DATA 1202

Spring 2024

Lecture 24

Designing Experiments

Announcements

- **Homework 12** due ~~tonight at 11pm~~ Friday at 11pm
- **Project 2**
  - Checkpoint due Friday (4/12)
  - Final deadline 4/19

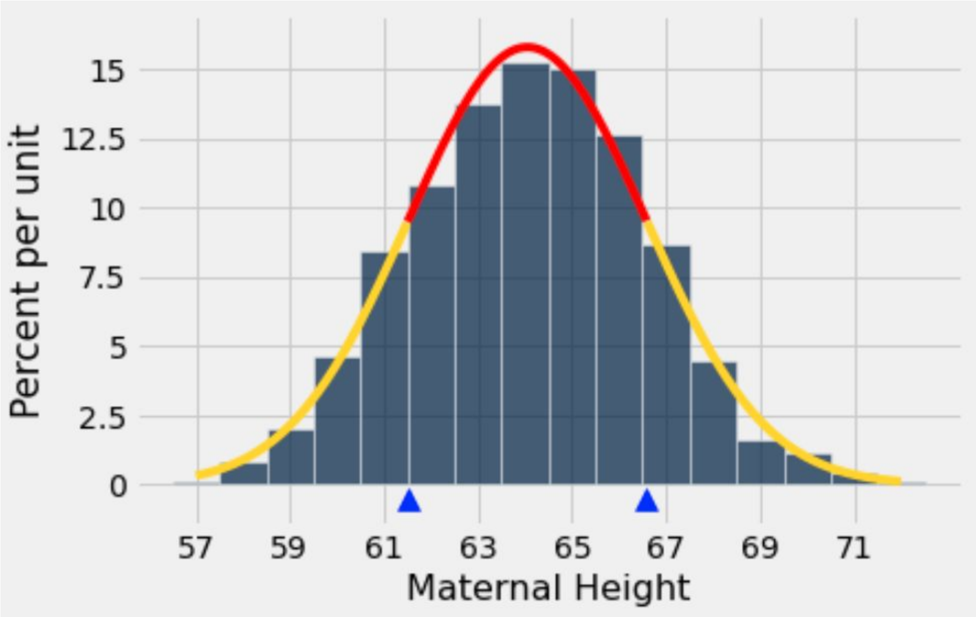
Weekly Goals

- Last week
  - The bell shaped curve and its relation to large random samples
- Monday
  - Central limit theorem
  - The variability in a random sample average
- **Today**
  - Constructing confidence intervals for sample means
  - Choosing the size of a random sample

Review: SD and Bell-Shaped Curves

If a histogram is bell-shaped, then

- Where is the average?
- What about SD?



Distribution of the Average of a Large Sample

CLT with More Details

If the sample is large and drawn at random with replacement:

Then, *regardless of the distribution of the population,*

- **the probability distribution of the sample average**
  - **is roughly normal**
  - What about mean and standard deviation?

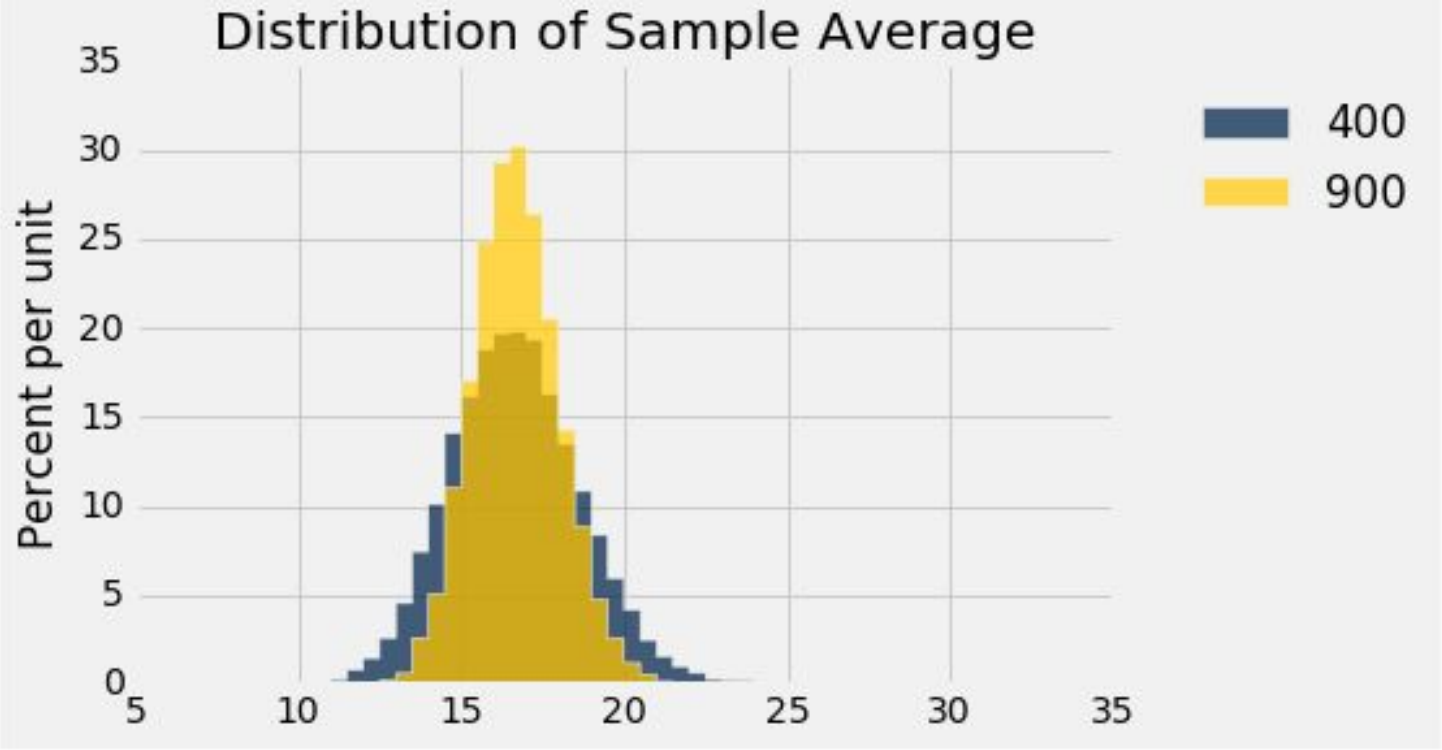
CLT with More Details

If the sample is large and drawn at random with replacement:

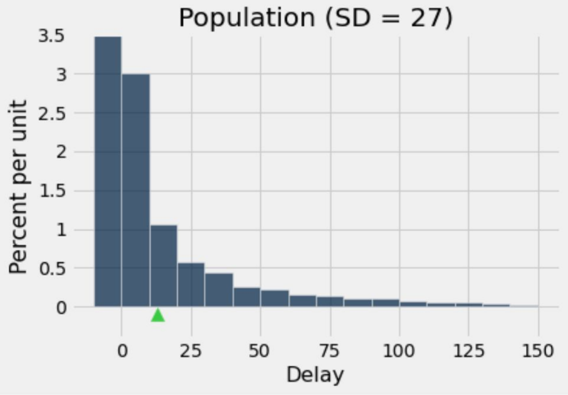
Then, *regardless of the distribution of the population,*

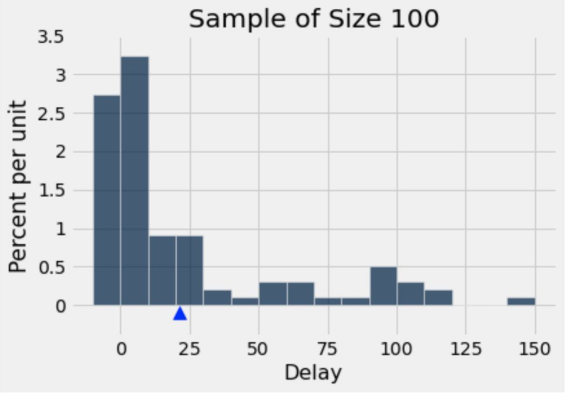
- **the probability distribution of the sample average**
  - **is roughly normal**
  - mean = population mean
  - SD = (population SD) / $\sqrt{\text{sample size}}$

Increasing Sample Size



Three Different SDs





Population of flight delays

- Population mean: ▲
- **Population SD**: 27 minutes

Random sample of 100 flights

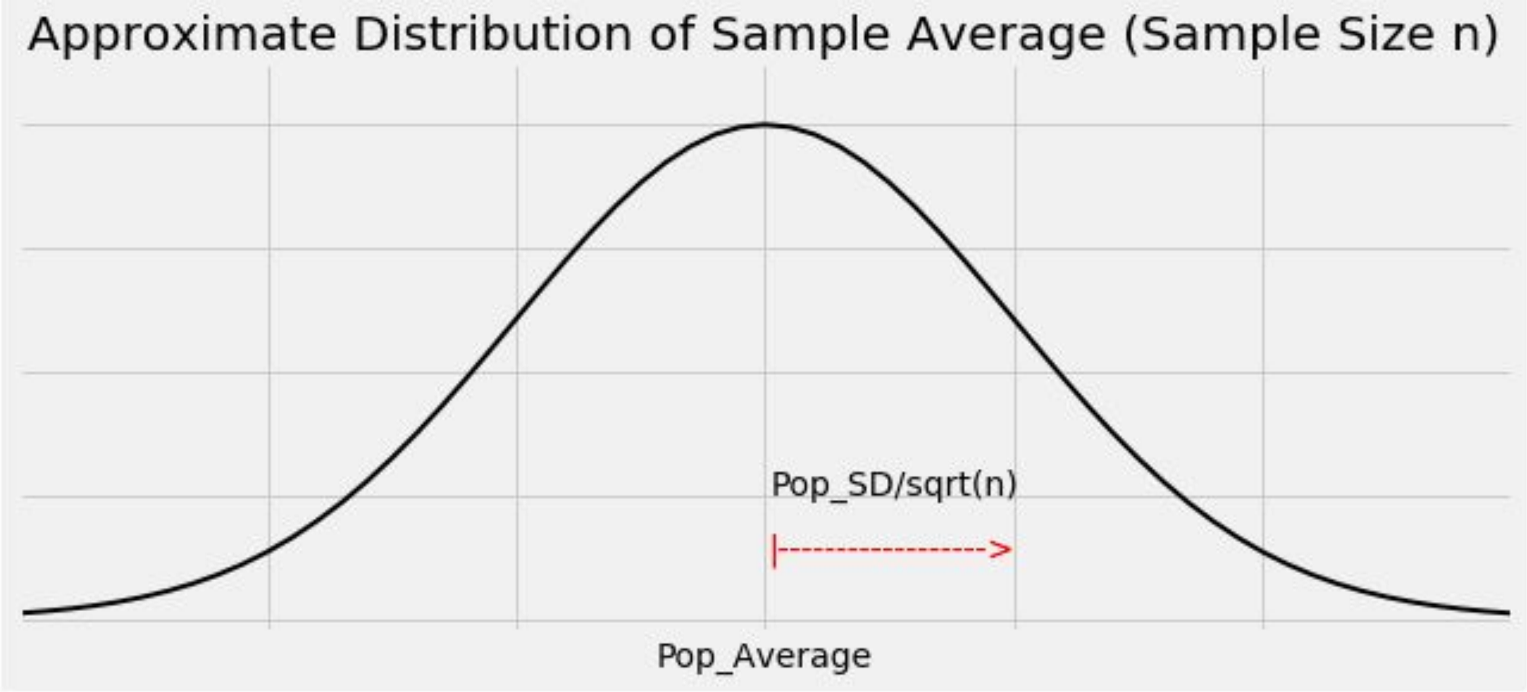
- Sample mean: ▲ (estimate of ▲)
- **Sample SD**: estimate of population SD

SD of sample average: 27/sqrt(100) = 2.7

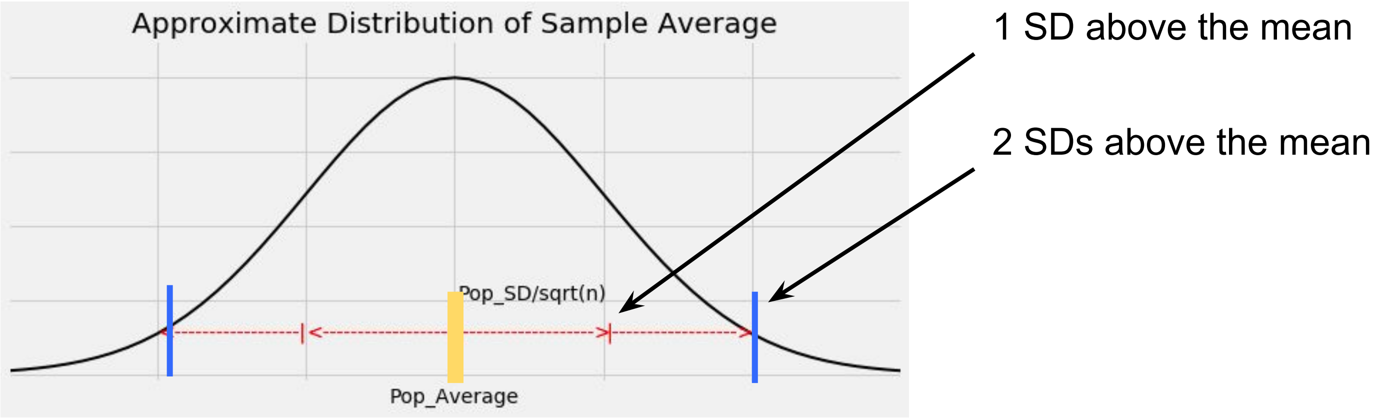
- If we calculated ▲ from 10,000 samples, their SD would be ~2.7

Confidence Intervals

Graph of the Distribution



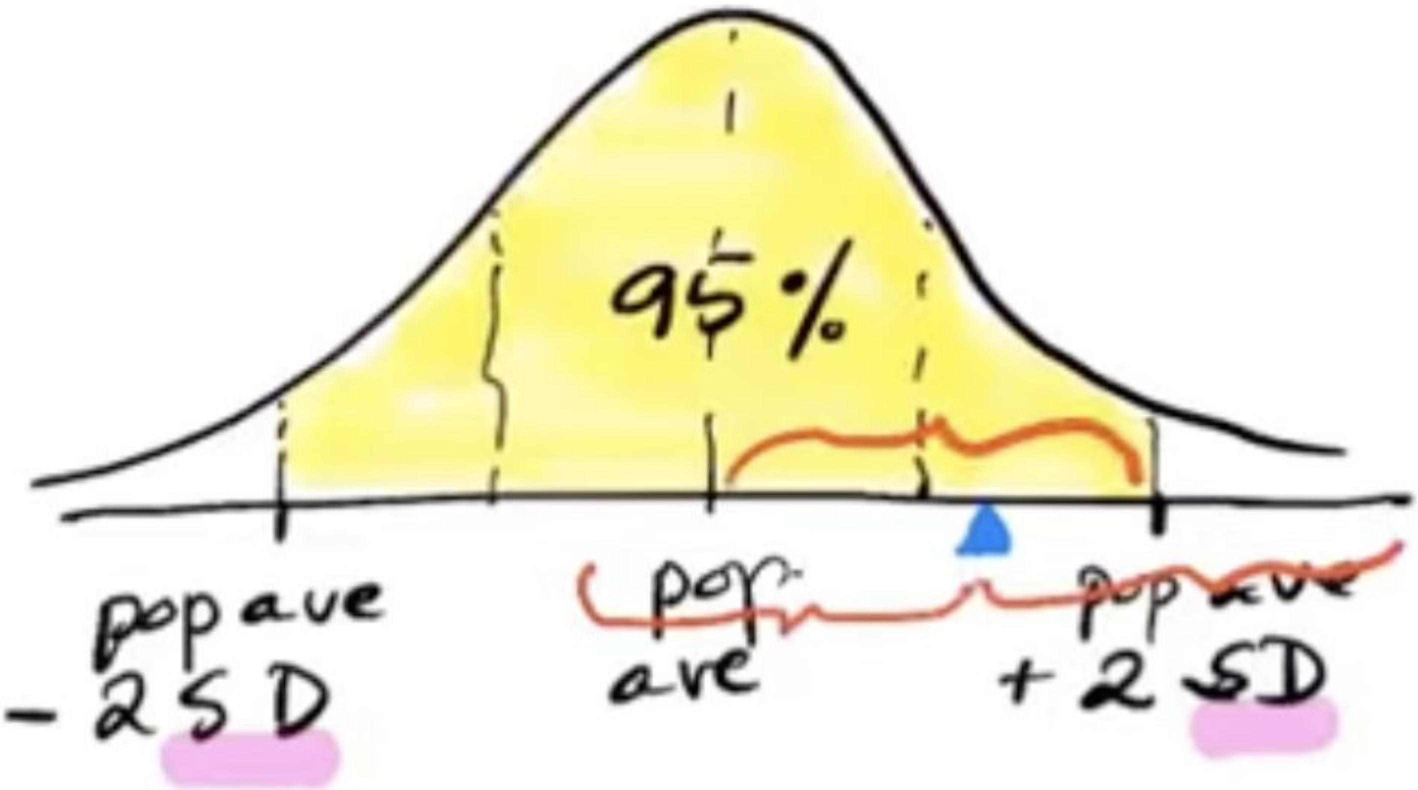
The Key to 95% Confidence



- For about 95% of all samples, the sample average and population average are within **2 SDs** of each other.
- **SD** = SD of sample average

  = (population SD) / $\sqrt{\text{sample size}}$

Constructing the Interval

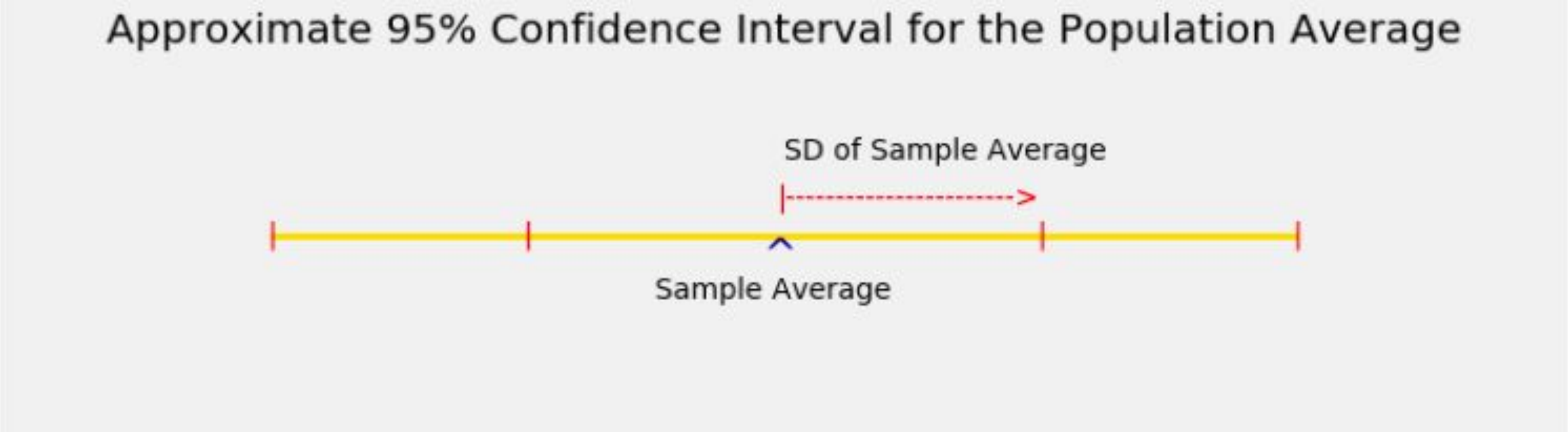


Constructing the Interval

For 95% of all samples,

- If you stand at the population average and look two **SDs** on both sides, you will find the sample average.
- Distance is symmetric.
- So if you stand at the sample average and look two **SDs** on both sides, you will capture the population average.

The Interval



(Demo)

Summarizing: construction of intervals

- 95% confidence interval for the sample mean
  - Sample_mean +/- 2*SD of the sample mean
- SD of the sample mean
  - (population SD) / $\sqrt{\text{sample size}}$

- But we don’t know the population SD
  - We can estimate it using the sample SD
  - Or overestimate it

Question

If we can make 95% confidence interval in this way:

- Sample_mean +/- 2*SD
- Then why do we need to make confidence intervals using bootstraps

This method only works for means and sums (as it is based on CLT) but bootstrap is a much more generalized approach which can work for other statistics like medians as well

Width of the Interval

Total width of a 95% confidence interval for the population average

= 4 * SD of the sample average

= 4 * (population SD) / $\sqrt{\text{sample size}}$

Sample Proportions

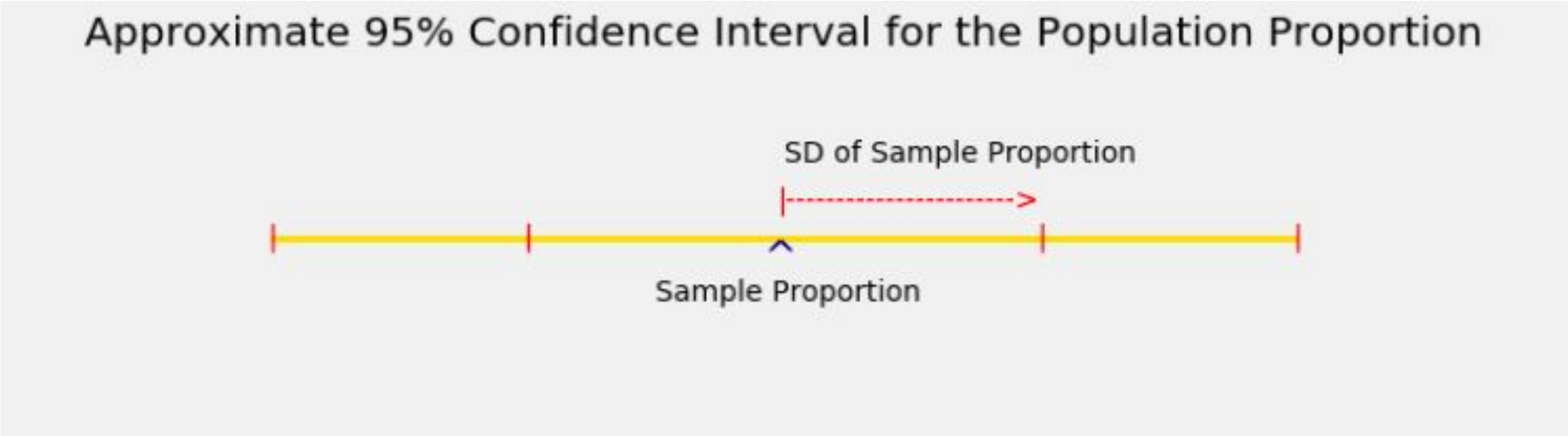
Proportions are Averages

- Data: 0 1 0 0 1 0 1 1 0 0 (10 entries)
- Sum = 4 = number of 1’s
- Average = 4/10 = 0.4 = proportion of 1’s

If the population consists of 1’s and 0’s (yes/no answers to a question), then:

- the population average is the proportion of 1’s in the population
- the sample average is the proportion of 1’s in the sample

Confidence Interval



Controlling the Width

- Total width of an approximate 95% confidence interval for a population proportion

  $= \quad 4 * (\text{SD of 0/1 population}) / \sqrt{\text{sample size}}$

- The narrower the interval, the more precise your estimate.
- Suppose you want the total width of the interval to be no more than 1%. How should you choose the sample size?

The Sample Size for a Given Width

$$0.01 = 4 * (\text{SD of 0/1 population}) / \sqrt{\text{sample size}}$$

- Left side: 1%, the max total width that you'll accept
- Right side: formula for the total width

$$\sqrt{\text{sample size}} = 4 * (\text{SD of 0/1 population}) / 0.01$$

(Demo)

“Worst Case” Population SD

- $\sqrt{\text{sample size}} = 4 * (\text{SD of 0/1 population}) / 0.01$
- SD of 0/1 population is at most 0.5
- $\sqrt{\text{sample size}} \geq 4 * 0.5 / 0.01$
- $\text{sample size} \geq (4 * 0.5 / 0.01) ** 2 = 40000$
- The sample size should be 40,000 or more

Discussion Question

Subscribe | SCIENTIFIC AMERICAN. | Cart 0 | Sign In | Stay Informed

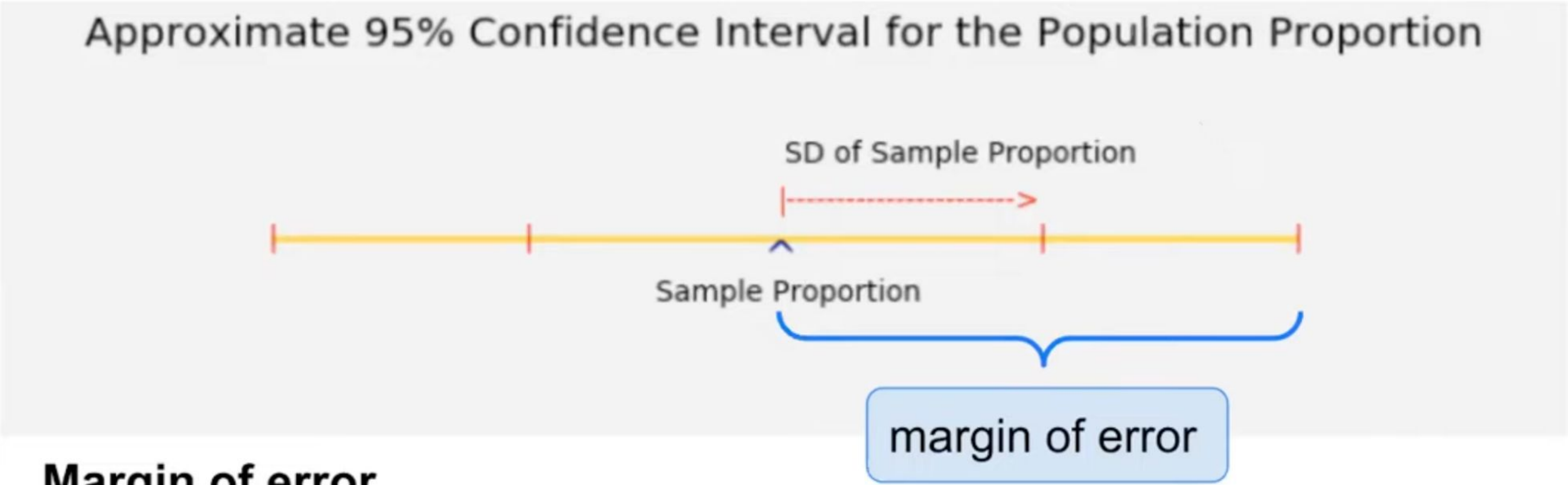
THE SCIENCES MIND HEALTH TECH SUSTAINABILITY EDUCATION VIDEO PODCASTS BLOGS PUBLICATIONS

THE SCIENCES

How can a poll of only 1,004 Americans represent 260 million people with only a 3 percent margin of error?

https://www.scientificamerican.com/article/howcan-a-poll-of-only-100/

Margin of Error in Polls



Margin of error

- Distance from the center to an end
- Half the width of the interval
- 2 * SD of sample proportion

Discussion Question

- 3% margin of error means **width of 6%**

  width = 4 * (0.5) / $\sqrt{1004}$

  width ≈ 0.063, so margin of error ≈ 3.15%

Discussion Question

- A researcher is estimating a population proportion based on a random sample of size 10,000.

Fill in the blank with a decimal:

- With chance at least 95%, the estimate will be correct to within ________________.

Discussion Question

- With chance at least 95%, the estimate will be correct to within **0.01**.

  width = 4 * (0.5) / $\sqrt{10000}$

  width = 0.02, so margin of error = 0.01

Discussion Question

- I am going to use a 68% confidence interval to estimate a population proportion.
- I want the total width of my interval to be no more than 2.5%.
- How large must my random sample be?

Discussion Question

- How large must my random sample be?

$0.025 = \mathbf{2} * (0.5) / \sqrt{\text{sample size}}$

$\sqrt{\text{sample size}} = 2 * (0.5) / 0.025$

sample size = 40**2 = 1600