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 6 posted
- Both due on Friday

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

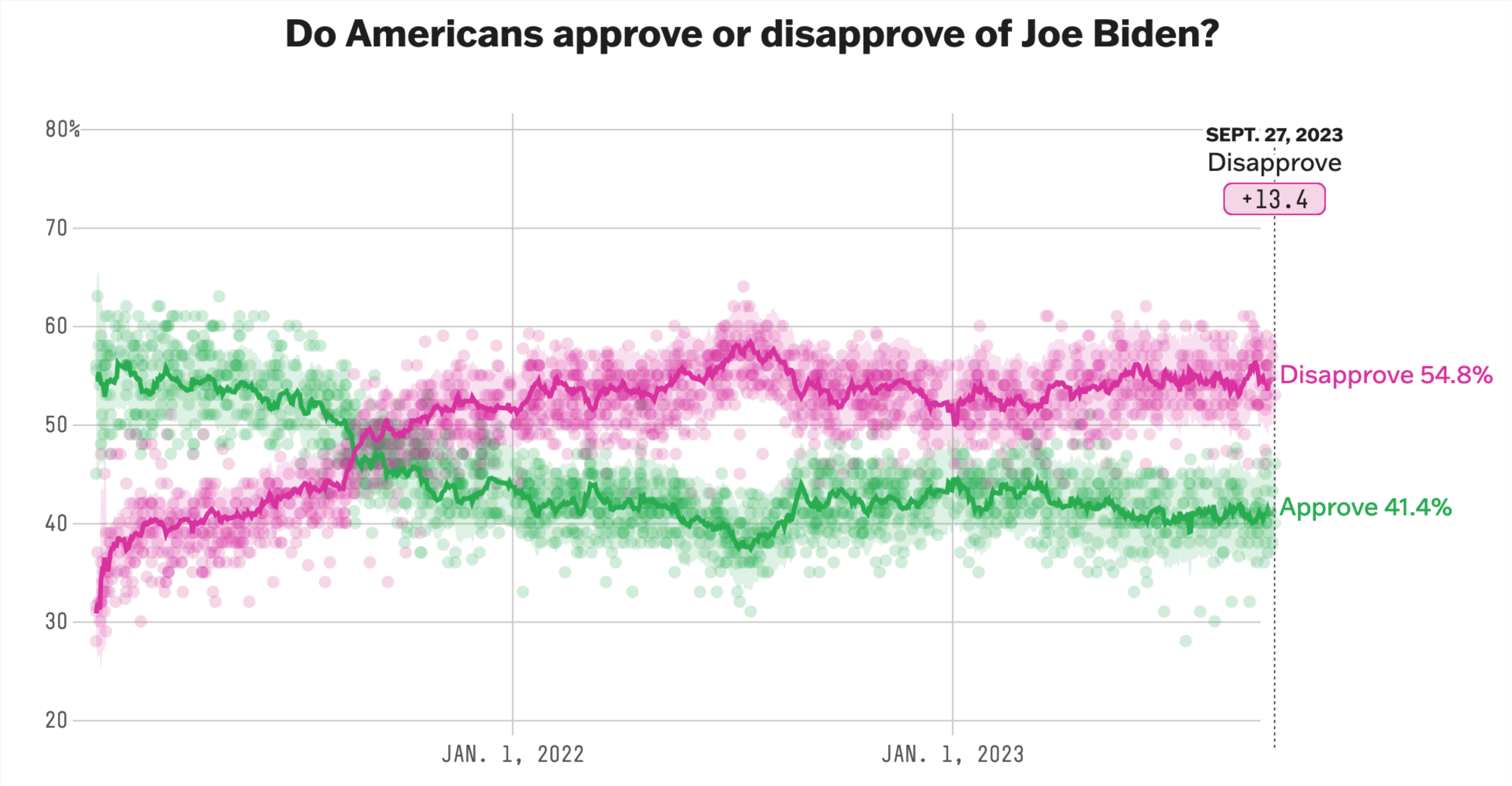
RANDOM SAMPLING

- 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

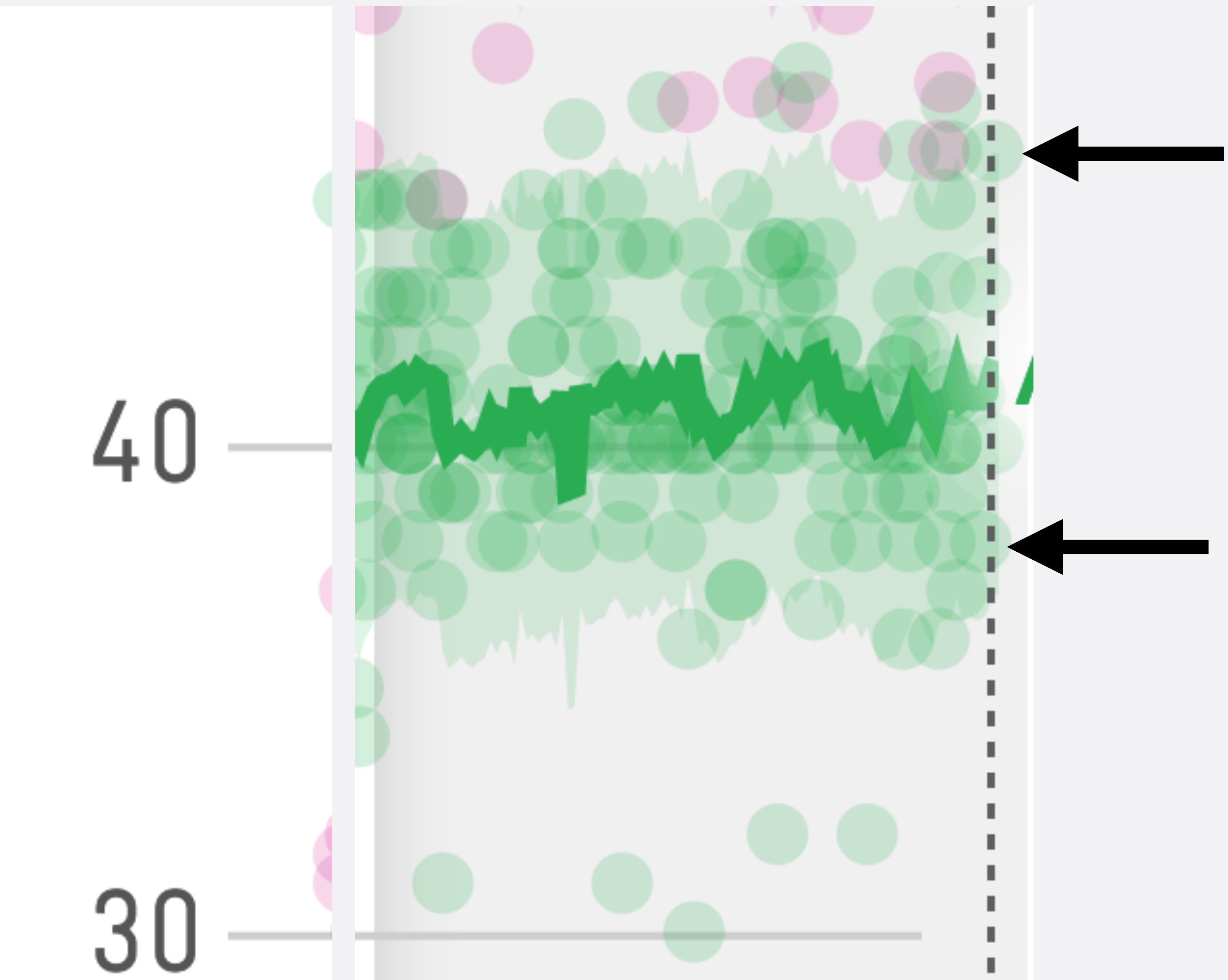
- But: random sampling introduces *random* sampling error
  - It is unlikely that our random sample looks *exactly* like the American population
  - e.g. by chance, we might draw more people that approve of Biden than is the case in the population
  - Or we might draw more people that disapprove of his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

RANDOM SAMPLING ERROR



- https://projects.fivethirtyeight.com/polls/approval/joe-biden/

RANDOM SAMPLING ERROR



One poll: 46%

Another poll (at roughly the same time): 38%

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

95% CONFIDENCE INTERVAL

$$95\%\ \text{CI} = [\ \textit{estimator} - 1.96 \times \text{standard error},\ \textit{estimator} + 1.96 \times \text{standard error}\ ]$$

where:

- *estimator* is a random variable across multiple hypothetical samples
- standard error is the estimated standard deviation of the estimator across multiple hypothetical samples.

CI SAMPLE MEAN

95% CONFIDENCE INTERVAL
FOR THE SAMPLE MEAN

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

where:

- $\overline{Y}$ is the sample mean of Y
- $\sqrt{var(Y)/n}$ is the standard error of the sample mean
- $var(Y)$ is the sample variance of Y
- n is the number of observations in the sample.

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:

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

where:

- $\overline{Y}_{\substack{\text{treatment}\\\text{group}}} - \overline{Y}_{\substack{\text{control}\\\text{group}}}$ is the difference-in-means estimator
- $\sqrt{var(Y_{\text{treatment}})/n_{\text{treatment group}} + var(Y_{\text{control}})/n_{\text{control group}}}$ is the standard error of the difference-in-means estimator
- $var(Y_{\text{treatment}})$ and $var(Y_{\text{control}})$ are the sample variances of Y under the treatment and control conditions
- $n_{\text{treatment group}}$ and $n_{\text{control group}}$ are the number of observations in the treatment and the control groups in the sample.

EXERCISE

- UA_survey.csv
- Compute difference-in-means for pro-Russian vote between those with and without access to Russian TV
- Compute the 95% confidence interval of that difference

POPULATION VS. SAMPLE, AGAIN

- Want to know: does Russian TV have effect on pro-Russian votes in the *population*?
- We only have data from a *random sample*
- Idea: Use relation between two variables in *sample* to make inference about relation between two variables in *population*
    - Of course, means we can make mistakes

NULL HYPOTHESIS

- In the population, there is *no relationship* between dependent and independent variable
  - H_0

ALTERNATIVE HYPOTHESIS

- There *is* a relationship between the independent and dependent variable in the population
  - H_a or H_1

ERRORS

	There Is A Relation In The Population	There Is No Relation In The Population
We Conclude There Is A Relation	✔	✘ Type I
We Conclude There Is No Relation	✘ Type II	✔

TYPE I ERROR

- We conclude there is a relationship between X and Y when in reality there is not
  - "Type I error"
  - We falsely reject H_0

TYPE II ERROR

- We conclude there is no relationship between X and Y when in reality there is
  - "Type II error"
  - We falsely do not reject H_0

DECISION

- It's really bad if we conclude there is a relationship when in reality there is not
- Type I error: falsely rejecting H_0
- We only want to reject H_0 based on our sample if chance of committing Type I error is relatively small
  - Typically: 5% or less