CORNELL UNIVERSITY FOUNDED A.D. 1865

DSFA

Spring 2021

Lecture 26

Designing Experiments

Announcements

- Project 2, Part 2, due Friday 5:59PM
- Prelim 2, April 20, 8:30PM-10PM in Kennedy 116 (here) for Ithaca-resident students
  - Coverage from Lecture 12 - Lecture 26 (today)
  - Review session on Saturday 3:30PM-5:30PM, room TBA
  - Review sheet and sample exam posted later today.
  - NB: The sample exam is not one I wrote, and is likely to be somewhat different than what I will do.
  - Table of functions included again, allowed a double-sided sheet of notes you make yourself

Questions from the Past Week

- How can we quantify natural concepts like “center” and “variability”?

- Why do many of the empirical distributions that we generate come out bell shaped?

- How is sample size related to the accuracy of an estimate?

How Far from the Average?

- Standard deviation (SD) measures roughly how far the data are from their average

- SD = root mean square of deviations from average
  5 4 3 2 1

- SD has the same units as the data

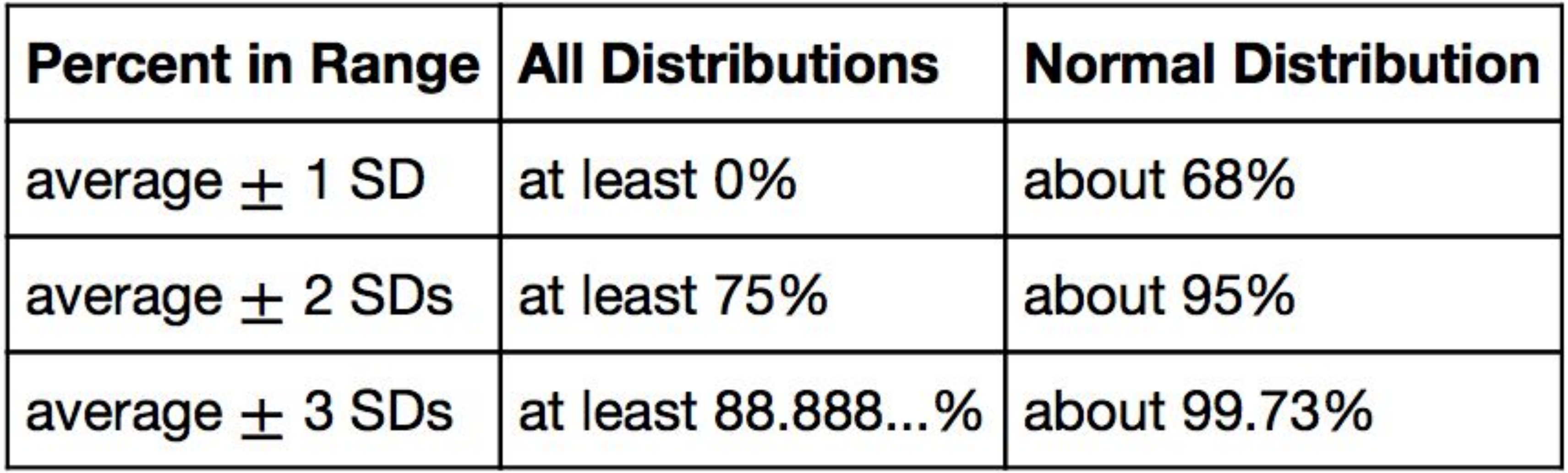
How Big are Most of the Values?

No matter what the shape of the distribution,
the bulk of the data are in the range “average ± a few SDs”

If a histogram is bell-shaped, then

- Almost all of the data are in the range “average ± 3 SDs”

Bounds and Normal Approximations



Percent in Range	All Distributions	Normal Distribution
average ± 1 SD	at least 0%	about 68%
average ± 2 SDs	at least 75%	about 95%
average ± 3 SDs	at least 88.888...%	about 99.73%

Central Limit Theorem

If the sample is

- large, and
- drawn at random with replacement,

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

the distribution of the sample sum
(or of the sample average) is roughly bell-shaped

(Demo)

Distribution of the Sample Average

- Imagine all possible random samples of the same size as yours. There are lots of them.

- Each of these samples has a mean.

- The **distribution of the sample average** is the distribution of the means of all the possible samples.

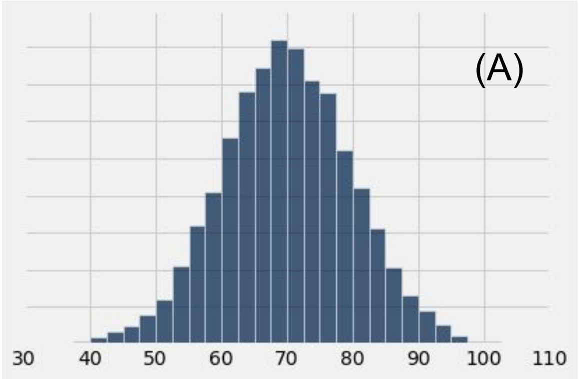
Distribution of the Sample Average

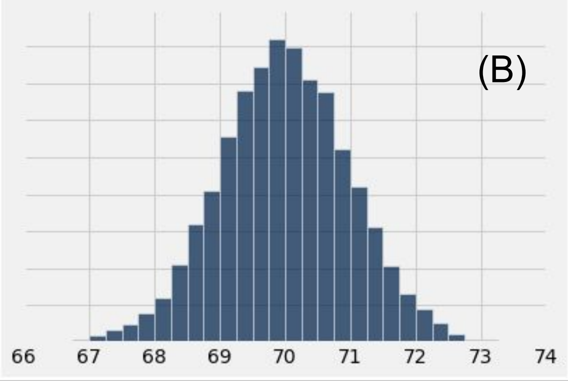
- Fix a large sample size.
- Draw all possible random samples of that size.
- Compute the average of each sample.
- You’ll end up with a lot of averages.
- The distribution of those is called the *distribution of the sample average.*
- It’s roughly normal, centered at the population average.
- SD = (population SD) / $\sqrt{\text{sample size}}$

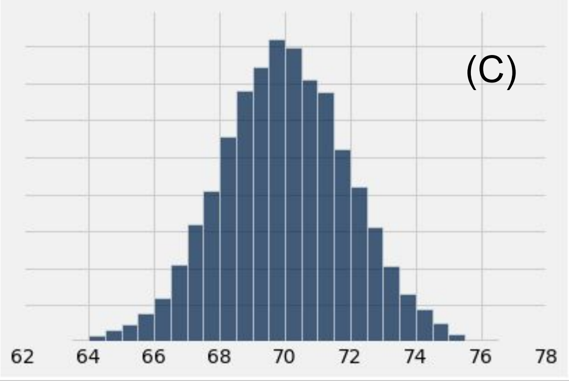
(Demo)

Discussion Question

A population has average 70 and SD 10. One of the histograms below is the empirical distribution of the averages of 10,000 random samples of size 100 drawn from the population. Which one?

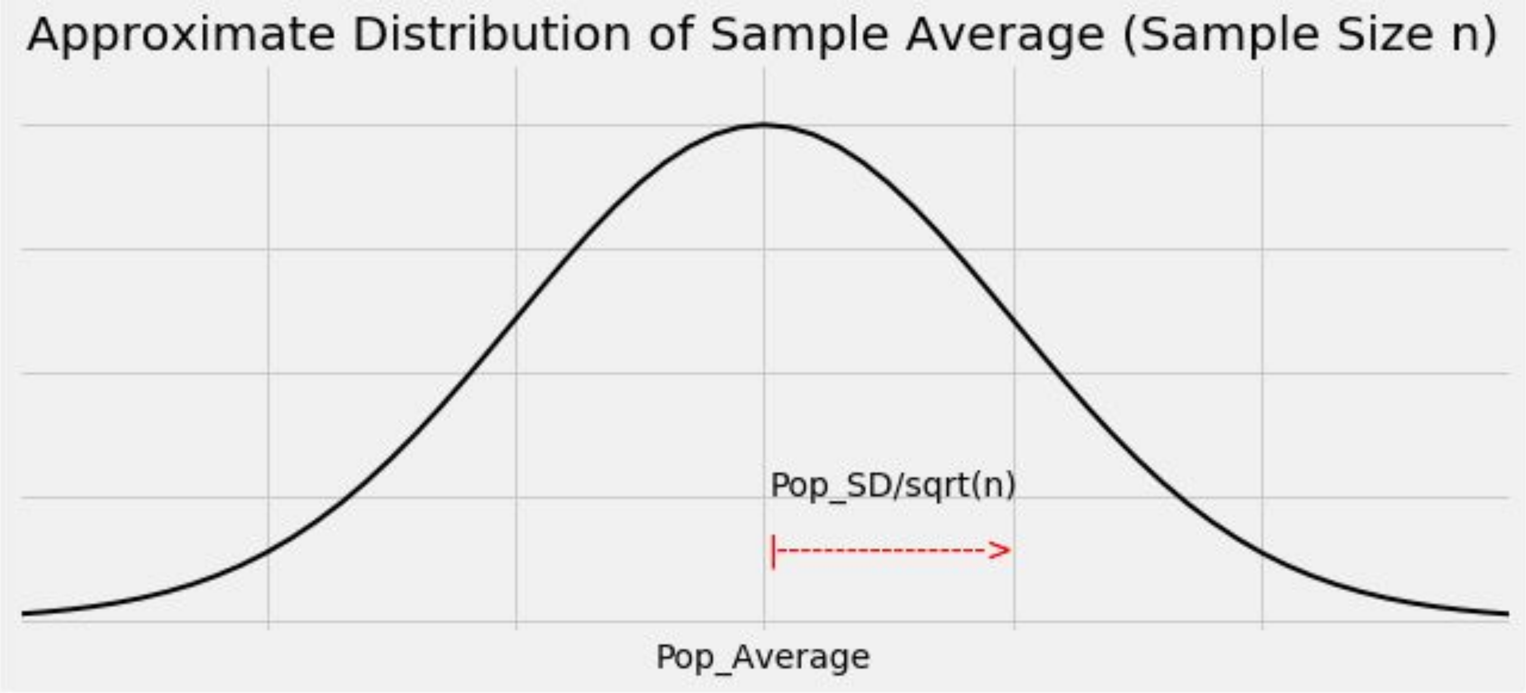




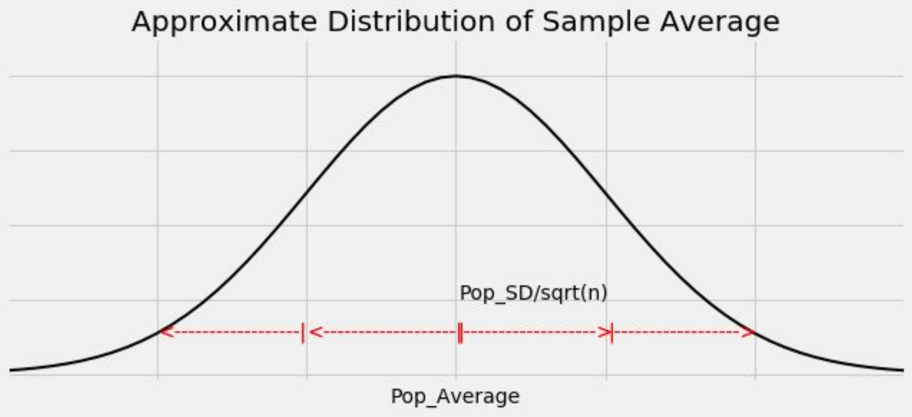


Confidence Intervals

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

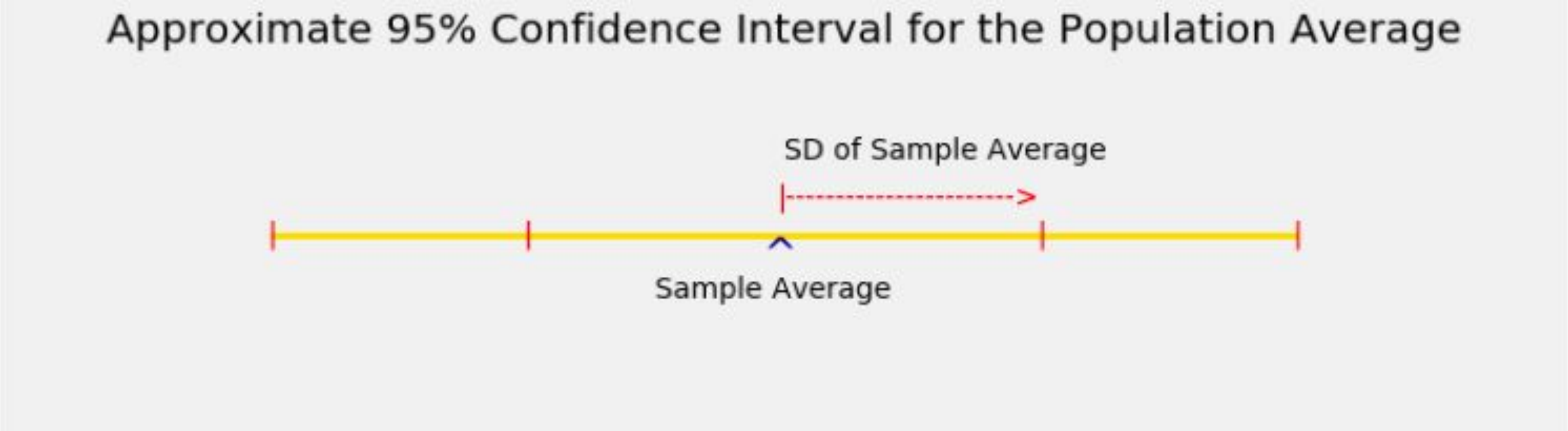
For 95% of all samples,

- If you stand at the population average and look two **SD**s on both sides, you will find the sample average.

- Distance is symmetric.

- So if you stand at the sample average and look two **SD**s on both sides, you will capture the population average.

The Interval



Approximate 95% Confidence Interval for the Population Average

Width of the Interval

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

= 4 * SD of the sample average

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

Example

Suppose I want to know the average flight delay to within ±2 minutes with 95% confidence. How big a sample size do I need?

4 * SD of sample average = 4 minutes

4 * population SD / √sample size = 4 minutes

√sample size = (4 * population SD) / 4 minutes

√sample size is about population SD (40), so sample size about 1600.

Sample Proportions

Proportions are Averages

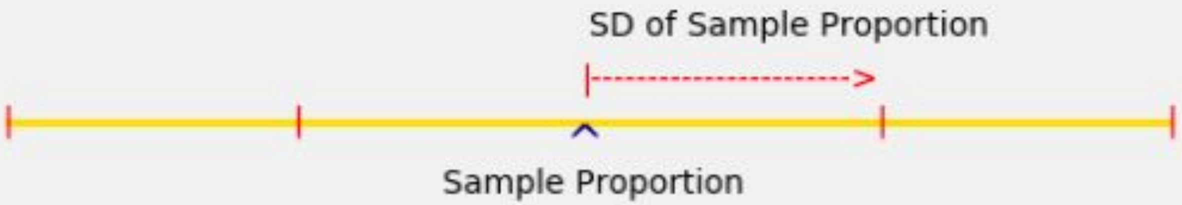
- Data: 0 1 0 0 1 0 1 1 0 0 (10 entries)
- Sum = 4 = number of 1’s
- Average = 4/10 = 0.4 = proportion of 1’s

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

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

- The narrower the interval, the more accurate your estimate.
- Suppose you want the total width of the interval to be no more than 3%. How should you choose the sample size?

The Sample Size for a Given Width

$$0.03 \;=\; 4 * (\text{SD of 0/1 population}) / \sqrt{\text{sample size}}$$

- Left hand side is 3%, the maximum total width that you will accept
- Right hand side is the formula for the total width

$$\sqrt{\text{sample size}} \;=\; 4 * (\text{SD of 0/1 population}) / 0.03$$

(Demo)

“Worst Case” Population SD

- $\sqrt{\text{sample size}}$ = 4 * (SD of 0/1 population) / 0.03

- SD of 0/1 population is at most 0.5

- $\sqrt{\text{sample size}}$ ≥ 4 * 0.5 / 0.03

- sample size ≥ (4 * 0.5 / 0.03) ** 2 = 4444.44

- The sample size should be 4445 or more

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 sample be?