Université Paris Dauphine
Département MIDO
Licence 3 – Statistical modelling

2022–2023

Practical n°3

stoehr@ceremade.dauphine.fr

Notations and Objectives. Let $X_1, \dots, X_n$ be a sequence of *i.i.d.* random variables with an **unknown** cumulative distribution function F and denote $x_1, \dots, x_n$ realisations of these random variables. The practical focuses on the empirical distribution of data $\widehat{F}_n$. The first objective is to study the properties of $\widehat{F}_n$ as an estimate of F for a given sample $x_1, \dots, x_n$. The second objective is to sample from it.

Exercise. A marketing brand studies the time spent by customers on its website. The file "time.csv" contains the duration (in seconds) for $n = 100$ customers.

1. Download "time.csv" in a folder "TP3" and import the data in R using the command `read.table` or `read.csv`.

The empirical distribution function. We are first interested in getting some information about the underlying unknown distribution F the data are coming from.

Definition

The empirical distribution function is a random function (as function of the random variables $X_1, \dots, X_n$) associated to the empirical measure and given for $y \in \mathbb{R}$ by

$$\widehat{F}_n(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq y\}} = \frac{\text{Card}\{X_i \mid X_i \leq y\}}{n}.$$

Its evaluation on a dataset $x_1, \dots, x_n$ brings information about the distribution contains in the data itself.

2. Write a function `cdf_hat(y, x)` which returns the empirical distribution at point y given a n-sample `x` $= (x_1, \dots, x_n)$.

$\widehat{F}_n$ is an estimate of the cumulative distribution F that generated the data. The first major result is that it converges pointwise.

Theorem. Strong Law of Large Numbers (SLLN)

For all y, $\widehat{F}_n(y)$ converges pointwise to the true cumulative distribution function $F(y)$ as almost surely (*a.s.*), *i.e.*,

$$\widehat{F}_n(y) \underset{n \to +\infty}{\longrightarrow} F(y) = \mathbb{P}\left[X_1 \leq y\right] \quad a.s.$$

3. Let assume the brand is interested in the probability that a customer spend less than 50 seconds on its website.

(a) Compute `cdf_hat(50, x)`.

(b) Write a function `cv_cdf_hat(y, x)` which returns $\left(\widehat{F}_1(y), \widehat{F}_2(y), \ldots, \widehat{F}_n(y)\right)$ for a given point `y` and a given sample $\mathtt{x} = (x_1, \ldots, x_n)$.

(c) Using a scatterplot, check if $\widehat{F}_n(50)$ converged or not with our sample size n.

$\widehat{F}_n$ is a random function and hence has a variability depending on the data it is based on, *i.e.*, $\widehat{F}_n$ changes when we use other data. The randomness of $\widehat{F}_n$ can be asymptoticaly characterised pointwise.

Theorem. Central Limit Theorem (CLT)

For all $y \in \mathbb{R}$, $\widehat{F}_n(y)$ has asymptotically normal distribution with the standard $\sqrt{n}$ rate of convergence, *i.e.*,

$$\sqrt{n}\left[\widehat{F}_n(t) - F(t)\right] \underset{n\to+\infty}{\xrightarrow{d}} \mathcal{N}(0, F(t)\{1 - F(t)\})$$

Remark. This result is a simple application of the Central Limit theorem to the sequence of *i.i.d.* random variables $\left(\mathbb{1}_{\{x_n \le t\}}\right)_{n\ge 1}$, that is a sequence of Bernoulli variables with parameter equal to $F(t)$.

4. **(a)** Show that

$$\sqrt{\frac{n}{\widehat{F}_n(t)(1-\widehat{F}_n(t))}}\left[\widehat{F}_n(t) - F(t)\right] \underset{n\to+\infty}{\xrightarrow{d}} \mathcal{N}(0,1).$$

(b) Assume that the asymptotic regime is valid. Compute the asymptotic confidence interval for $F(50)$ for 0.95 confidence level. Comment the result.

We actually have more than the pointwise convergence of the empirical distribution to the true cumulative distribution function.

Theorem. Glivenko-Cantelli theorem

$\widehat{F}_n$ converges uniformly to F over $\mathbb{R}$, *i.e.*,

$$\left\|\widehat{F}_n - F\right\|_\infty \underset{n\to+\infty}{\longrightarrow} 0 \quad a.s.$$

5. **(a)** Write a function `cdf_hat_vec(y, x)` which returns the empirical distribution function at multiple points $\mathtt{y} := (y_1, \ldots, y_k)$ given a n-sample $\mathtt{x} = (x_1, \ldots, x_n)$.

(b) Check if $\widehat{F}_n$ converges uniformly to the Gamma distribution with shape parameter $\alpha = 1.9$ and rate parameter $\beta = 0.04$.

We can further characterized the asymptotic behaviour of the sup-norm between $\widehat{F}_n$ and F. We also have a rate of convergence of $\sqrt{n}$ and the Dvoretzky–Kiefer–Wolfowitz inequality provides bound on the tail probabilities.

Theorem. Dvoretzky–Kiefer–Wolfowitz inequality

For all $\varepsilon \geq \sqrt{\frac{1}{2n}\log 2}$,

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

As per this inequality, we have another confidence interval for the true cumulative distribution function value $F(y)$, $y \in \mathbb{R}$. Given $\alpha \in]0,1[$,

$$\mathbb{P}\left[\max\left\{0, \widehat{F}_n(y) - \varepsilon\right\} \leq F(y) \leq \min\left\{\widehat{F}_n(y) + \varepsilon, 1\right\}\right] \geq 1 - \alpha, \quad \text{where} \quad \varepsilon = \sqrt{\frac{1}{2n}\log\frac{2}{\alpha}}.$$

6. Compute the confidence interval based on the Dvoretzky–Kiefer–Wolfowitz inequality for $F(50)$ and 0.95 confidence level. Comment the result.

Sampling from the empirical distribution function. Another key feature of the empirical distribution is that it is easy to sample from. The empirical distribution defines a discrete probability distribution on sample space $\{x_1, \ldots, x_n\}$. Put in other words, if $X^\star$ follows $\widehat{F}_n$, then $X^\star$ is taking values in $\{x_1, \ldots, x_n\}$ with probabilities

$$\mathbb{P}\left[X^\star = x_i\right] = \frac{1}{n}, \quad \text{for} \quad i \in [\![1, n]\!].$$

Remark. Above we did not take into account possible repetitions of values in $x_1, \ldots, x_n$. We consider that observations are ordered elements and that indices for the probability vector match indices of the observation. For instance, if we have observations $(x_1, x_2, x_3) = (1, 2, 1)$, the probability vector is $(p_1, p_2, p_3) = (1/3, 1/3, 1.3)$.

Property

In order to get a m-sample from $\widehat{F}_n$, we sample m times from $\{x_1, \ldots, x_n\}$ with replacement according to the vector of probabilities $(p_1, \ldots, p_n)$. The re-sampled sample of size m is denoted $(x_1^\star, \ldots, x_m^\star)$ and the associated empirical distribution function $\widehat{F}_n^\star$.

7. **(a)** For each of the following values, $m = 10$, $m = 50$ and $m = 100$, generate 100 samples $(x_1^\star, \ldots, x_m^\star)$ from $\widehat{F}_n$.

(b) Compute the percentile of each sample.

(c) Why is it important to have a resample of the same size as the original sample?