

# YData: An Introduction to Data Science

## Lecture 27: Sample Averages

Elena Khusainova & John Lafferty  
Statistics & Data Science, Yale University  
Spring 2021

Credit: [data8.org](https://data8.org)



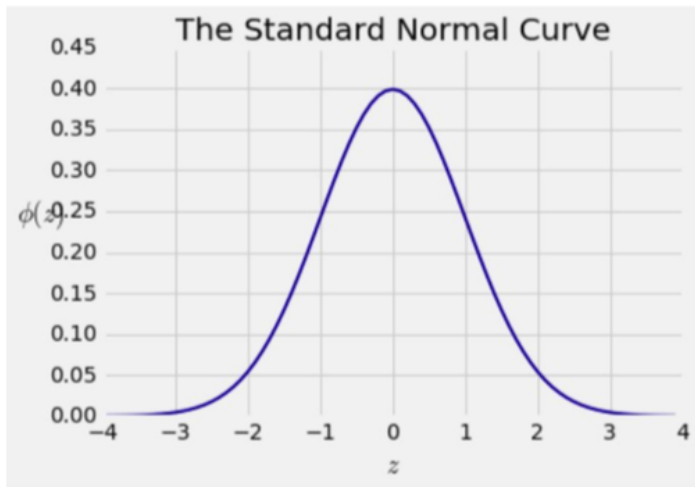
# Announcements

- Project 2 checkpoint is this Friday, 4/9
- For the checkpoint complete the first two parts – answers can be revised and changed later!
- If you are working with a partner, don't forget to indicate that on your submission

# Questions for This 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?

# Bell Curve



# Bounds and Normal Approximations

<b>Percent in Range</b>	<b>All Distributions</b>	<b>Normal Distribution</b>
average $\pm$ 1 SD	at least 0%	about 68%
average $\pm$ 2 SDs	at least 75%	about 95%
average $\pm$ 3 SDs	at least 88.888...%	about 99.73%

# Sample Averages

- The Central Limit Theorem describes how the normal distribution (a bell-shaped curve) arises in the context of random sampling.
- Many distributions we observed were not bell-shaped, but empirical distributions of sample averages were.
- We care about sample averages because they estimate population averages.

# 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.

# Shape of the Distribution

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

(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?

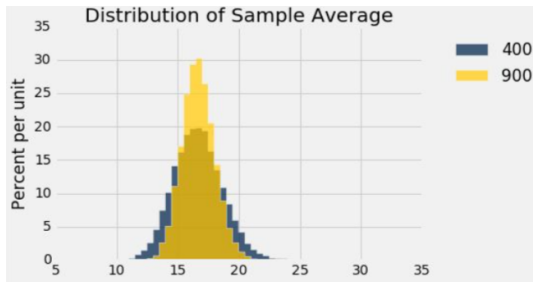
- Along with the center, the spread helps identify exactly which normal curve is the distribution of the sample average.
- The variability of the sample average helps us measure how accurate the sample average is as an estimate of the population average.
- If we want a specified level of accuracy, understanding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

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

(DEMO)

## 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.

## 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. A random sample of 900 households is taken.

Fill in the blanks and explain: There is about a 68% chance that the average annual income of the sampled households is in the range \$\_\_\_\_\_ plus or minus \$\_\_\_\_\_.