

Simulation-based inference - confidence intervals

Intro to Data Science

Shawn Santo

Announcements

- Hard deadline for Lab 05 is Friday
- Soft deadline for Homework 03 is Friday
- No office hours on March 9th and 10th
- No lab on March 9th

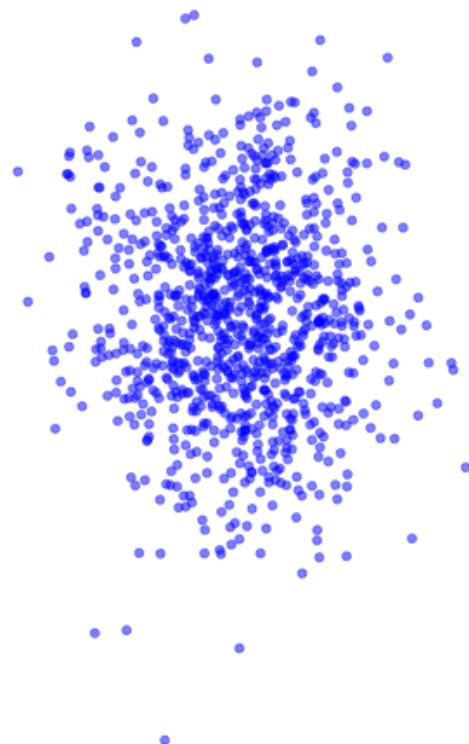
Today's agenda

- Create interval estimates using a simulation-based approach
- Interpret confidence intervals

Recall

Big picture

Population of interest



Sample



Terminology

Population: a group of individuals or objects we are interested in studying

Parameter: a numerical quantity derived from the population (almost always unknown)

Sample: a subset of our population of interest

Statistic: a numerical quantity derived from a sample

Common population parameters of interest and their corresponding sample statistic:

Quantity	Parameter	Statistic
Mean	μ	\bar{x}
Variance	σ^2	s^2
Standard deviation	σ	s
Median	M	\tilde{x}
Proportion	p	\hat{p}

Statistical inference

Statistical inference is the process of using sample data to make conclusions about the underlying population the sample came from.

- **Estimation:** estimating an unknown parameter based on values from the sample at hand
- **Testing:** evaluating whether our observed sample provides evidence for or against some claim about the population

Today we will focus on estimation.

Estimation

Point estimate

A point estimate is a single value computed from the sample data to serve as the "best guess", or estimate, for the population parameter.

Suppose we were interested in the population mean. What would be natural point estimate to use?

Quantity	Parameter	Statistic
Mean	μ	\bar{x}
Variance	σ^2	s^2
Standard deviation	σ	s
Median	M	\tilde{x}
Proportion	p	\hat{p}

What is the downside to using point estimates?

Confidence intervals

A plausible range of values for the population parameter is an interval estimate. One type of interval estimate is known as a **confidence interval**.



- If we report a point estimate, we probably won't hit the exact population parameter.
- If we report a range of plausible values, we have a good shot at capturing the parameter.

Variability of sample statistics

- In order to construct a confidence interval we need to quantify the variability of our sample statistic.
- For example, if we want to construct a confidence interval for a population mean, we need to come up with a plausible range of values around our observed sample mean.
- This range will depend on how precise and how accurate our sample mean is as an estimate of the population mean.
- Quantifying this requires a measurement of how much we would expect the sample mean to vary from sample to sample.

Suppose you randomly sample 50 students and 5 of them are left handed. If you were to take another random sample of 50 students, how many would you expect to be left handed? Would you be surprised if only 3 of them were left handed? Would you be surprised if 40 of them were left handed?

Quantifying the variability of a sample statistic

We can quantify the variability of sample statistics using

1. **simulation**: via bootstrapping (today);
2. **theory**: via Central Limit Theorem (later in the course).

Bootstrapping

Bootstrapping

- The term **bootstrapping** comes from the phrase "pulling oneself up by one's bootstraps", to help oneself without the aid of others.
- In this case, we are estimating a population parameter, and we'll accomplish it using data from only from the given sample.
- This notion of saying something about a population parameter using only information from an observed sample is the crux of statistical inference, it is not limited to bootstrapping.

Bootstrapping scheme

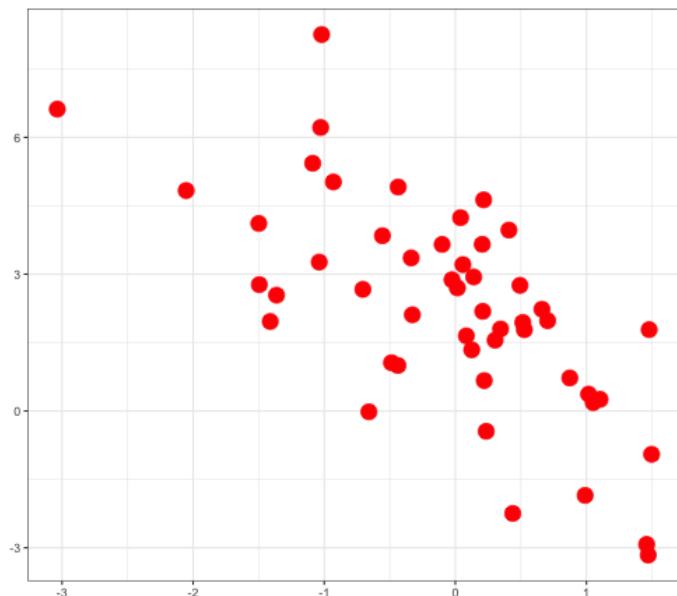
1. **Take a bootstrap sample** - a random sample taken with replacement from the original sample, of the same size as the original sample.
2. **Calculate the bootstrap statistic** - a statistic such as mean, median, proportion, slope, etc. computed from the bootstrap samples.
3. **Repeat steps (1) and (2) many times to create a bootstrap distribution** - a distribution of bootstrap statistics.
4. **Calculate the bounds of the XX% confidence interval** as the middle XX% of the bootstrap distribution.

Bootstrapping scheme (1 - 2) visualized

Bootstrapping scheme (1 - 2) animated

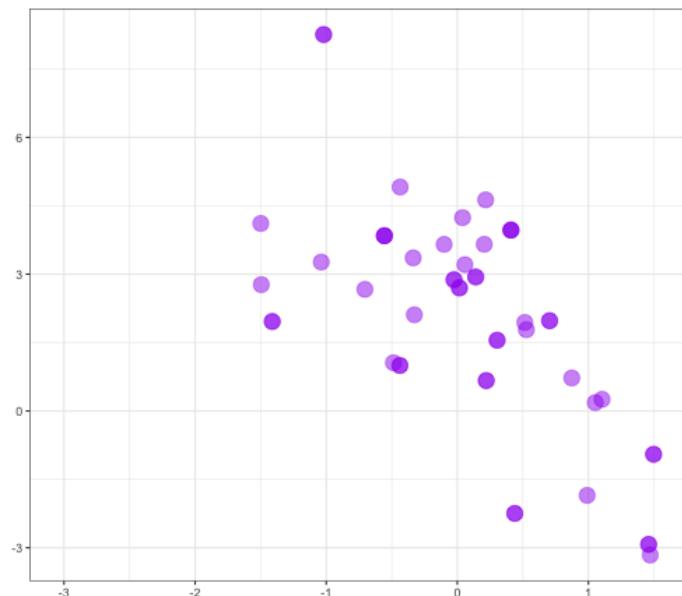
Original sample

Size: 50



Bootstrap sample: 1

Size: 50



For each bootstrap sample, we would compute our statistic of interest, e.g. correlation.

Let's explore these concepts

- Create your personal private repository by clicking
<https://classroom.github.com/a/tXyLEVzo>