PSC 400

SYRACUSE UNIVERSITY

DATA ANALYTICS FOR POLITICAL SCIENCE

QUANTIFYING UNCERTAINTY

ASSIGNMENTS

- Problem Set 3 posted
  - Q3: “which model fits the data better?” = R^2
- Review exercise posted

SAMPLE VS POPULATION

- What we are interested in: population parameter
  - Approval of J. Biden in American population
- What we can study: sample parameter
  - Approval of J. Biden in survey sample

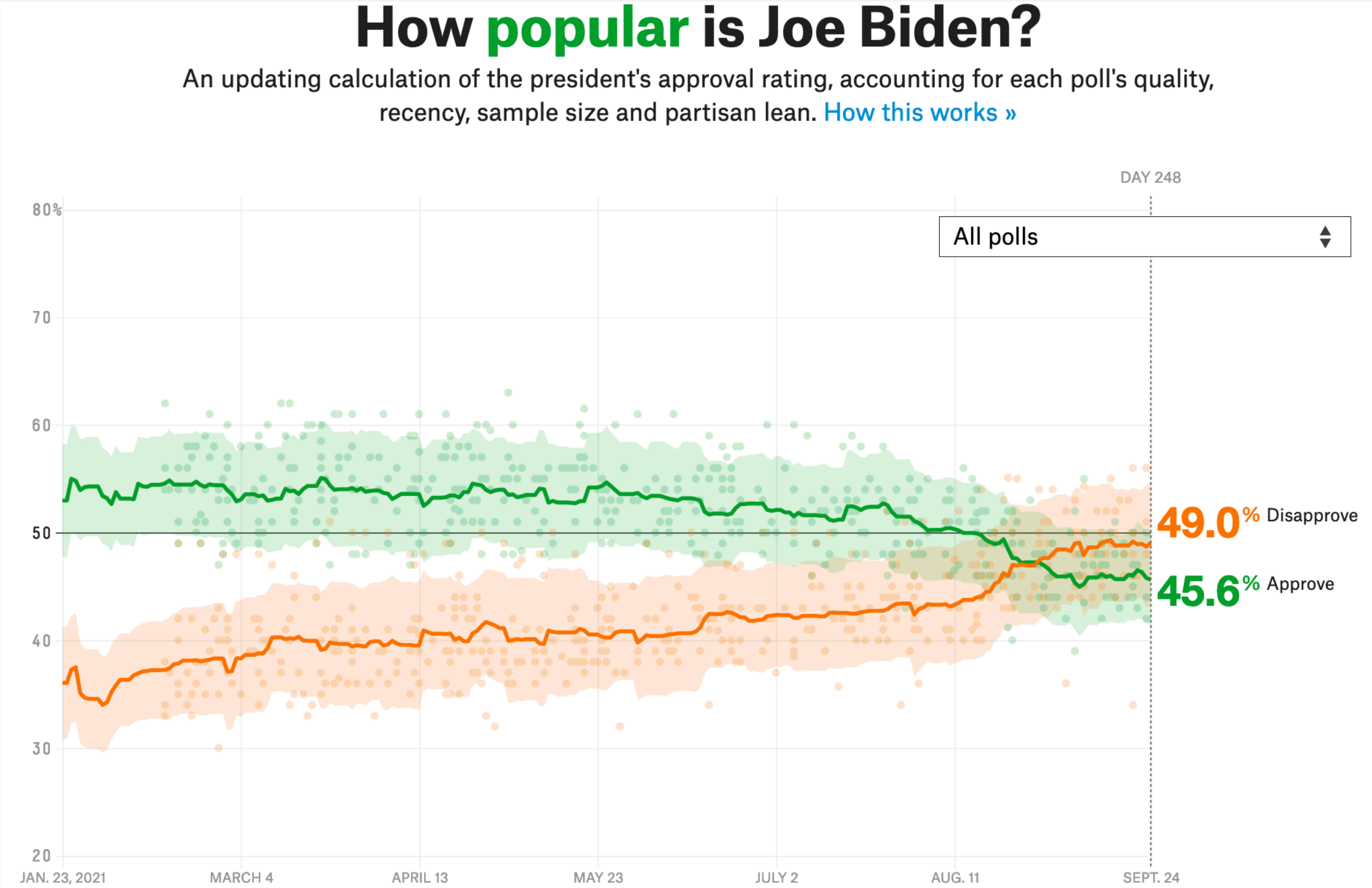
RANDOM SAMPLING

- A random sample of the population avoids *systematic* sampling error
- If we use random sampling, we can use our sample's characteristics to estimate the population's characteristics
  - e.g. can use 1000 randomly selected survey respondents to infer approval rating of J. Biden in American population

RANDOM SAMPLING ERROR

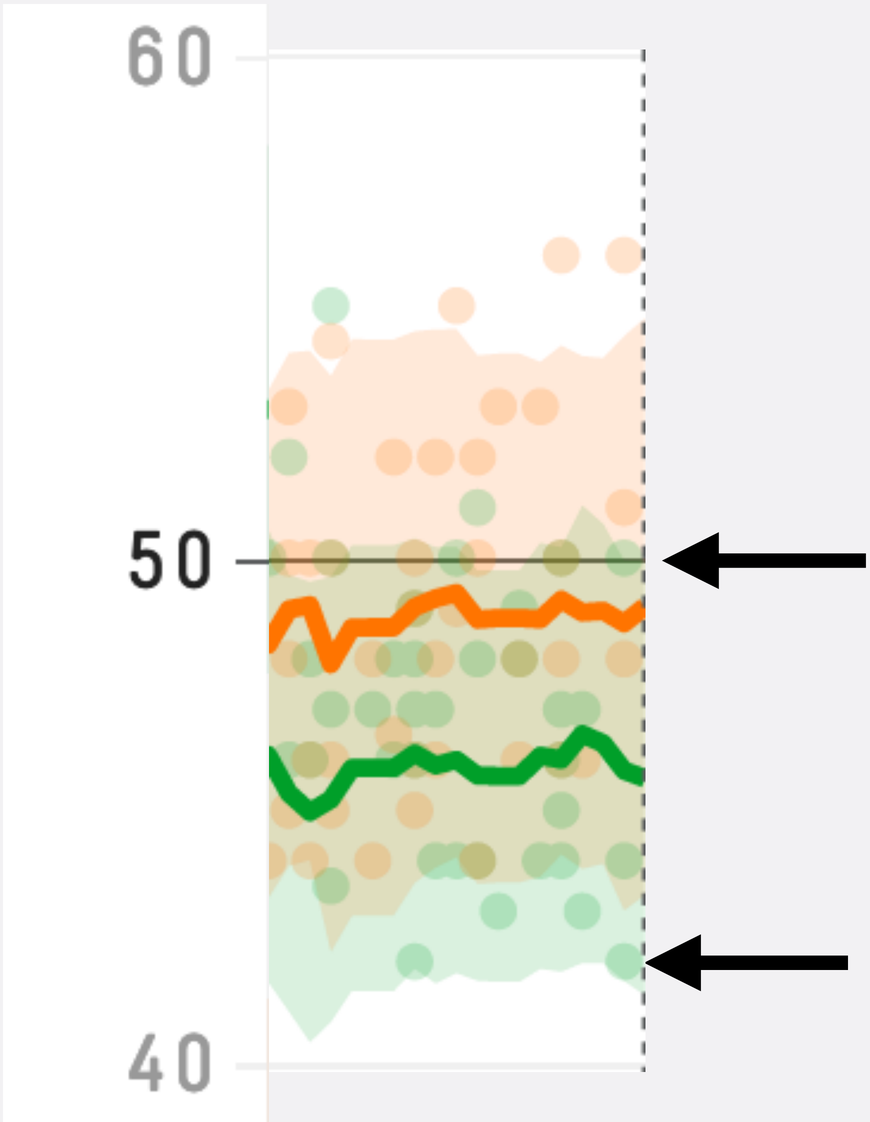
- But: random sampling introduces *random* sampling error
  - It is unlikely that our random sample looks *exactly* like the American population
  - e.g. by chance, we might draw more people that approve of Biden than is the case in the population
  - Or we might draw more people that disapprove of his performance than in the population

RANDOM SAMPLING ERROR



- https://projects.fivethirtyeight.com/biden-approval-rating

RANDOM SAMPLING ERROR



One poll: 50%

Another poll (at roughly the same time): 42%

RANDOM SAMPLING ERROR

- Random sampling introduces *random* sampling error
  - Example: Flipping a coin
  - For a fair coin, we know that Heads=50%, Tails=50%
  - We flip a coin 10 times:
    - We may get HHTHTTHTHT (5H, 5T)
    - We might also get HHHHHTHHHT (8H, 2T)
    - Or TTTHTTTTHT (2H, 8T)

THE PROBLEM

- Population parameter = Sample statistic + random sampling error

GOOD NEWS

- **We can figure out how large the random sampling error is**

CI SAMPLE MEAN

95% CONFIDENCE INTERVAL
FOR THE SAMPLE MEAN

$$\left[\ \overline{Y} - 1.96 \times \sqrt{\frac{var(Y)}{n}},\quad \overline{Y} + 1.96 \times \sqrt{\frac{var(Y)}{n}}\ \right]$$

CI DIFFERENCE IN MEANS

95% CONFIDENCE INTERVAL
FOR THE DIFFERENCE-IN-MEANS ESTIMATOR

LOWER LIMIT:

$$\overline{Y}_{\substack{\text{treatment}\\\text{group}}} - \overline{Y}_{\substack{\text{control}\\\text{group}}} - 1.96 \times \sqrt{\frac{var(Y_{\text{treatment}})}{n_{\text{treatment group}}} + \frac{var(Y_{\text{control}})}{n_{\text{control group}}}}$$

UPPER LIMIT:

$$\overline{Y}_{\substack{\text{treatment}\\\text{group}}} - \overline{Y}_{\substack{\text{control}\\\text{group}}} + 1.96 \times \sqrt{\frac{var(Y_{\text{treatment}})}{n_{\text{treatment group}}} + \frac{var(Y_{\text{control}})}{n_{\text{control group}}}}$$