Consistency of Empirical CDF

H.W Roh

We would like to show the consistency empirical CDF. Let $X_1, ..., X_n$ be i.i.d random variables with a distribution $F_X(x) = P(X \leq x)$. We define empirical CDF as

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

where I refers to indicator function, namely $I\{X_i \leq x\}$ is one if $X_i \leq x$ and zero otherwise. Note that this is simply the distribution function of a discrete random variable that places mass $1/n$ in the points $X_1, ..., X_n$. We would like to show the consistency property of this ECDF estimator.

Proposition. *For a fixed (but arbitrary) point $x \in \mathbb{R}$, we have that $n\hat{F}(x) = \sum_{i=1}^{n} I\{X_i \leq x\}$ has a binomial distribution with parameters n and success probability $p = F(x) = P(X \leq x)$. Therefore*

$$E[\hat{F}(x)] = p = F(x)$$
$$Var[\hat{F}(x)] = \frac{p(1-p)}{n}$$

This implies that $\hat{F}(x)$ converges in probability to $F(x)$ as $n \to \infty$, which means $\hat{F}(x)$ is a consistent estimator.

Proof. Let $Y_i = I\{X_i \leq x\}$ then Y_i are i.i.d with mean $F(x)$. Using weak law of large number,

$$\frac{1}{n}\sum_{i=1}^{n} Y_i \xrightarrow{p} E[Y_i] = F(x)$$

Equivalently, we can also show it by using Chebychev inequality. We can write it as

$$\begin{aligned}
P\left(\mid \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{X_{(i)} \leq x\} - P(X \leq x) \mid > \sqrt{\epsilon}\right) &= P\left(\mid \hat{F}(x) - F(x) \mid > \sqrt{\epsilon}\right) \\
&\leq \frac{E\left[\mid \hat{F}(x) - F(x) \mid^2\right]}{\epsilon} \\
&= \frac{E\left[\left(\hat{F}(x) - E(F_n)\right)^2\right]}{\epsilon} \\
&= \frac{Var\left[\hat{F}(x)\right]}{\epsilon} \\
&= \frac{p(1-p)}{n \cdot \epsilon} \\
&\leq \frac{1}{4n\epsilon}
\end{aligned}$$

■