STAT 383C: Statistical Modeling I **Fall 2015**

Lecture 24 — November 19

Lecturer: Purnamrita Sarkar *Scribe: Evan Ott*

Disclaimer: These scribe notes have been slightly proofread and may have typos etc.

Note: The latex template was borrowed from EECS, U.C. Berkeley.

Table of Contents

24.1 Bootstrap

This lecture, we revisited bootstrap, considering the theory of why it works, a scenario where it will fail (and the conditions to look for in general), some modified strategies to account for various scenarios, and then some of the computational considerations for the procedure.

24.1.1 Theory

We'll start with $X_1, \cdots, X_n \overset{\text{iid}}{\sim} f_{\mu,\sigma^2}$, where f_{μ,σ^2} is some distribution with mean μ and variance σ^2. If we want the distribution of the sample mean, we can do this with a bootstrapped sample:

$$X_1^*, \cdots X_n^* \overset{\text{iid}}{\sim} F_n \tag{24.1}$$

where F_n is the empirical distribution derived from the original data set (so, a discrete distribution with values $X_1, \cdots X_n$ and probability $1/n$ for each value).

The true confidence interval for the original data set is:

$$C_n = \left[\hat{\mu} - t_{1-\alpha/2},\ \hat{\mu} - t_{\alpha/2}\right] \tag{24.2}$$

where

$$H_n\left(t_{\alpha/2}\right) = P\left(|\hat{\mu} - \mu| \sqrt{n} \le t_{\alpha/2}\right) = \frac{\alpha}{2} \tag{24.3}$$

where $\hat{\mu}$ is the sample mean $\bar{X}$. Under the bootstrapped data, we have the approximate confidence interval:

$$C_n^* = \left[\hat{\mu} - \hat{t}_{1-\alpha/2},\ \hat{\mu} - \hat{t}_{\alpha/2}\right] \tag{24.4}$$

where

$$\hat{H}_n\left(\hat{t}_{\alpha/2}\right) = P\left(|\hat{\mu}^* - \hat{\mu}|\sqrt{n} \leq \hat{t}_{\alpha/2}|X_1, \ldots, X_n\right) = \frac{\alpha}{2} \tag{24.5}$$

where $\hat{\mu}$ is the sample mean $\bar{X}$ and $\hat{\mu}^*$ is the bootstrap mean $\bar{X}^*$. By construction of the true confidence interval, we know that

$$P(\mu \in C_n) = 1 - \alpha \tag{24.6}$$

And so we would like it if

$$P\left(\mu \in C_n^*|X_1, \cdots X_n\right) \to 1 - \alpha \tag{24.7}$$

Note that, while $P\left(\hat{\mu} \in C_n^*|X_1, \cdots X_n\right) = 1 - \alpha$, it is not necessary that μ also has the same coverage probability of belonging to C_n^*. Said another way, we want $\hat{t}_{1-\alpha/2} \to t_{1-\alpha/2}$ or $H_n \approx \hat{H}_n$ (where this approximation symbol means "close" with high probability)[1]. We can begin with the following **fact**:

Fact 24.1. *If* $\sup_z \left|\hat{H}_n(z) - H_n(z)\right| \approx \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$*, then* $\forall \beta \in (0,1)$*, we have* $\hat{t}_\beta - t_\beta \approx O(1/\sqrt{n})$.

We can think about the relationship between distributions as

$$\begin{array}{ccc} & \overset{\text{limit}}{\rightarrow} & \\ H_n & \cdots & H \\ \vdots & & \vdots \\ \hat{H}_n & \cdots & \hat{H} \text{ (this convergence is conditional on the given data)} \\ & \underset{\rightarrow}{\text{limit}} & \end{array} \tag{24.8}$$

Now, from the Central Limit Theorem we know that

$$H_n(z) = P\left(\frac{|\hat{\mu} - \mu|\sqrt{n}}{\sigma} \leq \frac{z}{\sigma}\right) \to \Phi(z/\sigma) \tag{24.9}$$

However, we know something much stronger. The Berry-Esseen Theorem (24.1) tells us that:

Theorem 24.1 (Berry-Esseen). *For* $X_1, \ldots, X_n \overset{i.i.d}{\sim} f_{\mu,\sigma^2}$*, such that the third moment* $\mu_3 = E|X|^3 < \infty$*:*

$$\sup_z \left|H_n(z) - \Phi\left(\frac{z}{\sigma}\right)\right| \leq \frac{const \cdot \mu_3}{\sqrt{n}\sigma^3} \tag{24.10}$$

[1] For those of you who are picky this is the O_P notation which you will probably learn in prob theory.

So now we want to show that:

$$\sup_z \left|\hat{H}_n(z) - \hat{H}(z)\right| \approx \mathcal{O}\left(\frac{1}{\sqrt{n}}\right) \tag{24.11}$$

We can analyze the bootstrapped data (Equation 24.1) by showing:

$$\forall i \qquad E\left[X_i^*|X_1, \cdots, X_n\right] = \sum_{i=1}^{n} \frac{X_1}{n} = \bar{X} = \hat{\mu} \tag{24.12}$$

The final step is true because the empirical distribution puts 1/n mass on each datapoint. Thus,

$$\begin{aligned} E\left[\hat{\mu}^*|X_1, \cdots, X_n\right] &= E\left[\frac{\sum_{i=1}^n X_i^*}{n}\middle| X_1, \cdots, X_n\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} E\left[X_i^*|X_1, \cdots, X_n\right] = \hat{\mu} \end{aligned}$$

Furthermore,

$$\begin{aligned} \forall i, \qquad \mathrm{var}(X_i^*|X_1, \ldots, X_n) = &= E\left((X_1^*)^2\middle|X_1, \cdots, X_n\right) - \left(E\left(X_1^*|X_1, \cdots, X_n\right)\right)^2 \\ &= E\left((X_1^*)^2\middle|X_1, \cdots, X_n\right) - \hat{\mu}^2 \\ &= \frac{\sum_{i=1}^n X_i^2}{n} - \hat{\mu}^2 = \hat{\sigma}^2 \end{aligned} \tag{24.13}$$

This clearly also means that,

$$\begin{aligned} \mathrm{var}\left(\hat{\mu}^*|X_1, \cdots, X_n\right) &= \mathrm{var}\left(\frac{\sum_{i=1}^n X_i^*}{n}\middle| X_1, \cdots, X_n\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{var}\left(X_i^*|X_1, \cdots, X_n\right) = \frac{\hat{\sigma}^2}{n} \end{aligned}$$

The point is that, if we condition on the data, we can say that:

$$X_1^*, \cdots, X_n^*|X_1, \cdots, X_n \sim f^*_{\hat{\mu}, \hat{\sigma}^2} \tag{24.14}$$

We can now again use the Berry Esseen theorem to show that:

$$P\left(\sqrt{n}\frac{|\hat{\mu}^* - \hat{\mu}|}{\hat{\sigma}} \leq \frac{z}{\hat{\sigma}}\middle| X_1, \cdots, X_n\right) \approx \Phi\left(\frac{z}{\hat{\sigma}}\right) \tag{24.15}$$

We are using the fact that $E[|X_i^3| \mid X_1, \ldots, X_n] = \sum_i |X_i|^3/n := \hat{\mu}_3$ is bounded. This can be shown because $\hat{\mu}_3$ converges to μ_3 which in turn is bounded.

So this gives

$$\begin{array}{ccc} & \overset{\text{limit}}{\rightarrow} & \\ H_n & \cdots & \Phi\left(\frac{z}{\sigma}\right) \\ \vdots & & \vdots \\ \hat{H}_n & \cdots & \Phi\left(\frac{z}{\hat{\sigma}_n}\right) \\ & \overset{\text{limit}}{\rightarrow} & \end{array} \tag{24.16}$$

So for the first step we have $X_1, \cdots, X_n \overset{\text{iid}}{\sim}$ dist. We want the distribution of $\sqrt{n}\left(\hat{\mu}^* - \hat{\mu}\right)$ conditioned on the data. Using taylor expansion it can also be shown that

$$sup_z|\Phi\left(\frac{z}{\sigma}\right) - \Phi\left(\frac{z}{\hat{\sigma}_n}\right)| = O(1/\sqrt{n}, .$$

since $\hat{\sigma}_n \to \sigma$. Thus by the triangle inequality, $H_n(z) \to \hat{H}_n(z)$ is bounded by $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$.

To show this, we have

$$\sup_z |\hat{H}_n(z) - H_n(z)| \leq \sup_z \left|\hat{H}_n(z) - \Phi\left(\frac{z}{\hat{\sigma}}\right)\right| + \sup_z \left|\Phi\left(\frac{z}{\hat{\sigma}}\right) - \Phi\left(\frac{z}{\sigma}\right)\right| + \sup_z \left|H_n(z) - \Phi\left(\frac{z}{\sigma}\right)\right|.$$

Each of the steps is bounded by $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$ so the total is also bounded by $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$. So $|\hat{H}_n(z) - H_n(z)|$ is bounded by $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$.

So $\hat{H}_n(z)$ is bootstrapped to find the parameters which is related to the number of times we bootstrap by $\mathcal{O}\left(\frac{1}{\sqrt{B}}\right)$.

24.1.2 Bootstrap rule of thumb

The rule of thumb for using bootstrap is this: If you are trying to estimate the distribution of a statistic, bootstrap will work if the statistic converges asymptotically to a normal distribution. There are exceptions to this, but it's a good rule of thumb.

24.1.3 Subsampling and related strategies

Subsampling is very sensitive to the choice of B. If B is too large or too small, we get poor results.

Jackknife is another method (somewhat reminiscent of LOOCV) to estimate the variance and bias of an estimator. The jackknife variance is

$$\text{var}(\hat{\sigma}) \approx \frac{n-1}{n} \sum_{i=1}^{n} \text{var}\{\left(\hat{\theta}^{(-i)} - \sum_j \theta^{(-j)}/n\right)^2 \tag{24.17}$$

so we take out one data point and use the rest to estimate θ. This is very much just like taking all possible subsamples of length $n-1$ and using those. It turns out that this estimate is not consistent. It provides reasonable estimates on smooth distributions, but that's about it. For example, you **cannot** use this to estimate the variance of the median or quantiles, whereas Bootstrap can easily estimate that.

To help, we can also do delete-d jackknife, which is very close to normal subsampling, but it uses all possible subsets of a given length – again this is similar to subsampling. Additionally, we can do m out of n bootstrap. We'll look at something similar shortly.

24.1.4 Dependent data distributions

One place where we expect bootstrap to fail is for conditional distributions such as time series. If we randomly select points, we lose all temporal structure, which defeats the purpose. Instead, we split the data into blocks of length L and use bootstrap on the blocks (keeping the data within each block intact to preserve most of the dependency), as seen in Figure 24.1

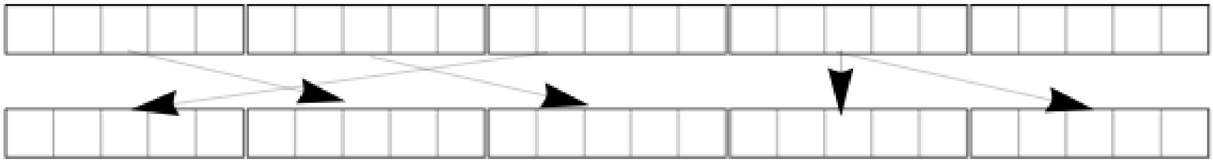

Figure 24.1. Graphic showing this method of bootstrap on time series data.

We could also do a moving bootstrap (sampling from $1 : B$ then $2 : B+1$, and so forth), or random walks, or random geometric lengths.

24.1.5 Computational considerations

Just like in linear regression where we were able to derive an efficient "online" solution where we are getting data in as a stream. We want to be able to not resample the whole distribution for every new data point. So, we can think about instead the number of times a particular data point is repeated in a bootstrapped sample:

$$P(X_1 \text{ appears i times}) = \text{Bin}\left(i, n, \frac{1}{n}\right) \to \text{Poisson}(i, 1) \tag{24.18}$$

So, for each new data point, we pick a random number of replications for that point from $Poisson(1)$. On average, we'll have n total data points in our new bootstrapped sample, and can keep generating the sample easily online.

We have another option which is the "bag of little bootstraps." We first obtain m without replacement samples of length b. From each of the m samples, we create r bootstrap samples

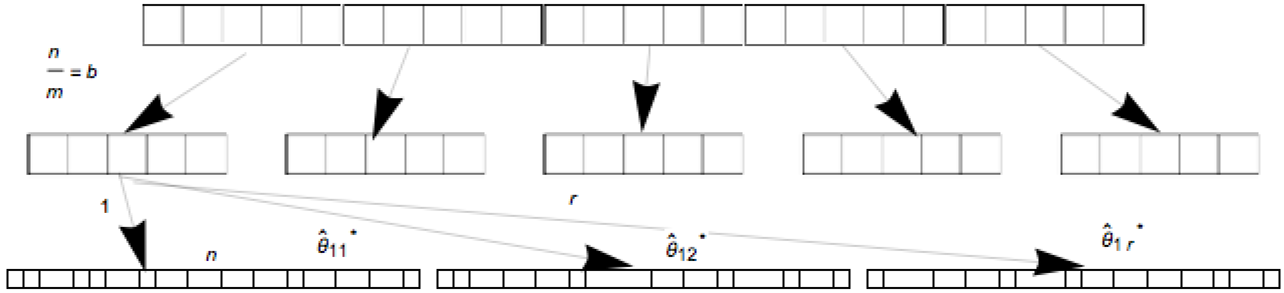

Figure 24.2. Demonstration of the bag of little bootstraps method. The original length n data is split into m subsamples (can be partitioned as above, or each sampled randomly) of length b. Each of these m blocks is used r times to create a sample of size n. These are used to estimate the parameter of interest.

of size n and use each of these to estimate $\hat{\theta}_{ij}^{*}$ for $1 \le i \le m$ and $1 \le j \le r$. This is most easily visualized in Figure 24.2

One nice feature of this algorithm is that the lowest layer of computation in the figure (creating the size n sample) can be stored very easily, storing the count of each data point in the sample. This means the data we store looks like Figure 24.3.

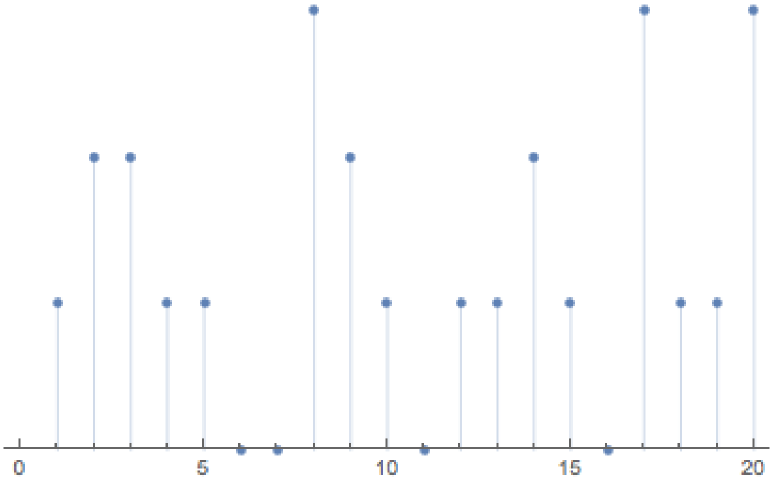

Figure 24.3. Graph indicating how data can be represented for the size n sample.