

Introduction to Data Science With Probability and Statistics

Lecture 20: Kernel Density Estimates & The Bootstrap

CSCI 3022 - Summer 2020

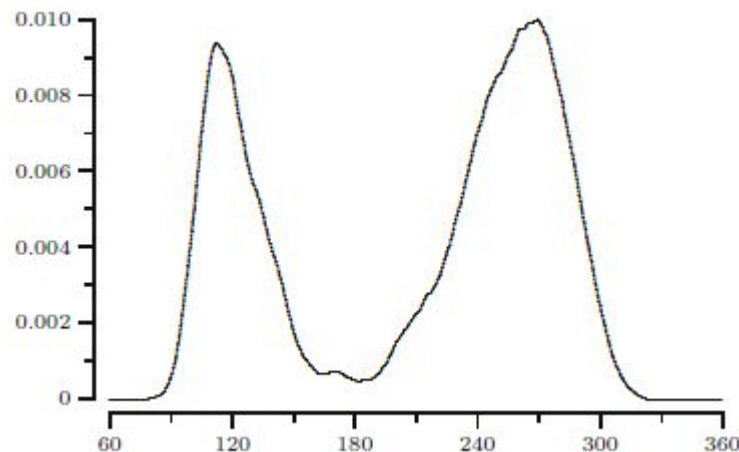
Sourav Chakraborty

Dept. of Computer Science

University of Colorado Boulder

Kernel density estimates

- We can graphically represent data in a more variegated plot by a so-called kernel density estimate.
- Kernel density estimate of a particular dataset is shown. The idea behind the construction of the plot is to “put a pile of sand” around each element of the dataset. At places where the elements accumulate, the sand will pile up.
- The actual plot is constructed by choosing a kernel K (reflecting shape of sand piles) and a bandwidth h (reflecting width of the sand piles).



Kernel density function conditions

A kernel K typically satisfies the following conditions:

(K1) K is a probability density, i.e., $K(u) \geq 0$ and $\int_{-\infty}^{\infty} K(u) du = 1$;

(K2) K is symmetric around zero, i.e., $K(u) = K(-u)$;

(K3) $K(u) = 0$ for $|u| > 1$.

Examples:-

1. *Epanechnikov kernel*

$$K(u) = \frac{3}{4} (1 - u^2) \quad \text{for } -1 \leq u \leq 1$$

3. *Normal Kernel*

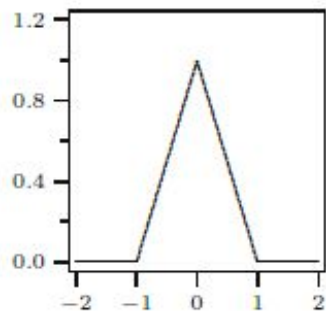
$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} \quad \text{for } -\infty < u < \infty.$$

2. *Triweight kernel*

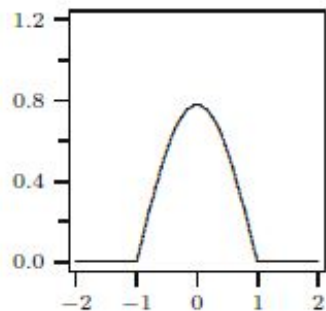
$$K(u) = \frac{35}{32} (1 - u^2)^3 \quad \text{for } -1 \leq u \leq 1$$



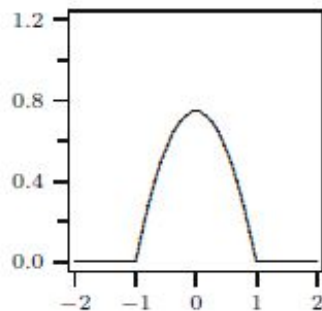
Examples of well-known kernels



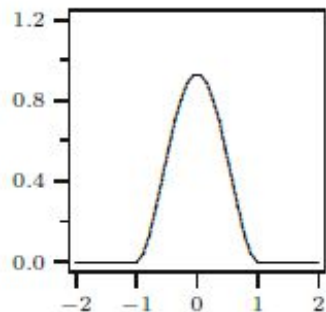
Triangular kernel



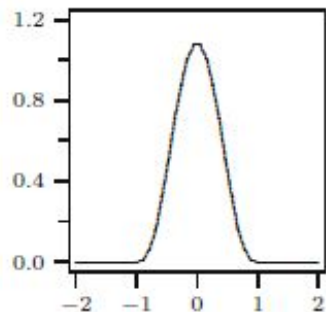
Cosine kernel



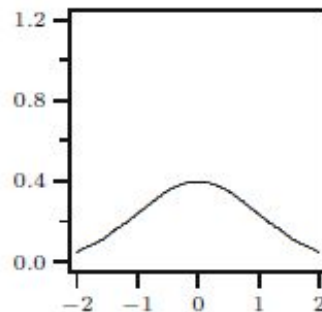
Epanechnikov kernel



Biweight kernel



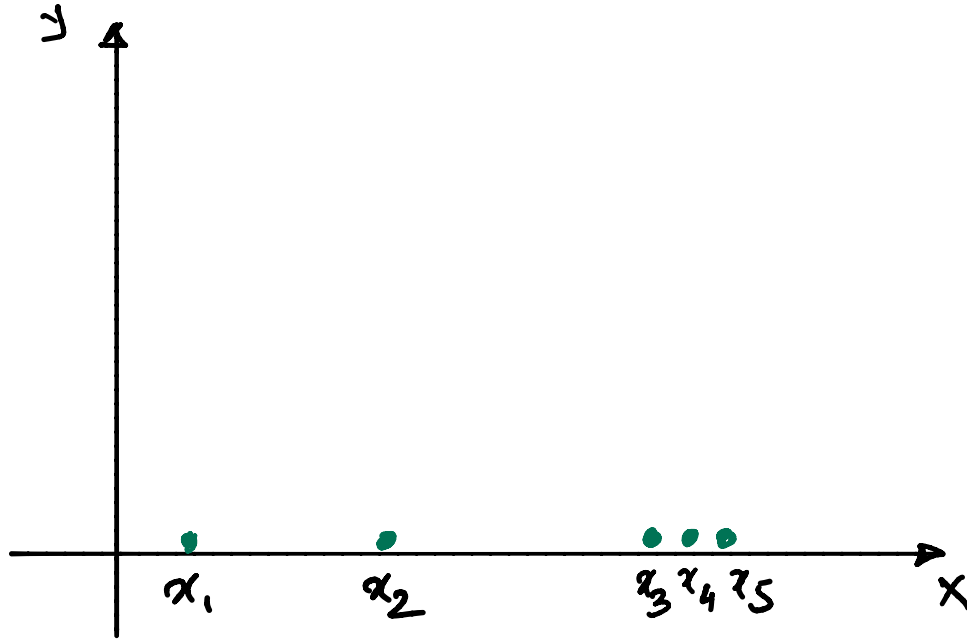
Triweight kernel



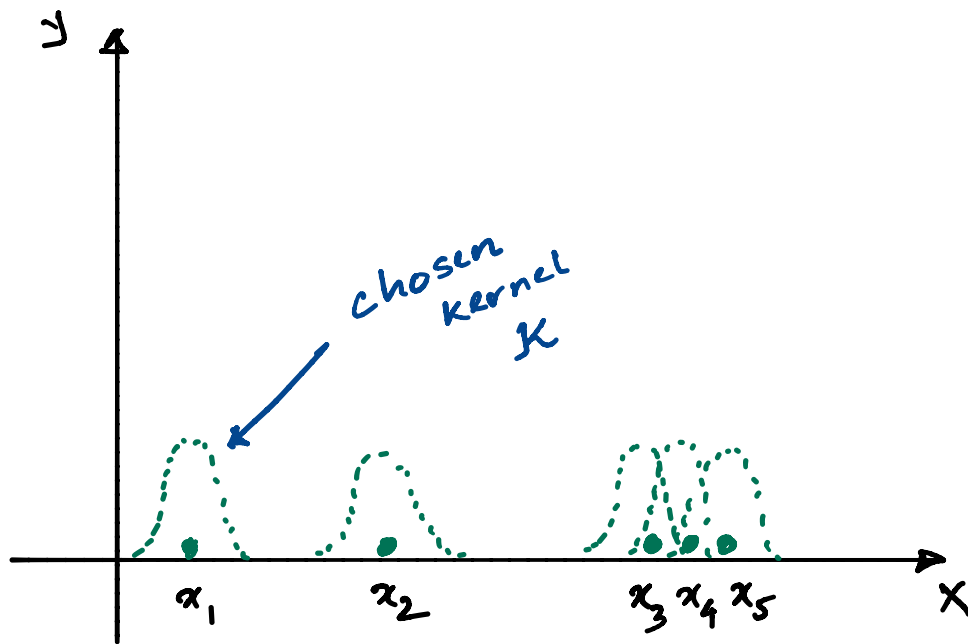
Normal kernel

Example construction

Dataset

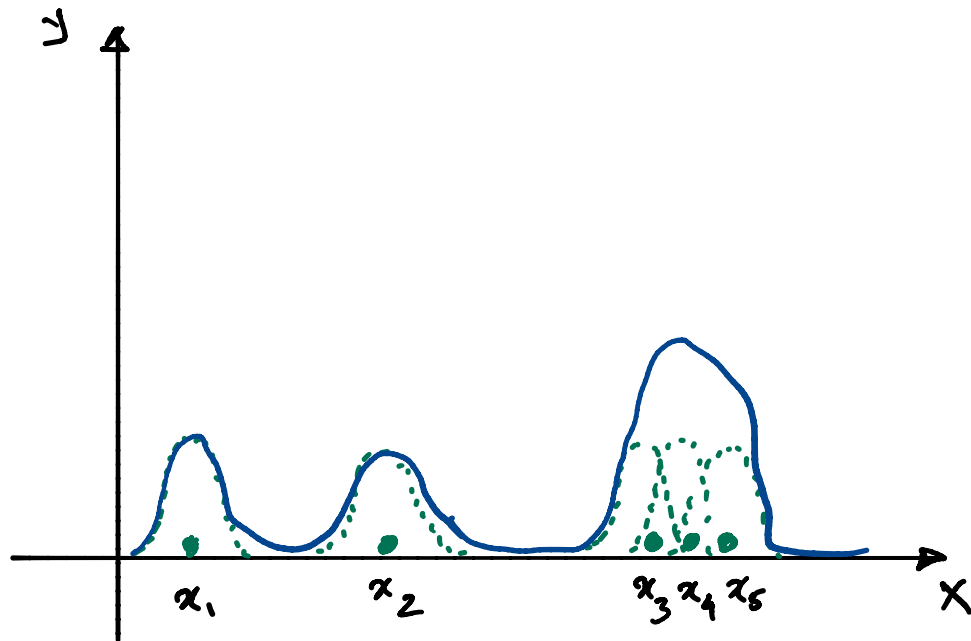


Example construction



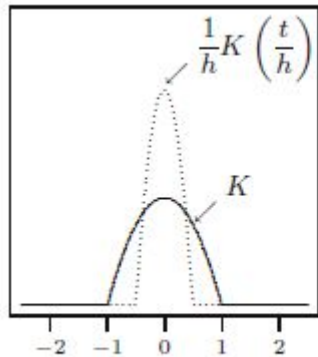
- We add the kernel on top of all the datapoints.

Example construction

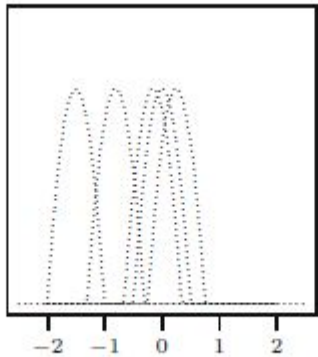


- Adding all the kernels .
- Normalizing to make the integral (sum) = 1.

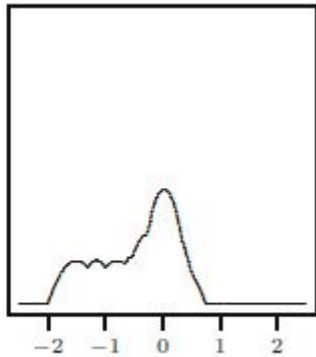
Construction of a kernel density estimate $f_{n,h}$



Kernel and scaled kernel



Shifted kernel



Kernel density estimate

$$f_{n,h}(t) \geq 0 \text{ and } \int_{-\infty}^{\infty} f_{n,h}(t) dt = 1.$$

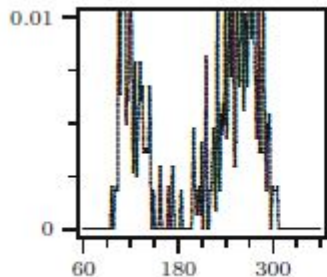
$$f_{n,h}(t) = \frac{1}{n} \left\{ \frac{1}{h} K\left(\frac{t-x_1}{h}\right) + \frac{1}{h} K\left(\frac{t-x_2}{h}\right) + \dots + \frac{1}{h} K\left(\frac{t-x_n}{h}\right) \right\}$$

$$\Rightarrow \boxed{f_{n,h}(t) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{t-x_i}{h}\right)}$$

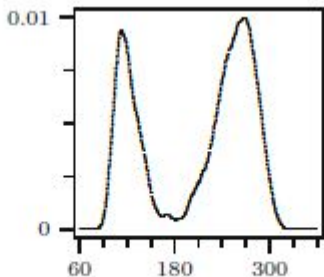
Choice of the bandwidth

The bandwidth h plays the same role for kernel density estimates as the bin width b does for histograms.

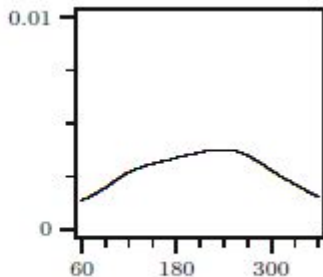
- Choosing the bandwidth too small will produce a curve with many isolated peaks.
- Choosing the bandwidth too large will produce a very smooth curve, at the risk of smoothing away important features of the data.



Bandwidth 1.8



Bandwidth 18



Bandwidth 180

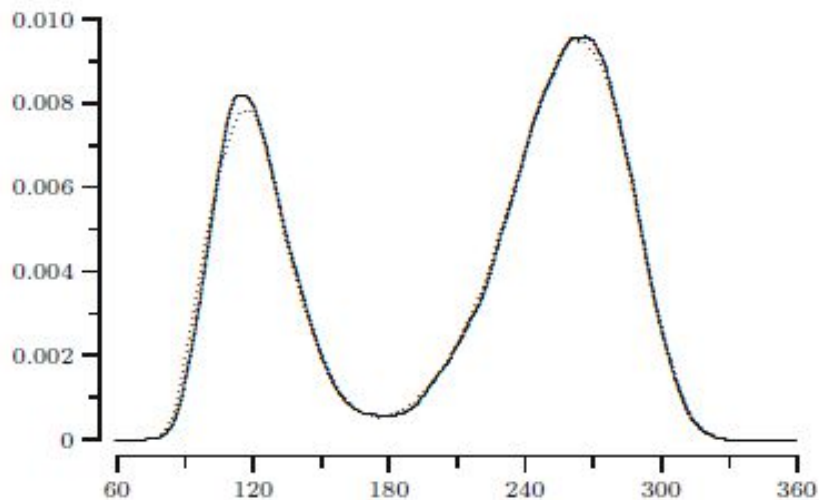
General guideline.

$$h = 1.06 s n^{-1/5}$$

Choice of the kernel

A kernel function determines the shape of the plot for kernel density function.

We can see different kernel representations of the Old faithful data below. Epanechnikov kernel (dotted), triweight kernel (solid line) both with bandwidth h as 24.



Bootstrap

- Bootstrap is a simulation procedure that approximates the distribution of the sample mean(or any other statistic) for *finite* sample size.
- The method is generally used when the sample size is small. As, large sample sizes give us the power to approximate the distribution of the sample mean by a normal distribution, via central limit theorem.



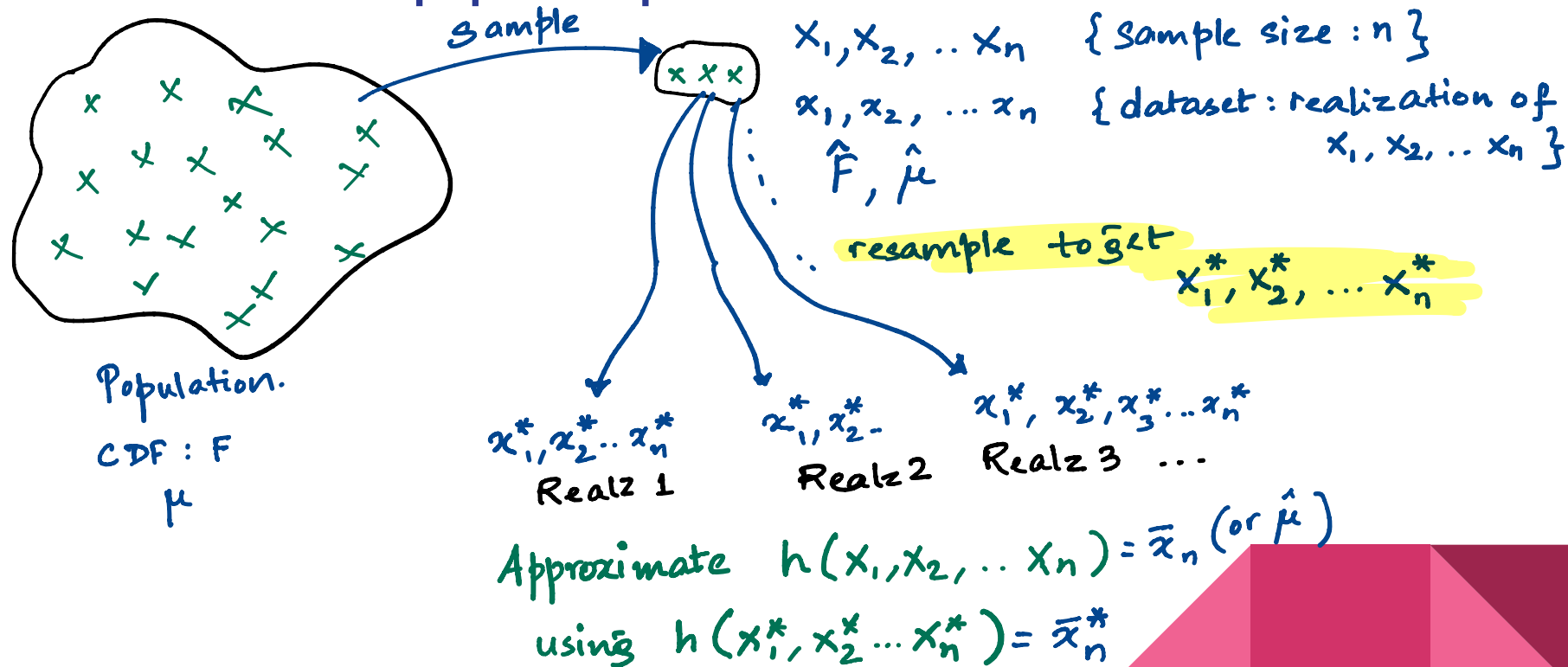
The bootstrap principle

BOOTSTRAP PRINCIPLE. Use the dataset x_1, x_2, \dots, x_n to compute an estimate \hat{F} for the “true” distribution function F . Replace the random sample X_1, X_2, \dots, X_n from F by a random sample $X_1^*, X_2^*, \dots, X_n^*$ from \hat{F} , and approximate the probability distribution of $h(X_1, X_2, \dots, X_n)$ by that of $h(X_1^*, X_2^*, \dots, X_n^*)$.

Example :

$$h(x_1, x_2, \dots, x_n) = \frac{x_1 + x_2 + \dots + x_n}{n} = \bar{x}_n$$

The bootstrap principle



The bootstrap principle

Describe how the bootstrap principle should be applied to approximate the distribution of $\text{Med}(X_1, X_2, \dots, X_n) - F^{\text{inv}}(0.5)$

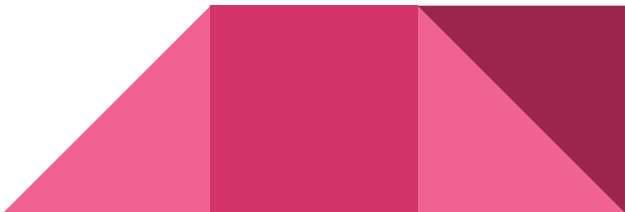
Note! we generally use centered sample mean : $\bar{X}_n - \mu$ for better estimation.

So corresponding centered sample median : $\text{Med}(x_1, \dots, x_n) - \underbrace{F^{-1}(0.5)}_{\text{pop. median}}$



The bootstrap principle

Describe how the bootstrap principle should be applied to approximate the distribution of $\text{Med}(X_1, X_2, \dots, X_n) - F^{\text{inv}}(0.5)$

1. Use x_1, x_2, \dots, x_n dataset to estimate \hat{F} .
 2. Resample $x_1^*, x_2^*, \dots, x_n^*$ from \hat{F} , multiple times.
 3. Approximate the prob. distribution of $\text{Med}(x_1, \dots, x_n) - F^{-1}(0.5)$ by $\text{Med}(x_1^*, \dots, x_n^*) - \hat{F}^{-1}(0.5)$
 4. Repeat 2 & 3 multiple times.
- 

The empirical bootstrap

EMPIRICAL BOOTSTRAP SIMULATION (FOR $\bar{X}_n - \mu$). Given a dataset x_1, x_2, \dots, x_n , determine its empirical distribution function F_n as an estimate of F , and compute the expectation

$$\mu^* = \bar{x}_n = \frac{x_1 + x_2 + \dots + x_n}{n}$$

corresponding to F_n .

1. Generate a bootstrap dataset $x_1^*, x_2^*, \dots, x_n^*$ from F_n .
2. Compute the centered sample mean for the bootstrap dataset:

$$\bar{x}_n^* - \bar{x}_n,$$

where

$$\bar{x}_n^* = \frac{x_1^* + x_2^* + \dots + x_n^*}{n}.$$

Repeat steps 1 and 2 many times.

when, you have no clue about F .



The empirical bootstrap

Describe the empirical bootstrap simulation for the centered sample median

$$\text{Med}(X_1, X_2, \dots, X_n) - F^{\text{inv}}(0.5)$$



The empirical bootstrap

Describe the empirical bootstrap simulation for the centered sample median

$$\text{Med}(X_1, X_2, \dots, X_n) - F^{\text{inv}}(0.5)$$

1. General empirical distribution func F_n from dataset $x_1, x_2 \dots x_n$, as an estimate for F .

2. Resample from F_n to get $x_1^*, x_2^*, \dots x_n^*$

3. Compute $\underbrace{\text{Med}_n^*}_{\text{Med}(x_1^*, x_2^*, \dots x_n^*)} - F_n^{-1}(0.5)$

$$\text{Med}(x_1^*, x_2^*, \dots x_n^*)$$

4. Repeat 2&3 multiple times.

An application of the empirical bootstrap

Consider a dataset modelled from a distribution F . Suppose we estimate the expectation μ corresponding to F by \bar{x}_n . Can we say how far away \bar{x}_n is from the “true” expectation μ ?

we want to know $\mathbb{P}(|\bar{x}_n - \mu| > k)$
↖ some threshold.

So;

$$\mathbb{P}(|\bar{x}_n - \mu| > k) \approx \mathbb{P}(|\bar{x}_n^* - \bar{x}_n| > k)$$



An application of the empirical bootstrap

Consider a dataset modelled from a distribution F . Suppose we estimate the expectation μ corresponding to F by \bar{x}_n . Can we say how far away \bar{x}_n is from the “true” expectation μ ?

Do resampling multiple times : $\bar{x}_{n,1}^* - \mu, \bar{x}_{n,2}^* - \mu, \dots, \bar{x}_{n,1000}^* - \mu$

$$\text{for } \mathbb{P}(|\bar{x}_n^* - \mu| > \kappa) = \frac{\text{number of times } |\bar{x}_{n,i}^* - \mu| > \kappa}{1000}$$



The parametric bootstrap

PARAMETRIC BOOTSTRAP SIMULATION (FOR $\bar{X}_n - \mu$). Given a dataset x_1, x_2, \dots, x_n , compute an estimate $\hat{\theta}$ for θ . Determine $F_{\hat{\theta}}$ as an estimate for F_{θ} , and compute the expectation $\mu^* = \mu_{\hat{\theta}}$ corresponding to $F_{\hat{\theta}}$.

1. Generate a bootstrap dataset $x_1^*, x_2^*, \dots, x_n^*$ from $F_{\hat{\theta}}$.
2. Compute the centered sample mean for the bootstrap dataset:

$$\bar{x}_n^* - \mu_{\hat{\theta}},$$

where

$$\bar{x}_n^* = \frac{x_1^* + x_2^* + \dots + x_n^*}{n}.$$

Repeat steps 1 and 2 many times.

from example if pop. is $\text{Exp}(\lambda)$ distribution. we can find $\hat{\lambda} = \frac{1}{\bar{x}_n}$.

All the best for Midterm!

