Th2: The Glivenko-Cantelli Theorem

Federico Detomaso

1 Meaning

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 behaviour 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.

2 Definition

Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with cumulative distribution function $F(x)$. The empirical distribution function, denoted by $F_n(x)$, is defined as:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq x\}}$$

where $\mathbf{1}_{\{A\}}$ is the indicator function of event A, which is 1 if A is true and 0 otherwise.

2.1 Theorem

The Glivenko-Cantelli Theorem states that, for any $\epsilon > 0$:

$$\lim_{n\to\infty} P\left(\sup_x |F_n(x) - F(x)| > \epsilon\right) = 0$$

In other words, the empirical distribution function $F_n(x)$ converges uniformly to the true distribution function $F(x)$ with probability 1.

3 Proof

The proof is based on the application of Hoeffding's inequality and the idea of empirical processes.

Define the empirical process $G_n(x) = \sqrt{n}(\sup_y |F_n(y) - F(y)|)$. Hoeffding's inequality yields:

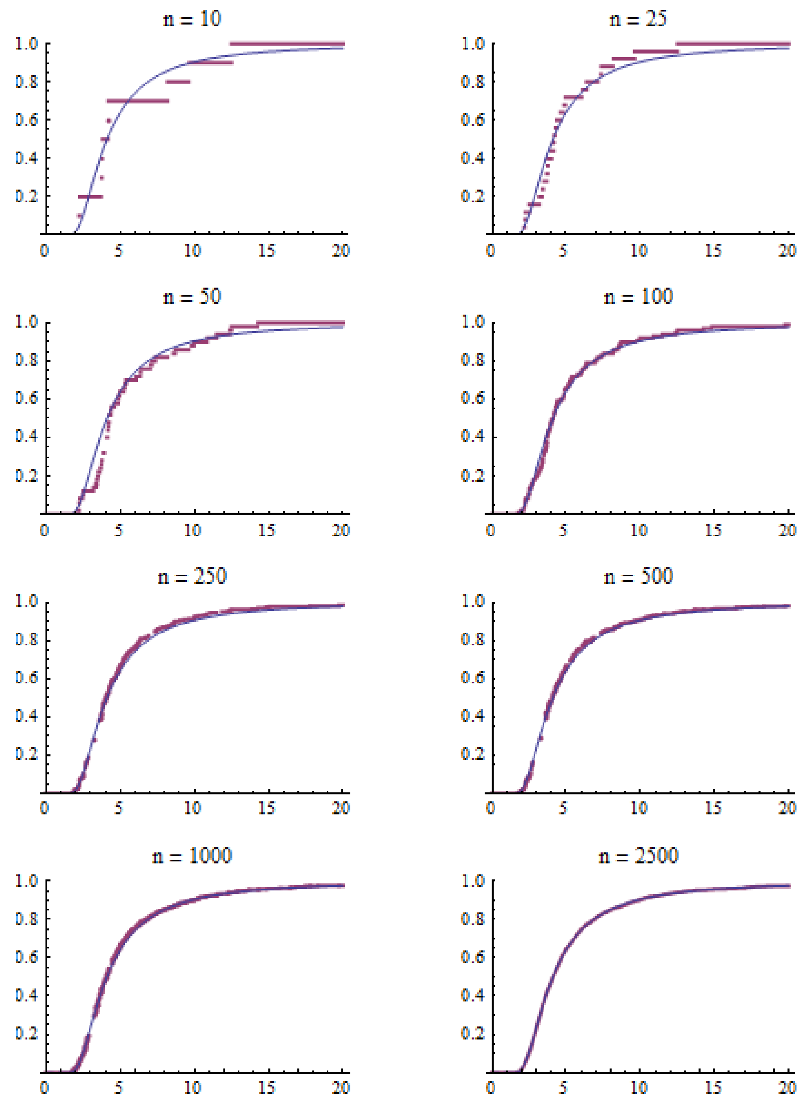
$$P(G_n(x) > t) \leq 2e^{-2nt^2}$$

Now, we want to bound the probability of the event $\{\sup_x |F_n(x) - F(x)| > \epsilon\}$. Using the union bound, we have:

$$P\left(\sup_x |F_n(x) - F(x)| > \epsilon\right) \leq P\left(G_n(x) > \frac{\epsilon}{\sqrt{n}}\right) \leq 2e^{-2n\left(\frac{\epsilon}{\sqrt{n}}\right)^2} = 2e^{-2\epsilon^2}$$

This probability converges to 0 as $n \to \infty$, and thus, the Glivenko-Cantelli Theorem is proved.

4 Simulations



Demonstrate that the empirical distribution tends to the theoretical distribution as sample size increases by comparing empirical and theoretical CDFs.

References

[1] https://en.wikipedia.org/wiki/Glivenko%E2%80%93Cantelli_theorem

[2] https://www.wolfram.com/mathematica/new-in-8/probability-and-statistics-solvers-and-properties/demonstrate-the-glivenko-cantelli-theorem.html