Measuring Uncertainty with Multiple Sources of Data

Sharon Lohr

June 10, 2019

sharonlohr.com

Official Statistics

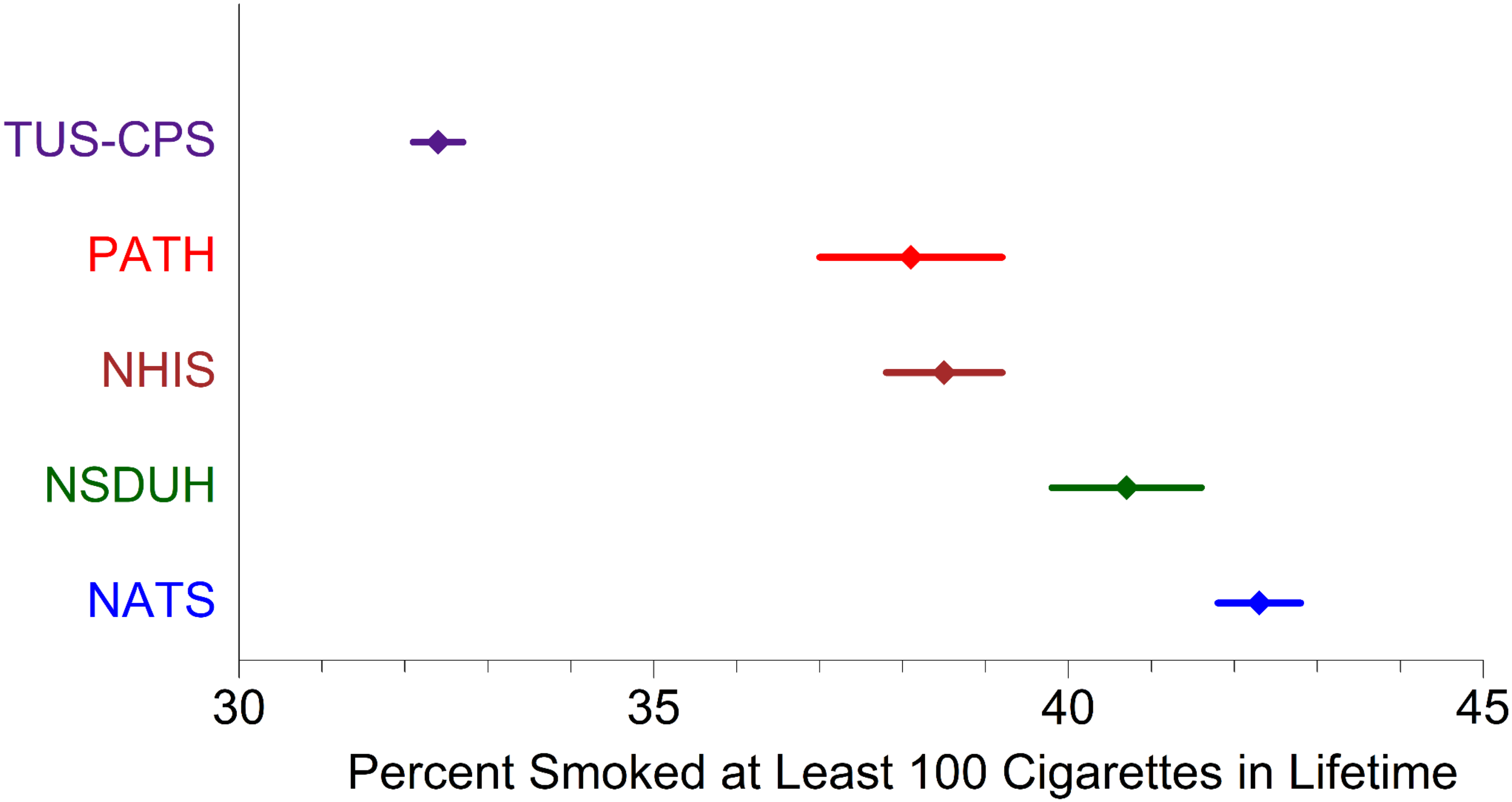
- Increased
  - Nonresponse to surveys
  - Demand for more granular data
    - Faster, more frequent
    - More geographic detail
  - Demand for more privacy
  - Intolerance for errors
- Decreased funding, personnel, ...

Use Multiple Sources

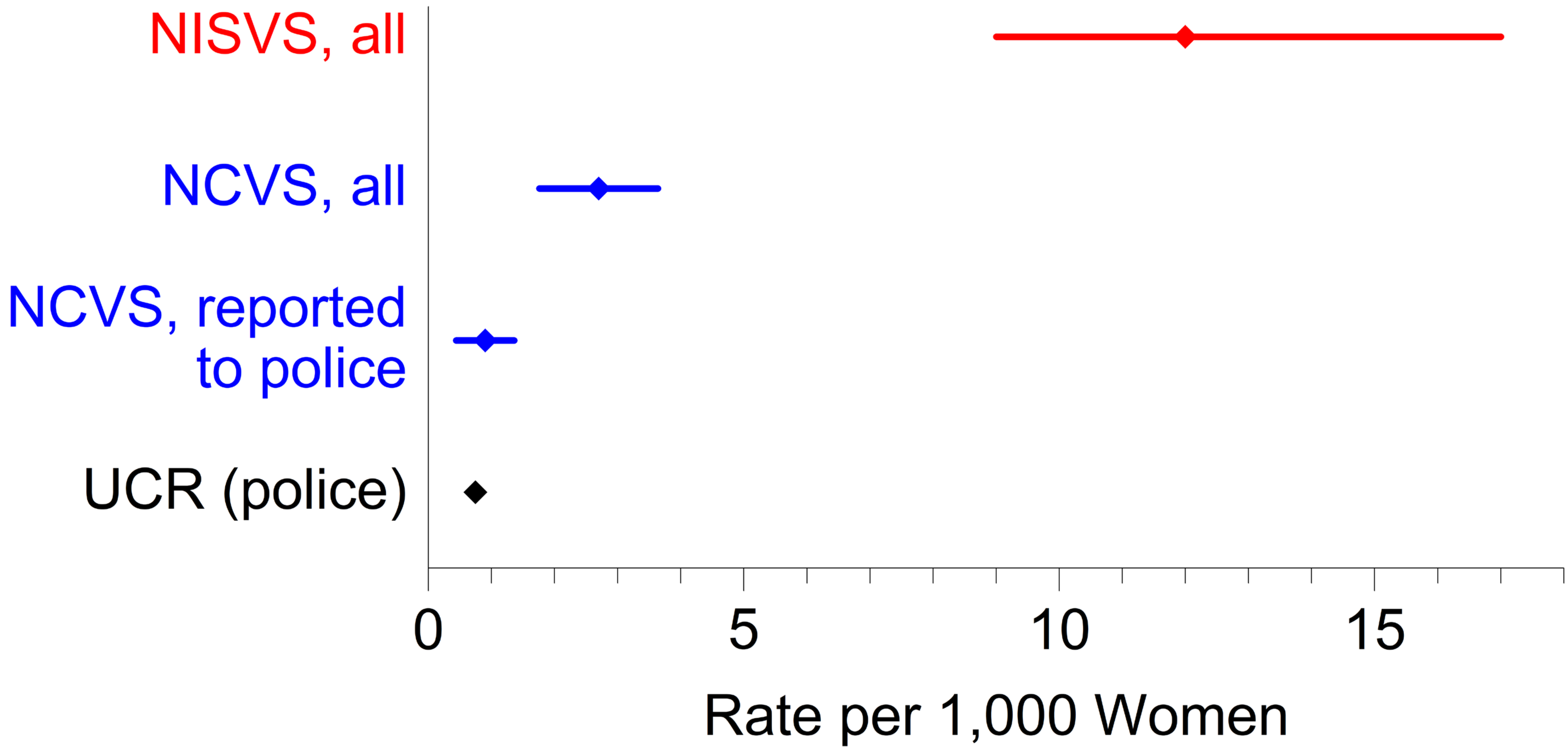
- Surveys
- Administrative Data (e.g. tax records)
- Sensor Data
- Social media, internet searches?

- How to combine?
- How to estimate uncertainty?

US Adult (age 18+) Smoking, 2014-15
Siegfried et al. (2017)



US rape/sexual assault rate, 2015



Lohr (2019) *Measuring Crime: Behind the Statistics*, CRC Press

Uniform Crime Reports (UCR)

- From police agencies
- Intended to be census
- No measures of uncertainty
- Errors from measurement, missing data are little studied
- Imputation method: from 1958

Uncertainty about Statistics from Combined Data

- Sampling error from sources
- Nonsampling error from sources
- Differences across sources
- Method used to combine

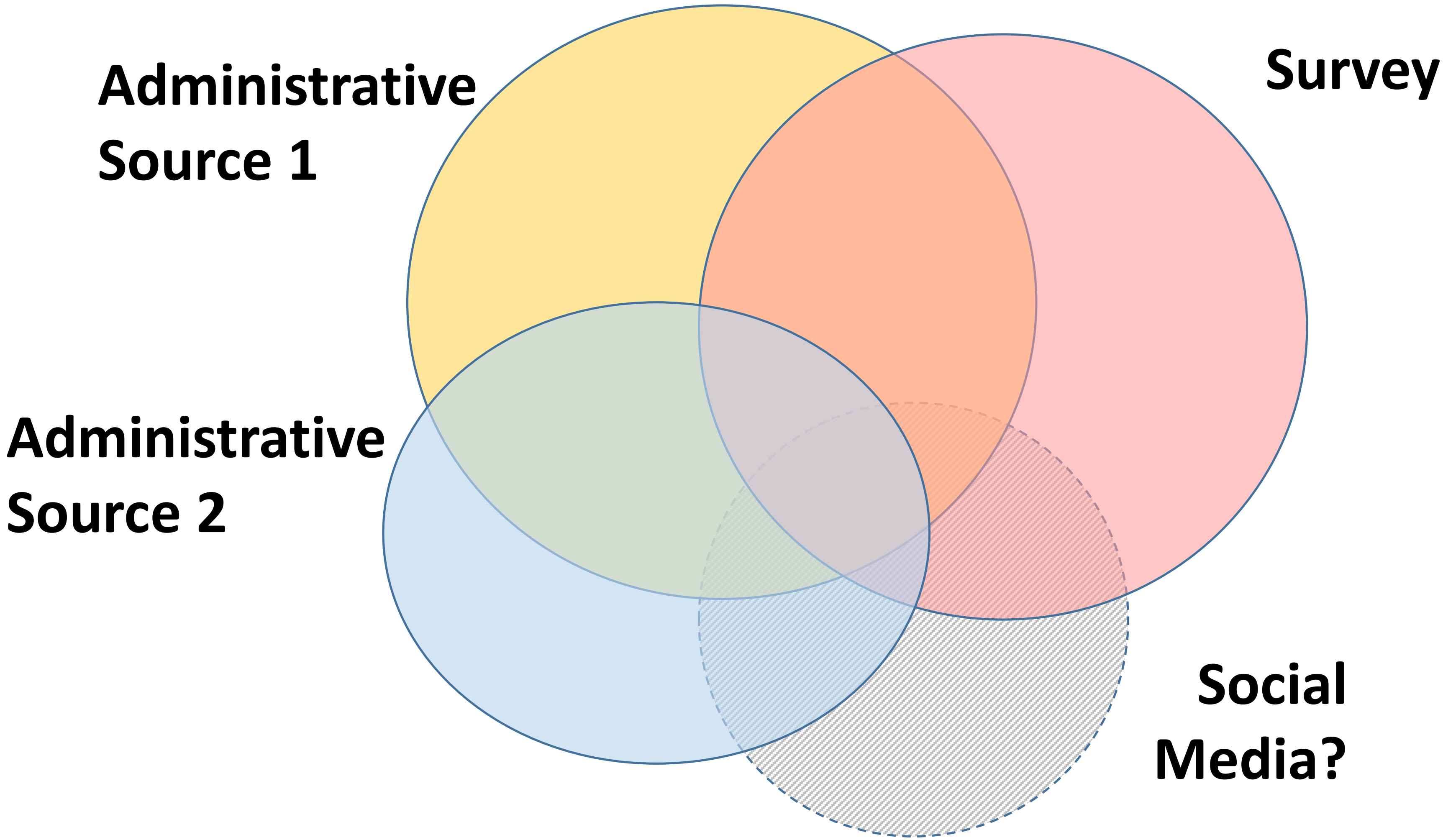
How to Combine?

- Lohr and Raghunathan (2017), *Statistical Science*
- *Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps* (National Academies of Sciences, 2017)

Methods

- Record Linkage
- Small Area Estimation
- Imputation
- Multiple Frame Methods
- Hierarchical Models
- Calibration

Multiple Frame Methods



Multiple Frame Methods

- Estimated total = sum of domains
- Traditional MF: $V(\hat{Y})$ is function of Cov(estimated domain totals) from each source
- Assumes
  - Each source has unbiased estimates
  - Domain classifications accurate
- Lohr (2011), Lin (2013)

Hierarchical Models

- Related to meta-analysis
- Manzi et al (2011) model for mean u_{dj} in domain d, source j:

$$u_{dj} = \underbrace{\theta_d}_{\text{domain mean}} + \underbrace{\delta_{dj}}_{\substack{\text{random effect} \\ \sim N(\Delta_j, \tau_j^2)}} + \underbrace{e_{dj}}_{\text{sampling error}}$$

- Lots of variations

Hierarchical Models Can

- Capture between-source variability
- Explicitly model bias
  - Need to define source or combination as unbiased
- Use prior information on source reliability, bias
- Include domain-level and record-level data

Hierarchical Models

- Strong assumptions on bias, model form
  - Do we have gold standard source?
- Survey weights, nonresponse, overlap
- Sensitive to prior information, model
- Model is explicit

Calibration

- Survey Data (y)
- Administrative Data (**x**)
- Adjust survey weights so

$$\text{Estimated Total of } \mathbf{x} \text{ from Survey, } \widehat{\boldsymbol{X}}$$
$$=$$
$$\text{Total of } \mathbf{x} \text{ from Admin Data, } \boldsymbol{X}$$

Calibration Uncertainty

- Assume $\boldsymbol{X}$ from admin data is known
- Assume “true” model is known
- Case: $\boldsymbol{X}$ = subpopulation counts

$$\hat{Y}_{ps} = \boldsymbol{X}'\hat{\bar{\boldsymbol{Y}}}, \qquad \hat{\bar{\boldsymbol{Y}}} = \left(\frac{\hat{Y}_1}{\hat{X}_1}, \dots, \frac{\hat{Y}_G}{\hat{X}_G}\right)$$

$$V\left(\hat{Y}_{ps}\right) \approx \boldsymbol{X}'V\left(\hat{\bar{\boldsymbol{Y}}}\right)\boldsymbol{X}$$

Dever & Valliant (2010, 2016)

- $\boldsymbol{X}$ measured with error

$$\hat{Y}_p = \widehat{\boldsymbol{X}}_{aux}{}'\widehat{\overline{\boldsymbol{Y}}}$$

$$V\left(\hat{Y}_p\right) \approx \boldsymbol{X}'V\left(\widehat{\overline{\boldsymbol{Y}}}\right)\boldsymbol{X} + \overline{\boldsymbol{Y}}'V\left(\widehat{\boldsymbol{X}}_{aux}\right)\overline{\boldsymbol{Y}}$$

Primary Poll Postmortems

POLITICS

The Atlantic

What the Polls Keep Missing in the Midterm Elections

There are multiple reasons why surveys have had a hard time capturing the success of this year's crop of insurgent Democrats.

The New York Times

Primary Season Was Full of Surprises. Here's Why the Polls Missed Some of Them.

Vox

Polling got Andrew Gillum's victory in Florida very wrong. 8 experts on how that happened.

W. Edwards Deming

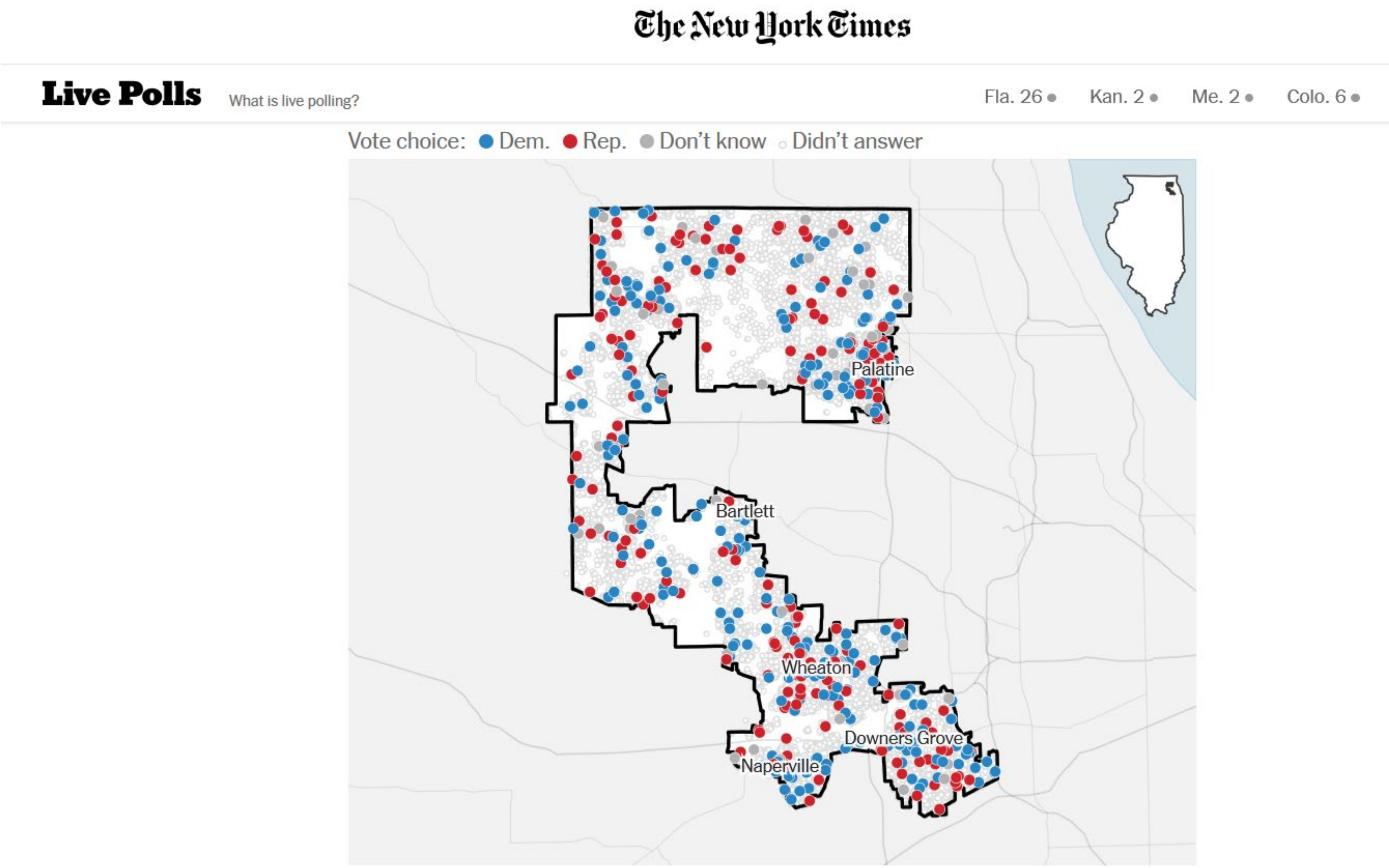
- Special Causes: factors that affect one survey
- Common Causes: systems features that affect **all** surveys

www.deming.org

Systems problems need systems solutions

New York Times Live Polls

- Illinois 6th Congressional District
- September 4-6, 2018
- Sampling Frame: Voter File
- 36,455 calls to likely voters
- 512 respondents
- 1.4% response rate



Source: https://www.nytimes.com/interactive/2018/upshot/elections-poll-il06-1.html

Poll Result

- Roskam (Republican, Incumbent)

  45% $\pm$ 4.7 %
- Casten (Democrat)

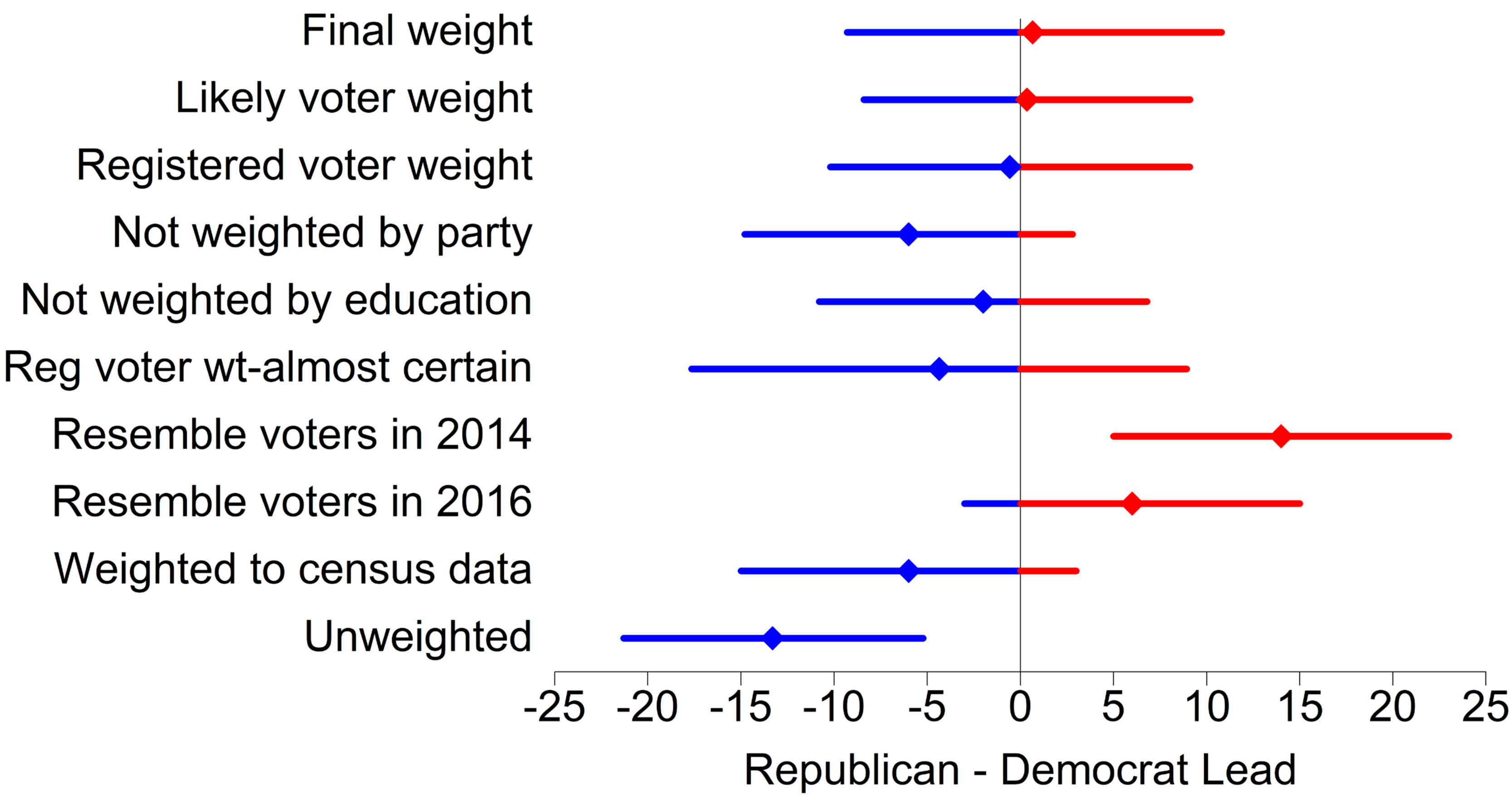
  44% $\pm$ 4.7 %
- Undecided 11%
- Republican Lead: 1% $\pm$ 9%,

But

- 1.4% Response Rate!
- 2 months before election!

- Strong assumptions for
  - Weighting
  - Who votes
  - What undecideds will do

Illinois CD 6 Race



Bayesian Model Averaging

- Hoeting et al. (1999); Lohr & Brick (2017)
- Models $M_1, \ldots, M_K$

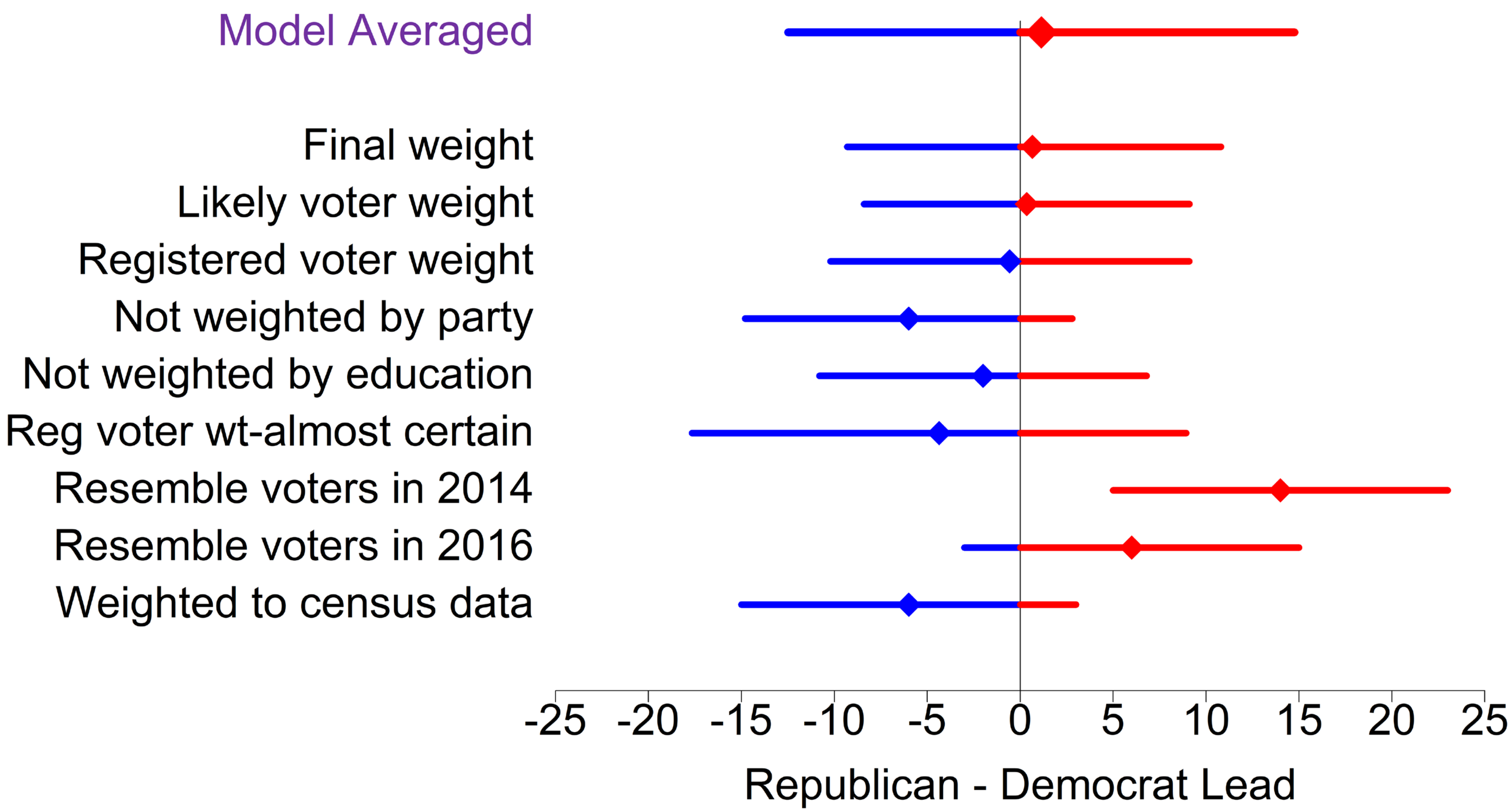
$$pr(Y \mid D) = \sum_{k=1}^{K} pr(Y \mid M_k, D)\, pr(M_k \mid D)$$

$pr(M_k \mid D) =$ posterior for model M_k

Inference

- Posterior mean
  - Weighted average of estimates
  - Weighted by $pr(M_k| \, D)$
- Posterior variance includes
  - Sampling variability
  - Model uncertainty

Illinois CD 6 Race



Model weights?

- Posterior model probabilities
- From past data
- “Past performance does not guarantee future results”

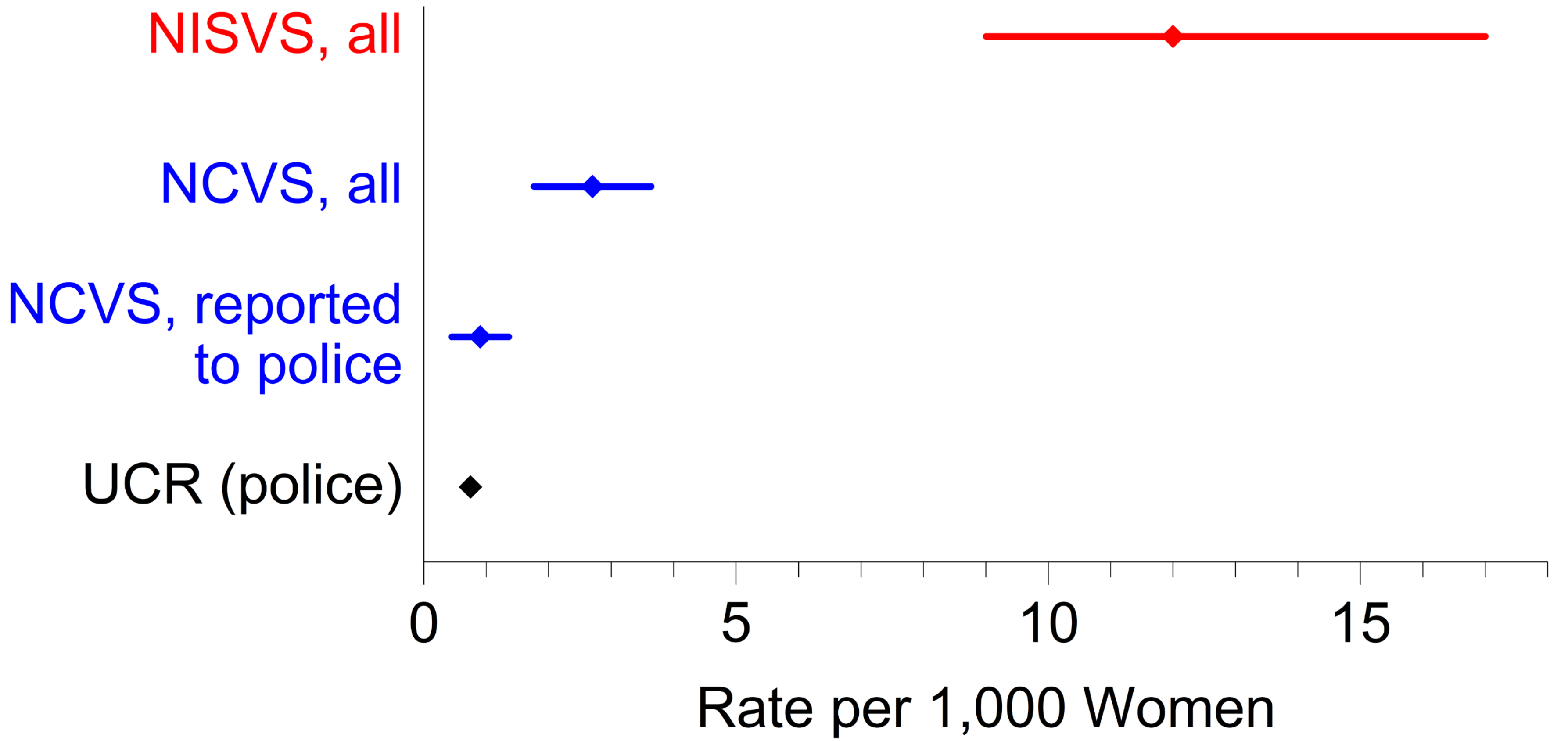
But it’s (gasp) Bayesian!

- I prefer design-based inference
  - Avoid model assumptions
  - No subjective priors
  - Elegant mathematical theory
- With nonresponse, **all** survey inference is Bayesian
  - Certainty prior on one model

Objections

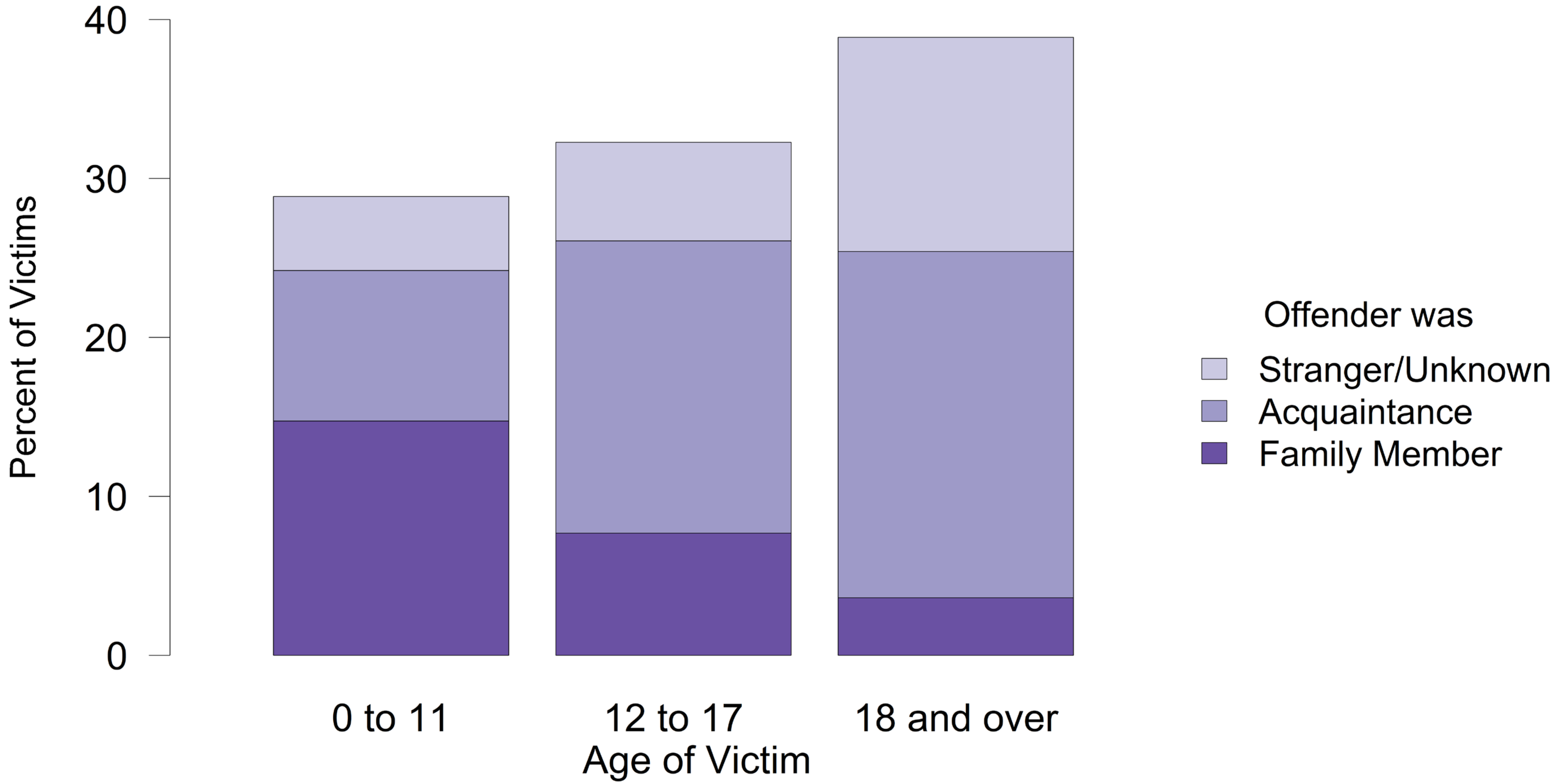
- Subjective
- Easy to cheat
  - Cherry-pick models
  - Incentives for survey-takers to have small measures of uncertainty
- Register priors before data collected?
- Make assumptions explicit

US rape/sexual assault rate, 2015

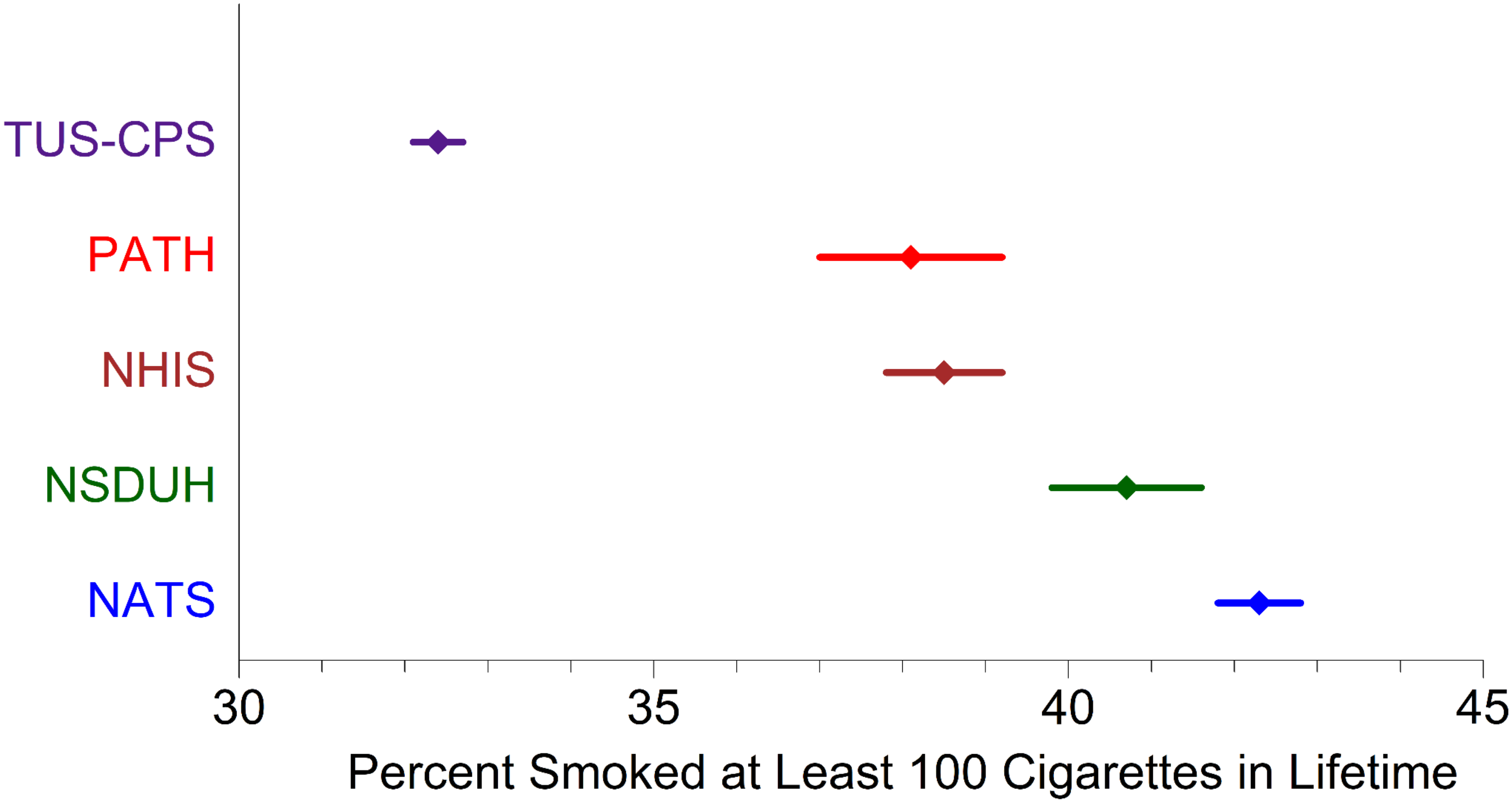


Lohr (2019) *Measuring Crime: Behind the Statistics*, CRC Press

National Incident-Based Reporting System, 2015 (1/3 of agencies)



Adult (age 18+) Smoking
Siegfried et al. (2017)



Zeroth Problem

- Colin Mallows, 1997 Fisher Lecture
- *American Statistician*, Feb 1998
- Consider relevance of data sources to the problem
- "Statistical arguments often fail because the basis for their assumptions is not spelled out."

Multiple sources

- Statistics from merged data
- Explore error properties
- Present alternative views
- Diversity is a strength

Inferences for combined data

- All use models for relationships among sources
- Depend on uncertainty measures for individual sources
  - Often underestimates
  - Inherited by combined estimate

Summary

- Use multiple sources to study quality
- Standard errors:
  - Systems-level problem
  - Include measurement, nonresponse
  - Variability from weighting models
- Industry standards
- Transparency