M3/M4S3 STATISTICAL THEORY II

THE GLIVENKO-CANTELLI LEMMA

Definition : The Empirical Distribution Function
Let $X_1, \ldots, X_n$ be a collection of i.i.d. random variables with cdf F_X. Then the *empirical distribution function* will be denoted $F_n(x)$, and defined for $x \in \mathbb{R}$ by

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{[X_i,\infty)}(x)$$

where $I_A(\omega)$ is the indicator function for set A.

If data $x_1, \ldots, x_n$ are available, then the *observed* or *estimated* empirical distribution function is denoted $\widehat{F}_n(x)$ and defined by

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{[x_i,\infty)}(x).$$

Note that for any **fixed** $x \in \mathbb{R}$, the Strong Law of Large Numbers ensures that

$$F_n(x) \xrightarrow{a.s.} F_X(x) \qquad \text{as } n \longrightarrow \infty$$

as

$$E[I_{[X_i,\infty)}(x)] = P[I_{[X_i,\infty)}(x) = 1] = P[X_i \leq x] = F_X(x).$$

This result is strengthened by the following Theorem.

Theorem 1.9 ***The Glivenko-Cantelli Theorem***
Let $X_1, \ldots, X_n$ be a collection of i.i.d. random variables with cdf F_X, and let $F_n(x)$ denote the empirical distribution function. Then, as $n \longrightarrow \infty$,

$$P\left[\sup_{x\in\mathbb{R}} |F_n(x) - F_X(x)| \longrightarrow 0\right] = 1$$

or equivalently

$$P\left[\lim_{n\longrightarrow\infty} \sup_{x\in\mathbb{R}} |F_n(x) - F_X(x)| = 0\right] = 1.$$

*that is, the convergence is **uniform in x**.*

Proof. Let $\epsilon > 0$. Then fix $k > 1/\epsilon$, and then consider "knot" points $\kappa_0, \ldots, \kappa_k$ such that

$$-\infty = \kappa_0 < \kappa_1 \leq \kappa_2 \leq \ldots \leq \kappa_{k-1} < \kappa_k = \infty$$

that define a partition of $\mathbb{R}$ into k disjoint intervals such that

$$F_X(\kappa_j^-) \leq \frac{j}{k} \leq F_X(\kappa_j) \qquad j = 1, \ldots, k-1$$

where, for each j,

$$F_X(\kappa_j^-) = P[X_j < \kappa_j] = F_X(\kappa_j) - P[X = \kappa_j].$$

Then, by construction, if $\kappa_{j-1} < \kappa_j$,

$$F_X(\kappa_j^-) - F_X(\kappa_{j-1}) \leq \frac{j}{k} - \frac{(j-1)}{k} = \frac{1}{k} < \epsilon.$$

Recall in the following that $F_n(x)$ is a **random** quantity. Now, by the Strong Law, we have pointwise convergence, so that, as $n \longrightarrow \infty$, for $j = 1, \ldots, k-1$.

$$F_n(\kappa_j) \xrightarrow{a.s.} F_X(\kappa_j) \qquad \text{and} \qquad F_n(\kappa_j^-) \xrightarrow{a.s.} F_X(\kappa_j^-).$$

Then it immediately follows that, for each j,

$$|F_n(\kappa_j^-) - F_X(\kappa_j^-)| \xrightarrow{a.s.} 0 \qquad \text{and} \qquad |F_n(\kappa_j^-) - F_X(\kappa_j^-)| \xrightarrow{a.s.} 0$$

as $n \longrightarrow \infty$, so looking at the maximum over all j,

$$\triangle_n = \max_{j=1,\ldots,k-1} \left\{ |F_n(\kappa_j) - F_X(\kappa_j)| \,, |F_n(\kappa_j^-) - F_X(\kappa_j^-)| \right\} \xrightarrow{a.s.} 0 \qquad \text{as } n \longrightarrow \infty.$$

For any x, find the interval within which x lies, that is, identify j such that

$$\kappa_{j-1} \leq x < \kappa_j.$$

Then we have

$$\begin{aligned} F_n(x) - F_X(x) &\leq F_n(\kappa_j^-) - F_X(\kappa_{j-1}) \leq F_n(\kappa_j^-) - F_X(\kappa_j^-) + \epsilon \\ F_n(x) - F_X(x) &\geq F_n(\kappa_{j-1}) - F_X(\kappa_j^-) \geq F_n(\kappa_{j-1}) - F_X(\kappa_{j-1}) - \epsilon \end{aligned}$$

and thus for any x,

$$F_n(\kappa_{j-1}) - F_X(\kappa_{j-1}) - \epsilon \leq F_n(x) - F_X(x) \leq F_n(\kappa_j^-) - F_X(\kappa_j^-) + \epsilon$$

and thus

$$|F_n(x) - F_X(x)| \leq \triangle_n + \epsilon \xrightarrow{a.s.} \epsilon \qquad \text{as } n \longrightarrow \infty.$$

Hence, as this holds for **arbitrary** x, it follows that

$$\sup_{x \in \mathbb{R}} |F_n(x) - F_X(x)| \xrightarrow{a.s.} \epsilon \qquad \text{as } n \longrightarrow \infty.$$

This holds for every $\epsilon > 0$; that is, if A_ϵ denotes the set of ω on which this convergence is observed, then $P(A_\epsilon) = 1$, and then by definition

$$A \equiv \bigcap_{\epsilon > 0} A_\epsilon \equiv \lim_{\epsilon \longrightarrow 0} A_\epsilon \qquad \Longrightarrow \qquad P(A) = P\left(\lim_{\epsilon \longrightarrow 0} A_\epsilon\right) = \lim_{\epsilon \longrightarrow 0} P(A_\epsilon) = 1$$

and it follows that

$$P\left[\lim_{n \longrightarrow \infty} \sup_{x \in \mathbb{R}} |F_n(x) - F_X(x)| = 0\right] = 1.$$