$\mathcal{F}$ a statistical model (set of dists)

$F \in \mathcal{F}$ unknown but fixed.

Get data $X_1, X_2, \dots X_n \sim F$ iid

Goal: Learn sth. about F (e.g. mean, var. parameter p, etc.)

Distribution Approximation

Suppose $X_1, X_2, \dots, X_n \sim F$. How can we estimate $F(x)$ for a given $x \in \mathbb{R}$?

- For each $x \in \mathbb{R}$, $y = F(x)$ is a parameter
- There are infinitely many parameters influencing F

$$F(x) = \mathbb{P}[X \leq x]$$

Let's look at the fraction of X_i below x

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

$$\mathbb{1}(\text{true}) = 1$$

$$\mathbb{1}(\text{false}) = 0$$

$$= \frac{\# X_i \leq x}{n}$$

$\hat{F}_n$ is called the empirical CDF.

$$\mathbb{1}(X \le x) = \begin{cases} 1 & X \le x \\ 0 & X > x \end{cases}$$

$$\left.\begin{aligned} \mathbb{P}[\mathbb{1}(X \le x) = 1] &= \mathbb{P}[X \le x] = F(x) \\ \mathbb{P}[\mathbb{1}(X > x) = 0] &= 1 - F(x) \end{aligned}\right\} \mathbb{1}(X \le x) \sim \mathrm{Ber}(F(x))$$

Theorem. For any $x \in \mathbb{R}$,

$$\left\{\begin{aligned} \mathbb{E}[\hat{F}(x)] &= F(x) \\ \mathbb{V}[\hat{F}(x)] &= \frac{F(x)(1 - F(x))}{n} \\ MSE &= \frac{F(x)(1 - F(x))}{n} \end{aligned}\right.$$

$$\hat{F}_n(x) \xrightarrow{P} F(x)$$

Theorem (Glivenko-Cantelli)

$$\left(\sup_x | \hat{F}_n(x) - F(x) \right) \xrightarrow{P} 0$$

Theorem (Dvoretzky-Kiefer-Wolfowitz)

For any $\varepsilon > 0$,

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

Find $L_n(x)$, $U_n(x)$ s.t.

$$\mathbb{P}\left[F(x) \in (L_n(x), U_n(x)) \ \forall x \right] \geq 1-\alpha$$

$$\mathbb{P}\left[\overbrace{\hat{F}_n(x) - \varepsilon}^{L_n(x)} \leq F(x) \leq \overbrace{\hat{F}_n(x) + \varepsilon}^{U_n(x)} \right]$$

$$= \mathbb{P}\left[\, | F(x) - \hat{F}_n(x) | \leq \varepsilon \right]$$

$$= 1 - \mathbb{P}\left[\, | \hat{F}_n(x) - F(x) | > \varepsilon \right] \qquad \text{(DKW)}$$

$$\geq 1 - 2e^{-2n\varepsilon^2}$$

Find ε st $\alpha = 2e^{-2n\varepsilon^2}$

$$\Rightarrow \varepsilon = \sqrt{\frac{1}{2n} \log\left(\frac{2}{\alpha}\right)}$$

$$L_n(x) = \hat{F}_n(x) - \sqrt{\frac{1}{2n} \log\left(\frac{2}{\alpha}\right)}$$

$$U_n(x) = \hat{F}_n(x) + \sqrt{\frac{1}{2n} \log\left(\frac{2}{\alpha}\right)}$$