Sampling Distributions

Galvanize

Objectives:

- Define population and sample
- Define independent and identically distributed
- Define sampling distribution of a statistic
- Visualize a sample - scatter plot, histogram, empirical CDF
- Bootstrapping - approximate the distribution of a statistic
- Bootstrapping - approximate a confidence interval for a population parameter

Review

- What is a random variable?
- What is the probability mass function
- What is the probability density function?
- What is the cumulative distribution function?
- What is the difference between the PMF and the PDF?
- How is the PDF related to the CDF?
- What is the difference between probability and statistics?
- What are the steps of hypothesis testing?
- What is a p-value?

Population vs Sample

Population: set of similar items or events

- Daily prices from the stock market
- Possible customers of insurance company
- Possible paths through a city

Sample: subset of individuals from within a statistical population

- October prices from the stock market
- Customers receiving quotes
- Routes found by randomly driving

Independent and Identically Distributed (i.i.d)

Independent: sample items are not connected to each other

Identically Distributed: sample items have the same distribution function

If X and Y are i.i.d. random variables:

- Any probabilistic statement about X and Y are the same
- Notation: $X \sim Y$

Which sample is not i.i.d.?

Population: set of similar items or events

- Daily prices from the stock market
- Possible customers of insurance company
- Possible paths through a city

Sample: subset of individuals from within a statistical population

- October prices from the stock market
- Customers receiving quotes
- Routes found by randomly driving

Which sample is not i.i.d.?

Population: set of similar items or events

- Daily prices from the stock market
- Possible customers of insurance company
- Possible paths through a city

Sample: subset of individuals from within a statistical population

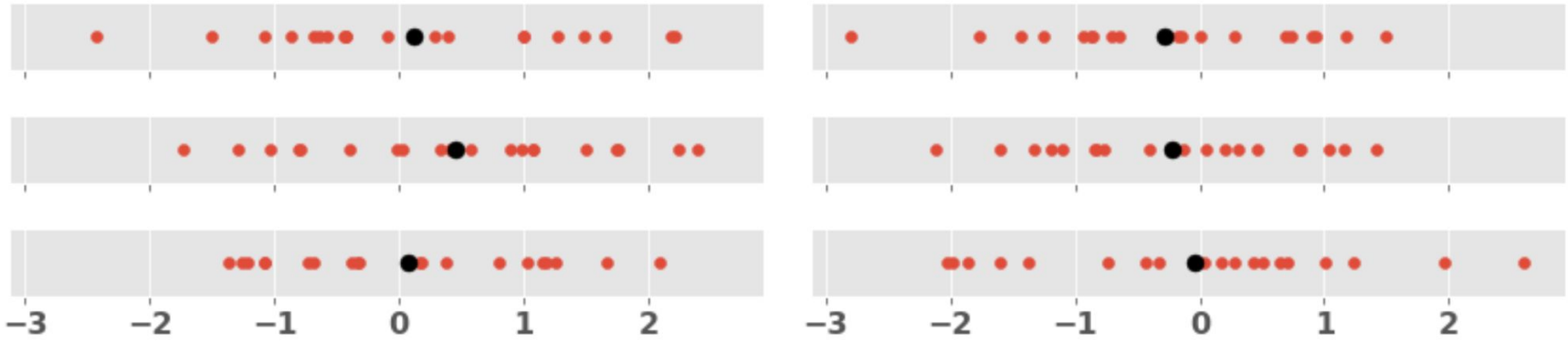
- October prices from the stock market
- Customers receiving quotes
- Routes found by randomly driving

Sampling Distributions

- Mathematically model a simple random sample:
  - Consider each individual data point drawn from our population as the outcome from its own random variable
  - i.i.d. sample >> sequence of random variables that are independent and identically distributed
  - Data collection provides data points where each one is drawn from its corresponding random variable
- Calculating a sample statistic
  - Function of a random sample: $T(X_1, X_2, X_3 \ldots X_n)$
  - Statistic can be computed once the random sample is drawn from the population
  - Drawing different random samples will result in different values of the statistic
  - Sample mean: $\frac{1}{n}\sum_i X_i$

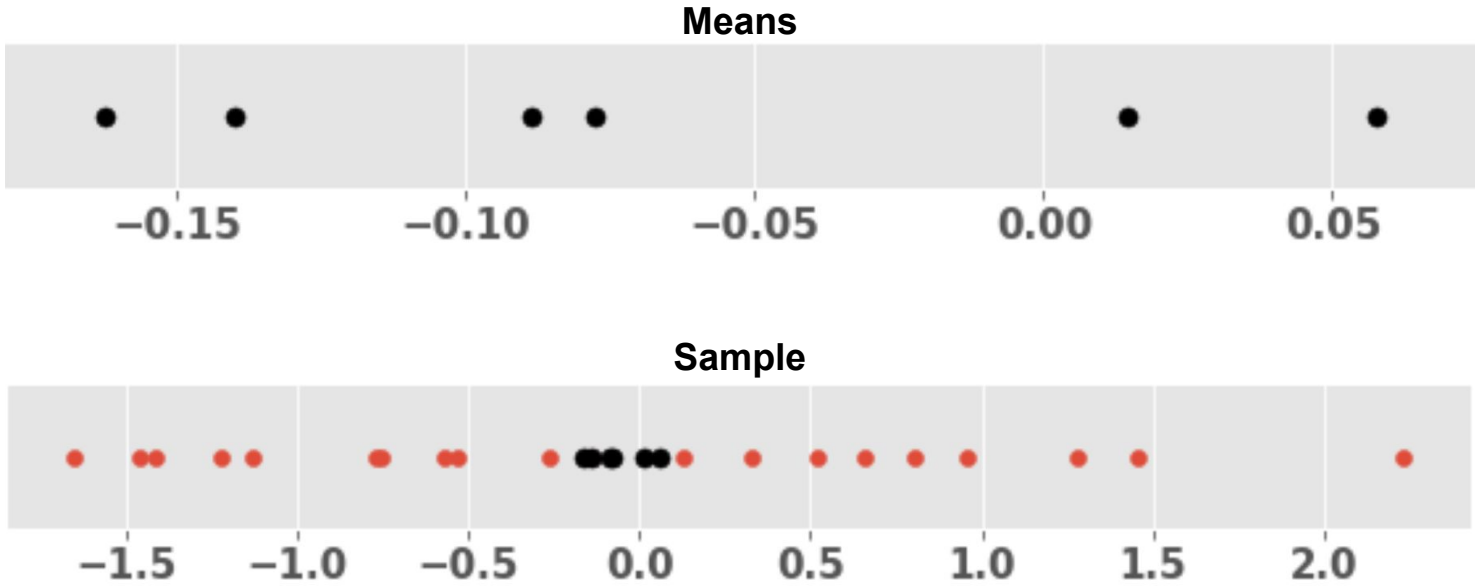
Sampling Distributions

Mean of six samples from the population



Sampling Distribution

Distribution of a sample statistic is more narrow than the distribution of the population



Sampling Distributions

Main Question:

How do we quantify the amount of variation of a sample statistic?

Ideal Answer:

1) Draw a number of i.i.d. points from the population
2) Compute the statistic
3) Record the statistic in a database
4) Repeat, repeat, repeat until the sun burns out
5) Determine the variation in your database

Breakout

1) Come up with one or two examples of situations where multiple samples would be difficult to draw from the population.
    a) What is the population?
    b) What is the sample?
    c) What statistic are you interested in?