Glivenko - Cantelli Theorem

Giacomo Babudri

December 2023

1 Introduction

The Glivenko-Cantelli theorem is a fundamental result in mathematical statistics that establishes the convergence properties of empirical distribution functions (EDFs) to their true distribution functions. In essence, the theorem provides a mathematical foundation for making inferences about an underlying population distribution based on observed sample data.

1.1 Theorem Statement

Formally, let $F_n(x)$ be the empirical distribution function based on a sample of size n, and let $F(x)$ be the true (but usually unknown) distribution function. The Glivenko-Cantelli theorem states that, almost surely, the empirical distribution function converges uniformly to the true distribution function as the sample size increases:

$$\sup_x |F_n(x) - F(x)| \longrightarrow 0 as n \longrightarrow$$

Here:

- $\sup_x$ denotes the supremum (the least upper bound) taken over all possible values of x,
- $|F_n(x) - F(x)|$ represents the pointwise difference between the empirical and true distribution functions,
- The notation $\longrightarrow 0 as n \longrightarrow$ indicates almost sure convergence, meaning that the convergence holds with probability 1.

2 Proof

Premise: For simplicity, consider 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 $j = 1, \ldots, m$. Each interval $(x_{j-1}, x_j]$ has length $\frac{1}{m}$.

Proof: Note that for every $x \in R$, there exists $j \in \{1, \ldots, m\}$ such that $x \in (x_{j-1}, x_j]$. Notice that:

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

Hence:

$$F_n(x) - F(x) \leq \frac{1}{m}.$$

Also notice that:

$$F_n(x_{j-1}) - F(x_{j-1}) \geq F_n(x) - F(x) \geq F_n(x_{j-1}) - F(x_{j-1}) - \frac{1}{m}.$$

Therefore:

$$F_n(x) - F(x) \geq -\frac{1}{m}.$$

Consequently:

$$|F_n(x) - F(x)| \leq \frac{1}{m}.$$

Thus, we can write:

$$\|F_n - F\|_\infty = \sup_{x \in R} |F_n(x) - F(x)| \leq \frac{1}{m}.$$

Since m is arbitrary, we can let it tend to infinity, and so $\|F_n - F\|_\infty \to 0$ as $m \to \infty$.

Conclusion: This proof demonstrates that the empirical distribution function $F_n(x)$ converges uniformly to the true distribution function $F(x)$ almost surely.

3 Applications

Economy:

In finance, particularly risk management, the Glivenko-Cantelli theorem is applied to model and estimate the distribution of financial returns. It provides a theoretical foundation for using historical data to simulate the distribution of future returns.

Machine Learning:

In machine learning, the Glivenko-Cantelli theorem is relevant when assessing the performance of models. For instance, in model validation, it assures practitioners that the discrepancy between predicted and observed outcomes converges uniformly to zero as the sample size increases.

Environment:

The theorem finds applications in environmental studies, such as analyzing the distribution of pollutant concentrations over time. It supports the use of empirical distribution functions to model and understand the probability distributions of environmental variables.

Telecomunication:

In the analysis of telecommunication networks, the Glivenko-Cantelli theorem can be applied to understand the distribution of communication link delays. This information is essential for optimizing network performance and predicting communication delays in various scenarios.

4 Simulation

In this simulation the purpose is to demonstrate, in a practical manner, the convergence of the EDF to the true CDF as the sample size increases. The Glivenko-Cantelli theorem ensures that, under certain conditions, the EDF provides a consistent estimate of the true distribution as more observations are included in the sample.

For this reason is useful bringing two different cases, one with a bigger sample size and the other one with a lower size.

First case: $N = 8000$. It's easy to notice that the representation of the function calculated is very close to the one empirical function, giving us a pratical proof of the theorem.

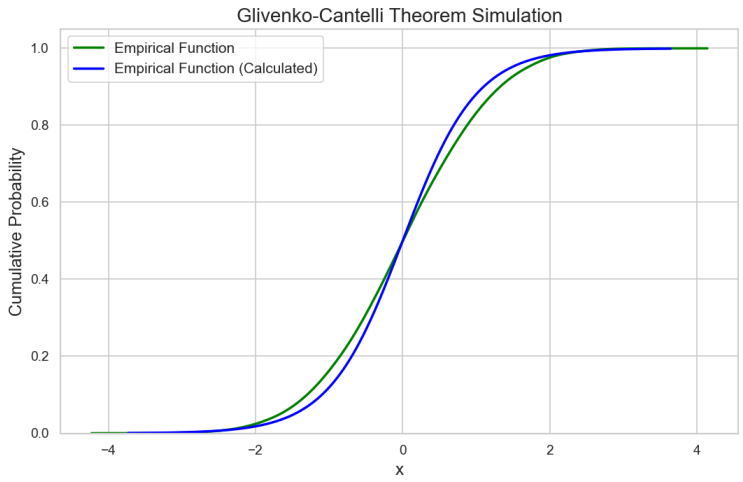

Figure 1: First case with N = 8000

Second case: N = 10. The representation it's unpredictable and far from the empirical function.

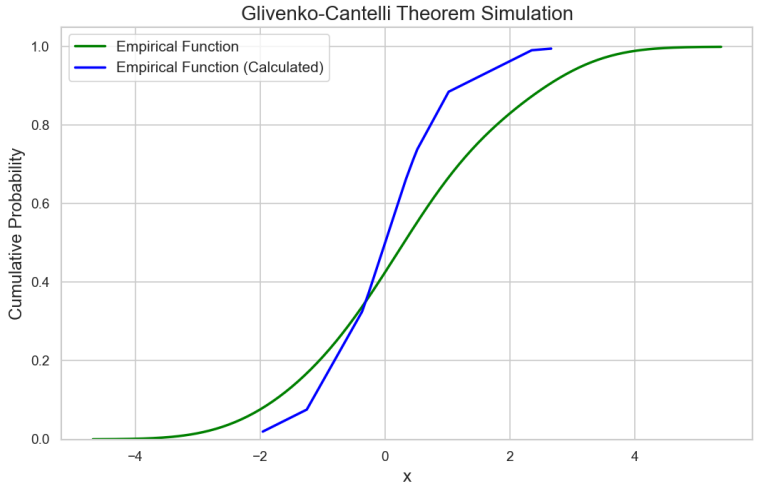

Figure 2: Second case with N = 10

The execution is built through a python code shown below:

```python
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

# Set style for a more pleasing appearance
sns.set(style='whitegrid')
plt.figure(figsize=(10, 6))

# Number of samples to generate
sample_size = 8000

# Generate a sequence of random variables from a different distribution (e.g., normal)
random_samples = np.random.normal(size=sample_size)

# Sort the samples
sorted_samples = np.sort(random_samples)

# Calculate the empirical distribution function (EDF)
edf = np.arange(1, sample_size + 1) / sample_size

# Define a different expected cumulative distribution function (CDF)
def true_cdf(x):
    return 0.5 * (1 + np.tanh(x))

# Visualize the convergence of EDF to CDF with Seaborn
sns.kdeplot(random_samples, cumulative=True, label='Empirical Function', color='green', linewidth=2)
sns.lineplot(x=sorted_samples, y=true_cdf(sorted_samples), label='Empirical Function (Calculated)', color='blue', linewidth=2)

# Aesthetics customization
plt.title('Glivenko-Cantelli Theorem Simulation', fontsize=16)
plt.xlabel('x', fontsize=14)
plt.ylabel('Cumulative Probability', fontsize=14)
plt.legend(fontsize=12)
plt.show()
```

Figure 3: Simulation's code