PSC 202
SYRACUSE UNIVERSITY

INTRODUCTION TO POLITICAL ANALYSIS

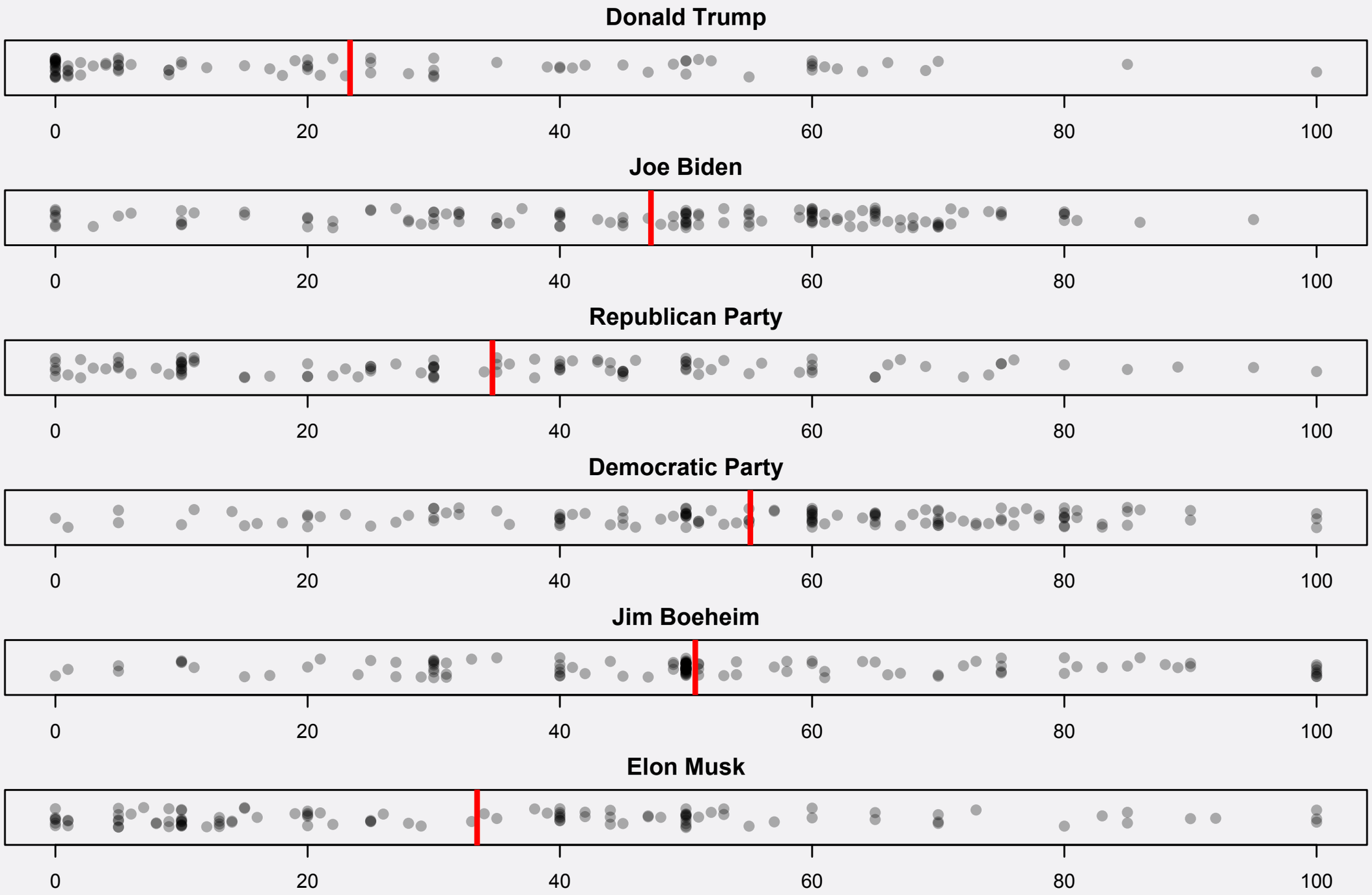
MORE SAMPLING AND SURVEYS, HYPOTHESES AND CAUSALITY

HOUSEKEEPING

- There will be sections on Friday
- Problem set 4 will be posted on Friday
- Also new reading quiz for Monday

SURVEY RESULTS

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



SURVEY RESULTS

- **How much do you agree with the Supreme Court's decision to overturn Roe v. Wade (which guaranteed the right to have an abortion)?**

	Number	Percentage
Strongly agree	9	6.8%
Somewhat agree	4	3.0%
Neither agree nor disagree	5	3.8%
Somewhat disagree	16	12.1%
Strongly disagree	98	74.2%

SURVEY RESULTS

- **How much do you agree with the following statement: Members of Congress should be more willing to compromise with members of the other party, even if they have to agree to things that go against the interest of their constituents**

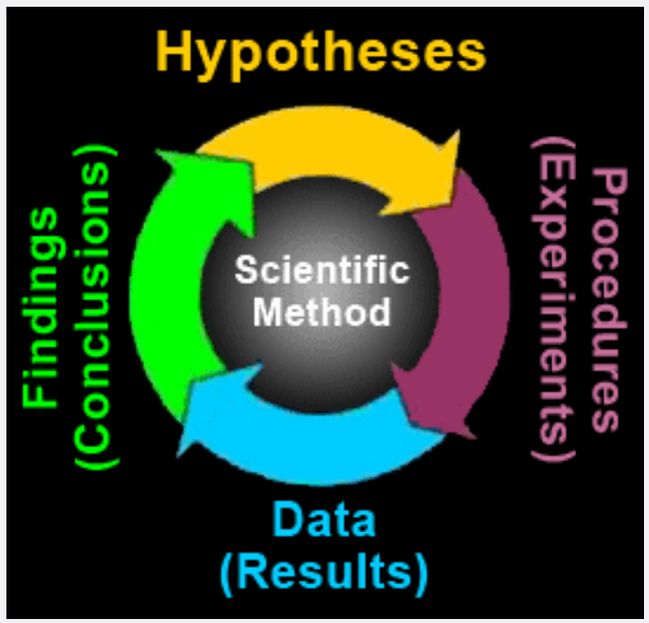
	Number	Percentage
Strongly agree	37	28.2%
Somewhat agree	56	42.7%
Neither agree nor disagree	21	16.0%
Somewhat disagree	10	7.6%
Strongly disagree	7	5.3%

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 JANUARY 25, 2023

Biden Averaged 41% Job Approval in His Second Year

Results for this Gallup poll are based on telephone interviews conducted Jan. 2-22, 2023, with a random sample of 1,011 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 41% approval rating among 1,011 respondents is close to the approval rating of *all* American voters?**

RECAP

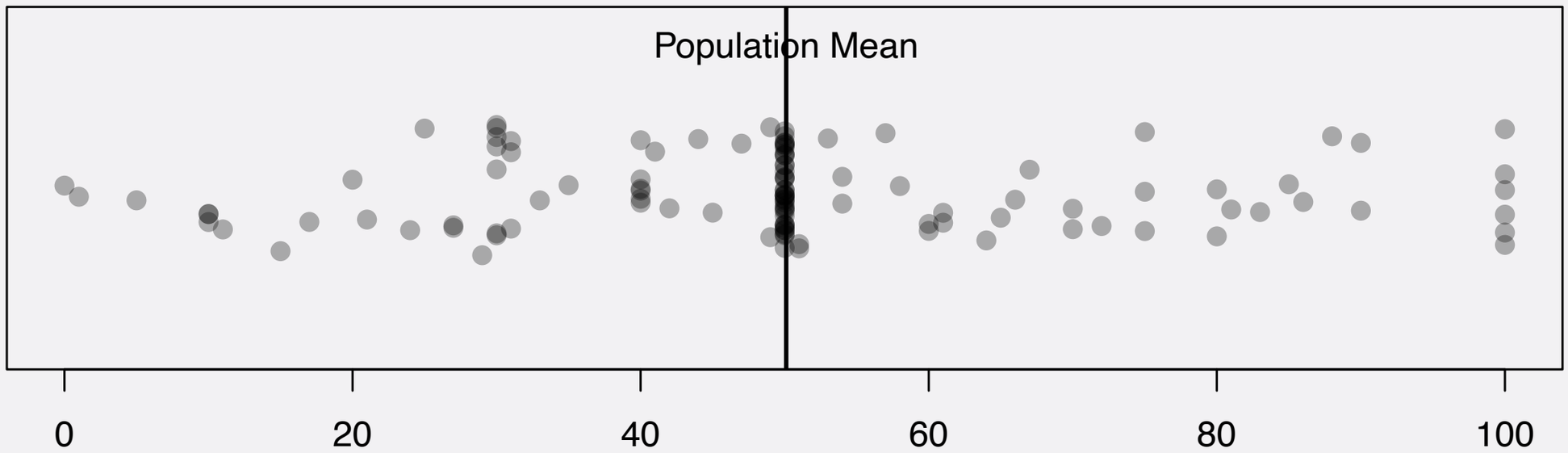
Unknown: Approval rating in population

Known: Approval rating in survey

- Population parameter = Sample statistic + random sampling error

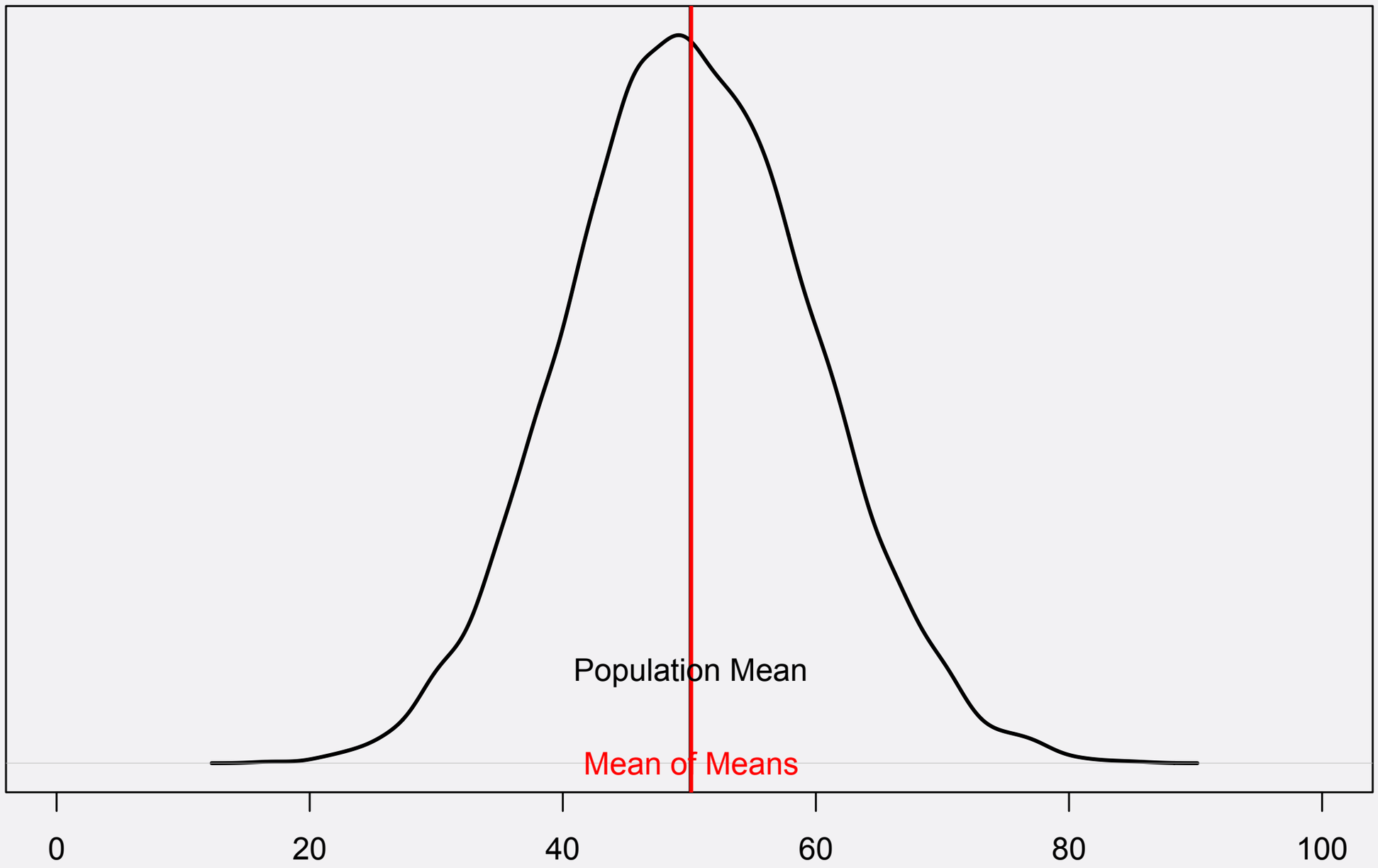
We can figure this out

RECAP

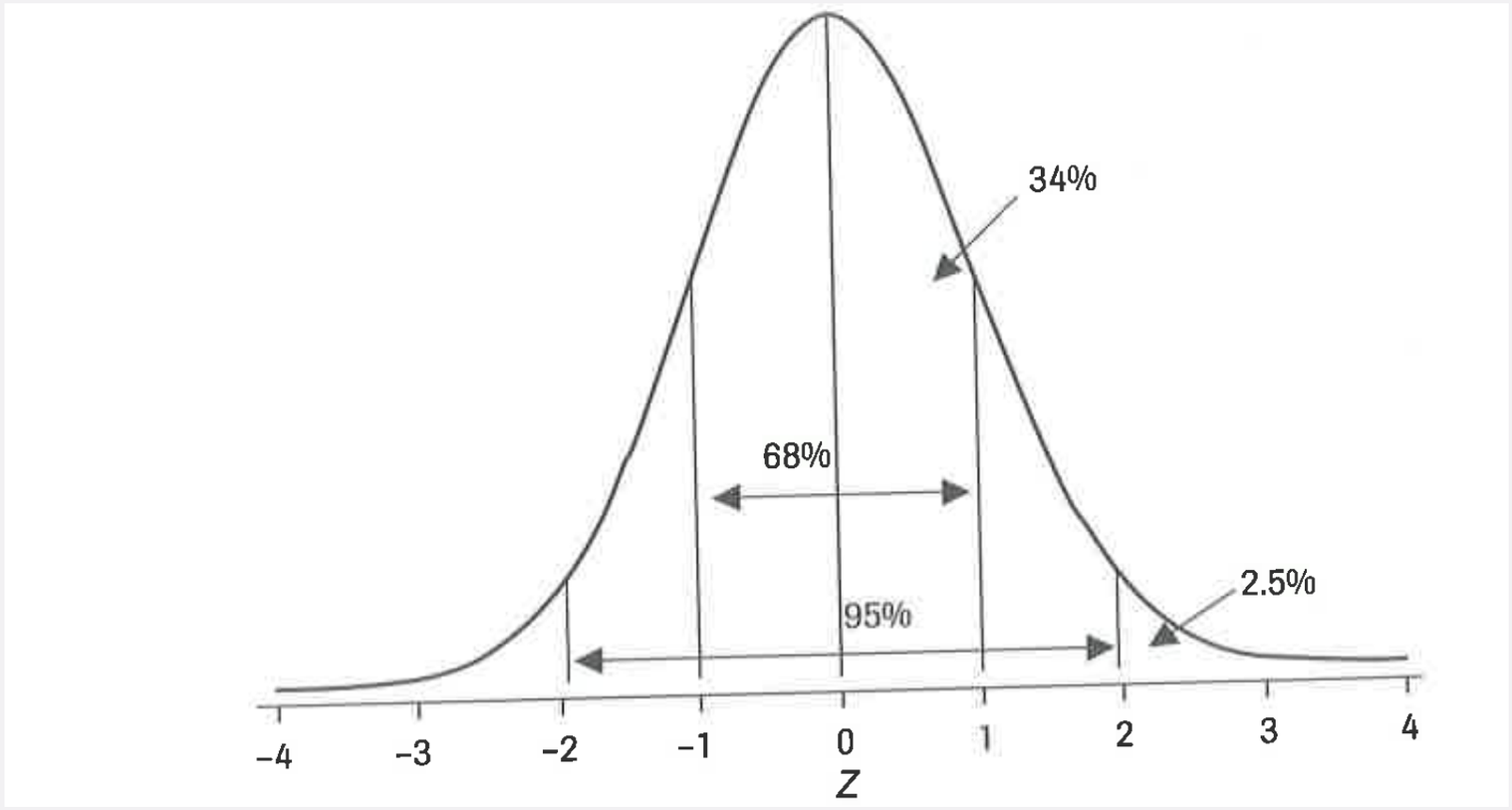


- Took random sample of 5 students
- Record average rating of those 5 students
- Do this thousands of times

AFTER 10,000 RANDOM SAMPLES

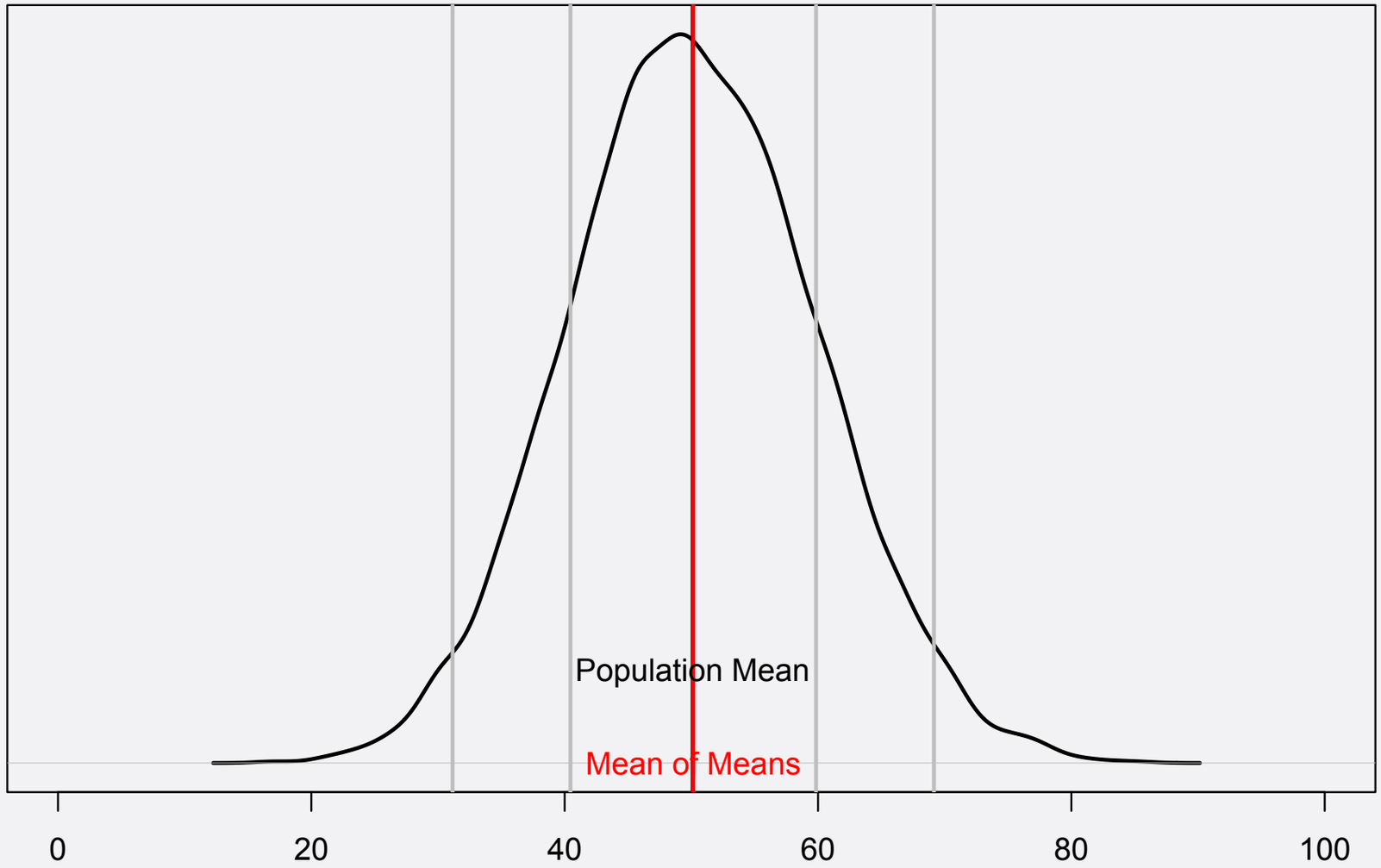


RECAP



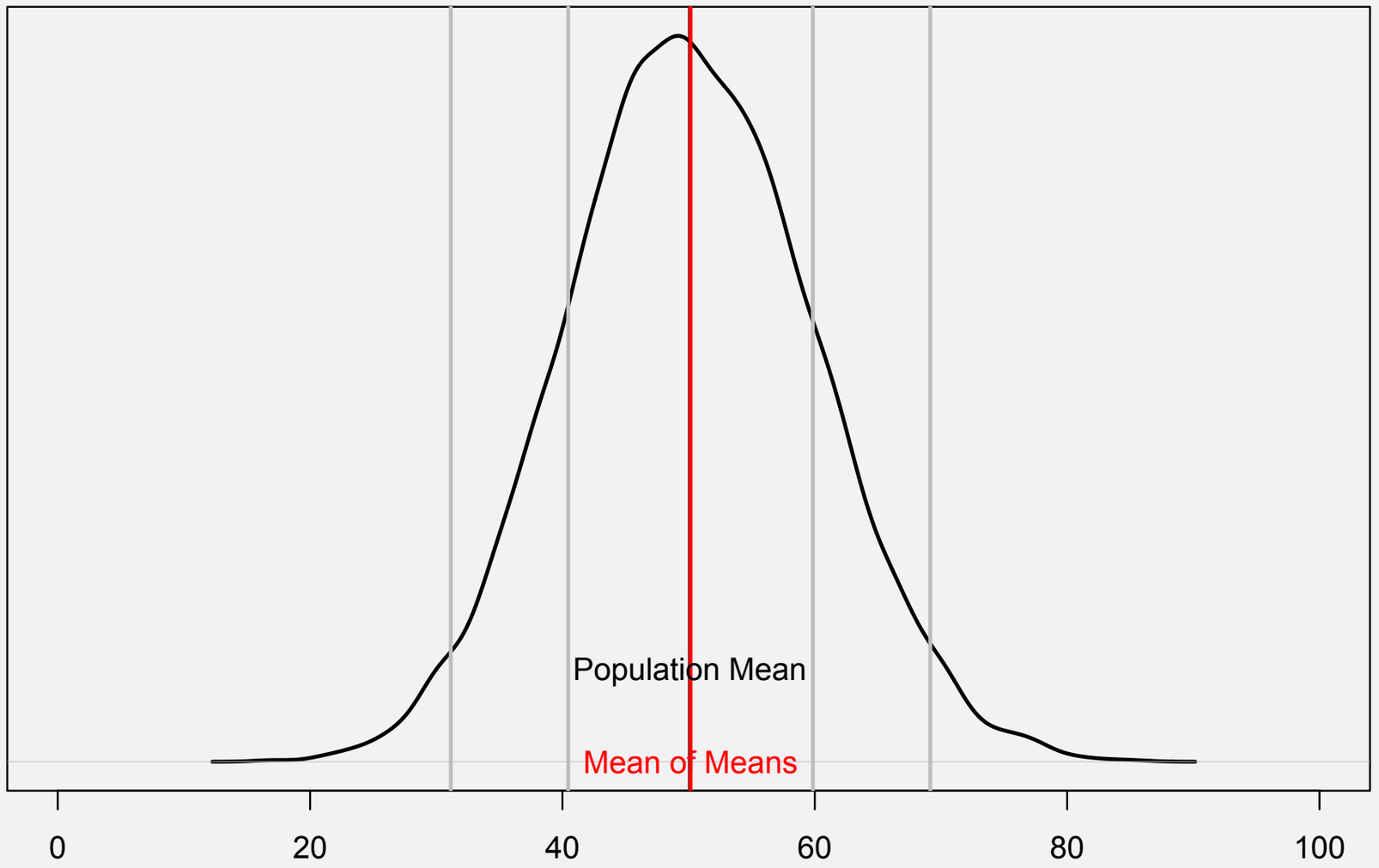
- If a variable follows a Normal distribution…
  - we know that 68% of observations are between mean ± 1SD, 95% between mean ± 1.96SD

RECAP



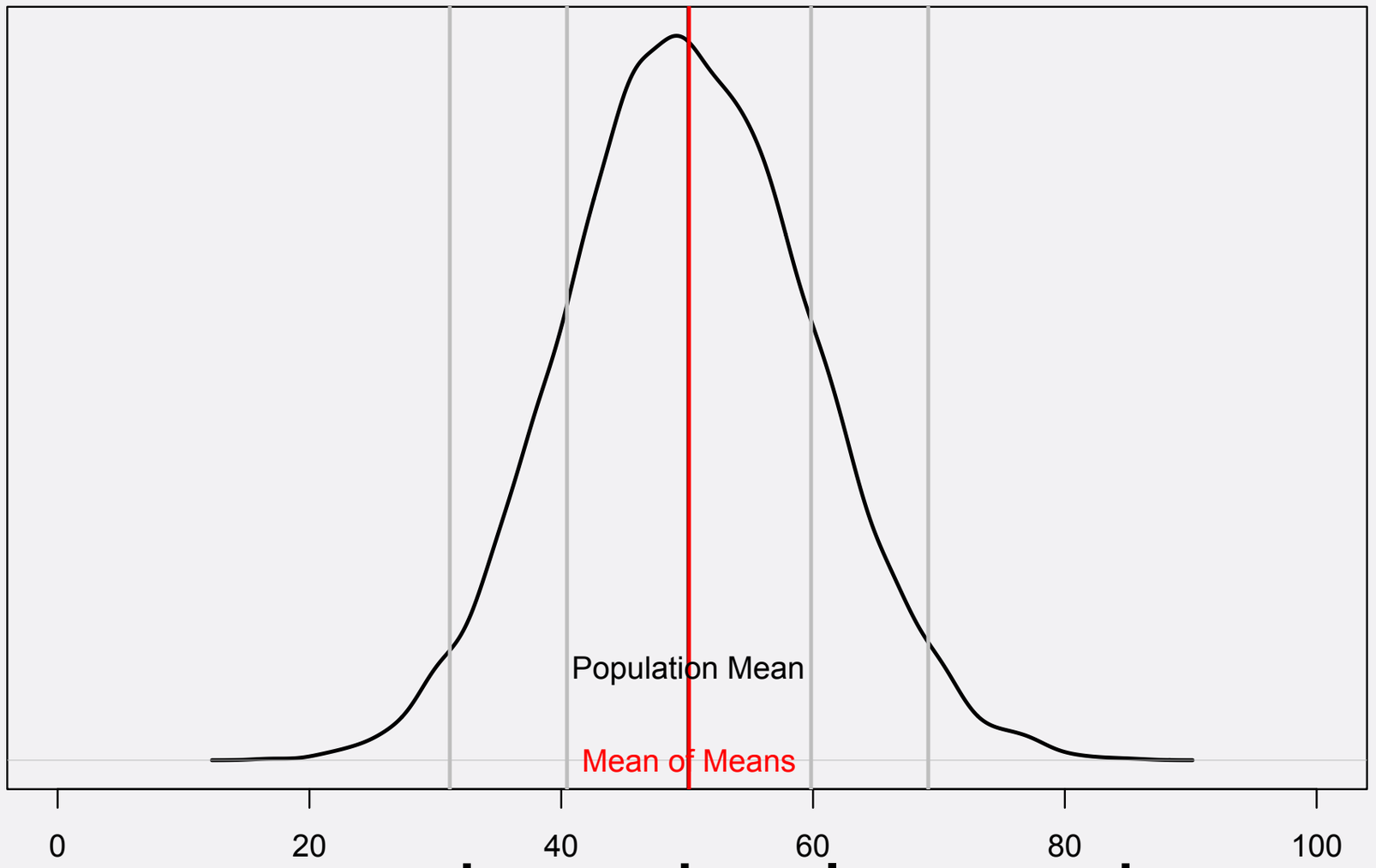
- If we take many random samples from a population
- and compute mean for each
- those means have a Normal distribution

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?

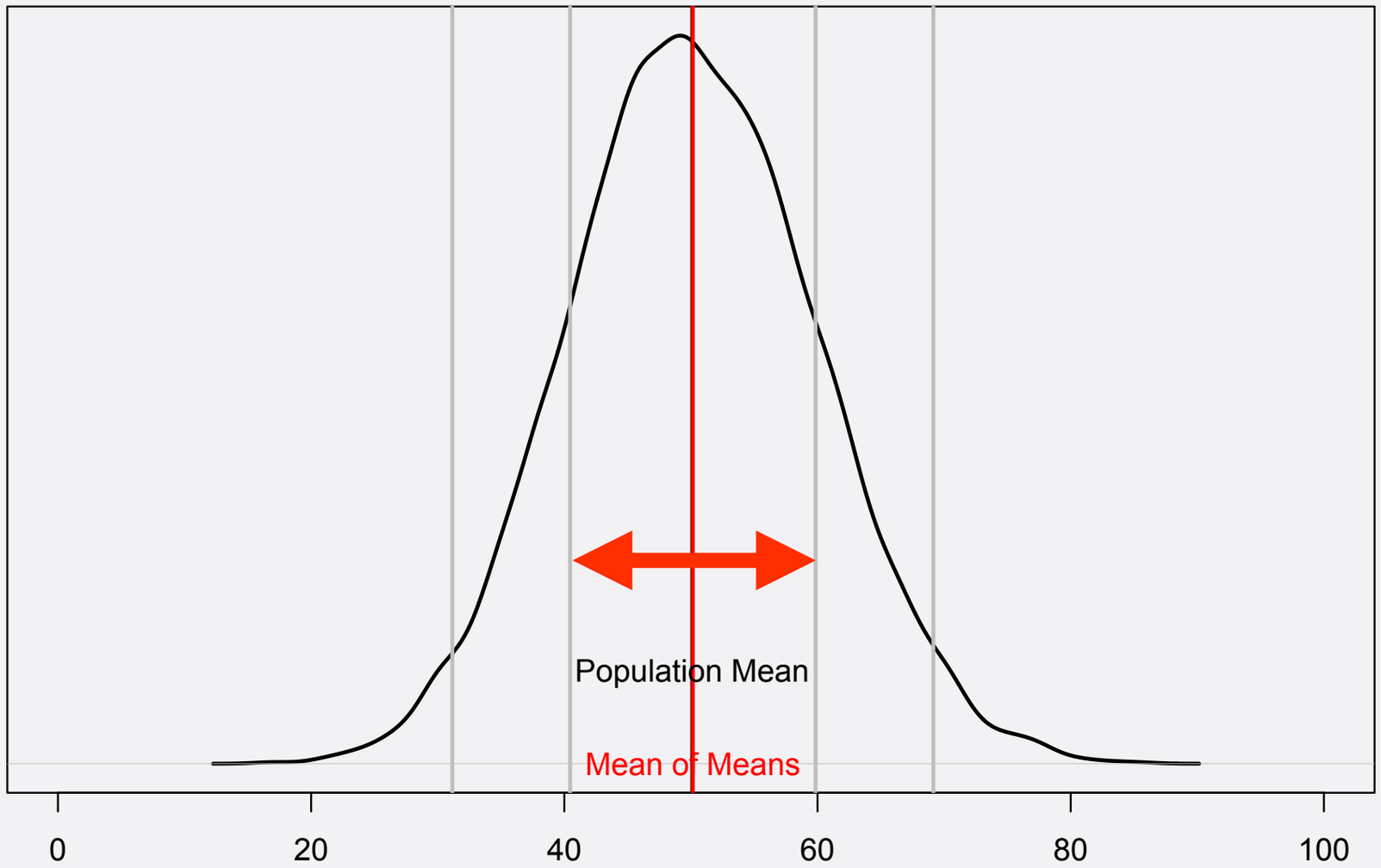


- Allows us to put a number on how large random measurement error is
- Which tells us how confident we can be that conclusions we draw from a sample hold in the population overall

HOW DOES THIS HELP?

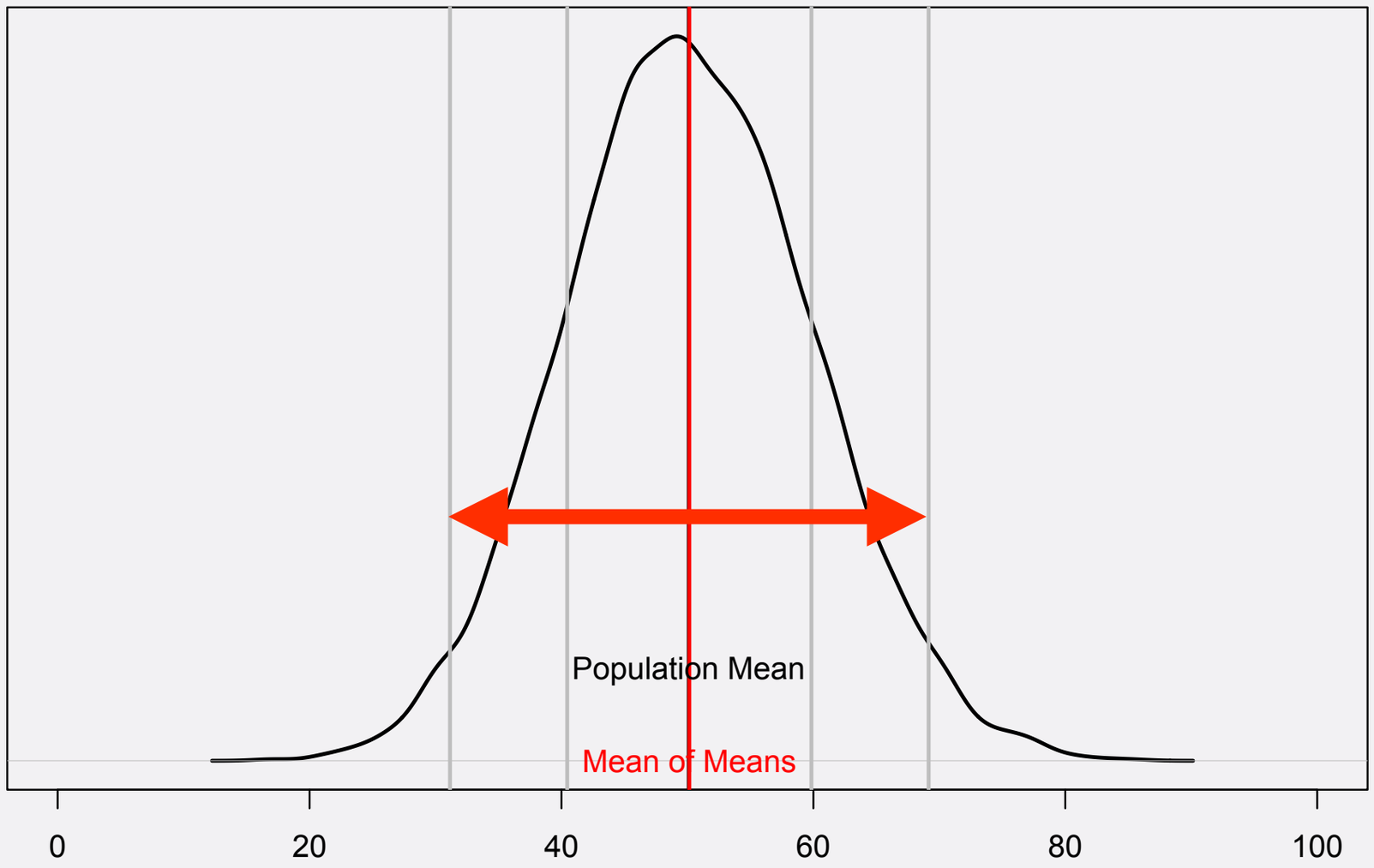
- How it helps, part 1:
- We can quantify how large the random sampling error in our sample is

RANDOM SAMPLING ERROR



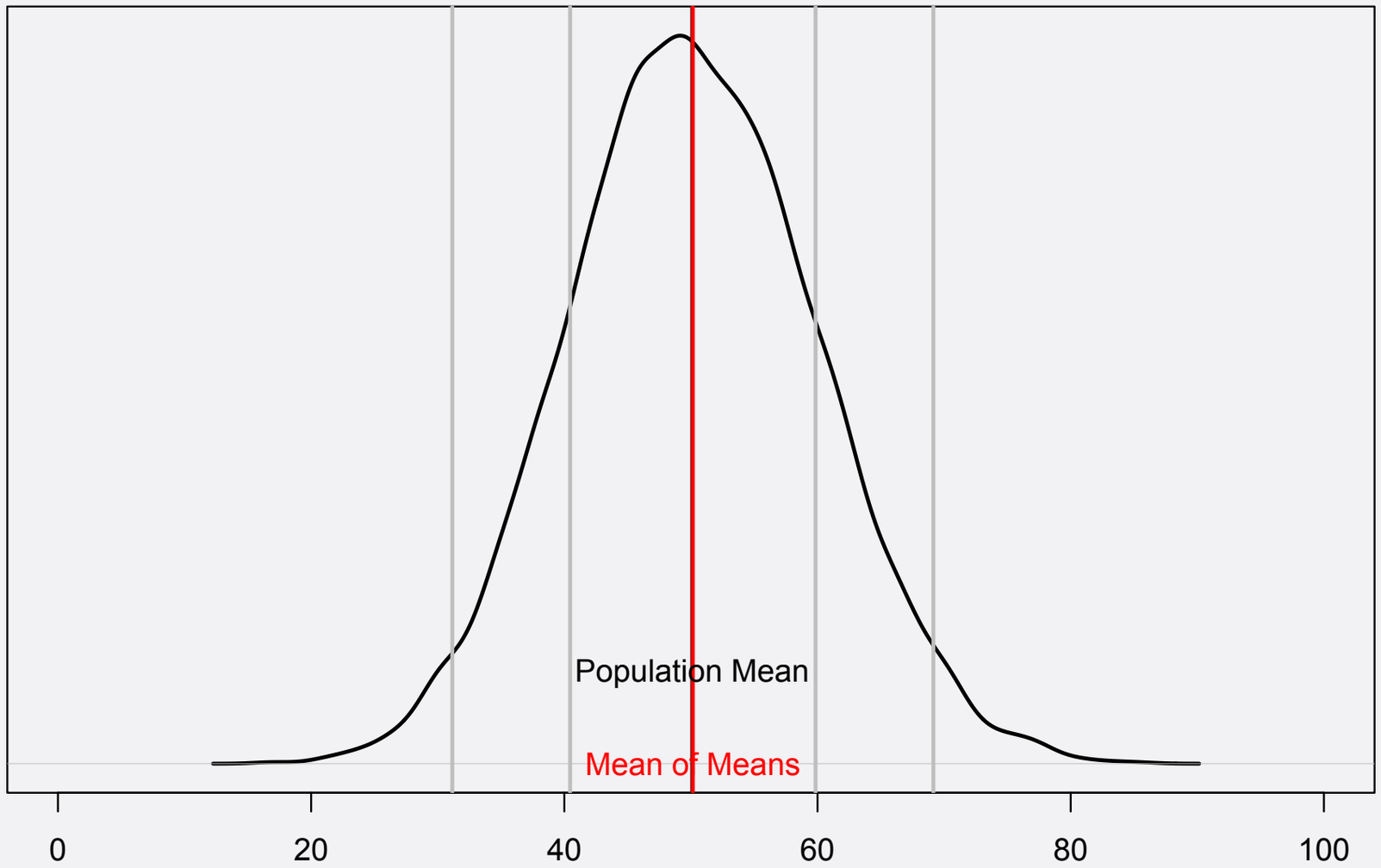
- 68% of sample means: population mean ± 1 standard deviations

RANDOM SAMPLING ERROR



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

RANDOM SAMPLING ERROR



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

STANDARD ERROR

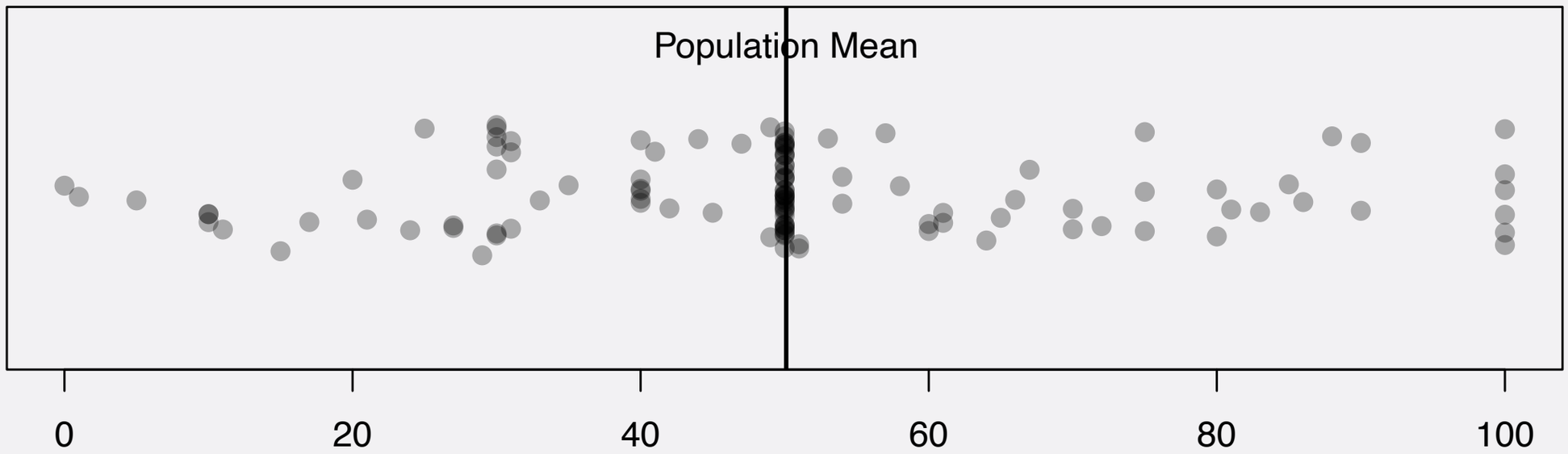
- The standard deviation of the distribution of sample means is called the *standard error of the sample mean*
- or simply *standard error*
- Standard error is a measure of the random sampling error

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



- Mean ($\bar{x}$): 50.1
- Standard deviation (s): 22.1
- Size of sample (n): 116

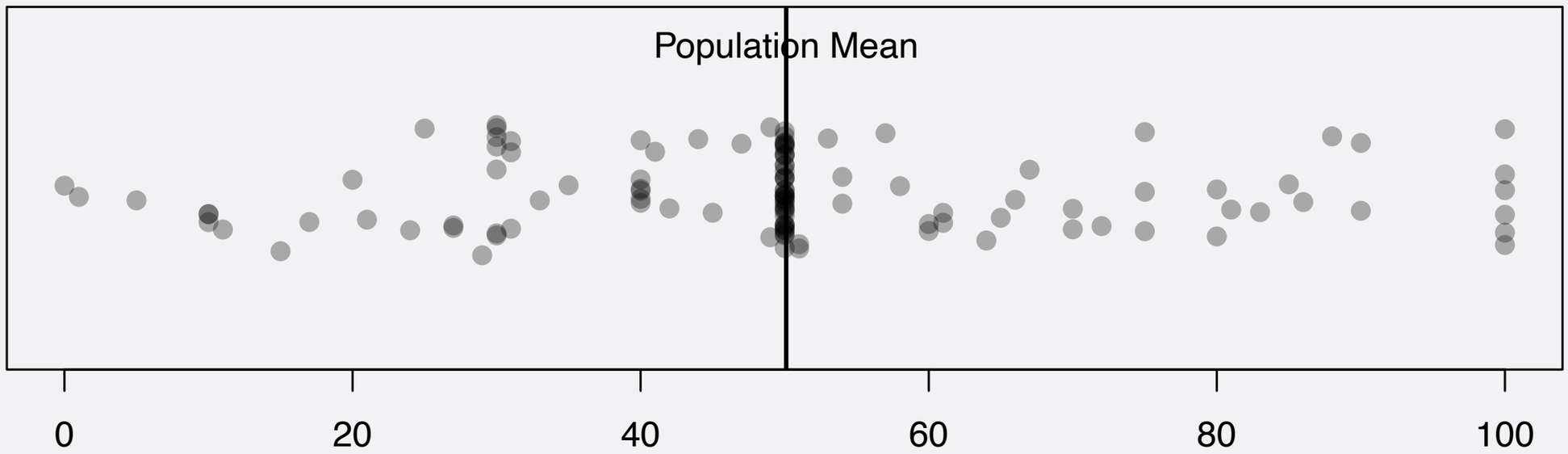
$$SE = \frac{s}{\sqrt{n}}$$

NOTE

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

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

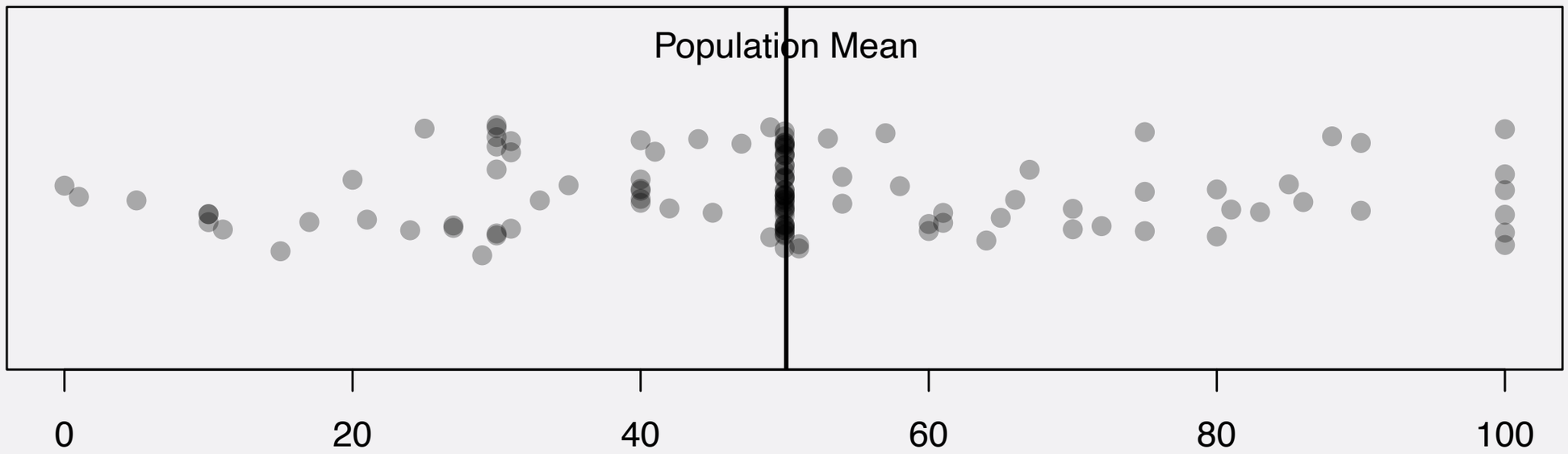
STANDARD ERROR



- Mean ($\bar{x}$): 50.1
- Standard deviation (s): 22.1
- Size of sample (n): 116

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

STANDARD ERROR



- **Mean ($\bar{x}$): 50.1**
- **Standard deviation (s): 22.1**
- **Size of sample (n): 116**

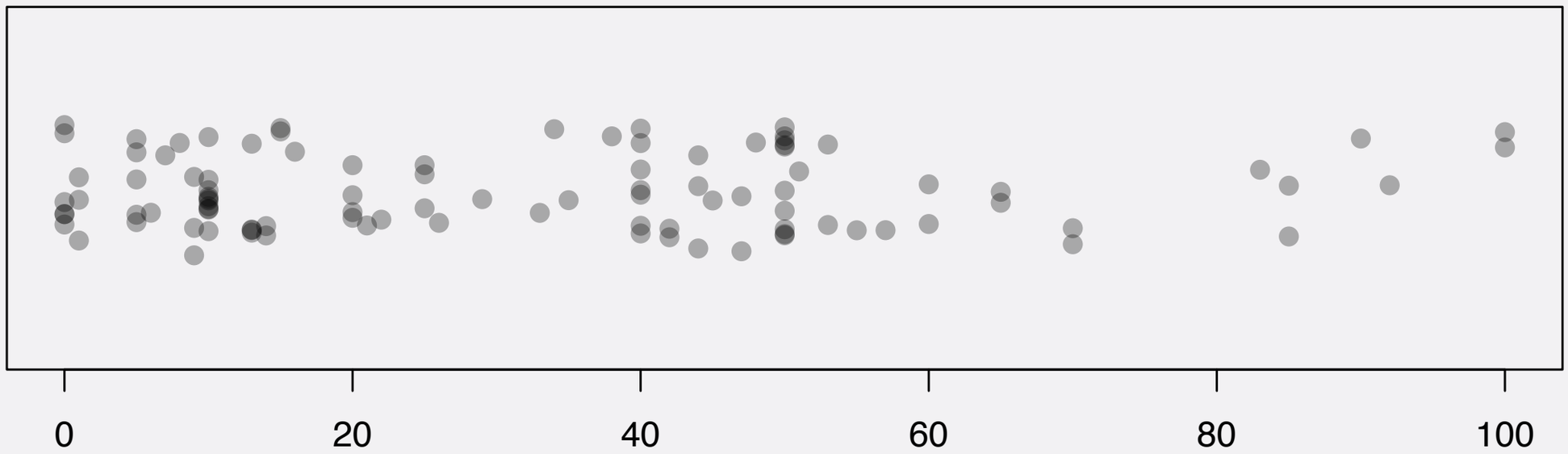
$$SE = \frac{s}{\sqrt{n}} = \frac{22.1}{\sqrt{116}} = 2.05$$

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

ELON MUSK



- **Mean ($\bar{x}$): 32.4**
- **Standard deviation (s): 25.6**
- **Size of sample (n): 116**

$$SE = \frac{s}{\sqrt{n}} = \frac{25.6}{\sqrt{116}} = 2.38$$

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

ELON MUSK

- **Mean ($\bar{x}$): 32.4**
- **Standard deviation (s): 25.6**
- **Size of sample (n): 116**

$$SE = \frac{s}{\sqrt{n}} = \frac{25.6}{\sqrt{116}} = 2.38$$

ELON MUSK

Let’s say we had interviewed 1160 people instead

- Mean ($\bar{x}$): 32.4
- Standard deviation (s): 25.6
- Size of sample (n): 1160

$$SE = \frac{s}{\sqrt{n}} = \frac{25.6}{\sqrt{1160}} = 0.75$$

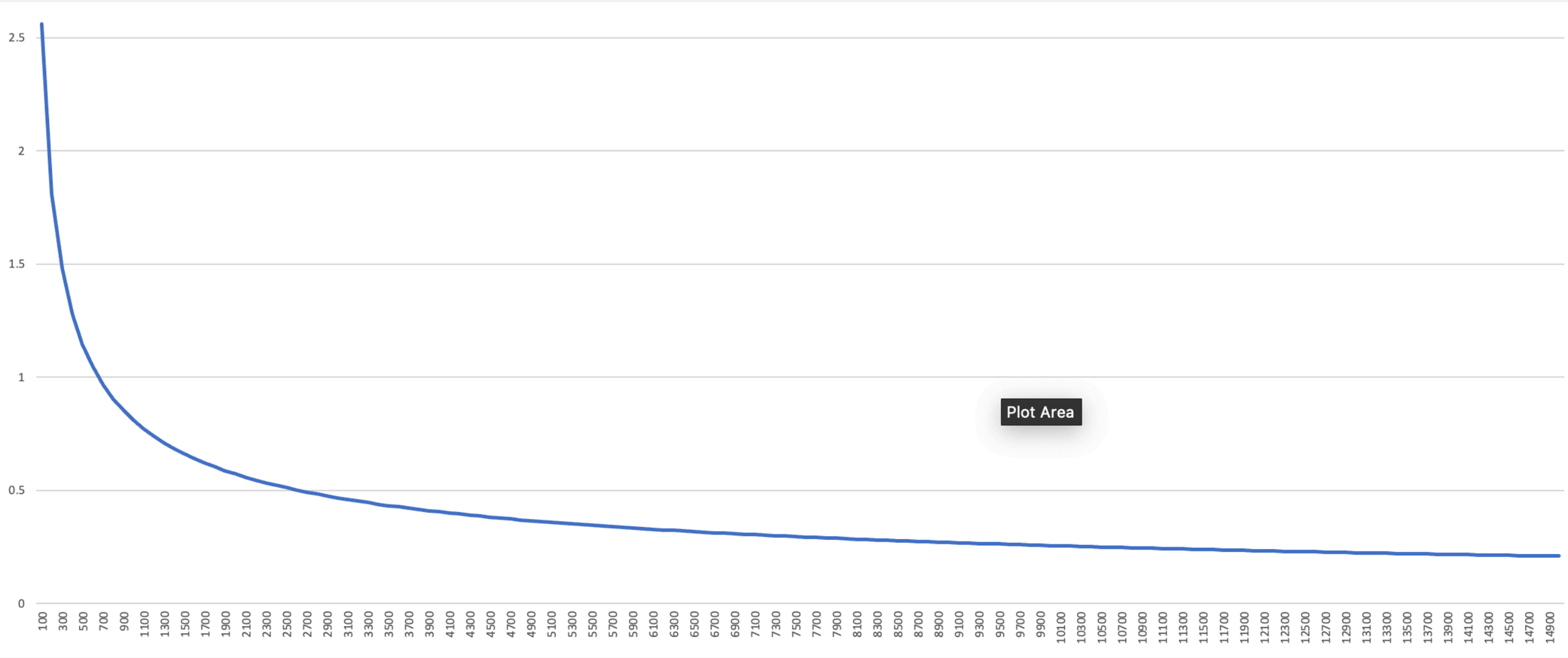
ELON MUSK

Let’s say we had interviewed 11600 people instead

- Mean ($\bar{x}$): 32.4
- Standard deviation (s): 25.6
- Size of sample (n): 11600

$$SE = \frac{s}{\sqrt{n}} = \frac{25.6}{\sqrt{11600}} = 0.24$$

SAMPLE SIZE



- Horizontal: n Vertical: SE

- Why surveys are often only 500-1000 people

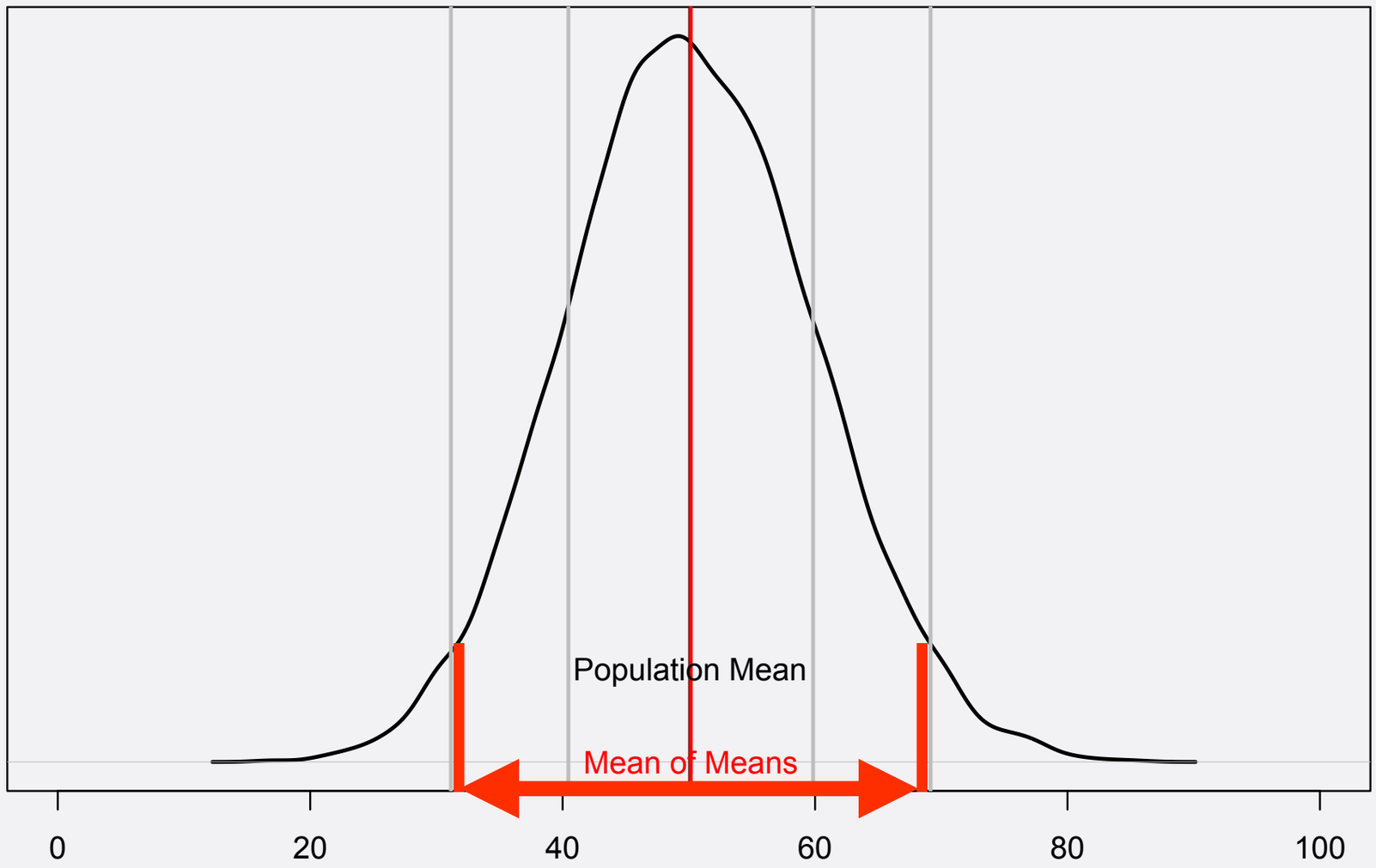
SUMMARY: STANDARD ERROR

- 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



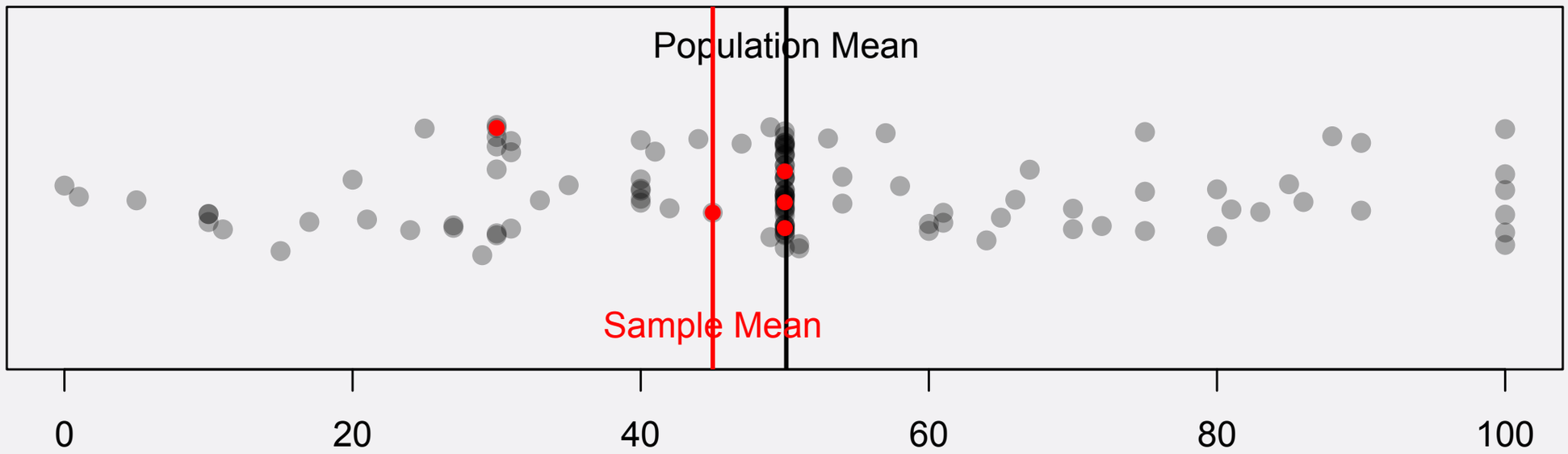
- We know that 95% of sample means are between population mean-1.96*SE and population mean+1.96*SE
- Based on this: 95% confidence interval

CONFIDENCE INTERVAL

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

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

OUR SURVEY

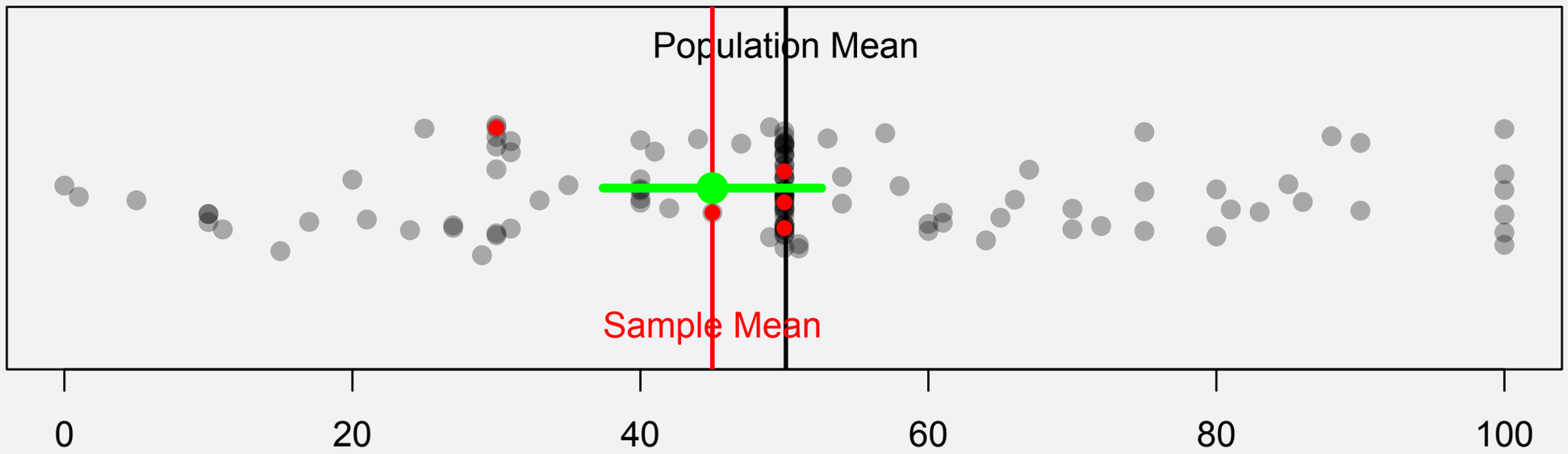


- **Random sample of 5 students**
- **Sample mean $\bar{x}$=45.0 (SE=3.87)**

Lower Bound : $\bar{x} - (1.96 \times SE) = 45.0 - (1.96 \times 3.87) = 37.4$

Upper Bound : $\bar{x} + (1.96 \times SE) = 45.0 - (1.96 \times 3.87) = 52.6$

OUR SURVEY



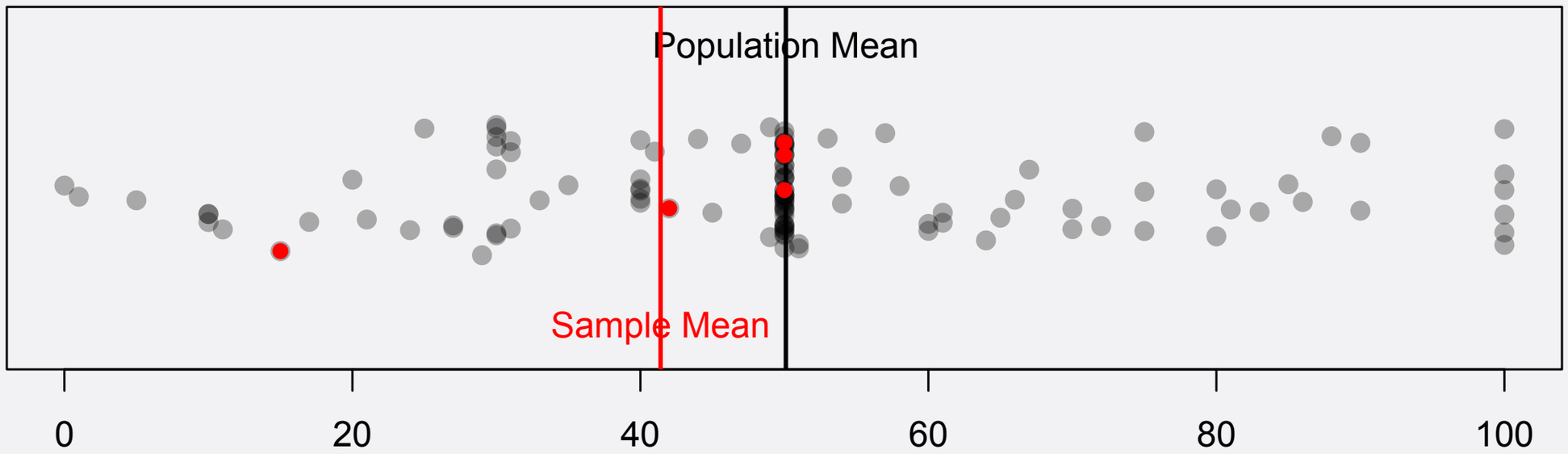
- **Random sample of 5 students**
- **Sample mean $\bar{x}$=45.0 (SE=3.87)**

Lower Bound : $\bar{x} - (1.96 \times SE) = 45.0 - (1.96 \times 3.87) = 37.4$

Upper Bound : $\bar{x} + (1.96 \times SE) = 45.0 - (1.96 \times 3.87) = 52.6$

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

OUR SURVEY

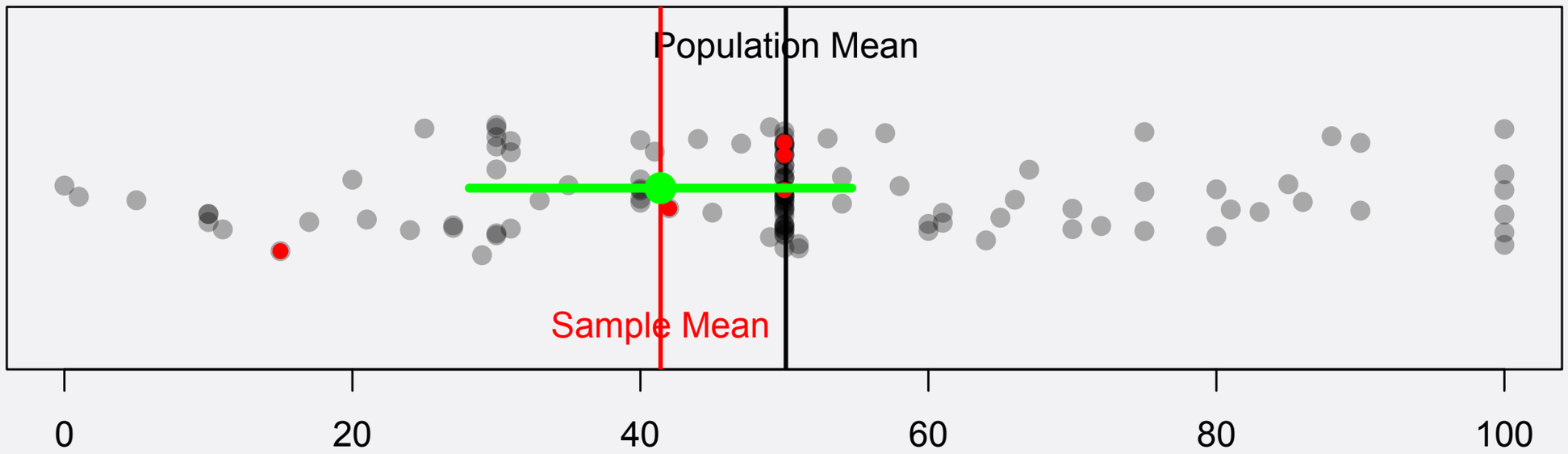


- **Another random sample of 5 students**
- **Sample mean $\bar{x}$=41.4 (SE=6.8)**

Lower Bound : $\bar{x} - (1.96 \times SE) = 41.4 - (1.96 \times 6.8) = 28.1$

Upper Bound : $\bar{x} + (1.96 \times SE) = 41.4 + (1.96 \times 6.8) = 54.7$

OUR SURVEY



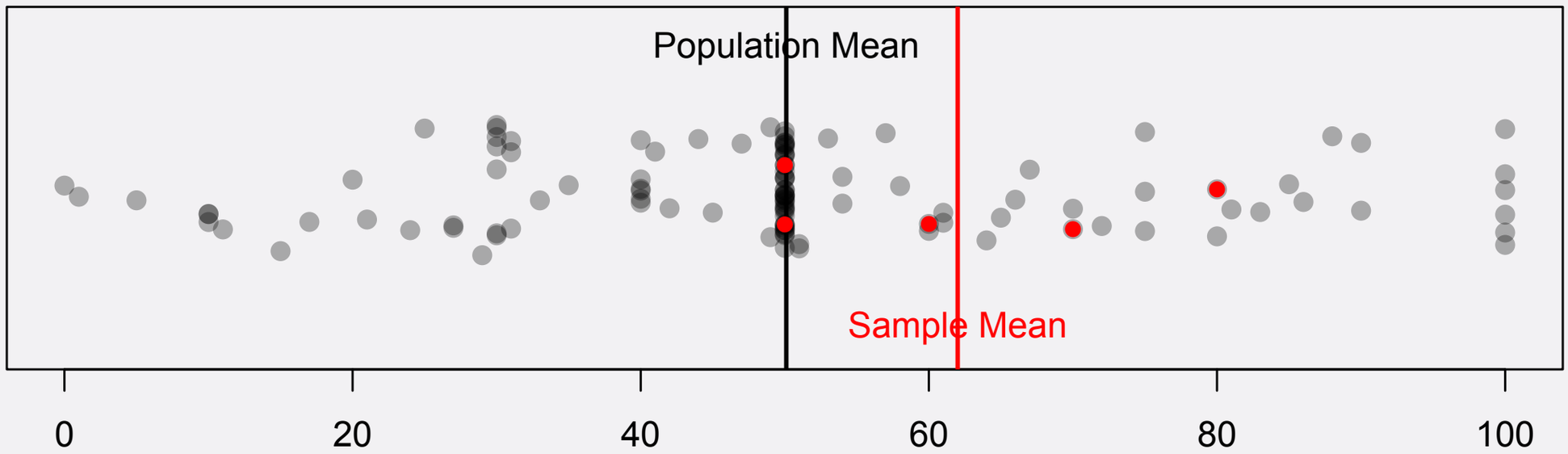
- **Another random sample of 5 students**
- **Sample mean $\bar{x}$=41.4 (SE=6.8)**

Lower Bound : $\bar{x} - (1.96 \times SE) = 41.4 - (1.96 \times 6.8) = 28.1$

Upper Bound : $\bar{x} + (1.96 \times SE) = 41.4 + (1.96 \times 6.8) = 54.7$

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

OUR SURVEY

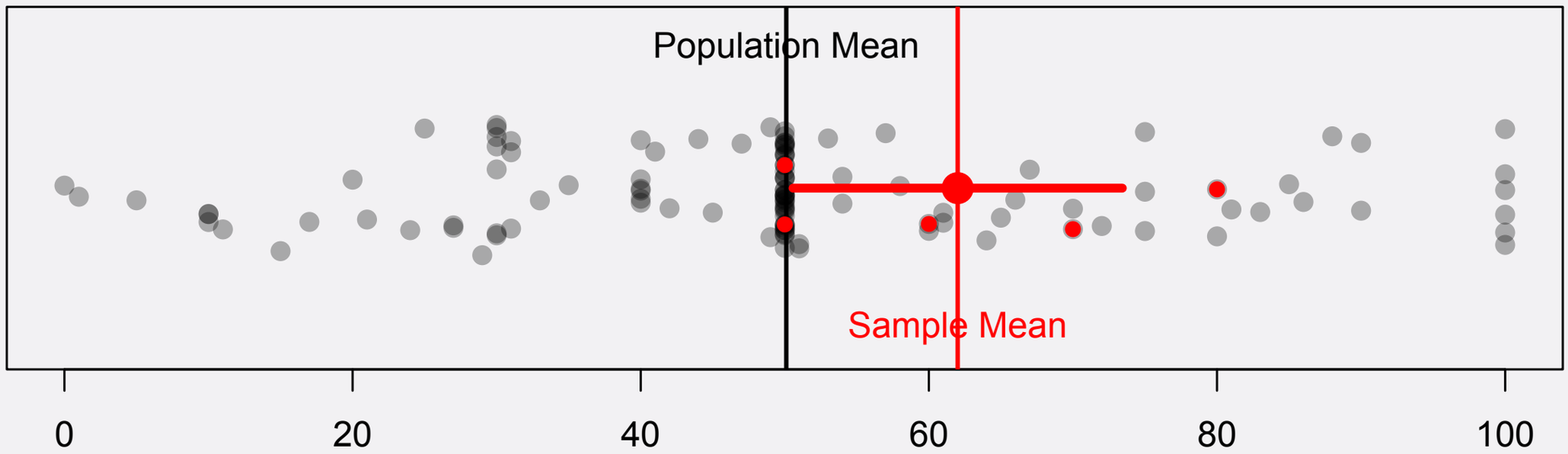


- **Another random sample of 5 students**
- **Sample mean $\bar{x}$=62.0 (SE=5.8)**

Lower Bound : $\bar{x} - (1.96 \times SE) = 62 - (1.96 \times 5.8) = 50.6$

Upper Bound : $\bar{x} + (1.96 \times SE) = 62 + (1.96 \times 5.8) = 73.4$

OUR SURVEY



- **Another random sample of 5 students**
- **Sample mean $\bar{x}$=62.0 (SE=5.8)**

Lower Bound : $\bar{x} - (1.96 \times SE) = 62 - (1.96 \times 5.8) = 50.6$

Upper Bound : $\bar{x} + (1.96 \times SE) = 62 + (1.96 \times 5.8) = 73.4$

- **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 JANUARY 25, 2023

Biden Averaged 41% Job Approval in His Second Year

Results for this Gallup poll are based on telephone interviews conducted Jan. 2-22, 2023, with a random sample of 1,011 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: 1,011 respondents (randomly selected)
  - Found $\bar{x}$ = 41
  - SE=2.05

CONFIDENCE INTERVAL

- **Poll: 1,011 respondents (randomly selected)**
  - **Found $\bar{x} = 41$**
  - **SE=2.05**
- **95% CI:**

Lower Bound : $\bar{x} - 1.96 \times SE = 41 - 1.96 \times 2.05 = 37$

Upper Bound : $\bar{x} + 1.96 \times SE = 41 + 1.96 \times 2.05 = 45$

CONFIDENCE INTERVAL

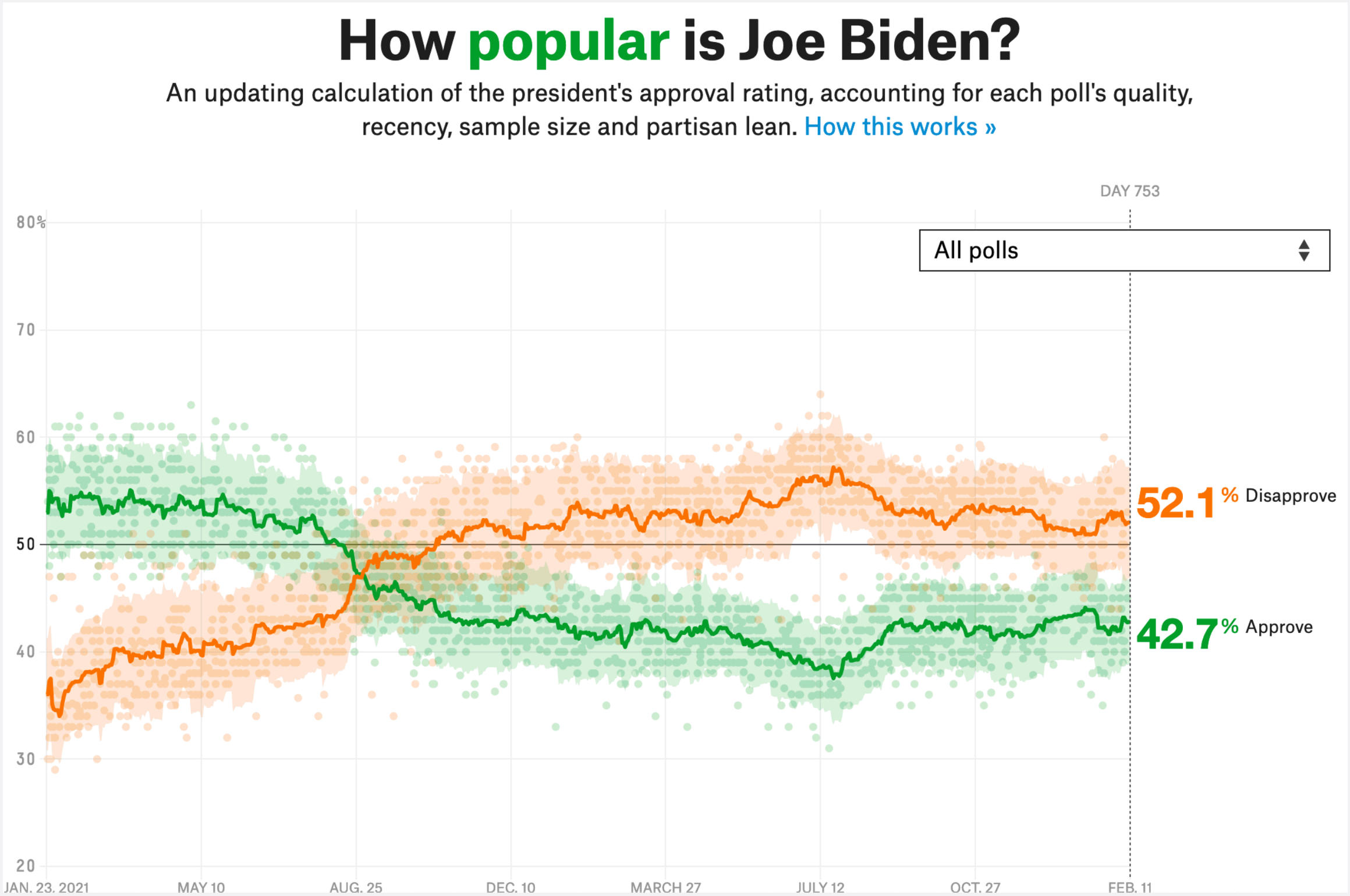
- **So: 95% confidence interval is**

$$41 \pm 4 = (37, 45)$$

Results for this Gallup poll are based on telephone interviews conducted Jan. 2-22, 2023, with a random sample of 1,011 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 true approval rating in population**

REPEATED POLLS



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

RECAP

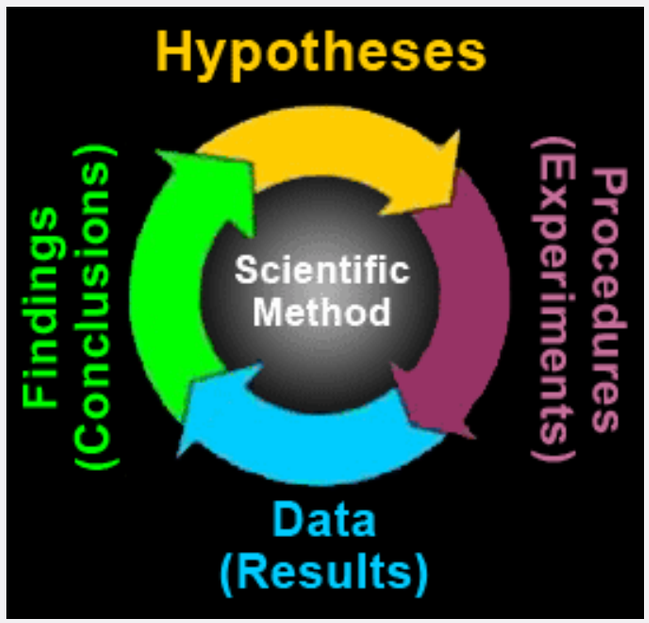
- Core problem: We are interested in population parameters, but usually only have a (random) sample
  - Is what we find in our sample representative of what is going on in the population?
- Answer: We can never be sure that it is, but we can give a *probability* of how sure we are that it is
  - 95%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

TODAY

- Finishing up Sampling and Surveys
- Hypotheses and causality

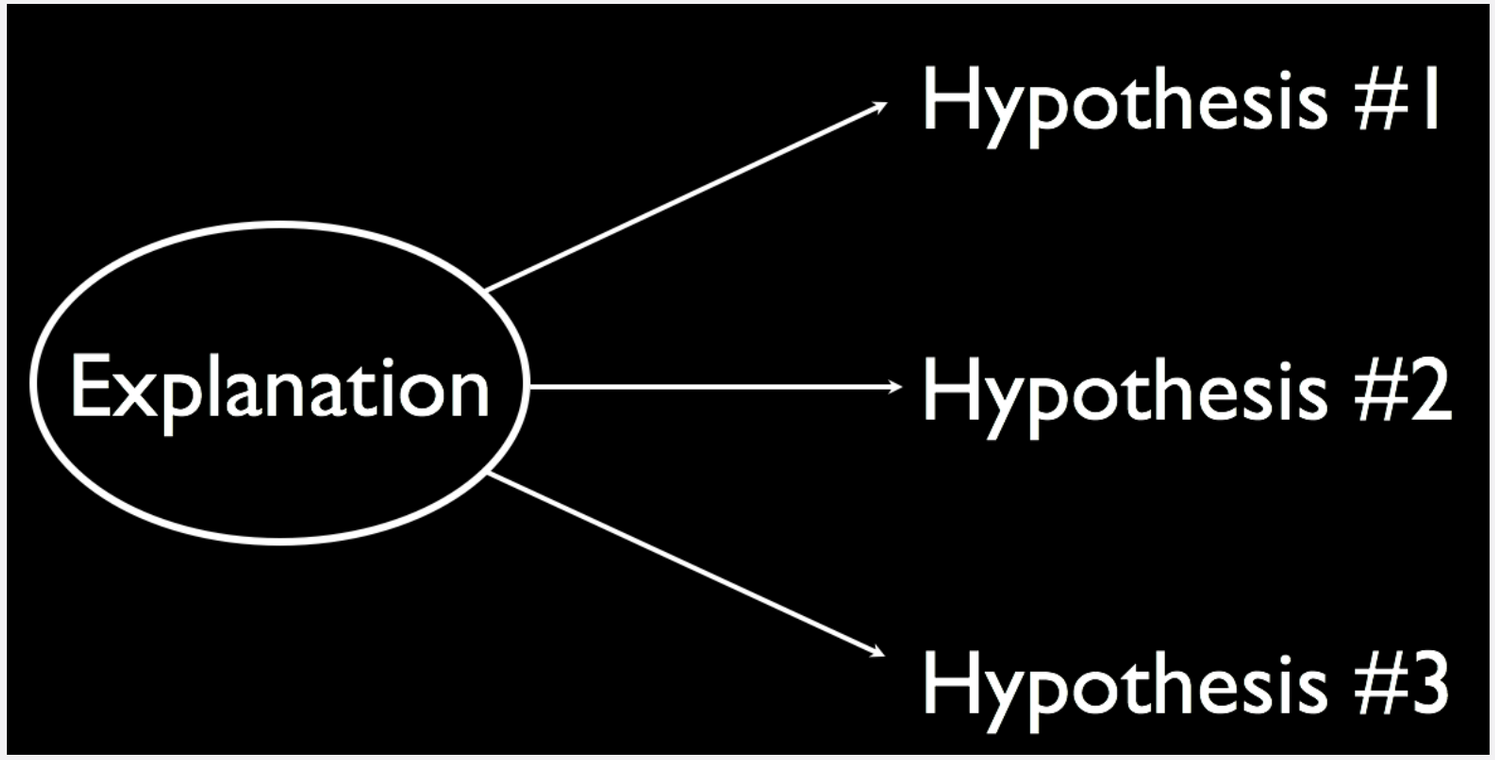
RESEARCH PROCESS

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



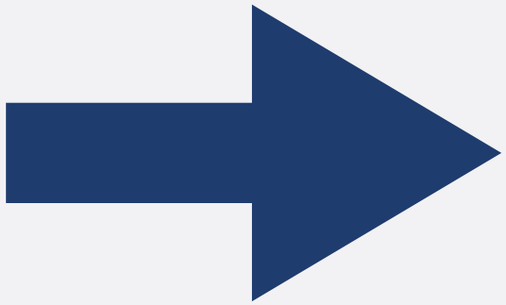
HYPOTHESES AND THEORY

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



THEORY

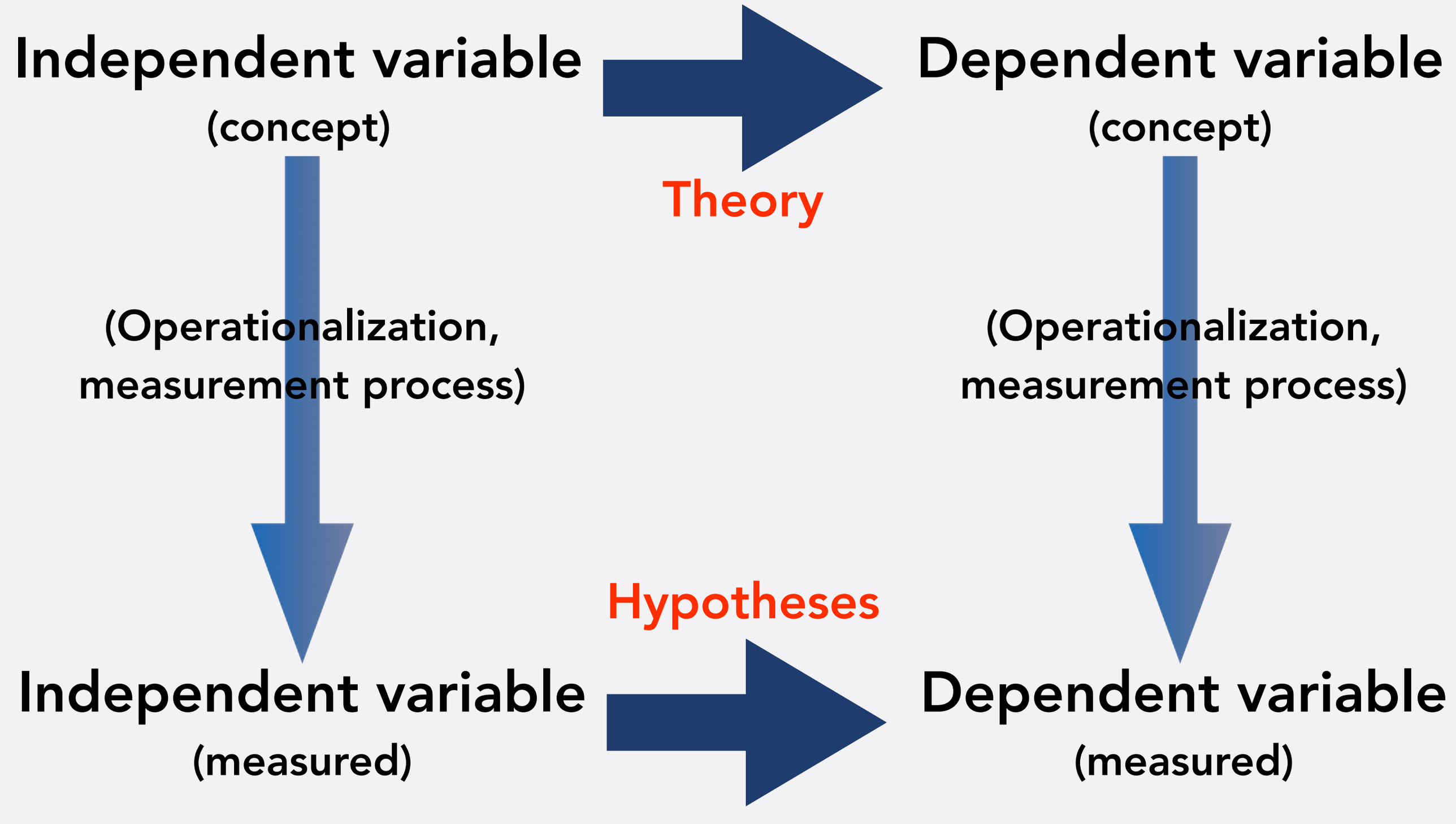
Independent variable
(concept)



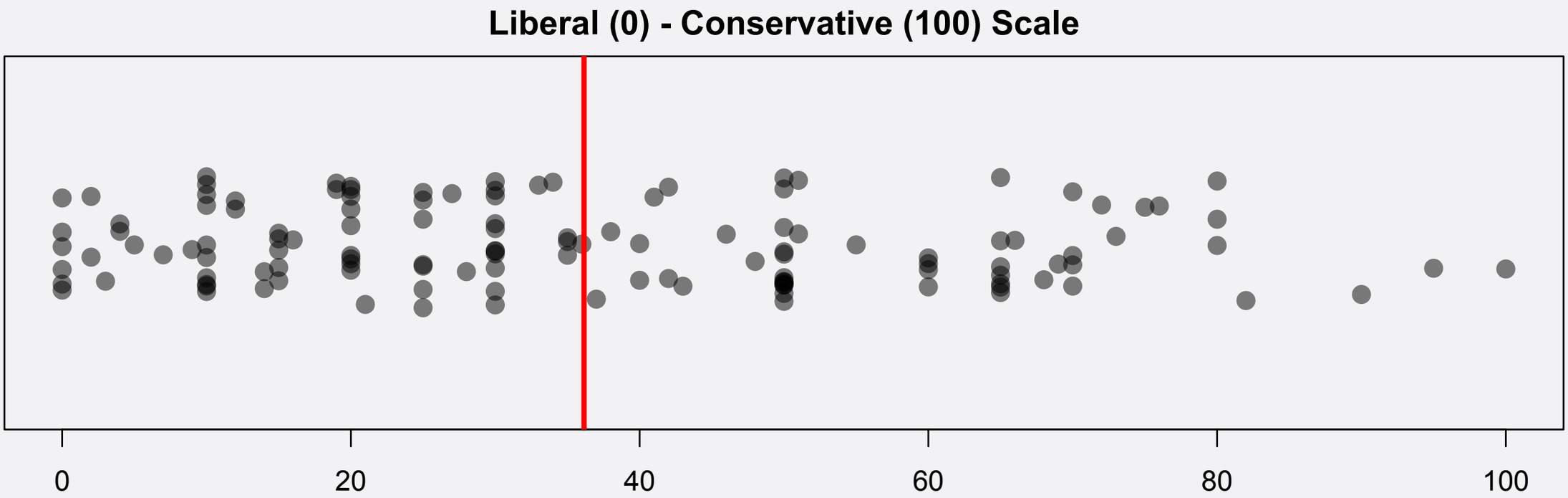
Theory

Dependent variable
(concept)

MEASUREMENT



TODAY'S EXAMPLE



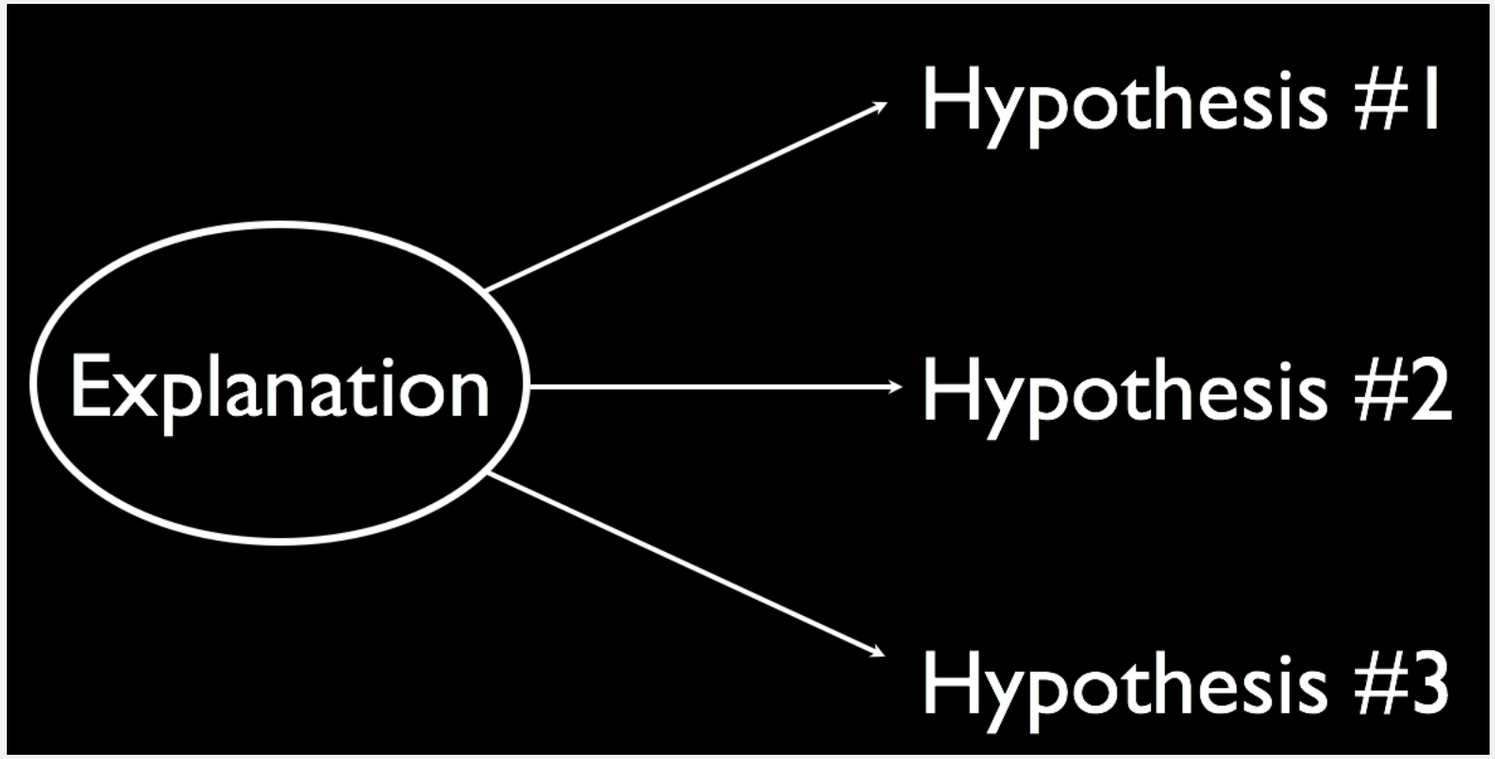
- **Why are some students in 202 liberal and others conservative? What determines which ideology students have?**

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**

GOOD HYPOTHESES

- Involves two variables
  - dependent and independent variable
- Relationship between the variables is clearly specified and measurable
- Unit of analysis is clear
- Hypothesis is testable
  - falsifiable

TEMPLATE

- In a comparison of [units of analysis], those having [one value of the independent variable] will be more likely to have [one value of the dependent variable] than will those having [a different value on the independent variable].

TEMPLATE

- In a comparison of individuals, those having conservative parents will be more likely to be conservative than will those having liberal parents.