Lab Session-9: The Bootstrap

MATH350 – Statistical Inference

Dr. Tanujit Chakraborty, Ph.D. from ISI Kolkata.
Assistant Professor in Statistics at Sorbonne University.
Mail me at tanujitisi@gmail.com

Course Webpage: https://www.ctanujit.org/SI.html

Code available at https://github.com/tanujit123/MATH350

Bootstrap Meme

Statistician
True value of θ
Bootstrap
Researcher who wants to obtain confidence intervals without knowing the estimator's variance

Image from reddit.com

What is the bootstrap?

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

There are many variants of the bootstrap; it is more of an idea underlying a collection of methods, rather than one single method.

Simulating the standard error

Typical question of interest: Given $X_1, \ldots, X_n \overset{IID}{\sim} f(x \mid \theta)$, what is the standard error of an estimator $\hat{\theta}$ for θ?

Use asymptotic theory to study the sampling distribution and

variance of $\hat{\theta}$, when n is large.

The simulation approach: Repeatedly simulate

$X_1^*, \ldots, X_n^* \overset{\text{IID}}{\sim} f(x \mid \theta)$, compute $\hat{\theta}^*$ from $X_1^*, \ldots, X_n^*$, and take the empirical standard deviation of $\hat{\theta}^*$ across simulations.

The bootstrap principle

We can’t actually simulate $X_1^*, \ldots, X_n^* \overset{IID}{\sim} f(x \mid \theta)$ in practice, because we don’t know θ to begin with.

The bootstrap idea: Simulate $X_1^*, \ldots, X_n^*$ from an *estimate* of the true data distribution.

This is a plugin principle analogous to how we use $I(\hat{\theta})$ for $I(\theta)$ when estimating the standard error of the MLE. Here, we "plug in" an estimate of the data distribution for the true data distribution, and then simulate new data from this estimate.

The name comes from the English saying, “To pull oneself up by one’s own bootstraps".

Example: Bootstrap standard error in the Poisson model

Numbers of alpha particles emitted by a sample of Americium-241 in 10-second intervals (Rice[1] Chapter 8)

Berkson (1966) conducted a careful analysis of data obtained from the National Bureau of Standards. The source of the alpha particles was americium 241. The experimenters recorded 10220 times between successive emissions. The observed mean emission rate (total number of emissions divided by total time) was 0.8392 emissions per sec.

The first two columns of the following table display the counts, n, that were observed in 1207 intervals, each of length 10 sec. In 18 of the 1207 intervals, there were 0, 1, or 2 counts; in 28 of the intervals, there were 3 counts, etc.

[1] Rice, John A. Mathematical statistics and data analysis.

Example: Bootstrap standard error in the Poisson model

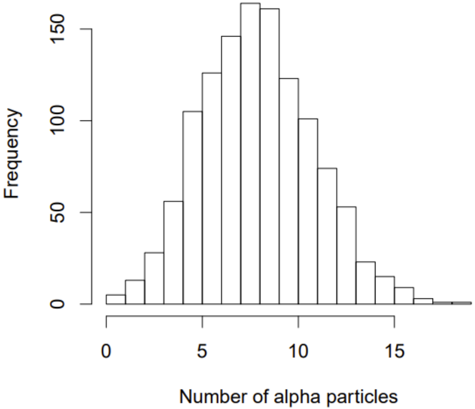
Numbers of alpha particles emitted by a sample of Americium-241 in 10-second intervals (Rice[2] Chapter 8)

n	Observed	Expected
0-2	18	12.2
3	28	27
4	56	56.5
5	105	94.9
6	126	132.7
7	146	159.1
8	164	166.9
9	161	155.6
10	123	130.6
11	101	99.7
12	74	69.7
13	53	45
14	23	27
15	15	15.1
16	9	7.9
17+	5	7.1

[2] Rice, John A. Mathematical statistics and data analysis.

Example: Bootstrap standard error in the Poisson model

Numbers of alpha particles emitted by a sample of Americium-241 in 10-second intervals (Rice[3] Chapter 8):



[3] Rice, John A. Mathematical statistics and data analysis.

Example: Bootstrap standard error in the Poisson model

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 bootstrap: Repeatedly simulate

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

compute $\hat{\lambda}^* = \bar{X}^*$ for each simulation, and compute the empirical standard deviation of $\hat{\lambda}^*$ across simulations.

Example: Bootstrap standard error in the Poisson model

```
# Input: Data vector X
X = 2:17
obs = c(18,28,56,105,126,146,164,161,
      123,101,74,53,23,15,9,5)
n = sum(obs)
lambda_hat = sum(X*obs)/n

# 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 parametric bootstrap

The method on the preceding slides is called the **parametric bootstrap**. Suppose, more generally, we are interested in the standard error of some statistic $T := T(X_1, \ldots, X_n)$.

1. Fit a parametric model $f(x \mid \theta)$ to $X_1, \ldots, X_n$ using an estimate $\hat{\theta}$ (say, the MLE)
2. For $i = 1, 2, \ldots, B$:
   a. Simulate $X_1^*, \ldots, X_n^* \overset{\text{IID}}{\sim} f(x \mid \hat{\theta})$
   b. Compute the statistic $T^* := T(X_1^*, \ldots, X_n^*)$ on the data $X_1^*, \ldots, X_n^*$
3. Return the empirical standard deviation of T^* across the B simulations

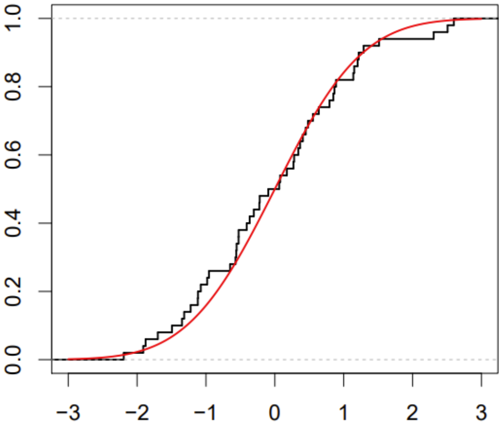
This is called the parametric bootstrap because the estimated distribution from which we simulate new data is obtained by fitting a parametric model $f(x \mid \theta)$.

The non-parametric bootstrap

A different method of performing the bootstrap is to "estimate" the true data distribution by the **empirical distribution** of the data, which is the discrete distribution that places mass $\frac{1}{n}$ at each of the observed data values $X_1, \dots, X_n$.

That is, given the observed data $X_1, \dots, X_n$, this is the distribution of a random variable that can equal each of these observed values with probability $\frac{1}{n}$.
Q: What is the CDF of this empirical distribution?

The non-parametric bootstrap



A: The empirical CDF $F_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i \leq t\}$, which equals the fraction of data values $\leq t$. This estimates the true CDF $F(t)$.

The non-parametric bootstrap

Simulating IID samples $X_1^*, \ldots, X_n^*$ from the empirical distribution of the data amounts to sampling, *with replacement*, n values from $X_1, \ldots, X_n$.

(Note that it is highly likely for some of the values $X_1^*, \ldots, X_n^*$ to be the same, even if the original values $X_1, \ldots, X_n$ were all distinct.)

This method of simulation is called the **non-parametric bootstrap**.

The non-parametric bootstrap

Suppose we are interested in the standard error of a statistic $T := T(X_1, \ldots, X_n)$. The nonparametric bootstrap does the following:

1. For $i = 1, 2, \ldots, B$:
    a. Simulate $X_1^*, \ldots, X_n^*$ as n samples with replacement from the original data $X_1, \ldots, X_n$.
    b. Compute the statistic $T^* = T(X_1^*, \ldots, X_n^*)$ on the data $X_1^*, \ldots, X_n^*$.
2. Return the empirical standard deviation of T^* across the B simulations.

There is no assumption of a parametric model!

Example: Bootstrap standard error in the Poisson model

```r
# Input: Data vector X
X = 2:17
obs = c(18,28,56,105,126,146,164,161,
       123,101,74,53,23,15,9,5)
n = sum(obs)
lambda_hat = sum(X*obs)/n

# 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))
```

MLE in a misspecified model revisited

Let's consider what happens when we fit the Poisson (λ) model to data $X_1, \ldots, X_n$ that do not follow a Poisson distribution. We compute the MLE $\hat{\lambda} = \bar{X}$.

The true standard error of this MLE is $\sigma/\sqrt{n}$ where σ is the standard deviation of the true distribution for the X_i 's.

The Fisher information is $1/\lambda$, so the plugin Fisher information estimate of the standard error is $\sqrt{\hat{\lambda}/n} = \sqrt{\bar{X}/n}$. This is incorrect if the mean of the distribution of $X_i's$ is not the same as its variance.

MLE in a misspecified model revisited

The sandwich estimate of the standard error of $\hat{\lambda}$ estimates separately $\mathrm{Var}[z(X,\lambda)]$ and $\mathbb{E}\left[z'(X,\lambda)\right]$:

$$\begin{aligned} \log f(x \mid \lambda) &= x \log \lambda - \lambda - \log x! \\ z(x,\lambda) &= \frac{\partial}{\partial \lambda} \log f(x \mid \lambda) = \frac{x}{\lambda} - 1 \\ z'(x,\lambda) &= \frac{\partial^2}{\partial \lambda^2} \log f(x \mid \lambda) = -\frac{x}{\lambda^2} \end{aligned}$$

Sample variance of $z\left(X_1,\bar{X}\right),\ldots,z\left(X_n,\bar{X}\right)$: $S_X^2/\bar{X}^2$
Sample mean of $z'\left(X_1,\bar{X}\right),\ldots,z'\left(X_n,\bar{X}\right)$: $-1/\bar{X}$.

So the sandwich estimate of the standard error of $\hat{\lambda}$ is $S_X/\sqrt{n}$, which is a correct estimate of $\sigma/\sqrt{n}$.

MLE in a misspecified model revisited

For a simulated sample $X_1, \ldots, X_{100} \overset{\text{IID}}{\sim}$ Geometric(0.3):

Fisher information estimate $\sqrt{\bar{X}/n}$: 0.14

Sandwich estimate $S_X/\sqrt{n}$: 0.22

Parametric bootstrap: 0.14

Non-parametric bootstrap: 0.22

Even if the statistic T of interest is motivated by a parametric model (for example, $T = \hat{\theta}$ is the MLE in this model), the nonparametric bootstrap may be used to estimate the standard error of T to guard against model misspecification.

The real world and the bootstrap world

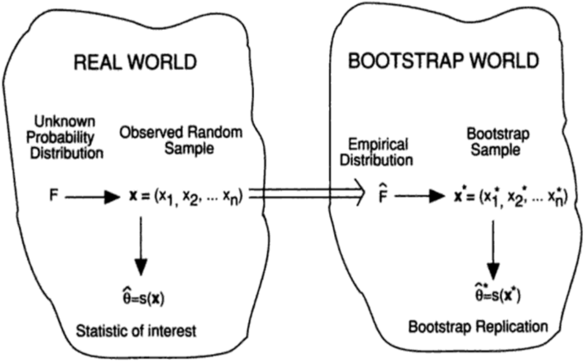

REAL WORLD
Unknown Probability Distribution
Observed Random Sample
$F \rightarrow \mathbf{x} = (x_1, x_2, \ldots x_n)$
$\hat{\theta} = s(\mathbf{x})$
Statistic of interest
BOOTSTRAP WORLD
Empirical Distribution
Bootstrap Sample
$\hat{F} \rightarrow \mathbf{x}^* = (x_1^*, x_2^*, \ldots x_n^*)$
$\hat{\theta}^* = s(\mathbf{x}^*)$
Bootstrap Replication

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

Bootstrap confidence intervals

There are many ways of using the bootstrap to construct a confidence interval for θ, using an estimator $\hat{\theta}$. We will discuss three simple methods in this class.

1. The normal interval:

Let $\hat{\theta} := \hat{\theta}\,(X_1, \ldots, X_n)$ be the estimate computed on the original data, and let $\widehat{\text{se}}$ be the bootstrap estimate of the standard error of $\hat{\theta}$. Construct a $100(1-\alpha)\%$ confidence interval as

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

Rationale: Here we replace an asymptotic estimate of the standard error of $\hat{\theta}$ with a bootstrap estimate.

It is valid if the distribution of $\hat{\theta}$ is approximately normal around θ.

Bootstrap confidence intervals

2. The percentile interval:

Let $\hat{\theta}^{*(\alpha/2)}$ and $\hat{\theta}^{*(1-\alpha/2)}$ be the $\alpha/2$ and $1-\alpha/2$ quantiles of the simulated values of $\hat{\theta}^*$. (If we performed B bootstrap simulations, these are the $(\alpha/2 \times B)^{\text{th}}$ and $((1-\alpha/2) \times B)^{\text{th}}$ ordered values of $\hat{\theta}^*$.) Construct a $100(1-\alpha)\%$ confidence interval as

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

Rationale: If $\hat{\theta}$ is close to θ, then the simulated distribution of $\hat{\theta}^*$ should be close to the theoretical distribution of $\hat{\theta}$.

Bootstrap confidence intervals

3. The "basic bootstrap" interval:

Estimate the distribution of $\hat{\theta} - \theta$ by the simulated distribution of $\hat{\theta}^* - \hat{\theta}$: The simulated $\alpha/2$ and $(1 - \alpha/2)$ quantiles of $\hat{\theta}^* - \hat{\theta}$ are $q_{\alpha/2} := \hat{\theta}^{*(\alpha/2)} - \hat{\theta}$ and $q_{1-\alpha/2} := \hat{\theta}^{*(1-\alpha/2)} - \hat{\theta}$. Since $\theta = \hat{\theta} - (\hat{\theta} - \theta)$, construct a $100(1 - \alpha)\%$ confidence interval for θ as

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

Rationale: The deviations of $\hat{\theta}^*$ from $\hat{\theta}$ in the "Bootstrap World" should approximate the deviations of $\hat{\theta}$ from θ in the "Real World".

Comparison of intervals

- If the distribution of $\hat{\theta}^*$ around $\hat{\theta}$ is symmetric, then the basic bootstrap interval and the percentile interval are equivalent $\left(\text{because } \hat{\theta}^{*(\alpha/2)} + \hat{\theta}^{*(1-\alpha/2)} \approx 2\hat{\theta}\right)$.
- If in addition, the distribution of $\hat{\theta}^*$ around $\hat{\theta}$ is normal, then these are equivalent to the normal interval.
- If the sampled values of $\hat{\theta}^*$ do not appear normally distributed around $\hat{\theta}$, then the normal interval should not be used.
- Rice[4] sticks to 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."

[4] Rice, John A. Mathematical statistics and data analysis.

Comparison of intervals

- Argument for the basic bootstrap interval: Suppose $\hat{\theta}$ is a positively biased estimate of θ. Then we expect $\hat{\theta}^*$ to be a positively biased estimate of $\hat{\theta}$. Hence the percentile interval "worsens" the bias of $\hat{\theta}$, whereas the basic bootstrap interval corrects for it.
- Argument for the percentile interval: It is invariant under reparametrization - let $\eta = g(\theta)$ where g is an increasing function. If we compute the percentile interval $\left[\theta^{*(\alpha/2)}, \theta^{*(1-\alpha/2)}\right]$ and then reparametrize, we get the interval $\left[g\left(\theta^{*(\alpha/2)}\right), g\left(\theta^{*(1-\alpha/2)}\right)\right]$ for η. If we reparametrize first by $\eta^* = g\left(\theta^*\right)$ and then compute the interval, we get $\left[\eta^{*(\alpha/2)}, \eta^{*(1-\alpha/2)}\right]$, which is the same thing. This doesn't hold for the basic bootstrap interval: The quantiles of $g\left(2\hat{\theta} - \hat{\theta}^*\right)$ are not the same as the quantiles of $2g(\hat{\theta}) - g\left(\hat{\theta}^*\right)$.

Other intervals

There are other bootstrap intervals with theoretical and empirical support for having more accurate coverage:

- Studentized bootstrap intervals - estimate the distribution of $\frac{\hat{\theta}-\theta}{\widehat{\mathrm{se}}(\hat{\theta})}$ using the simulated distribution of $\frac{\hat{\theta}^*-\hat{\theta}}{\widehat{\mathrm{se}}^*\left(\hat{\theta}^*\right)}$.
- Bias-corrected and accelerated intervals - explicitly adjust for the bias and skewness of the bootstrap distribution.

Moving Forward

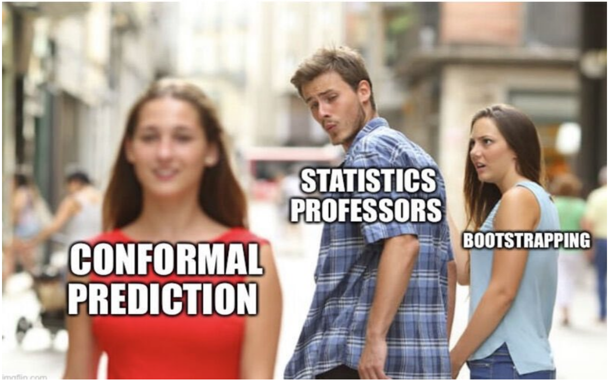


Image from https://twitter.com/predict_addict