S&DS 242/542: Theory of Statistics

Lecture 20: Uncertainty quantification using the bootstrap

What is the bootstrap?

The **bootstrap** (Efron, 1979) refers to a simulation-based approach to quantify the uncertainty of statistical estimates.

It may be used to estimate the standard error of a statistic, or to construct a confidence interval for an estimated parameter.

There are two common versions of the bootstrap: the **parametric bootstrap** and the **nonparametric bootstrap**. We will discuss both methods in this lecture.

Bootstrap standard errors

Simulating the standard error

Typical question of interest: We fit a model $f(x \mid \theta)$ to data $X_1, \ldots, X_n$. How can we estimate the standard error of $\hat{\theta}$?

Previous lectures: Compute a theoretical distribution for $\hat{\theta}$, or a normal approximation for this distribution when n is large. Then estimate the standard deviation of this distribution.

The simulation approach: Repeatedly simulate new data

$$X_1^*, \ldots, X_n^*$$

In each simulation, compute $\hat{\theta}^*$ using $X_1^*, \ldots, X_n^*$. Use the standard deviation of $\hat{\theta}^*$ across simulations.

The bootstrap idea

We'd like to simulate

$$X_1^*, \ldots, X_n^*$$

from the same distribution as our actual data $X_1, \ldots, X_n$. The challenge is that we can't actually do this in practice, because we don't know the distribution of $X_1, \ldots, X_n$ to begin with.

The bootstrap idea: Simulate $X_1^*, \ldots, X_n^*$ from an *estimate* of the true data distribution. The name comes from the old English saying, "To pull oneself up by one's bootstraps."

The parametric and nonparametric bootstraps differ in how to estimate the true data distribution, which determines how to simulate $X_1^*, \ldots, X_n^*$.

The parametric bootstrap

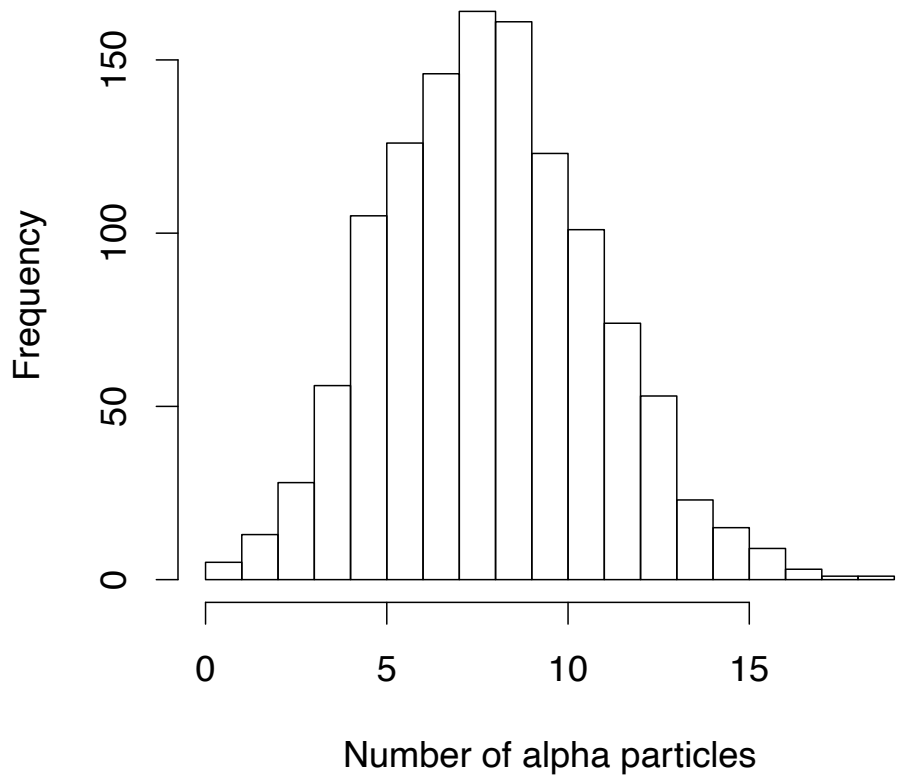
In the parametric bootstrap, we assume a parametric model $X_1, \ldots, X_n \overset{IID}{\sim} f(x \mid \theta)$. We estimate the parameter θ by $\hat{\theta}$, and simulate

$$X_1^*, \ldots, X_n^* \overset{IID}{\sim} f(x \mid \hat{\theta}).$$

This is a "plug-in" principle, quite analogous to how we have been using $I(\hat{\theta})$ in place of $I(\theta)$ to estimate the variance of the MLE in parametric models.

Example: Parametric bootstrap for the Poisson MLE

Numbers of alpha particles emitted by a sample of Americium-241 in 10-second intervals (Rice Chapter 8):



Example: Parametric bootstrap for the Poisson MLE

Fitting a Poisson(λ) model to this data, the MLE is $\hat{\lambda} = \bar{X} = 8.37$. What is the standard error of this estimate?

Using asymptotic theory (either by CLT or Fisher information):

$$\sqrt{n}(\hat{\lambda} - \lambda) \to \mathcal{N}(0, \lambda).$$

We can estimate the standard error as $\sqrt{8.37/n} = 0.083$.

Using the parametric bootstrap: Repeatedly simulate

$$X_1^*, \ldots, X_n^* \overset{IID}{\sim} \text{Poisson}(8.37),$$

compute $\hat{\lambda}^* = \frac{1}{n}(X_1^* + \ldots + X_n^*)$ for each simulation, and compute the empirical standard deviation of $\hat{\lambda}^*$ across simulations.

Example: Parametric bootstrap for the Poisson MLE

```
# Input: Data vector X
lambda_hat = mean(X)
n = length(X)

# Perform 100000 bootstrap simulations
B = 100000
lambda_hat_star = numeric(B)
for (i in 1:B) {
  X_star = rpois(n,lambda_hat)
  lambda_hat_star[i] = mean(X_star)
}
print(sd(lambda_hat_star))
```

We obtain the same answer, 0.083.

The nonparametric bootstrap

In the nonparametric bootstrap, we do not assume any parametric model when simulating new data. Instead, we sample

$$X_1^*, \ldots, X_n^*$$

independently *with replacement* from the original data values $X_1, \ldots, X_n$.

- The sample size for $X_1^*, \ldots, X_n^*$ in each bootstrap simulation is still n, the sample size of the original data.
- It is likely to have repeated values in $X_1^*, \ldots, X_n^*$, since we may sample the same value $X_1, \ldots, X_n$ more than once.
- It is also likely that some original values $X_1, \ldots, X_n$ will not appear in $X_1^*, \ldots, X_n^*$. For large n, typically 63.2% of the values in $X_1, \ldots, X_n$ will appear in $X_1^*, \ldots, X_n^*$. (This is a different 63.2% of samples in each bootstrap simulation.)

Example: Nonparametric bootstrap for the Poisson MLE

```
# Input: Data vector X
n = length(X)

# Perform 100000 bootstrap simulations
B=100000
lambda_hat_star = numeric(B)
for (i in 1:B) {
  X_star = sample(X, size=n, replace=TRUE)
  lambda_hat_star[i] = mean(X_star)
}
print(sd(lambda_hat_star))
```

We obtain an estimated standard error of 0.085. For this data, this is very close to the previous value 0.083, computed using the parametric bootstrap or asymptotic theory.

Rationale for the nonparametric bootstrap

The nonparametric bootstrap also simulates

$$X_1^*, \ldots, X_n^*$$

as IID draws from an estimate of the true data distribution.

This estimated distribution is the **empirical distribution** of the observed data, which is the discrete distribution that places mass $\frac{1}{n}$ at each of the observed data values $X_1, \ldots, X_n$.

A sample X_i^* drawn from this empirical distribution is equally likely to be any of the values $X_1, \ldots, X_n$. Thus IID samples $X_1^*, \ldots, X_n^*$ from this empirical distribution are exactly n samples drawn with replacement from the original data $X_1, \ldots, X_n$.

The empirical distribution

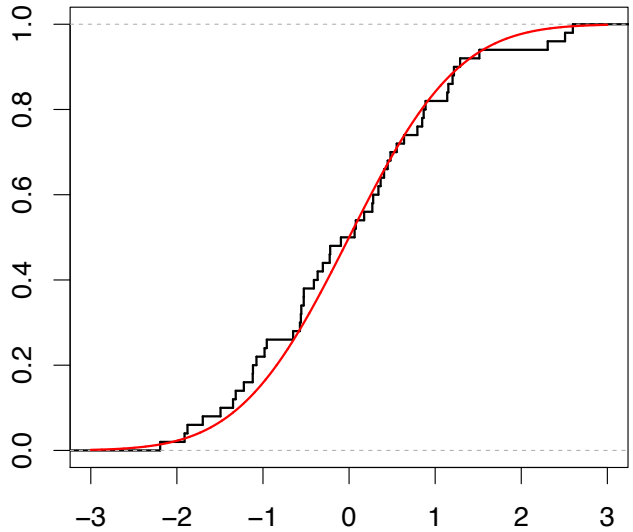
Estimating the true distribution by the empirical distribution is quite different from how we've discussed estimation thus far (which was to assume a model, and estimate parameters in this model).

There are some obvious ways in which the empirical distribution differs from the true distribution:

- Even if the true distribution were continuous, the empirical distribution is always discrete.
- It doesn't make sense to compare the mode, maximum value, and minimum value of the true distribution with the mode, maximum, and minimum of the empirical distribution.

However, the empirical distribution is an accurate estimate of the true distribution in other ways, for example in terms of its CDF, mean, and variance.

The empirical CDF



The CDF of the true distribution $\mathcal{N}(0,1)$ is in red. The CDF of the empirical distribution of $X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}(0,1)$ is in black. They are very close. This is because the empirical CDF is

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{X_i \leq t\}.$$

As $n \to \infty$, this converges to $\mathbb{P}[X_i \leq t] = F(t)$, for every value t.

The empirical mean and moments

Since the empirical distribution of $X_1, \ldots, X_n$ places mass $\frac{1}{n}$ on each value, its mean is the sample mean $\bar{X} = \frac{1}{n}(X_1 + \ldots + X_n)$. As $n \to \infty$, this converges to the mean of the true distribution, $\mathbb{E}[X]$.

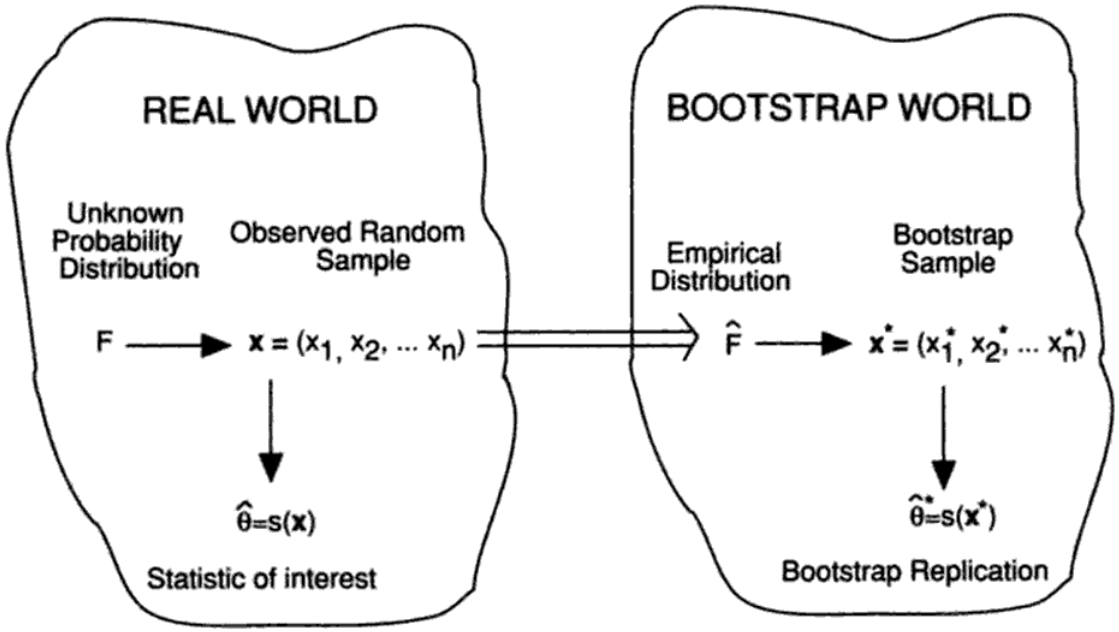
More generally, for any function $g(x)$, the expectation of $g(x)$ under the empirical distribution is

$$\frac{1}{n} \sum_{i=1}^{n} g(X_i).$$

As $n \to \infty$, this converges to the expectation of $g(x)$ under the true distribution, $\mathbb{E}[g(X)]$.

Taking $g(x) = x^k$, each k^{th} moment of the empirical distribution is close to the moment of the true distribution, for large n. In particular, the empirical distribution is a good approximation to the true distribution in terms of mean, variance, skewness, kurtosis, ...

Depiction of the nonparametric bootstrap



It makes sense to use the nonparametric bootstrap for statistics $s(\mathbf{X})$ whose behavior is similar under F and $\hat{F}$ (e.g. sample mean, sample variance, sample average of some function $g(x)$).

The nonparametric bootstrap should not be used for statistics like the sample mode or the maximum value $\max(X_1, \ldots, X_n)$.

Image from Efron and Tibshirani, *An Introduction to the Bootstrap*, 1993.

The bootstrap and model misspecification

Example: Suppose we have integer count data $X_1, \ldots, X_n \overset{IID}{\sim} g(x)$. We fit a Poisson($\lambda$) model by computing the MLE $\hat{\lambda} = \bar{X}$.

The Fisher information in this model is $I(\lambda) = \frac{1}{\lambda}$. So the plug-in Fisher information estimate of the standard error is

$$\sqrt{\frac{1}{nI(\hat{\lambda})}} = \sqrt{\frac{\hat{\lambda}}{n}} = \sqrt{\frac{\bar{X}}{n}}.$$

By the CLT, the true standard error is $\sqrt{\frac{\text{Var}_g[X]}{n}}$. If $g(x)$ is indeed a Poisson distribution, then $\text{Var}_g[X] = \lambda$ so $\sqrt{\frac{\hat{\lambda}}{n}}$ is an accurate estimate of this standard error for large n.

In general, this is accurate only if the variance of the true distribution is equal to or close to its mean.

The bootstrap and model misspecification

Suppose $X_1, \ldots, X_{100} \overset{IID}{\sim} \text{Geometric}(0.3)$, but we fit this Poisson(λ) model. Let $\bar{X}$ and S^2 be the sample mean and sample variance.

Estimated standard errors for the Poisson MLE $\hat{\lambda}$:

Fisher information estimate $\sqrt{\bar{X}/n}$:	0.14
Sample-variance based estimate $\sqrt{S^2/n}$:	0.22
Parametric bootstrap (assuming Poisson model):	0.14
Nonparametric bootstrap:	0.22

- The nonparametric bootstrap guards against model misspecification. The parametric bootstrap does not.
- Even if our statistic is motivated by a parametric model (for example, the MLE in this model), we may still wish to use the nonparametric bootstrap to estimate its standard error, to guard against model misspecification.

Bootstrap confidence intervals

Bootstrap confidence intervals

In addition to estimating standard errors, the bootstrap may also be used to construct confidence intervals.

There are many ways to do this. In this lecture, we will discuss and compare the three simplest approaches:

- The normal interval
- The percentile interval
- The "basic bootstrap" interval

These may all be applied using either the parametric or the nonparametric bootstrap.

The normal bootstrap interval

Let $\hat{\theta}$ be an estimate of $\theta \in \mathbb{R}$. Let $\widehat{\text{se}}$ be the (parametric or nonparametric) bootstrap estimate of the standard error of $\hat{\theta}$.

We may construct a $(1-\alpha)$-confidence interval for θ as

$$\hat{\theta} \pm z^{(\alpha/2)} \, \widehat{\text{se}}$$

where $z^{(\alpha/2)}$ is the upper-$\alpha/2$ point of the standard normal.

- This method is most similar to how we constructed confidence intervals in previous lectures. It simply replaces the theoretical standard error estimate with the bootstrap estimate.
- This interval is valid if the distribution of $\hat{\theta}$ is approximately normal around θ. This holds for many estimators $\hat{\theta}$ of interest, for large n.

The percentile bootstrap interval

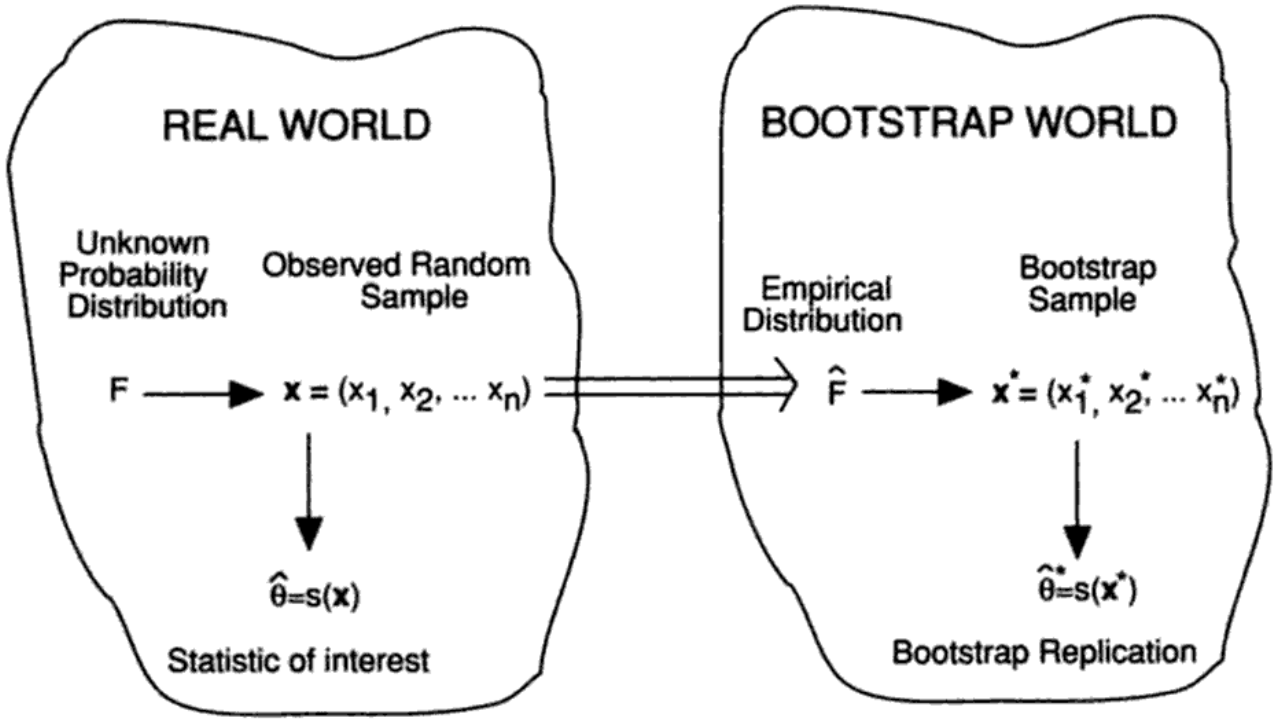
Suppose we performed B bootstrap simulations. Let $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_B$ be the values of $\hat{\theta}$ computed in these B simulations.

Let $\hat{\theta}^{*(\alpha/2)}$ and $\hat{\theta}^{*(1-\alpha/2)}$ be the empirical $\alpha/2$ and $1-\alpha/2$ quantiles of these simulated values. Construct a $(1-\alpha)$-confidence interval for θ as

$$[\hat{\theta}^{*(\alpha/2)}, \hat{\theta}^{*(1-\alpha/2)}].$$

- Rationale: Simulating $X^*_1, \ldots, X^*_n$ from an estimate of the true data distribution, the distribution of $\hat{\theta}^* = \hat{\theta}(X^*_1, \ldots, X^*_n)$ should be close to the distribution of $\hat{\theta}(X_1, \ldots, X_n)$. Then the quantiles for $\hat{\theta}^*$ should be close to the quantiles for $\hat{\theta}$.
- The quantiles of $\hat{\theta}$ form a valid confidence interval for θ if $\hat{\theta}$ is symmetrically distributed around θ. (Otherwise, the justification for this approach is perhaps unclear.)

The "basic bootstrap" interval



- In the Real World, θ is the unknown parameter and $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ is the estimate.
- In the Bootstrap World, $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ plays the role of the unknown parameter, and $\hat{\theta}^* = \hat{\theta}(X_1^*, \ldots, X_n^*)$ is the estimate.

The "basic bootstrap" interval

Let $q^{(\alpha/2)}$ and $q^{(1-\alpha/2)}$ be the $\alpha/2$ and $1-\alpha/2$ quantiles of

$$\hat{\theta}_1^* - \hat{\theta}, \ldots, \hat{\theta}_B^* - \hat{\theta}$$

across the B bootstrap simulations.

Use this to approximate the true distribution of $\hat{\theta} - \theta$. Observe that

$$\hat{\theta} - \theta \in [q^{(\alpha/2)}, q^{(1-\alpha/2)}] \Longleftrightarrow \theta \in [\hat{\theta} - q^{(1-\alpha/2)}, \hat{\theta} - q^{(\alpha/2)}].$$

Then construct a $(1-\alpha)$-confidence interval for θ as

$$[\hat{\theta} - q^{(1-\alpha/2)}, \hat{\theta} - q^{(\alpha/2)}].$$

If $q^{(\alpha/2)}$ and $q^{(1-\alpha/2)}$ were the true quantiles of $\hat{\theta} - \theta$, then this would have the desired coverage probability of $1-\alpha$.

Basic bootstrap versus percentile interval

The quantiles $q^{(\alpha/2)}$ and $q^{(1-\alpha/2)}$ of $\hat{\theta}^* - \hat{\theta}$ are equivalently

$$q^{(\alpha/2)} = \hat{\theta}^{*(\alpha/2)} - \hat{\theta}, \qquad q^{(1-\alpha/2)} = \hat{\theta}^{*(1-\alpha/2)} - \hat{\theta}.$$

Then the basic bootstrap interval $[\hat{\theta} - q^{(1-\alpha/2)}, \hat{\theta} - q^{(\alpha/2)}]$ is

$$[2\hat{\theta} - \hat{\theta}^{*(1-\alpha/2)}, 2\hat{\theta} - \hat{\theta}^{*(\alpha/2)}].$$

This is not the same as the percentile interval

$$[\hat{\theta}^{*(\alpha/2)}, \hat{\theta}^{*(1-\alpha/2)}].$$

If the distribution of $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_B$ is symmetric around $\hat{\theta}$, then these are the same because $\hat{\theta}^{*(\alpha/2)} + \hat{\theta}^{*(1-\alpha/2)} = 2\hat{\theta}$.

Basic bootstrap versus percentile interval

- Suppose $\hat{\theta}$ has positive bias for θ, and $\hat{\theta}^*$ has equal positive bias for $\hat{\theta}$. Then the percentile interval has twice this bias, and may not cover θ if the bias is severe. However, the basic bootstrap interval corrects for the bias, since the bias in $\hat{\theta}^* - \hat{\theta}$ is subtracted from the original bias of $\hat{\theta}$.
- Rice advocates for the basic bootstrap interval, and says of the percentile interval: "Although this direct equation of quantiles of the bootstrap sampling distribution with confidence limits may seem initially appealing, its rationale is somewhat obscure."

Pros and cons of the bootstrap

The bootstrap is one of the most widely used methods in practice for quantifying statistical uncertainty.

- It is easy to apply to complex statistics, and does not require theoretical approximation of the sampling distribution.
- The nonparametric bootstrap is a simple way to obtain standard errors that are valid under model misspecification.
- It can be computationally prohibitive, especially when the statistic of interest is hard to compute. This statistic must be recomputed for each bootstrap simulation.
- There are certain statistics (e.g. $\max(X_1, \ldots, X_n)$) that do not have similar behaviors under the true distribution F and under the empirical distribution F_n, for which the nonparametric bootstrap should not be applied.