Glivenko-Cantelli Theorem

The Glivenko-Cantelli theorem is a fundamental result of the theory of probability and statistics.

It states that, given a sequence of independent and identically distributed random variables, their **empirical distribution function** will converge uniformly to the **theoretical cumulative distribution** function as the number of observations tends to infinity.

An **empirical distribution** is a representation of the probability distribution of a set of observed data and attributes an empirical probability to each observed value. This empirical probability is the relative frequency of a value compared to the total observations.

The Glivenko-Cantelli theorem states that the difference between the empirical distribution (based on observed data) and the true probability distribution becomes smaller and smaller as the number of observations increases, following a uniform convergence.

This result is of great practical importance because it provides a way to establish how well an empirical distribution approximates the true theoretical distribution of the data.

Glivenko-Cantelli Theorem Example

To better understand this theorem, let's imagine we have a fair six-sided die. Suppose we roll this die an increasing number of times and record the results.

The theoretical probability distribution for rolling a fair die is uniform, with a probability of 1/6 for each side.

If we rolled the die 10 times, we could get for example:
3, 6, 1, 4, 2, 6, 5, 3, 1, 2.

The empirical distribution in this case would be:

- Face 1: empirical frequency = 2/10 = 0.2
- Face 2: empirical frequency = 2/10 = 0.2
- Face 3: empirical frequency = 2/10 = 0.2
- Face 4: empirical frequency = 1/10 = 0.1
- Face 5: empirical frequency = 1/10 = 0.1
- Face 6: empirical frequency = 2/10 = 0.2

With the Glivenko-Cantelli theorem, as the number of dice rolls increases, the empirical distribution of relative frequencies for each face will converge uniformly to the uniform theoretical probability distribution (1/6 for each face).

So, as you increase the number of launches, the empirical distribution will get closer and closer to the expected theoretical distribution.

Glivenko-Cantelli Theorem Formula

Suppose we have a sequence of independent and identically distributed random variables $X_1, X_2, \ldots, X_n$ with a theoretical distribution function $F(x)$.

The empirical distribution $F_n(x)$ is the distribution function constructed from the observed data.

The theorem states that, given an increasing number of observations n, the difference between the empirical distribution $F_n(x)$ and the theoretical probability distribution $F(x)$ will tend to zero when considered in terms of uniform convergence.

Formally, this is expressed as:

$$sup_{x \in R} |\ F_n(x) \ - F(x)\ | \to^P 0$$

Where sup denotes the supreme (minimum upper bound) of the true interval, $\to^P$ denotes convergence in probability and $|\ F_n(x) \ - F(x)\ |$ represents the difference between the empirical distribution and the theoretical distribution.

Summing up, the Glivenko-Cantelli theorem states that the difference between the empirical distribution of the data and the true theoretical distribution will reduce to zero as the number of observations increases, following uniform convergence. This is critical in statistical analysis because it provides a way to evaluate how well the empirical data approximates the theoretical distribution.

References:

The Glivenko-Cantelli theorem - *Wikipedia*
https://en.wikipedia.org/wiki/Glivenko%E2%80%93Cantelli_theorem