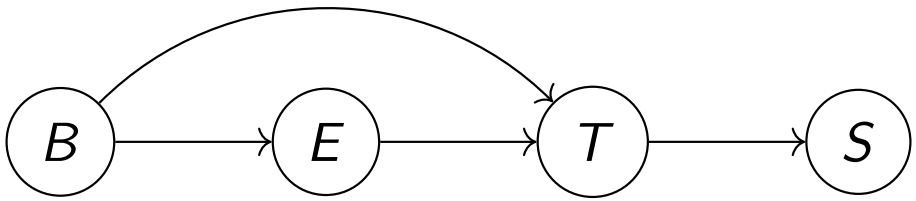
Here $g(\cdot)$ models the **epidemic growth** in Wuhan before the lockdown.

Assumption 2: Disease progression independent of travel

$$f_S(s \mid b, e, t) = \begin{cases} \nu \cdot h(s - t), & \text{if } t < s < \infty, \\ 1 - \nu, & \text{if } s = \infty. \end{cases}$$

Here $h(\cdot)$ is the density of the **incubation period** $S - T$ (for symptomatic cases).

Graphical model representation



- This is in temporal/causal order if we view E as the **planned** traveling date.
- Assumption 1 restricts the density of T given B, E.
- Assumption 2 says that $S \perp\!\!\!\perp B, E \mid T$.

Parametric assumptions

To ease the interpretation and simply the likelihood functions, we assume

Assumption 3: Exponential growth

$$g(t) = g_{\kappa,r}(t) \stackrel{\Delta}{=} \kappa \cdot \exp(rt), \ t \leq L,$$

Assumption 4: Gamma-distributed incubation period

$$h(s-t) = h_{\alpha,\beta}(s-t) \stackrel{\Delta}{=} \frac{\beta^\alpha}{\Gamma(\alpha)}(s-t)^{\alpha-1}\exp\{-\beta(s-t)\}.$$

- Assumptions 3 & 4 are relaxed in a Bayesian nonparametric analysis (see the paper).

Wuhan-exported cases

The event of observing Wuhan-exported cases can be written as

$$\mathcal{D} = \{(b, e, t, s) \in \mathcal{P} \mid b \leq t \leq e \leq L, t \leq s < \infty\}.$$

This makes three further restrictions on $\mathcal{P}$:

1. $B \leq T \leq E$, because we only use cases who contracted the virus during their stay in Wuhan;
2. $E \leq L$, because the case can only be observed if they left Wuhan before the travel ban;
3. $S < \infty$, because we only consider COVID-19 cases who showed symptoms.

Which likelihood function?

For a moment, let's pretend the time of transmission T is observed.

✗ Sample from $\mathcal{P}$

$$\prod_{i=1}^{n} f(B_i, E_i, T_i, S_i)$$

✓ Sample from $\mathcal{D}$ (Unconditional likelihood)

$$\prod_{i=1}^{n} f(B_i, E_i, T_i, S_i \mid \mathcal{D}), \text{ where } f(b, e, t, s \mid \mathcal{D}) \triangleq \frac{f(b, e, t, s) \cdot 1_{\{(b,e,t,s)\in\mathcal{D}\}}}{\mathbb{P}\big((B, E, T, S) \in \mathcal{D}\big)}.$$

✓ Sample from $\mathcal{D}$ (Conditional likelihood)

$$\prod_{i=1}^{n} f(T_i, S_i \mid B_i, E_i, \mathcal{D}), \text{ where } f(t, s \mid b, e, \mathcal{D}) \triangleq \frac{f(t, s \mid B = b, E = e) \cdot 1_{\{(b,e,t,s)\in\mathcal{D}\}}}{\mathbb{P}\big((B, E, T, S) \in \mathcal{D} \mid B = b, E = e\big)}.$$

Unobserved T

In reality, the time of transmission T is unobserved. We can either treat T as a latent variable and use e.g. an EM algorithm, or use the **integrated likelihood**:

Unconditional likelihood

$$L_{\text{uncond}}(\theta) = \prod_{i=1}^{n} \int f\left(B_i, E_i, t, S_i \mid \mathcal{D}\right) dt,$$

where $\theta = (f_B(\cdot), f_E(\cdot \mid \cdot), g(\cdot), h(\cdot))$.

Conditional likelihood

$$L_{\text{cond}}(\theta) = \prod_{i=1}^{n} \int f\left(t, S_i \mid B_i, E_i, \mathcal{D}\right) dt,$$

where $\theta = (g(\cdot), h(\cdot))$.

The conditional likelihood is less efficient because it does not use information in $f(b, e \mid \mathcal{D})$; but it is robust to misspecifying the travel models $f_B(\cdot), f_E(\cdot \mid \cdot)$.

Conditional likelihood function

Proposition

Under Assumptions 1–4,

$$L_{\text{cond}}(r,\alpha,\beta) = \begin{cases} r^n\left(\dfrac{\beta}{\beta+r}\right)^{n\alpha} \cdot \displaystyle\prod_{i=1}^{n} \dfrac{\exp(rS_i)\left[H_{\alpha,\beta+r}(S_i-B_i) - H_{\alpha,\beta+r}((S_i-E_i)_+)\right]}{\exp(rE_i)-\exp(rB_i)}, & \text{for } r>0, \\ \displaystyle\prod_{i=1}^{n} \dfrac{H_{\alpha,\beta}(S_i-B_i) - H_{\alpha,\beta}((S_i-E_i)_+)}{E_i-B_i}, & \text{for } r=0, \end{cases}$$

where $H_{\alpha,\beta}(\cdot)$ is the CDF of Gamma(α,β) and $(\cdot)_+ = \max(\cdot, 0)$.

- This does not depend on ν (proportion of symptomatic cases) and κ (baseline transmission).
- When $r = 0$, this reduces to the likelihood function in Reich et al. (2009) *Statistics in Medicine*, 28.
- The unconditional likelihood function assuming "stable travel" can be found in the paper.

Results

Location	Sample size	Doubling time (in days)	Incubation period Median	Incubation period 95% quantile
		Conditional likelihood		
China - Hefei	34	2.1 (1.2–3.7)	4.3 (2.9–6.0)	12.0 (9.1–17.3)
China - Shaanxi	53	1.7 (1.0–2.8)	4.5 (3.1–6.2)	14.6 (11.5–19.8)
China - Shenzhen	129	2.2 (1.7–3.0)	3.5 (2.8–4.3)	11.2 (9.5–13.6)
China - Xinyang	74	2.3 (1.5–3.5)	6.8 (5.4–8.2)	16.4 (13.8–20.1)
China - Other	42	2.0 (1.1–3.4)	5.1 (3.6–6.7)	12.3 (9.8–16.4)
International	46	2.1 (1.4–3.4)	3.8 (2.5–5.3)	10.9 (8.4–15.1)
All locations	378	2.1 (1.8–2.5)	4.5 (4.0–5.0)	13.4 (12.2–14.8)
		Unconditional likelihood		
China - Hefei	34	1.8 (1.4–2.4)	4.1 (2.8–5.5)	11.9 (9.0–17.2)
China - Shaanxi	53	2.5 (2.0–3.1)	5.3 (3.9–6.8)	15.0 (12.0–20.0)
China - Shenzhen	129	2.4 (2.1–2.8)	3.6 (2.9–4.3)	11.3 (9.6–13.7)
China - Xinyang	74	2.4 (2.0–2.9)	6.8 (5.6–8.1)	16.4 (13.9–20.2)
China - Other	42	2.1 (1.7–2.8)	5.3 (4.0–6.6)	12.4 (10.0–16.4)
International	46	2.0 (1.6–2.6)	3.7 (2.5–5.0)	10.8 (8.4–15.1)
All locations	378	2.3 (2.1–2.5)	4.6 (4.1–5.1)	13.5 (12.3–14.9)

(Point estimates obtained by MLE. Confidence intervals obtained by inverting LRT.)

What's wrong with simple exponential growth?

✗ Density of S in $\mathcal{P}$

It is reasonable to assume incidence of symptom onset is growing exponentially in **Wuhan-exposed population** $\mathcal{P}$:

$$f(s \mid \mathcal{P}) \overset{\propto}{\sim} \exp(rs), \text{ for } s \leq L.$$

But the observations are from the **Wuhan-exported cases** $\mathcal{D}$.

✓ Density of S in $\mathcal{D}$

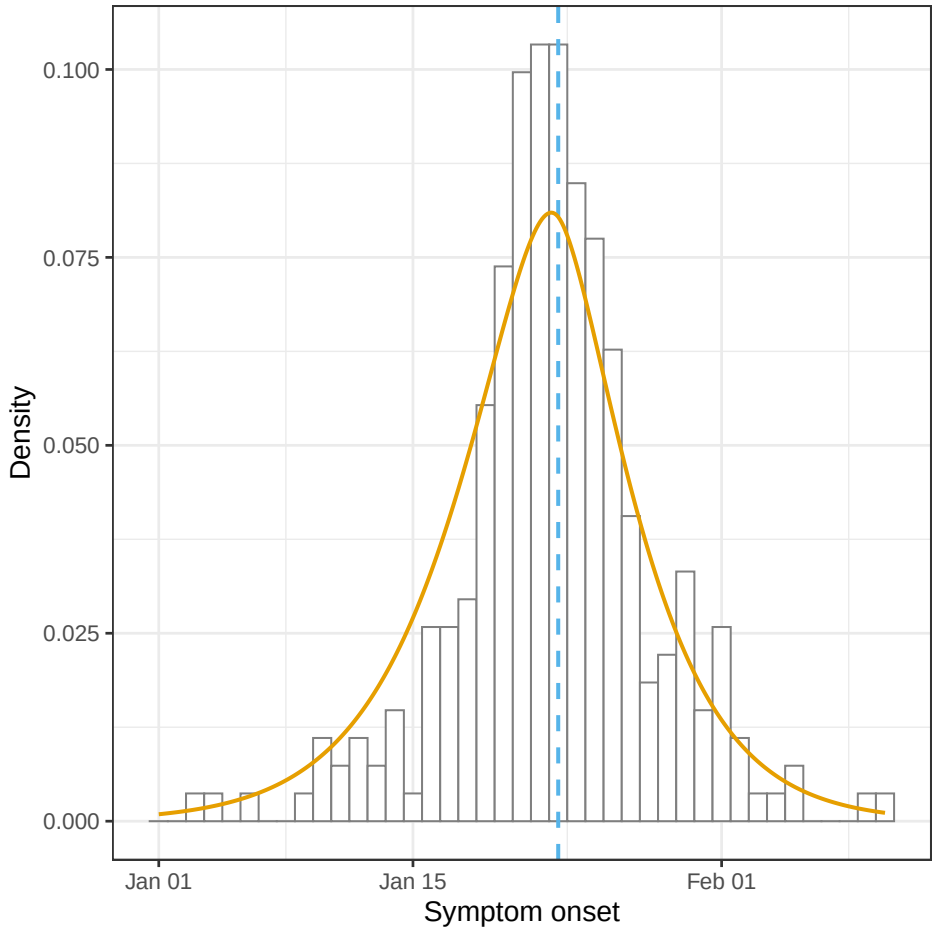
Under Assumptions 1–4 and reasonable approximations,

$$f(t \mid \mathcal{D}) \overset{\propto}{\sim} \exp(rt)\,(L - t)\,1_{\{t \leq L\}},$$

We can further derive the theoretical $f_S(s \mid \mathcal{D})$; in particular,

$$f_S(s \mid \mathcal{D}) \overset{\propto}{\sim} \exp(rs)\left(L + \frac{\alpha}{\beta + r} - s\right), \text{ for } s \leq L.$$

Illustration of the selection bias

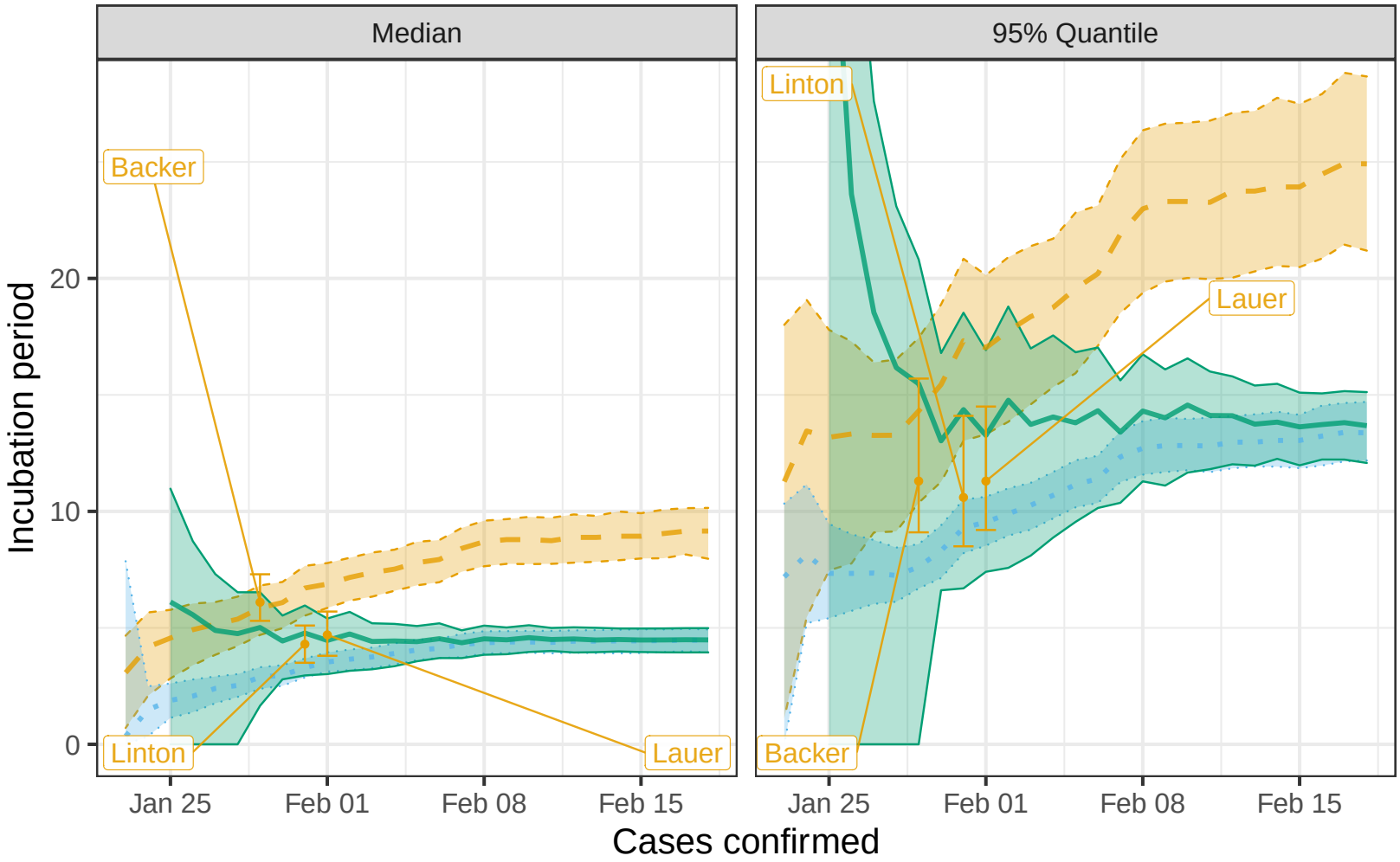


- Histogram: Symptom onsets of Wuhan-exported cases;
- Orange curve: Theoretical fit $f_S(s \mid \mathcal{D})$ using the MLE of (r, α, β).
- Blue dashed line: January 23, 2020 (time L).

Incubation period estimates

An experiment

- For each day between January 23 and February 18, obtain the subset of cases confirmed by that day.
- Fit the parametric BETS model by using one of the following likelihoods:
  1. **Adjusted for nothing:** $L_{\text{cond}}(0, \alpha, \beta)$ (likelihood function in Reich et al. (2009) used in other studies).
  2. **Adjusted for growth:** $L_{\text{cond}}(r, \alpha, \beta)$.
  3. **Adjusted for growth and right-truncation:** $L_{\text{cond,trunc}}(r, \alpha, \beta; M)$ (conditional on $S \leq M$).
- Obtain point estimates by MLE and CIs by nonparametric Bootstrap.
- Compare with previous studies:
  1. Backer, J. A. et al. *Eurosurveillance*, 25(5), 2020. PubMed: 32046819.
  2. Lauer, S. A. et al. *Annals of Internal Medicine*, 2020. PubMed: 32150748.
  3. Linton, N. M. et al. *Journal of Clinical Medicine*, 9(2), 2020. PubMed: 32079150.



Ignore epidemic growth $\Longrightarrow$ Overestimate incubation period.
Ignore right-truncation $\Longrightarrow$ Underestimate incubation period.

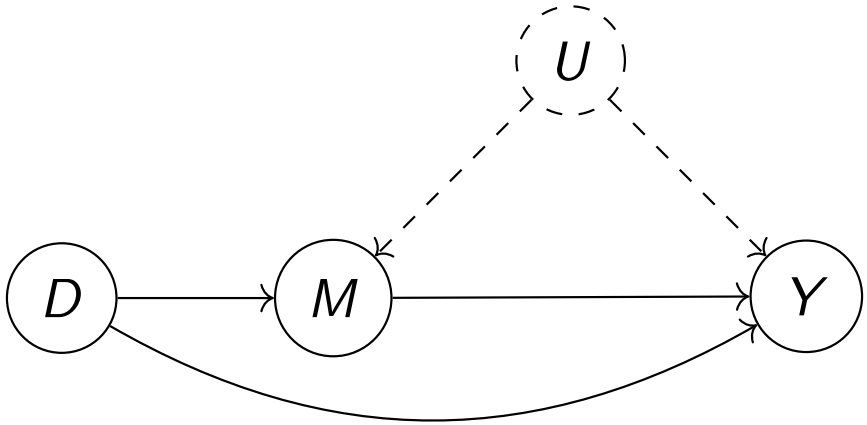
Questions about the first study?

Second example: Racial bias in policing

This work is motivated by

- Knox, D., et al. (2020) Administrative records mask racially biased policing. *American Political Science Review* 114(3).
- Gaebler, J., et al. (2020) A causal framework for observational studies of discrimination. arXiv:2006.12460.

Setup in Knox et al.



- D: binary, 1 means minority.
- M: binary, 1 means police detainment.
- Y: binary, 1 means use of force.

Key challenges

1. Only observe data with $M = 1$ in police admin data.
2. There can be unmeasured M-Y confounders.

$\Longrightarrow$ Collider bias (when conditioning on $M = 1$) in influential studies.

What can be learned from police admin data?

Let $Y(d)$ be the potential outcome for race $D = d$.

Two methods in Knox et al.

1. Partial identification of

$$\text{ATE}_{M=1} = \mathbb{E}[Y(1) - Y(0) \mid M = 1],$$
$$\text{ATT}_{M=1} = \mathbb{E}[Y(1) - Y(0) \mid M = 1, D = 1].$$

2. Identification of $\text{ATE} = \mathbb{E}[Y(1) - Y(0)]$:

Key assumptions in Knox et al.

1. **Mandatory reporting:** $Y(M = 0) = 0$ and all police stops are recorded.
2. **Treatment ignorability:** $D \perp\!\!\!\perp M(d), Y(d, m)$.
3. **Mediator monotonicity:** $M(1) \geq M(0)$. (Not needed for ATE.)

Our results

1. $\text{ATE}_{M=1}$ and $\text{ATT}_{M=1}$ can be difficult to interpret: They may have a different sign even if the natural direct and indirect effects have the same sign.
2. ATE estimation requires estimating the magnitude of $\mathbb{P}(M = 1)$:

$$\begin{aligned}\text{ATE} =& \mathbb{E}[Y \mid D = 1, M = 1]\mathbb{P}(M = 1 \mid D = 1) \\ & - \mathbb{E}[Y \mid D = 0, M = 1]\mathbb{P}(M = 1 \mid D = 0).\end{aligned}$$

This can be circumvented by considering the risk ratio:

$$\text{RR} = \frac{\mathbb{E}[Y(1)]}{\mathbb{E}[Y(0)]} = \underbrace{\frac{\mathbb{E}[Y \mid D = 1, M = 1]}{\mathbb{E}[Y \mid D = 0, M = 1]}}_{\text{naive estimator}} \cdot \underbrace{\left\{\frac{\mathbb{P}(D = 1 \mid M = 1)}{\mathbb{P}(D = 0 \mid M = 1)}\right\} \Big/ \left\{\frac{\mathbb{P}(D = 1)}{\mathbb{P}(D = 0)}\right\}}_{\text{selection bias factor}}.$$

How large is the selection bias?

$$\mathrm{RR} = \frac{\mathbb{E}[Y(1)]}{\mathbb{E}[Y(0)]} = \underbrace{\frac{\mathbb{E}[Y \mid D=1, M=1]}{\mathbb{E}[Y \mid D=0, M=1]}}_{\text{naive estimator}} \cdot \underbrace{\left\{\frac{\mathbb{P}(D=1 \mid M=1)}{\mathbb{P}(D=0 \mid M=1)}\right\} \Big/ \left\{\frac{\mathbb{P}(D=1)}{\mathbb{P}(D=0)}\right\}}_{\text{selection bias factor}}.$$

- Police admin data: NYPD stop-and-frisk.
- We estimated $\mathbb{P}(D=1)$ using two external surveys.

External dataset	Estimated risk ratio	95% Confidence interval
	Naive estimator	
None	1.29	1.28–1.30
	Adjusted for selection bias	
CPS	13.6	12.8–14.3
PPCS	32.3	31.3–33.3
PPCS (Large Metro)	16.7	15.4–18.4

- The selection bias could be $>$ **10-fold!!**

Summary

- **Ridiculously large selection bias** in naive analyses of two topical problems.
- These examples bring discredit on our professions—statistics, epidemiology, social science, data science, ...
- Things are much better in well established research topics, but we cannot be complacent.
- The only solution (I think): **Start from the first principles**.
- Causal inference researchers are uniquely well positioned.
- Act now!!