MS&E 125: Intro to Applied Statistics

The Bootstrap

Professor Udell

Management Science and Engineering
Stanford

April 24, 2023

Announcements

- hw3 due Tuesday
- in-class quiz on Wednesday
- project proposal due Friday
- keep up the good participation! we can keep the zoom/async option as long as > 25 people are in the classroom

Outline

How to construct confidence interval?

- (last class) normal approximation with analytic formula for standard error
- use a normal approximation with bootstrap estimate for standard error
- use bootstrap quantiles

How to construct confidence interval?

- (last class) normal approximation with analytic formula for standard error
- use a normal approximation with bootstrap estimate for standard error
- use bootstrap quantiles

now suppose we have no model, only data $X_1, \ldots, X_n$

- can't compute analytic formula for standard error
- can't resample from the distribution

how to estimate uncertainty?

Motivating question

a **100 year flood** is a flood that has a 1% chance of occurring each year.

how can we estimate a "100 year flood" level using only data from one year?

Outline

Motivation

Empirical distribution

Bootstrap

Independent random variables

Definition

random variables X and Y are **independent** if

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

for all x and y.

(given the probability distributions of each), the value of X doesn't tell you anything about Y

Definition

random variables X and Y are **independent and identically distributed** (iid) if they are independent and $P(X = x) = P(Y = x)$ for all x.

Independent vs dependent examples

independent random variables:

- the amount of rainfall in two different cities
- the outcome of a coin toss
- the number of goals scored in a soccer match
- the closing stock price of two different companies
- the performance of a student on two different tests

dependent random variables:

- the number of cars sold by a dealership in one month compared to the previous month
- the amount of time it takes to complete a task versus the number of people working on it
- the height of a person compared to their weight
- the speed of a car compared to the amount of fuel it consumes
- the cost of a product compared to its demand

poll!

Empirical distribution

- given iid data $X_1, \ldots, X_n$,
- estimate the (CDF of the) distribution of X
- by the (CDF of the) **empirical distribution**

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq x\}},$$

the fraction of the data that is less than or equal to x.

Plug-in estimator

a **plug-in estimator** estimates a statistic θ (any function of the data) by plugging in the empirical distribution:

$$\hat{\theta}_n = \theta(\hat{F}_n).$$

examples:

- mean: $\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$
- standard deviation: $\hat{\theta}_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\theta}_n)^2}$

Plug-in estimator

a **plug-in estimator** estimates a statistic θ (any function of the data) by plugging in the empirical distribution:

$$\hat{\theta}_n = \theta(\hat{F}_n).$$

examples:

- mean: $\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$
- standard deviation: $\hat{\theta}_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\theta}_n)^2}$

how to estimate error or produce confidence intervals?

Outline

Bootstrap

idea:

- can't sample from the **model**
- instead, sample from the **data**

Definition

a **bootstrap sample** B_n is a sample of size n drawn **with replacement** from the data $X_1, \ldots, X_n$

$$\mathcal{B}_n = \{X_{i_1}, \ldots, X_{i_n}\},$$

where $i_1, \ldots, i_n$ are chosen uniformly at random from $\{1, \ldots, n\}$.

bootstrap **resamples** the data

Bootstrap

idea:

- can't sample from the **model**
- instead, sample from the **data**

Definition

a **bootstrap sample** B_n is a sample of size n drawn **with replacement** from the data $X_1, \ldots, X_n$

$$\mathcal{B}_n = \{X_{i_1}, \ldots, X_{i_n}\},$$

where $i_1, \ldots, i_n$ are chosen uniformly at random from $\{1, \ldots, n\}$.

bootstrap **resamples** the data

Q: How does the bootstrap sample differ from the original data?

Bootstrap

idea:

- can't sample from the **model**
- instead, sample from the **data**

Definition

a **bootstrap sample** B_n is a sample of size n drawn **with replacement** from the data $X_1, \dots, X_n$

$$\mathcal{B}_n = \{X_{i_1}, \dots, X_{i_n}\},$$

where $i_1, \dots, i_n$ are chosen uniformly at random from $\{1, \dots, n\}$.

bootstrap **resamples** the data

Q: How does the bootstrap sample differ from the original data?
A: Some data points are repeated, others are omitted

Demo: The bootstrap

https://colab.research.google.com/github/stanford-mse-125/demos/blob/main/bootstrap.ipynb

Ideal: sample from the model

for $k = 1, \ldots$

- sample new $X_i^k \sim P$, $i = 1, \ldots, n$, iid
  to form dataset $\mathcal{D}_k$
- estimate $\hat{\theta}_k = \theta(\mathcal{D}_k)$

Q: How sensitive is the prediction to the data set $\mathcal{D}$?

Ideal: sample from the model

for $k = 1, \ldots$

- sample new $X_i^k \sim P$, $i = 1, \ldots, n$, iid
  to form dataset $\mathcal{D}_k$
- estimate $\hat{\theta}_k = \theta(\mathcal{D}_k)$

Q: How sensitive is the prediction to the data set $\mathcal{D}$?
A: Look at histogram of $\{\theta_k\}_k$

Ideal: sample from the model

for $k = 1, \ldots$

- sample new $X_i^k \sim P$, $i = 1, \ldots, n$, iid
  to form dataset $\mathcal{D}_k$
- estimate $\hat{\theta}_k = \theta(\mathcal{D}_k)$

Q: How sensitive is the prediction to the data set $\mathcal{D}$?
A: Look at histogram of $\{\theta_k\}_k$
Q: Can we compute a **confidence interval** for the statistic θ?

Ideal: sample from the model

for $k = 1, \ldots$

- sample new $X_i^k \sim P$, $i = 1, \ldots, n$, iid
  to form dataset $\mathcal{D}_k$
- estimate $\hat{\theta}_k = \theta(\mathcal{D}_k)$

Q: How sensitive is the prediction to the data set $\mathcal{D}$?
A: Look at histogram of $\{\theta_k\}_k$
Q: Can we compute a **confidence interval** for the statistic θ?
A: Look at 95% confidence bound for $\{\theta_k\}_k$

Bootstrap: sample from the data

given dataset $\mathcal{D}$, for $k = 1, \ldots$

- sample $X_i^k \sim P$, $i = 1, \ldots, n$ **with replacement** from $\mathcal{D}$ to form dataset $\mathcal{D}_k$
- estimate $\hat{\theta}_k = \theta(\mathcal{D}_k)$

Q: How sensitive is the prediction to the data set $\mathcal{D}$?

Bootstrap: sample from the data

given dataset $\mathcal{D}$, for $k = 1, \ldots$

- sample $X_i^k \sim P$, $i = 1, \ldots, n$ **with replacement** from $\mathcal{D}$ to form dataset $\mathcal{D}_k$
- estimate $\hat{\theta}_k = \theta(\mathcal{D}_k)$

Q: How sensitive is the prediction to the data set $\mathcal{D}$?
A: Look at histogram of $\{\theta_k\}_k$

Bootstrap: sample from the data

given dataset $\mathcal{D}$, for $k = 1, \ldots$

- sample $X_i^k \sim P$, $i = 1, \ldots, n$ **with replacement** from $\mathcal{D}$ to form dataset $\mathcal{D}_k$
- estimate $\hat{\theta}_k = \theta(\mathcal{D}_k)$

Q: How sensitive is the prediction to the data set $\mathcal{D}$?
A: Look at histogram of $\{\theta_k\}_k$
Q: Can we compute a **confidence interval** for the statistic θ?

Bootstrap: sample from the data

given dataset $\mathcal{D}$, for $k = 1, \ldots$

- sample $X_i^k \sim P$, $i = 1, \ldots, n$ **with replacement** from $\mathcal{D}$ to form dataset $\mathcal{D}_k$
- estimate $\hat{\theta}_k = \theta(\mathcal{D}_k)$

Q: How sensitive is the prediction to the data set $\mathcal{D}$?
A: Look at histogram of $\{\theta_k\}_k$
Q: Can we compute a **confidence interval** for the statistic θ?
A: Look at 95% confidence bound for $\{\theta_k\}_k$

Bootstrap estimator for the variance

pick a function $h : \mathcal{D} \to \mathbf{R}$.
we want to estimate how much h varies when applied to finite data sets from the same distribution.

- resample $\mathcal{D}_1, \dots, \mathcal{D}_K$ from $\mathcal{D}$
- compute $h(\mathcal{D}_1), \dots, h(\mathcal{D}_K)$
- estimate the mean $\hat{\mu}_h = \frac{1}{K} \sum_{k=1}^{K} h(\mathcal{D}_k)$
- estimate the variance

$$\hat{\sigma}_h = \sqrt{\frac{1}{K} \sum_{k=1}^{K} (h(\mathcal{D}_k) - \hat{\mu}_h)^2}$$

Demo: The bootstrap

https://colab.research.google.com/github/stanford-mse-125/demos/blob/main/bootstrap.ipynb

Bootstrap confidence intervals

two ways to compute bootstrap confidence intervals:

- normal approximation:
  - use the bootstrap to estimate the variance of the statistic
- percentiles of bootstrapped distribution

Why does bootstrap work?

sample X_i^k with replacement from $\mathcal{D}$

$$
\begin{aligned}
& \mathbb{P}\left(X_1^1 = x\right) \\
= & \sum_{i=1}^{n} \mathbb{P}(\text{picked } X_i \text{ from } \mathcal{D} \text{ and was equal to } x) \\
= & \sum_{i=1}^{n} \mathbb{P}(\text{picked } X_i \text{ from } \mathcal{D})\, \mathbb{P}(X_i = x) \\
= & \sum_{i=1}^{n} \frac{1}{n}\, \mathbb{P}(x) \\
= & n\frac{1}{n}\, \mathbb{P}(x) \\
= & \mathbb{P}(x)
\end{aligned}
$$

Why does bootstrap work?

sample X_i^k with replacement from $\mathcal{D}$

$$\begin{aligned}
& \mathbb{P}\left(X_1^1 = x\right) \\
= & \sum_{i=1}^{n} \mathbb{P}(\text{picked } X_i \text{ from } \mathcal{D} \text{ and was equal to } x) \\
= & \sum_{i=1}^{n} \mathbb{P}(\text{picked } X_i \text{ from } \mathcal{D})\,\mathbb{P}(X_i = x) \\
= & \sum_{i=1}^{n} \frac{1}{n}\,\mathbb{P}(x) \\
= & n\frac{1}{n}\,\mathbb{P}(x) \\
= & \mathbb{P}(x)
\end{aligned}$$

so X_i^k has the same distribution as X_i (before conditioning on the data)

Why does bootstrap work?

$\mathcal{D}_k$ each have the same distribution as $\mathcal{D}$. So for any function $h : \mathcal{D} \to \mathbf{R}$,

$$\mathbb{E}_{\mathcal{D}} \frac{1}{K} \sum_{k=1}^{K} h(\mathcal{D}_k) = \mathbb{E}_{\mathcal{D}} h(\mathcal{D})$$

References

- The Bootstrap: http://www.stat.cmu.edu/~larry/=stat705/Lecture13.pdf. Wasserman, CMU Stat 705.