Resampling

ESS 469/569

University of Washington

What do we mean by resampling?

In the context of data science, resampling refers to the process of generating one or more *new* samples from some sample[1].

Essentially: "repeated draws from a sample."

[1] Here, sample refers to a subset (i.e., *statistical sample*) of some population.

Why would we want to resample data?

To answer this question, we should consider the challenges associated with statistical sampling.

Why would we want to resample data?

To answer this question, we should consider the challenges associated with statistical sampling.

Among other benefits, resampling enables us to understand uncertainties associated with our sample (and derived statistics).

Resampling techniques are central to data science

You will see resampling—in various forms—time and time again when working with ML/AI.

There are many kinds of resampling methods

Including:

- Randomization/permutation
- Bootstrap
- Jackknife
- Cross-validation[2]

[2]We will revisit cross-validation when we start to examine the validity of models.

Randomization

Given two populations, A and B, and a statistic, x, how might we determine if x_A differs from x_B?

Randomization

Given two populations, A and B, and a statistic, x, how might we determine if x_A differs from x_B?

We can randomly assign observations to either A or B, calculate $x_A - x_B$ and repeat to build a distribution of differences.

Randomization

Given two populations, A and B, and a statistic, x, how might we determine if x_A differs from x_B?

We can randomly assign observations to either A or B, calculate $x_A - x_B$ and repeat to build a distribution of differences.

If you use every possible combination of rearranging data, then you are permutation testing.

Some important concepts to consider

You might come across the terms *with* and *without* replacement. The former means that you can draw the sample multiple times. The latter means that, once a sample is drawn, it cannot be redrawn.

Some important concepts to consider

You might come across the terms *with* and *without* replacement. The former means that you can draw the sample multiple times. The latter means that, once a sample is drawn, it cannot be redrawn.

When resampling, you want a sufficiently large number of iterations.

Some important concepts to consider

You might come across the terms *with* and *without* replacement. The former means that you can draw the sample multiple times. The latter means that, once a sample is drawn, it cannot be redrawn.

When resampling, you want a sufficiently large number of iterations.

Unless you use an outside source of random data, all of your random methods are pseudo-random.

Bootstrap

Let us assume that you have some data, D, that describe some true population well (which is not always true!).

Bootstrap

Let us assume that you have some data, D, that describe some true population well (which is not always true!).

Let us also assume that you've calculated some parameter, x, using D. You are interested in how x might vary.

Bootstrap

Let us assume that you have some data, D, that describe some true population well (which is not always true!).

Let us also assume that you've calculated some parameter, x, using D. You are interested in how x might vary.

In such a case, you might decided to repeatedly generate new samples, each time drawing from D independently and **with replacement**. For each new sample, you could calculate a value for x. The distribution of these values can provide additional information about x.

Bootstrap

Let us assume that you have some data, D, that describe some true population well (which is not always true!).

Let us also assume that you've calculated some parameter, x, using D. You are interested in how x might vary.

In such a case, you might decided to repeatedly generate new samples, each time drawing from D independently and **with replacement**. For each new sample, you could calculate a value for x. The distribution of these values can provide additional information about x.

This process is known as bootstrapping.

Jackknife

Given a dataset, D and a target statistic/parameter, x, you can systematically leave out one observation from D and calculate a pseudovalue, k, of the target of interest.

Jackknife

Given a dataset, D and a target statistic/parameter, x, you can systematically leave out one observation from D and calculate a pseudovalue, k, of the target of interest.

As before, you can use the collection of k values to better estimate x.

Jackknife

Given a dataset, D and a target statistic/parameter, x, you can systematically leave out one observation from D and calculate a pseudovalue, k, of the target of interest.

As before, you can use the collection of k values to better estimate x.

This so-called jackknife estimate predates many other resampling techniques.

Moving on: Monte Carlo methods

So far, we have been taking draws from samples of populations, where the true distribution of data is unknown.

Moving on: Monte Carlo methods

So far, we have been taking draws from samples of populations, where the true distribution of data is unknown.

But what if we have knowledge about some underlying distribution?

Moving on: Monte Carlo methods

So far, we have been taking draws from samples of populations, where the true distribution of data is unknown.

But what if we have knowledge about some underlying distribution?

Monte Carlo methods are one class of resampling techniques. When implementing a Monte Carlo approach, you take draws from predefined distributions in order to better answer some question.

Moving on: Monte Carlo methods

So far, we have been taking draws from samples of populations, where the true distribution of data is unknown.

But what if we have knowledge about some underlying distribution?

Monte Carlo methods are one class of resampling techniques. When implementing a Monte Carlo approach, you take draws from predefined distributions in order to better answer some question.

These draws can be independent or not (e.g., Markov Chains).

Moving on: Monte Carlo methods

So far, we have been taking draws from samples of populations, where the true distribution of data is unknown.

But what if we have knowledge about some underlying distribution?

Monte Carlo methods are one class of resampling techniques. When implementing a Monte Carlo approach, you take draws from predefined distributions in order to better answer some question.

These draws can be independent or not (e.g., Markov Chains).

Monte Carlo approaches have many applications, from approximations of complex functions to error propagation.