Empirical distribution function and fundamental theorem of statistics

Olga Klopp

Consider the sampling on $\mathbb{R}$: we observe

$$X_1, \dots, X_n$$

i.i.d. random variables following the law $\mathbb{P}_X$.

Rem. : As the law of observations $(X_1, \dots, X_n)$ is $\mathbb{P}_X^{\otimes n}$, to give a model here (for the sampling model) is equivalent to give a model for $\mathbb{P}_X$.

For example: $\mathbb{P}_X \in \{\mathcal{N}(\theta, 1) : \theta \in \mathbb{R}\}$

Fundamental Question: If we consider the "total" model $= \mathbb{P}_X \in \{$ all the laws on $\mathbb{R}\}$, is it possible to know **exactly** $\mathbb{P}_X$ when n, the number of observations, goes to ∞?

1 Empirical distribution function

Recall: To know the law of a random variable X it is enough to know its distribution function F and the distribution function is easier to study in a statistical context. So, we observe

$$X_1, \dots, X_n \sim_{i.i.d.} F,$$

F **any, unknown** distribution function.

Question: Is it possible to find exactly F when n goes to ∞?

Idea: We will try to estimate F. Let $x \in \mathbb{R}$, then $F(x) = \mathbb{P}[X \leq x]$ is the probability that X will take a value less than or equal to x. We will then count the numbers of X_i which are smaller than x and divide it by n:

$$\frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x).$$

Definition 1. **Empirical distribution function** *associated with n-sample $(X_1, \dots, X_n)$:*

$$\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I\big(X_i \leq x\big),\ x \in \mathbb{R}\,.$$

This cumulative distribution function is a step function that jumps up by $1/n$ at each of the n data points. Its value at any point x is the fraction of observations of the measured variable that are less than or equal to x.

It's a random variable!

2 Fundamental theorem of statistics

2.1 Asymptotic properties of $\widehat{F}_n$

For any fixed $x \in \mathbb{R}$, the empirical distribution function $\widehat{F}_n(x)$ converges almost surly to the true distribution function $F(x)$:

$$\boxed{\widehat{F}_n(x) \xrightarrow{a.s.} F(x) \text{ when } n \to \infty}$$

This is a consequence of the **strong law of large numbers** applied to the sequence of i.i.d. r.v. $\big(I(X_i \le x)\big)_i$. We say that $\widehat{F}_n(x)$ is **strongly consistent** estimator of $F(x)$.

Theorem 1 (Glivenko-Cantelli)**.**

$$\left\|\widehat{F}_n - F\right\|_\infty \xrightarrow{a.s.} 0 \text{ when } n \to \infty$$

Also called the **Fundamental Theorem of Statistics**.

Interpretation: With an infinite number of data, one can reconstruct exactly the distribution function F and therefore exactly determine the law of observations.

R Example Glivenko-Cantelli

Let $x \in \mathbb{R}$. We know that if $n \to \infty$ then

$$\widehat{F}_n(x) \xrightarrow{a.s.} F(x)$$

Question : What is the rate of convergence of $F_n(x)$ to $F(x)$?

Tool : **Central-limit theorem** applied to the sequence of i.i.d. r.v. $\big(I(X_i \le x)\big)_i$:

$$\boxed{\sqrt{n}\big(\widehat{F}_n(x) - F(x)\big) \xrightarrow{d} \mathcal{N}\big(0, F(x)(1 - F(x))\big)}$$

We say that $\widehat{F}_n(x)$ is **asymptotically normal** with **asymptotic variance** $F(x)(1 - F(x)$. CLT implies that

$$\mathbb{P}\left[\left|\widehat{F}_n(x) - F(x)\right| \ge c_\alpha \frac{\sigma(F)}{\sqrt{n}}\right] \to \int_{|x|>c_\alpha} \exp(-x^2/2)\frac{dx}{\sqrt{2\pi}} = \alpha$$

for any $0 < \alpha < 1$, when $n \to \infty$. Here $\sigma(F) = F(x)(1 - F(x))$ and $c_\alpha = \Phi^{-1}(1 - \alpha/2)$.

- Attention! this **does not** provide a confidence interval: $\sigma(F) = F(x)^{1/2}\big(1 - F(x)\big)^{1/2}$ is unknown!
- Solution : replace $\sigma(F)$ by $\sigma(\widehat{F}_n) = \widehat{F}_n(x)^{1/2}\big(1 - \widehat{F}_n(x)\big)^{1/2}$ (that we observe), thanks to **Slutsky's theorem**.

Proposition 1. *For any* $\alpha \in (0, 1)$,

$$\mathcal{I}_{n,\alpha}^{\texttt{asymp}} = \left[\widehat{F}_n(x) \pm \frac{\widehat{F}_n(x)^{1/2}\big(1 - \widehat{F}_n(x)\big)^{1/2}}{\sqrt{n}}\Phi^{-1}(1 - \alpha/2)\right]$$

is an asymptotic confidence interval for $F(x)$ *at the confidence level* $1 - \alpha$ *:*

$$\mathbb{P}\left[F(x) \in \mathcal{I}_{n,\alpha}^{\texttt{asymp}}\right] \to 1 - \alpha.$$

R Example Confidence interval $F(x)$.

Theorem 2 (Kolmogorov-Smirnov's Theorem)**.** *Let* X *be a random variable with cdf* F *that we suppose continuous and* $(X_n)_n$ *sequence of i.i.d. r.v. all having the same law as* X*. Then,*

$$\sqrt{n}\left\|\widehat{F}_n - F\right\|_\infty \xrightarrow{d} K$$

where K *is a random variable such that for any* $x \in \mathbb{R}$

$$\mathbb{P}[K \le x] = 1 - 2\sum_{k=1}^{\infty}(-1)^{k+1}\exp(-2k^2x^2).$$

- Useful for Kolmogorov-Smirnov test
- non-asymptotic version of this result: when F is continuous, the law of $\left\|\widehat{F}_n - F\right\|_\infty$ is independent of F

In general, statistical results can be classified into two categories:

1. A result when n goes to infinity is an asympotic result
2. A result when n is fixed is a non-asympotic result

2.2 Non-asymptotic estimation of $F(x)$ by $\widehat{F}_n(x)$

Given a (small) $0 < \alpha < 1$ we want to find ε, as small as possible, so that

$$\mathbb{P}\left[|\widehat{F}_n(x) - F(x)| \geq \varepsilon\right] \leq \alpha.$$

Using Chebyshev we get

$$\begin{aligned}
\mathbb{P}\left[|\widehat{F}_n(x) - F(x)| \geq \varepsilon\right] &\leq \frac{1}{\varepsilon^2}\mathrm{Var}\big[\widehat{F}_n(x)\big] \\
&= \frac{F(x)\big(1 - F(x)\big)}{n\varepsilon^2} \\
&\leq \frac{1}{4n\varepsilon^2} \\
&\leq \alpha
\end{aligned}$$

Leads to

$$\boxed{\varepsilon = \frac{1}{2\sqrt{n\alpha}}}$$

Conclusion: for any $\alpha > 0$,

$$\mathbb{P}\left[|\widehat{F}_n(x) - F(x)| \geq \frac{1}{2\sqrt{n\alpha}}\right] \leq \alpha.$$

Terminology 1. *The interval*

$$\boxed{\mathcal{I}_{n,\alpha} = \left[\widehat{F}_n(x) \pm \frac{1}{2\sqrt{n\alpha}}\right]}$$

is a $1 - \alpha$ confidence interval for $F(x)$.

Proposition 2 (Hoeffding's inequality)**.** *$Y_1, \ldots, Y_n$ i.i.d. r.v. such that $a \leq Y_1 \leq b$ a.s.. Then,*

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} Y_i - \mathbb{E}\,Y_1\right| \geq t\right] \leq 2\exp\left(-\frac{2nt^2}{(a-b)^2}\right)$$

Application: set $Y_i = I(x_i \leq x)$. From Hoeffding inequality We can deduce

$$\mathbb{P}\left[\big|\widehat{F}_n(x) - F(x)\big| \geq \varepsilon\right] \leq 2\exp(-2n\varepsilon^2).$$

We solve in ε:

$$2\exp(-2n\varepsilon^2) = \alpha,$$

that is

$$\boxed{\varepsilon = \sqrt{\frac{1}{2n}\log\frac{2}{\alpha}}}.$$

2.2.1 Chebyshev vs Hoeffding comparison

New confidence interval

$$\boxed{\mathcal{I}_{n,\alpha}^{\texttt{hoeffding}} = \left[\widehat{F}_n(x_0) \pm \sqrt{\frac{1}{2n}\log\frac{2}{\alpha}}\right]},$$

to compare with

$$\mathcal{I}_{n,\alpha}^{\texttt{chebyshev}} = \left[\widehat{F}_n(x_0) \pm \frac{1}{2\sqrt{n\alpha}}\right].$$

- Same order of magnitude with respect to n: $1/\sqrt{n}$
- Significant gain in the limit $\alpha \to 0$: from $1/\alpha$ to $\log(1/\alpha)$. The "risk taking" becomes marginal compared to the number of observations.
- **Optimality of this approach?**

Comparison of the lengths of the 3 confidence intervals:

- Chebyshev (non-asymptotic) $\frac{2}{\sqrt{n}}\frac{1}{2}\frac{1}{\sqrt{\alpha}}$
- **Hoeffding (non-asymptotic)** $\frac{2}{\sqrt{n}}\sqrt{\frac{1}{2}\log\frac{2}{\alpha}}$
- CLT (asymptotic) $\frac{2}{\sqrt{n}}\widehat{F}_n(x_0)^{1/2}\big(1-\widehat{F}_n(x_0)\big)^{1/2}\Phi^{-1}(1-\alpha/2)$.
- The smallest length is provided by the CLT. But the length of the confidence interval provided by the Hoeffding's inequality comparable to CLT in n and α (in the limit $\alpha \to 0$)

2.2.2 Non-asymptotic version of Kolmogorov-Smirnov

Let X be a random variable with cdf F that we suppose **continuous** and $(X_n)_n$ sequence of i.i.d. r.v. all having the same law as X. Let $\widehat{F}_n$ be the corresponding empiric distribution function

Proposition 3 (Dvoretsky-Kiefer-Wolfowitz inequality)**.** *For any $\varepsilon > 0$.*

$$\mathbb{P}\big[\sup_{x\in\mathbb{R}}\big|\widehat{F}_n(x) - F(x)\big| \geq \varepsilon\big] \leq 2\exp\big(-2n\varepsilon^2\big).$$

- Difficult result (theory of empirical processes).
- We can construct confidence regions with results similar to those obtained for a fixed value of x :

$$\mathbb{P}\Big[\forall \mathbf{x} \in \mathbb{R}, F(x) \in \big[\widehat{F}_n(x) \pm \sqrt{\tfrac{1}{2n}\log\tfrac{2}{\alpha}}\big]\Big] \geq 1-\alpha$$

Remark: Kolmogorov-Smirnov Test it is used for testing if a sample comes from a distribution with a given cdf or to test if two samples come from the same distribution.