PSC 202

SYRACUSE UNIVERSITY

INTRODUCTION TO POLITICAL ANALYSIS

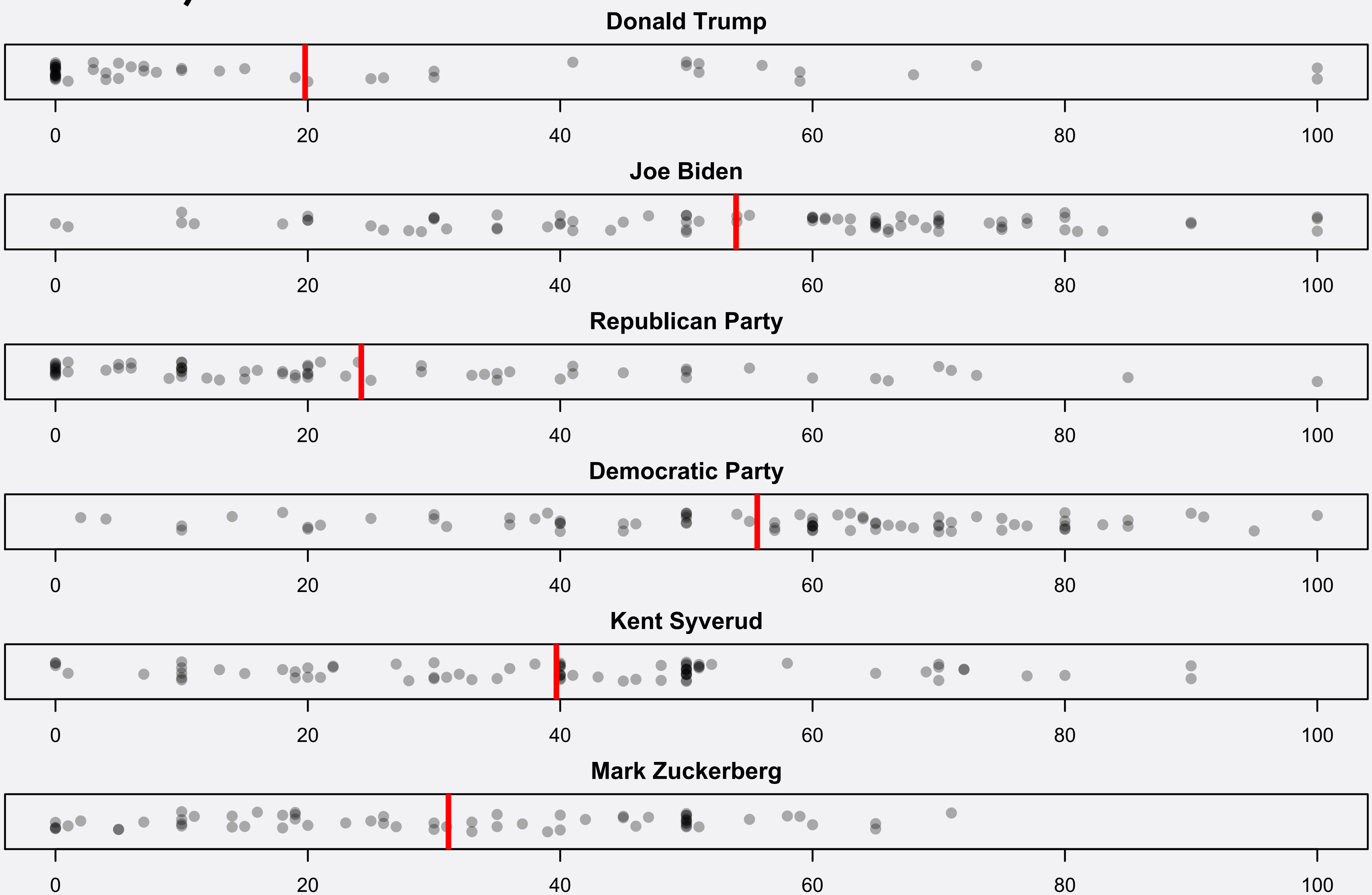
MORE SAMPLING AND SURVEYS, HYPOTHESES AND CAUSALITY

SURVEY

- Response rate: 89%
  - Needed 85% to get extra participation credit for whole class

SURVEY RESULTS

- **Feeling thermometers (0=cold/unfavorable, 100=warm/favorable)**



SURVEY RESULTS

- "The involvement of the US in Afghanistan has been beneficial"

	Number	Percentage
Strongly agree	1	1.2%
Somewhat agree	13	14.6%
Neither agree nor disagree	8	9.8%
Somewhat disagree	33	40.2%
Strongly disagree	27	32.9%

SURVEY RESULTS

- "The federal government should mandate that everyone has to be vaccinated against Covid-19 (unless they have a medical or religious exemption)."

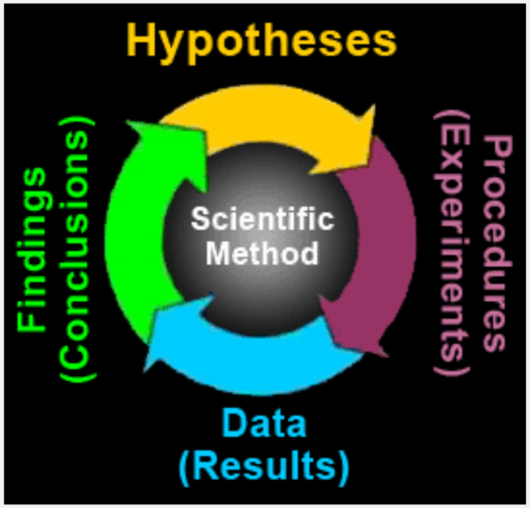
	Number	Percentage
Strongly agree	44	53.6%
Somewhat agree	21	25.6%
Neither agree nor disagree	6	7.3%
Somewhat disagree	4	4.9%
Strongly disagree	7	8.5%

TODAY AND NEXT MONDAY

- Finishing up Sampling and Surveys
- Hypotheses and causality

RESEARCH PROCESS

- Formulate research question
- Propose explanation/theory, hypotheses
- Data collection process
- Use data to evaluate hypotheses
- Reassess explanation



RECAP

POLITICS SEPTEMBER 22, 2021

Biden's Approval Rating Hits New Low of 43%; Harris' Is 49%

Results for this Gallup poll are based on telephone interviews conducted Sept. 1-17, 2021, with a random sample of 1,005 adults, aged 18 and older, living in all 50 U.S. states and the District of Columbia. For results based on the total sample of national adults, the margin of sampling error is ±4 percentage points at the 95% confidence level. All reported margins of sampling error include computed design effects for weighting.

- **How confident can we be that the 43% approval rating among 1,005 respondents is close to the approval rating of *all* American voters?**

RECAP

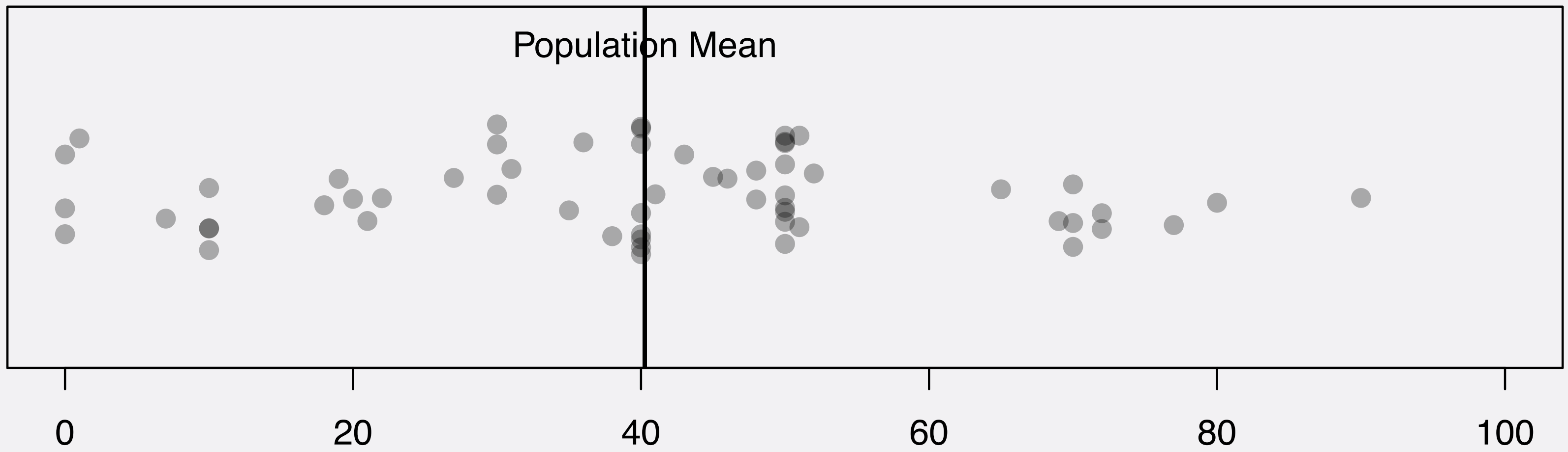
Unknown:
Approval rating in
population

Known: Approval
rating in survey

- Population parameter = Sample statistic + random sampling error

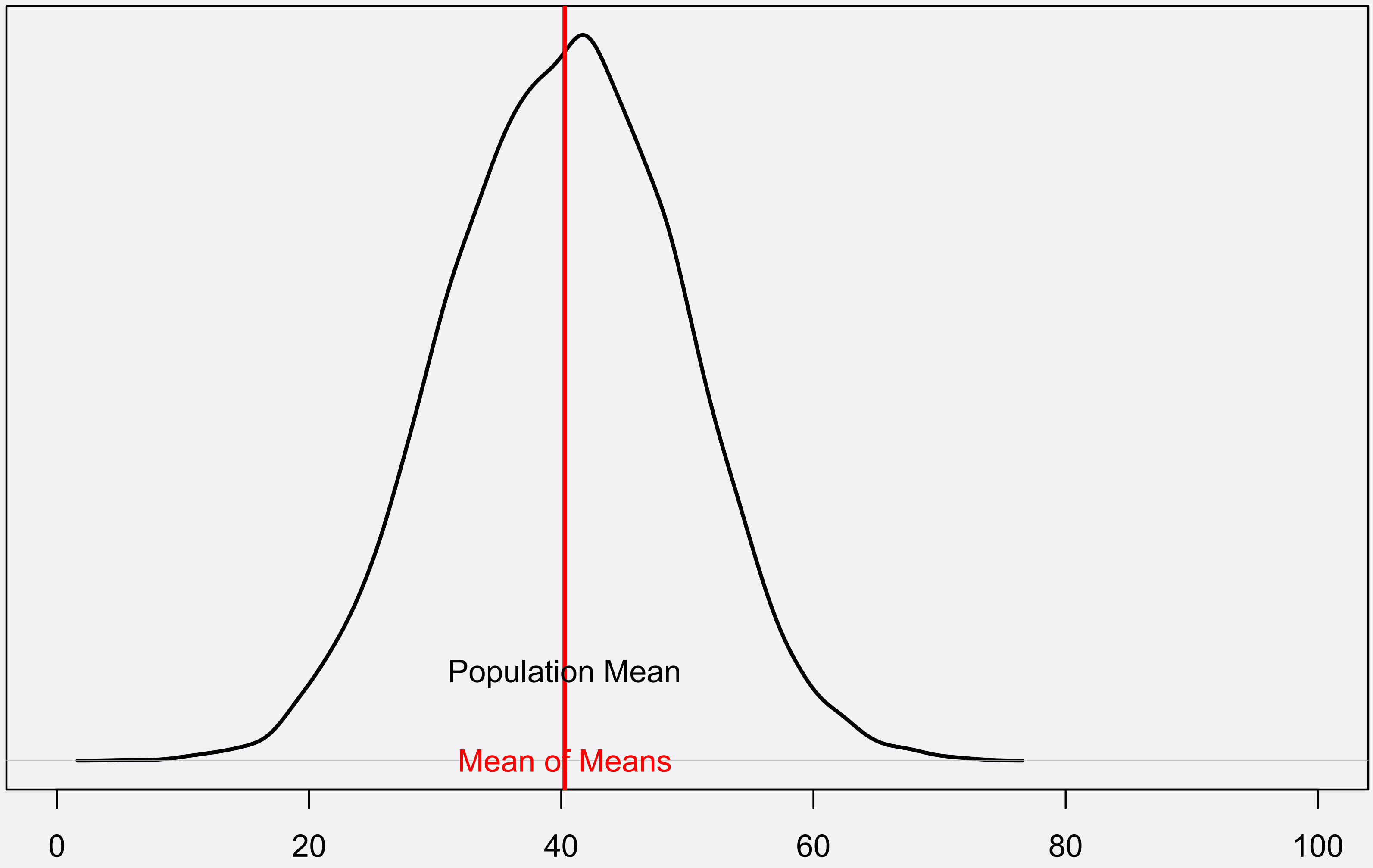
We can figure
this out

RECAP

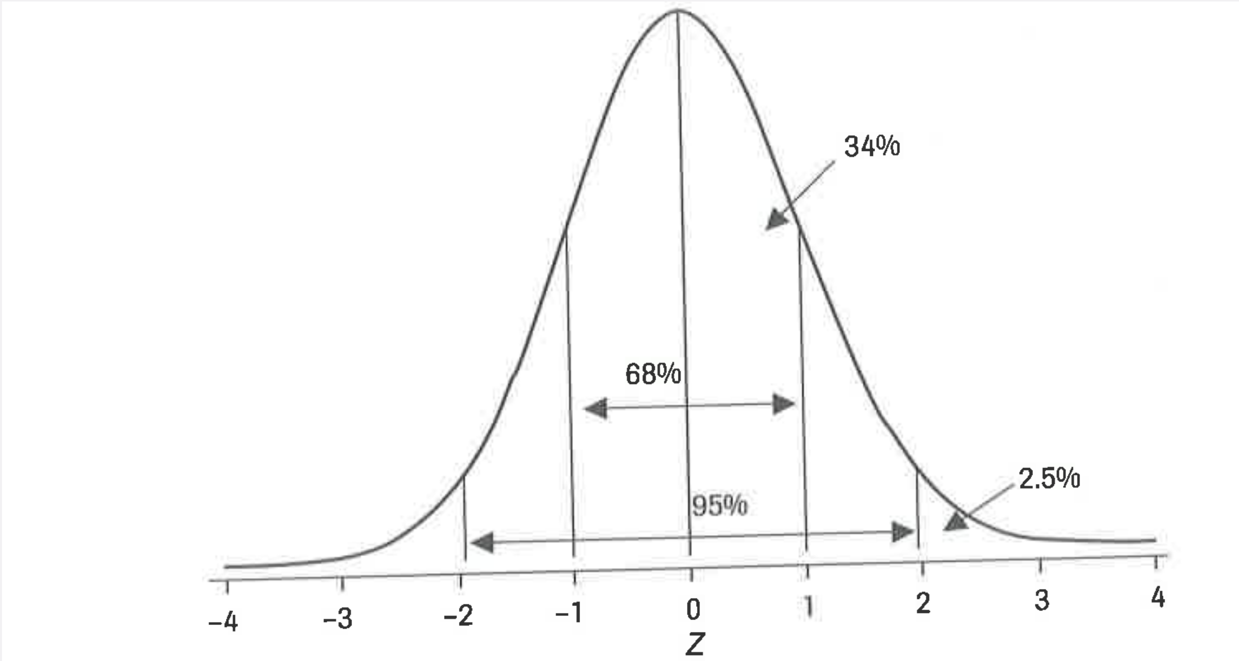


- Took random sample of 5 students
- Record average rating of those 5 students
- Do this thousands of times

AFTER 10,000 RANDOM SAMPLES

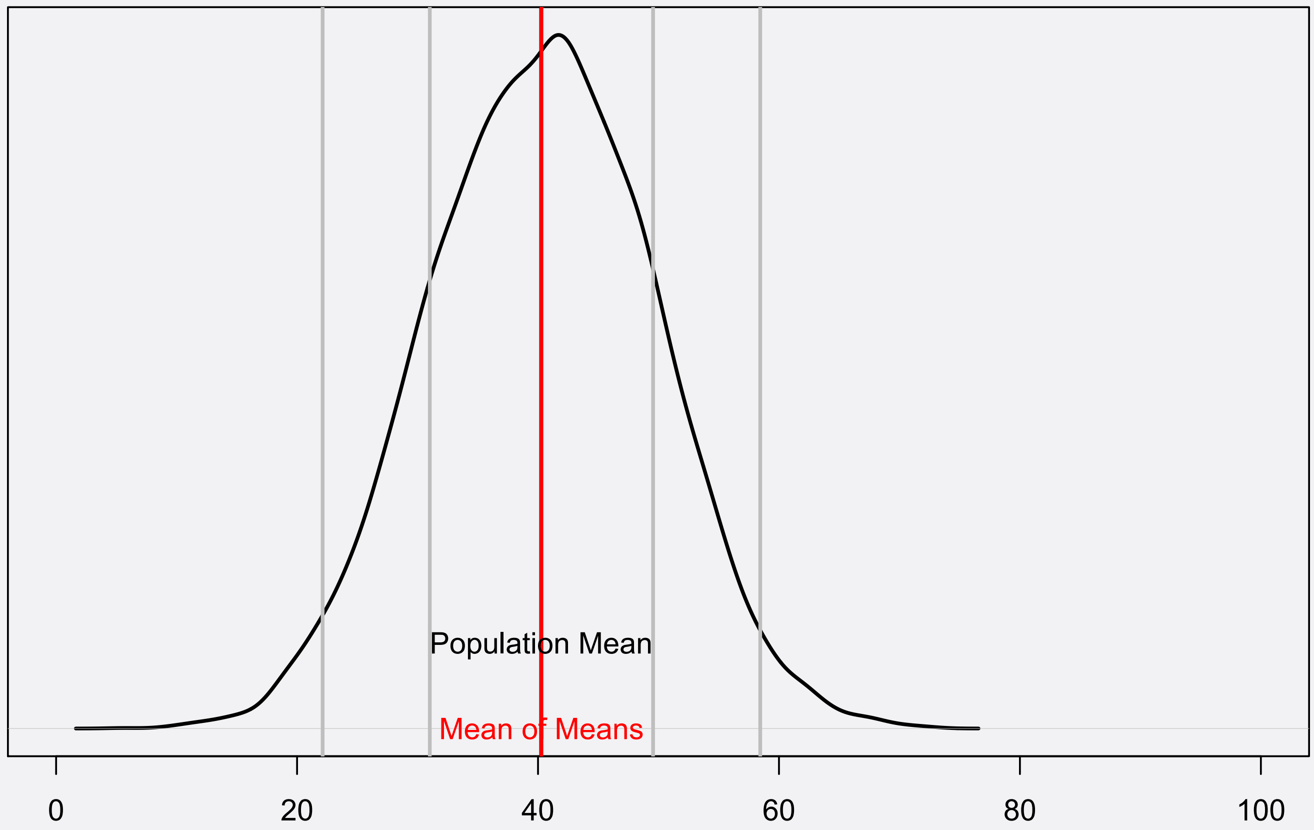


RECAP



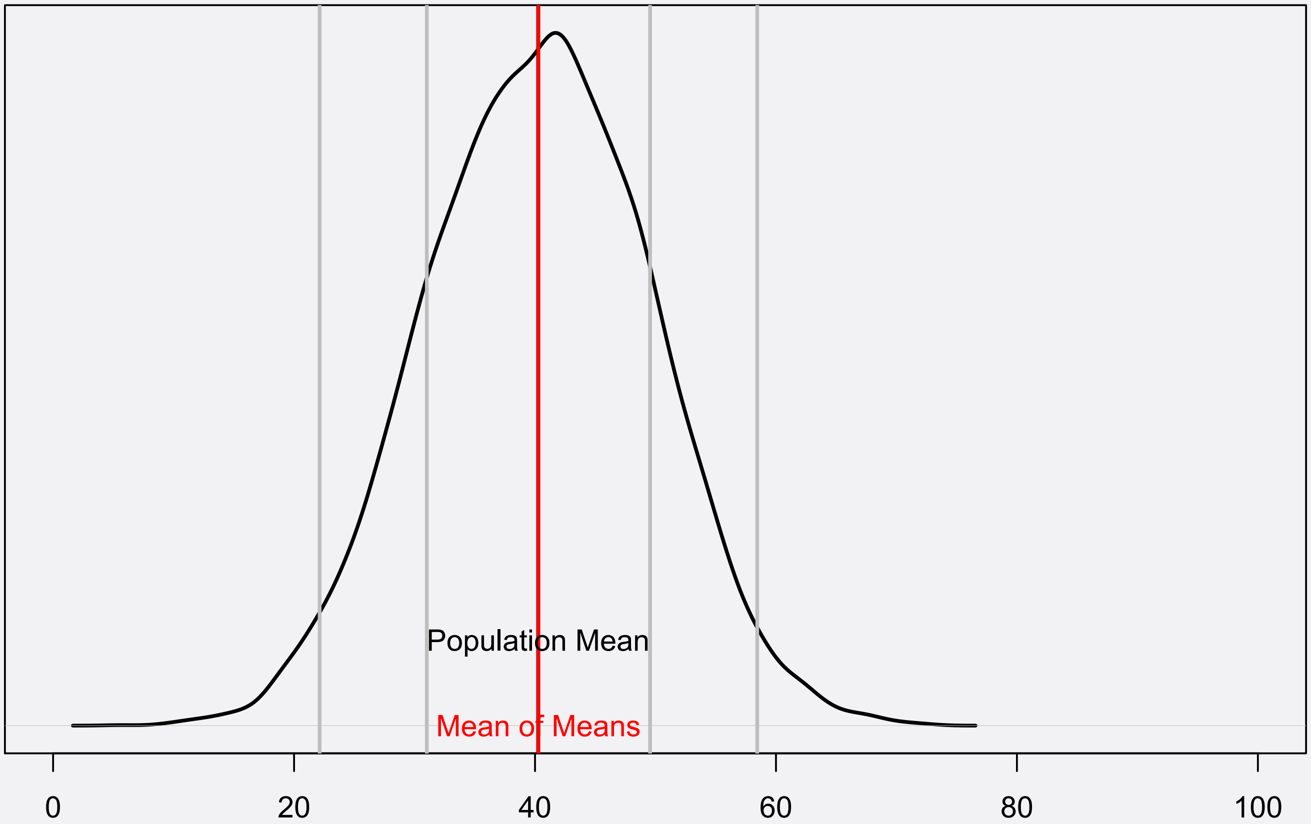
- If a variable follows a Normal distribution…
    - we know that 68% of observations are between mean ± 1SD, 95% between mean ± 1.96SD

WHAT DOES THIS HAVE TO DO WITH RANDOM SAMPLES?



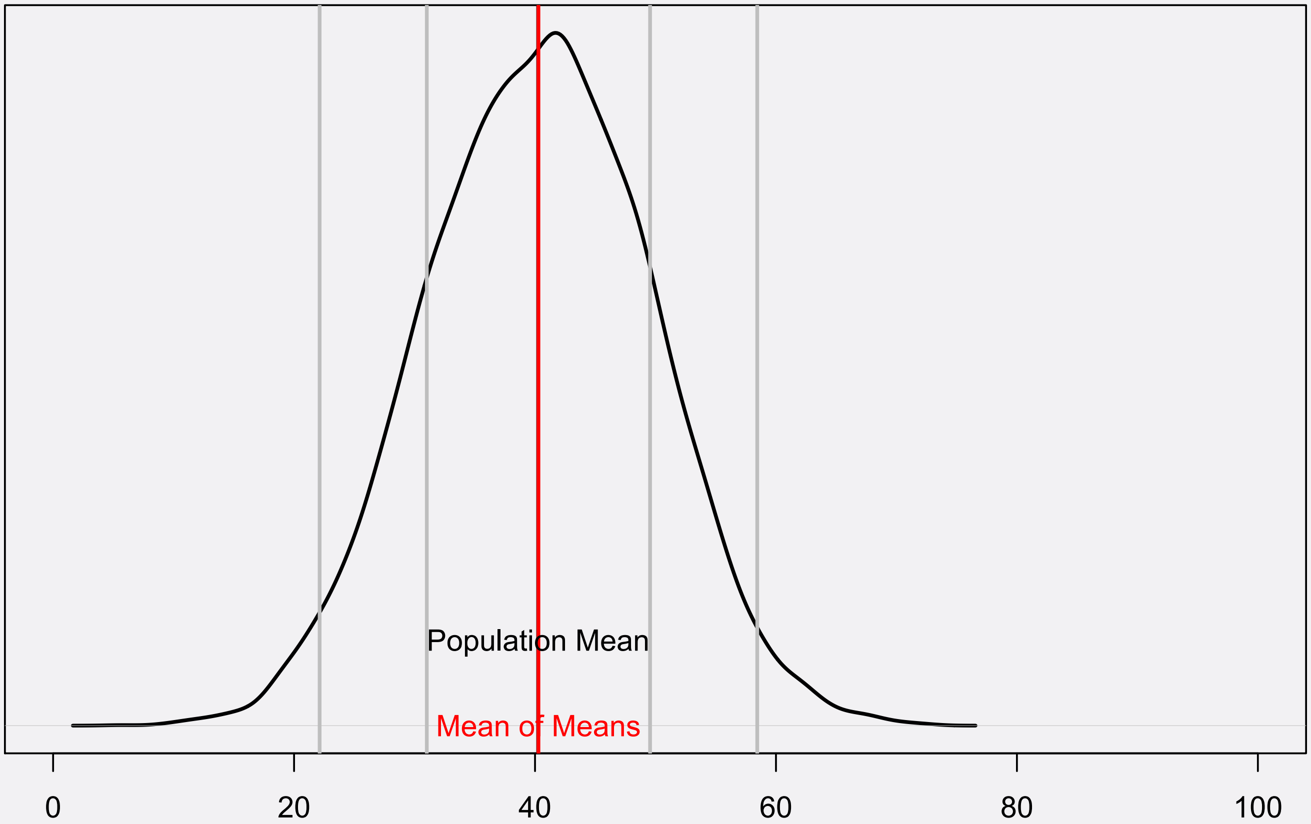
- If we take many random samples from a population
- and compute mean for each
- those means have a Normal distribution

WHAT DOES THIS HAVE TO DO WITH RANDOM SAMPLES?



- **68% of sample means will be within mean ± 1SD, 95% between mean ± 1.96SD**

WHAT DOES THIS HAVE TO DO WITH RANDOM SAMPLES?

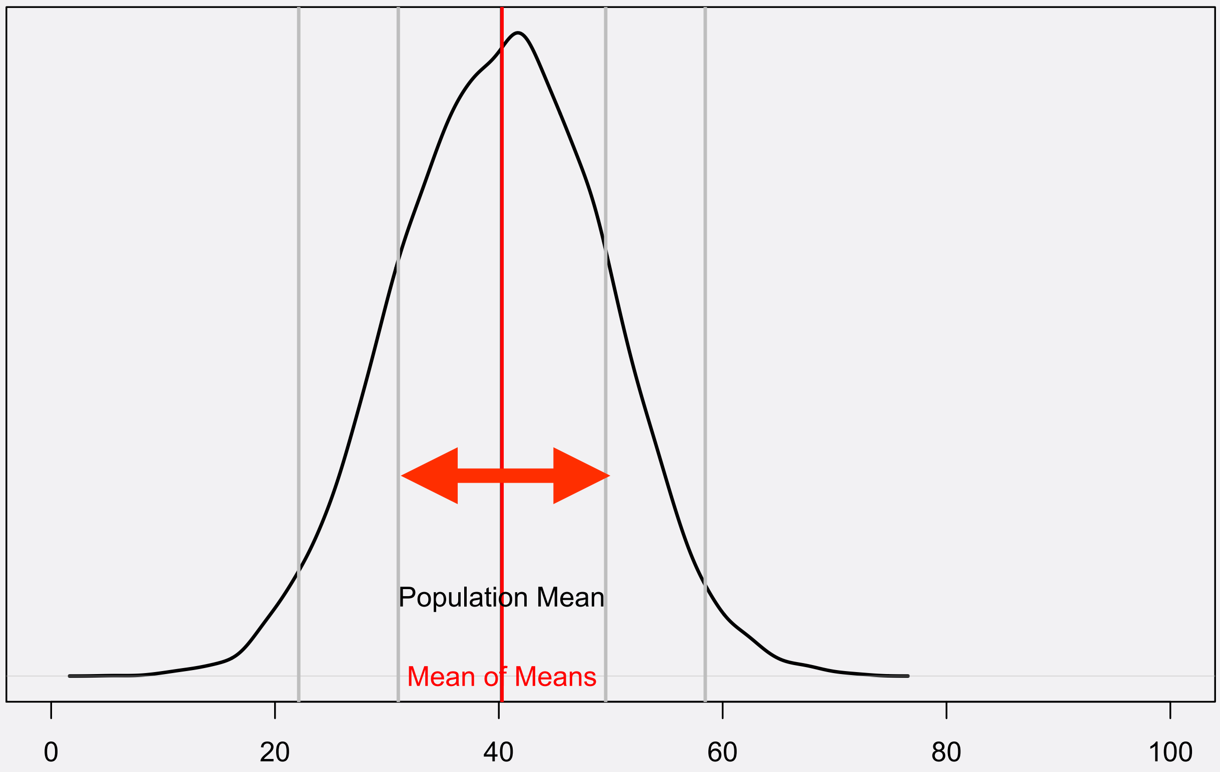


- Allows us to put a number on how large random measurement error is
- Which tells us how confident we can be that conclusions we draw from a sample hold in the population overall

HOW DOES THIS HELP?

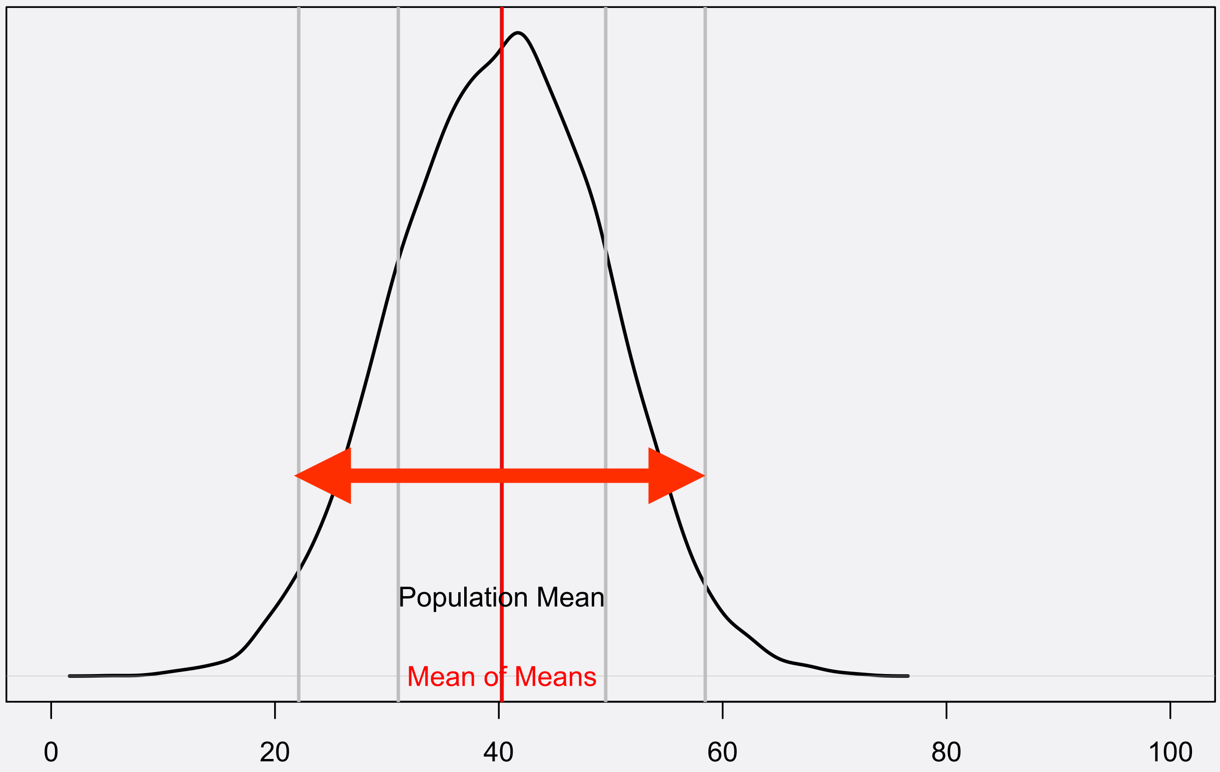
- How it helps, part 1:
- We can quantify how large the random sampling error in our sample is

RANDOM SAMPLING ERROR



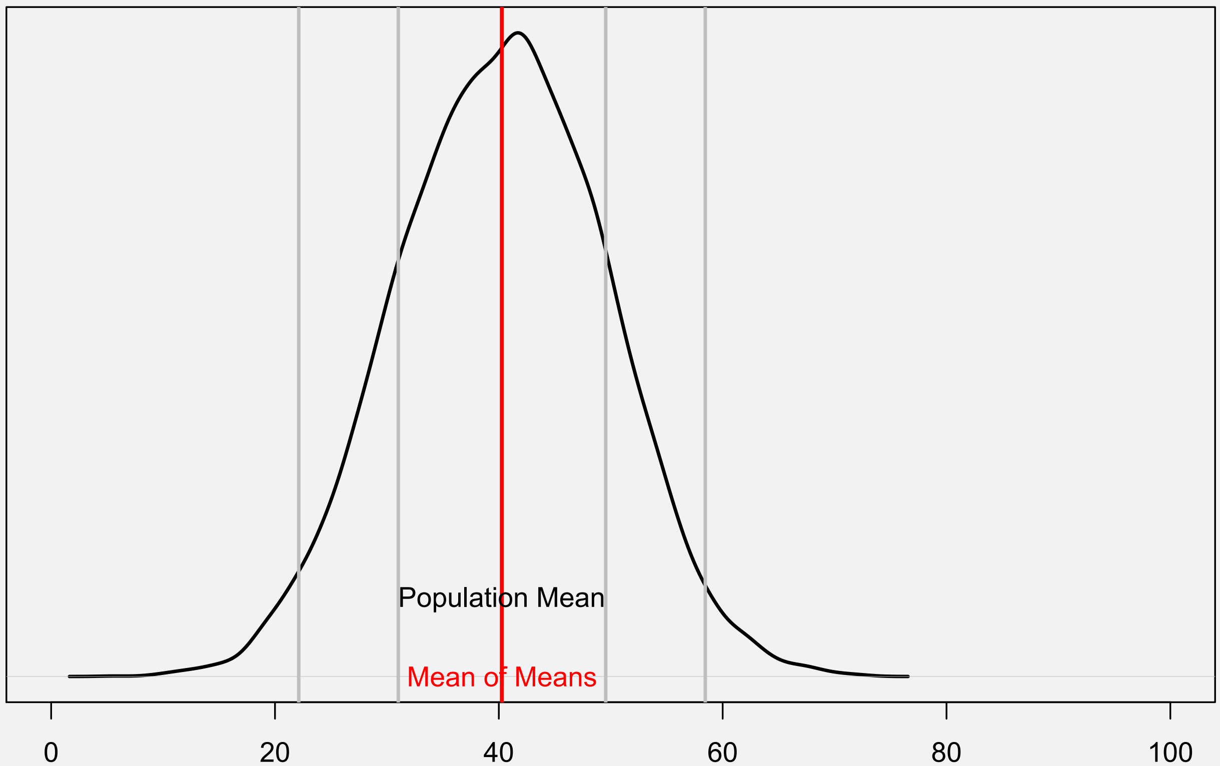
- **68% of sample means: population mean ± 1 standard deviations**

RANDOM SAMPLING ERROR



- **95% of sample means: population mean ± 1.96 standard deviations**

RANDOM SAMPLING ERROR



- **What is the standard deviation here?**
- **If we know, we can tell how large the sampling error is**

STANDARD ERROR

- The standard deviation of the distribution of sample means is called the *standard error of the sample mean*
  - or simply *standard error*
  - Standard error is a measure of the random sampling error

STANDARD ERROR

$$SE = \frac{s}{\sqrt{n}}$$

- s: standard deviation in our random sample
- n: size of our sample
- The larger SE, the larger the random sampling error

STANDARD ERROR

$$SE = \frac{s}{\sqrt{n}}$$

- SE is larger if *s* is large
- Random sampling error is larger if there is a lot of variation in population

STANDARD ERROR

$$SE = \frac{s}{\sqrt{n}}$$

- **SE is larger if *n* is small**
- **Random sampling error is larger if we sample a small number of people**

STANDARD ERROR

- **Let's say: 500 respondents (randomly selected)**
- **We got: $\bar{x}$ = 64.0 (s=14.0)**

$$\text{SE} = \frac{s}{\sqrt{n}} = \frac{14}{\sqrt{500}} = 0.626$$

NOTE

- Population size: N
- Population mean: μ
- Population standard deviation: σ

- Sample size: n
- Sample mean: $\bar{x}$
- Sample standard deviation: s

STANDARD ERROR

- **Let's say: 500 respondents (randomly selected)**
- **We got: $\bar{x}$ = 64.0 (s=14.0)**

$$\mathrm{SE} = \frac{s}{\sqrt{n}} = \frac{14}{\sqrt{500}} = 0.626$$

SAMPLE SIZE

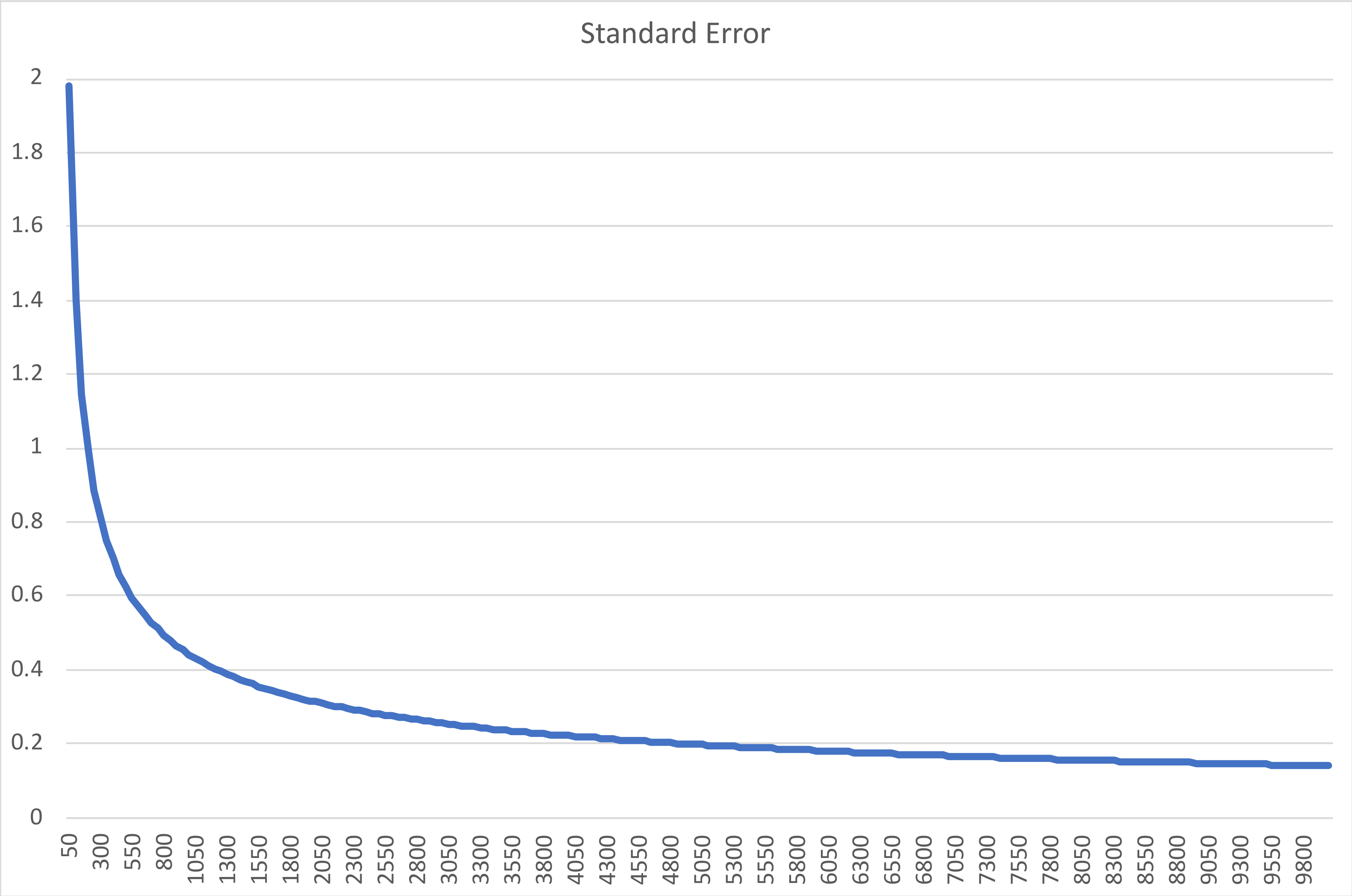
- **Let's say: 2000 respondents (randomly selected)**
- **We got: $\bar{x}$ = 64.0 (s=14.0)**

$$\text{SE} = \frac{s}{\sqrt{n}} = \frac{14}{\sqrt{2000}} = 0.313$$

SAMPLE SIZE

- The larger the sample, the smaller the standard error of the sample mean
  - Larger survey = less random measurement error

SAMPLE SIZE



- **Why surveys are often only 1000 people**

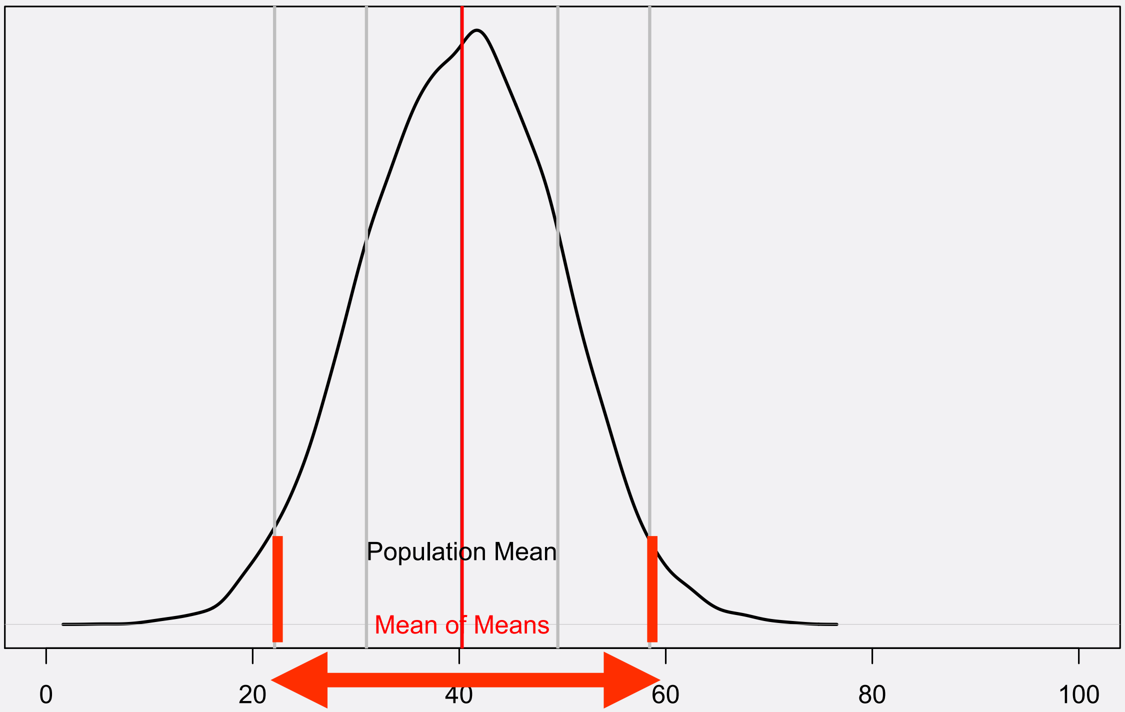
SUMMARY: STANDARD ERROR

- How large is random sampling error?
- If we draw many random samples from a population and record their means…
  - They are normally distributed with mean=population mean
  - Measure of random sampling error: one standard deviation of that distribution ("standard error")
  - The standard error is $s/\sqrt{n}$

HOW DOES THIS HELP?

- How it helps, part 2:
- For random samples, we can provide an interval that likely contains the true mean μ

CONFIDENCE INTERVAL



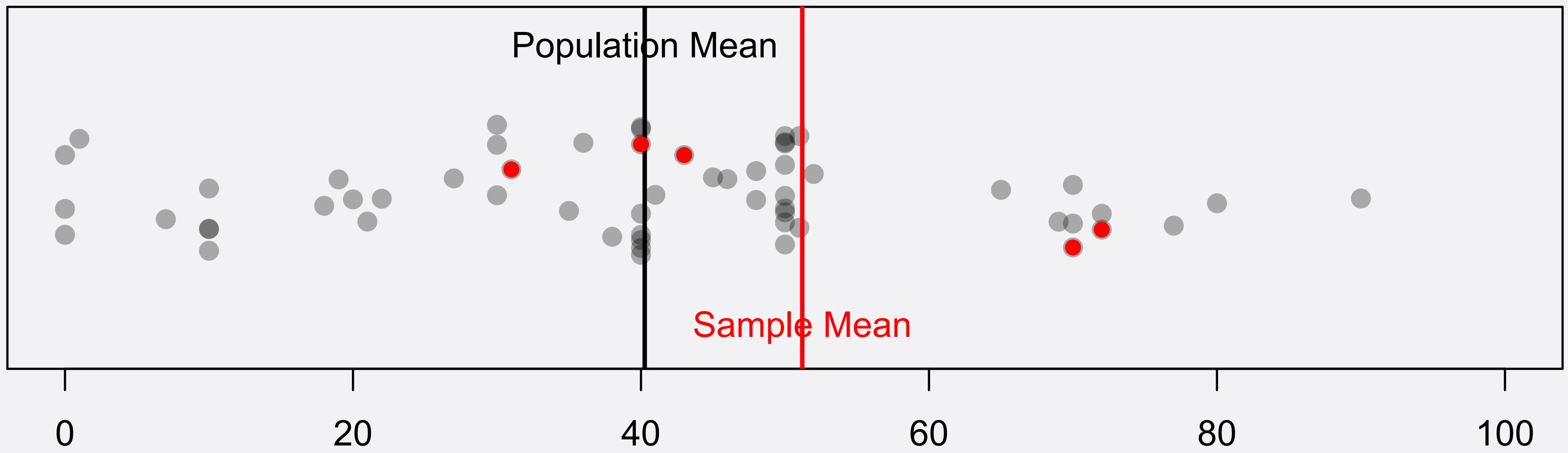
- We know that 95% of sample means are between population mean-1.96*SE and population mean+1.96*SE
- Based on this: 95% confidence interval

CONFIDENCE INTERVAL

- **95% confidence interval**
  - **Sample mean +/- (1.96 x standard error)**

$$\text{95\% CI} = \bar{x} \pm (1.96 \times \text{SE})$$

OUR SURVEY

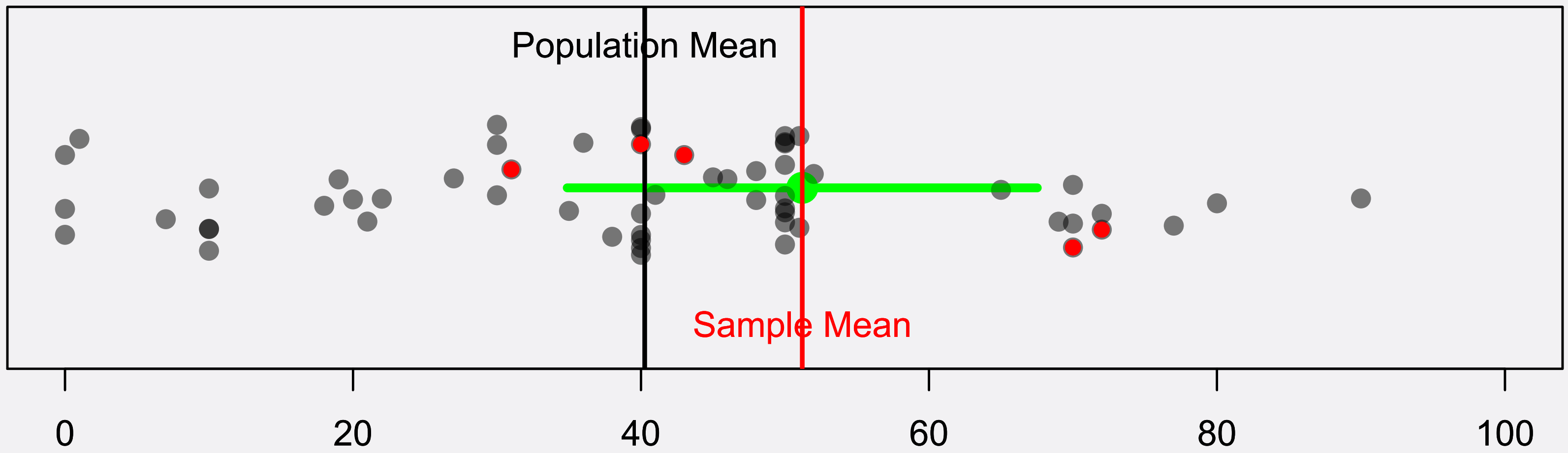


- **Random sample of 5 students**
- **Sample mean x̄=51.2 (SE=8.3)**

$$\text{Lower Bound}: \bar{x} - (1.96 \times SE) = 51.2 - (1.96 \times 8.3) = 34.9$$

$$\text{Upper Bound}: \bar{x} + (1.96 \times SE) = 51.2 + (1.96 * 8.3) = 67.5$$

OUR SURVEY



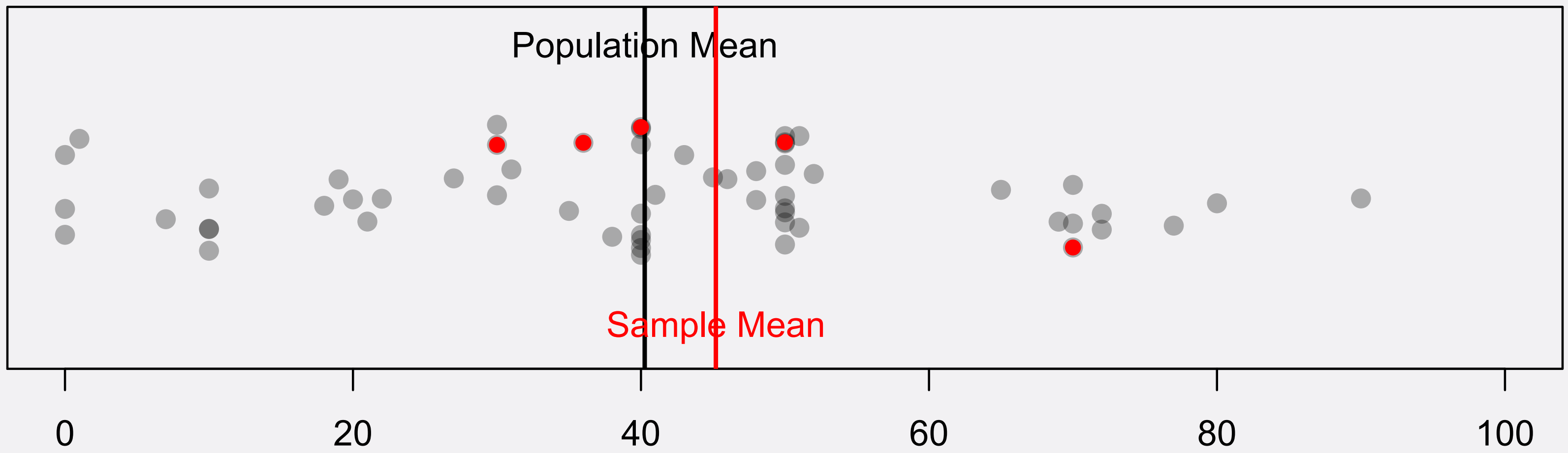
- **Random sample of 5 students**
- **Sample mean x̄=51.2 (SE=8.3)**

Lower Bound : $\bar{x} - (1.96 \times SE) = 51.2 - (1.96 \times 8.3) = 34.9$

Upper Bound : $\bar{x} + (1.96 \times SE) = 51.2 + (1.96 * 8.3) = 67.5$

- **95% CI *contains* true population mean**

OUR SURVEY

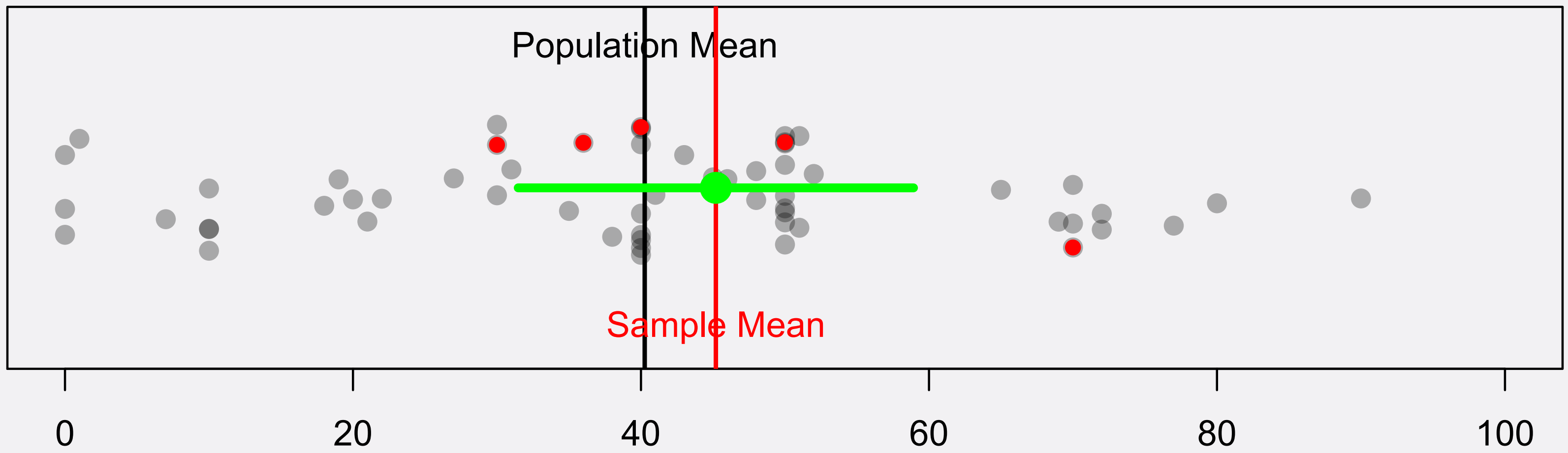


- **Another random sample of 5 students**
- **Sample mean x̄=45.2 (SE=7.0)**

$$\text{Lower Bound}: \bar{x} - (1.96 \times SE) = 45.2 - (1.96 \times 7.0) = 31.5$$

$$\text{Upper Bound}: \bar{x} + (1.96 \times SE) = 45.2 + (1.96 \times 7.0) = 58.9$$

OUR SURVEY



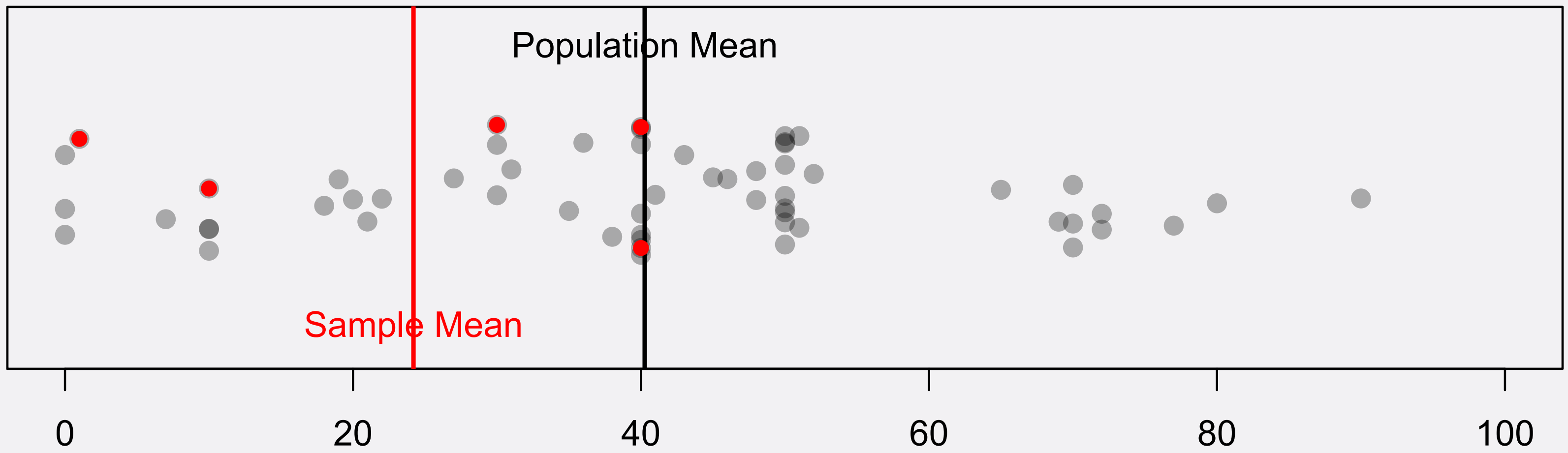
- **Another random sample of 5 students**
- **Sample mean x̄=45.2 (SE=7.0)**

$$\text{Lower Bound}: \bar{x} - (1.96 \times SE) = 45.2 - (1.96 \times 7.0) = 31.5$$

$$\text{Upper Bound}: \bar{x} + (1.96 \times SE) = 45.2 + (1.96 \times 7.0) = 58.9$$

- **95% CI *contains* true population mean**

OUR SURVEY

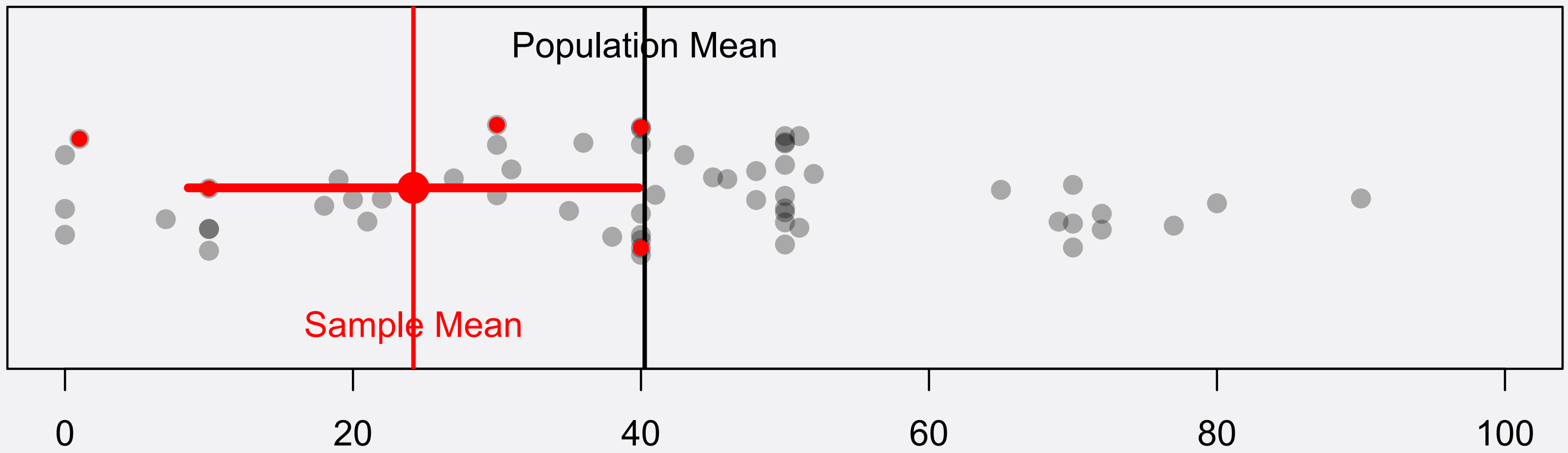


- **Another random sample of 5 students**
- **Sample mean x̄=24.2 (SE=8.0)**

$$\text{Lower Bound} : \bar{x} - (1.96 \times SE) = 24.2 - (1.96 \times 8.0) = 8.5$$

$$\text{Lower Bound} : \bar{x} + (1.96 \times SE) = 24.2 - (1.96 \times 8.0) = 39.9$$

OUR SURVEY



- **Another random sample of 5 students**
- **Sample mean x̄=24.2 (SE=8.0)**

Lower Bound : $\bar{x} - (1.96 \times SE) = 24.2 - (1.96 \times 8.0) = 8.5$

Lower Bound : $\bar{x} + (1.96 \times SE) = 24.2 - (1.96 \times 8.0) = 39.9$

- **95% CI *does not contain* true population mean**

CONFIDENCE INTERVAL

- If we do this many times:
- 95% of the confidence intervals will contain true population mean

CONFIDENCE INTERVAL

- **95% confidence interval: Interval around sample mean that would contain true population mean in 95% of repeated samples**

WHAT WE CAN DO WITH THIS

POLITICS SEPTEMBER 22, 2021

Biden's Approval Rating Hits New Low of 43%; Harris' Is 49%

Results for this Gallup poll are based on telephone interviews conducted Sept. 1-17, 2021, with a random sample of 1,005 adults, aged 18 and older, living in all 50 U.S. states and the District of Columbia. For results based on the total sample of national adults, the margin of sampling error is ±4 percentage points at the 95% confidence level. All reported margins of sampling error include computed design effects for weighting.

- **What does this mean?**

CONFIDENCE INTERVAL

- Poll: 1,005 respondents (randomly selected)
  - Found $\bar{x} = 43$
  - $SE=2.05$

CONFIDENCE INTERVAL

- **Poll: 1,005 respondents (randomly selected)**
  - **Found x̄ = 43**
  - **SE=2.05**
- **95% CI:**

Lower Bound : $\bar{x} - 1.96 \times SE = 43 - 1.96 \times 2.05 = 39.0$

Upper Bound : $\bar{x} + 1.96 \times SE = 43 + 1.96 \times 2.05 = 47.0$

CONFIDENCE INTERVAL

- **So: 95% confidence interval is**

$$43 \pm 4 = (39, 47)$$

Results for this Gallup poll are based on telephone interviews conducted Sept. 1-17, 2021, with a random sample of 1,005 adults, aged 18 and older, living in all 50 U.S. states and the District of Columbia. For results based on the total sample of national adults, the margin of sampling error is ±4 percentage points at the 95% confidence level. All reported margins of sampling error include computed design effects for weighting.

- **If we do many polls of Biden's approval rating and report 95% CI for each, 95% of them will contain (unknown) true approval rating in population**

RECAP

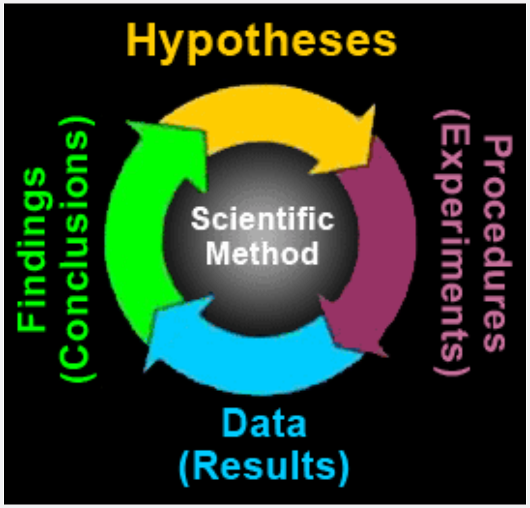
- Core problem: We are interested in population parameters, but usually only have a (random) sample
  - Is what we find in our sample representative of what is going on in the population?
- Answer: We can never be sure that it is, but we can give a *probability* of how sure we are that it is
  - 95% confidence interval

TODAY

- Finishing up Sampling and Surveys
- Hypotheses and causality

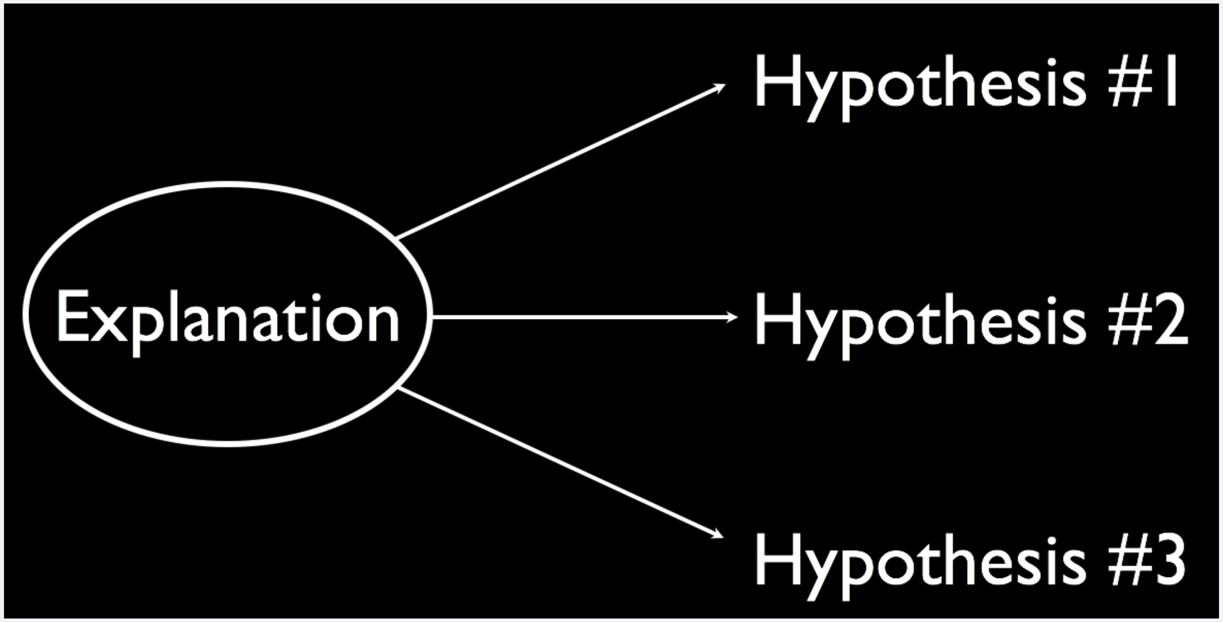
RESEARCH PROCESS

- Formulate research question
- Propose explanation/theory, hypotheses
- Data collection process
- Use data to evaluate hypotheses
- Reassess explanation



HYPOTHESES AND THEORY

- Explanation/Theory: (Simplified) description of how social reality works
- Hypotheses: Statements what, if the theory is true, we should observe in *our* data



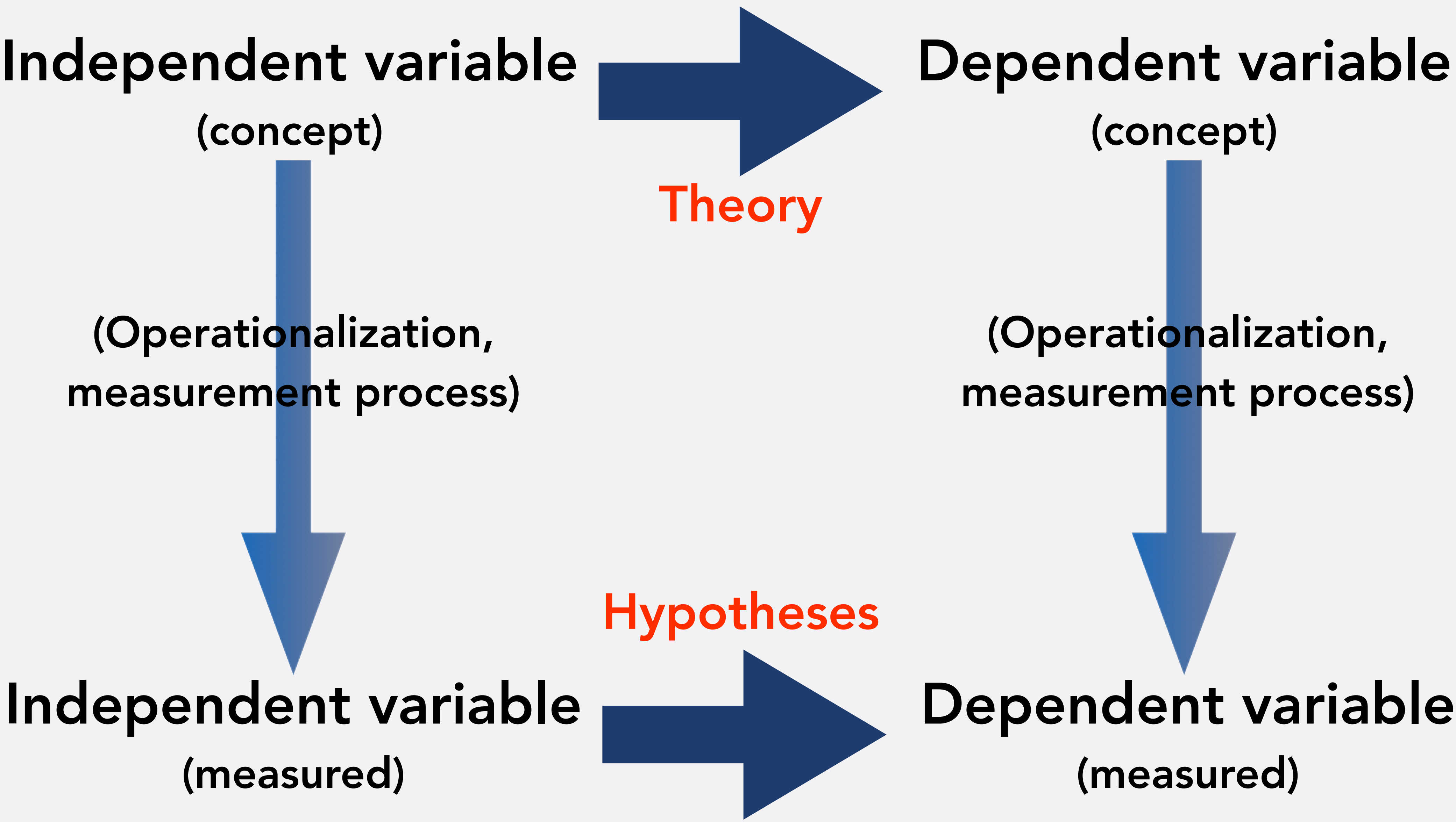
THEORY

Independent variable
(concept)

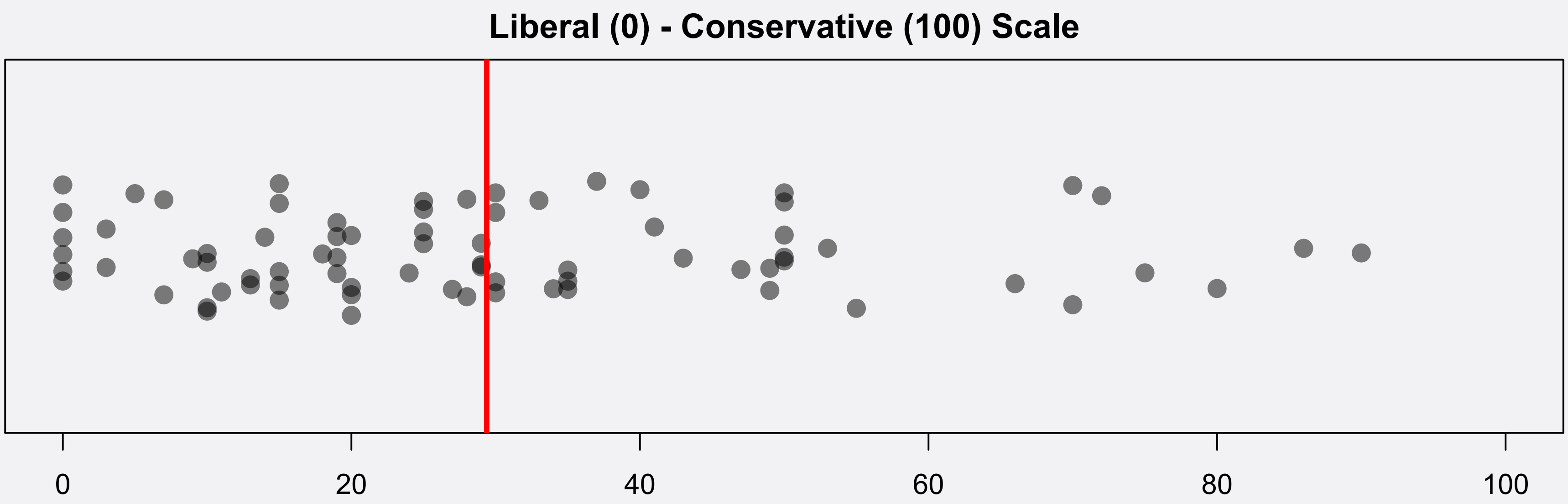
Dependent variable
(concept)

Theory

MEASUREMENT



TODAY'S EXAMPLE



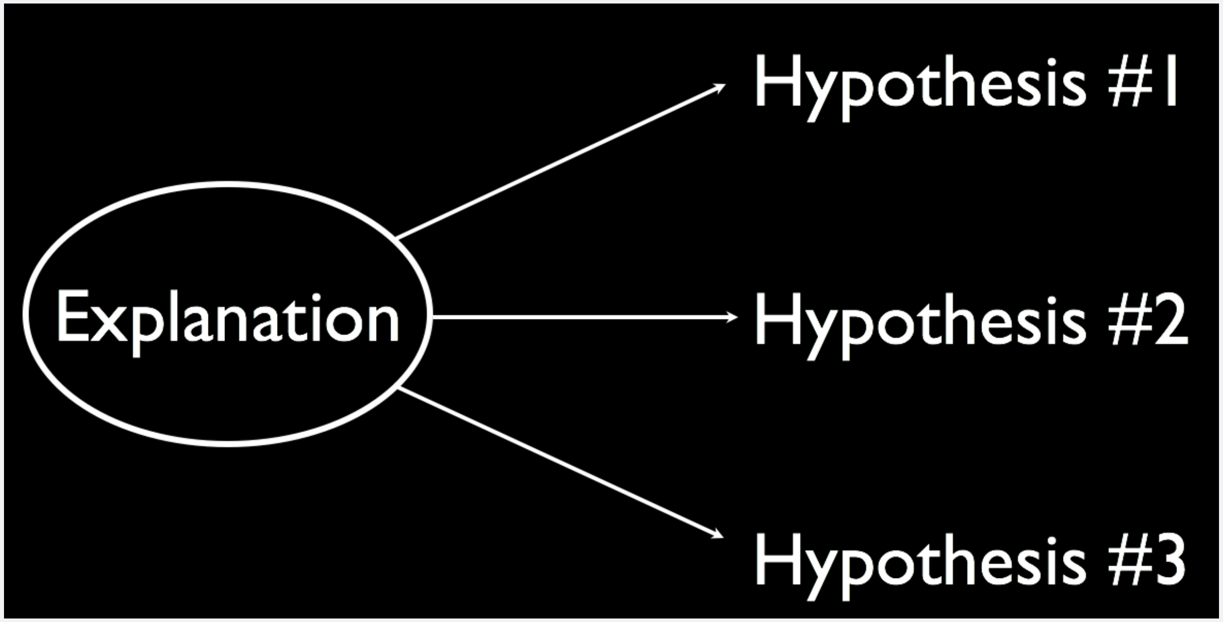
- **Why are some students in 202 liberal and others conservative? What determines which ideology students have?**

PROPOSE EXPLANATION/ THEORY, HYPOTHESES

- Explanation/Theory: broad statement about how, and why the world works in a specific way
  - Example: People's ideology is influenced by their upbringing

PROPOSE EXPLANATION/ THEORY, HYPOTHESES

- Hypotheses: *empirically* testable statements that follows from a theory



PROPOSE EXPLANATION/ THEORY, HYPOTHESES

- Hypotheses: *Empirically testable* statements that follows from a theory
  - Hypothesis 1: Students whose parents are conservative are on average more conservative than students with liberal parents

PROPOSE EXPLANATION/ THEORY, HYPOTHESES

- Hypotheses: *Empirically testable* statements that follows from a theory
  - Hypothesis 2: Students who grew up in a conservative area are on average more conservative than students who grew up in a liberal area

PROPOSE EXPLANATION/ THEORY, HYPOTHESES

- **Hypotheses: *Empirically testable* statements that follows from a theory**
  - **Hypothesis 3: Students who attended a STEM-focused high school are on average more conservative than those who attended a Liberal Arts-focused high school**