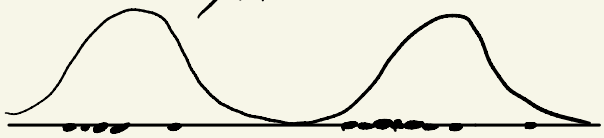
GLIVENKO–CANTELLI THEOREM

- Basic problem in statistics: what can we learn about the unknown distribution from a random sample?

- Mean: Let $X, X_1, X_2, \dots$ be iid rv's with finite mean μ.



SLLN: $\frac{1}{n}\sum_{1}^{n} X_i \xrightarrow{a.s.} \mu$.

- Entire distribution? i.e. is

$$F(x) = P\{X \le x\} \qquad \text{(cdf)}$$

$$\overset{?}{\approx} F_n(x) = \frac{1}{n}\#\{i=1,\dots,n : X_i \le x\} \qquad \text{(empirical cdf)} \ ?$$

$$\forall x \in \mathbb{R}: \quad F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\mathbb{1}_{\{X_i \le x\}}}_{\text{iid, mean} = P\{X \le x\}} \xrightarrow[\text{SLLN}]{a.s.} P\{X \le x\} = F(x)$$

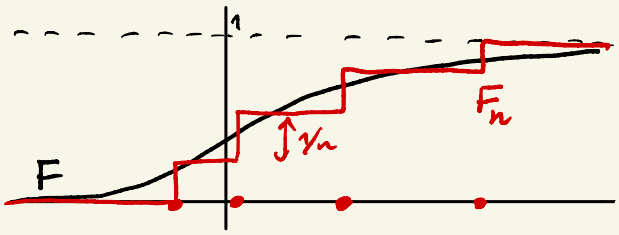
- Improvement from pointwise to uniform convergence:

THM [Glivenko–Cantelli '1933] Let $X_1, X_2, \dots$ be iid rv's. Then

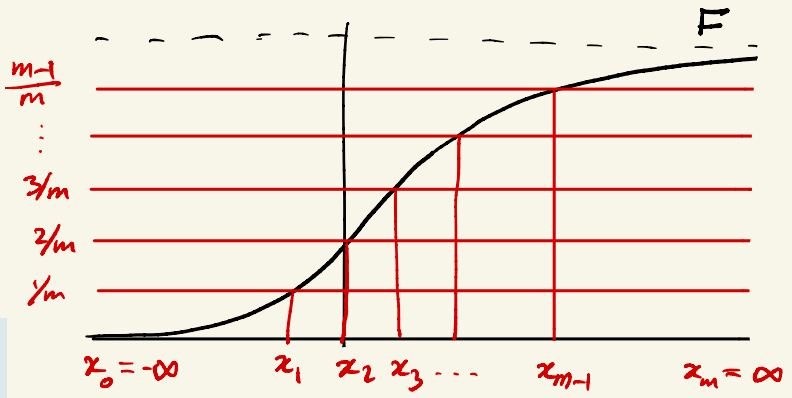
$$\|F_n - F\|_\infty = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \xrightarrow{a.s.} 0$$

- GCT allows the statistician to interact with data:

estimate $P\{a \le X \le b\}$ where a, b depend on data.



Proof ① For continuous F:



Fix $m \in \mathbb{N}$, choose $x_j \in \mathbb{R} \cup \{\pm\infty\}$:

$$F(x_j) = \frac{j}{m}, \quad j = 0, \dots, m.$$

x_0 can be $-\infty$, x_m can be $+\infty$
$F(-\infty) = 0$, $F(+\infty) = 1$
well defined since F is continuous and
$F(x) \to 0$ as $x \to -\infty$, $F(x) \to 1$ as $x \to \infty$

Then

$$F(x_j) - F(x_{j-1}) = \frac{1}{m}, \quad j = 1, \dots, m. \qquad (*)$$

- SLLN $\Rightarrow \forall j$, $|F_n(x_j) - F(x_j)| \xrightarrow{a.s.} 0$ (see p. 100) $\Rightarrow$

$$\varepsilon_n := \max_j |F_n(x_j) - F(x_j)| \xrightarrow{a.s.} 0 \qquad (**)$$

- $\forall x \in \mathbb{R} \quad \exists j: \quad x_{j-1} \le x \le x_j \Rightarrow$

$$\begin{aligned} F_n(x) - F(x) &\le F_n(x_j) - F(x_{j-1}) && \text{(monotonicity)} \\ &\le F_n(x_j) - F(x_j) + \frac{1}{m} && \text{(by } *) \\ &\le \varepsilon_n + \frac{1}{m} && \text{(by } **) \end{aligned}$$

Similarly,

$$\begin{aligned} F_n(x) - F(x) &\ge F_n(x_{j-1}) - F(x_j) && \text{(monotonicity)} \\ &\ge F_n(x_{j-1}) - F(x_{j-1}) + \frac{1}{m} && \text{(by } *) \\ &\ge -\varepsilon_n + \frac{1}{m} && \text{(by } **). \end{aligned}$$

$\Rightarrow$

$$|F_n(x) - F(x)| \le \varepsilon_n + \frac{1}{m} \quad \forall x \in \mathbb{R} \Rightarrow$$

$$\|F_n - F\|_\infty \le \varepsilon_n + \frac{1}{m} \quad \forall n \in \mathbb{N}$$

$\overset{(**)}{\Rightarrow}$ with prob. 1, $\limsup_n \|F_n - F\|_\infty \le \frac{1}{m} \quad \forall m \in \mathbb{N}$

with prob. 1, $\|F_n - F\|_\infty \to 0$. $\square$

② For arbitrary F,

Use right continuity:

$$F(x^+) := \lim_{y \downarrow x} F(y) = F(x)$$

Also,

$$F(x^-) := \lim_{y \uparrow x} F(x) = \lim_{y \uparrow x} P\{X \le y\} = P\{X < x\}$$ (by continuity of probability)

Fix $m \in \mathbb{N}$, define

$$x_j := \inf\left\{x : F(x) \ge \frac{j}{m}\right\}, \quad j = 0, \dots, m$$

(as before, x_0 can be $-\infty$, x_m can be $+\infty$)

- Claim:
$$F(x_j^-) - F(x_{j-1}) \le \frac{1}{m}, \quad j = 1, \dots, m. \qquad (*)$$

$F(x_j^-) = \lim_{x \uparrow x_j} F(x) \le \frac{j}{m}$ (∧ def of x_j)

$F(x_{j-1}) = \lim_{x \downarrow x_{j-1}} F(x) \ge \frac{j-1}{m}$ (∨ def of x_{j-1})

Moreover, by right continuity,

$$\Rightarrow F(x_j^-) - F(x_{j-1}) \ge \frac{j}{m} - \frac{j-1}{m} = \frac{1}{m}.$$

- SLLN $\Rightarrow$
$$\varepsilon_n := \max_j \left|F_n(x_j) - F(x_j)\right| \xrightarrow{a.s.} 0 \quad \leftarrow \text{(as before)}$$
$$\delta_n := \max_j \left|F_n(x_j^-) - F(x_j^-)\right| \xrightarrow{a.s.} 0 \qquad (**)$$

$F_n(x_j^-) = \lim_{y \uparrow x} \frac{1}{n} \sum_1^n P\{X_n \le y\} = \frac{1}{n} \sum_1^n P\{X_n < x\} \xrightarrow{SLLN} P\{X < x\} = F(x_j^-)$

- $\forall x \in \mathbb{R} \ \exists j: \ x_{j-1} \le x \le x_j \Rightarrow$

$$F_n(x) - F(x) \le F_n(x_j^-) - F(x_{j-1}) \quad \text{(monotonicity)}$$
$$\le F_n(x_j^-) - F(x_j^-) + \frac{1}{m} \quad (\text{by } *)$$
$$\le \delta_n + \frac{1}{m} \quad (\text{by } **).$$

Similarly,

$$F_n(x) - F(x) \ge F_n(x_{j-1}) - F(x_j^-) \quad \text{(monotonicity)}$$
$$\ge F_n(x_{j-1}) - F(x_{j-1}) - \frac{1}{m} \quad (\text{by } *)$$
$$\ge -\varepsilon_n - 1/m \quad (\text{by } **)$$

- $\Rightarrow |F_n(x) - F(x)| \le \max(\varepsilon_n, \delta_n) + 1/m$. Finish the proof as in ①. □

Remarks:

1. $F_n(x) = \frac{1}{n} \#\{i \le n : X_i \le x\}$ is the cdf of the prob. measure

$$\mu_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$$

This random meas. μ_n is called the empirical distribution.

2. GCT can be stated as follows:

$$\sup_B \left| \frac{1}{n} \#\{i \le n : X_i \in B\} - P\{X \in B\} \right| \xrightarrow{a.s.} 0$$

where the sup is over all intervals $B = (-\infty, x]$.

$\Rightarrow$ same is true for intervals $B = (a, b]$ or $[a, b]$ (Why?)

Question: is this true for all Borel sets B ?

No: $B = \{X_1, \dots, X_n\}$ $\quad |1 - 0| \not\to 0$

What families of sets B is this true for?

Ans: $\forall$ family with finite VC dimension (see e.g. my book)

$\downarrow$

ML

3. Rate of convergence in GCT is $O(1/\sqrt{n})$:

Dvoretzky-Kiefer-Wolfowicz inequality '1956:

$$\mathbb{P}\{\|F_n - F\|_\infty > \varepsilon\} \le 2e^{-n\varepsilon^2}$$

↑ [Massart 1990]

4. $\Rightarrow$ Kolmogorov-Smirnov test for whether 2 samples come from the same distribution (e.g. if a drug has any effect):

(in this case, $\|F_n - G_n\|_\infty \le \|F_n - F\|_\infty + \underbrace{\|F - G\|_\infty}_{0} + \|G_n - G\|_\infty = O(1/\sqrt{n})$)