DATA 1202

Spring 2024

Lecture 23

Sample Means

Announcements

- **Homework 12** due ~~Wed. at 11pm~~ Friday at 11pm
- **Project 2 - Checkpoint** due on Friday

Weekly Goals

- Last Week
  - The bell shaped curve and its relation to large random samples
  - Introduced Central limit theorem
- **Today**
  - Central limit theorem
  - The variability in a random sample average
- Wednesday
  - Choosing the size of a random sample

Central Limit Theorem

Sample Averages

- The Central Limit Theorem describes how the normal distribution (a bell-shaped curve) is connected to random sample averages.
- We care about sample averages because they estimate population averages.

Central Limit Theorem

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

Distribution of the Sample Average

Why is There a Distribution?

- You have only one random sample, and it has only one average.
- But **the sample could have come out differently**.
- And then the sample average might have been different.
- So there are many possible sample averages.

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

(Demo)

Specifying the Distribution

Suppose the random sample is large.

- We have seen that the distribution of the sample average is roughly bell shaped.

- Important questions remain:
  - Where is the center of that bell curve?
  - How wide is that bell curve?

Center of the Distribution

The Population Average

The distribution of the sample average is roughly a bell curve centered at the population average.

Variability of the Sample Average

Why Is This Important?

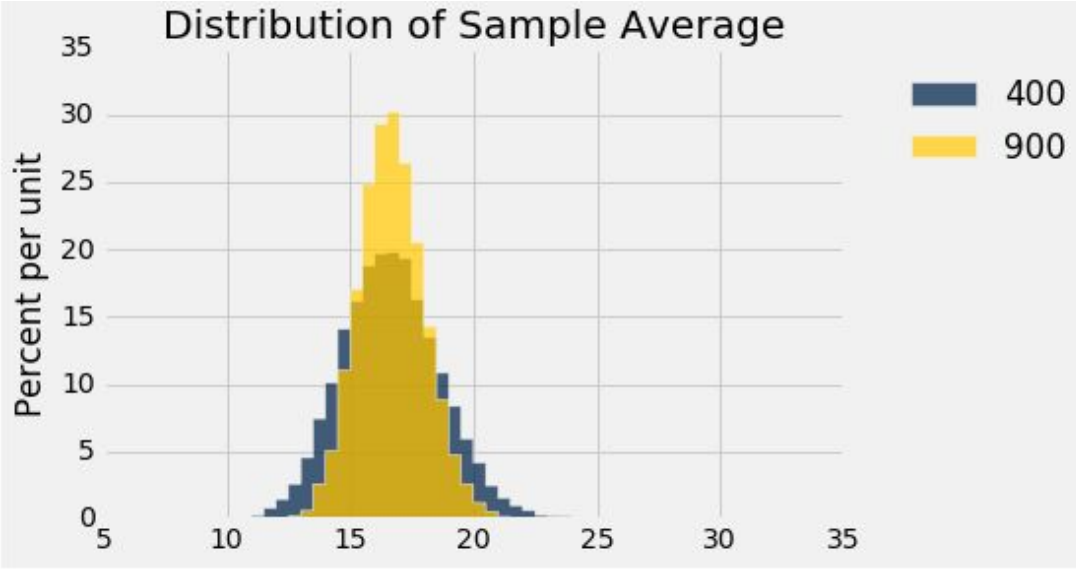
- Along with the center, the spread helps identify exactly which normal curve is the distribution of the sample average.
- The variability of the sample average helps us measure how accurate the sample average is as an estimate of the population average.
- If we want a specified level of accuracy, understanding the variability of the sample average helps us work out how large our sample has to be.

(Demo)

Discussion Question

The gold histogram shows the distribution of _____ values, each of which is _____.

(a) 900
(b) 10,000
(c) a randomly sampled flight delay
(d) an average of flight delays



The Two Histograms

- The gold histogram shows the distribution of 10,000 values, each of which is an average of 900 randomly sampled flight delays.
- The blue histogram shows the distribution of 10,000 values, each of which is an average of 400 randomly sampled flight delays.
- Both are roughly bell shaped.
- The larger the sample size, the narrower the bell.

(Demo)

Variability of the Sample Average

- The distribution of all possible sample averages of a given size is called the *distribution of the sample average.*
- We approximate it by an empirical distribution.
- By the CLT, it's roughly normal:
  - Center = the population average
  - SD = (population SD) / $\sqrt{\text{sample size}}$

Discussion Question

A city has 500,000 households. The annual incomes of these households have an average of $65,000 and an SD of $45,000. The distribution of the incomes [pick one and explain]:

(a) is roughly normal because the number of households is large.
(b) is not close to normal.
(c) may be close to normal, or not; we can't tell from the information given.

Central Limit Theorem

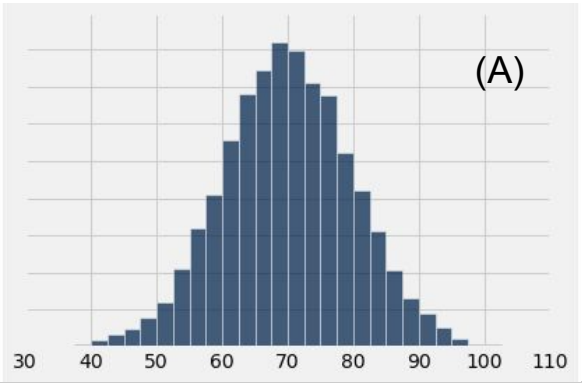
If the sample is large and drawn at random with replacement,

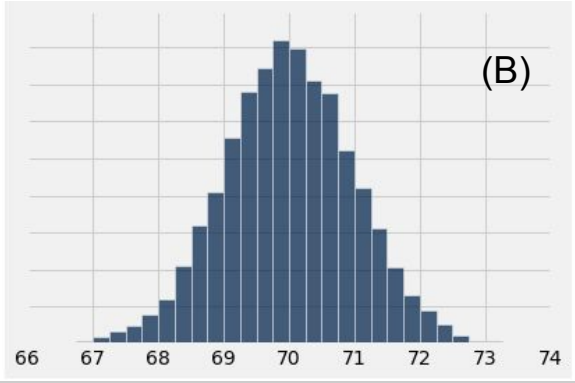
Then, *regardless of the distribution of the population,*

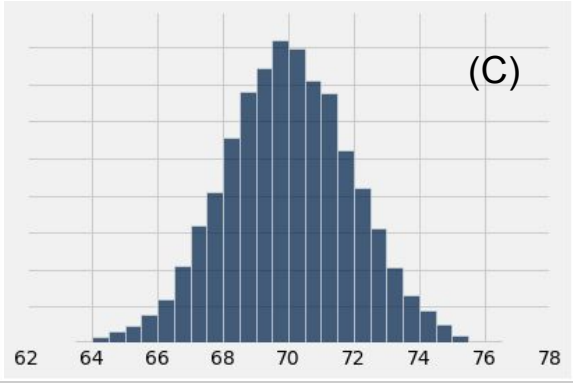
- **the probability distribution of the sample average:**
  - is roughly normal
  - mean = population mean
  - SD = (population SD) / $\sqrt{\text{sample size}}$

Discussion Question

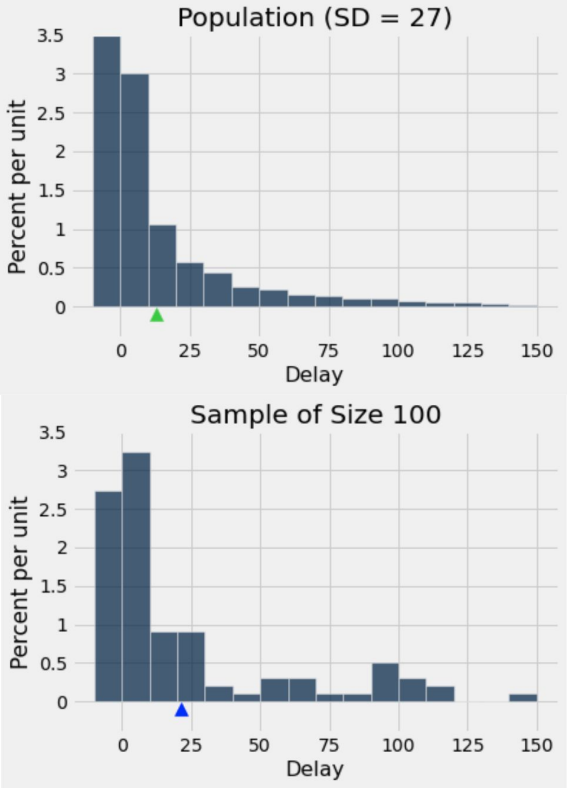
A population has average 70 and SD 10. One of the histograms below is the empirical distribution of the averages of 10,000 random samples of size 100 drawn from the population. Which one?







Three Different SDs



Population of flight delays

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

Random sample of 100 flights

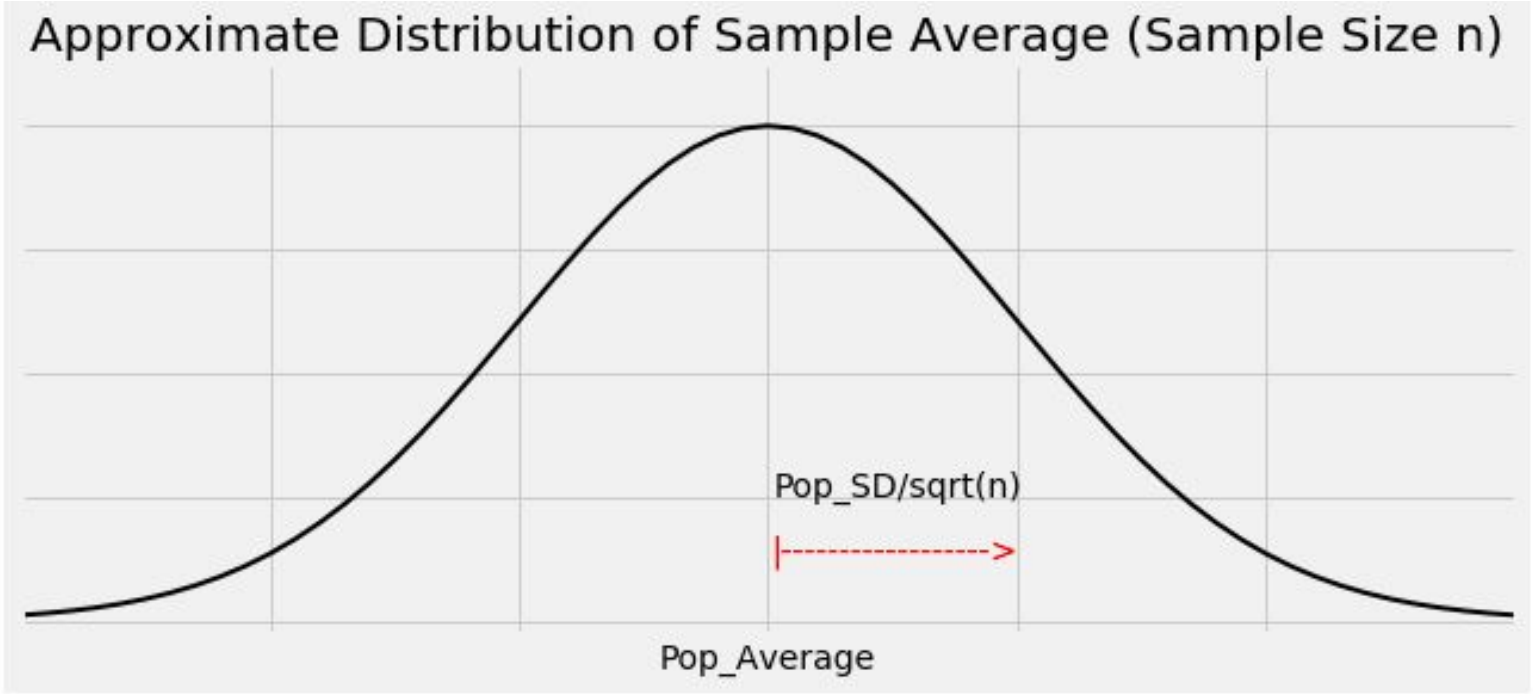
- Sample mean: ▲ (estimate of ▲)
- **Sample SD**: estimate of population SD

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

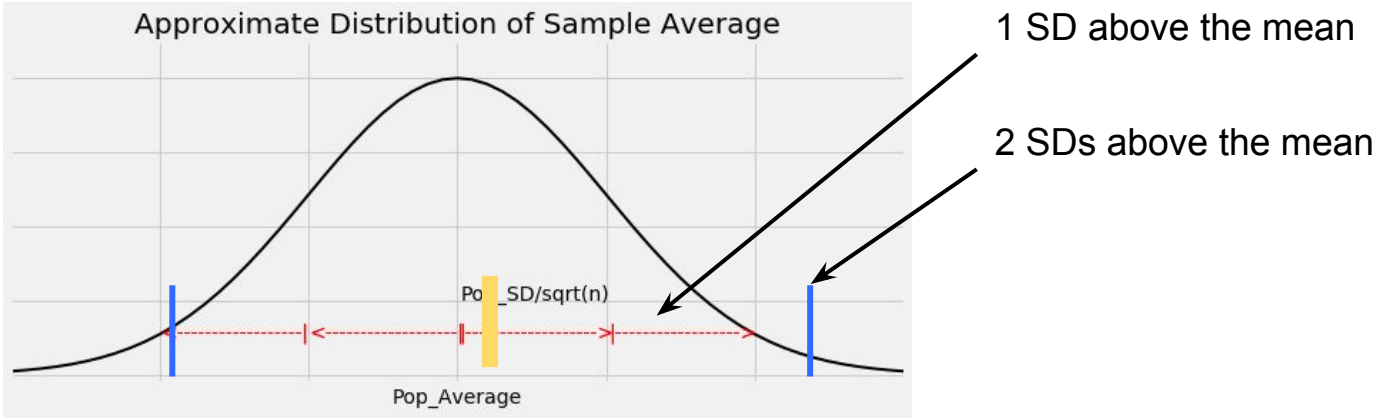
- If we instead calculated ▲ from 10,000 samples, their SD would be ~0.27

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 SD**s of each other.
- **SD** = SD of sample average

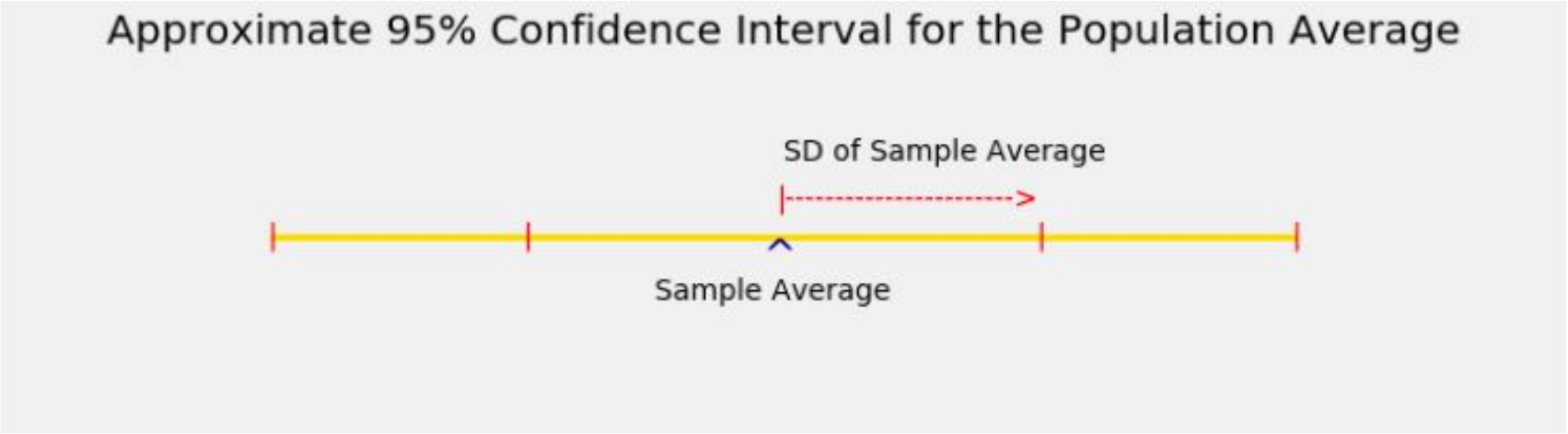
  = (population SD) / $\sqrt{\text{sample size}}$

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



Width of the Interval

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

= 4 * SD of the sample average

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