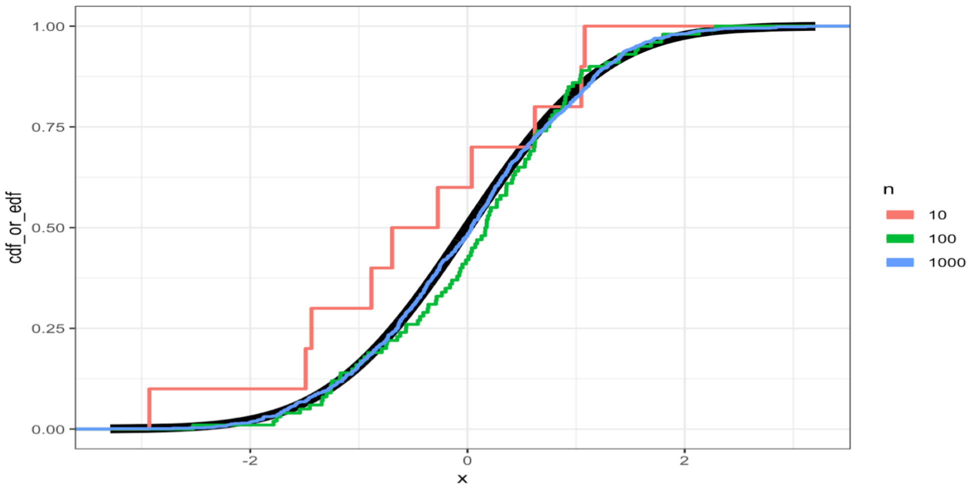
GLIVENKO-CANTELLI THEOREM



- What is the meaning of the statement: "A random sample describes the population"?

- Can we explain it in probability theory terms?
- Answer is: "YES".

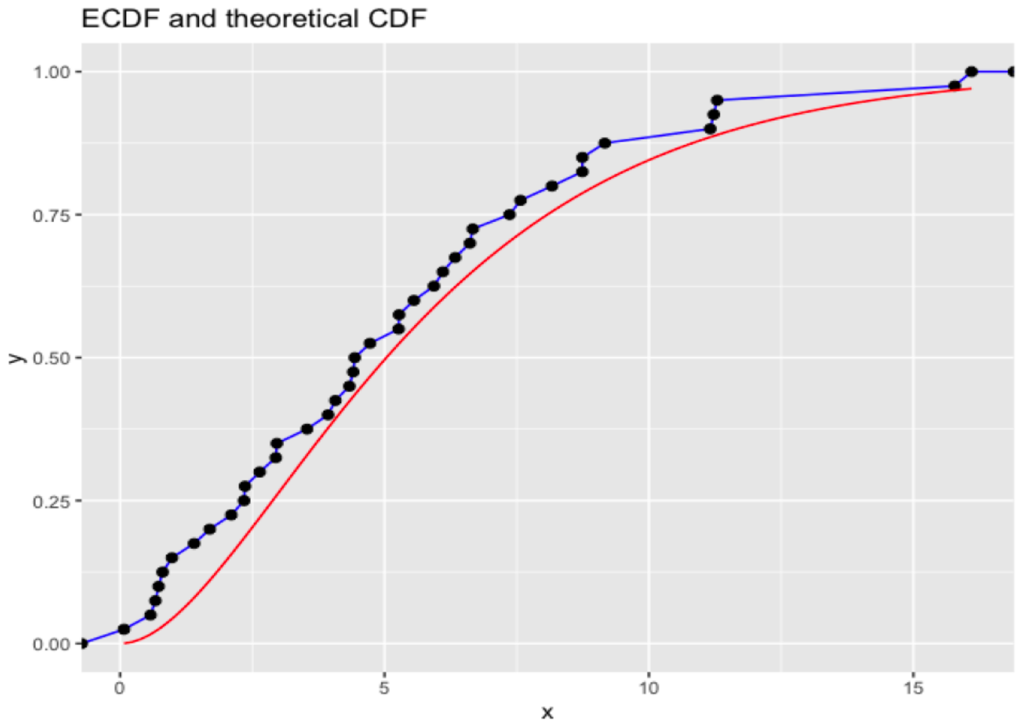
- And this is because of the famous Glivenko-Cantelli Theorem Sometimes

it is called Fundamental Theorem of Statistics.

Definition :

Let $\{X_1, X_2, \ldots, X_n\}$ be an i.i.d. sequence of random variables with distribution function $F(a)$ on $\mathbb{R}$. Then the empirical distribution function, denoted by $F_n : \mathbb{R} \to [0,1]$, is defined as:

$$F_n(a) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq a\}} = \frac{\text{number of } X_1, X_2, \ldots, X_n \text{ that are } \leq a}{n}.$$

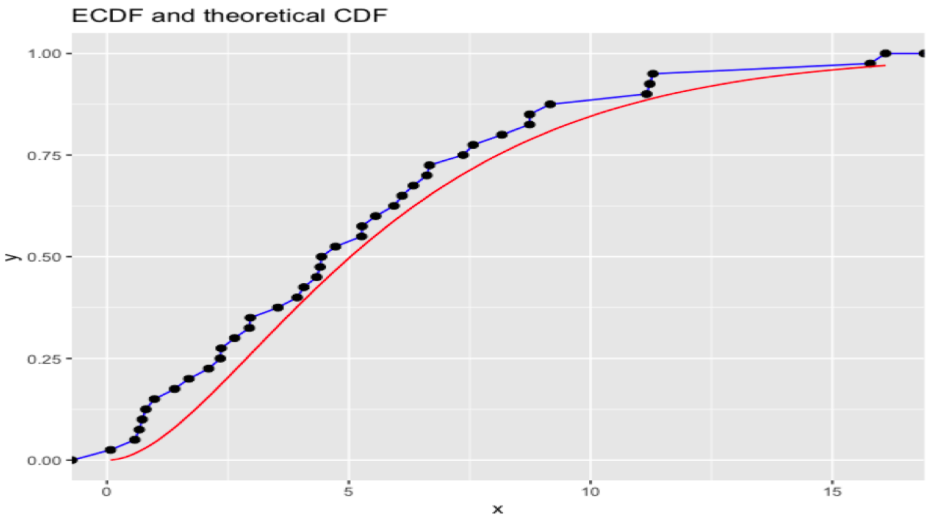


Before we move to the theorem lets look at if we fixed one particular $a_0 \in \mathbb{R}$. We want to know how far the blue curve from the red curve. What happens to $F_n(a_0)$?

By strong law of large number we can say that, $F_n(a_0) \to F(a_0)$ $a.s.$ for any fixed $a_0 \in \mathbb{R}$. Because $\mathbb{1}_{\{X_i \leq a\}}$ is a binomial random variable with probability $P = P(X_i \leq a_0)$ so if a_0 fixed then P is fixed. This is by the law of large number converges to $F(a_0)$.

- Recall Uniform convergence

- Two functions converge uniformly if the distance over the whole space uniformly get smaller, so the maximal distance between these two curves get close to zero.



Glivenko-Cantelli Theorem

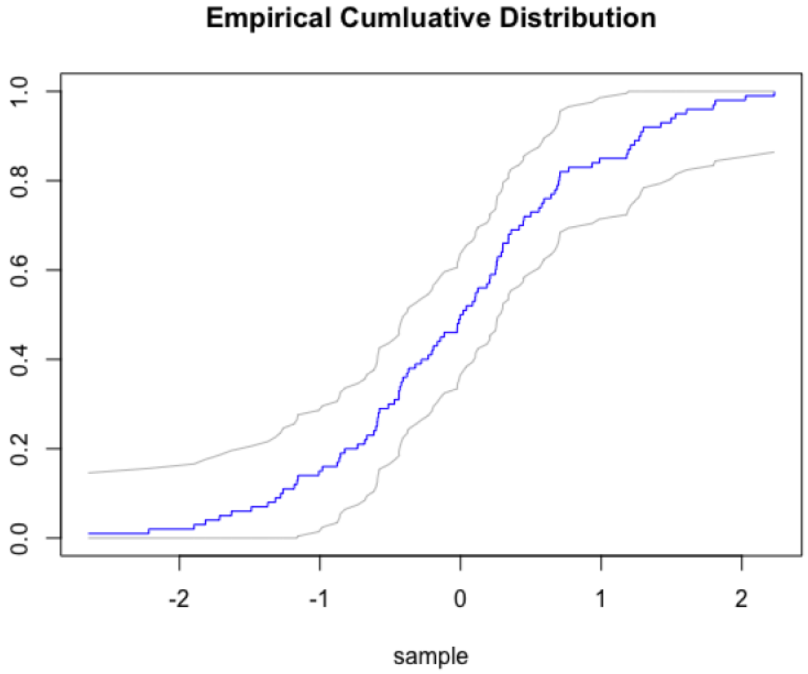
Let $X_1, \ldots, X_n$ be iid random variables with cdf $F(a)$. Let $F_n(a)$ be the empirical CDF induced by the sample

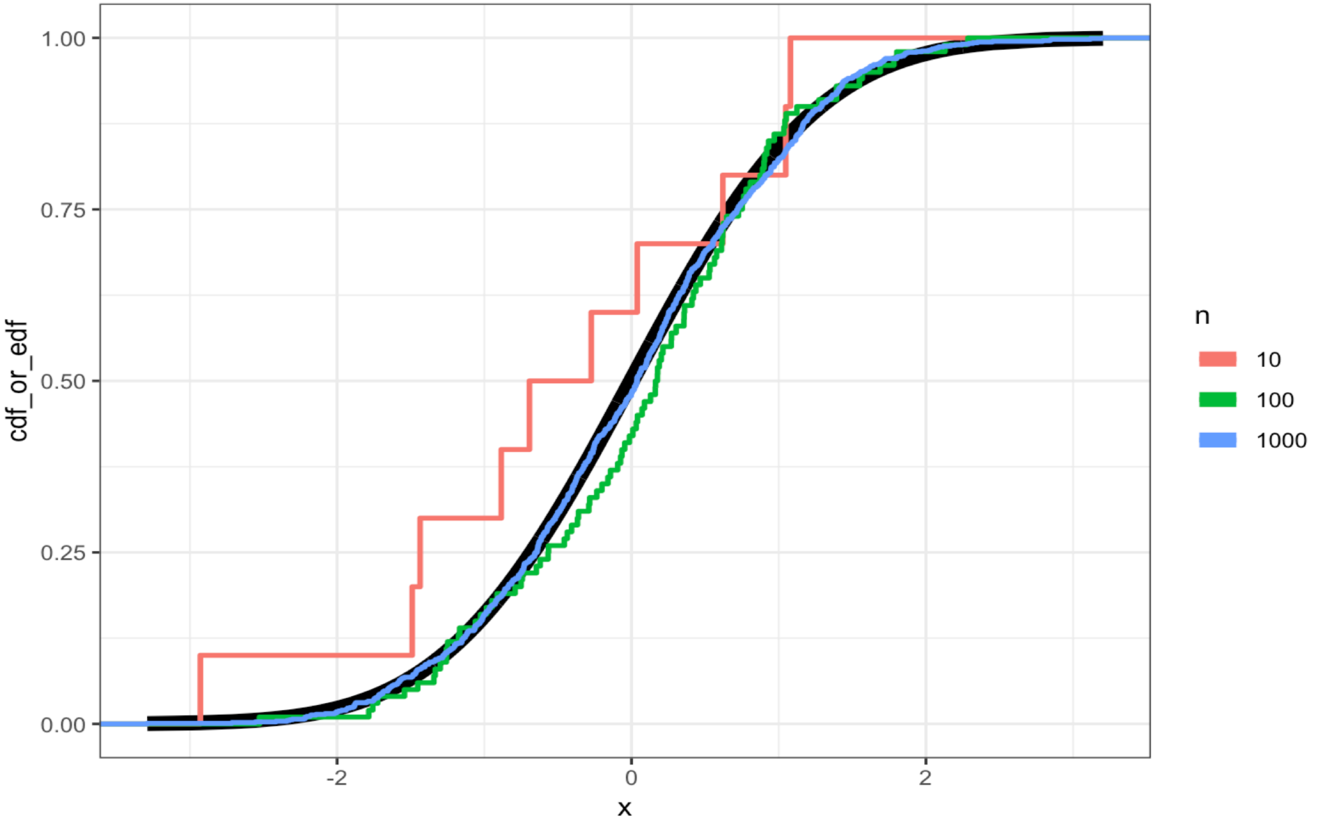
$$F_n(a) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq a\}}$$

Then

$$P\left(\sup_{a \in \mathbb{R}} |F_n(a) - F(a)| > \epsilon\right) \leq 8(n+1) \exp\left(-\frac{n\epsilon^2}{32}\right)$$

In particular, $\sup_{a \in \mathbb{R}} |F_n(a) - F(a)| \to 0$ $a.s.$ as $n \to \infty$.



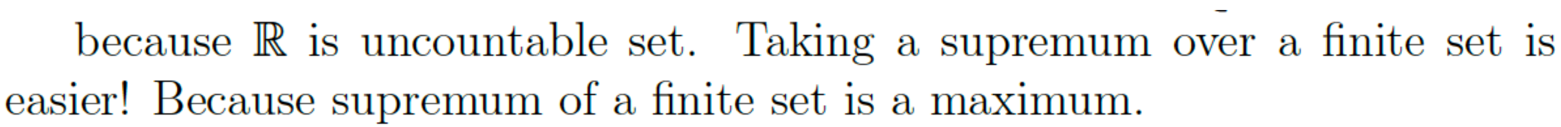


As the sample size grows the empirical cdf reach the cdf of the real distribution.

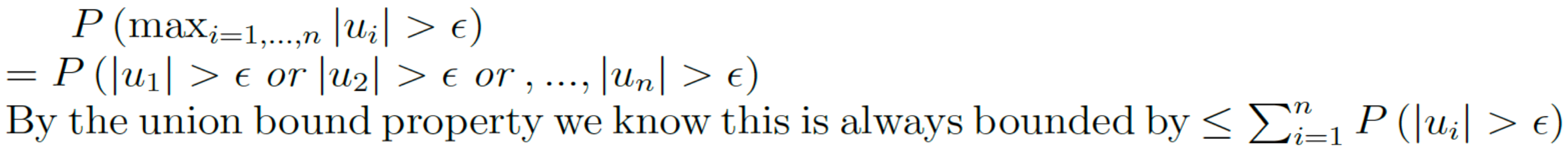
Few Remarks:

Observe, by the law of large number $P(|F_n(a) - F(a)| > \epsilon) \to 0$ for any fixed a_0 but our problem we need to look at the expression $P\left(\sup_{a\in\mathbb{R}} |F_n(a) - F(a)| > \epsilon\right)$

- Why it is difficult!



because $\mathbb{R}$ is uncountable set. Taking a supremum over a finite set is easier! Because supremum of a finite set is a maximum.



$P\left(\max_{i=1,\ldots,n}|u_i| > \epsilon\right)$

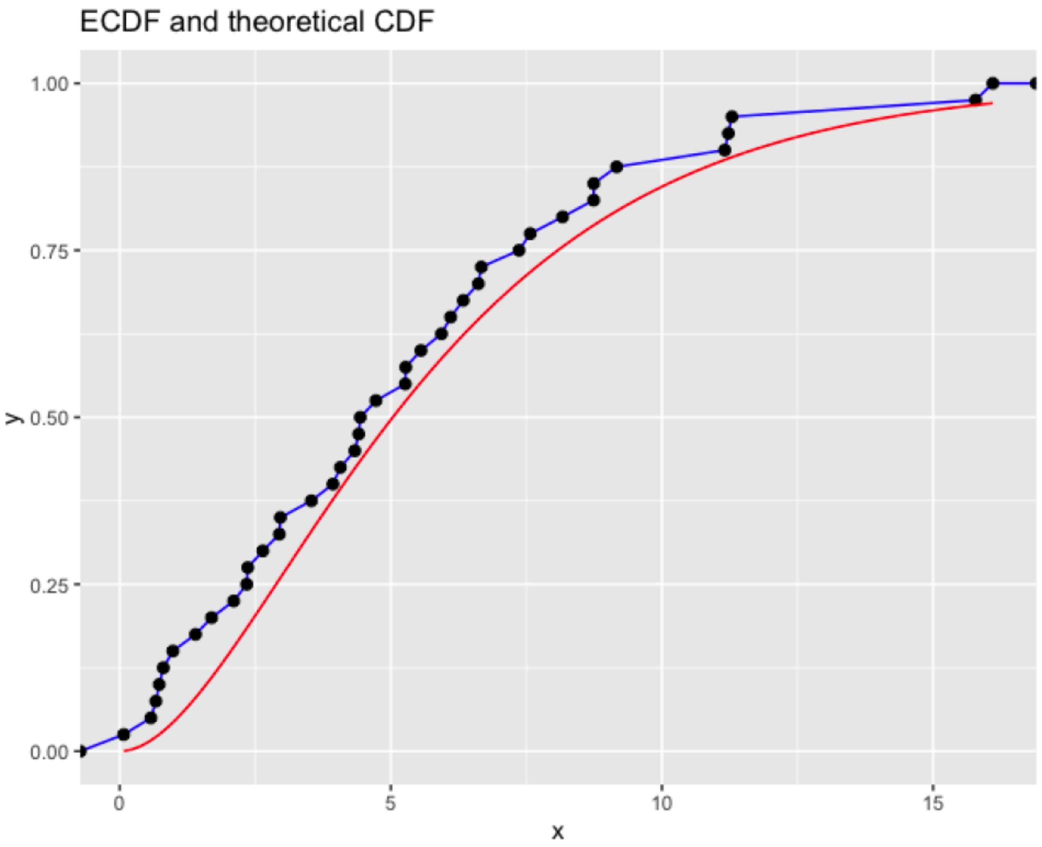
$= P\left(|u_1| > \epsilon\ or\ |u_2| > \epsilon\ or\ , \ldots, |u_n| > \epsilon\right)$

By the union bound property we know this is always bounded by $\leq \sum_{i=1}^{n} P\left(|u_i| > \epsilon\right)$

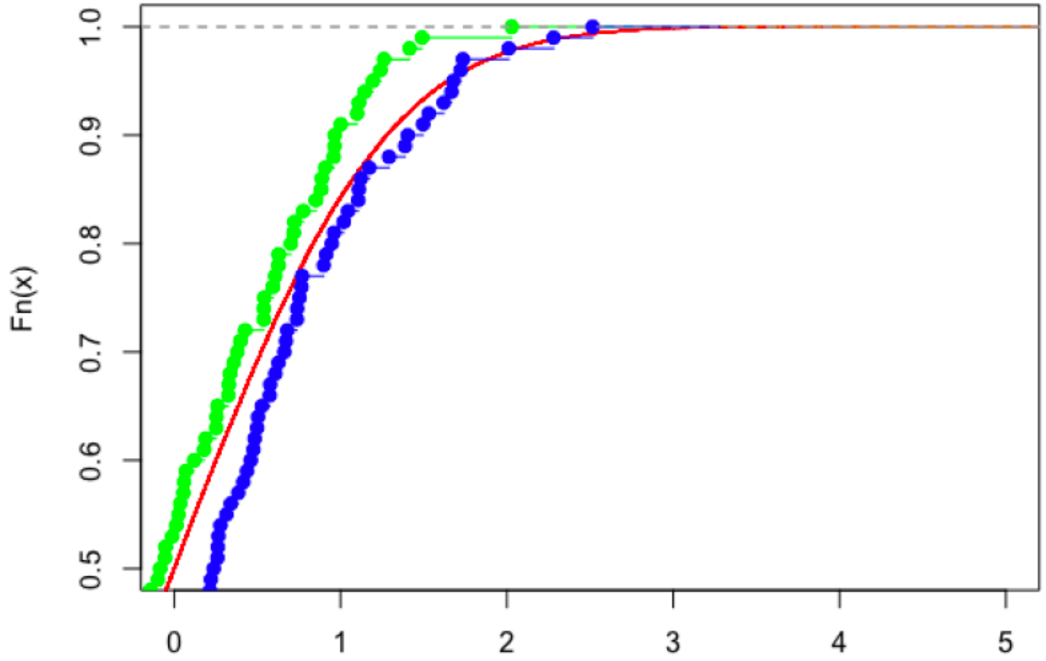
- The trick of the proof:

- How could we achieve that ?

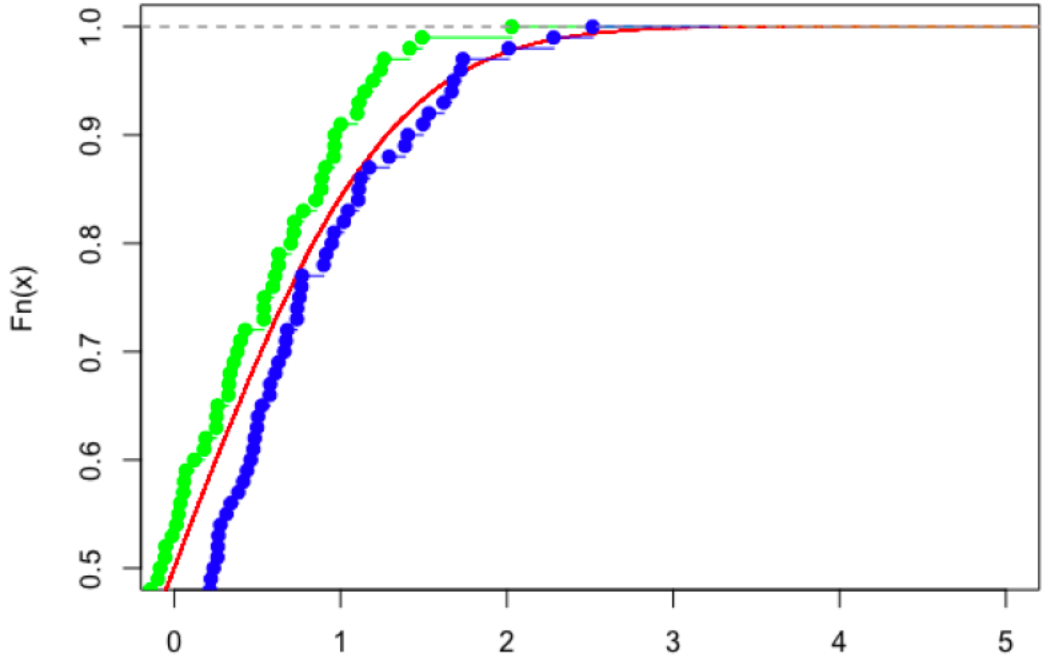
we are interested in the difference between the blue curve and red curve.
If we want to take the sup over the difference between the two curves, F takes infinitely many variables.



What would happen if we introduce another empirical distribution function for another sample. The green function called F'_n induced by a **ghost sample.**

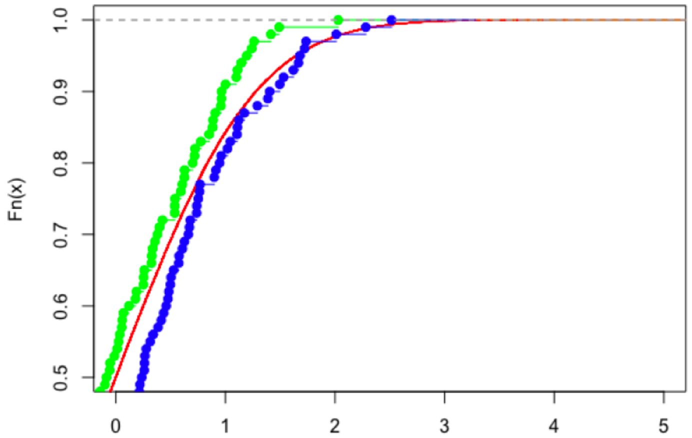


Assume we have a second sample $X'_1, ..., X'_n$, it has the same number of elements as first sample , completely independent from sample 1 and generated from the same distribution.

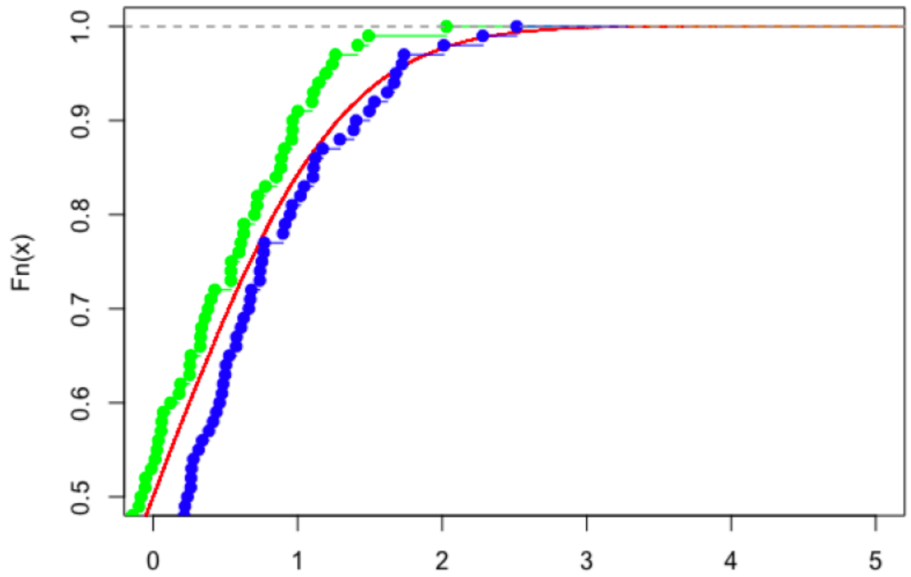


Ghost sample: is a sample we had in our mind but it does not really exist. Our goal: is to prove that the blue curve converge to the red and what holds for the blue will hold for the green because they both generated from the same distribution.

If the blue curve and red curve are close then the green curve and red curve will be also closed. Then we can bound the distance between the red and blue by twice the distance between green and blue.



- Now, If we want to look at the distances between the green and the blue curves there are only finitely many distances between them and this is what we want to achieve by introducing a ghost sample.



- The proof consists of several key steps: Assume $n\varepsilon^2 > 2$

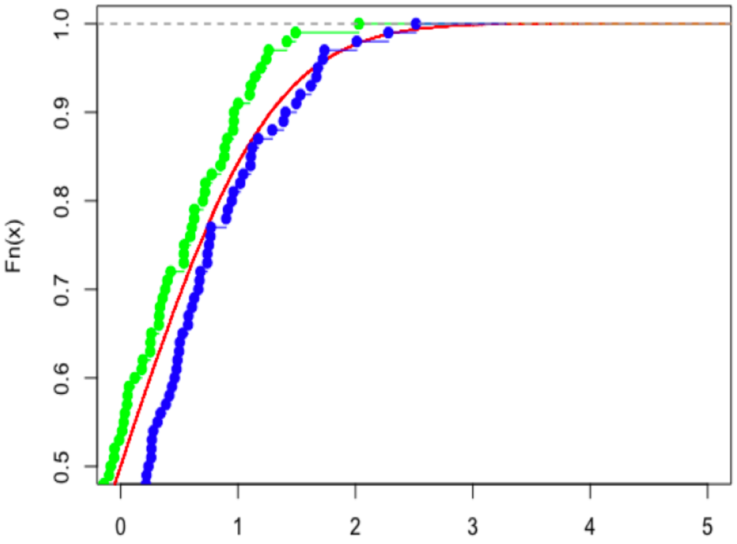
- Step1: symetrization by ghost sample. Assume $X_1', ..., X_n' \sim F$ Dentoe by F_n' the empirical CDF induced by ghost sample.

  Now,it is easy to prove:

$$P\left(\sup_{a\in\mathbb{R}} |F_n(a) - F(a)| > \epsilon\right)$$

$$\leq 2P\left(\sup_{a\in\mathbb{R}} |F_n(a) - F_n'(a)| > \epsilon/2\right)$$

  This step called symetrization by ghost sample, because what we have now is a term that does not depend on the true CDF instead it's now depends on two empirical CDF.



- Step 2 : split this last term into two terms,

$$|F_n(a) - F_n'(a)| = |\frac{1}{n}\sum_{i=1}^{n}(\mathbb{1}_{\{X_i \le a\}} - \mathbb{1}_{\{X_i' \le a\}})| \qquad (*)$$

Let $\sigma_1, ..., \sigma_n$ be iid random variables,independent of $X_1, ..., X_n, X_1', ..., X_n'$ with $P(\sigma_i = 1) = P(\sigma_i = -1) = 1/2$. Such random variables are called Rademacher random variables.

Observe that the distribution of * is the same as the distribution of the following,

$$|\frac{1}{n}\sum_{i=1}^{n}\sigma_i(\mathbb{1}_{\{X_i \le a\}} - \mathbb{1}_{\{X_i' \le a\}})| \qquad (**)$$

By the definition of $X_1, ..., X_n, X_1', ..., X_n'$ and $\sigma_1, ..., \sigma_n$.

Now we have from symmetrization lemma

$$\leq 2P\left(\sup_{a\in\mathbb{R}} |F_n(a) - F'_n(a)| > \epsilon/2\right)$$

this is the same as,

$$= 2P\left(\sup_{a\in\mathbb{R}} \left|\frac{1}{n}\sum_{i=1}^{n} \sigma_i(\mathbb{1}_{\{X_i\leq a\}} - \mathbb{1}_{\{X'_i\leq a\}})\right| > \epsilon/2\right)$$

$$\leq 2P(\sup_{a\in\mathbb{R}} |\frac{1}{n}\sum_{i=1}^{n} \sigma_i(\mathbb{1}_{\{X_i\leq a\}})| > \epsilon/4) + 2P(\sup_{a\in\mathbb{R}} |\frac{1}{n}\sum_{i=1}^{n} \sigma_i(\mathbb{1}_{\{X'_i\leq a\}})| > \epsilon/4)$$

Now what we have achieved is we divided this complicated probability into two probabilities and these two terms are exactly the same except we have in the first term X_i and the second X_i'. Since both X_i, X_i'

come from the same distribution, so instead of having this two terms we can write it as 4 times the first probability

$$= 4P(\sup_{a\in\mathbb{R}} |\frac{1}{n}\sum_{i=1}^{n} \sigma_i(\mathbb{1}_{\{X_i\leq a\}}| > \epsilon/4)$$

observe, we have an expression here that contains only empirical distribution function with finitely many steps.

- Step 3: exploit finite structure. For this step we fix $X_1, ..., X_n$, and fixing here means conditioning on $X_1, ..., X_n$. Conditioning intuitively means we fix into a particular value, so we want to look at the probability that something happens conditioning on the fact that $X's$ take certain values.

- note that the vector the random variables $\mathbb{1}_{\{X_1 \leq a\}}, ..., \mathbb{1}_{\{X_n \leq a\}}$ for fixed a can have at most $n + 1$.
  Thus conditioned on $X_1, ..., X_n$, the supremum is just a maximum over at most $n + 1$ random variables.

To compute the following probability we have,

$$P\left(\sup_{a\in\mathbb{R}} \frac{1}{n}\left|\sum_{i=1}^{n} \sigma_i \mathbb{1}_{\{X_1\le a\}},\right| > \frac{\epsilon}{4} \mid X_1,\ldots,X_n\right)$$

By applying union bound we obtain

$$\le (n+1)\sup_{a\in\mathbb{R}} P\left(\frac{1}{n}\left|\sum_{i=1}^{n} \sigma_i \mathbb{1}_{\{X_i\le a\}}\right| > \frac{\epsilon}{4} \mid X_1,\ldots,X_n\right)$$

where the sup is outside of the probability. As we notice the only random variable here is σ_i. The next step is to find an exponential bound for the RHS.

- Step 4: Hoeffding's inequality for Rademacher variables:
  With $X_1, \ldots, X_n$ fixed, we left with Rademacher random variables,$\sum_{i=1}^{n} \sigma_i \mathbb{1}_{\{X_i \leq a\}}$ is a sum of n independent zero mean random variables between [-1,1].

  Thus, by Hoeffding's inequality lemma for Rademacher variables we have,

$$P\left(\frac{1}{n}\left|\sum_{i=1}^{n} \sigma_i \mathbb{1}_{\{X_i \leq a\}}\right| > \frac{\epsilon}{4} \mid X_1, \ldots, X_n\right)$$
$$\leq 2 \exp\left(-\frac{n\epsilon^2}{32}\right).$$

Now we combine things together from all steps to achieved GC theorem.

Glivenko-Cantelli proof*

Since $F_n(a)$ and $F_n'(a)$ are piecewise constant functions, thus $\mid F_n(a) - F_n'(a) \mid$ has at most $(2n+1)$ different values when $-\infty < a < \infty$.
Step 2: Turn infinite many "Sup" to finite many "Max", corresponding to values.

$$2P\left(\sup_{-\infty<a<\infty} \mid F_n(a)\ - F_n'(a) \mid > \frac{\epsilon}{2}\right) = 2P\left(\max_{a=a_1,\ldots,a_{2n+1}} |F_n(a) - F_n'(a)| > \frac{\epsilon}{2}\right).$$

$$= 2P\left(\bigcup_{i=1}^{2n+1} |F_n\,(a_i) - F_n'\,(a_i)| > \frac{\epsilon}{2}\right)$$

$$\leq 2\sum_{i=1}^{2n+1} P\left(|F_n\,(a_i) - F_n'\,(a_i)| > \frac{\epsilon}{2}\right) \textit{ (By union bound)}$$

Step 3: Hoeffding's Inequality Suppose $Y_1^\star, \ldots, Y_n^\star$ are independent with $EY_i^\star = 0($ Mean 0) and $c_i \leq Y_1^\star \leq d_i$(bounded) then,

$$\forall \eta > 0, \quad P\left(|Y_1^\star + Y_2^\star + \ldots + Y_n^\star| > \eta\right) \leq 2e^{\frac{-2\eta^2}{\Sigma_{i=1}^n (d_i - c_i)^2}}$$

Let

$$Y_i^\star = \frac{1}{n}\left(I_{[X_i \leq a]} - I_{[X_i' \leq a]}\right)$$

then we have

$$-\frac{1}{n} \leq Y_i^\star \leq \frac{1}{n}$$

and $E\left(Y_i^\star\right) = 0$. Thus Hoeffding's Inequality can be applied to $\mid F_n\left(a_i\right) - F_n'\left(a_i\right)\mid$, with $\eta = \frac{\epsilon}{2}$

$$2\sum_{i=1}^{2n+1} P\left(\left|F_n\left(t_i\right) - F_n'\left(a_i\right)\right| > \frac{\epsilon}{2}\right) \leq (8n+4)e^{\frac{-n\epsilon^2}{8}}$$

$$\rightarrow 0 \quad as\, n \rightarrow \infty$$

Generalizations

Many generalizations are possible.
1. The random variables $X_1, X_2, \cdots, X_n$ need only be independent; and do not have to be identically distributed. The limiting distribution is then $\bar{F}_n(a) = 1/n \sum F_i(a)$. (The limit is always obtained by replace the random variables by the expectations)
2. The constant $1/n$ may be replaced by other constants or a sequence of n constants: $c_1, c_2, \cdots, c_n$. The result will be

$$P\left(\sup_{-\infty<a<\infty} \sum_{i=1}^{n} |c_i I\left[X_i \leq a\right] - c_i F_i(a)| > \epsilon\right) \leq (8n{+}4) \exp\left[-\frac{\epsilon^2}{8\sum_{i=1}^{n} 1/c_i^2}\right];$$

3. The limit do not have to be distribution functions. Any bounded non random function will do. In particular a sub-distrbution function.

$$\sup_a \sum_{i=1}^{n} c_i \left| I_{[X_i \leq a, \delta_i = 1]} - U_i(a) \right|$$

where $U_i(a) = E I_{[X_i \leq a, \delta_i = 1]}$.