Lecture 2, part 2:

The Bootstrap

The Bootstrap

- The **bootstrap** is a method for estimating standard errors and confidence intervals.

- Let $T_n = g(X_1, \ldots, X_n)$ be a **statistic**, and we want to know $\mathbb{V}_F(T_n)$, the variance of T_n – it depends on the unknown distribution function F, thus the subscript.

- The idea of bootstrap has two steps:

Step 1: Estimate $\mathbb{V}_F(T_n)$ with $\mathbb{V}_{\widehat{F}}(T_n)$

Step 2: Approximate $\mathbb{V}_{\widehat{F}}(T_n)$ using **simulation** (because we don't always have a formula for it with empirical CDF, $\widehat{F}$)

The Bootstrap

Some reasoning behind bootstrap is:

- Suppose we draw an IID sample $Y_1, \ldots, Y_B$ from a distribution G. By the **law of large numbers**,

$$\overline{Y}_n = \frac{1}{B} \sum_{j=1}^{B} Y_j \xrightarrow{P} \int y \, dG(y) = \mathbb{E}(y) \quad \text{as} \quad B \to \infty$$

so if we draw a large sample, sample mean is a good estimate for $\mathbb{E}(Y)$. In a simulation, we can make B **as large as we like**.

In fact, for any function $h(Y)$ with **finite mean** its sample mean will converge (in P) to $\mathbb{E}(h(Y))$ – and so will the variance $\mathbb{V}(F)$.

The Bootstrap

So, how do we simulate? Basically, we simulate $X_1^*, \ldots X_n^*$ from $\widehat{F_n}$, and then compute $T_n^* = g(X_1^*, \ldots, X_n^*)$ – this makes one draw from the distribution of T_n.

How do we simulate $X_1^*, \ldots X_n^*$ from $\widehat{F_n}$? Notice that $\widehat{F_n}$ puts mass $1/n$ at each data point $X_1, \ldots, X_n$, so:

> drawing an observation from $\widehat{F_n}$ is **equivalent to drawing one point at random** from the original dataset!
>
> So to simulate $X_1^*, \ldots X_n^* \sim \widehat{F_n}$, one just **draws n observations with replacement from $X_1, \ldots, X_n$**

The Bootstrap

Here's how we **estimate variance** with of T_n with bootstrap:

1. Draw $X_1^*, \ldots X_n^* \sim \widehat{F}_n$

2. Compute $T_n^* = g(X_1^*, \ldots, X_n^*)$

3. Repeat steps 1 and 2, B times, to get $T_{n,1}^*, \ldots, T_{n,B}^*$

4. Let $V_{\text{boot}} = \frac{1}{B} \sum_{b=1}^{B} \left(T_{n,b}^* - \frac{1}{B} \sum_{r=1}^{B} T_{n,r}^* \right)^2$

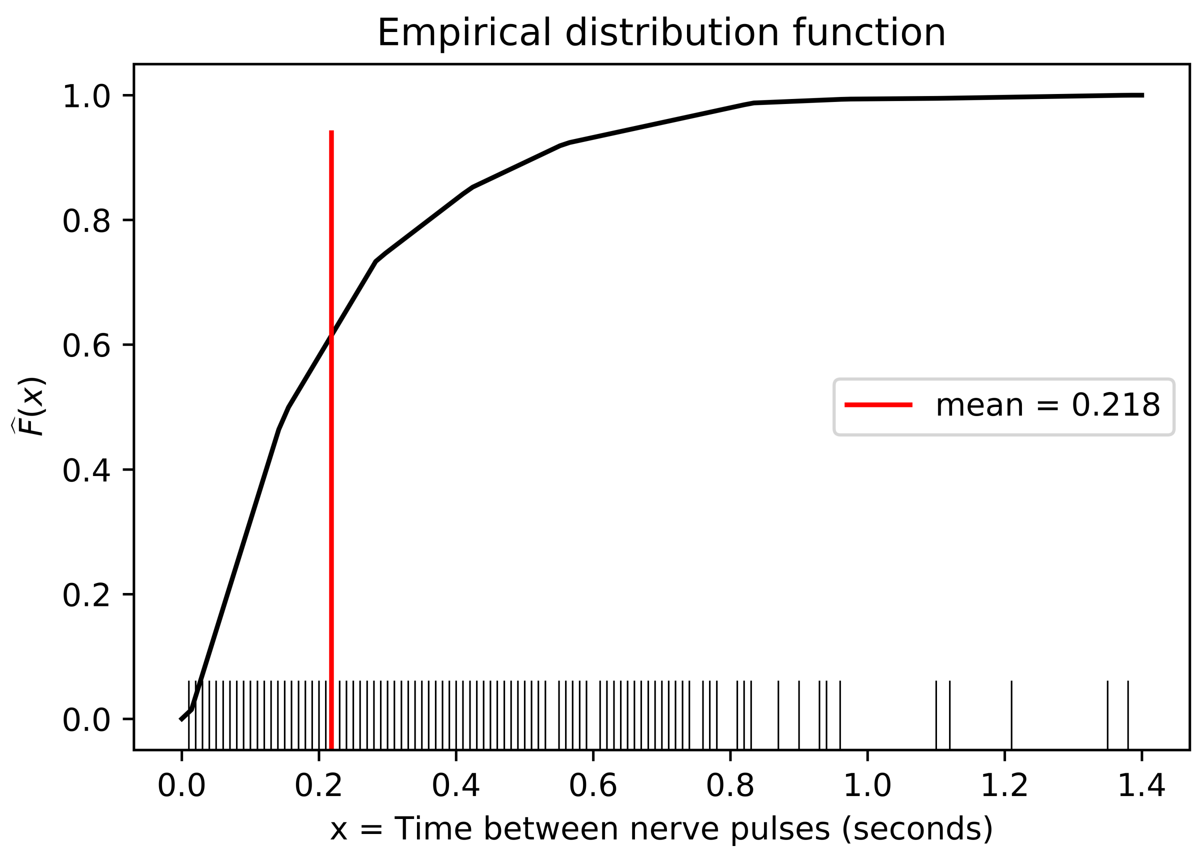
The Bootstrap

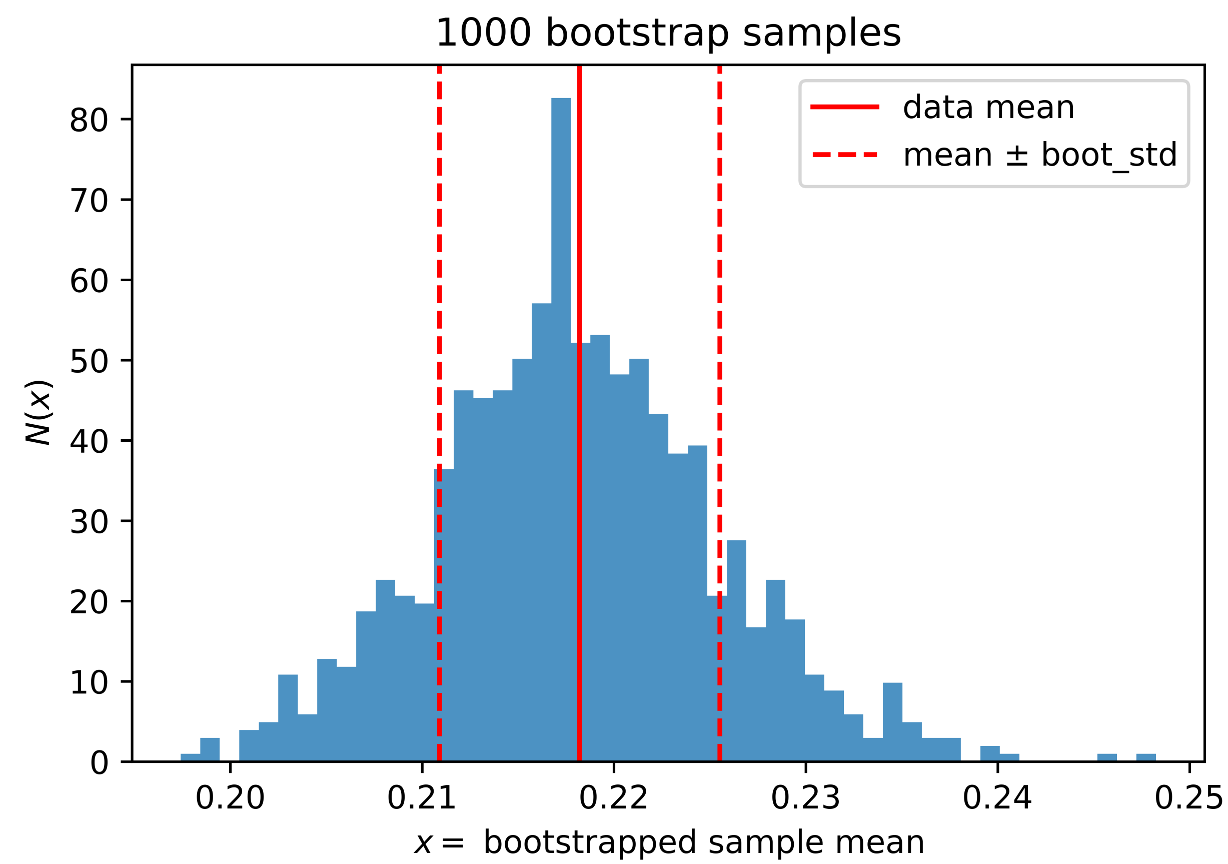
Here’s how to **sample with replacement** with Numpy.random:

```
In [1]: data = [1,3,3,4,7,9]

In [2]: from numpy.random import choice

In [3]: choice(data,3)
Out[3]: array([3, 3, 7])

In [4]: choice(data,3)
Out[4]: array([3, 1, 9])

In [5]: choice(data,3)
Out[5]: array([3, 9, 7])

In [6]: choice(data,3)
Out[6]: array([1, 3, 1])
```

The Bootstrap

Example 1: Let's get back to our nerve firing data and compute bootstrapped s.e. of the mean:





The Bootstrap: Confidence intervals

Now, with bootstrap giving us an estimate for the s.e. – how do we construct **confidence intervals**? One way is:

- **The Normal interval**:

$$T_n \pm z_{\alpha/2}\, \hat{\text{se}}_{boot}$$

  This interval is accurate if the distribution of T_n is close to normal, which is not always the case

Other methods rely on finding the quantiles of the “bootstrap distribution” itself, so we’ll cover them in practice on the seminar.

The Jackknife

- Before bootstrap was invented (by Bradley Efron, in 1979), there was another method for computing standard errors – the **jackknife** (due Quenouille, 1949). It is **less computationally expensive** than bootstrap, but is **less general**.

- Let $T_n = T(X_1, \ldots, X_n)$ be a statistic and $T_{(-i)}$ denote the statistic with i-th **observation removed**. Let $\overline{T}_n = n^{-1} \sum_{i=1}^{n} T_{(-i)}$ then the jackknife estimate of Var(T_n) is

$$v_{jack} = \frac{n-1}{n} \sum_{i=1}^{n} \left(T_{(-i)} - \overline{T}_n \right)^2$$

The Jackknife

- It can be shown that v_{jack} is a consistent estimate of Var(T_n) – in the sense that v_{jack}/Var$(T_n) \xrightarrow{P} 1$.

- However, unlike the bootstrap, it does not produce consistent estimates of the standard error of the sample quantiles.