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

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

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

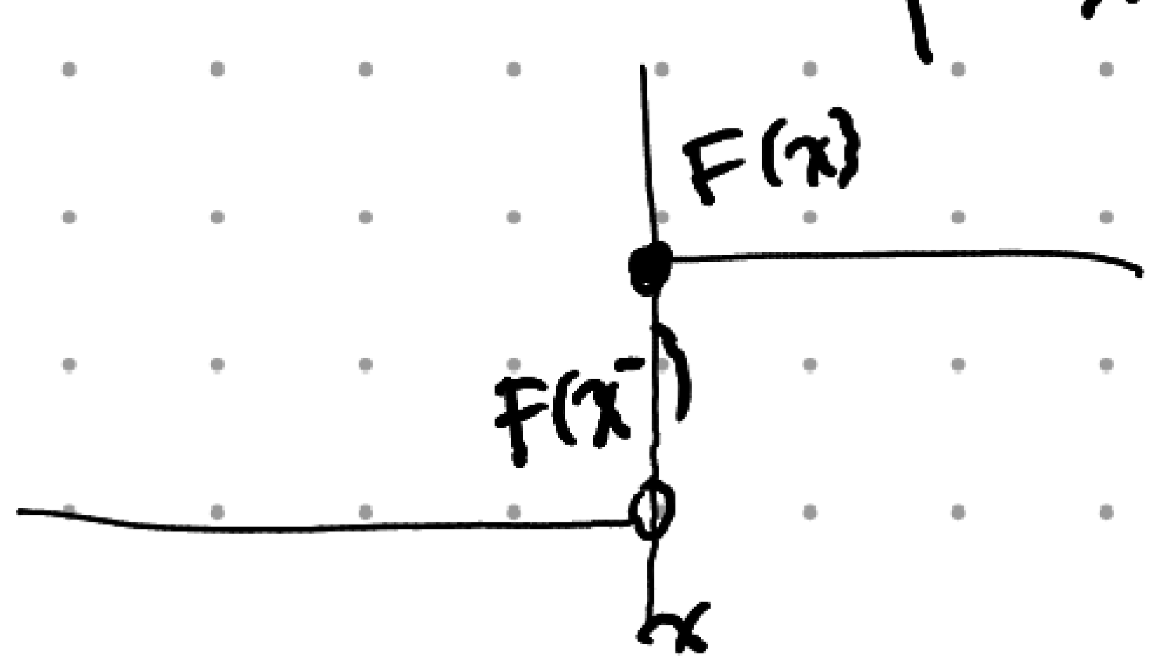
$$\mathbb{V}[\hat{F}_n(x)] = \frac{F(x)(1-F(x))}{n}$$

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

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 <u>statistical functional</u> $T(F)$ 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 <u>linear functional</u>

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

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

Ex. $\mu = T(F) = \int x \, dF(x)$ (linear)

Plug in estimator:

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

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

Plug in estimator:

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2$$

$$= \left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - \left(2\frac{1}{n}\sum_{i=1}^{n} X_i \bar{X}_n\right) + \bar{X}_n^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(X_i^2 - 2X_i\bar{X}_n + \bar{X}_n^2\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

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

Bootstrapping

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

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

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

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

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

Idea

1. Approximate T_n with T_n^*, where

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

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

<u>Ex.</u> $\hat{T}_n = \frac{1}{n}(\hat{X}_1 + \cdots + \hat{X}_n)$

$$\mathbb{V}[T_n^* | X_1, \ldots, X_n] = \frac{1}{n^2} \sum_{i=1}^{n} \mathbb{V}[X_i^* | X_1, \ldots, X_n]$$

$$= \frac{1}{n} \mathbb{V}[X_1^* | X_1, \ldots, X_n]$$

$$\mathbb{E}[X_1^* | X_1, \ldots, X_n] = \int x \, d\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}_n$$

$$\mathbb{E}[X_1^{*2} | X_1, \ldots, X_n] = \int x^2 \, d\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} X_i^2$$

$$\mathbb{V}[T_n^* | X_1, \ldots, X_n] = \frac{1}{n}\left(\left(\frac{1}{n} \sum_{i=1}^{n} X_i^2\right) - \left(\frac{1}{n} \sum_{i=1}^{n} X_i\right)^2 \right)$$

$$= \hat{\sigma}_n^2$$

so we don't even need to approximate this quantity.

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

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