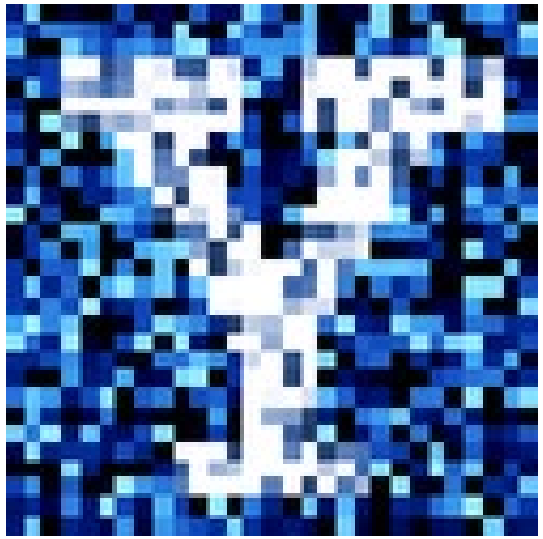
YData: Introduction to Data Science



Lecture 27: sample averages

Overview

Review

Continuation examining the normal distribution

The Central Limit Theorem

Sampling distributions

Announcements

Project 2 is due on Friday

Homework 8 has been posted

- It is due on Sunday

Questions

How can we quantify natural concepts like "center" and "variability" of data?

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?

Review: Measures of central tendency and spread

Measures of central tendency

- The average (or mean)
- The median

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Measures of variability

- The standard deviation

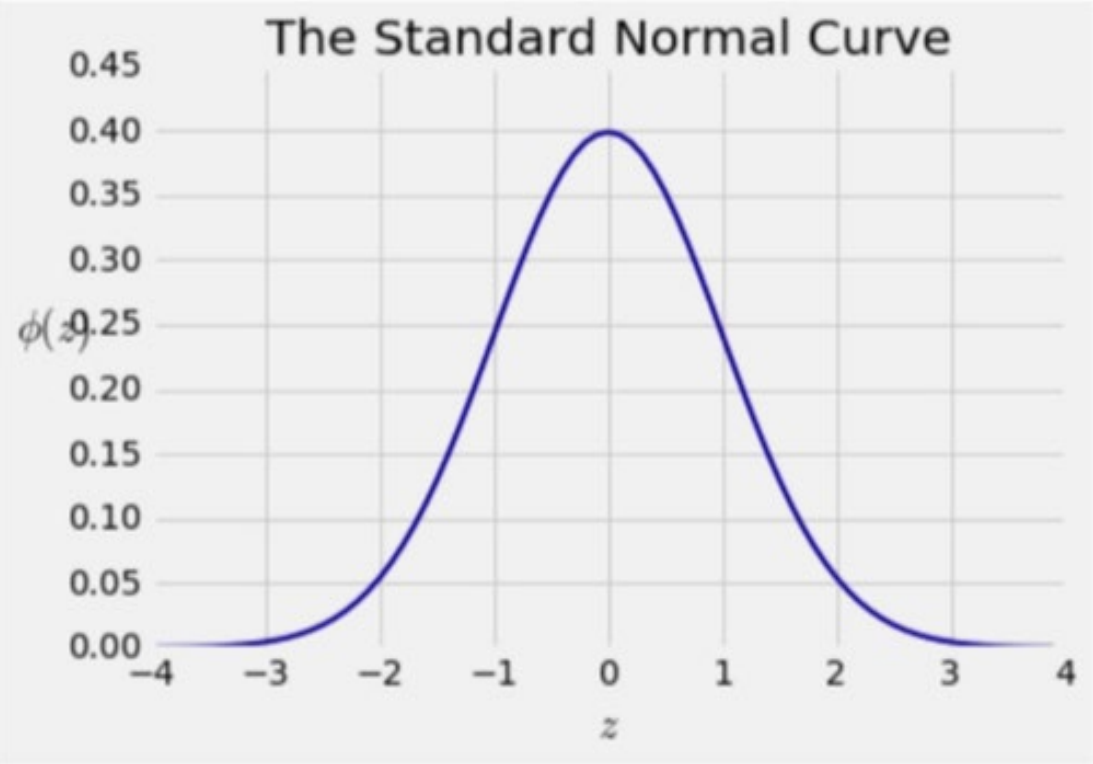
$$SD = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Standardized units allow us to compare data on different scales

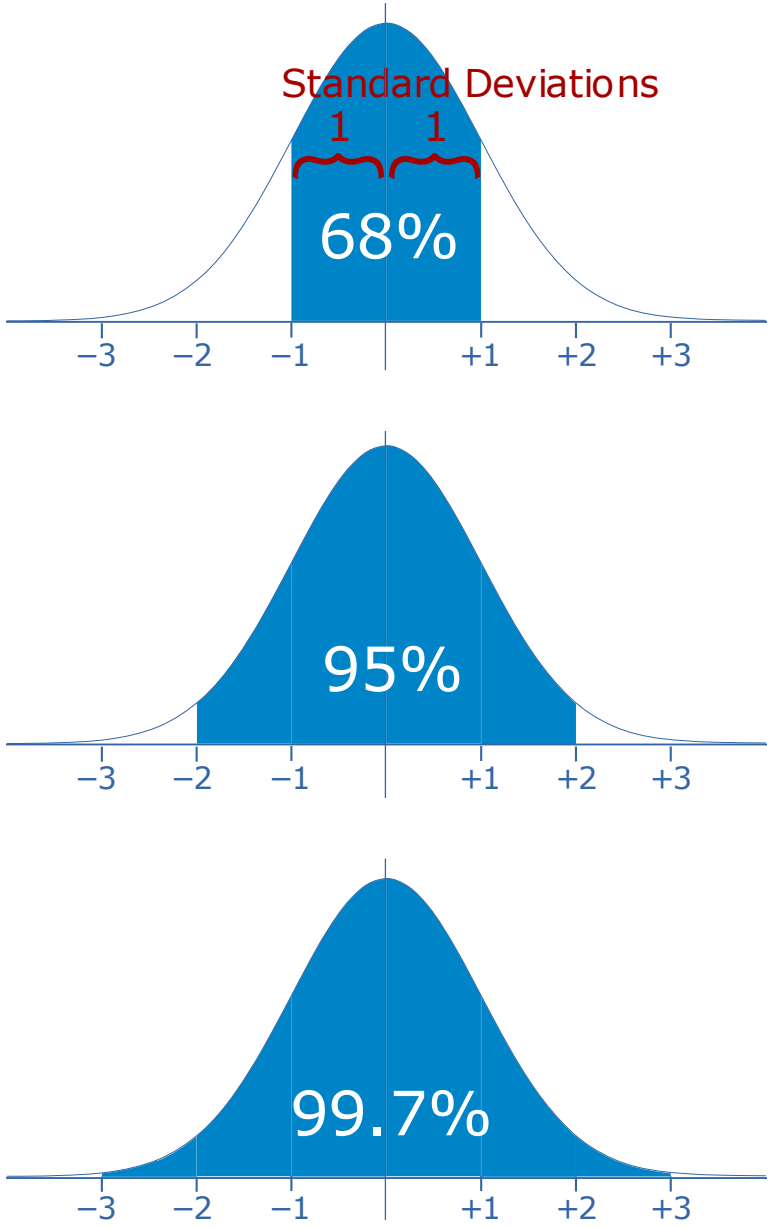
- Z-scores tell us how many standard deviations a point is from the mean

$$\text{z-score}(x_i) = \frac{x_i - \bar{x}}{SD}$$

The standard normal curve



$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$



Bounds and normal approximations

Chebyshev's Inequality: 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"

Percent in Range	All Distributions	Normal Distribution
Average ± 1 SDs	at least 0%	About 68%
Average ± 2 SDs	at least 75%	About 95%
Average ± 3 SDs	at least 88.88%	About 99.73%

Let's explore this in Jupyter!

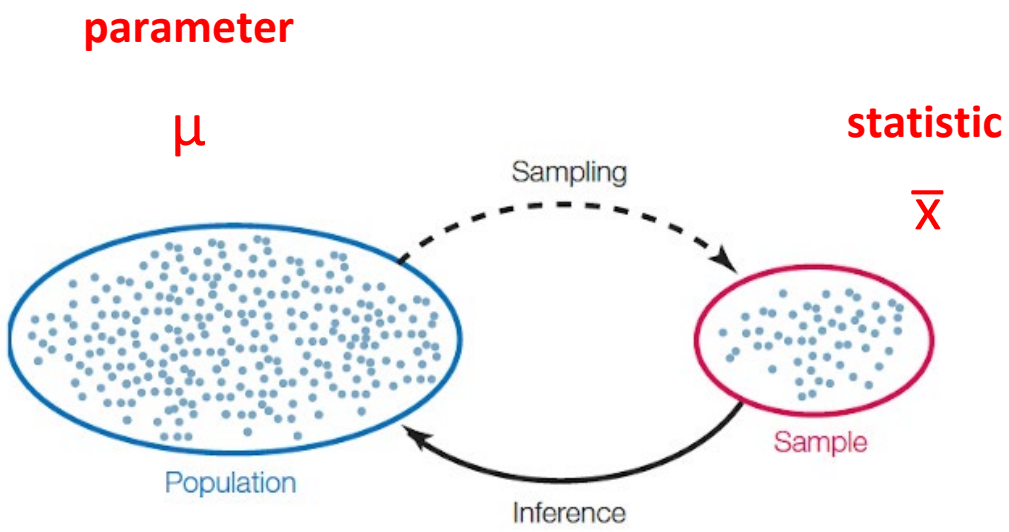
The Central Limit Theorem

A distribution of sample averages…

If you have only one random sample, then there is only one sample average ($\bar{x}$).

But the sample could have come out differently

- And then the sample average might have been different
- So, theoretically, we can think about many potentially possible sample averages that could have occurred from the same population.



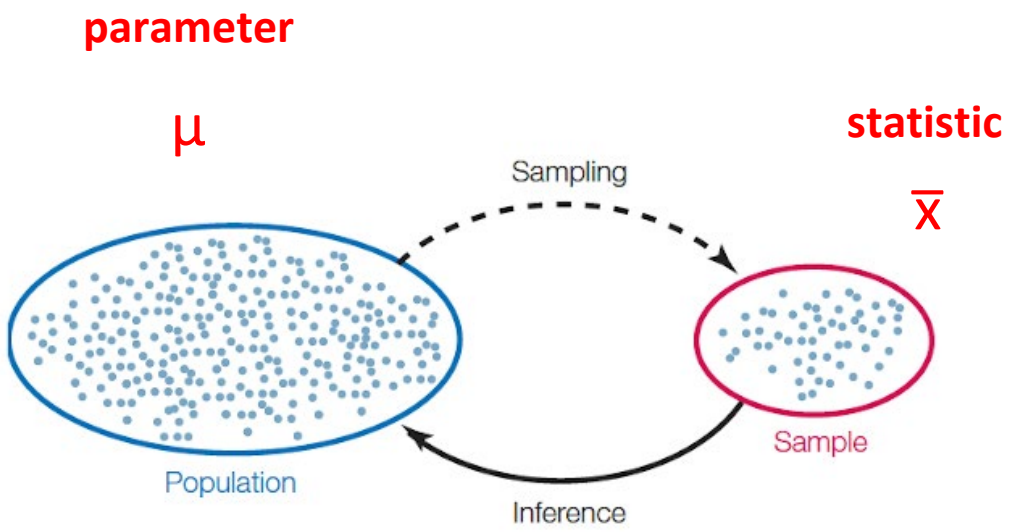
Distribution of the sample average

Imagine all possible random samples of the same size as yours (size n)

- 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



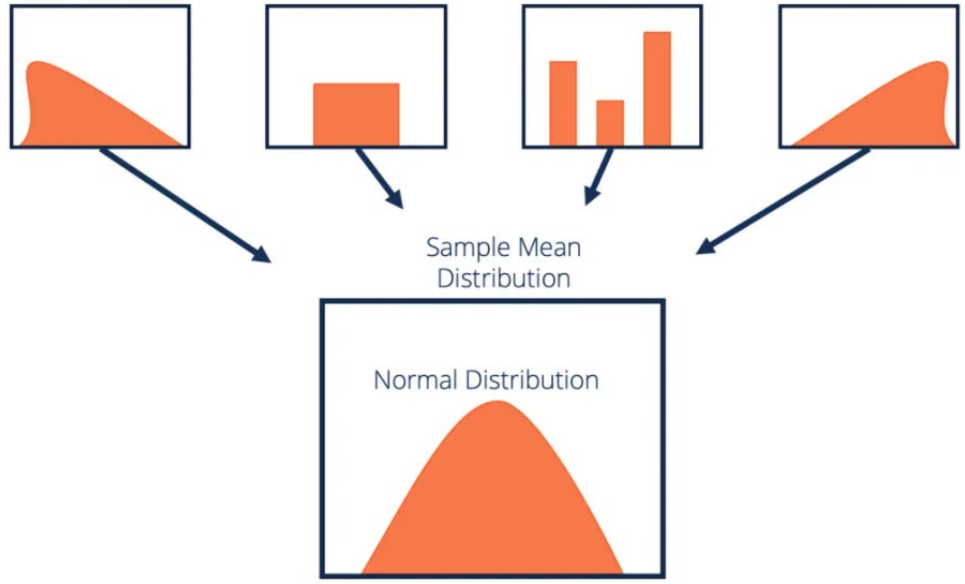
The Central Limit Theorem

If the sample is:

- large, and
- drawn at random with replacement….

Then, *<u>regardless of the distribution of the population</u>*,

the probability distribution of the sample sum (or of the sample average) is roughly normal



Let's explore this in Jupyter!

Sampling distributions

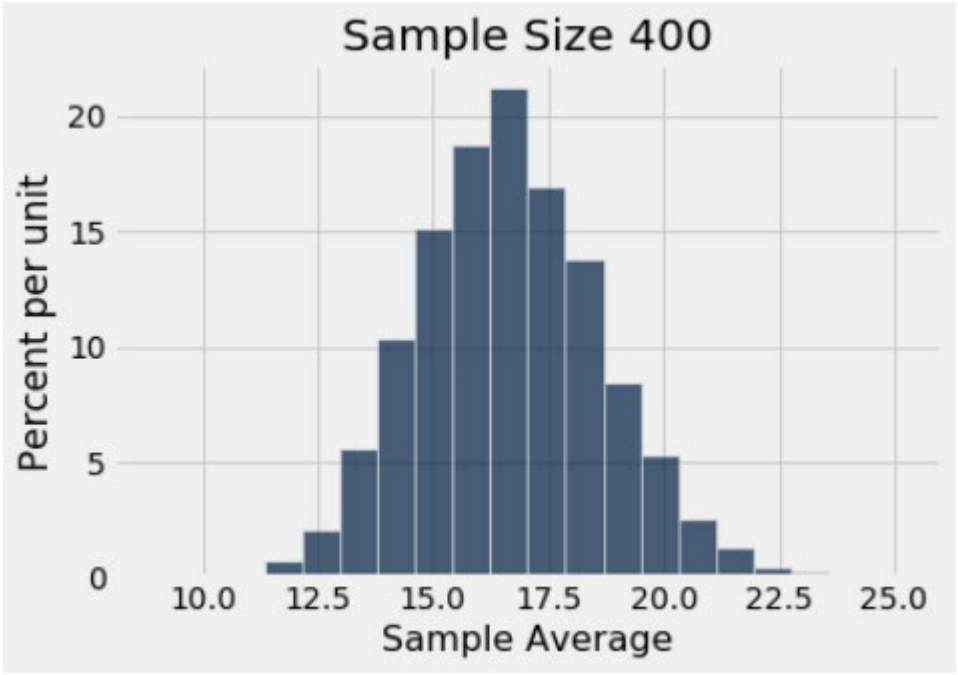
Specifying the sampling distribution

Suppose the random sample is large

We have seen that the "sampling distribution" of the sample average is roughly bell shaped

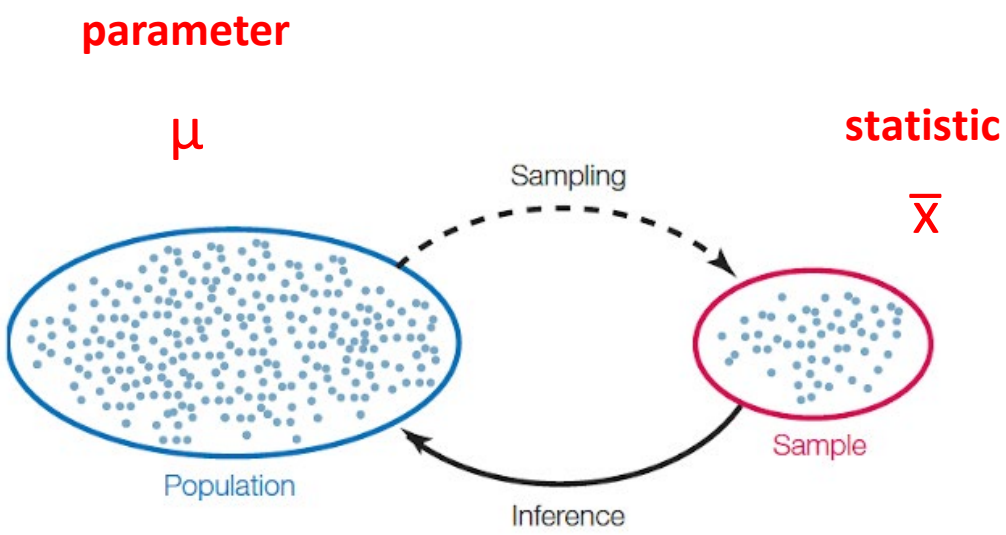
Important questions remain:

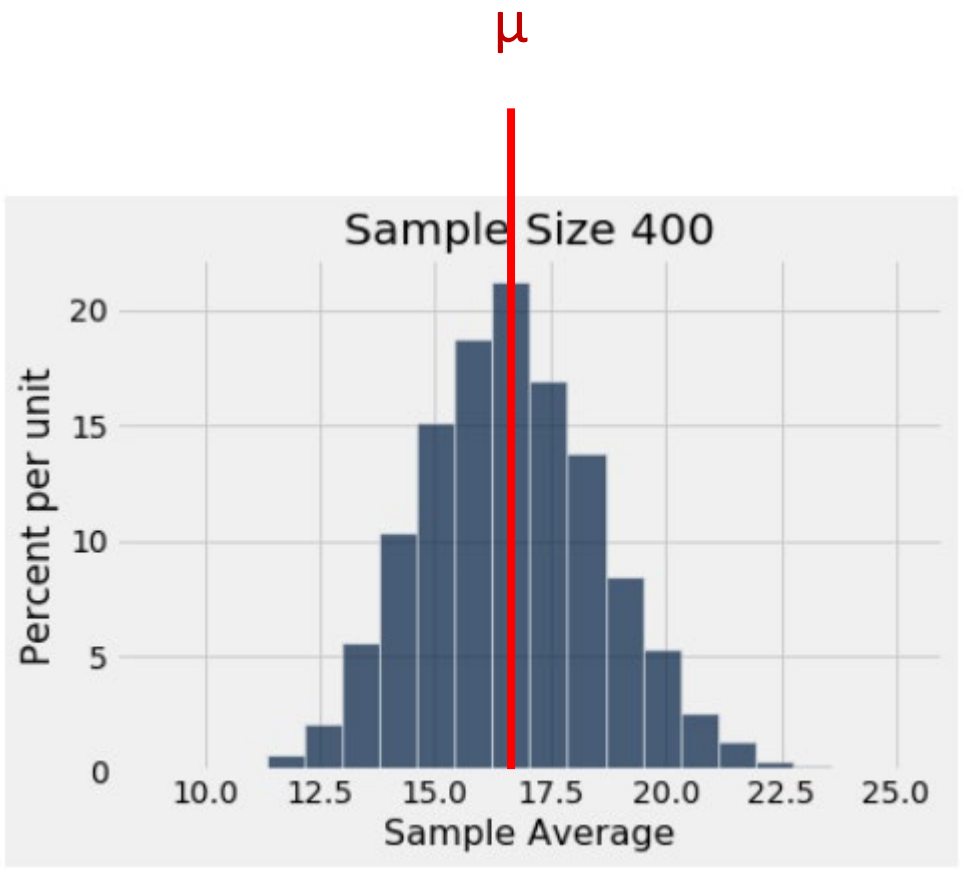
- Where is the center of that bell curve?
- How wide is that bell curve?



The center of the sampling distribution

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



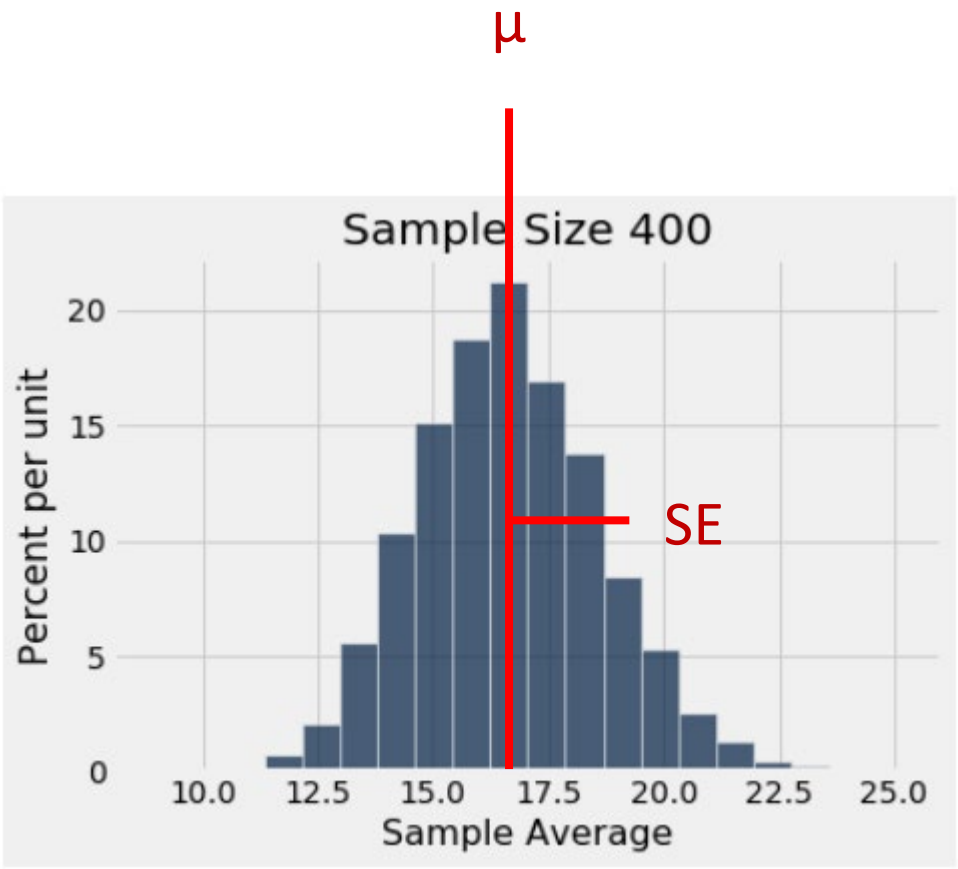


The variability of the sample average

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.



Let's explore this in Jupyter!

Discussion question

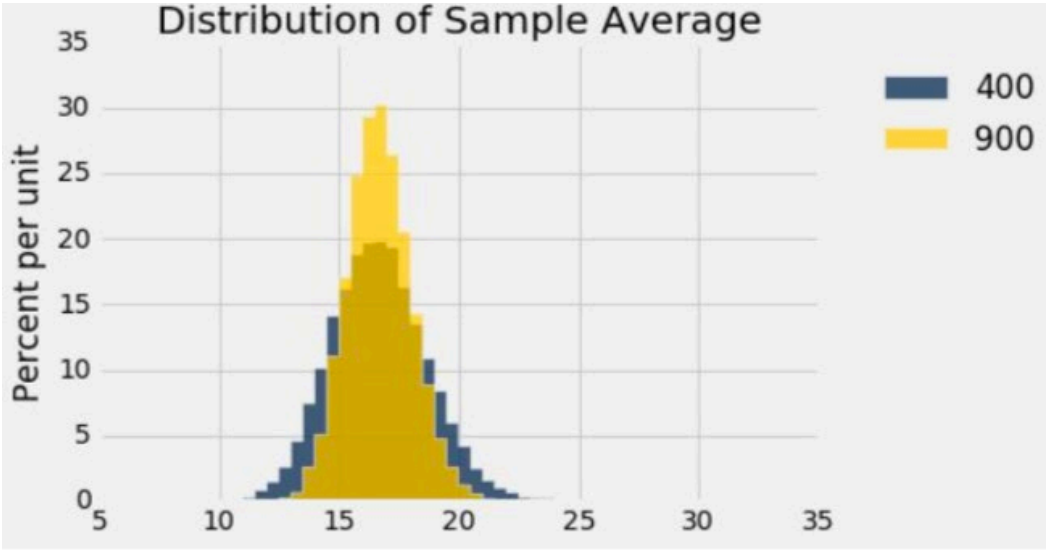
The gold histogram shows the sampling distribution of n = _____ values, each of which is _____.

(a) 900

(b) 10,000

(c) a randomly sampled flight delay

(d) an average of flight delays



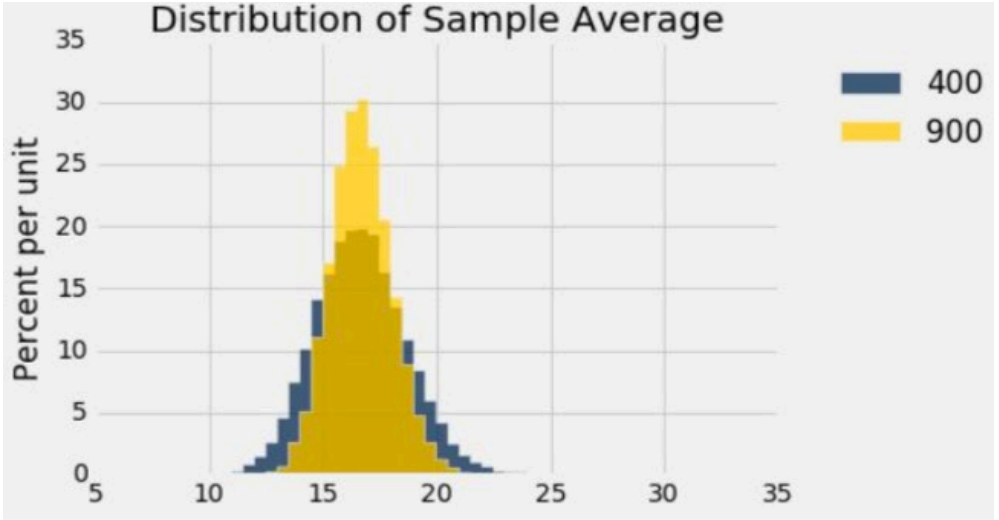
Two approximate sampling distributions

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.



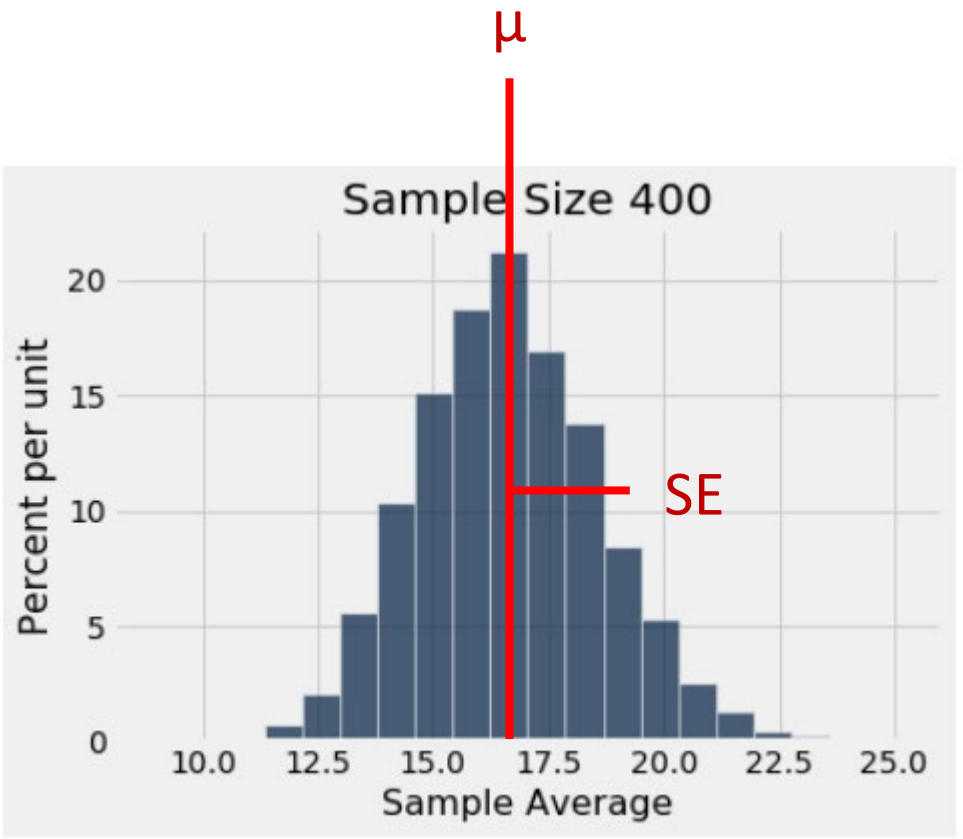
Variability of the sampling distribution

The distribution of all possible sample averages of a given size (n) is called the sampling distribution of the sample average.

We approximate it by an empirical distribution

By the CLT, it's roughly normal:

- Center: the population average (μ)
- Spread: $SE = \frac{population\ SD}{\sqrt{sample\ size}}$



Let's explore this in Jupyter!

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 $_____ .