Glivenko–Cantelli theorem

Wendy Rosettini

November 2023

1 Introduction

Glivenko–Cantelli Theorem

In the theory of probability, the Glivenko–Cantelli theorem (sometimes referred to as the Fundamental Theorem of Statistics), named after Valery Ivanovich Glivenko and Francesco Paolo Cantelli, determines the asymptotic behavior of the empirical distribution function as the number of independent and identically distributed observations grows.

The uniform convergence of more general empirical measures becomes an important property of the Glivenko–Cantelli classes of functions or sets. The Glivenko–Cantelli classes arise in Vapnik–Chervonenkis theory, with applications to machine learning. Applications can be found in econometrics making use of M-estimators.

Statement

Assume that $X_1, X_2, \ldots, X_n$ are independent and identically distributed random variables in $\mathbb{R}$ with a common cumulative distribution function $F(x)$. The empirical distribution function for $X_1, \ldots, X_n$ is defined by

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

where I_C is the indicator function of the set C. For every (fixed) x, $F_n(x)$ is a sequence of random variables which converge to $F(x)$ almost surely by the strong law of large numbers. Glivenko and Cantelli strengthened this result by proving uniform convergence of F_n to F.

Theorem

$$\|F_n - F\|_\infty = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \longrightarrow 0$$

almost surely.

This theorem originates with Valery Glivenko and Francesco Cantelli, in 1933.

Remarks

If X_n is a stationary ergodic process, then $F_n(x)$ converges almost surely to $F(x) = \mathbb{E}[1_{X_1 \leq x}]$. The Glivenko–Cantelli theorem gives a stronger mode of convergence than this in the iid case. An even stronger uniform convergence result for the empirical distribution function is available in the form of an extended type of law of the iterated logarithm.

Proof

For simplicity, consider a case of a continuous random variable X. Fix $-\infty = x_0 < x_1 < \ldots < x_{m-1} < x_m = \infty$ such that $F(x_j) - F(x_{j-1}) = \frac{1}{m}$ for $j = 1, \ldots, m$. Now for all $x \in \mathbb{R}$, there exists $j \in \{1, \ldots, m\}$ such that $x \in [x_{j-1}, x_j]$. Note that

$$F_n(x) - F(x) \leq F_n(x_j) - F(x_{j-1}) = F_n(x_j) - F(x_j) + \frac{1}{m},$$

$$\mathrm{F}_n(x) - F(x) \geq F_n(x_{j-1}) - F(x_j) = F_n(x_{j-1}) - F(x_{j-1}) - \tfrac{1}{m}.$$

Therefore,

$$\|F_n - F\|_\infty = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \leq \max_{j \in \{1,\ldots,m\}} |F_n(x_j) - F(x_j)| + \frac{1}{m}.$$

Since $\max_{j \in \{1,\ldots,m\}} |F_n(x_j) - F(x_j)| \to 0$ a.s. by the strong law of large numbers, we can guarantee that for any positive ϵ and any integer m such that $\frac{1}{m} < \epsilon$, we can find N such that for all $n \geq N$, we have $\max_{j \in \{1,\ldots,m\}} |F_n(x_j) - F(x_j)| \leq \epsilon - \frac{1}{m}$ a.s. Combined with the above result, this further implies that $\|F_n - F\|_\infty \leq \epsilon$ a.s., which is the definition of almost sure convergence. Source(*https* : *//en.wikipedia.org/wiki/Glivenko*

2 Simulations

The chart illustrates a simulation of the Glivenko-Cantelli theorem applied to a continuous uniform distribution over the interval $[-1, 1]$. This theorem states that the empirical cumulative distribution function (ECDF) of a dataset, derived from the data itself, converges uniformly to the true cumulative distribution function (CDF) of the underlying distribution as the sample size increases.

In the specific context of this chart:

- **Blue Line (True Distribution):** Represents the actual CDF of the continuous uniform distribution over the interval $[-1, 1]$. This line provides a reference for the theoretical shape of the cumulative distribution.
- **Red Lines (Empirical Distributions):** Represent the ECDFs obtained from different simulations. Each simulation involves sampling independent

random variables from the uniform distribution over $[-1, 1]$ and constructing the ECDF based on these samples. The overlap of the red lines indicates how the ECDFs converge toward the true CDF as the number of simulations increases.

In essence, the chart offers an intuitive visualization of how the ECDF estimated from various datasets gradually approaches the true CDF, thus supporting the Glivenko-Cantelli theorem for the specified uniform distribution interval.

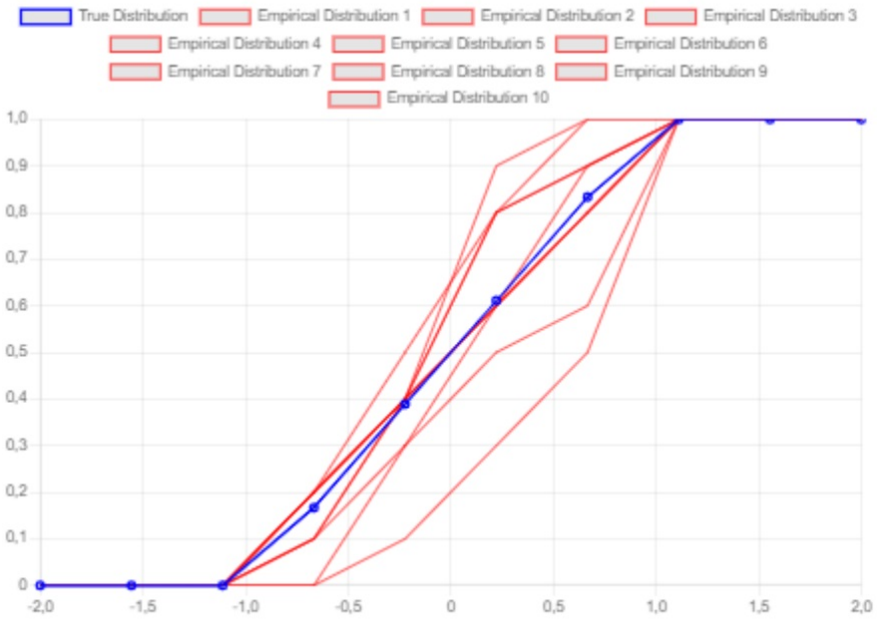

Figure 1: Glivenko–Cantelli theorem

You can find my code on my webpage[1]

[1] http://wendy.altervista.org/cantelli.html