Given data $X_1, X_2, \ldots, X_n \sim F$ iid, define

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_i \le x)$$

$$\mathbb{E}[\hat{F}_n(x)] = F(x)$$

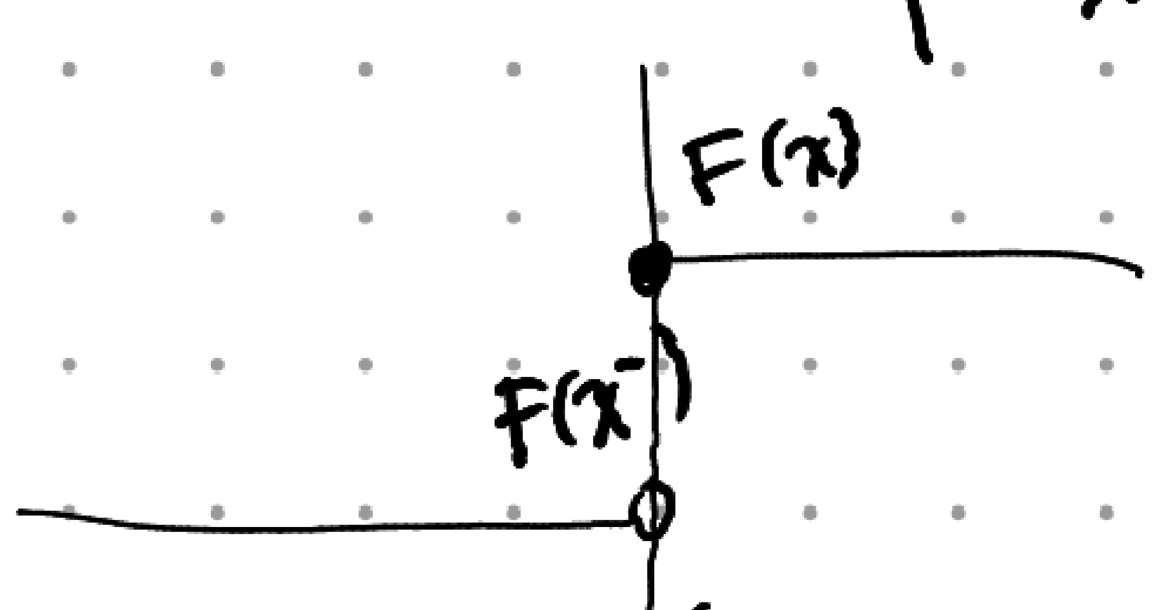
$$\mathbb{V}[\hat{F}_n(x)] = \frac{F(x)(1-F(x))}{n}$$

$$\mathbb{P}\left[\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| > \varepsilon\right] \le 2e^{-2n\varepsilon^2}$$

Notation

F a distribution function

$$\int r(x)\,dF(x) = \begin{cases} \int r(x) f(x)\,dx & f(x) = F'(x) \\ \sum_x r(x) f(x) & f(x) = F(x) - F(x^-) \end{cases}$$



Statistical Functionals

Def. A statistical functional $T(F)$ is any function of a distribution F

Ex. $\mu = \int x \, dF(x) = \int x f(x) \, dx$

$$\sigma^2 = \int (x-\mu)^2 \, dF(x) = \int (x-\mu)^2 f(x) \, dx$$

Def The plug-in estimate of $\theta = T(F)$ is

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

Def If $T(x) = \int r(x) \, dF(x)$ for some func $r(x)$ (not depending on F), then $T(x)$ is a linear functional

Theorem The plug in estimator for a linear functional $T(x) = \int r(x) \, dF(x)$ is

$$T(\hat{F}_n) = \int r(x) \, d\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} r(X_i)$$

$\underline{\text{Ex.}}$ $\mu = T(F) = \int x \, dF(x)$ (linear)

Plug in estimator:

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} X_i = \overline{X}_n$$

$\underline{\text{Ex.}}$ $\sigma^2 = T(F) = \int x^2 \, dF(x) - \left(\int x \, dF(x) \right)^2$ (not linear)

Plug in estimator:

$$\begin{aligned}
\hat{\sigma}_n^2 &= \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n} \sum_{i=1}^{n} X_i \right)^2 \\
&= \left(\frac{1}{n} \sum_{i=1}^{n} X_i^2 \right) - \left(2 \frac{1}{n} \sum_{i=1}^{n} X_i \overline{X}_n \right) + \overline{X}_n^2 \\
&= \frac{1}{n} \sum_{i=1}^{n} \left(X_i^2 - 2 X_i \overline{X}_n + \overline{X}_n^2 \right) \\
&= \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2
\end{aligned}$$

Compare with: $\hat{S}_n = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$

Bootstrapping

$X_1, X_2, \dots, X_n \sim F$ iid.

Suppose $T_n = g(X_1, X_2, \dots, X_n)$ is a statistic.

How can we estimate $\mathbb{V}[T_n]$?

Ex. $T_n = \bar{X}_n = \frac{1}{n}(X_1 + \dots + X_n)$

$\mathbb{V}[T_n] = \frac{\sigma^2}{n}$, but we don't know σ^2 !!!

Idea

1. Approximate T_n with T_n^*, where

$$T_n^* = g(X_1^*, X_2^*, \dots, X_n^*), \quad X_1^*, X_2^*, \dots, X_n^* \sim \hat{F}_n$$

2. Compute (or approximate) $\mathbb{V}[T_n^* \mid X_1, \dots, X_n]$

Variance estimation

Suppose we cannot compute $\mathbb{V}[T_n^* \mid X_1, \ldots, X_n]$.

We can estimate it by sampling.

1. Draw $X_1^*, \ldots, X_n^* \sim \hat{F}_n$
2. Compute $T_n^* = g(X_1^*, \ldots, X_n^*)$
3. Repeat steps 1, 2 to get

   $T_{n,1}^*, T_{n,2}^*, \ldots, T_{n,B}^*$

4. Use estimate

$$\underbrace{v_{boot} = \frac{1}{B} \sum_{b=1}^{B} \left(T_{n,b}^* - \underbrace{\left(\frac{1}{B} \sum_{r=1}^{B} T_{n,r}^* \right)}_{\text{sample mean}} \right)^2}_{\text{sample var}}$$