



DATA 8
Fall 2022

Lecture 28

Designing Experiments

Announcements

- **Lab 8** is due tonight at 11pm
 - **Homework 9** has been released and is due next Wednesday (11/2) at 11pm
 - **Project 2** will be released this afternoon
 - Checkpoint next Friday (11/4)
 - Final deadline on November 11th
 - **Get started early and utilize OH/Project Party!**
-

Weekly Goals

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

Distribution of the Average of a Large Sample

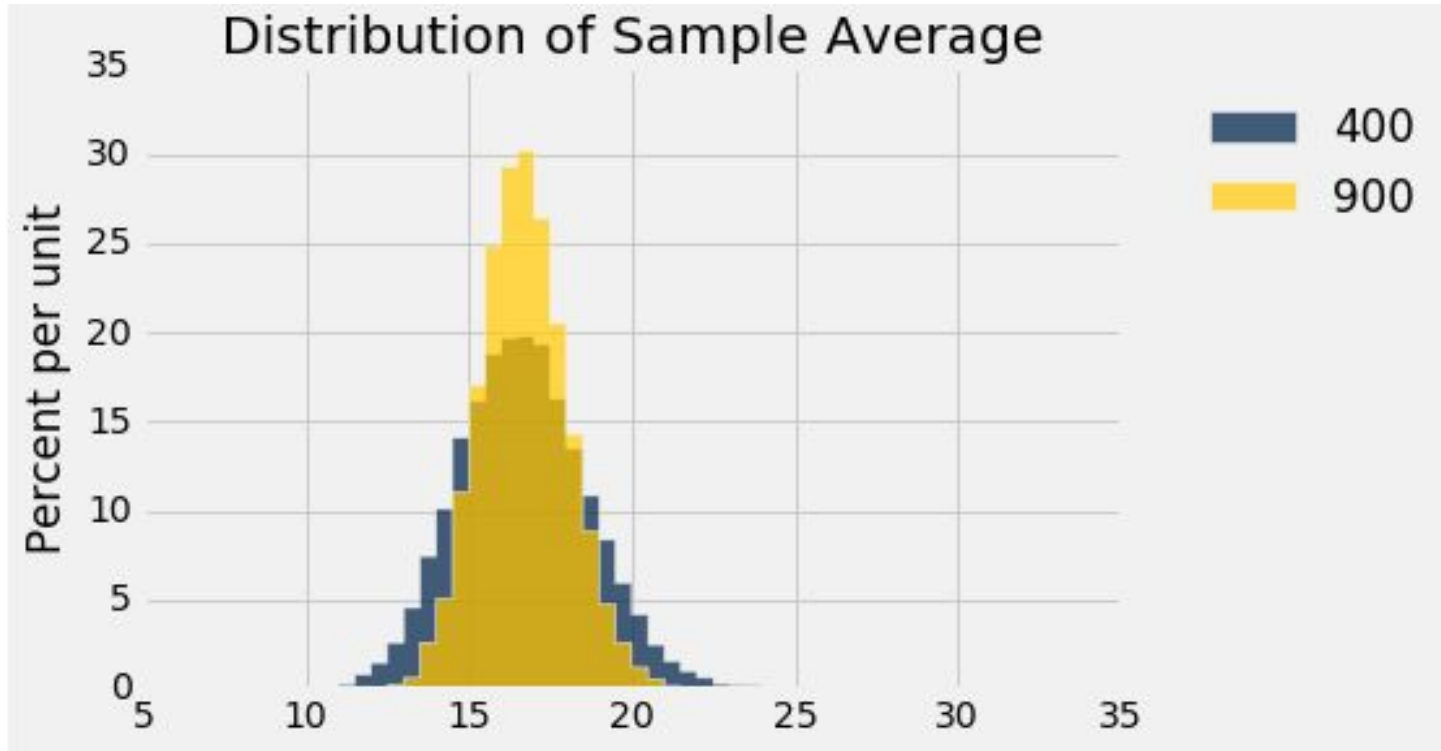
CLT with More Details

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

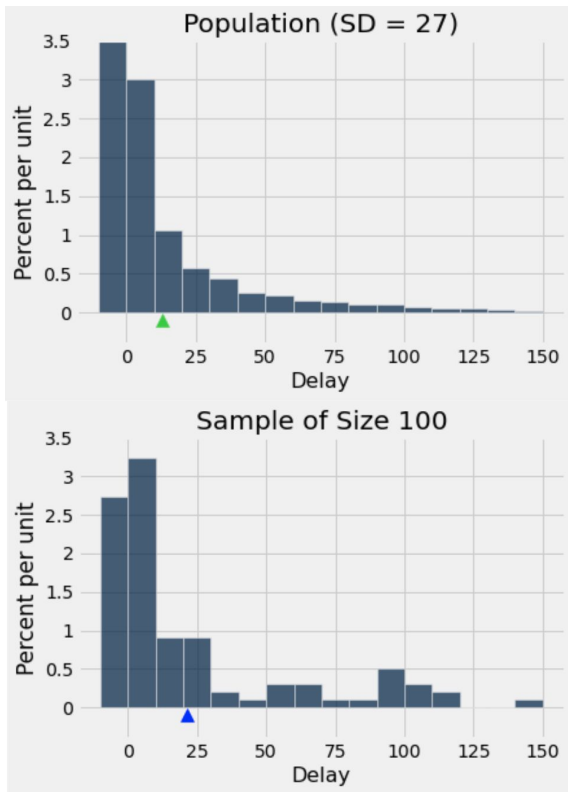
Then, *regardless of the distribution of the population,*

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

Increasing Sample Size



Three Different SDs



Population of flight delays

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

Random sample of 100 flights

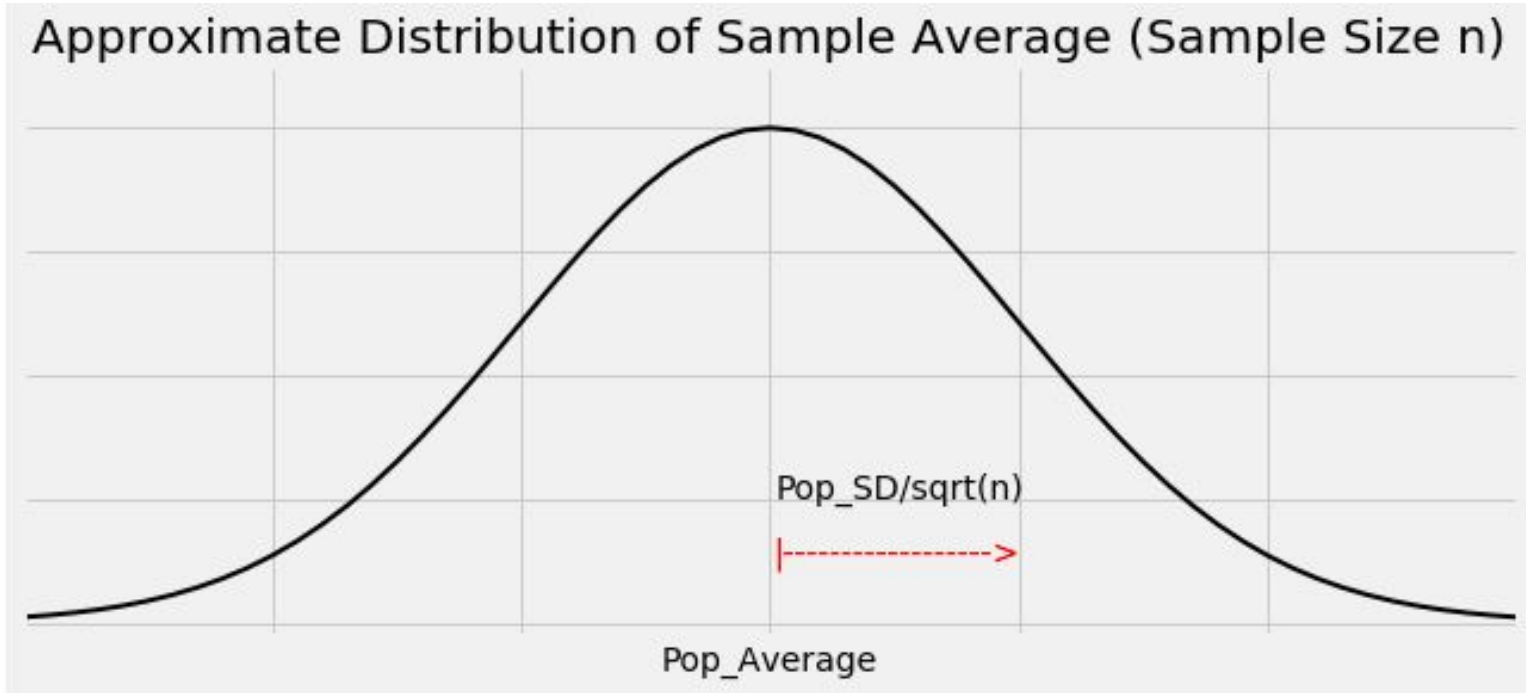
- Sample mean: ▲ (estimate of ▲)
- **Sample SD**: estimate of population SD

SD of sample average: $27/\sqrt{100} = 2.7$

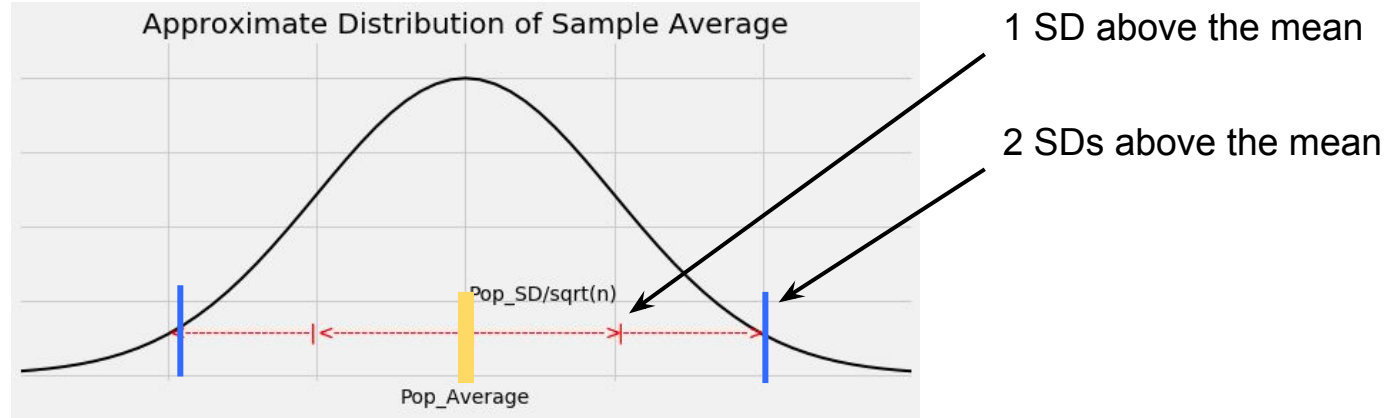
- If we calculated ▲ from 10,000 samples, their SD would be ~ 2.7

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) / $\sqrt{\text{sample size}}$

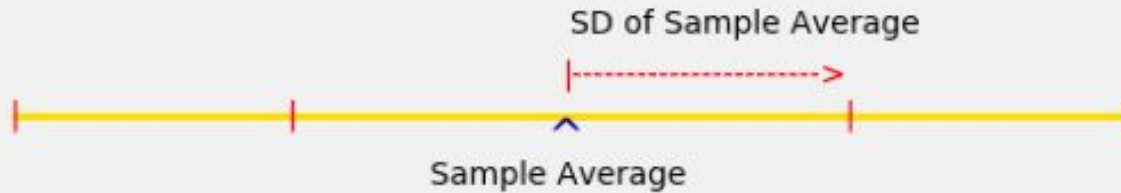
Constructing the Interval

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

Approximate 95% Confidence Interval for the Population Average



(Demo)

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

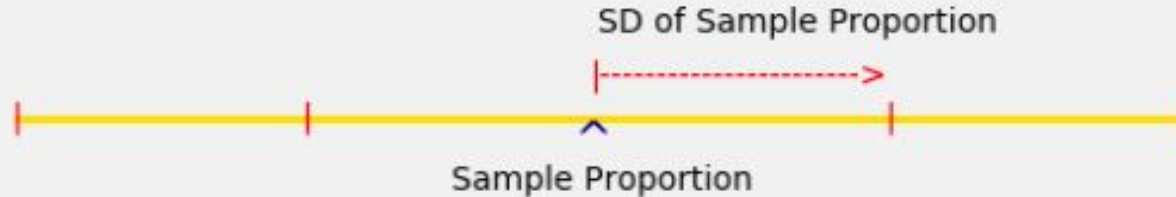
- Data: 0 1 0 0 1 0 1 1 0 0 (10 entries)
- Sum = 4 = number of 1's
- Average = $4/10 = 0.4$ = proportion of 1's

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

Approximate 95% Confidence Interval for the Population Proportion



Controlling the Width

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

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

- 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/>

Discussion Question

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

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

width ≈ 0.063 , so margin of error $\approx 3.15\%$

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**.

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

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?

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

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

$$\text{sample size} = 40^{**}2 = 1600$$
