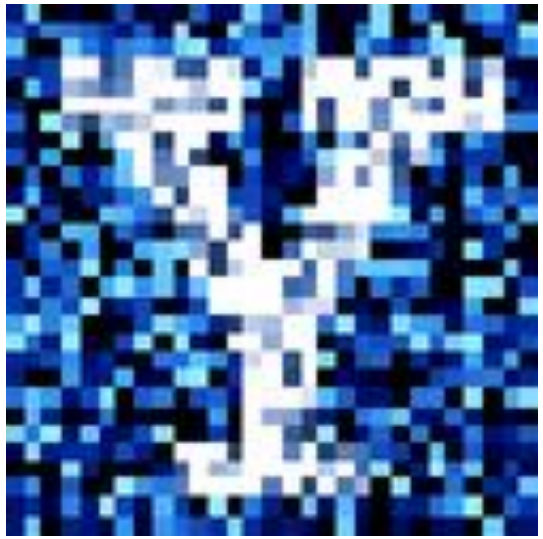
YData: Introduction to Data Science



Lecture 24: The bootstrap and confidence intervals

Overview

Quick review of percentiles

Estimation

The Bootstrap

Confidence intervals

If there is time:

- Interpreting confidence intervals
- Connections between confidence intervals and hypothesis tests

Announcements

Project 2 has been posted!

- It is due on Friday, April 8th

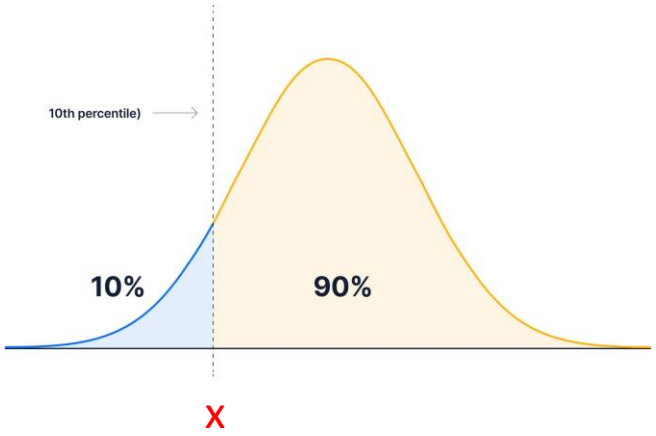
Homework 7 has is optional to give you more time to work on your project

- Homework 8 (next week's homework) is not optional, so try to finish project 2 early!

Review Percentiles

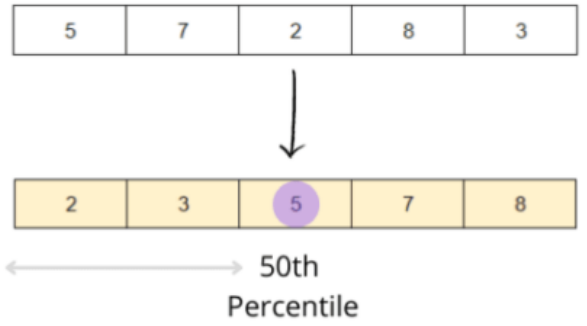
The percentile function

The p^{th} percentile is the smallest value in a set that is as large or larger than p% of the elements in the set



Function in the datascience module: percentile(p, values)

- p: a number between 0 and 100
- values: an array or list of values



For a percentile that does not exactly correspond to an element, take the next greater element instead

- sidenote: percentile functions can be defined slightly differently, but this is the definition used in the datascience package

Estimation

Inference: Estimation

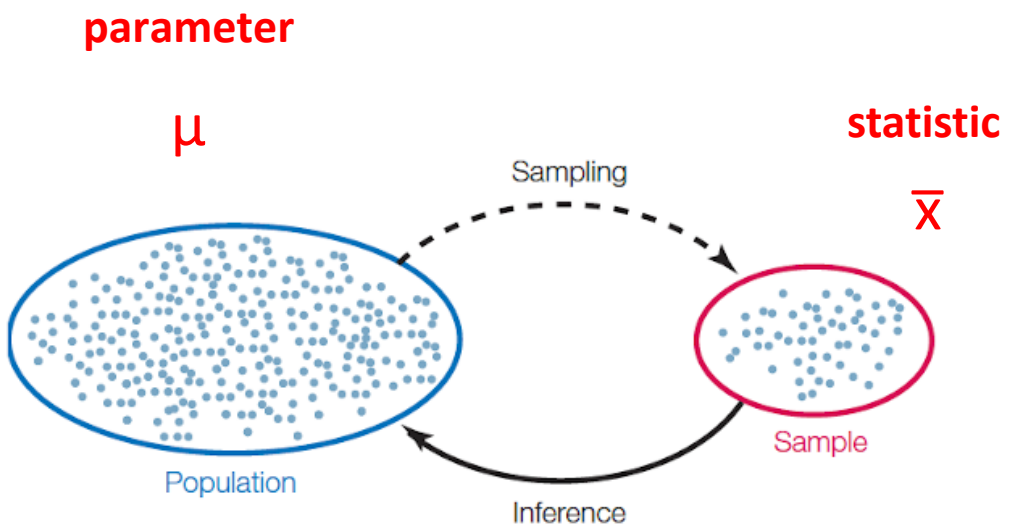
What is the value of an unknown parameter?

If you have data on the whole population:

- Just calculate the parameter value and you're done

If you only have a random sample from the population

- Use a statistic as a point estimate of the parameter
  - Best guess at the parameter value



Let's explore this in Jupyter!

Variability of the estimate

One sample → One estimate

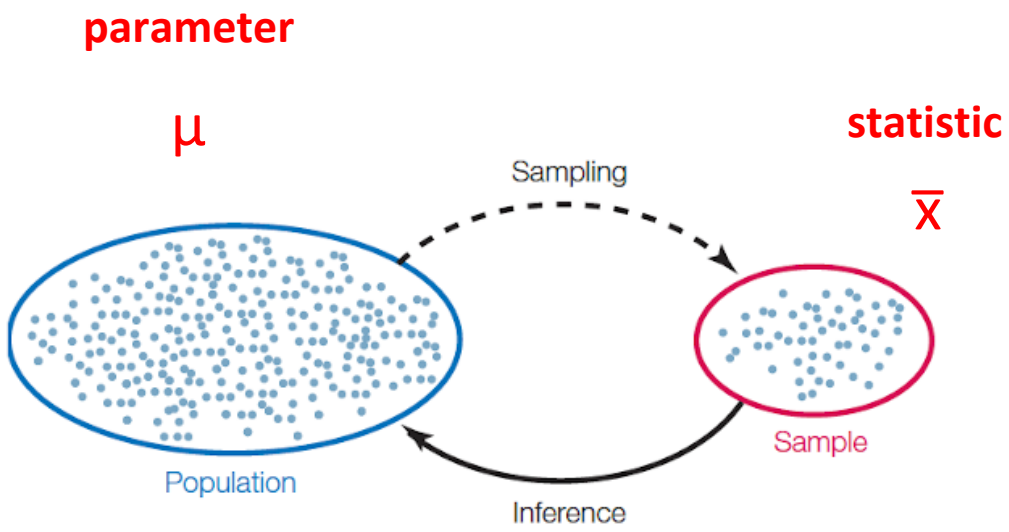
But the random sample could have come out differently

- And so the estimate could have been different

Main question: How different could the estimate have been?

The variability of the estimate tells us something about how accurate the estimate is:

estimate = parameter + error



Where to get another sample?

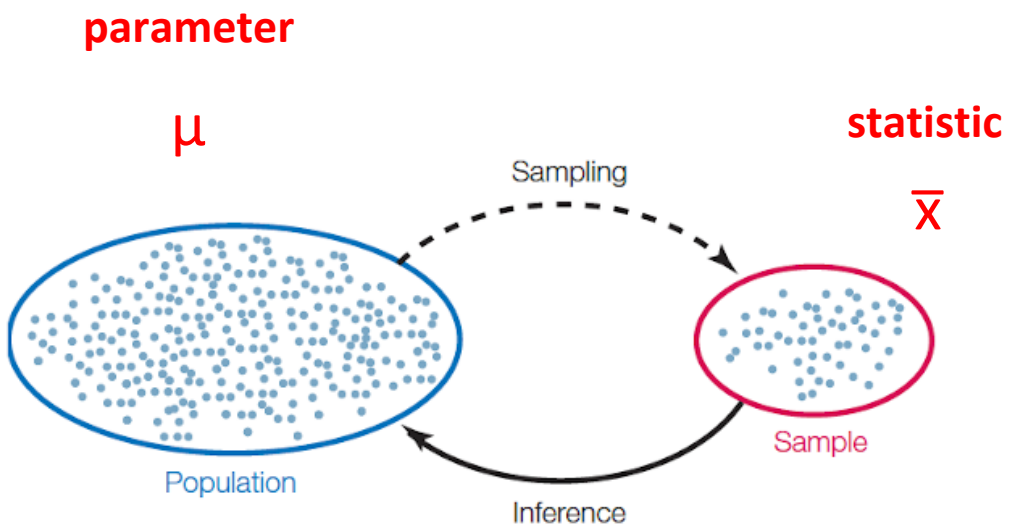
One sample → One estimate

To get many values of the estimate, we needed many random samples

Can't go back and sample again from the population:

- Too costly in terms of time and money

Stuck?



The Bootstrap

The Bootstrap

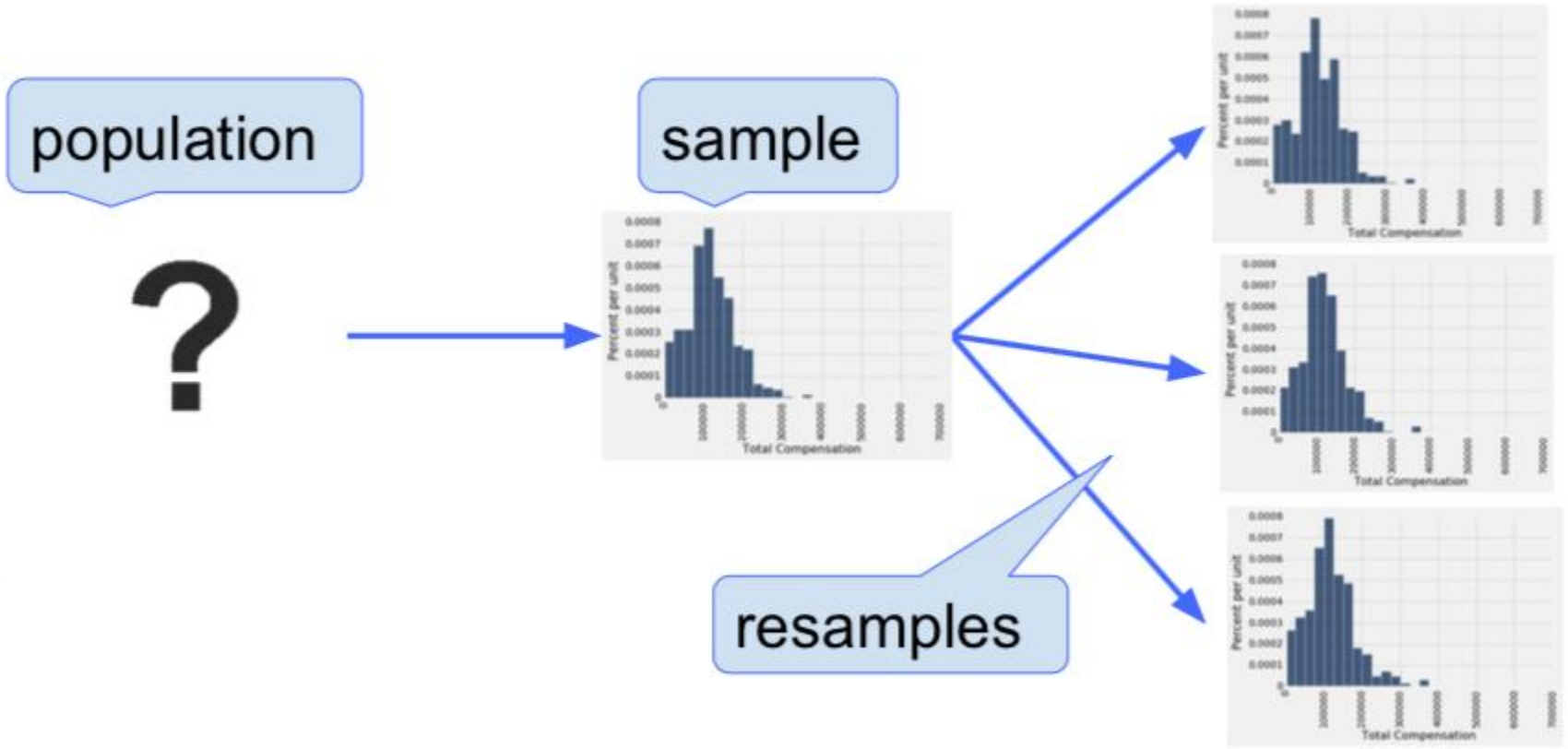
A technique for estimating confidence by simulating repeated random sampling

All that we have is the original sample

- ... which is large and random
- Therefore, it probably resembles the population

So we sample at random from the original sample!

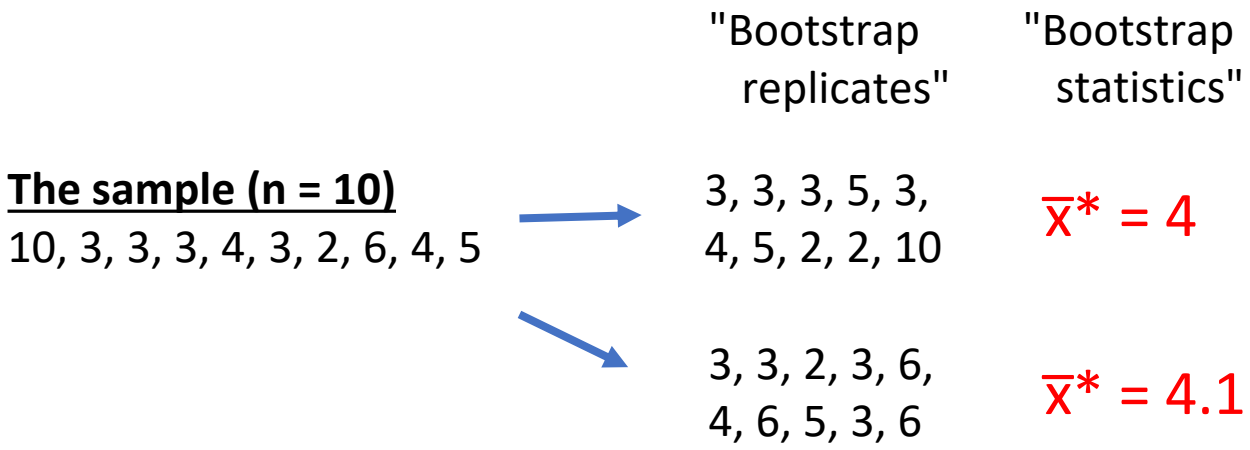
How the Bootstrap works



Key to resampling

From the original sample:

- draw at random
- with replacement
- as many values as the original sample contained



The size of the new sample has to be the same as the original one, so that we are replicating the process of drawing samples from the population

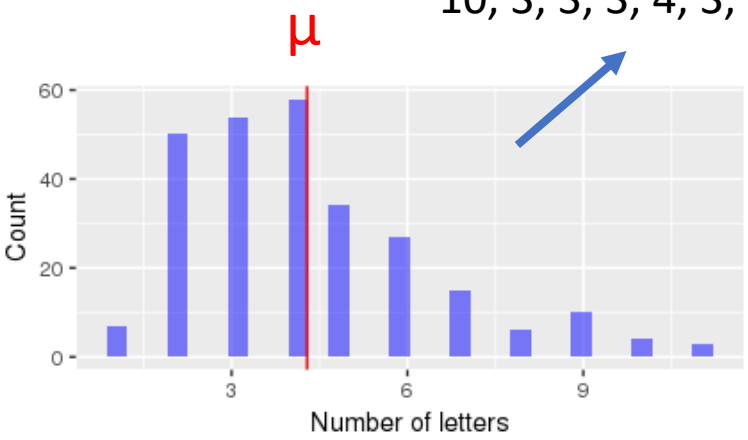
The amount variability across the bootstrap statistics tells us how well our real statistic is at estimating a parameter of interest (e.g., μ)

- Large amount of variability, we can't trust our statistic to be a good estimate for the parameter

Bootstrap distribution illustration

The sample (n = 10)

10, 3, 3, 3, 4, 3, 2, 6, 4, 5



3, 3, 3, 5, 3,
4, 5, 2, 2, 10

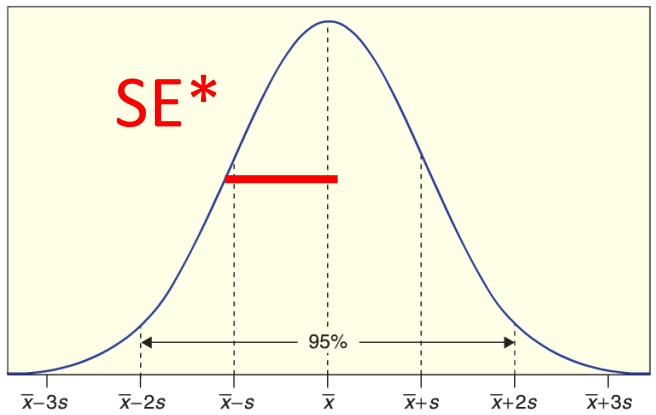
$\bar{x}^* = 4$

3, 3, 2, 3, 6,
4, 6, 5, 3, 6

$\bar{x}^* = 4.1$

5, 3, 2, 3, 3,
3, 10, 3, 4, 3

$\bar{x}^* = 3.9$



Bootstrap distribution!

Let's explore this in Jupyter!

Confidence intervals

Interval estimate based on a margin of error

An **interval estimate** give a range of plausible values for a population parameter.

Example: 42% of American approve of Biden's job performance, plus or minus 3%

How do we interpret this?

Says that the population parameter lies somewhere between 39% to 45%

- i.e., if they sampled all voters the true population proportion would be likely be in this range

Confidence Intervals

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

- i.e., if the estimation were repeated many times, the interval will have the parameter x% of the time

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

Think ring toss…

Parameter exists in the ideal world

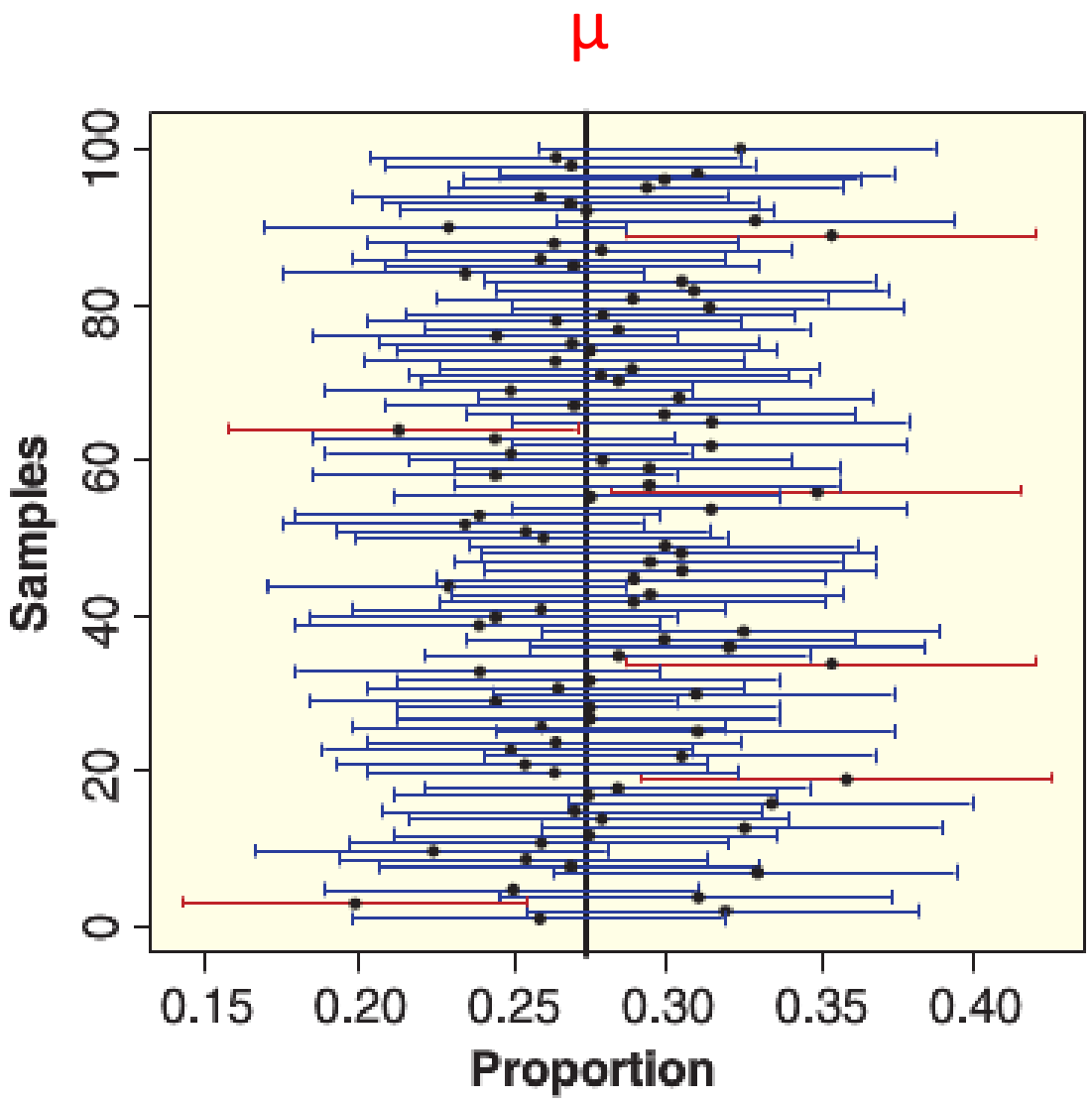
We toss intervals at it

95% of those intervals capture the parameter

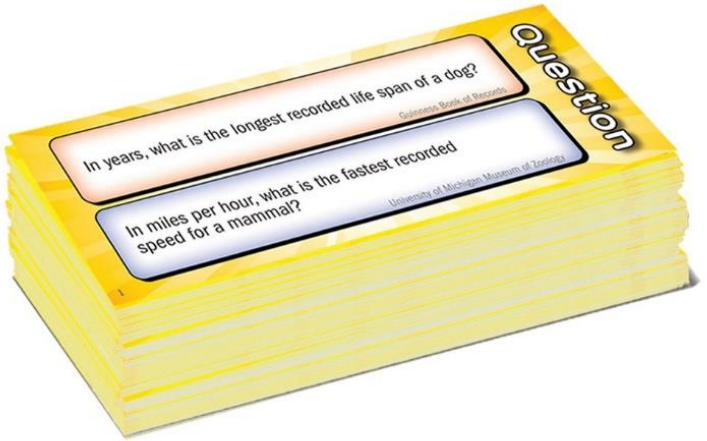
Confidence Intervals

For a **confidence level** of 95%...

95% of the **confidence intervals** will have the parameter in them



Wits and Wagers:
90% confidence interval estimator



I will ask 10 questions that have numeric answers

Please come up with a range of values that contains the true value in it for 9 out of the 10 questions

- i.e., be a 90% confidence interval estimator

Tradeoff between interval size and confidence level

There is a tradeoff between the **confidence level** (percent of times we capture the parameter) and the **confidence interval size**

Let's explore this in Jupyter!

Interpreting confidence intervals

By our calculation, an approximate 95% confidence interval for the average age of the mothers in the population is (26.9, 27.6) years.

True or False:

- About 95% of the mothers in the population were between 26.9 years and 27.6 years old.

When not to use the Bootstrap

If you're trying to estimate very high or very low percentiles, or min and max

If you're trying to estimate any parameter that's greatly affected by rare elements of the population

If the probability distribution of your statistic is not roughly bell shaped (the shape of the empirical distribution will be a clue)

If the original sample is very small

Confidence Intervals for Testing

Using a CI for testing

Null hypothesis: Population average = x

Alternative hypothesis: Population average ≠ x

Cutoff for P-value: p%

Method:

- Construct a (100-p)% condence interval for the population average
- If x is not in the interval, reject the null
- If x is in the interval, can't reject the null

Let's explore this in Jupyter!