Bootstrapping Statistical Quantities and Models

David Carlson

March 28, 2022

Overview

- We have not covered this yet

Overview

- We have not covered this yet
- Even though it is not in the book or plan, it is an important concept to understand

Overview

- We have not covered this yet
- Even though it is not in the book or plan, it is an important concept to understand
- Bootstrapping is a general approach to statistical inference based on building a sampling distribution for a statistic by resampling from the data at hand

Overview

- We have not covered this yet
- Even though it is not in the book or plan, it is an important concept to understand
- Bootstrapping is a general approach to statistical inference based on building a sampling distribution for a statistic by resampling from the data at hand
- We can use it for population statistics or modeling statistics

What It Is

- Any test or metric that relies on random sampling *with replacement* from observed data

What It Is

- Any test or metric that relies on random sampling *with replacement* from observed data
- Mimics sampling process

What It Is

- Any test or metric that relies on random sampling *with replacement* from observed data
- Mimics sampling process
- Allows estimation of the sampling distribution of almost any statistic using random sampling methods

What It Is

- Any test or metric that relies on random sampling *with replacement* from observed data
- Mimics sampling process
- Allows estimation of the sampling distribution of almost any statistic using random sampling methods
- Estimates the properties of an estimator (such as its variance) by measuring those properties when sampling from an approximating distribution (e.g., empirical distribution function of observed data)

What It Is

- Any test or metric that relies on random sampling *with replacement* from observed data
- Mimics sampling process
- Allows estimation of the sampling distribution of almost any statistic using random sampling methods
- Estimates the properties of an estimator (such as its variance) by measuring those properties when sampling from an approximating distribution (e.g., empirical distribution function of observed data)
- When obs i.i.d. (we often make this assumption anyways) construct resamples *with replacement* of observed data and of equal size to observed data

Approach

- inference about a population from sample data (sample $\rightarrow$ population) can be modeled by resampling the sample data and performing inference about a sample from resampled data (resampled $\rightarrow$ sample)

Approach

- inference about a population from sample data (sample $\rightarrow$ population) can be modeled by resampling the sample data and performing inference about a sample from resampled data (resampled $\rightarrow$ sample)
- As the population is unknown, the true error in a sample statistic against its population value is unknown

Approach

- inference about a population from sample data (sample $\rightarrow$ population) can be modeled by resampling the sample data and performing inference about a sample from resampled data (resampled $\rightarrow$ sample)
- As the population is unknown, the true error in a sample statistic against its population value is unknown
- In bootstrap-resamples, the 'population' is in fact the sample, and this is known; hence the quality of inference of the 'true' sample from resampled data (resampled $\rightarrow$ sample) is measurable

Approach

- inference about a population from sample data (sample $\rightarrow$ population) can be modeled by resampling the sample data and performing inference about a sample from resampled data (resampled $\rightarrow$ sample)
- As the population is unknown, the true error in a sample statistic against its population value is unknown
- In bootstrap-resamples, the 'population' is in fact the sample, and this is known; hence the quality of inference of the 'true' sample from resampled data (resampled $\rightarrow$ sample) is measurable
- If sample is reasonable approximation to population of interest, quality of inference on population of interest can be inferred

Example of Population Stat (the mean)

- Assume we are interested in the average (or mean) height of people worldwide

Example of Population Stat (the mean)

- Assume we are interested in the average (or mean) height of people worldwide
- We cannot measure all the people in the global population, so instead, we sample only a tiny part of it, and measure that

Example of Population Stat (the mean)

- Assume we are interested in the average (or mean) height of people worldwide
- We cannot measure all the people in the global population, so instead, we sample only a tiny part of it, and measure that
- Assume the sample is of size N; that is, we measure the heights of N individuals

Example of Population Stat (the mean)

- Assume we are interested in the average (or mean) height of people worldwide
- We cannot measure all the people in the global population, so instead, we sample only a tiny part of it, and measure that
- Assume the sample is of size N; that is, we measure the heights of N individuals
- From that single sample, only one estimate of the mean can be obtained

Example of Population Stat (the mean)

- Assume we are interested in the average (or mean) height of people worldwide
- We cannot measure all the people in the global population, so instead, we sample only a tiny part of it, and measure that
- Assume the sample is of size N; that is, we measure the heights of N individuals
- From that single sample, only one estimate of the mean can be obtained
- In order to reason about the population, we need some sense of the variability of the mean that we have computed

Example of Population Stat (the mean)

- Assume we are interested in the average (or mean) height of people worldwide
- We cannot measure all the people in the global population, so instead, we sample only a tiny part of it, and measure that
- Assume the sample is of size N; that is, we measure the heights of N individuals
- From that single sample, only one estimate of the mean can be obtained
- In order to reason about the population, we need some sense of the variability of the mean that we have computed
- The simplest bootstrap method involves taking the original data set of heights, and sampling from it to form a new sample (called a 'resample' or bootstrap sample) that is also of size N

Example of Population Stat (the mean), cont.

- The bootstrap sample is taken from the original by using sampling with replacement

Example of Population Stat (the mean), cont.

- The bootstrap sample is taken from the original by using sampling with replacement
- Assuming N is sufficiently large, for all practical purposes there is virtually zero probability that it will be identical to the original 'real' sample

Example of Population Stat (the mean), cont.

- The bootstrap sample is taken from the original by using sampling with replacement
- Assuming N is sufficiently large, for all practical purposes there is virtually zero probability that it will be identical to the original ‘real’ sample
- This process is repeated a large number of times (typically 1,000 or 10,000 times), and for each of these bootstrap samples, we compute its mean (each of these is called bootstrap estimates)

Example of Population Stat (the mean), cont.

- The bootstrap sample is taken from the original by using sampling with replacement
- Assuming N is sufficiently large, for all practical purposes there is virtually zero probability that it will be identical to the original 'real' sample
- This process is repeated a large number of times (typically 1,000 or 10,000 times), and for each of these bootstrap samples, we compute its mean (each of these is called bootstrap estimates)
- We now can create a histogram of bootstrap means

Example of Population Stat (the mean), cont.

- The bootstrap sample is taken from the original by using sampling with replacement
- Assuming N is sufficiently large, for all practical purposes there is virtually zero probability that it will be identical to the original 'real' sample
- This process is repeated a large number of times (typically 1,000 or 10,000 times), and for each of these bootstrap samples, we compute its mean (each of these is called bootstrap estimates)
- We now can create a histogram of bootstrap means
- This histogram provides an estimate of the shape of the distribution of the sample mean from which we can answer questions about how much the mean varies across samples

Example of Population Stat (the mean), cont.

- The bootstrap sample is taken from the original by using sampling with replacement
- Assuming N is sufficiently large, for all practical purposes there is virtually zero probability that it will be identical to the original 'real' sample
- This process is repeated a large number of times (typically 1,000 or 10,000 times), and for each of these bootstrap samples, we compute its mean (each of these is called bootstrap estimates)
- We now can create a histogram of bootstrap means
- This histogram provides an estimate of the shape of the distribution of the sample mean from which we can answer questions about how much the mean varies across samples
- The method here, described for the mean, can be applied to almost any other statistic or estimator (e.g., model parameters)

The Practice for Estimators

- Suppose we have N observations / rows

The Practice for Estimators

- Suppose we have N observations / rows
- We want to estimate β's and their variability (think OLS but works for any GLM)

The Practice for Estimators

- Suppose we have N observations / rows
- We want to estimate β's and their variability (think OLS but works for any GLM)
- We sample *with replacement* N rows of the data, estimate β's, and store

The Practice for Estimators

- Suppose we have N observations / rows
- We want to estimate β's and their variability (think OLS but works for any GLM)
- We sample *with replacement* N rows of the data, estimate β's, and store
- Repeat many times (again, typically 1,000 or 10,000)

The Practice for Estimators

- Suppose we have N observations / rows
- We want to estimate β's and their variability (think OLS but works for any GLM)
- We sample *with replacement* N rows of the data, estimate β's, and store
- Repeat many times (again, typically 1,000 or 10,000)
- We then look at the collected β estimates and make inferences using their empirical distributions

The Practice for Estimators

- Suppose we have N observations / rows
- We want to estimate β's and their variability (think OLS but works for any GLM)
- We sample *with replacement* N rows of the data, estimate β's, and store
- Repeat many times (again, typically 1,000 or 10,000)
- We then look at the collected β estimates and make inferences using their empirical distributions
- E.g., for CIs, we look at the 0.95 interval of collected β's

The Practice for Estimators

- Suppose we have N observations / rows
- We want to estimate β's and their variability (think OLS but works for any GLM)
- We sample *with replacement* N rows of the data, estimate β's, and store
- Repeat many times (again, typically 1,000 or 10,000)
- We then look at the collected β estimates and make inferences using their empirical distributions
- E.g., for CIs, we look at the 0.95 interval of collected β's
- Moves away from the idea that β's are fixed points, and starts to integrate distributional inferences

Advantages

- The bootstrap is quite general

Advantages

- The bootstrap is quite general
- Because it does not require distributional assumptions (such as normally distributed errors), the bootstrap can provide more accurate inferences when the data are not well behaved or when the sample size is small

Advantages

- The bootstrap is quite general
- Because it does not require distributional assumptions (such as normally distributed errors), the bootstrap can provide more accurate inferences when the data are not well behaved or when the sample size is small
- It is possible to apply the bootstrap to statistics with sampling distributions that are difficult to derive, even asymptotically

Advantages

- The bootstrap is quite general
- Because it does not require distributional assumptions (such as normally distributed errors), the bootstrap can provide more accurate inferences when the data are not well behaved or when the sample size is small
- It is possible to apply the bootstrap to statistics with sampling distributions that are difficult to derive, even asymptotically
- It is relatively simple to apply the bootstrap to complex data-collection plans (such as stratified and clustered samples)

Advantages

- The bootstrap is quite general
- Because it does not require distributional assumptions (such as normally distributed errors), the bootstrap can provide more accurate inferences when the data are not well behaved or when the sample size is small
- It is possible to apply the bootstrap to statistics with sampling distributions that are difficult to derive, even asymptotically
- It is relatively simple to apply the bootstrap to complex data-collection plans (such as stratified and clustered samples)
- Bootstrap is also an appropriate way to control and check the stability of the results

Advantages

- The bootstrap is quite general
- Because it does not require distributional assumptions (such as normally distributed errors), the bootstrap can provide more accurate inferences when the data are not well behaved or when the sample size is small
- It is possible to apply the bootstrap to statistics with sampling distributions that are difficult to derive, even asymptotically
- It is relatively simple to apply the bootstrap to complex data-collection plans (such as stratified and clustered samples)
- Bootstrap is also an appropriate way to control and check the stability of the results
- Although for most problems it is impossible to know the true confidence interval, bootstrap is asymptotically more accurate than the standard intervals obtained using sample variance and assumptions of normality

Advantages

- The bootstrap is quite general
- Because it does not require distributional assumptions (such as normally distributed errors), the bootstrap can provide more accurate inferences when the data are not well behaved or when the sample size is small
- It is possible to apply the bootstrap to statistics with sampling distributions that are difficult to derive, even asymptotically
- It is relatively simple to apply the bootstrap to complex data-collection plans (such as stratified and clustered samples)
- Bootstrap is also an appropriate way to control and check the stability of the results
- Although for most problems it is impossible to know the true confidence interval, bootstrap is asymptotically more accurate than the standard intervals obtained using sample variance and assumptions of normality
- Bootstrapping is also a convenient method that avoids the cost of repeating the experiment to get other groups of sample data

Disadvantages

- Bootstrapping depends heavily on the estimator used and, though simple, ignorant use of bootstrapping will not always yield asymptotically valid results and can lead to inconsistency

Disadvantages

- Bootstrapping depends heavily on the estimator used and, though simple, ignorant use of bootstrapping will not always yield asymptotically valid results and can lead to inconsistency
- Although bootstrapping is (under some conditions) asymptotically consistent, it does not provide general finite-sample guarantees

Disadvantages

- Bootstrapping depends heavily on the estimator used and, though simple, ignorant use of bootstrapping will not always yield asymptotically valid results and can lead to inconsistency
- Although bootstrapping is (under some conditions) asymptotically consistent, it does not provide general finite-sample guarantees
- The result may depend on the representative sample

Disadvantages

- Bootstrapping depends heavily on the estimator used and, though simple, ignorant use of bootstrapping will not always yield asymptotically valid results and can lead to inconsistency
- Although bootstrapping is (under some conditions) asymptotically consistent, it does not provide general finite-sample guarantees
- The result may depend on the representative sample
- The apparent simplicity may conceal the fact that important assumptions are being made when undertaking the bootstrap analysis (e.g. independence of samples or large enough of a sample size) where these would be more formally stated in other approaches

Disadvantages

- Bootstrapping depends heavily on the estimator used and, though simple, ignorant use of bootstrapping will not always yield asymptotically valid results and can lead to inconsistency
- Although bootstrapping is (under some conditions) asymptotically consistent, it does not provide general finite-sample guarantees
- The result may depend on the representative sample
- The apparent simplicity may conceal the fact that important assumptions are being made when undertaking the bootstrap analysis (e.g. independence of samples or large enough of a sample size) where these would be more formally stated in other approaches
- Bootstrapping can be time-consuming

Disadvantages

- Bootstrapping depends heavily on the estimator used and, though simple, ignorant use of bootstrapping will not always yield asymptotically valid results and can lead to inconsistency
- Although bootstrapping is (under some conditions) asymptotically consistent, it does not provide general finite-sample guarantees
- The result may depend on the representative sample
- The apparent simplicity may conceal the fact that important assumptions are being made when undertaking the bootstrap analysis (e.g. independence of samples or large enough of a sample size) where these would be more formally stated in other approaches
- Bootstrapping can be time-consuming
- Not terribly common in social science substantive papers (if a top journal, need almost no explanation, lower-tier journals will require a bit of discussion)