557: MATHEMATICAL STATISTICS II

THE GLIVENKO-CANTELLI LEMMA

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 $\widehat{F}_n(x)$, and defined for $x \in \mathbb{R}$ by

$$\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{[X_i, \infty)}(x)$$

where $\mathbb{1}_A(\omega)$ 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 defined by

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

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

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

as

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

This result is strengthened by the following Theorem.

Theorem. ***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 $\widehat{F}_n(x)$ denote the empirical distribution function. Then, as $n \longrightarrow \infty$,

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

or equivalently

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

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

Proof. Let $\epsilon > 0$. Then fix $k > 1/\epsilon$, and then consider points $t_0, \ldots, t_k$ such that

$$-\infty = t_0 < t_1 < t_2 < \ldots < t_{k-1} < t_k = \infty$$

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

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

where for $t \in \mathbb{R}$, $F_X(t^-) = \lim_{s \longrightarrow t^-} F_X(s) = P[X < t] = F_X(t) - P[X = t]$. Then, by construction

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

Recall that $\widehat{F}_n(x)$ is a **random** quantity for each x. Now, by the Strong Law, we have pointwise convergence, so that, as $n \longrightarrow \infty$, for $j = 1, \ldots, k-1$.

$$\widehat{F}_n(t_j) \xrightarrow{a.s.} F_X(t_j) \qquad \text{and} \qquad \widehat{F}_n(t_j^-) \xrightarrow{a.s.} F_X(t_j^-)$$

or equivalently for each j,

$$|\widehat{F}_n(t_j^-) - F_X(t_j^-)| \xrightarrow{a.s.} 0 \qquad \text{and} \qquad |\widehat{F}_n(t_j^-) - F_X(t_j^-)| \xrightarrow{a.s.} 0$$

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

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

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

$$t_{j-1} \leq x < t_j.$$

Then we have

$$\widehat{F}_n(t_{j-1}) \leq \widehat{F}_n(x) \leq \widehat{F}_n(t_j^-) \qquad \text{and} \qquad F_X(t_{j-1}) \leq F_X(x) \leq F_X(t_j^-)$$

so that, as from above $F_X(t_j^-) - F_X(t_{j-1}) < \epsilon$,

$$\widehat{F}_n(x) - F_X(x) \;\leq\; \widehat{F}_n(t_j^-) - F_X(t_{j-1}) \;\leq\; \widehat{F}_n(t_j^-) - F_X(t_j^-) + \epsilon$$

$$\widehat{F}_n(x) - F_X(x) \;\geq\; \widehat{F}_n(t_{j-1}) - F_X(t_j^-) \;\geq\; \widehat{F}_n(t_{j-1}) - F_X(t_{j-1}) - \epsilon$$

and thus for any x,

$$\widehat{F}_n(t_{j-1}) - F_X(t_{j-1}) - \epsilon \leq \widehat{F}_n(x) - F_X(x) \leq \widehat{F}_n(t_j^-) - F_X(t_j^-) + \epsilon$$

and thus

$$\begin{aligned} \left| \widehat{F}_n(x) - F_X(x) \right| &\leq \max \left\{ |\widehat{F}_n(t_{j-1}) - F_X(t_{j-1})|, |\widehat{F}_n(t_j^-) - F_X(t_j^-)| \right\} + \epsilon \\ &\leq \triangle_n + \epsilon \xrightarrow{a.s.} \epsilon \qquad \text{as } n \longrightarrow \infty. \end{aligned}$$

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

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

This holds for every $\epsilon > 0$: 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

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

and the convergence is uniform in x. ∎