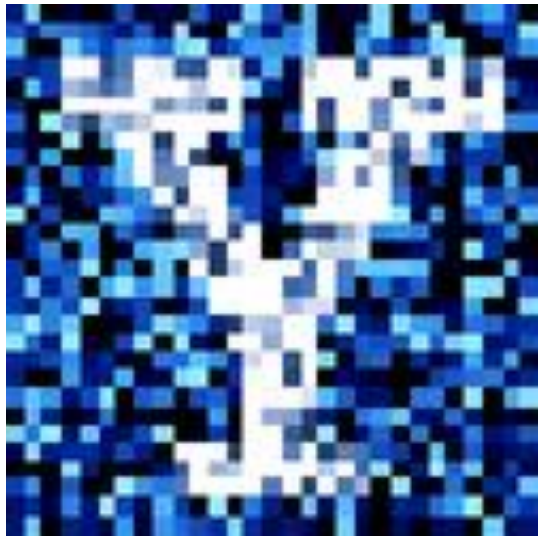
YData: Introduction to Data Science



Lecture 28: designing experiments

Overview

Review: The Central Limit Theorem

Sampling distributions

Confidence intervals for a mean revisited

If there is time:

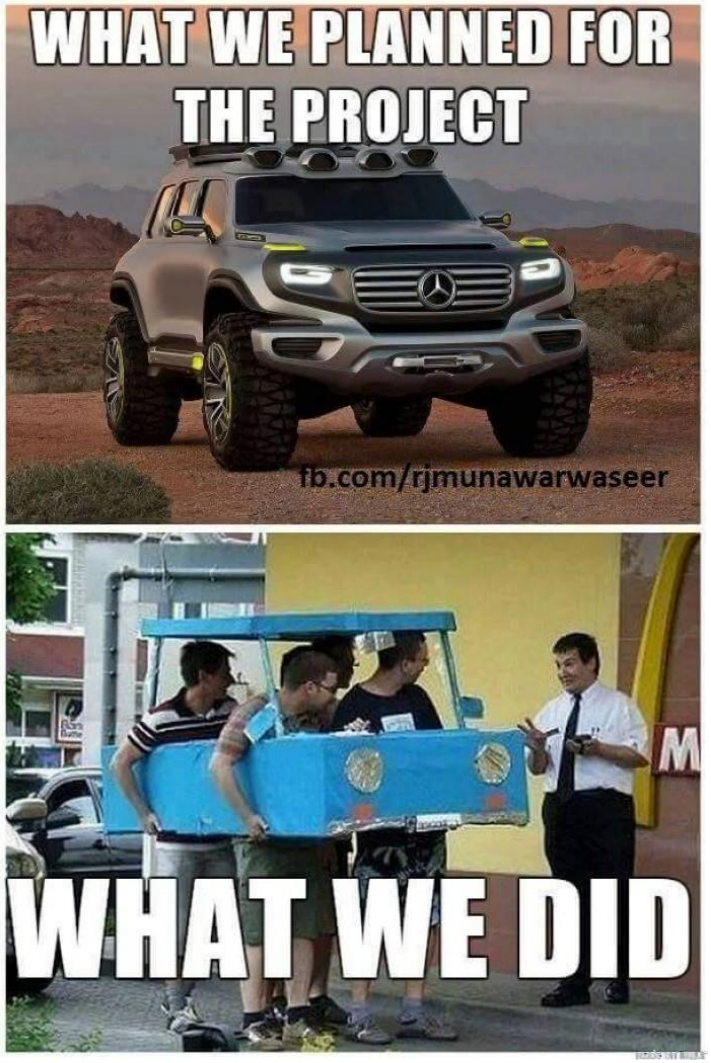
- Confidence intervals for proportions

Announcements

Project 2 is due on Friday

Homework 8 has been posted

- It is due on Sunday



Questions

- [x] 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: The Central Limit Theorem

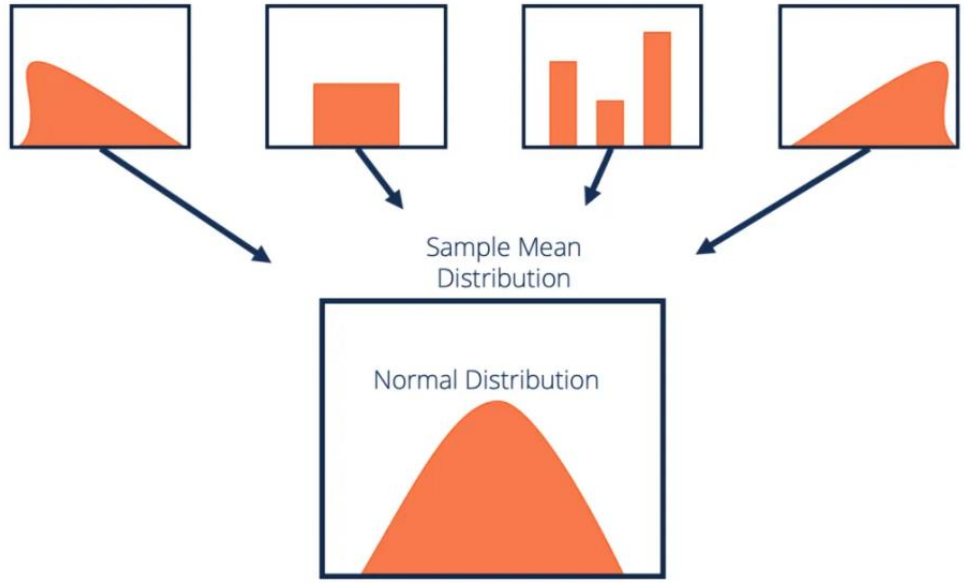
The Central Limit Theorem

If the sample is:

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

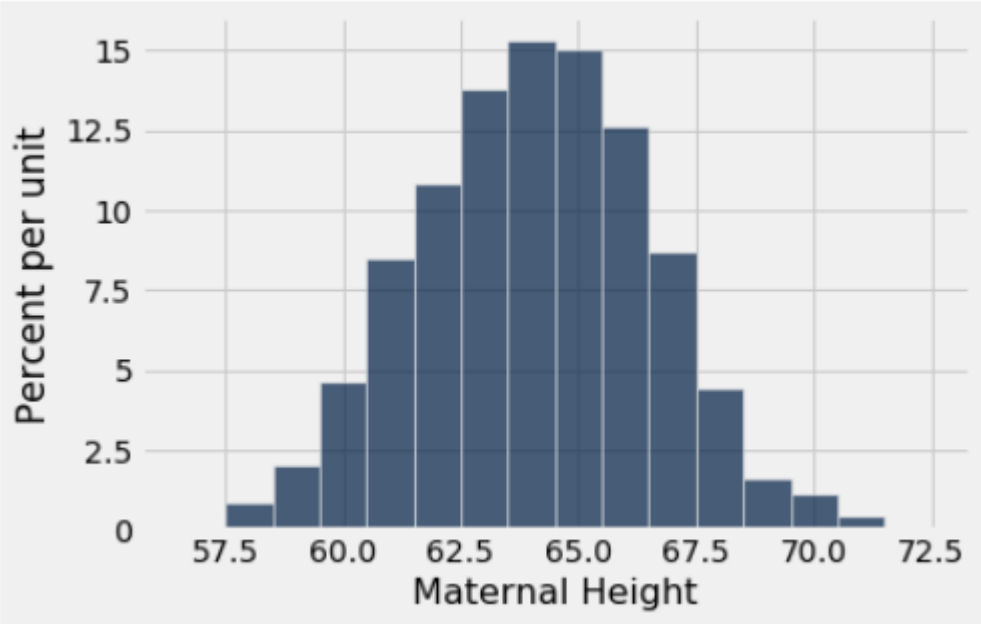
Then, *<u>regardless of the distribution of the population</u>*,

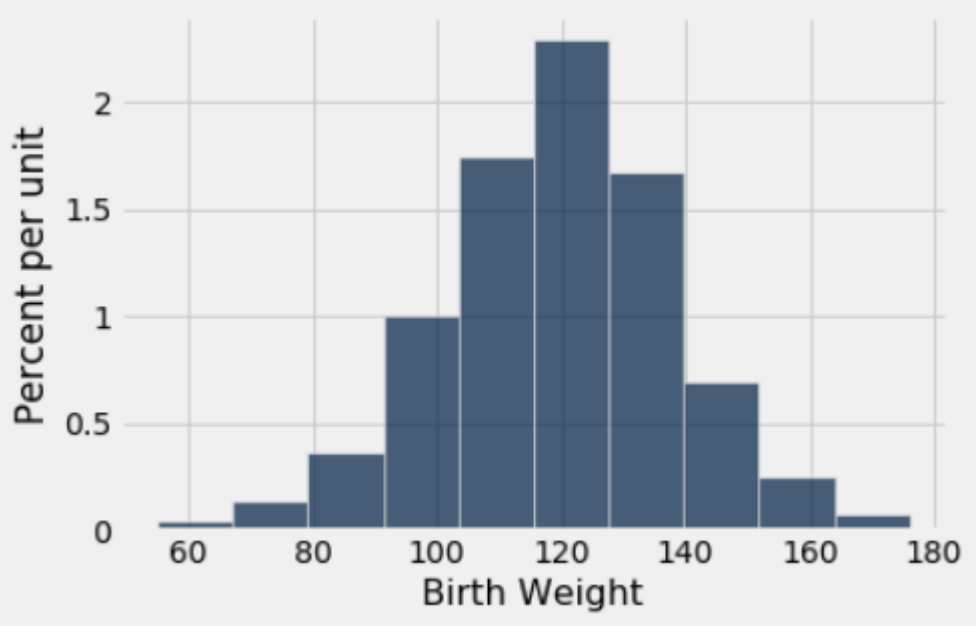
the probability distribution of the sample sum (or of the sample average) is roughly normal



Discussion Question

Q: Why does the distribution of mother heights and baby weights appear to be normal?





Questions

- [x] How can we quantify natural concepts like "center" and "variability" of data?
- [x] 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?

Sampling distributions

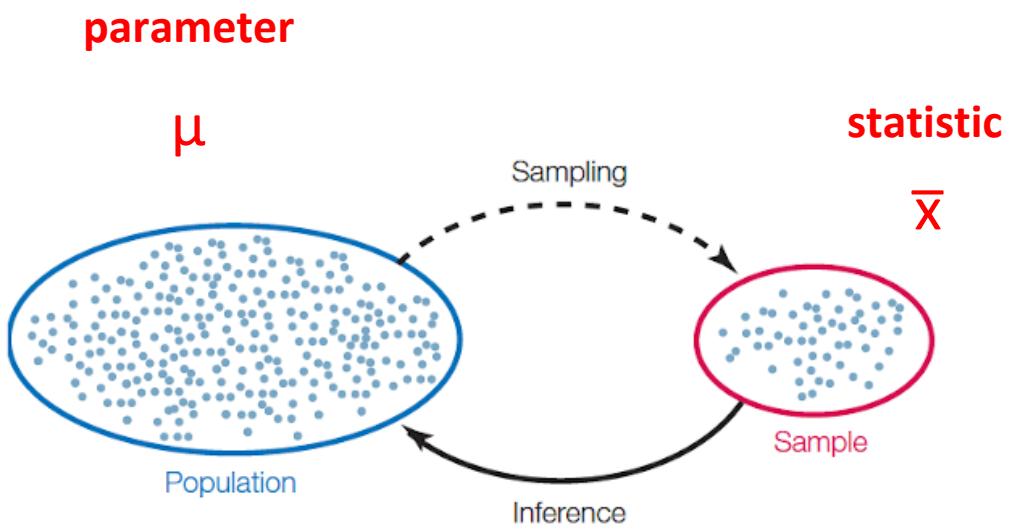
Distribution of the sample average

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

- There are lots of them

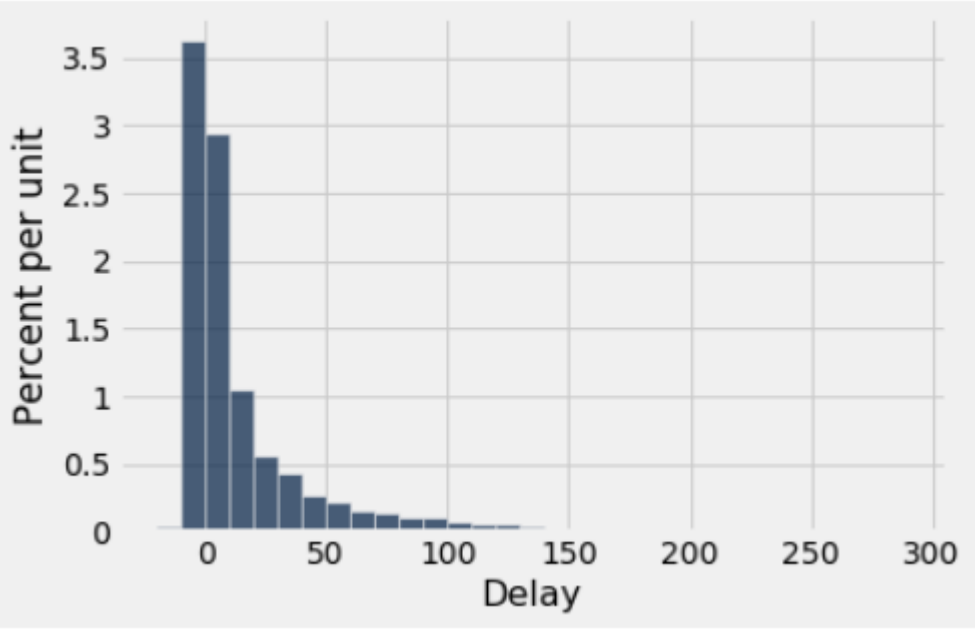
Each of these samples has an average

The distribution of the sample average is the distribution of the averages of all the possible samples



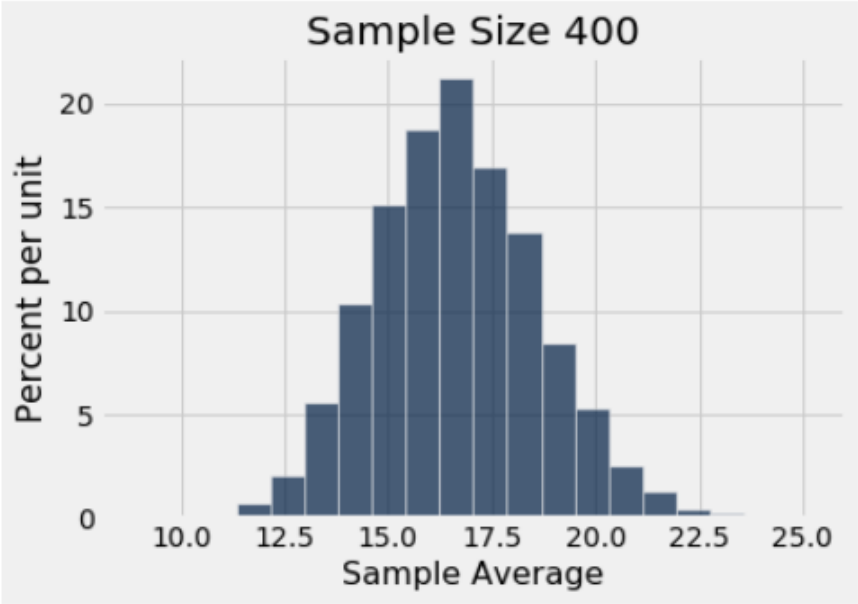
The Central Limit Theorem: Flight delays

Flight delays population



Take many sample of size n = 400 and calculate means ($\bar{x}$'s)

Flight delays sampling distribution



Specifying the sampling distribution

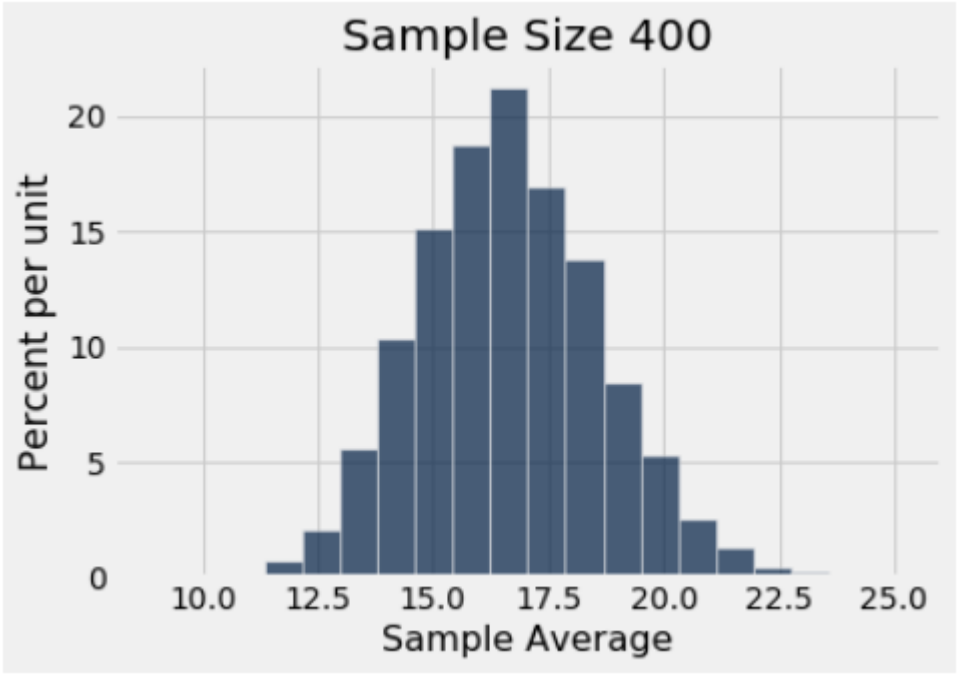
Suppose the random sample is large

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

Important questions remain:

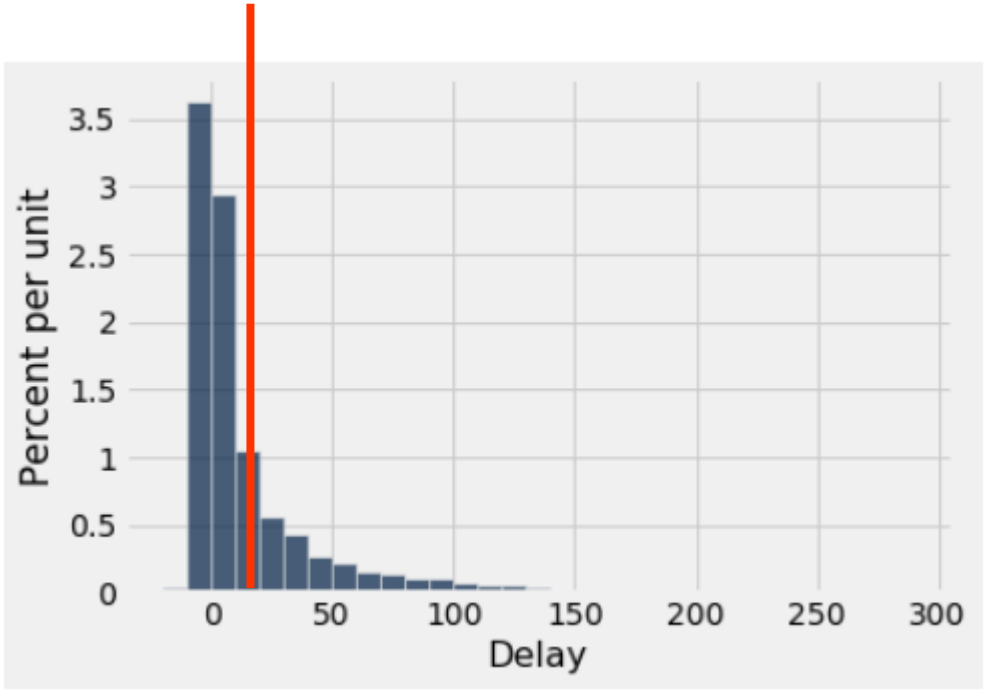
- Where is the center of that bell curve?
- How wide is that bell curve?

Flight delays sampling distribution



The Central Limit Theorem: Flight delays

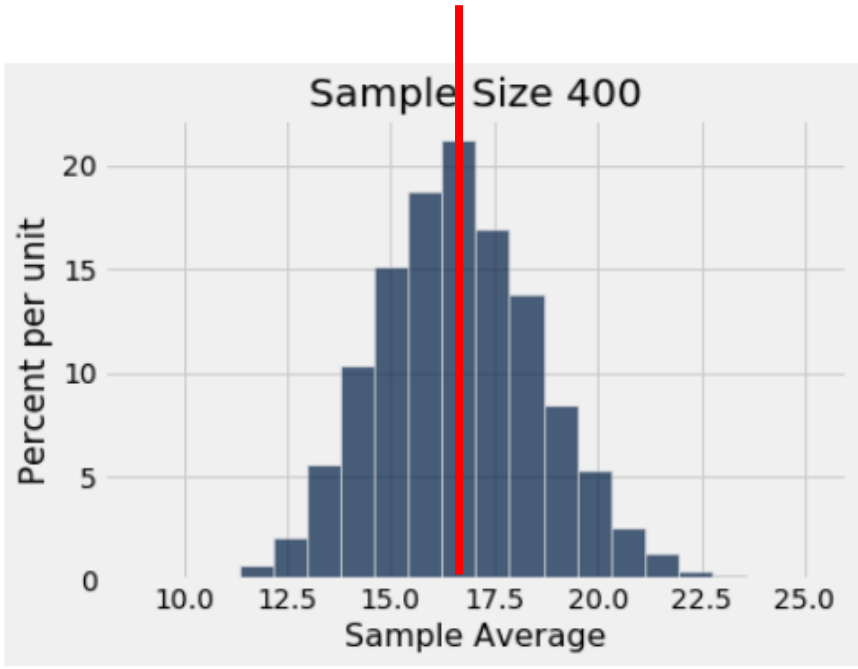
μ = 16.7



Flight delays population

Take many sample of size n = 400 and calculate means ($\bar{x}$'s)

μ = 16.7



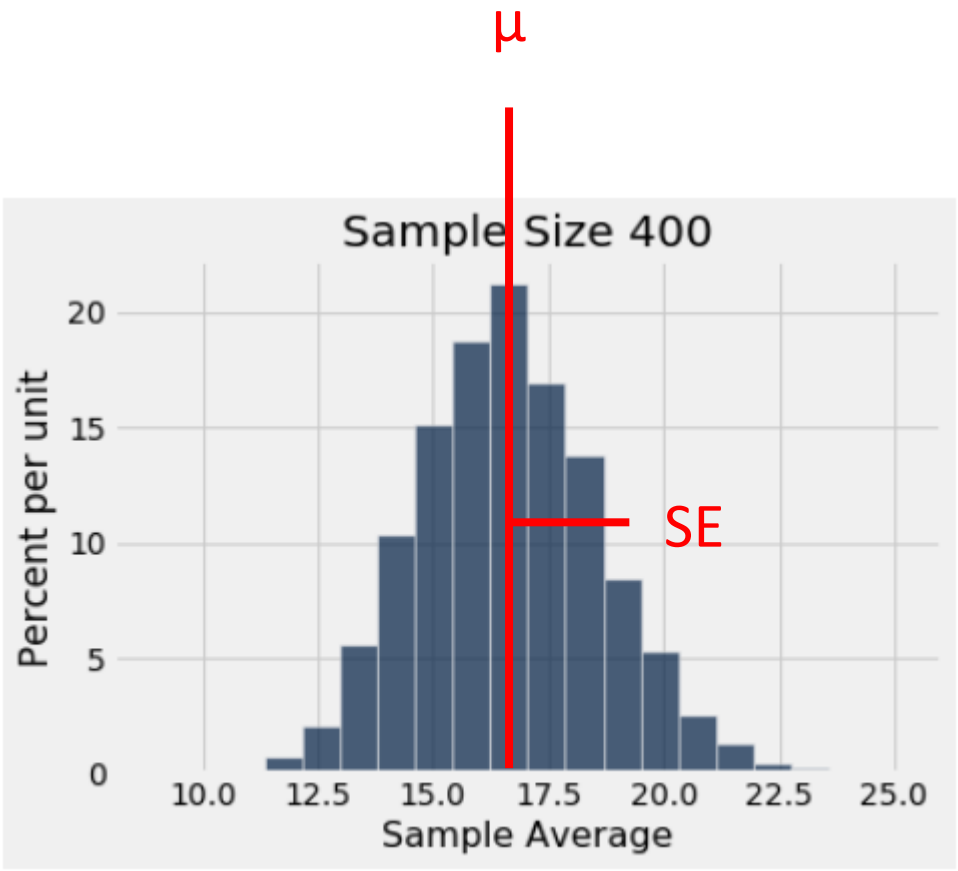
Flight delays sampling distribution

The sampling distribution is ***centered at the population average***

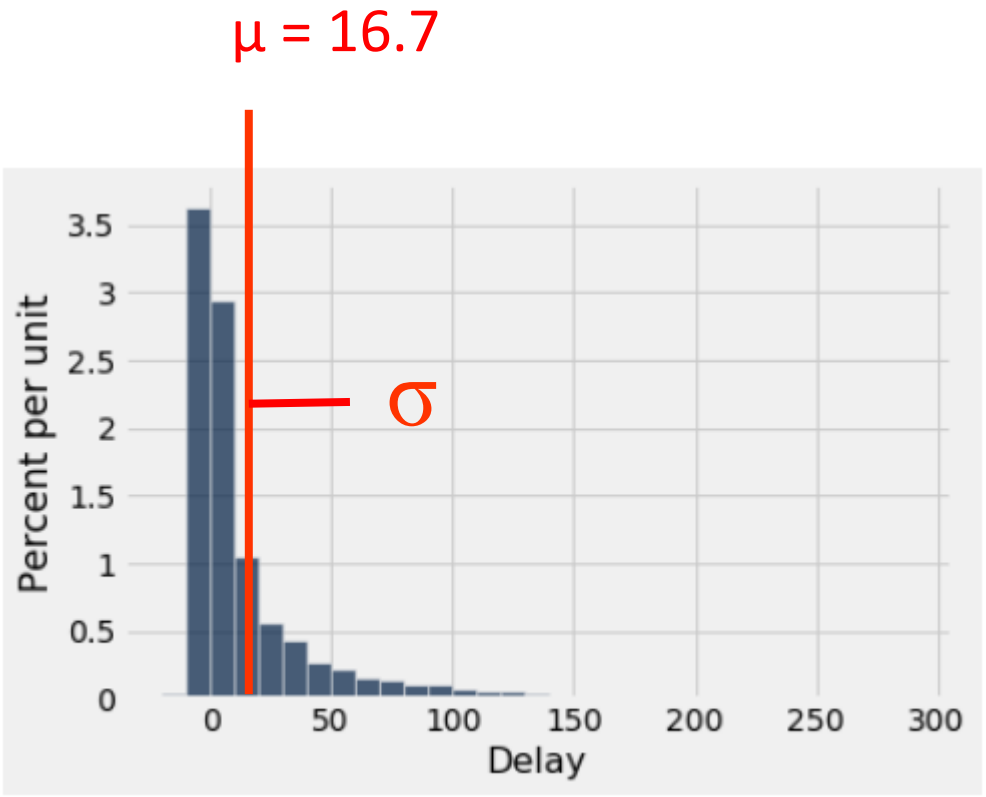
The variability of the sample average

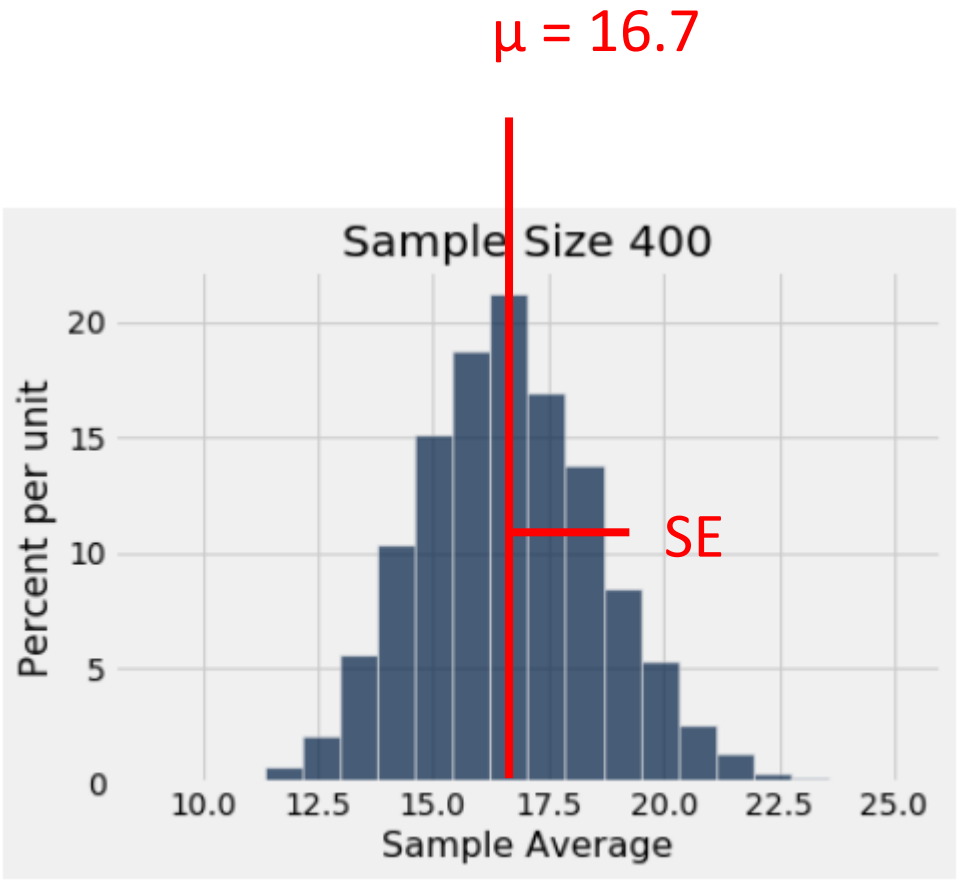
Along with the center, the spread helps identify exactly which normal curve is the distribution of the sample average.

The variability of the sample average helps us measure how accurate the sample average is as an estimate of the population average.



The variability of the sample average





Q: How is the population standard deviation (σ) related to the sampling distribution SE?

Let's explore this in Jupyter!

Discussion question

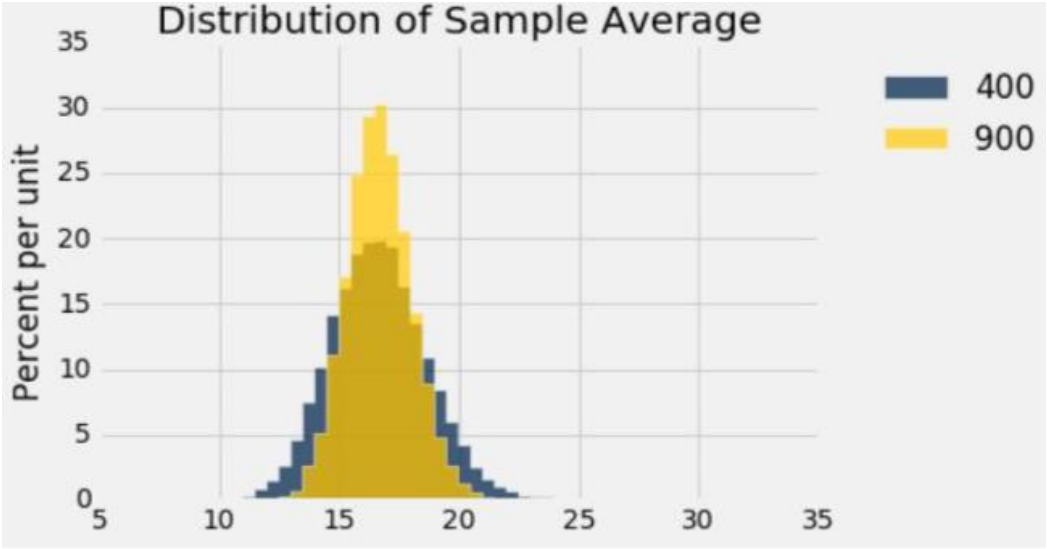
The gold histogram shows the sampling distribution of n = __[a or b]__ values, each of which is __[c or d]__.

(a) 900

(b) 10,000

(c) a randomly sampled flight delay

(d) an average of flight delays



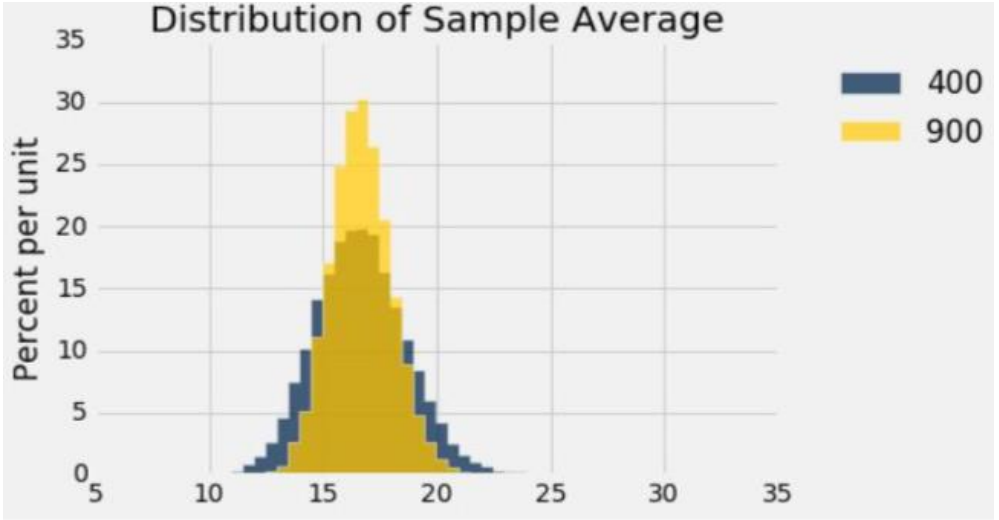
Two approximate sampling distributions

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

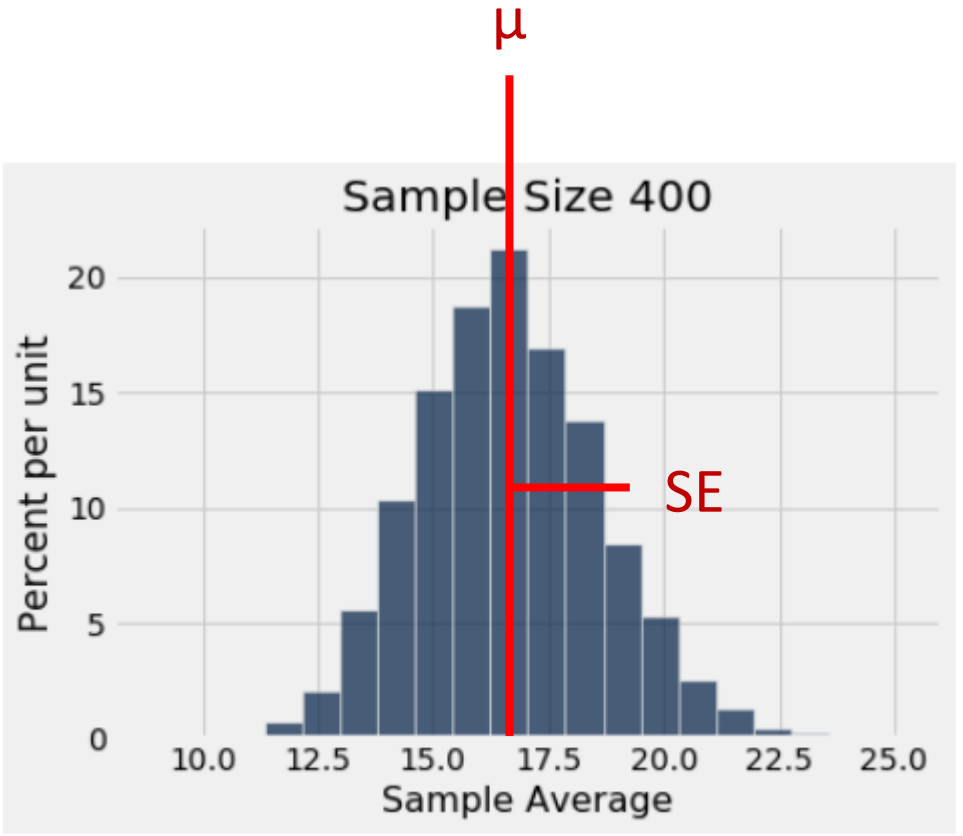
The distribution of all possible sample averages of a given size (n) is called the ***sampling distribution*** of the sample average.

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 $_____ .

Questions

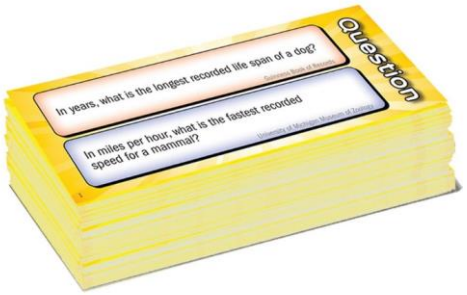
- How can we quantify natural concepts like "center" and "variability" of data?
- Why do many of the empirical distributions that we generate come out bell shaped?
- How is sample size related to the accuracy of an estimate?

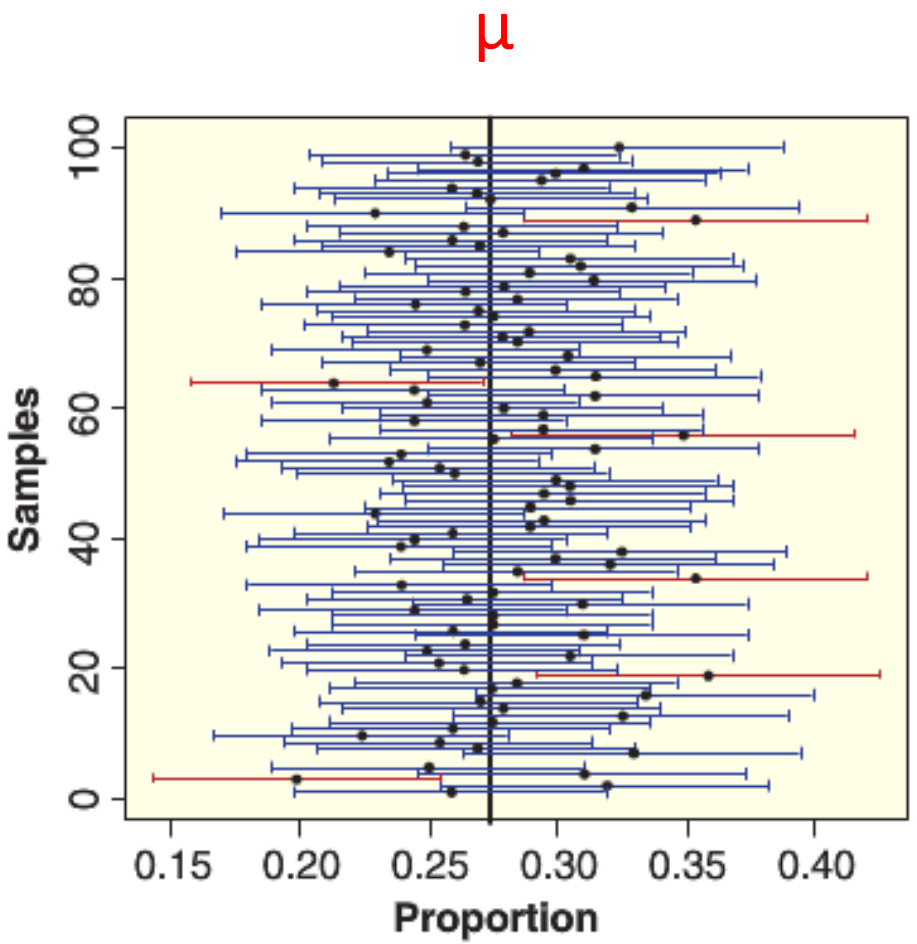
Confidence intervals

Recall: confidence intervals

A **confidence interval** is an interval computed by a method that will contain the *parameter* a specified percent of times

The **confidence level** is the percent of all intervals that contain the parameter

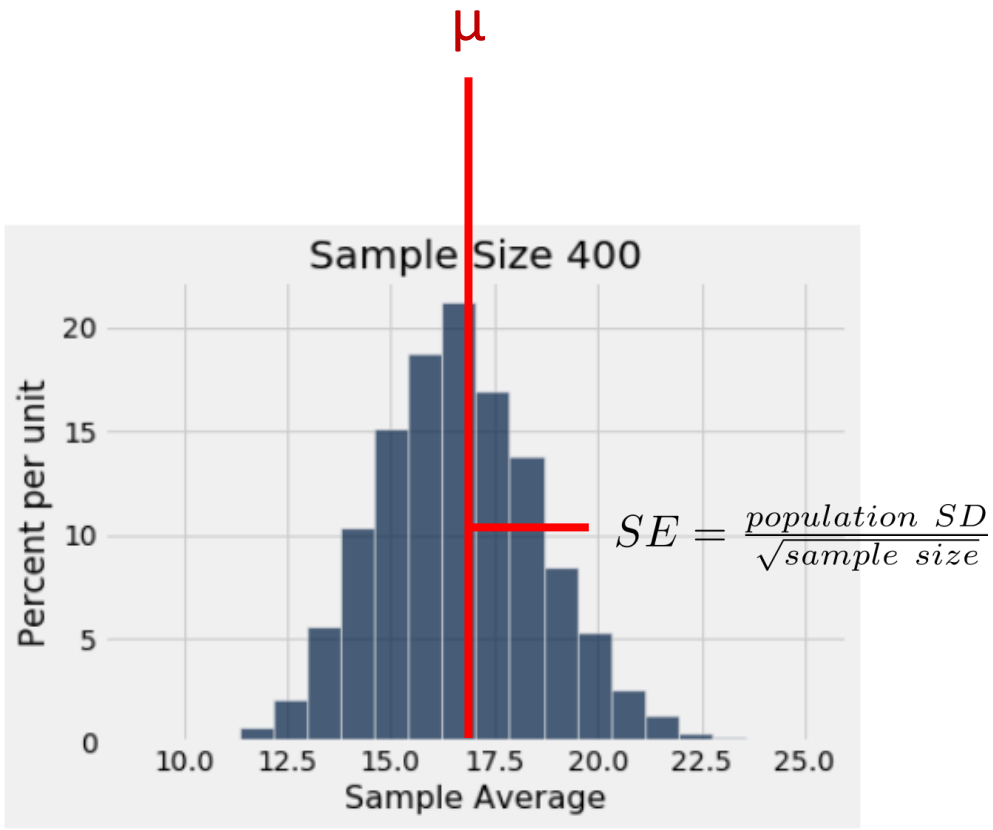




Variability of the sampling distribution

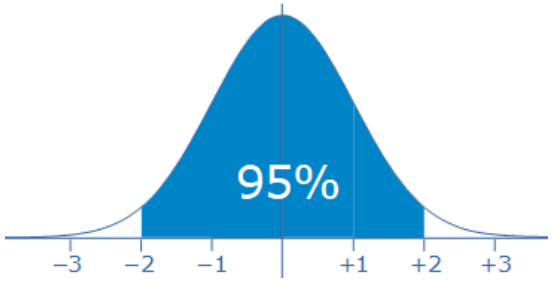
Recall our sampling distribution is roughly normal:

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



What percent of our statistics lie within 2 standard deviations (i.e., 2 SE) of the mean?

- 95% of our statistics in the sampling distribution lie within 2 SE of the mean

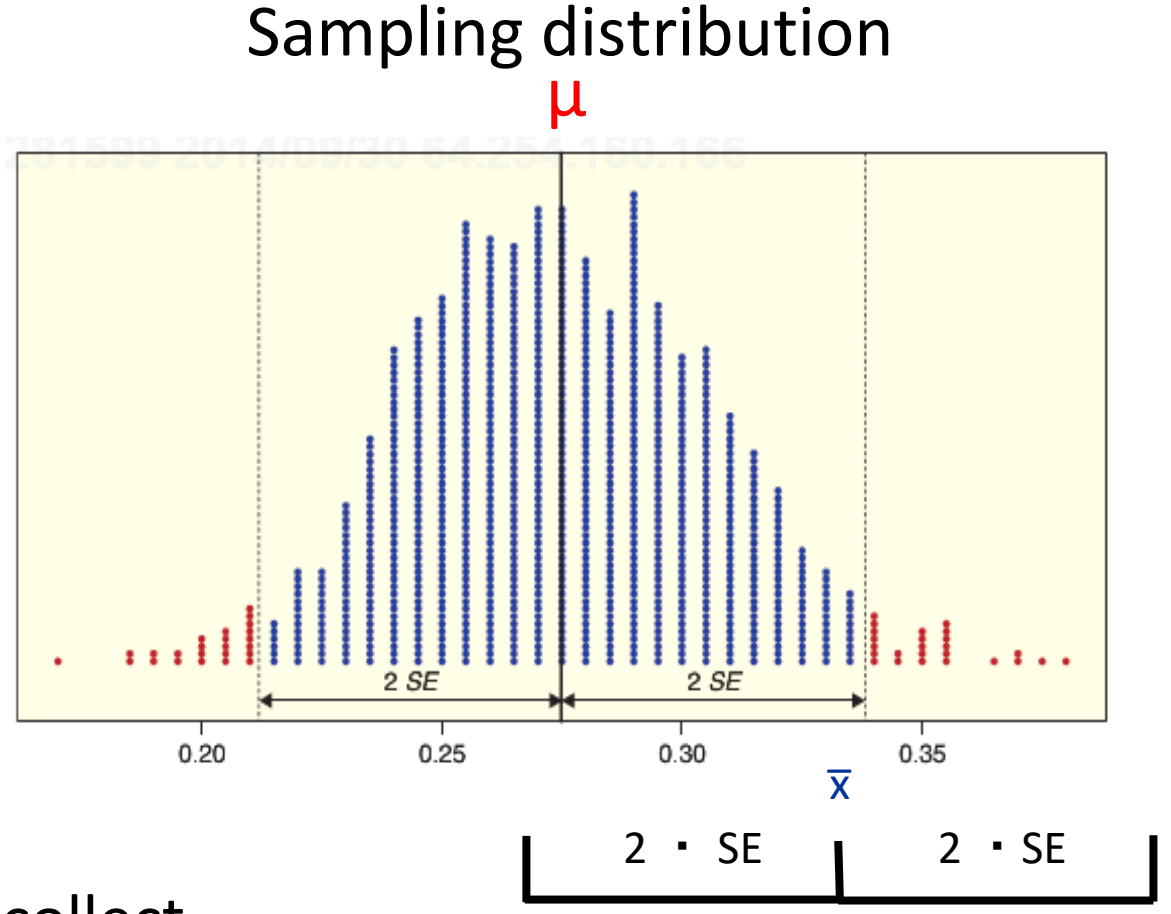


Constructing confidence intervals

We can construct 95% confidence intervals for a population mean μ using:

$$\bar{x} \pm 2 \cdot SE$$

Why does this work?



95% percent of the sample means $\bar{x}$ we collect will be within $\pm\ 2\ \cdot SE$ of the population mean μ

- So $\bar{x} \pm 2 \cdot SE$ will overlap with μ 95% of the time

Let's explore this in Jupyter!

Sample proportions

Proportions are averages

Suppose we had the following data and we wanted to calculate the proportion of cats ($\hat{p}_{cat}$):

Categorical data: "dog", "cat", "fish", "dog", "cat", "dog", "cat", "cat", "fish", "dog"

We can code data: 0 1 0 0 1 0 1 1 0 0

We can calculate the proportion based on taking the average of the coded data

Since we are dealing with averages, the central limit theorem applies!

A conservative estimate for the SE is: $SE = \frac{.5}{\sqrt{n}}$

Let's explore this in Jupyter!