ESTIMATING THE CDF

Let $X_1, \dots, X_n$ be iid with continuous cdf F.

Defn: The Empirical cdf is the function

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

Result: For a fixed $x_0 \in \mathbb{R}$ we have

$$\sqrt{n}\left(\hat{F}_n(x_0) - F(x_0)\right) \xrightarrow{D} N\left(0,\ F(x_0)[1 - F(x_0)]\right).$$

Proof: Note that

$$\mathbb{E}\ \mathbb{1}(X_1 \le x_0) = F(x_0)$$
$$\mathrm{Var}\ \mathbb{1}(X_1 \le x_0) = F(x_0)[1 - F(x_0)],$$

Then the result follows directly from the Central Limit Theorem.

Result: For F any continuous cdf, we have

(Wass, pg 141) (i) $\sup_{x \in \mathbb{R}} \left| \hat{F}_n(x) - F(x) \right| \to 0$ w.p. 1 (Glivenko-Cantelli)

(Wass, pg 14) (ii) $P\left(\sup_{x \in \mathbb{R}} \left| \hat{F}_n(x) - F(x) \right| \le \varepsilon \right) \ge 1 - 2e^{-2n\varepsilon^2}$ (DKW)

[See Massart (1990)] (iii) $\sqrt{n}\left(\sup_{x \in \mathbb{R}} \left| \hat{F}_n(x) - F(x) \right| \right) \xrightarrow{D} \sup_{t \in [0,1]} |B_0(t)|$,

where $\{B_0(t) : t \in [0,1]\}$ is a Brownian bridge (defined below).

Defn: A Wiener process B is a random function in the space $C[0,1]$ of continuous functions on $[0,1]$ which satisfies
(Serf, pg 14)

(a) $B(0) = 0$ with probability 1.

(b) $B(t) \sim N(0, t)$, $t \in (0,1]$

(c) For $0 \le t_0 \le t_1 \le \cdots \le t_k \le 1$, the increments

$$B(t_1) - B(t_0), \ldots, B(t_k) - B(t_{k-1})$$

are mutually independent.

Also called "Standard Brownian Motion."

Result: (Allows us to simulate Brownian motion) For each $n \ge 1$, let

$$B_n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor tn \rfloor} Z_i, \qquad Z_1, \ldots, Z_n \overset{ind}{\sim} N(0,1).$$

Then B_n converges to B by a theorem called Donsker's Thm.

Defn: A Brownian bridge is the random function in $C[0,1]$ given by
(Athla, pg 375)

$$B_0(t) = B(t) - tB(1),$$

where B is a standard Brownian motion.

Remark: The "bridge" begins and ends at $0 = B_0(0) = B_0(1)$.

Exercise: Build confidence bands for F:

- with DKW result.
- with KS result (simulate to get needed quantile).

Solution: To use the DKW result, we write

$$P\left(\sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F(x)| \le \varepsilon\right) \ge 1 - 2e^{-2n\varepsilon^2} = 1-\alpha$$

Now we have

$$1 - 2e^{-2n\varepsilon^2} = 1-\alpha$$

$$\iff$$

$$-2n\varepsilon^2 = \log\left(\frac{\alpha}{2}\right)$$

$$\iff$$

$$\varepsilon = \sqrt{\frac{\log(2)-\log(\alpha)}{2n}}$$

So $(1-\alpha)100\%$ confidence bands may be constructed as

$$\hat{F}_n(x) \pm \sqrt{\frac{\log(2)-\log(\alpha)}{2n}} \quad \text{for all } x\in\mathbb{R}.$$

From the KS result we would construct the interval

$$\hat{F}_n(x) \pm \frac{1}{\sqrt{n}} B_{0,\alpha/2} \quad \text{for } x\in\mathbb{R},$$

where $B_{0,\alpha/2}$ represents the upper $\alpha/2$ quantile of $\sup_{t\in[0,1]} |B_0(t)|$.

Result: For a Brownian bridge B_0, we have

[See Massart (1990)] $$P\left(\sup_{t\in[0,1]} |B_0(t)| \le x\right) = 1 - 2\sum_{i=1}^{\infty} (-1)^{i+1} \exp\left(-2i^2x^2\right).$$

Exercise: Check accuracy of simulated quantiles of $\sup_{t\in[0,1]} |B_0(t)|$.