Tutorial course N° 2

NOTION OF CONVERGENCE, CENTRAL LIMIT THEOREM (CLT) AND LAW OF LARGE NUMBERS (LLN)

Exercise 0. Change the working directory and create a script named `Lab2.R` that you will use to save all the results of this session.

1. Simulation of probability distributions

It is very useful to generate realizations of various probability laws. This can be done in `R` using the functions of the form `r`*func*`(n,param)` where *func* indicates the probability law, `n` is the sample size (number of realizations) and `param` are the parameters of the law. Herewith some examples :

Law	Function `R`
uniform $U[0,1]$	`runif(n)`
uniform $U[a,b]$	`runif(n, a, b)`
normal $\mathcal{N}(0,1)$	`rnorm(n)`
normal $\mathcal{N}(m,\sigma^2)$	`rnorm(n, m, sd)`
Poisson $\mathcal{P}(\lambda)$	`rpois(n, lambda)`
binomiale $\text{Bin}(k,p)$	`rbinom(n, k, p)`
exponential $\mathcal{E}(\theta)$	`rexp(n, theta)`
Gamma $\Gamma(a,b)$	`rgamma(n, a, b)`
Cauchy(θ)	`rcauchy(n, 0, theta)`

Exercise 1.

(1) Run the following commands and interpret the output :

```
> runif(5)
> rnorm(5)
> rnorm(5, 10, 1)
```

(2) Generate 300 observations of the exponential law $\mathcal{E}(4)$, save the values in a vector named `data.exp`.

(3) Generate 300 observations of the Poisson law $\mathcal{P}(2)$, save the values in a vector named `data.pois`.

The functions of the form `r`*func* (with *func*=`norm` or `unif`...) have associated functions :

— `d`*func* `(x,arguments)` : returns the probability density at point `x` if *func* is a continuous distribution. For discrete distributions, the probability of taking the value `x` is returned.
— `p`*func* `(x,arguments)` : returns the value of the cumulative distribution function in `x`,
— `q`*func* `(a,arguments)` : returns the quantile of order `a`,

(4) Calculate the value of the density of a standard normal distribution at point 0.

(5) Calculate the probability $\mathbb{P}(X \leq 0)$ for $X \sim \mathcal{N}(0, 5)$.

(6) Evaluate the density of the exponential law $\mathcal{E}(4)$ at 100 points belonging to the interval $[0, 4]$. Plot the histogram of the sample data `data.exp` using the function `hist()`. Overlay the theoretical density on the histogram of data `data.exp` (using the `lines()` function). Interpret the result.

(7) Calculate the probabilities $\mathbb{P}(X = k)$ of a random variable X following the Poisson distribution $\mathcal{P}(2)$ for all $k \in \{0, 1, \ldots, 10\}$. After drawing the bar graph for the sample data `data.pois` by running the command :

```
> plot(table(data.pois)/300)
```

Overlay these probabilities and interpret the result.

2. Central limit theorem (CLT)

Exercise 2.

(1) Consider the game of tossing a balanced coin. Let us assign 1 point for heads and 0 for tails. We are interested in the sum of points after n draws. Execute the following code and comment on the obtained results.

```
berTLC = function(n)
{
  X=seq(0,n)
  p=dbinom(X,n,0.5)
  return(p)
}

par(mfrow=c(2,5))
plot(seq(0,1),c(0.5,0.5),ylim=c(0,0.6),xlim=c(0,2),type="l",xlab="X",
ylab="Freq",main="n=1",col="red")
for(n in seq(2,10))
{
  plot(seq(0,n),berTLC(n),ylim=c(0,0.6),xlim=c(0,n+1),type="l",xlab="X",
  ylab="Freq",col="red", main=bquote(paste(,"n=",.(n))))
 }
```

* Now consider the game of rolling a balanced six-sided die and the sum of the values ? ?rolled on the top side, after n throws. Plot the graphs analogously to the previous exercise for n ranging from 1 to 4. Comment on the results.

(2) Create a function `tclbernoulli` which takes as input two parameters : an integer `N` and the parameter `p` of the Bernoulli distribution. This function simulates 1000 vectors $(X_1, \ldots, X_N)$ where X_i follows the Bernoulli distribution with parameter `p` and calculates for each vector the value of the following random variable :

$$S_N = \sqrt{N} \frac{\bar{X}_N - p}{\sqrt{p(1-p)}} ,$$

where $\bar{X}_N = \frac{1}{N} \sum_{i=1}^{N} X_i$. Notice that $\bar{X}_N$ is an estimator of p.

(3) Let be `p` $= 0.2$. We want to evaluate the probability $\mathbb{P}(S_N \in (-1.96, 1.96))$ for different values of `N`. To do this, call the `tclbernoulli` function for `N`$=$ 10, 100 and 1000, and calculate for each value of `N` the proportion of the realizations of S_N which belong to the interval $(-1.96, 1.96)$. What can you observe ? We remind that the quantile 97.5% of the standard normal distribution is about 1.96.

(4) Let $(X_1, \dots, X_n)$ be a sample of i.i.d. (independent and identically distributed) random variables of a law F. The empirical cumulative distribution function F_n is defined by

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i \leq t\} \quad \forall t\,.$$

In R the predefined function to calculate the empirical cumulative distribution function of a vector `x` is the `ecdf()` function. Use the function `plot.ecdf()` to plot the empirical cumulative distribution function of the vector returned by the function `tclbernoulli` for `p` =0.2 and `N` = 10, `N` = 100 and `N` = 1000. Overlay the theoretical cumulative distribution function of a standard normal N (0,1) law for comparison. What result of the probability theory is illustrated here ?

3. Law of large numbers (LLN)

Some functions in R that we will use :

— `mean(x)` : calculates the empirical mean of the vector `x`.

— `cumsum(x)` : returns a vector of the same length as `x` of the cumulative sums, that is, the i-th entry is the sum of the i first values of `x`. For example,

```
> cumsum(c(1,5,3,2))
[1]  1  6  9 11
```

Exercise 3.

(1) Which theorem of the probability theory is illustrated by the following code :

```
Nfin <- 5000
 X <- rexp(Nfin,2)
 Y <- cumsum(X)/1:Nfin
 plot(1:Nfin,Y,type = "l",ylim = c(0,1),xlab="n",ylab="empirical mean")
 for (i in 2:50){
         X <- rexp(Nfin,2)
         Y <- cumsum(X)/1:Nfin
         lines(1:Nfin,Y,col = i)
         }
```

(2) Create the following function :

```
lgnexpo <- function(N){
  moy=rep(0,100)
  for (i in 1:100){
    moy[i]=mean(rexp(N,2))
  }
 return(moy)}
}
```

Analyze the code of this function and plot the boxplots of the vectors returned by the function for N = 100, 1000 and 10000. Comment on the results. What result of the probability theory is illustrated by this phenomenon ?

(3) * Write a function `lgncauchy()` that performs the same operations as the `lgnexpo()` for realizations of the Cauchy distribution. Compare the boxplots of the outputs of this function for different values of `N`. You can set the options outline = TRUE, then outline = FALSE in the boxplot function. What do you notice ? Explain the observed phenomenon.

4. Poisson distribution and Binomial distribution*

The Poisson law can be considered as the limit of a series of Binomial laws. More precisely, let $\lambda > 0$ be a real number and a sequence $p_n \in (0,1)$ such that

$$np_n \to \lambda\ , \quad n \to \infty\ .$$

Let $Y_n \sim \mathcal{B}(n, p_n)$ be a sequence of independent random variables and $Z \sim \mathcal{P}(\lambda)$. So, for all $k \in \mathbb{N}$,

$$\lim_{n\to\infty} \mathbb{P}(Y_n = k) = \lim_{n\to\infty} \binom{n}{k} p_n^k (1-p_n)^{n-k} = \mathbb{P}(Z = k) = \frac{\lambda^k}{k!} \mathrm{e}^{-\lambda}\ .$$

The purpose of this exercise is to illustrate this property.

Exercise 4. Set $\lambda = 8$.

(1) Generate four samples of size 1000, whose elements are i.i.d. following a binomial law $\mathcal{B}(n, p_n)$ of parameter $n \in \{10, 20, 30, 100\}$ and $p_n = \lambda/n$, respectively.

(2) Calculate the probabilities $\mathbb{P}(Z = k)$ for $k = 0, 1, \ldots, M$, where M denotes the maximum value in the four samples of the binomial distribution.

(3) Draw the bar charts for the four samples of binomial distribution. Overlay each time the theoretical probabilities of the Poisson distribution. Interpret the results.

(4) Plot the empirical cumulative distribution functions of the four samples and overlay the theoretical cumulative distribution function of the Poisson law. To plot a staircase function use the option `type='s'`. Interpret the plots.