PSC 202
SYRACUSE UNIVERSITY

INTRODUCTION TO POLITICAL ANALYSIS

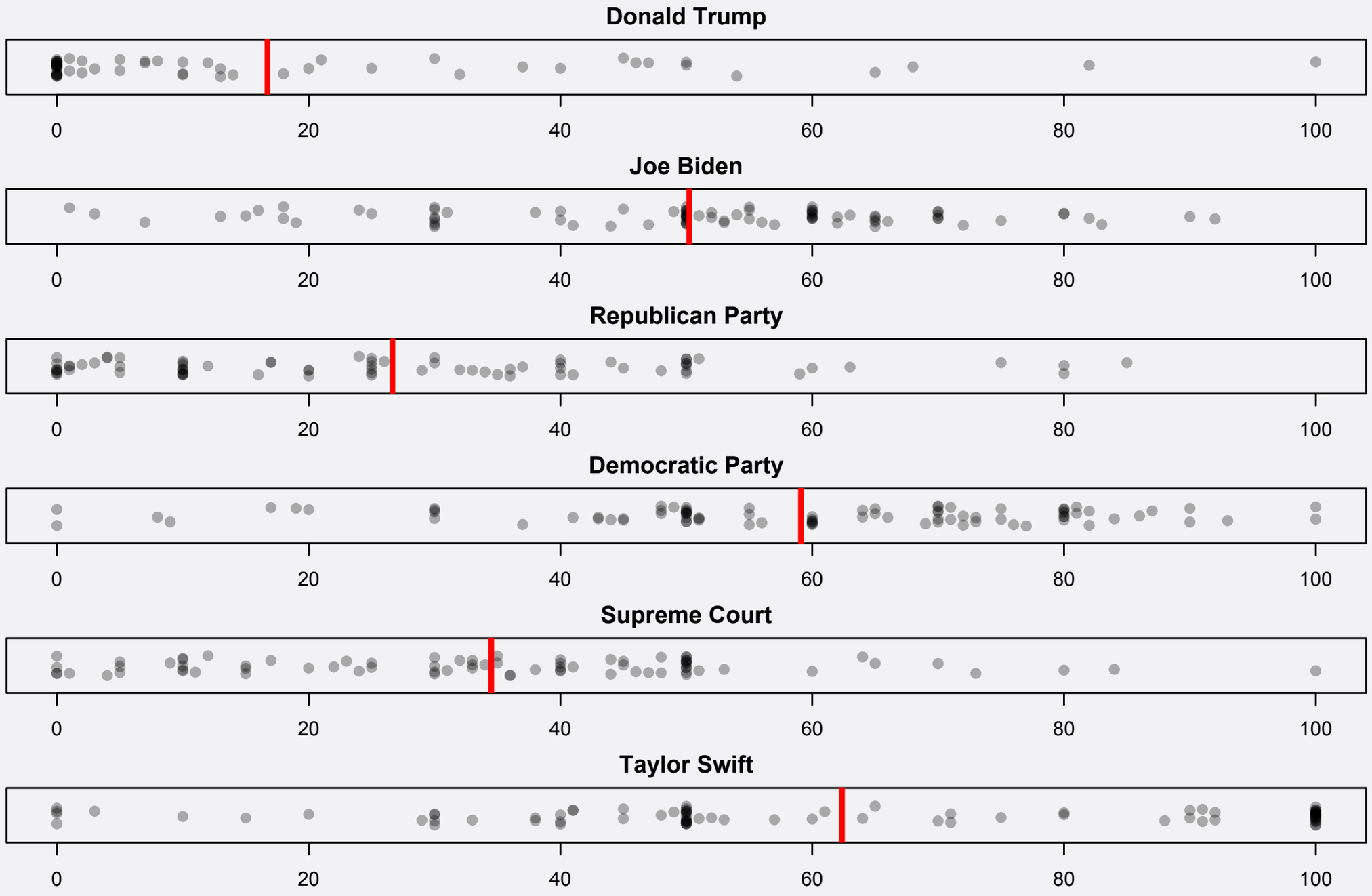
MORE SAMPLING AND SURVEYS, HYPOTHESES AND CAUSALITY

HOUSEKEEPING

- Sections on Friday
- Problem set 4 will be posted tomorrow
- 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 Dobbs v. Jackson decision, which overturned the right to have an abortion (by overturning Roe v. Wade)?**

	Number	Percentage
Strongly agree	0	0.0%
Somewhat agree	7	8.0%
Neither agree nor disagree	4	4.5%
Somewhat disagree	5	5.7%
Strongly disagree	72	81.8%

SURVEY RESULTS

- **In your view, should immigration into the United States be kept at its present level, increased, or decreased?**

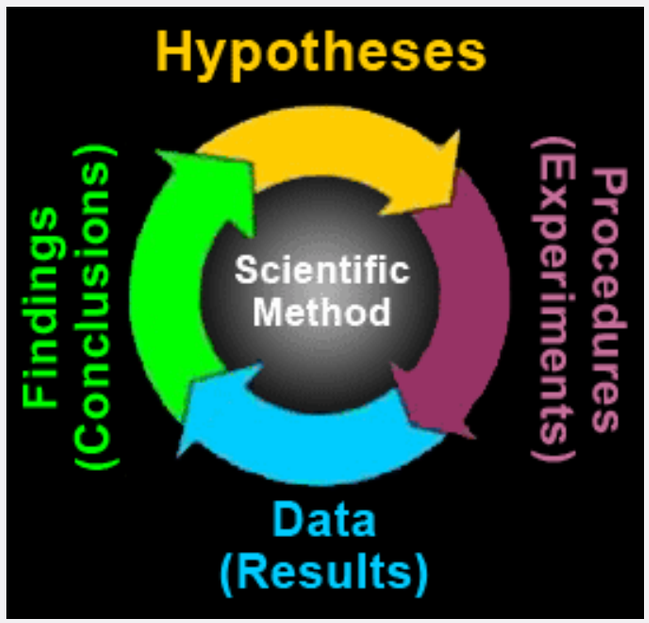
	Number	Percentage
Increased	32	36.4%
Present Level	45	51.1%
Decreased	11	12.5%

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



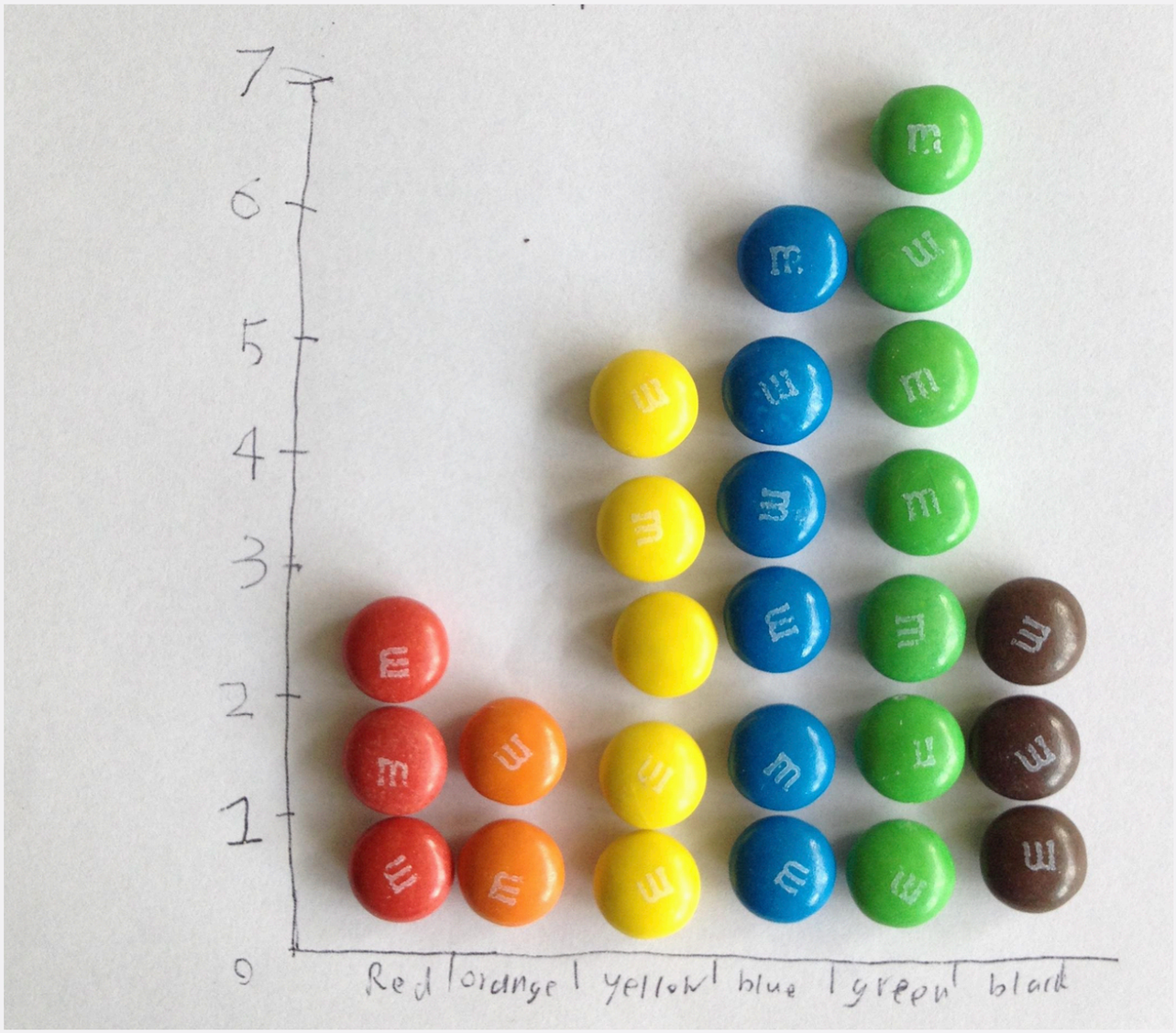
RECAP

- Population: the entire universe of objects to which our hypothesis applies
- Sample: the subset of the population that we study in order to make inferences about the full population

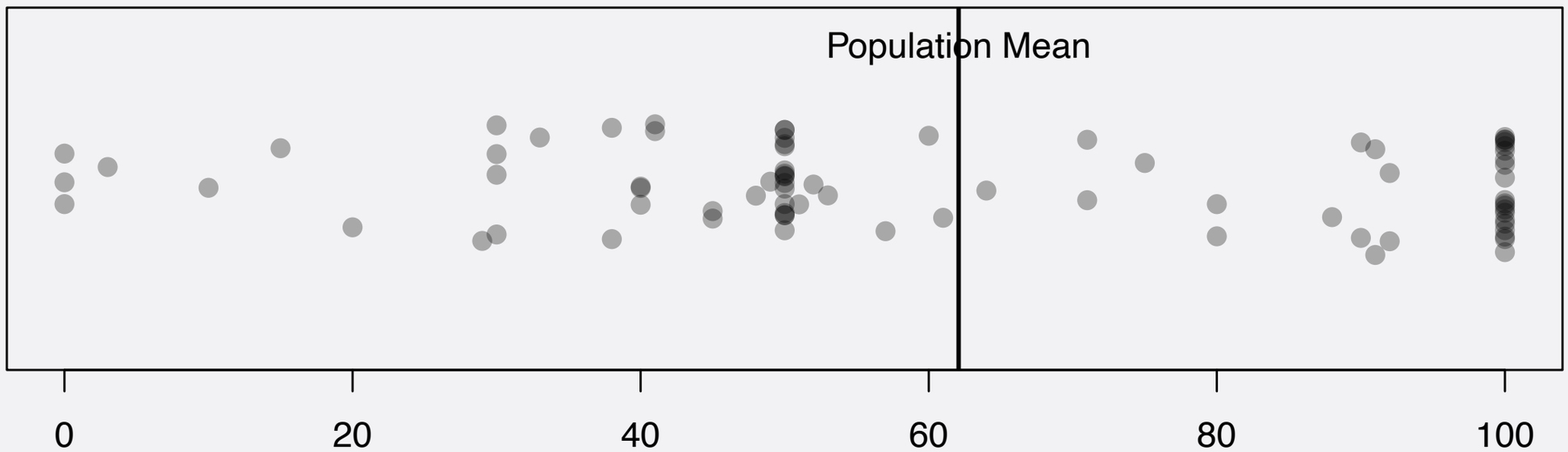
- We often use random sample of the population to learn about the population

RANDOM SAMPLING ERROR

- But: random sampling introduces *random* sampling error

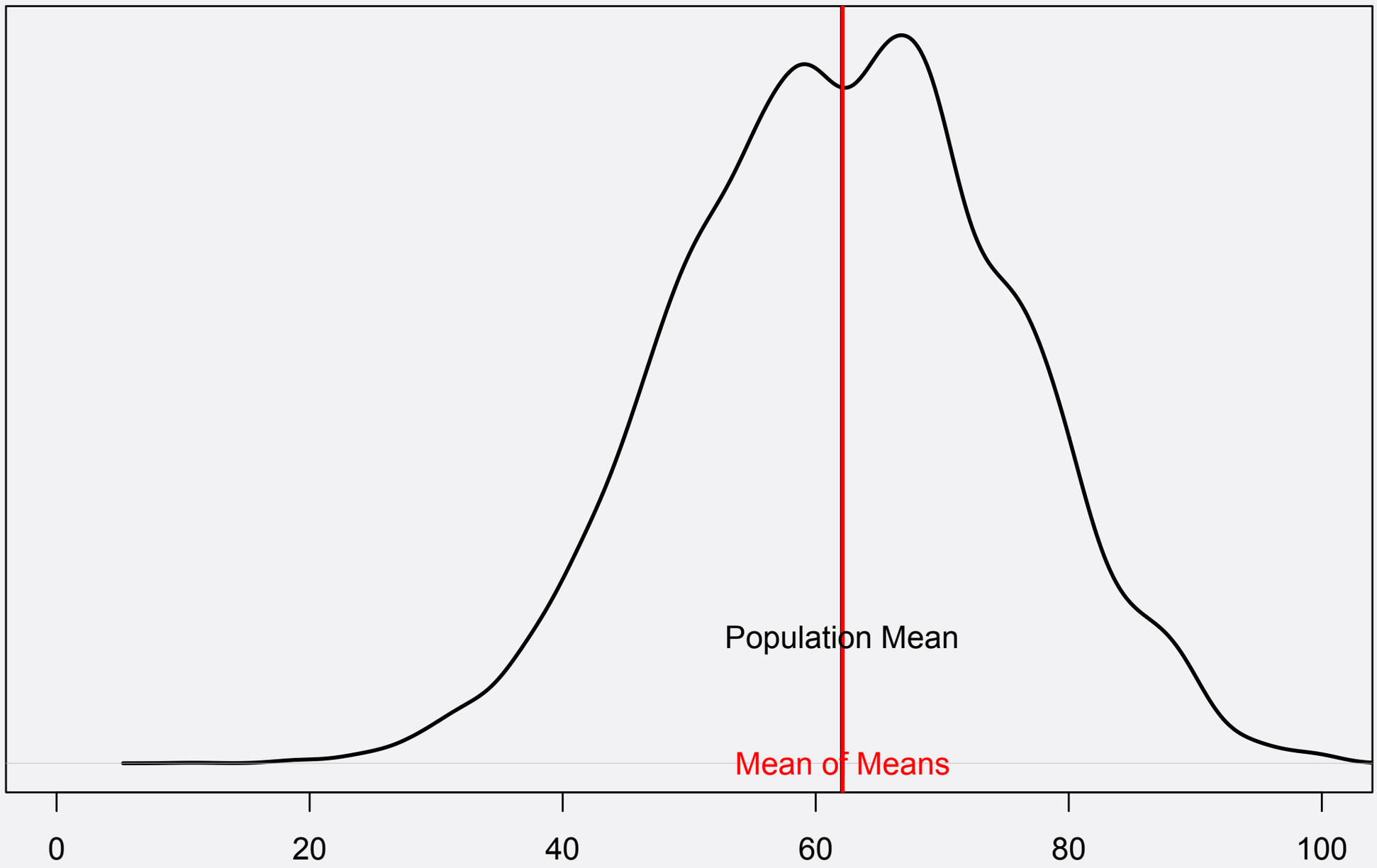


OUR SURVEY



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

AFTER 10,000 RANDOM SAMPLES

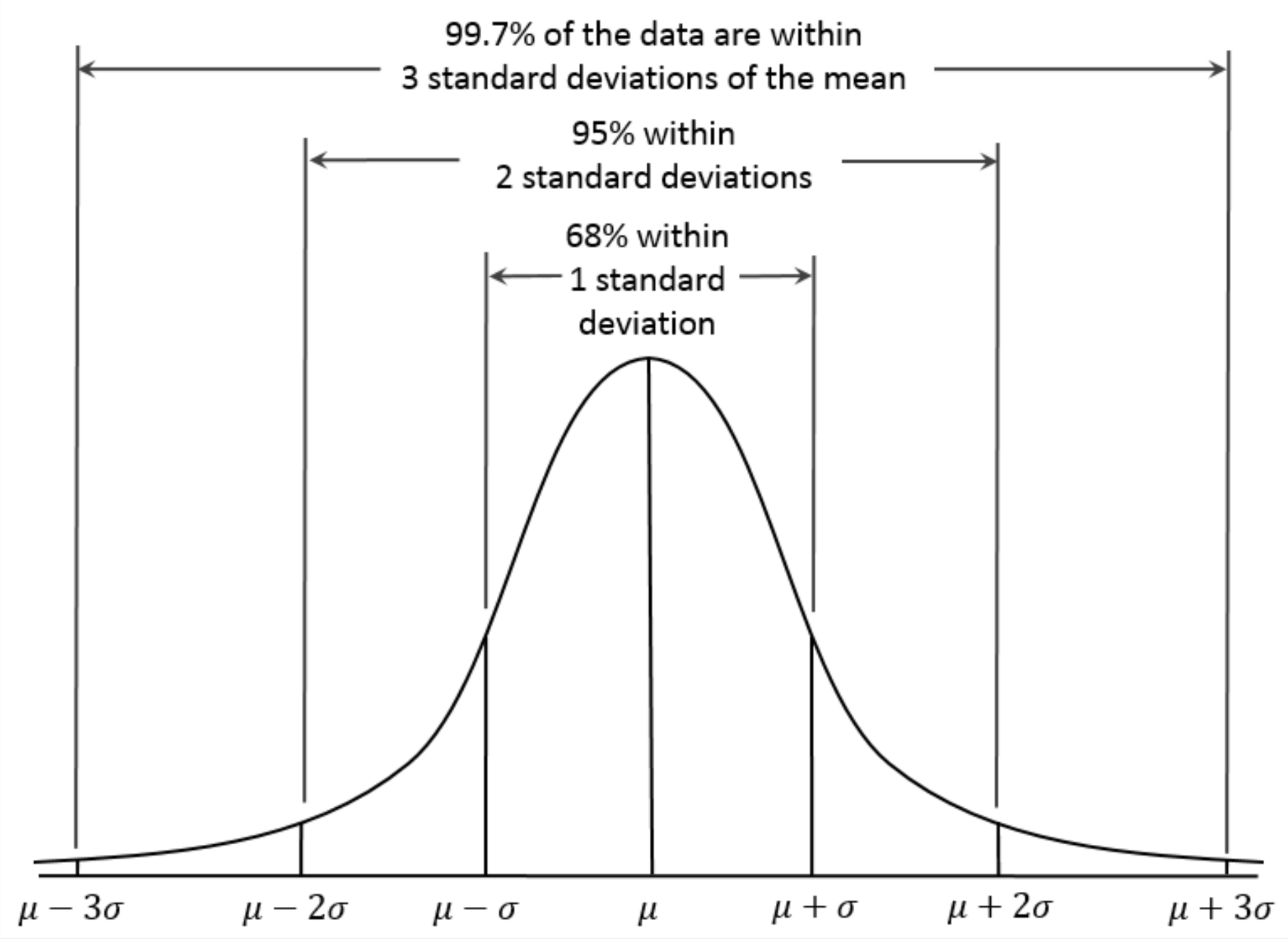


CENTRAL LIMIT THEOREM

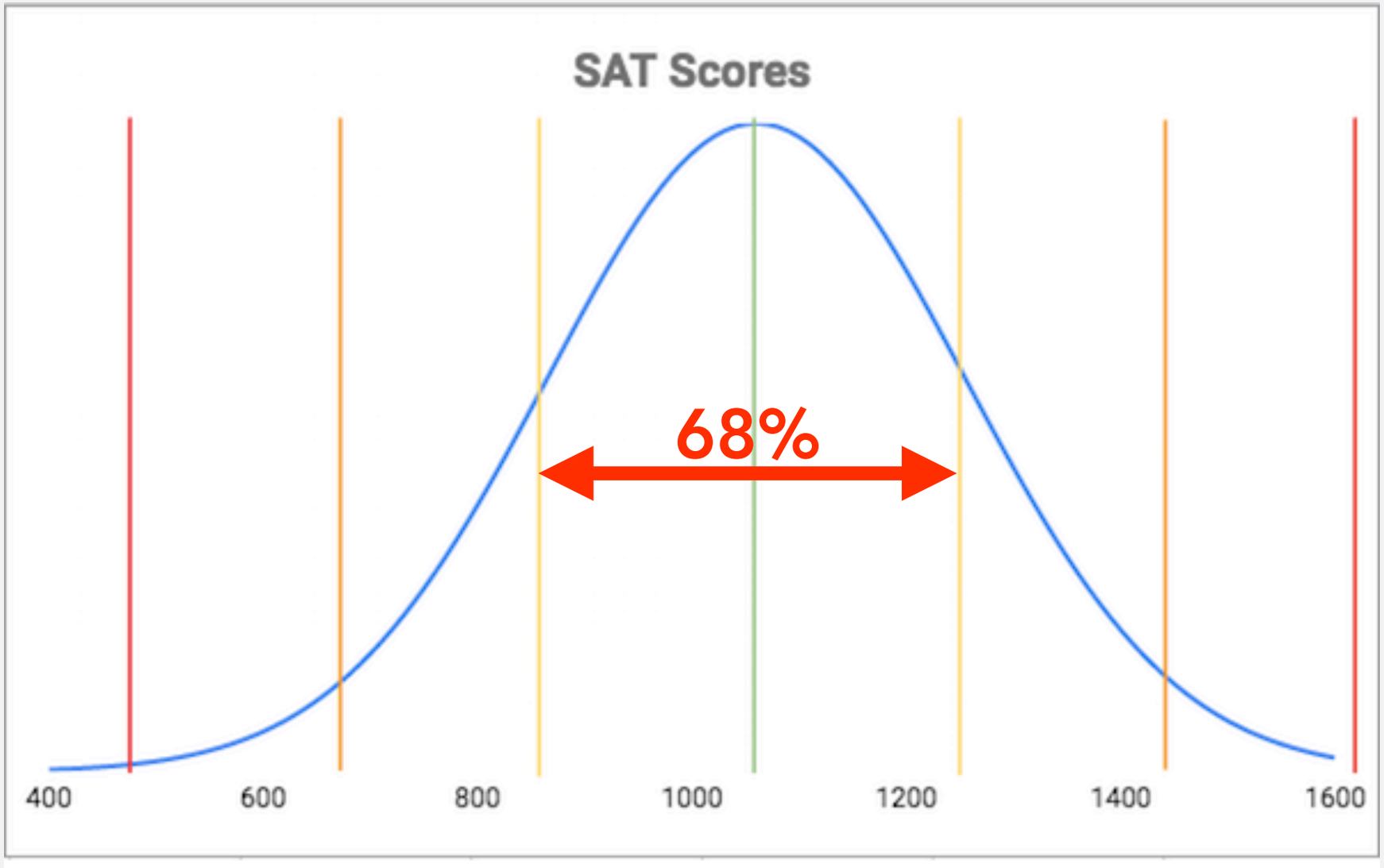
- If we take many random samples from a population and compute the sample mean for each sample…
  - The mean of the sample means will be equal to the population mean
    - No *systematic* sampling error
- The sample means will have a Normal distribution
  - *Random* sampling error follows a Normal distribution
  - Things that follow a Normal distribution behave in very predictable ways

NORMAL DISTRIBUTION

- Important characteristic of *all* Normal distributions:

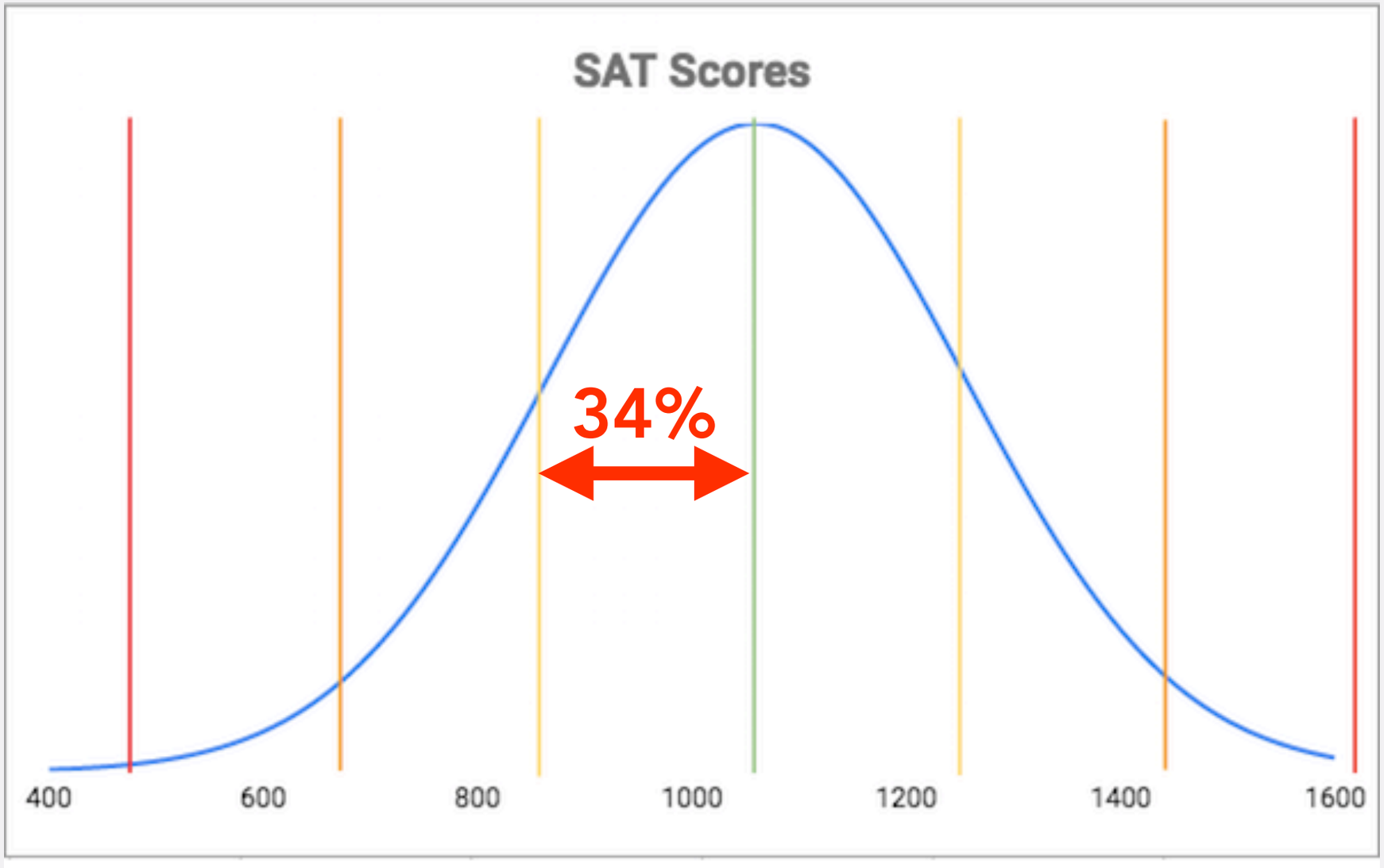


NORMAL DISTRIBUTION



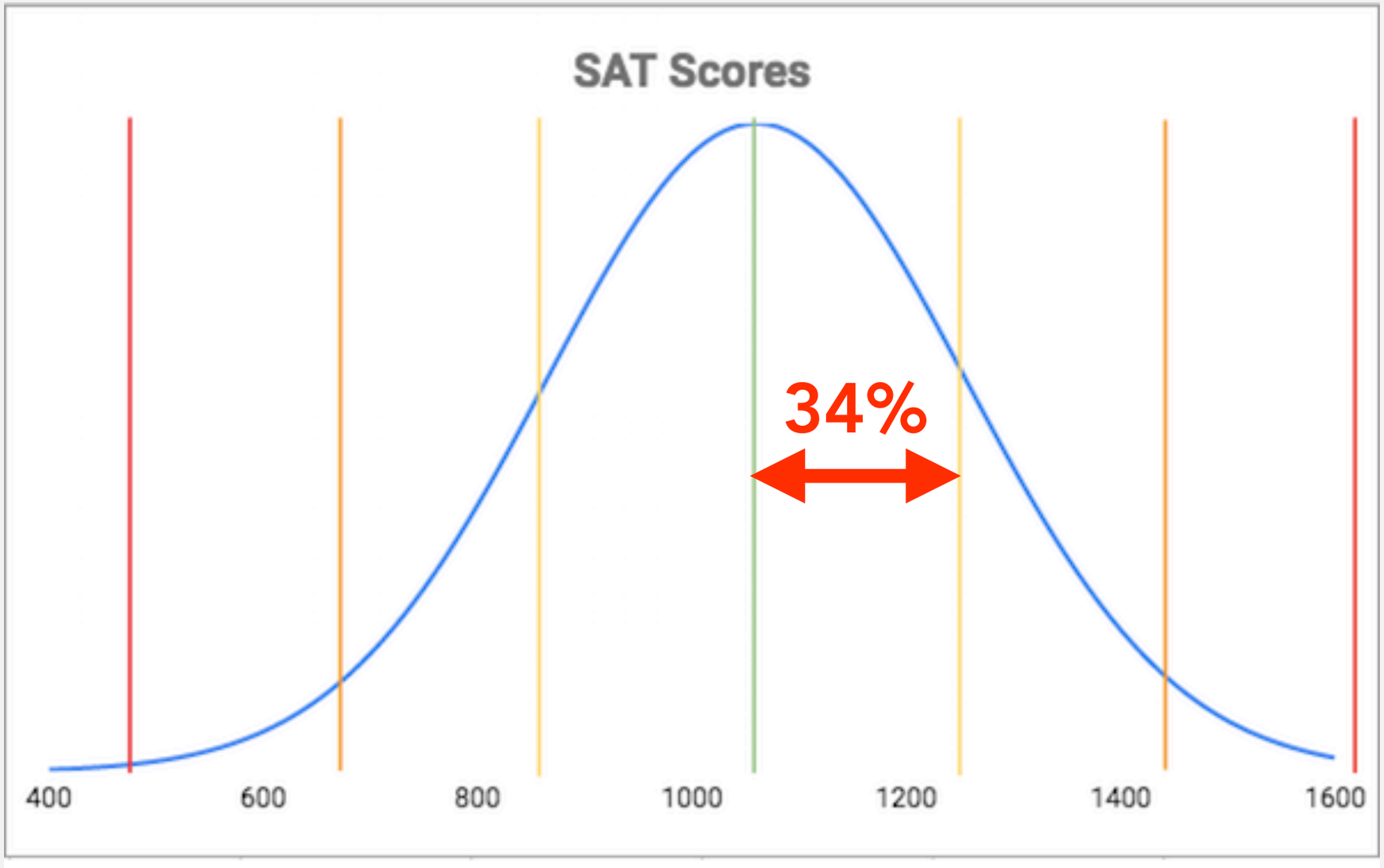
- Mean: 1060; Standard deviation: 195
- So: 68% of SAT scores between 1060-195 and 1060+195 (865-1255)

NORMAL DISTRIBUTION



- Mean: 1060; Standard deviation: 195
- So: 34% of SAT scores between 1060-195 and 1060 (865-1060)

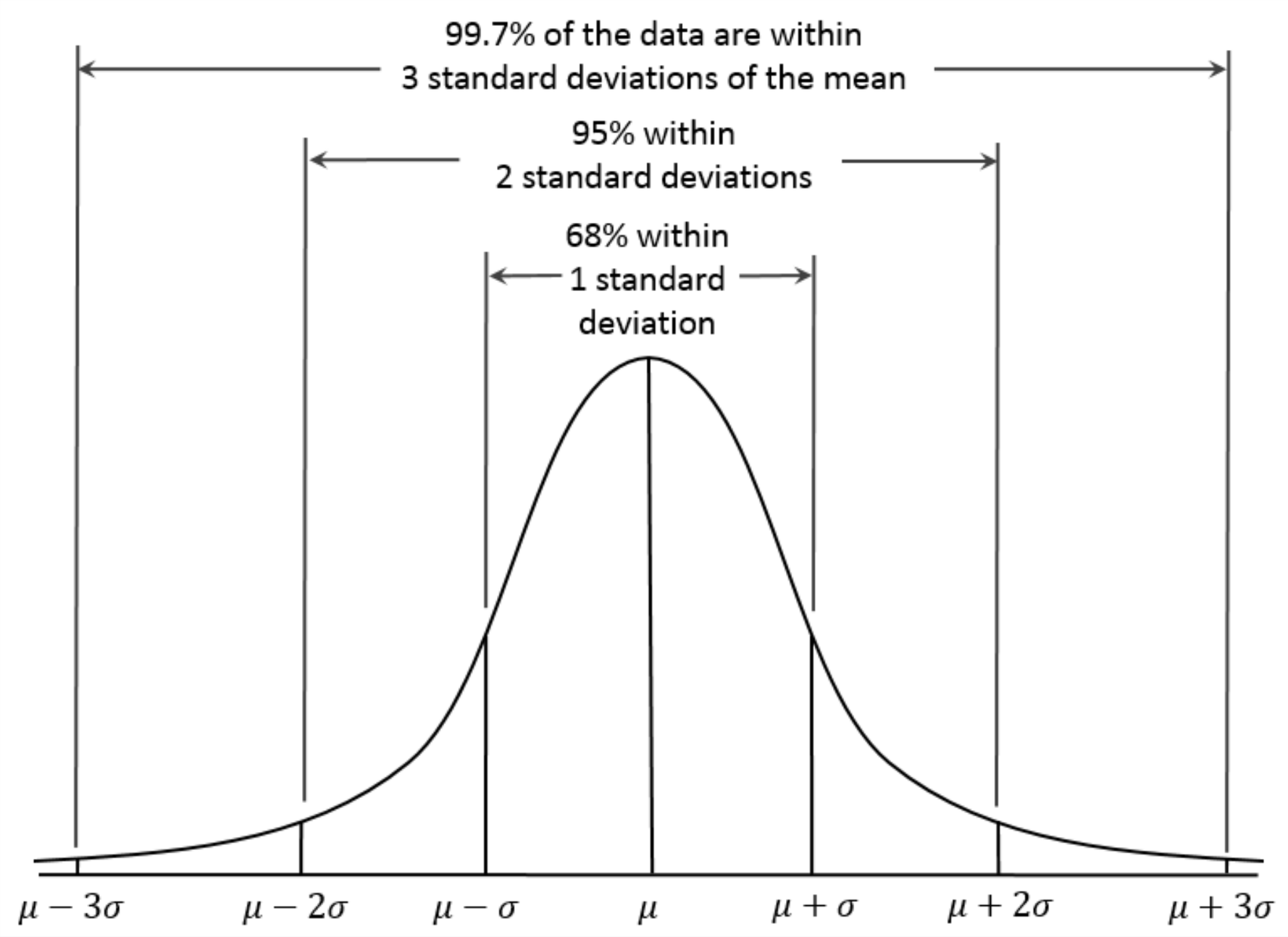
NORMAL DISTRIBUTION



- Mean: 1060; Standard deviation: 195
- So: 34% of SAT scores between 1060 and 1060+195 (1060-1255)

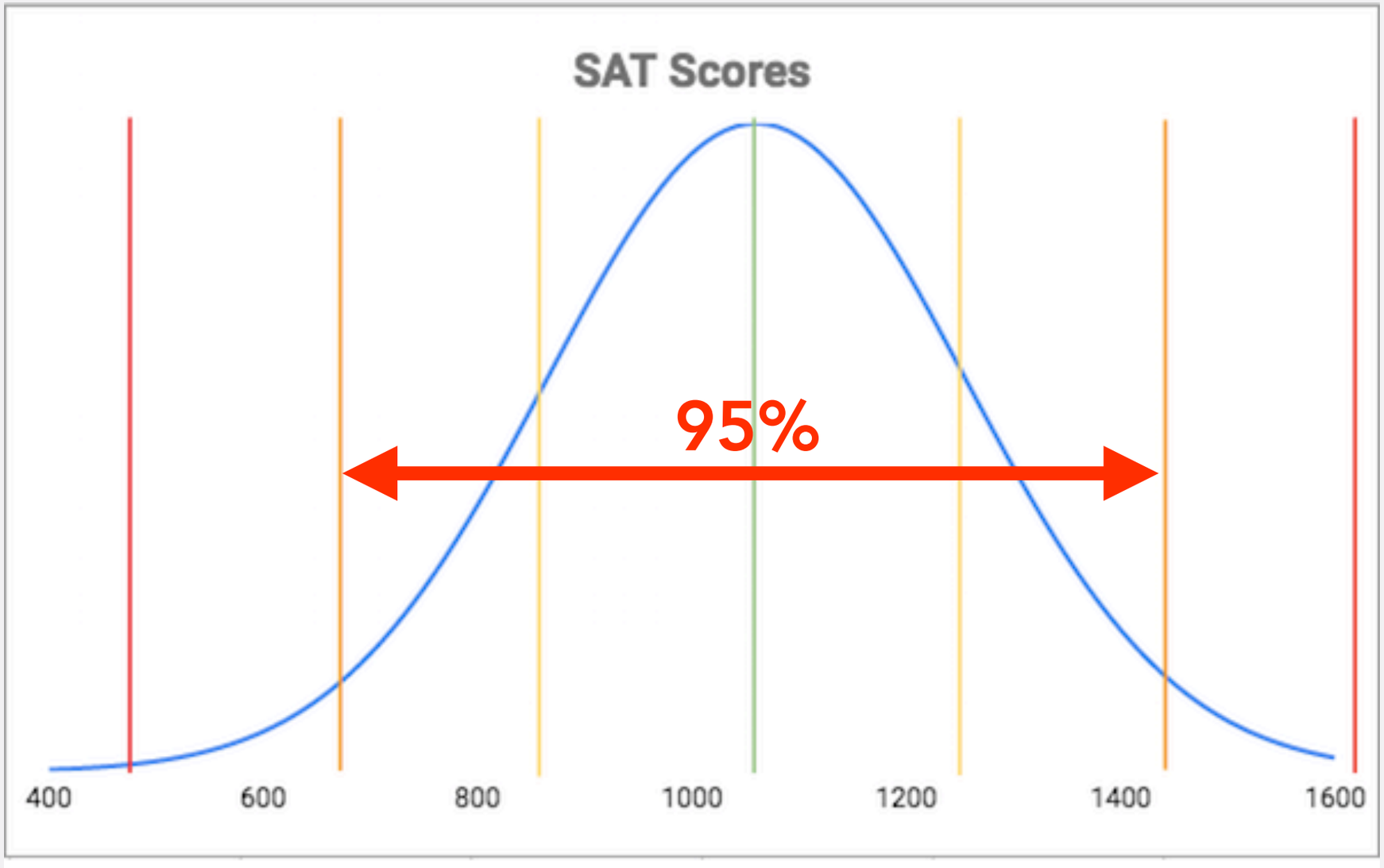
NORMAL DISTRIBUTION

- Important characteristic of *all* Normal distributions:



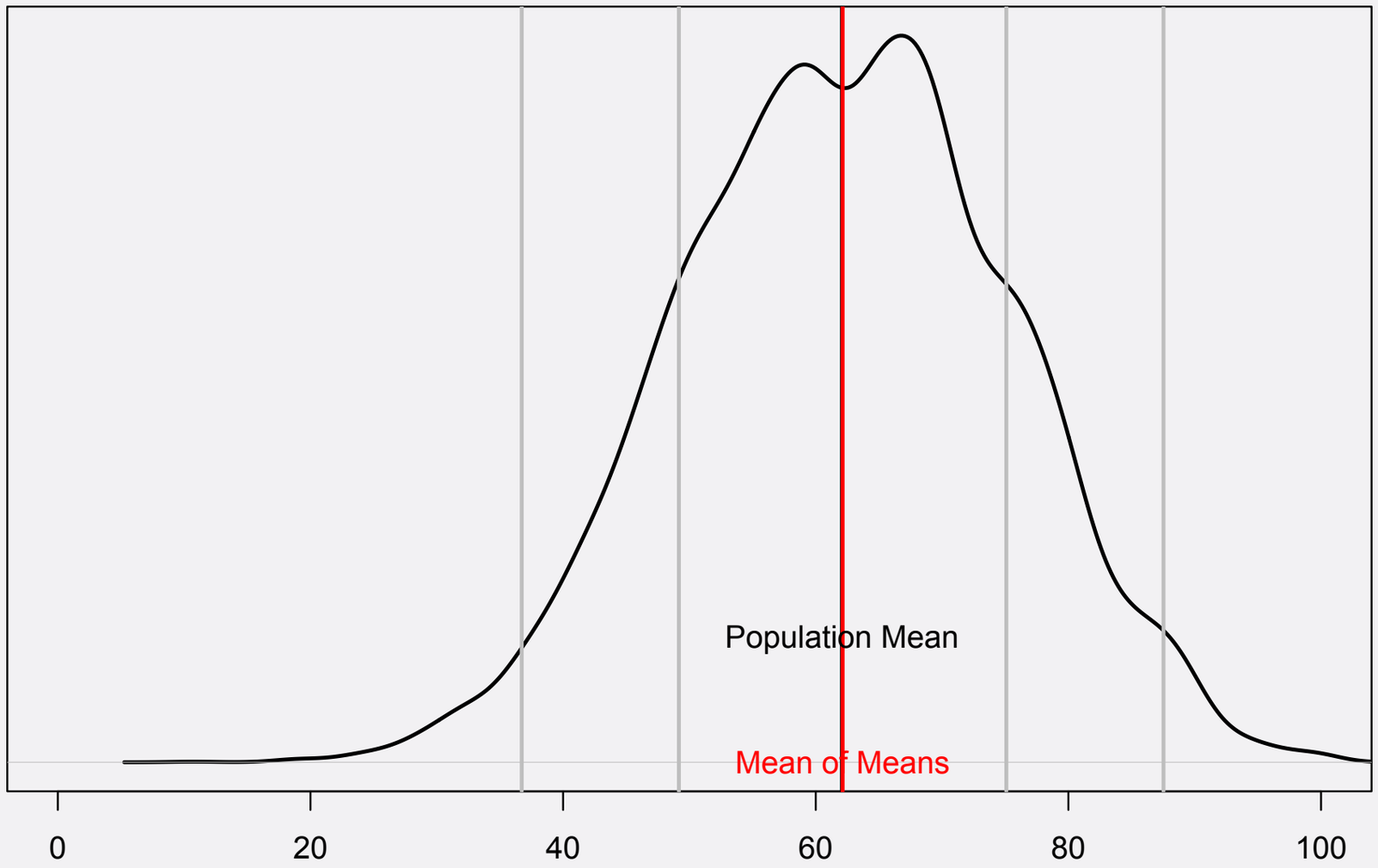
- Actually: 95% within 1.96 standard deviations

NORMAL DISTRIBUTION



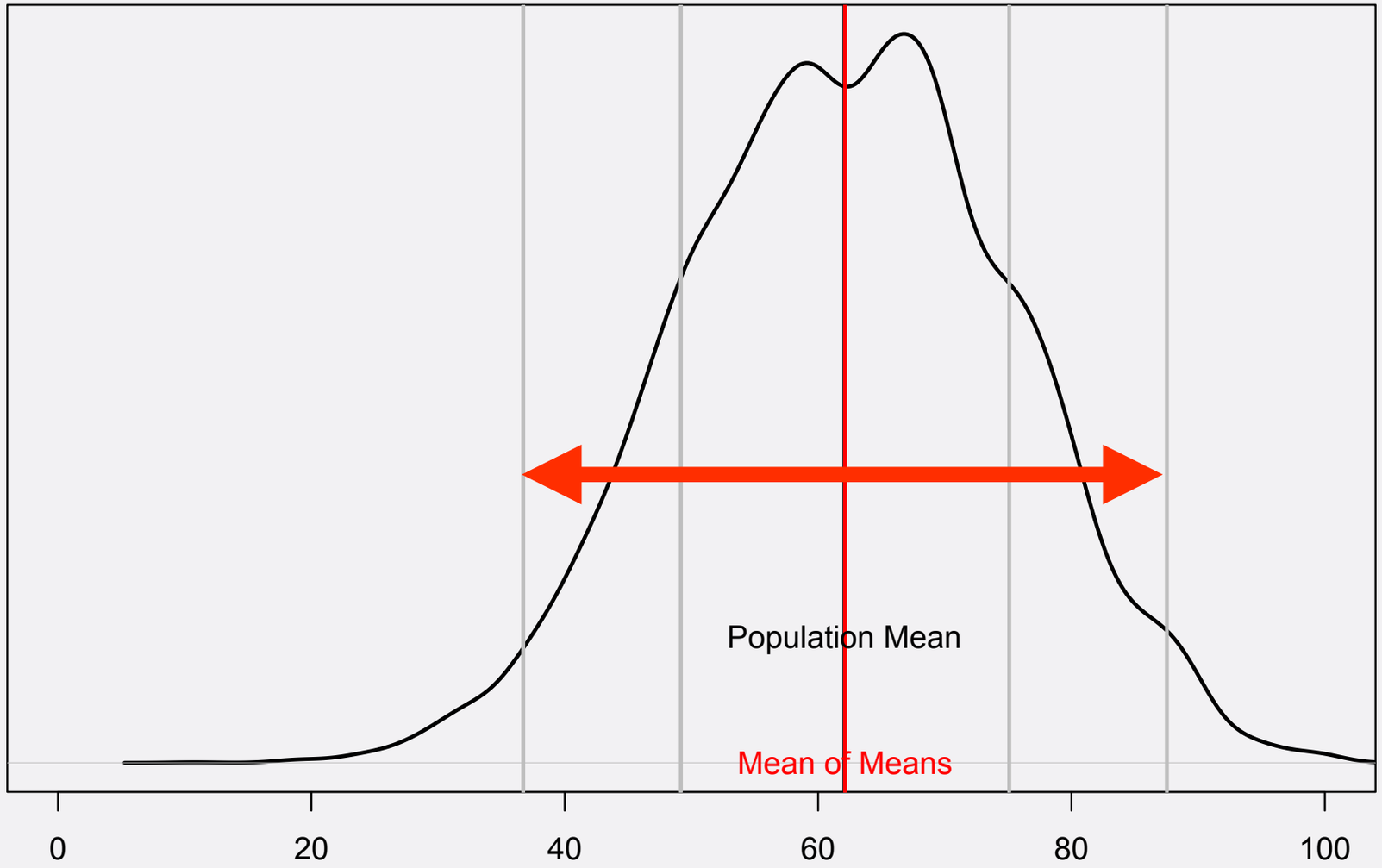
- Mean: 1060; Standard deviation: 195
- So: 95% of SAT scores between 1060-(1.96*195) and 1060+(1.96*195) [678-1442]

WHAT DOES THIS HAVE TO DO WITH RANDOM SAMPLES?



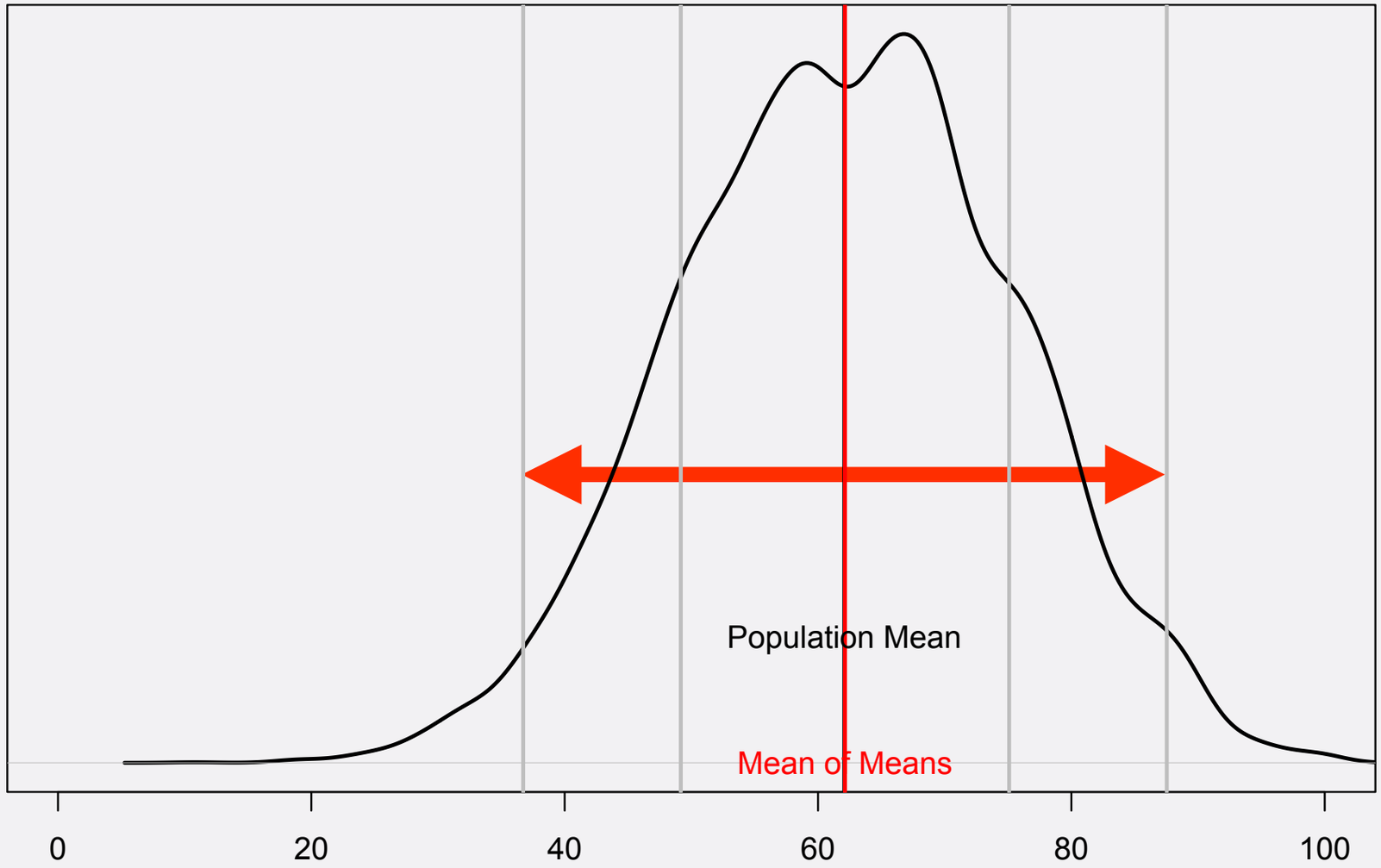
- If we take many random samples (=polls) from a population
- and compute mean for each
- those means have a Normal distribution

RANDOM SAMPLING ERROR



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

RANDOM SAMPLING ERROR



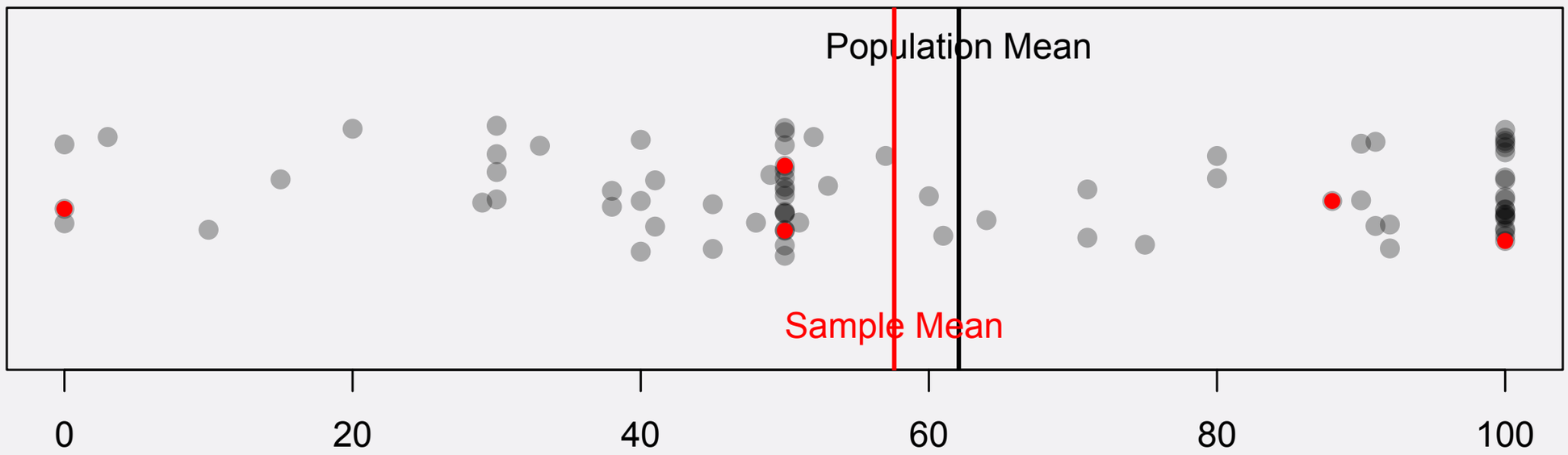
- What is the standard deviation?
- If we know, we can tell how large the random sampling error is

STANDARD ERROR

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

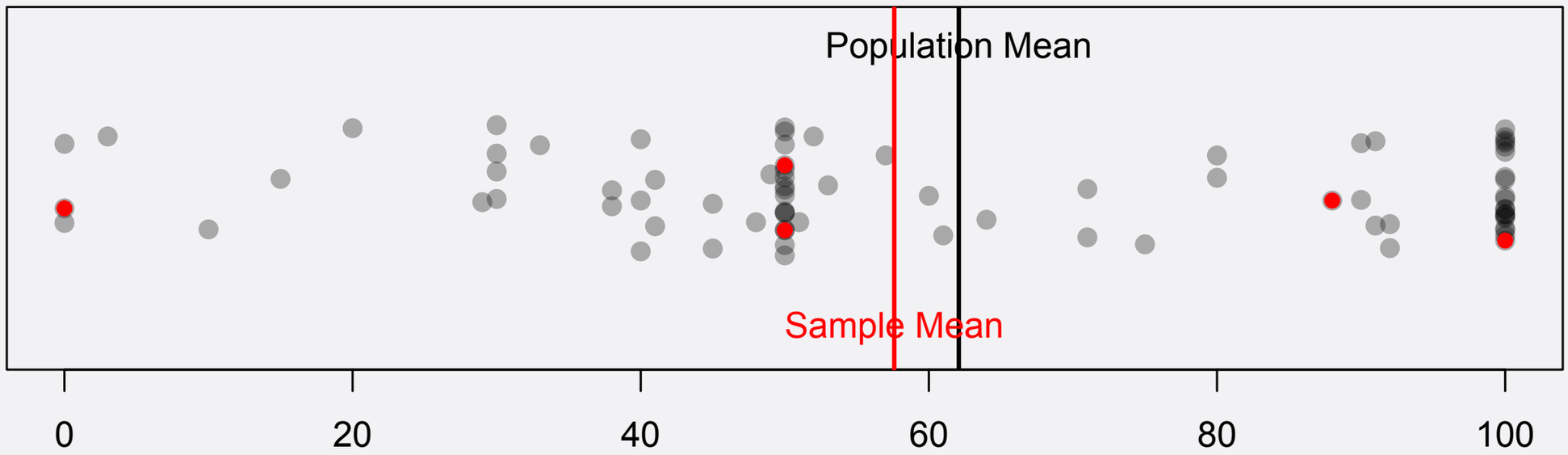
- s: standard deviation in our random sample
- n: size of our sample
- The larger SE, the larger the random sampling error

OUR SURVEY



- Random sample:
  - $\bar{x}$=57.6 ($\bar{x}$ means sample mean)
  - s=17.5 (s means sample standard deviation)

OUR SURVEY



- **Random sample:**
  - **$\bar{x}$=57.6 ($\bar{x}$ means sample mean)**
  - **s=17.5 (s means sample standard deviation)**

$$\text{SE} = \frac{s}{\sqrt{n}} = \frac{17.5}{\sqrt{5}} = 7.8$$

STANDARD ERROR

- Let's say we had 50 respondents (randomly selected) instead

$$\text{SE} = \frac{s}{\sqrt{n}} = \frac{17.5}{\sqrt{50}} = 2.5$$

STANDARD ERROR

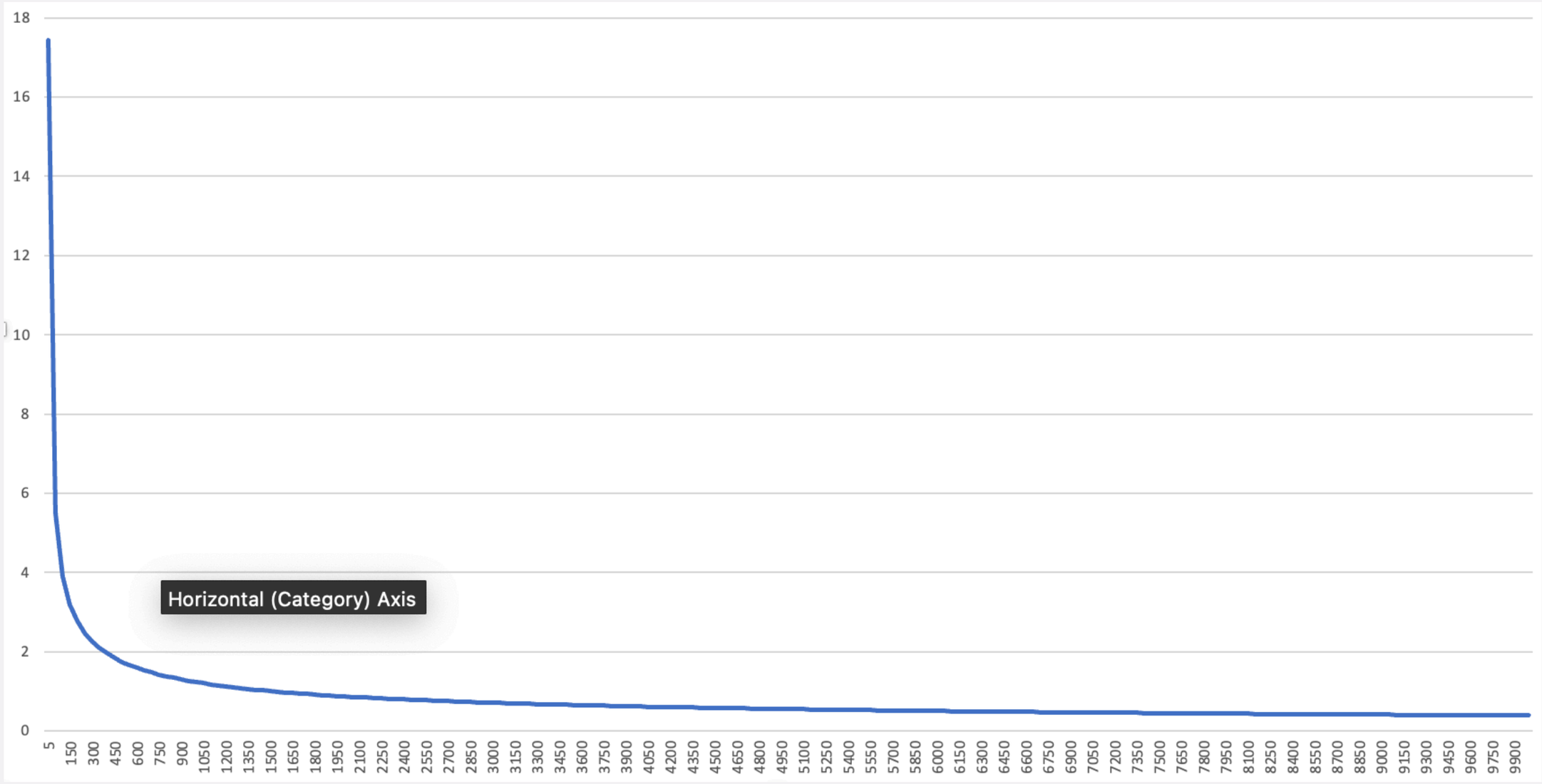
- Let's say we had 500 respondents (randomly selected) instead

$$\text{SE} = \frac{s}{\sqrt{n}} = \frac{17.5}{\sqrt{500}} = 0.8$$

SAMPLE SIZE

- The larger the sample, the smaller the standard error of the sample mean
- Larger survey = less random measurement error
- But: Doubling of sample size does NOT cut standard error in half

SAMPLE SIZE



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

SUMMARY: STANDARD ERROR

- Ok, the standard error tells us how much random sampling error there is
- What can we do with that?

HOW DOES THIS HELP?

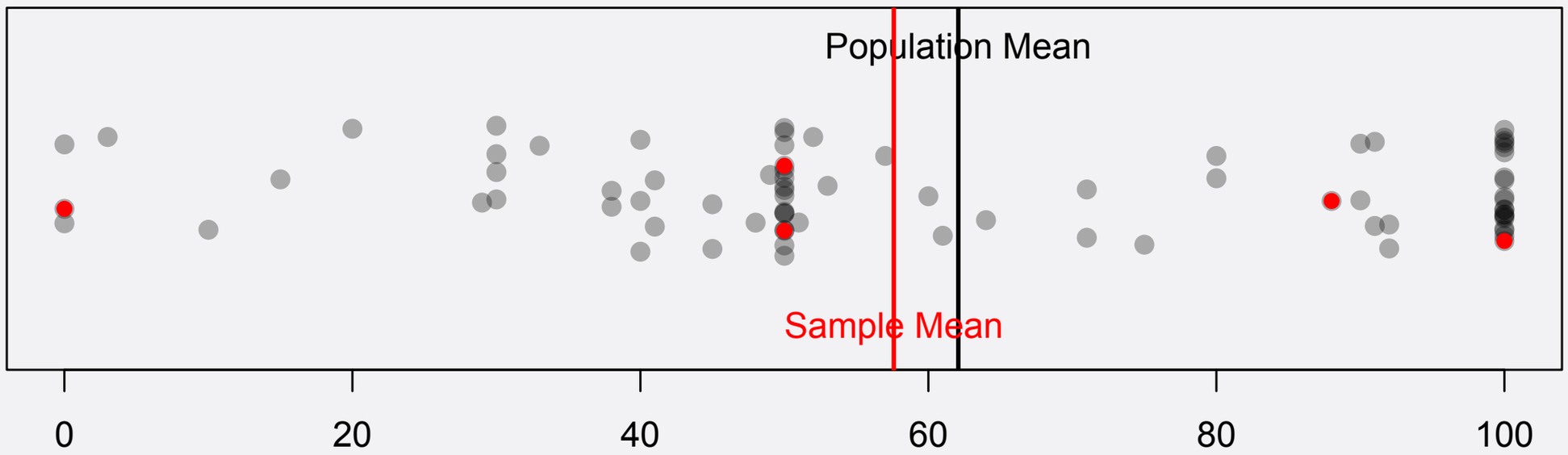
- For random samples, we can provide an interval that likely contains the true population mean

CONFIDENCE INTERVAL

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

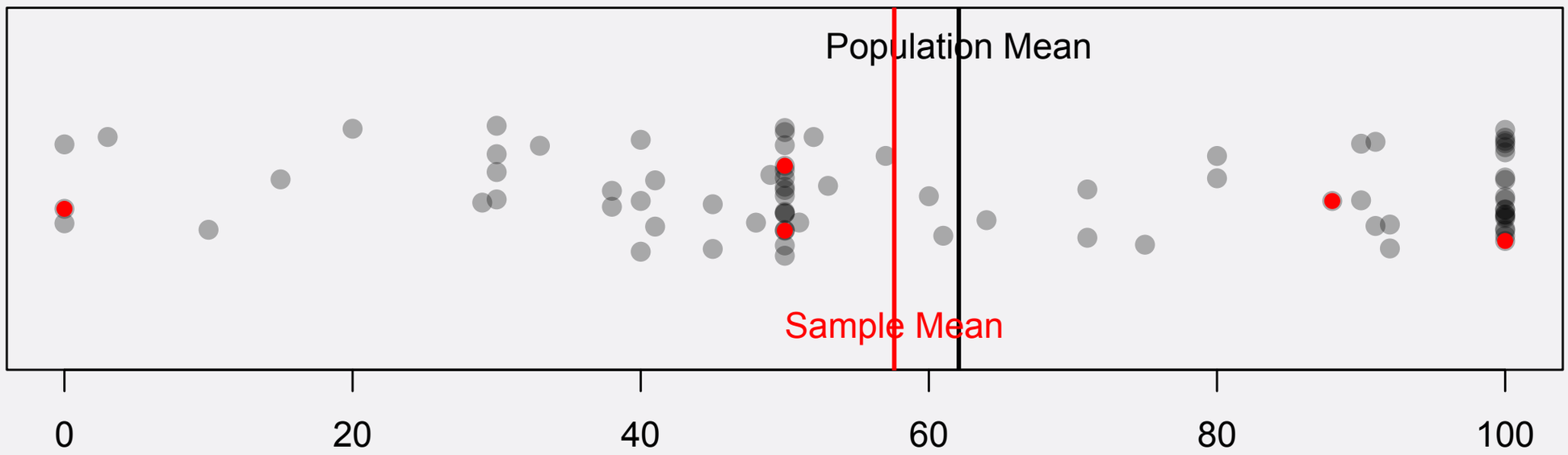
$$95\%\ \text{CI} = \bar{x} \pm (1.96 \times \text{SE})$$

OUR SURVEY



- Random sample of 5 students
- Sample mean $\bar{x}$=57.6 (SE=17.5)

OUR SURVEY

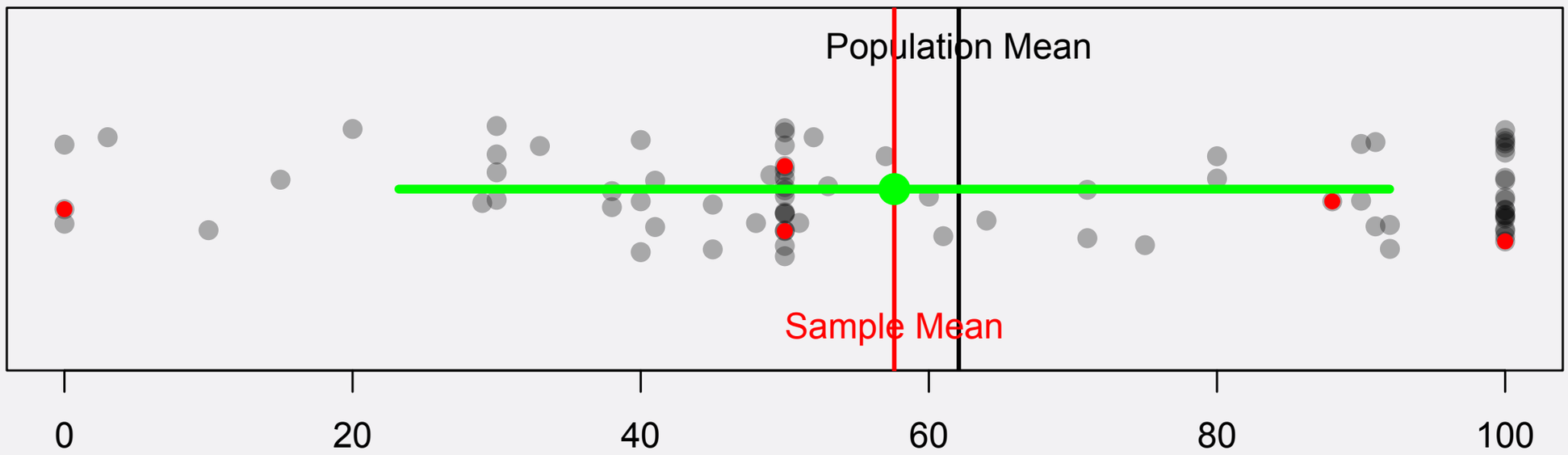


- **Random sample of 5 students**
- **Sample mean $\bar{x}$=57.6 (SE=17.5)**

$$\bar{x} - (1.96 \times SE) = 57.6 - (1.96 \times 17.5) = 23.3$$

$$\bar{x} + (1.96 \times SE) = 57.6 + (1.96 \times 17.5) = 91.9$$

OUR SURVEY



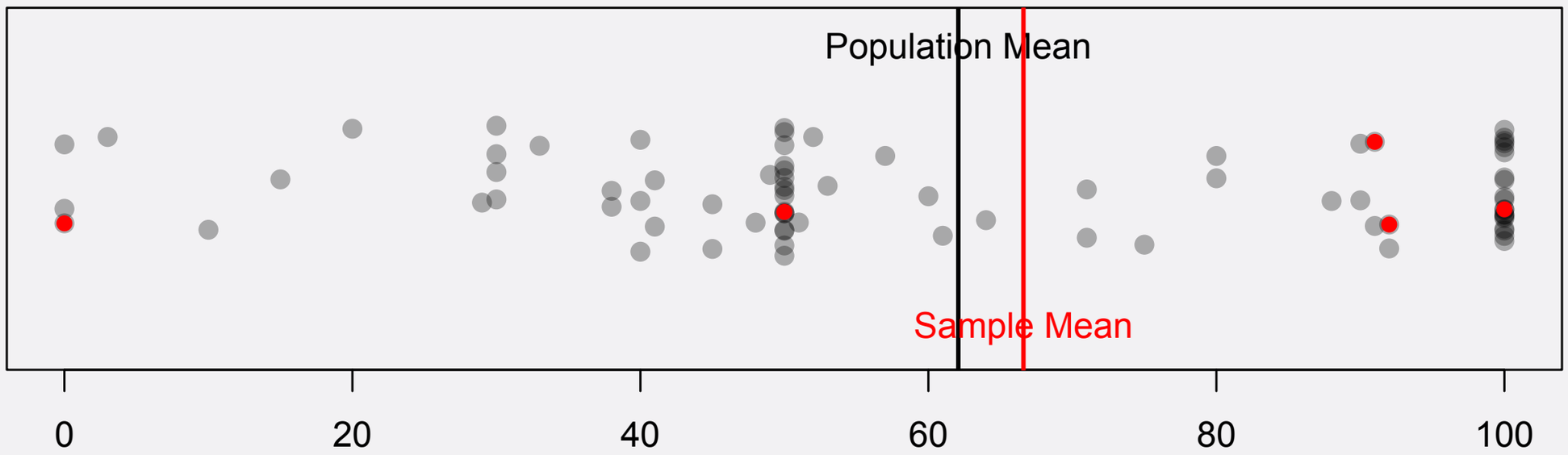
- **Random sample of 5 students**
- **Sample mean $\bar{x}$=57.6 (SE=17.5)**

$$\bar{x} - (1.96 \times SE) = 57.6 - (1.96 \times 17.5) = 23.3$$

$$\bar{x} + (1.96 \times SE) = 57.6 + (1.96 \times 17.5) = 91.9$$

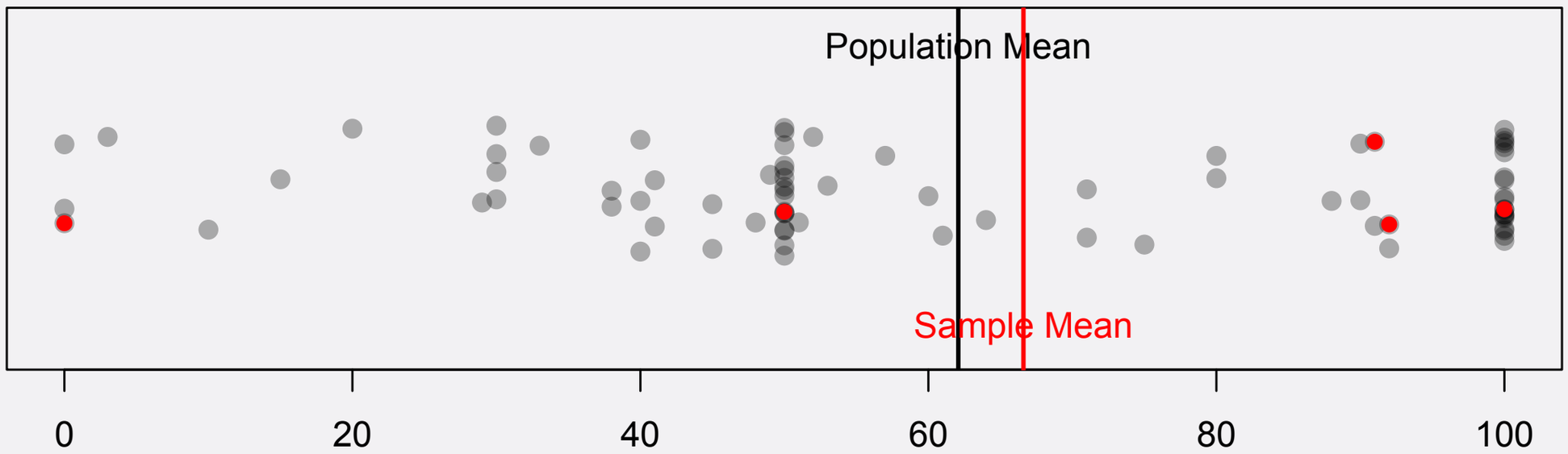
- **95% CI *contains* true population mean**

OUR SURVEY



- Another random sample of 5 students
- Sample mean $\bar{x}$=66.6 (SE=18.8)

OUR SURVEY

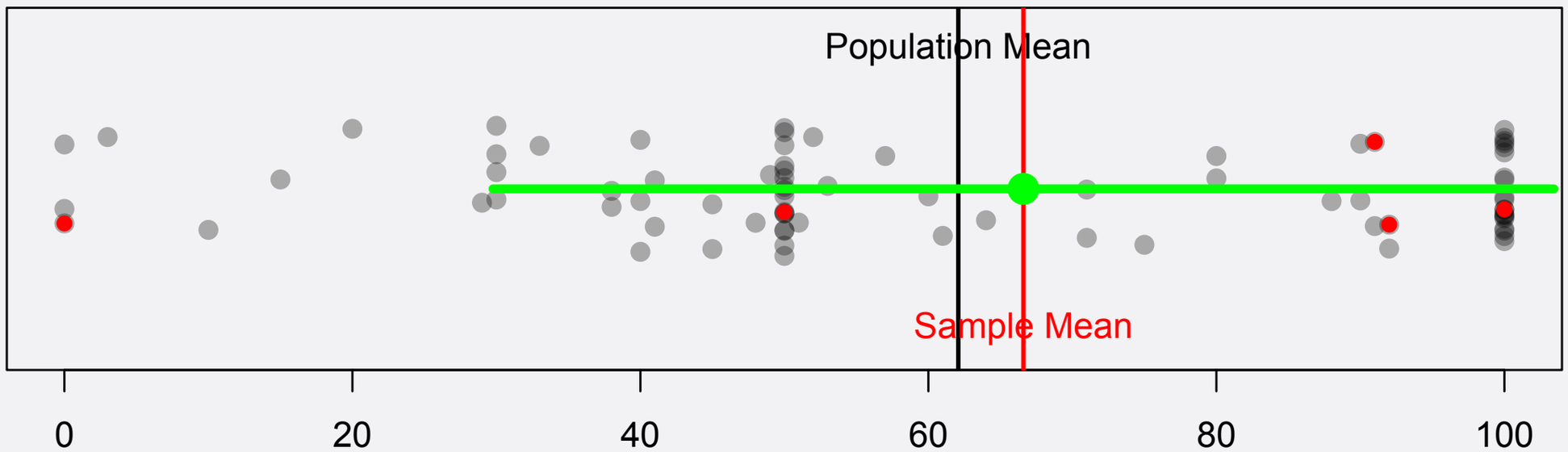


- **Another random sample of 5 students**
- **Sample mean $\bar{x}$=66.6 (SE=18.8)**

$$\bar{x} - (1.96 \times SE) = 66.6 - (1.96 \times 18.8) = 29.8$$

$$\bar{x} + (1.96 \times SE) = 66.6 + (1.96 \times 18.8) = 103.4$$

OUR SURVEY



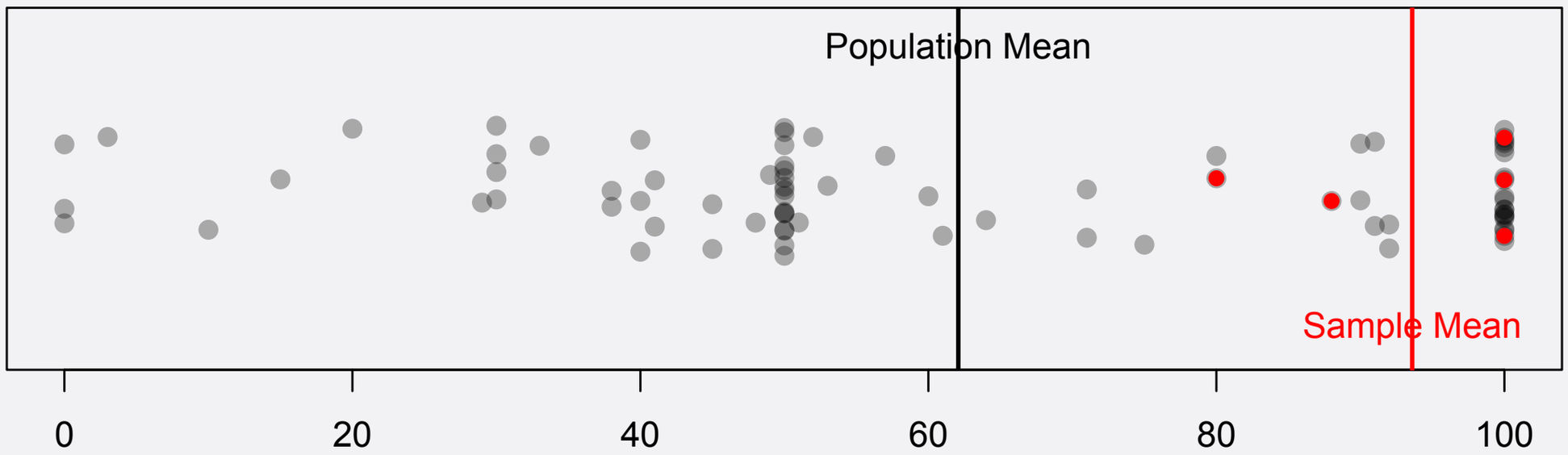
- Another random sample of 5 students
- Sample mean $\bar{x}$=66.6 (SE=18.8)

$$\bar{x} - (1.96 \times SE) = 66.6 - (1.96 \times 18.8) = 29.8$$

$$\bar{x} + (1.96 \times SE) = 66.6 + (1.96 \times 18.8) = 103.4$$

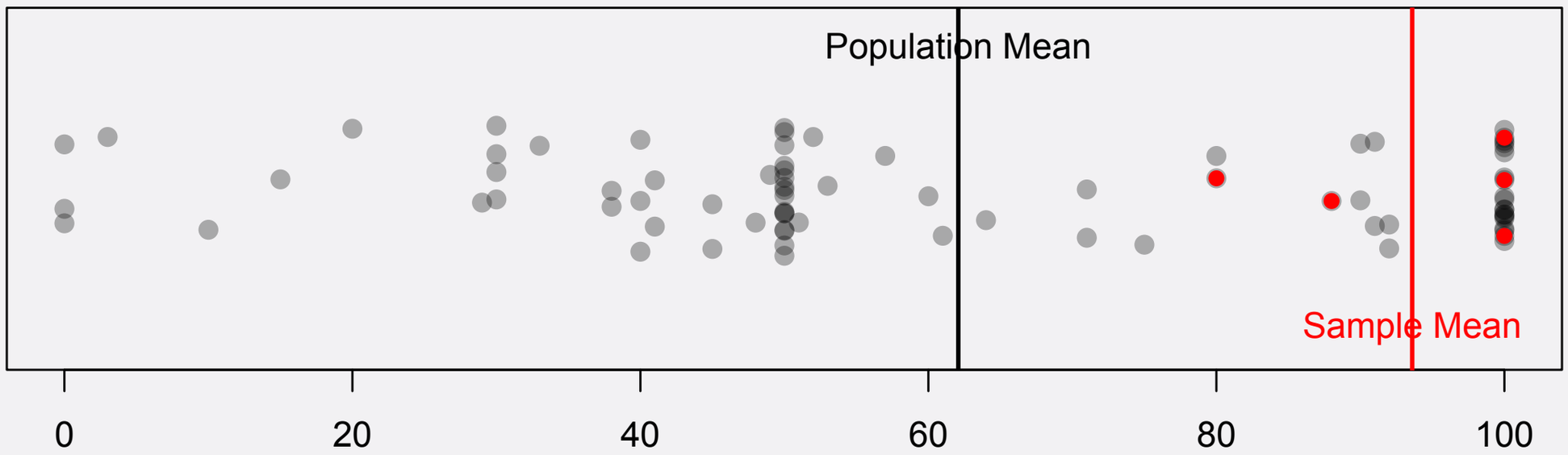
- 95% CI *contains* true population mean

OUR SURVEY



- Yet another random sample of 5 students
- Sample mean $\bar{x}$=93.6 (SE=4.1)

OUR SURVEY

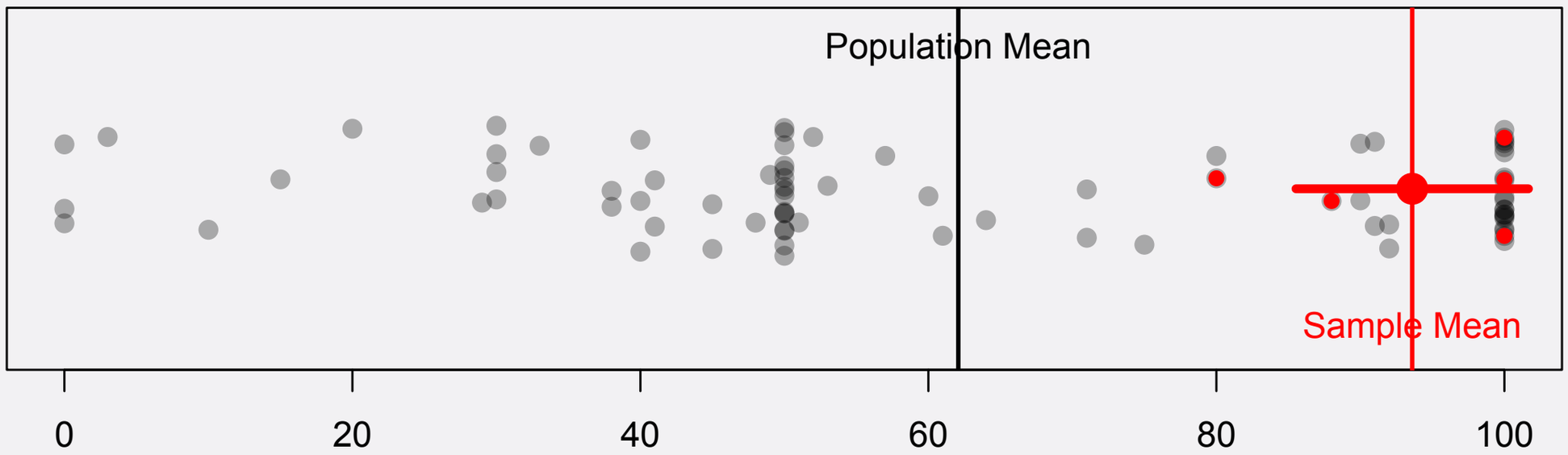


- **Yet another random sample of 5 students**
- **Sample mean $\bar{x}$=93.6 (SE=4.1)**

$$\bar{x} - (1.96 \times SE) = 93.6 - (1.96 \times 4.1) = 85.4$$

$$\bar{x} + (1.96 \times SE) = 93.6 + (1.96 \times 4.1) = 101.6$$

OUR SURVEY



- Yet another random sample of 5 students
- Sample mean $\bar{x}$=93.6 (SE=4.1)

$$\bar{x} - (1.96 \times SE) = 93.6 - (1.96 \times 4.1) = 85.4$$

$$\bar{x} + (1.96 \times SE) = 93.6 + (1.96 \times 4.1) = 101.6$$

- 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:

$$\bar{x} - 1.96 \times SE = 41 - 1.96 \times 2.05 = 37.0$$

$$\bar{x} + 1.96 \times SE = 41 + 1.96 \times 2.05 = 45.0$$

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.

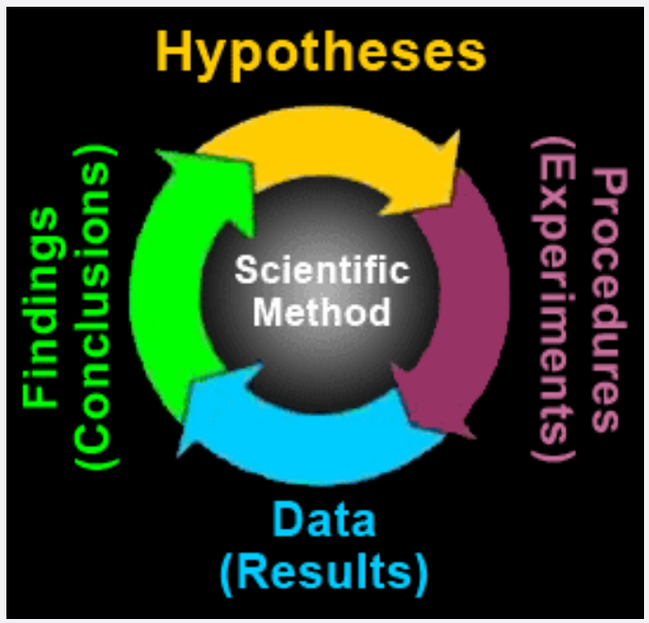
- **If we do many polls of Biden's approval rating and report 95% CI for each, 95% of them will contain (unknown) true approval rating in population**

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

NEXT STEP

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

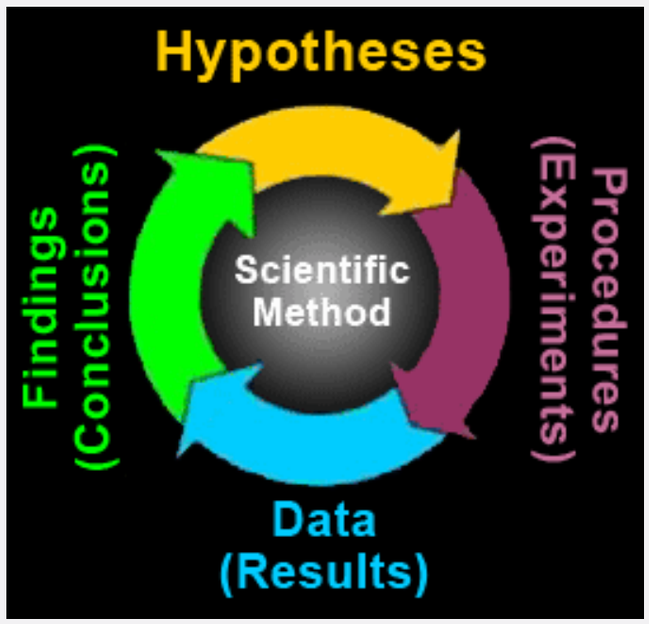


TODAY

- Finishing up Sampling and Surveys
- Hypotheses and causality

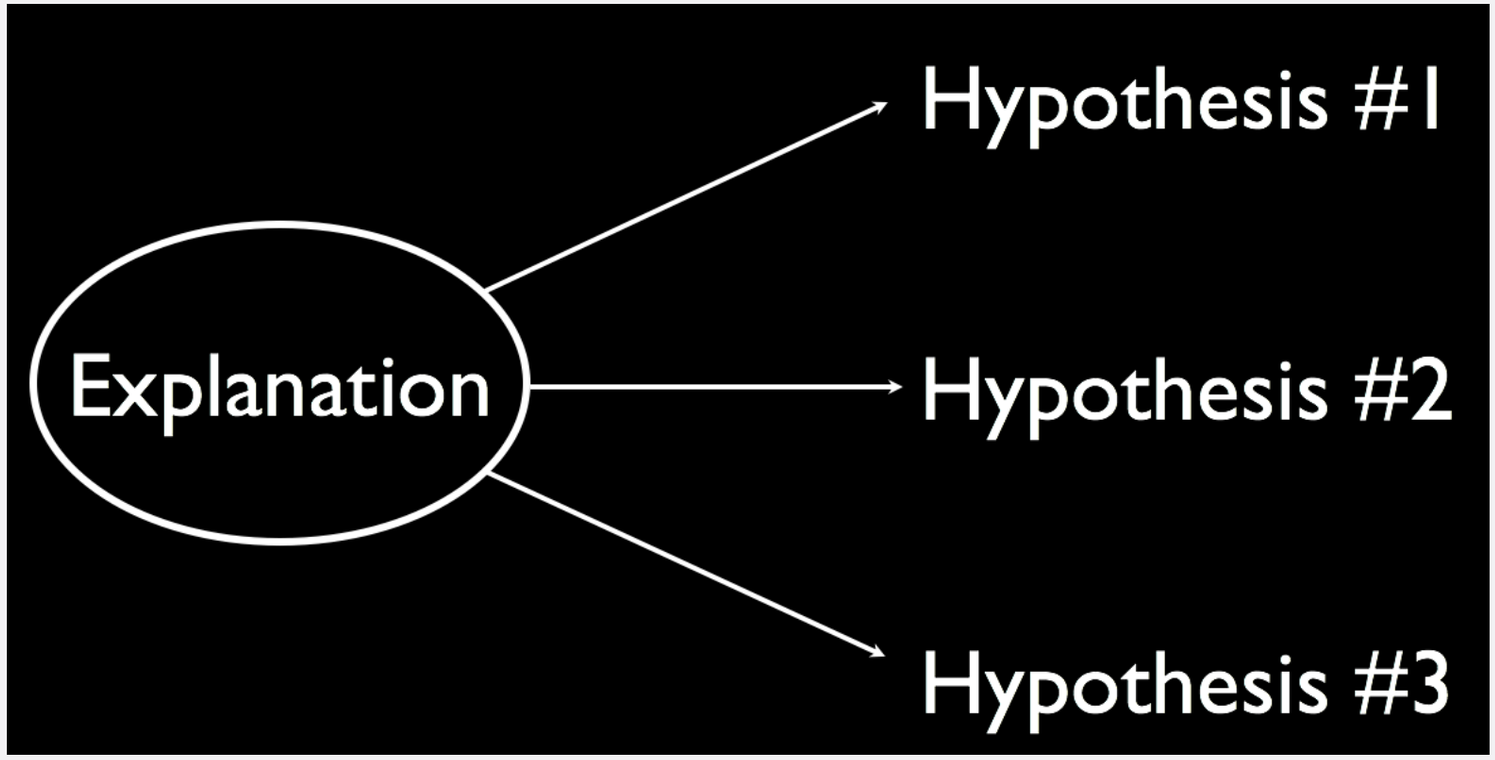
RESEARCH PROCESS

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



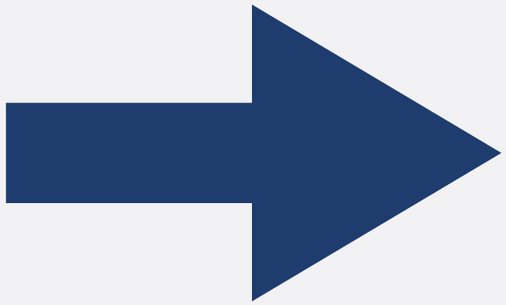
HYPOTHESES AND THEORY

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



THEORY

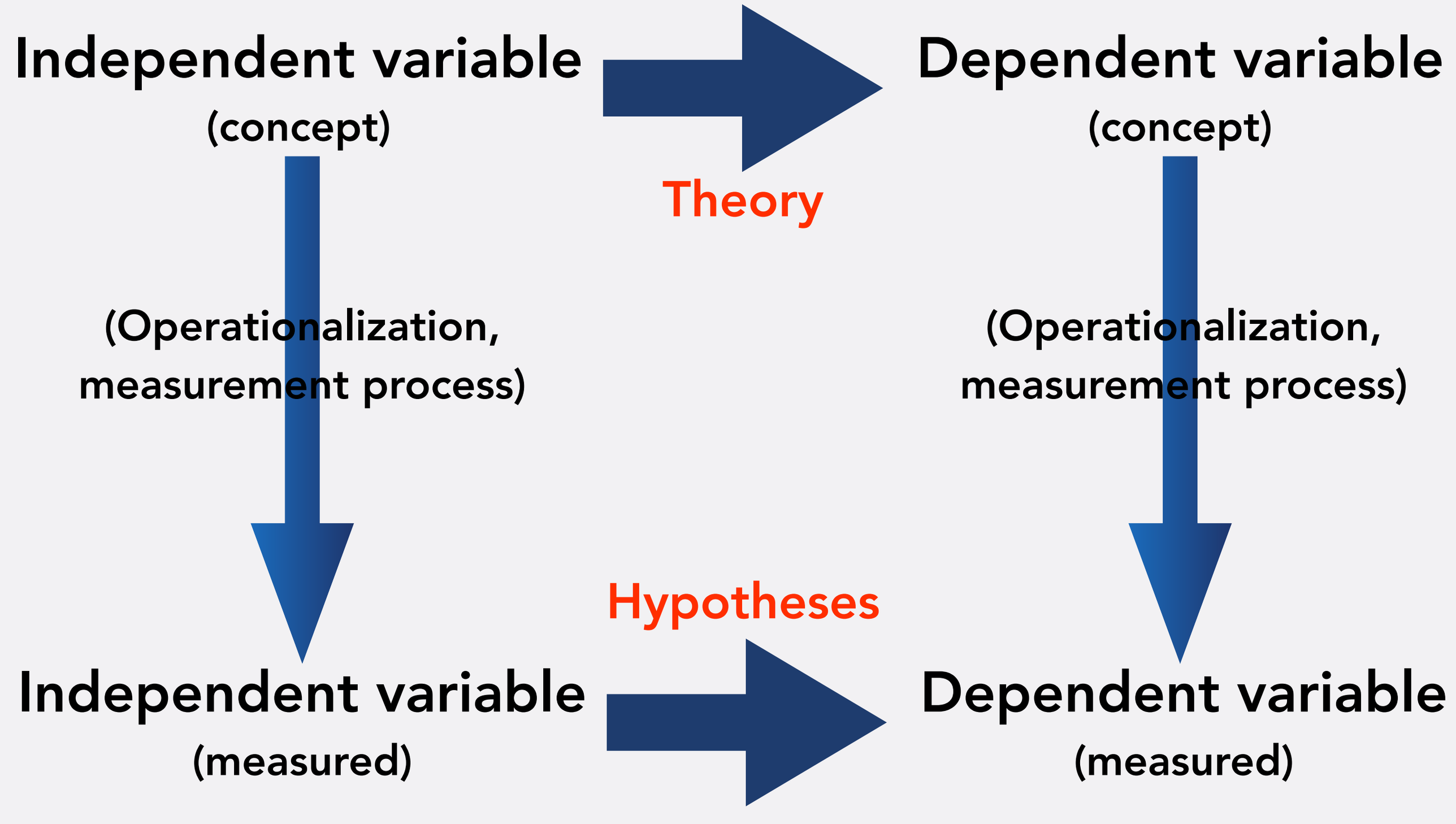
Independent variable
(concept)



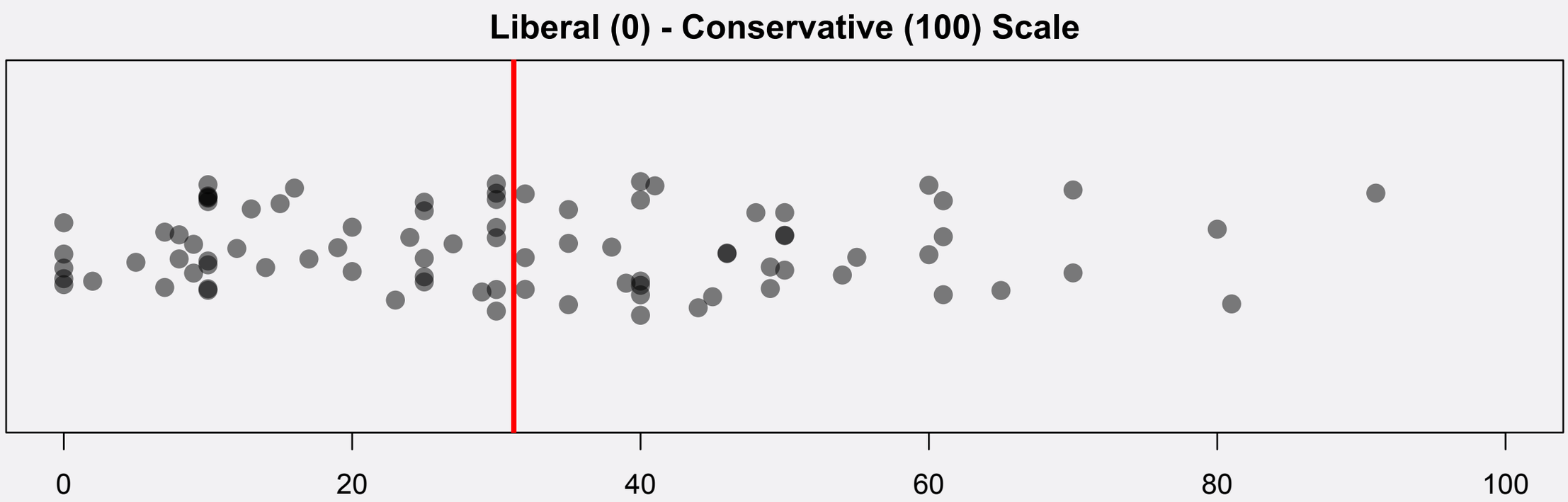
Theory

Dependent variable
(concept)

MEASUREMENT



TODAY'S EXAMPLE



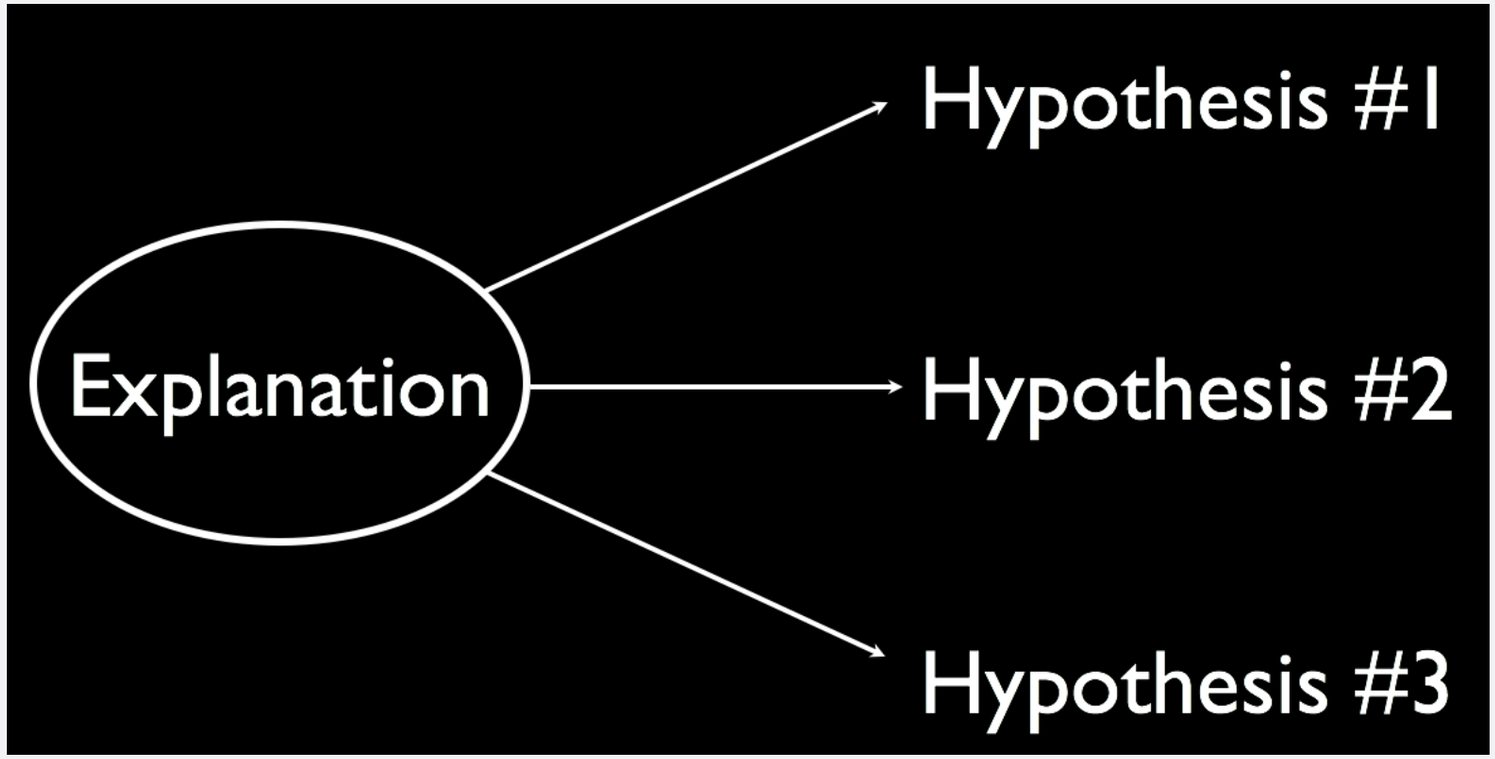
- **Why are some students in 202 liberal and others conservative? What determines which ideology students have?**

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

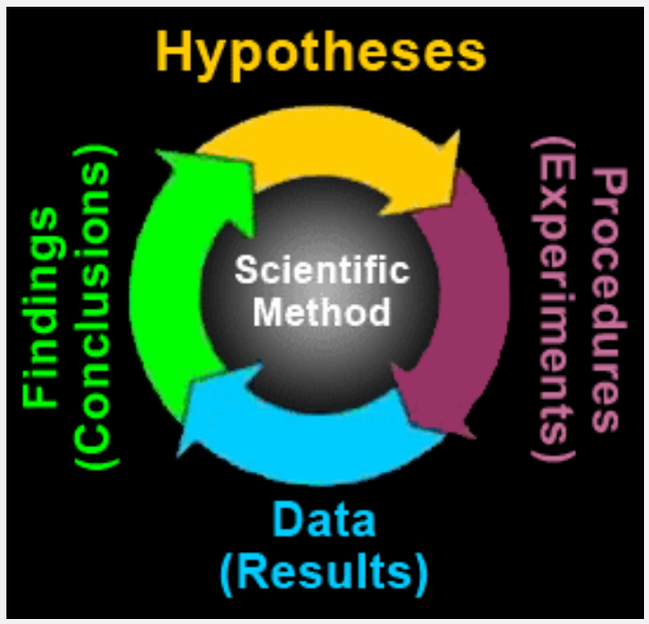
- In a comparison of [units of analysis], those having [one value of the independent variable] will be more likely to have [one value of the dependent variable] than will those having [a different value on the independent variable].

TEMPLATE

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

RESEARCH PROCESS

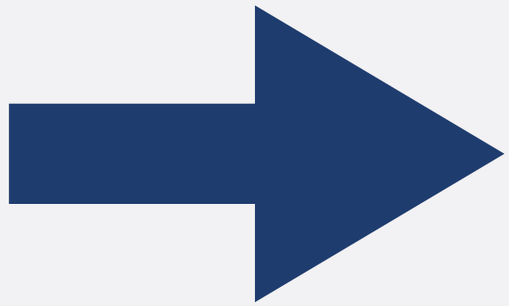
- Formulate research question
- Propose explanation/theory, hypotheses
- Data collection process
- Use data to evaluate hypotheses
- Reassess explanation



EVALUATING HYPOTHESES

- How to establish a causal relationship with data?

CAUSALITY

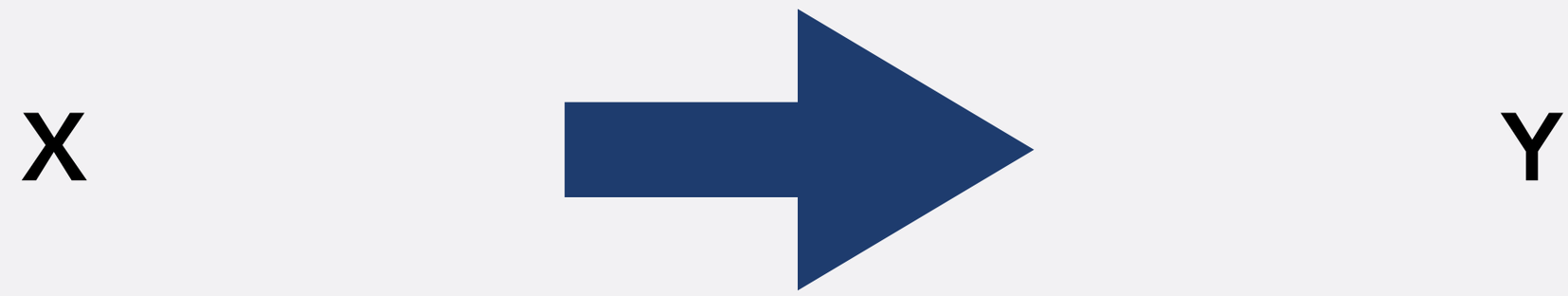
Independent variable  Dependent variable

- Most of our theories: relationship between a single cause (independent variable) and a single effect (dependent variable)
  - “bivariate” relationship (2 variables)

CAUSALITY

- But: World is more complicated than that
  - Almost never the case that *only* one independent variable causes variation in the dependent variable
  - Theories are simplified descriptions how world works, but when we test them with data, we need to take the complexity of the real world into account

HURDLES TO CAUSALITY



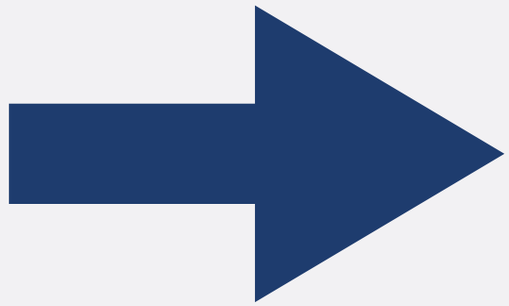
- X=independent variable; Y=dependent variable
- Want to establish that X is a cause of Y
- Four hurdles along the route to establish causality

HURDLE 1

- Is there a credible causal mechanism that connects X to Y?
  - Talked about this when looking at theory
  - Need a convincing process/mechanism that logically suggests why X might be a cause of Y

HURDLE 1

Parents' liberal-conservative orientation



Liberal-Conservative

HURDLE 1

Parents' liberal-conservative orientation → Liberal-Conservative

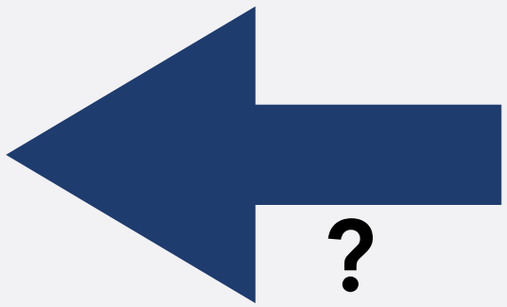
- Example
  - Parents transmit their values to their children (e.g. dinner table conversations, emphasis what is important and right or wrong, parents' political activism)

HURDLE 2

- Can we rule out the possibility that Y could cause X?
  - e.g. your theory is that smoking (X) causes depression (Y)
  - But: People who suffer from depression may be more likely to smoke

HURDLE 2

Parents' liberal-conservative orientation



Liberal-Conservative

?

- Likely or not?