Empirical Cumulative Distribution Function (eCDF)

Hanxiao Du

1 Introduction

In statistics, an empirical distribution function (commonly also called an empirical Cumulative Distribution Function, eCDF) is the distribution function associated with the empirical measure of a sample.

Definition 1 (Empirical distribution function)**.** *Let $(X_1, \cdots, X_n)$ be n independent, identically distributed real random variables with the common cumulative distribution function $F(t)$. Then, the empirical distribution function is defined as:*

$$\hat{F}_n(t) = \frac{\textit{number of elements in the sample} \leq t}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i \leq t) \tag{1}$$

where $\mathbb{I}(X_i \leq t)$ is the indicator of event $X_i \leq t$.

Theorem 1 ($\hat{F}_n(t)$ is an unbiased estimator of $F(t)$)**.**

Proof. Let $(X_1, \cdots, X_n)$ be n independent, identically distributed real random variables with the common cumulative distribution function $F(t)$. For a fixed t, since that $\mathbb{I}(X_i \leq t) = 1$ with probability $F(t)$ and $\mathbb{I}(X_i \leq t) = 0$ with probability $1 - F(t)$ by the definition of $F(t)$, the indicator $\mathbb{I}(X_i \leq t)$ is a Bernoulli random variable with parameter $p = F(t)$. Therefore, $\forall i \in [n]$,

$$\mathbb{E}[\mathbb{I}(X_i \leq t)] = \mathbb{P}[X_i \leq t] = F(t) \tag{2}$$

Furthermore,

$$\begin{aligned}
\mathbb{E}[\hat{F}_n(t)] &= \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i \leq t)] && (3)\\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\mathbb{I}(X_i \leq t)] && (4)\\
&= \frac{1}{n}\sum_{i=1}^{n} F(t) && (5)\\
&= F(t) && (6)
\end{aligned}$$

Thus, $\hat{F}_n(t)$ is an unbiased estimator of $F(t)$.

■