Université Paris Dauphine — Année 2012-2013
U.F.R. Mathématiques de la Décision / L3
Mat-6-3 — NOISE

Exercise sheet 0 :

Initiation to R

Preliminary steps

- To start R, you either open a terminal window and lauch R from the command line, or call R or Rkward by clicking on the appropriate icon.
- **How to include comments** : put ♯ before the comments.
- **On-line help** : use `?xx` to get the documentation on the function `xx` and its uses. Do not forget this highly helpful shortcut to `help(xx)`.
- Always make sure to **save your code in a file** to avoid disasters (and to be prepared for the exam). On a terminal window, the basic instruction `source("code.R")` loads all the functions defined in this file `code.R` and execute any relevant R code. When using Rkward, there is a workspace window which can be edited by a straightforward editor and from which pieces of code can be loaded and executed. This workspace must be labelled in an recognisable way and periodically saved during the working session. In the sad event you erase this file, the commands of the current session can be found in the hidden file `.Rhistory` and the output is saved in the corresponding `.RData`. Ask your instructor before handling those files.
- **All the answerts to the following exercises are provided in the reference manual "Initiation to R" by Robin Ryder and Jean-Michel Marin, available on your account, which should be read and mastered by the first fortnight of the course. It is highly recommended to test all instructions on a machine.**

1 Object manipulation

1.1 Vector manipulation

1. Create a vector `v1=( 1, 4, -3, 78, 9)`.
2. Display `v1`, then display only the 3rd component of `v1`.
3. Create `v2` that contains the 2nd and 4th terms of `v1`.
4. Create `v3` that contains the 2nd up to 4th terms of `v1`.

5. Create `v4` by concatenating `v1` et 12, then `v5` by concatenating `v2` and `v3`.
6. Multiply `v1` by 2, then only its 3rd term by 10.
   Note : you can implement the same principe when adding, subtracting, etc.
7. Add the two vectors `v2` and `v3`, then the two vectors `v1` and `v5`, which sizes differ. What is the result ?
8. Derive the sum and product of all the components of `v1`.
9. Determine the number of elements in the vector `v1`.
10. Transpose the vector `v1`.
11. Compute the scalar product between the vectors `v1` and `v5`.

1.2 Useful functions

Create the following vectors :

1. `x1=(1 1 1 1 1 1 1 1 1 1)` and `x2=( 1 2 3 4 1 2 3 4 1 2 3 4)`, using the property that `x1` is made of 10 repetitions of the integer `1` while `x2` is made of 3 repetitions of the vector `(1 2 3 4)`.
2. Find two alternative codes to construct `x3=( 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0)`.
3. Operate a random equiprobable draw with no replication of 5 elements from the vector `x3`.
   What should you modify to allow repetition ? to modify the probabilities into `p=( 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.2)` ?

1.3 Matrices

1. Create the matrices

```
 > m1                                  and      > m1bis
     [,1] [,2] [,3] [,4] [,5]                        [,1] [,2] [,3] [,4] [,5]
[1,]    1    4    7   10   13                   [1,]    1    2    3    4    5
[2,]    2    5    8   11   14                   [2,]    6    7    8    9   10
[3,]    3    6    9   12   15                   [3,]   11   12   13   14   15
```

   by a transform of the vector `(1 2 3 4 5 6 7 8 9 10 11 12 13 14 15)`,

   then the matrices

```
 > m2                                  and      > m3
     [,1] [,2] [,3] [,4] [,5]                        [,1] [,2] [,3] [,4] [,5]
[1,]    1    3    5    7    9                   [1,]    3    6    9   12   15
[2,]    2    4    6    8   10                   [2,]    4    7   10   13   16
[3,]    5    8   11   14   17
```

2. What happens to the R command building a matrix out of a vector with the inappropriate number of elements ?

   Test this instance by calling `m4=matrix((1:10), nrow=4,ncol=5)`.

3. Compute the sum, the term-by-term product, and the matricial product of the matrices `m1` and `m1bis`.

   What occurs if `m1bis` is replaced with `m3` ?

4. Extract some elements from `m1` : the $(1, 3)$ element, the 1st row, the 3rd column, both 1st and 3rd columns, all rows but the 2nd.

5. Exhibit the elements of `m1` that are larger than 10, replace them by 10.

6. Concatenate `m1` and `m1bis` verticaly, then horizontaly.

7. Compute the sum of the rows, then of the columns, of `m1`.

8. Create a matrix `msquare` as follows,

```
> msquare
     [,1] [,2] [,3] [,4]
[1,]    1    5    9   13
[2,]    2    6   10   14
[3,]    3    7   11   15
[4,]    4    8   12   16
```

   then derive its eigenvectors and eigenvalues.

1.4 Lists

1. Create a list object made of `x1,m1,a=TRUE`.
2. Extract the vector from this list by its name, then by its position in the list.

2 Probability distributions

This section is essential, practice again and again to avoid confusing `dnorm` with `rnorm`, or mixing the position of the parameters... Use `help` or `arg` when not sure.

After practising the instruction provided in the R manual about distributions :

1. Simulate a sample of 100 r.v.'s with a uniform $\mathcal{U}([0, 10])$ distribution.
2. Compute the value of a normal density at $x = 2$ when its mean is 5 and its variance 4.
3. Determine the 50% quantile of a Poisson distribution with mean parameter 2.
4. Find the cdf of a standard Cauchy distribution in $x = 1$.

3 Functions

Once again, make sure to store and save your own functions in appropriate files like `mycode.R`. Under a terminal window, use the instruction `source(''mycode.R'')` to load and execute those functions!!!

1. Write a function called `mymean` that returns the empirical average of a vector of arbitrary length n of normal $\mathcal{N}(0,1)$ r.v.'s. Give the values of `mymean(10)`, `mymean(100)`,`mymean(1000)`.
2. Modify this function towards the computation of the average of a sample of size n from a $\mathcal{N}(\mu,\sigma^2)$ distribution for μ and σ additional parameters of the function. Apply for $n=10^4$, $\mu=5$ and $\sigma=2$.
3. Write another function `moments` that outputs both the average and the empirical variance of a given sample.

4 Loops, etc...

1. **For :** Write a function that returns all integers from 1 to n.
2. **If :** Write a function that produces a r.v. X uniform over $[0,1]$ and outputs X if $X>0.5$, 0.5 otherwise.
3. **While :** Write a function that produces a r.v. X with distribution a truncated $N(0,1)$ distribution over $(-\infty,2)$.

5 Histograms

The function `hist` is used to produce a rudimentary approximation of the density of an iid sample $x_1,\ldots,x_n$).

1. Recover all the arguments one can use when calling `hist` and separate the necessary arguments from the optional ones.
2. Describe the elements of the list return by `hist(x)` and identify those you do not understand.
3. Explain why the grey area can have an area equal to either one or n, depending on the choice of a specific option.
4. Given that `hist(x)$density` provides a sequence of weights over the intervals defined by `{hist(x)$breaks`, show how to plot the density approximation given by `hist(x)`
5. The number of intervals in the histogram approximation is specified by the argument `nclass}` of `hist(x)`. For a sample of 100 points from a $\mathcal{N}(0,1)$ distribution, plot the evolution of the density approximation as `nclass}` increases.

Université Paris Dauphine Année 2011-2012
U.F.R. Mathématiques de la Décision
Stat. Exploratoire

Exercise sheet # 1

Simulation of random variables : cdf inversion, Box-Müller algorithm et accept-reject algorithm

1 Generic inversion

Fundamental principle : Bases

Given a uniform r.v. U over $[0,1)$ and a cdf F_X corresponding to the r.v. X, $F_X^{-1}(U)$ has the same distribution as X

Proof. Obviously, $P(F_X^{-1}(U) \leq x) = P(U \leq F_X(x)) = F_X(x)$.
Note. When F_X is not (strictly) increasing, and hence non-invertible, we define the generalised inverse by

$$F_X^{-1}(u) = \inf\{x; F_X(x) \geq u\}$$

Exercise 1 : An illustration of the inversion technique

1. Write an R function that simulate a sample $(X_1, \ldots, X_n)$ with size n such that the X_i's are i.i.d. distributed from an exponential distribution with parameter λ, when using the cdf inversion technique.
2. Simulate a sample of size 10^4 from an exponential distribution with parameter 4 using this function. Demonstrate graphically that the histogram of the resulting sample fits the exponential density modulo the Monte Carlo variations.
3. Repeat the above question for the Cauchy distribution.

2 Box-Müller transform

Box-Müller transform : Basics

If $U_1, U_2 \sim_{i.i.d.} \mathcal{U}[0,1]$, then, when

$$X_1 = \sqrt{-2\,\log(U_1)}\,\sin(2\pi U_2) \quad X_2 = \sqrt{-2\,\log(U_1)}\,\cos(2\pi U_2)$$

X_1 and X_2 are i.i.d. $\mathcal{N}(0,1)$.

Exercise 2 : Application of the Box-Müller transform and Cauchy distribution

Take a Cauchy $\mathcal{C}(0,1)$ random variable X with density

$$f(x) = \frac{1}{\pi}\frac{1}{1+x^2}$$

1. Show (or accept) that we can simulate realisations of X via the Box-Müller algorithm, thanks to the following property : if X_1 , X_2 are i.i.d with distribution $\mathcal{N}(0,1)$, then $\frac{X_1}{X_2} \sim \mathcal{C}(0,1)$.
2. Study the evolution of the empirical average

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

   when $X_i \sim_{i.i.d.} \mathcal{C}(0,1)$ and $n \geq 1$ increases from 1 to 10^4. What is your intuition about the observed phenomenon ?
3. Show that the distribution $\mathcal{C}(0,1)$ has no mean. What is the consequence on $\bar{X}_n$?
4. *Take-home problem* : Write a Monte Carlo experiment that would establish that $\bar{X}_n$ is also a Cauchy $\mathcal{C}(0,1)$ random variable, for all $n \geq 1$

3 Accept-reject algorithm

Accept-reject algorithm : Basics

One aims at generating a realisation of the random variable X with a distribution represented by the density f.

1. Obtain a density g that can be simulated and such that $\sup_x \frac{f(x)}{g(x)} = M$. ($M \in]1, < \infty[$)
2. Generate

$$Y_1, Y_2, \ldots \sim_{i.i.d.} g\,, \qquad U_1, U_2, \ldots \sim_{i.i.d.} \mathcal{U}([0,1])$$

3. Take $X = Y_k$ where

$$k = \inf\{n\,; U_n \leq f(Y_n)/Mg(Y_n)\}$$

The random variable resulting from the above is distributed from f_X.

Exercise 3 : Application of the Accept-reject algorithm

1. Using the Accept-reject method, generate a realisation of a $\mathcal{N}(0,1)$ distribution using only `rcauchy`.
2. Show that the constante M is $\sqrt{2\pi}e^{-1/2}$ by simulation.
3. Illustrate by a graph the accuracy of your algorithm.
4. Change the bound M and check its influence on the waiting time till an acceptance.
5. Using the Accept-reject method, generate a realisation of a random variable with density

$$f(x) = \frac{2}{5}(2+\cos(x))e^{-x}$$

   using only the function `rexp`. Establish the validity of your algorithm by graphical means.

Exercise 4 : Take-home problem : truncated variable generation

Consider a Gaussian random variable X that is centred, with variance 1 and restricted to the support $[a, b]$ avec $0 < b$

1. Give the density of this random variable and find the normalising constant.
2. Plot in R the probability $\mathbb{P}(Y \in [0, b])$ when $Y \sim \mathcal{N}(0, 1)$ and a and b vary.
3. Evaluate the efficiency of the algorithm that simulates $Y \sim \mathcal{N}(0, 1)$ until $Y \in [a, b]$
4. <u>Consider the case $a = 0$</u>. Propose an Accept-reject method, based on exponential $\mathcal{E}(\lambda)$ distributions. Optimise in λ and write the corresponding R function.

Université Paris Dauphine Année 2012-2013
U.F.R. Mathématiques de la Décision
Stat. Exploratoire

Exercise Sheet # 2

Monte Carlo Methods

1 Monte Carlo Integration

Monte Carlo integration : Bases

Let X be a random variable of density f and let h be a function defined on the support of X and such that $\int |h(x)|f(x)dx < \infty$. We want to evaluate

$$\mathfrak{I} = \int h(x)f(x)dx = \mathbb{E}_f[h(X)]\,.$$

In many situations, this integral cannot be explicitely calculated. A numerical approximation can be computed by Monte Carlo integration.
In principle : following the law of large numbers, if $X_1, \dots, X_n$ are independent et identically distributed random variables of density f, then

$$\lim_{n\to\infty} \hat{\mathfrak{I}}_n = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n h(X_i) = \mathfrak{I}, \quad \text{a.s.}.$$

In practice : simply simulate an n-sample $X_1, \dots, X_n \sim f$ and approximate $\mathfrak{I}$ with $\hat{\mathfrak{I}}_n$.

Convergence of $\hat{\mathfrak{I}}_n$: Let $\hat\sigma_n^2(h(X)) = \frac{1}{n}\sum_{i=1}^n (h(X_i) - \hat{\mathfrak{I}}_n)^2$ be the variance estimator of $h(X)$ and suppose that $\int |h(x)|^2 f(x)dx < \infty$. Following the Central Limit Theorem, we have

$$\lim_{n\to\infty} \sqrt{n}\frac{\hat{\mathfrak{I}}_n - \mathfrak{I}}{\hat\sigma_n(h(X))} = \mathcal{N}(0,1) \quad (\mathcal{L})\,,$$

that is $\hat{\mathfrak{I}}_n \sim \mathcal{N}\left(\mathfrak{I}, \frac{1}{n}\hat\sigma_n^2(h(X))\right)$ for large n values. Calling $q_{1-\alpha/2}$ the $(1-\frac{\alpha}{2})$-quantile of the normal distribution $\mathcal{N}(0,1)$, we are able to compute

$$\lim_{n\to\infty} \mathbb{P}\left(\mathfrak{I} \in \left[\hat{\mathfrak{I}}_n - q_{1-\alpha/2}\frac{1}{\sqrt{n}}\hat\sigma_n(h(X)),\, \hat{\mathfrak{I}}_n + q_{1-\alpha/2}\frac{1}{\sqrt{n}}\hat\sigma_n(h(X))\right]\right) = (1-\alpha)\%$$

thus providing the $(1-\alpha)$ *asymptotic confidence interval* for $\mathfrak{I}$.

Remark : Be careful not to mix up the following quantities :

(i) the variance of X, estimated by $\hat\sigma_n^2(X) = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2$;

(ii) the variance of $h(X)$, estimated by $\hat\sigma_n^2(h(X)) = \frac{1}{n}\sum_{i=1}^n (h(X_i) - \hat{\mathfrak{I}}_n)^2$;

(iii) the variance of $\hat{\mathfrak{I}}_n$, estimated by $\dfrac{\hat\sigma_n^2(h(X))}{n}$.

Exercise 1 : Application of the Monte Carlo method
Let us consider a random variable $X \sim \mathcal{G}amma(a,b)$ with probability density

$$f_{a,b}(x) = \frac{b^a x^{a-1}}{\Gamma(a)} \exp(-bx)\mathbb{I}_{x>0}$$

and set $a = 4$ et $b = 1$.

1. Simulate a sample of $n = 1000$ realization of X.
2. Compute an estimation of the expected value and variance of X by Monte Carlo method, then give an estimation of the variance of the expected value.
3. Compute the approximated values of the distribution function $F_X(x)$ in $x = 2$ and $x = 5$ by means of a simulation method.
4. Give the approximated values of the 85%, 90% and 95% quantiles of the law of X.

Exercise 2 : Application of the Monte Carlo method (2)
Let us consider consider a random variable X whose probability density is *proportional* to the following function :

$$(2 + \sin^2(x)) \exp\left(-\left(2 + \cos^3(3x) + \sin^3(2x)\right) x\right) \mathbf{1}_{\mathbb{R}^+}(x).$$

1. Verify that $\cos^3(3x) + \sin^3(2x) > -\frac{7}{4}$ for all $x \in [0, 2\pi]$, and build an algorithm to generate the realizations of X .
2. Compute an estimation of the expected value and of the variance of X by a simulation method.
3. Compute the approximated value of the distribution function $F_X(x)$ of X for $x \in (0.5, 1, 1.5, 5, 10, 15)$ and an approximation of the 85%, 90% et 95% quantiles of the law of X .

2 Monte Carlo Intergration with Importance Sampling

Importance Sampling : Bases

Again, we want to approximate $\mathfrak{I} = \int h(x)f(x)dx$. We introduce the following alternative representation :

$$\mathfrak{I} = \int h(x)f(x)dx = \int h(x)\frac{f(x)}{g(x)}g(x)dx$$

where g is such that $\int \left| h(x)\frac{f(x)}{g(x)} \right| g(x)dx < \infty$.

Consequence : If $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} g$, following the law of large numbers

$$\hat{\mathfrak{I}}_n = \frac{1}{n}\sum_{i=1}^{n} h(Y_i)\frac{f(Y_i)}{g(Y_i)} \longrightarrow \mathfrak{I} \quad \text{a.s.}.$$

In practice : Simulate an n-sample $Y_1, \ldots, Y_n \sim g$ and approximate $\mathfrak{I}$ by $\frac{1}{n}\sum_{i=1}^{n} h(Y_i)\frac{f(Y_i)}{g(Y_i)}$.

Advantages :
- It works for all g such that $\text{supp}(g) \supset \text{supp}(f)$.
- Simple laws g can be chosen.
- Possible improvement of the variance of the estimator of $\mathfrak{I}$.
- Simulations $\{Y_i\}_{i=1,\ldots,N} \sim g$ can be recycled for other densities f.

Exercise 3 : Application of the Importance Sampling

We want to evaluate the integral

$$I = \int_2^{\infty} \frac{1}{\pi(1+x^2)} dx\,.$$

1. Analytically calculate the value of I.
2. By a direct simulation method, compute an approximation of I, $\widehat{I}_{1,n}$, by means of a sample of n simulations. Give the corresponding 95% confidence interval for I.
3. Show that $I = \frac{1}{2} - \int_0^2 \frac{1}{\pi(1+x^2)} dx$ and propose a new approximation of I, $\widehat{I}_{2,n}$. Give the corresponding 95% confidence interval for I.
4. Show that $I = \int_0^{1/2} \frac{y^{-2}}{\pi(1+y^{-2})} dy$ and propose a new approximation of I, $\widehat{I}_{3,n}$. Give the corresponding 95% confidence interval for I.
5. Plot $\widehat{I}_{1,n}$ as a function of n (n varying between 1 et 10000). Add the curves corresponding to $\widehat{I}_{2,n}$ and $\widehat{I}_{3,n}$ as functions of n and the line corresponding to I.

Exercise 4 : Application of the Importance Sampling (2)

Let us consider a random variable X, whose probability density is proportional to the following function :

$$(2+\sin^2(x)) \exp\left(-\left(3+\cos^3(3x)\right) x\right) \mathbf{1}_{\mathbb{R}^+}(x).$$

The density of X is only known up to multiplicative factor. Compute by a simulation method an approximated value of this factor.

Université Paris Dauphine — Année 2012-2013
U.F.R. Mathématiques de la Décision
Stat. Exploratoire

Exercise Sheet # 3

The Distribution Function

1 Definition of the empirical distribution function

Empirical distribution function : bases

Definition : Let $(X_1, X_2, ..., X_n)$ be an n-sample of independent and identically distributed random variables of distribution function F. Without any further hypothesis on F, this function can be estimated at every point t by menas of the empirical distribution function $\widehat{F_n}$:

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

Remark : $\widehat{F_n}(t)$ is a nonparametric, unbiased estimator of $F(t)$. $\widehat{F_n}(t)$ is a step function, i.e. :

$$\hat{F}_n(t) = \begin{cases} 0 & \text{if } t < X_{(1)} \\ \frac{1}{n} & \text{if } X_{(1)} \leq t < X_{(2)} \\ \vdots & \\ \frac{i}{n} & \text{if } X_{(i)} \leq t < X_{(i+1)} \\ \vdots & \\ 1 & \text{if } t \geq X_{(n)} \end{cases}$$

where $(X_{(1)}, X_{(2)}, \ldots, X_{(n)})$ corresponds to the set of values of X sorted in ascending order.

In Exercises 1 to 3, we will consider $\boldsymbol{X}$ as an n-sample of law $\mathcal{N}(0,1)$.

Exercise 1 : Computation and graphical representation of $\widehat{F_n}$

1. Write a function that computes $\widehat{F_n}(t)$ starting from an n-sample $\boldsymbol{X}$.
2. Plot $\widehat{F_n}$ (take $n = 100$) together with the curve representing F.

Homeworks : The same exercise with a sample drawn from $\mathcal{E}(1)$.

2 Asymptotic behavior of the empirical distribution function

Reminders : Consequences of the Strong Law of Large Numbers and the Central Limit Theorem

Strong Law of Large Numbers (SLLN) : At each point t, $\widehat{F_n}(t)$ is the proportion of observations smaller than t, *i.e.* an estimator of $P(X \leq t)$. As a consequence of SLLN, we have that

$$\forall\ t,\ \widehat{F_n}(t) \xrightarrow{ps} F(t), n \to \infty$$

Central Limit Theorem (CLT) : We notice that $\mathbb{I}_{\{X \leq t\}} \sim \mathcal{B}er(F(t))$ and we easily obtain that $\mathbb{V}(\widehat{F_n}(t)) = \frac{1}{n}F(t)(1-F(t))$. Following the CLT, we get :

$$\sqrt{n}(\widehat{F_n}(t) - F(t)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, F(t)(1-F(t))),\ n \to \infty$$

Consequence : The application of the CLT and the SLLN yields a confidence interval for $F(t)$. Let $q_{1-\alpha/2}$ be the $(1-\alpha/2)$-quantile of $\mathcal{N}(0,1)$. Then

$$\lim_{n\to\infty} \mathbb{P}\left(F(t) \in \left[\widehat{F_n}(t) \pm q_{1-\alpha/2}\frac{1}{\sqrt{n}}\sqrt{\widehat{F_n}(t)(1-\widehat{F_n}(t))} \right] \right) = (1-\alpha)\%\,.$$

Exercise 2 : Verification of the LLN and of the CLT

1. LLN : Study the convergence of $\widehat{F_n}$ towards F. (Use the values $30, 50, 100, 500$ for n.)
2. CLT : Graphically check that $\sqrt{n}(\widehat{F_n}(t) - F(t)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, F(t)(1-F(t)))$. (Use the same value for n as in the previous point and $t = 0$.) Give the corresponding $CI_{.95}$.

Homeworks The same exercise with a sample drawn from $\mathcal{E}(1)$ and $t = 2$.

3 Precision of the distribution function estimator

Precision of the estimator : Bases

Definition : For a confidence interval of the form

$$CI_{1-\alpha}(I) = \left(\widehat{I} \pm q_{1-\alpha/2}\sigma_n(\widehat{I})\right),$$

with the value of α fixed, the precision corresponds to the interval length, that is

$$p = 2q_{1-\alpha/2}\sigma_n(\widehat{I}).$$

Consequence : By estimating $\sigma_n(\widehat{I})$ as a function of n, we are able to find the sample size needed to obtain a given precision at a fixed α value.

Example of approximation : For a law which is symmetric in a, $F(a) = 1/2$, thus $F(t)(1 - F(t)) \approx \frac{1}{4}$ for t close to a. As a consequence, for t close to a, the variance of $\widehat{F_n}(t)$, can be approximated by $\frac{1}{4n}$ ($\sigma_n(\widehat{I}) \approx \frac{1}{2\sqrt{n}}$).

Exercise 3 : Determination of the sample size needed to obtain a given precision

1. Always using a sample $\boldsymbol{X}$ drawn from a normal law $\mathcal{N}(0,1)$, compute an approximation of the variance of $\widehat{F_n}(t)$ when $t \approx 0$.
2. Derive the sample size $n^\star$ needed to obtain a precision of $10e-3$ for $\alpha = 95\%$. Check by a Monte Carlo experiment that the obtained precision is sufficient.

Homeworks : Build a $CI_{.95}$ for $F(t)$ in $t = 2$ based on a sample of size $n^\star$. Is the precision of this estimator smaller or larger than $10e-3$? Why ?

Hint : Look at the variation of the function $g(x) = x(1-x)$ on the interval $[0,1]$.

4 Generation of an m-sample of distribution function $\widehat{F_n}$

Generation of an m-sample of distribution function $\widehat{F_n}$

Let $X_1, \cdots, X_n$ be an n-sample of distribution function F^X. We are able to compute the empirical distribution function $\widehat{F_n^X}$ from this sample. This distribution function $\widehat{F_n^X}$ defines a new probability law whose support consists of $\{X_1, \cdots, X_n\}$ only.

Aim : We want to generate an m-sample $Y_1, \cdots, Y_m$ of distribution function $F^Y = \widehat{F_n^X}$.

Consequence : $Y_i, i = 1, \cdots, m$ may only take the values in $\{X_1, \cdots, X_n\}$ and

$$\mathbb{P}(Y_i = X_k) = \frac{1}{n} \quad \forall k = 1 \dots n, \forall i = 1 \dots m\,.$$

In practice : we sample with replacement from the equiprobable sample $X_1, \cdots, X_n$ using the `R` function `sample`(...).

Beware : The distribution function of Y is $\widehat{F_n^X}$, its empirical distribution function is $\widehat{F_m^Y} = \frac{1}{m}\sum_{i=1}^{m} \mathbb{I}_{\{Y_i \le t\}}$ and we have that

$$\forall\ t,\ \widehat{F_m^Y}(t) \xrightarrow{as} F_n^X(t), m \to \infty\,.$$

Exercise 4 : Re-sampling from a known empirical distribution function

Let X be an n-sample of $\mathcal{N}(0,1)$ with $n = 30$.

1. Plot the empirical distribution function of X, $\widehat{F_n^X}$.
2. Simulate a sample $Y_1, \cdots, Y_m$ from the distribution function $\widehat{F_n^X}$.
3. Plot the empirical distribution function of Y, $\widehat{F_m^Y}$.
4. Graphically check that $\widehat{F_m^Y}$ gets closer to $\widehat{F_n^X}$ when m becomes large (use the values $30, 50, 100, 500$ for m).

Université Paris Dauphine Année 2012-2013
U.F.R. Mathématiques de la Décision
Stat. Exploratoire

Exercise Sheet #4

Bootstrap

The objective of this TP is to present the Bootstrap re-sampling method. This method allows, when classical statistical methods are not available, to solve usual inferential problelms (bias, variance, mean square error of an estimator, confidence intervals, hypothesis testing, ..) .

Bootstrap is an inferential technique based on a succession of re-samplings. Let X be a real random variable with cumulative distribution function F unknown : $F(x) = P(X \leq x)$. Let $(X_1, ..., X_n)$ be a sample from the law of X and F_n the associated empirical distribution function : $F_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}_{X_i \leq x}$.

We're interested in θ, a parameter of the law of X. θ can be written as a functional of F as $\theta = t(F)$. A natural estimator for $\theta = t(F)$ is than given by $\widehat{\theta} = t(\widehat{F}_n) = T(X_1, \ldots, X_n)$.

Exemples :

1. If $\theta = E[h(X)] = \int h(x)dF(x)$, where h is a function from $\mathbb{R}$ in $\mathbb{R}$,
   $\hat{\theta} = \int h(x)d\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n h(X_i)$
2. If $\theta = Var[h(X)] = E[(h(X) - E[h(X)])^2] = \int h(x)^2 dF(x) - \left(\int h(x)dF(x)\right)^2$ where h is a function from $\mathbb{R}$ in $\mathbb{R}$,
   $\hat{\theta} = \int h(x)^2 d\widehat{F}_n(x) - \left(\int h(x)d\widehat{F}_n(x)\right)^2 = \frac{1}{n}\sum_{i=1}^n h(X_i)^2 - \left[\frac{1}{n}\sum_{i=1}^n h(X_i)\right]^2$
3. If θ is the median value of X, $\hat{\theta} = (X_{(n/2)} + X_{(n/2+1)})/2$ if n is odd, $X_{(n+1)/2}$ if n is even, where $(X_{(1)}, ..., X_{(n)})$ is the order statistic associated with $X_1, ..., X_n$ (i.e. the increasingly ordered sample) .
4. If θ is the quantile of level $1 - \alpha$ of the law of X ,$\hat{\theta} = X_{([(1-\alpha)n]+1)}$.

We'd like to estimate bias, variance or mean square error of a given estimator $\hat{\theta} = T(X_1, ..., X_n)$, obtain confidence interval on θ, etc...

1 Estimation of the bias of $\hat{\theta} = T(X_1, ..., X_n)$.

We call bias of $\hat{\theta}$ the quantity $E[\hat{\theta}] - \theta$. This bias is in general unknown because it depends on F, that is unknown too. We'd like to etimate it based on just one observation $(X_1, \ldots X_n)$. There are several possible cases :
For a start we suppose that F **and** θ **are known** (this is just an academic exercise because if this was true, we wouldn't need to estimate θ !). So we could :

- compute $E[\hat{\theta}] - \theta$ analytically and the problem is solved,
- if we can't compute $E[\hat{\theta}] - \theta$ analytically we could resort to a Monte Carlo technique. More precisely,
  1. We simulate B n-samples $(X_1^l, ..., X_n^l)$ with distribution F.
  2. For each sample, we compute $\hat{\theta}^l = T(X_1^l, ..., X_n^l)$.
  3. In the end, we obtain an estimate of the bias $E[\hat{\theta}] - \theta$ with : $\frac{1}{B}\sum_{l=1}^{B} \hat{\theta}^l - \theta$

But, in a realistic setting, both F and θ are unknown and the previous procedure is then unavailable.

The Bootstrap method consist in replacing in the previous Monte Carlo scheme F with $\widehat{F}_n$ and θ with $\hat{\theta}$.

Remember : In sheet # 3, we learned how to simulate an n-sample according to $\widehat{F}_n$: it amounts to randomly sample with replacement n variables from the observations $X_1, \dots X_n$

Finally we can write the Bootsrap procedure to estimate the bias as :

Bootsrap procedure to estimate the bias

1. Compute $\hat{\theta}$ from the sample $X_1 \dots X_n$.
2. For $l = 1 \dots B$,
   (a) Simulate an n-sample $(X_1^{*l}, ..., X_n^{*l})$ according to $\widehat{F}_n$ i.e. extract a random sample with replacement of length n from the observations $(X_1, ..., X_n)$: `sample(X,n,replace=TRUE)`.
   (b) For each new sample, compute $\hat{\theta}^{*l} = T(X_1^{*l}, ..., X_n^{*l})$
3. We obtain an estimation of the bias $E[\hat{\theta}] - \theta$ with : $\frac{1}{B}\sum_{l=1}^{B} \hat{\theta}^{*l} - \hat{\theta}$

Exercise 1

We're interested in $\widehat{\sigma_n^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$, the natural estimator for the variance $V(X) = \sigma^2$.

1. Suppose that X follow the distribution $\mathcal{N}(0, 1)$.
   (a) Simulate a 100 - sample $(X_1, \dots, X_{100})$ and save it into the vector XX.
   (b) Compute analytically the bias $b = E[\widehat{\sigma_n^2}] - \sigma^2$.
   (c) Evaluate this bias by the Monte Carlo method. Design a graph that explain how this appoximation change as a function of the iteration number B. Add to this graph an horizontal line of ordinate b.
2. Now we still take $(X_1, \dots, X_{100})$, stocked in the XX vector, but we suppose that the observations have an unknown distribution. Estimate the bias with a Bootstrap procedure. As before, design a graph that explains how this appoximation change as a function of the iteration number B.
3. Compare the three methods.

Remarque : The same procedure can be utilized to estimate the variance, the mean square error of an estimator, ...

2 Bootstrap confidence intervals

Let $\hat{\theta}$ be an estimator of θ . We'd like to find a confidence interval for θ i.e we're searching $q_1(\hat{\theta})$ and $q_2(\hat{\theta})$ such that $\mathcal{P}\left[q_1(\hat{\theta}) \leq \theta \leq q_2(\hat{\theta})\right] = 1 - \alpha$ (α fixed).

Remark on confidence intervals (see the L2 stat course) :

1. Let $(X_1, \ldots, X_n)$ a sample <u>from a normal distribution</u> $\mathcal{N}(\theta, \sigma^2)$, then $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}_n$ is the natural estimator of θ .
   (a) If σ^2 is known, let Φ be the cumulative distribution function of a standard normal law and q such that $\Phi(q) = 0.975$. Than $I_c = [\hat{\theta} - q\frac{\sigma}{\sqrt{n}}, \hat{\theta} + q\frac{\sigma}{\sqrt{n}}]$ is an exact 95%-level confidence interval for θ.
   (b) If σ^2 is unknown, under the hypothesis of normality for the observations, let $\widehat{S}_n = \frac{1}{n-1}\sum_{i-1}^{n}(X_i - \overline{X}_n)^2$ be an estimator of the variance. Than $\frac{\overline{X}_n - \theta}{\sqrt{\widehat{S}_n/n}} \sim \mathcal{T}(n-1)$ Thus, let Φ_T be the cumulative distribution function of a Student's T with $n-1$ degrees of freedom and q such that $\Phi_T(q) = 0.975$ than $I_c = \left[\overline{X}_n - q\sqrt{\widehat{S}_n/n}, \overline{X}_n + q\sqrt{\widehat{S}_n/n},\right]$ is a 95%-level confidence interval for θ.
2. Now, suppose that $(X_1, \ldots, X_n)$ is an n-sample of X with <u>unknow distribution</u> such that $E[X] = \theta$. As before, $\hat{\theta} = \overline{X}_n$ is a natural estimator for θ.
   (a) If σ^2 is known, the Central-Limit theorem enables us able to give an <u>asymptotic</u> confidence interval.
   (b) If σ^2 is unknown, let $\widehat{S}_n$ be a consistent estimator of σ^2. Thanks to Slutsky's Lemma and the Central-Limit theorem, we have convergence for $\frac{\overline{X}_n - \theta}{\sqrt{\widehat{S}_n/n}}$ towards a standard normal law. We can easly than obtain <u>asymptotic</u> confidence intervals.

Problems :

- The previous results are applicable only when the Central-Limit theorem applies.
- These results are asymptotic and thus they are reliable only in the presence of a large number of observations.
- We need a consistent estimator of the variance for them to apply.

The <u>percentiles Bootstrap</u> procedure allows us to overcome these problems. Its principle is to approximate the distribution function of the estimator $\hat{\theta} = T(X_1, ..., X_n)$ with its empirical distribution function obtained with a Bootstrap sample. The bounds of the confidence interval are then obtained from this empirical distribution function.

Bootstrap Percentiles procedure

Let $(X_1, \ldots, X_n)$ be an observed n-sample.

1. For $l = 1 \ldots B$,
   (a) Simulate an n-sample $(X_1^{*l}, ..., X_n^{*l})$ with distribution $\widehat{F}_n$ i.e. extract a random sample with replacement of length n from the observations $(X_1, ..., X_n)$:

   `sample(X,n,replace=TRUE).`

   (b) Compute $\hat{\theta}^{*l} = T(X_1^{*l}, ..., X_n^{*l})$
2. The sample $(\hat{\theta}^{*l})_{l=1...B}$ leads to an approximation of the distribution function of $\hat{\theta}$. Compute q_1 and q_2 using the function `quantile`.

Exercise 2

We're interested in the expected value of μ of a normal distributed sample.

1. Simulate an n-sample $(X_1, ..., X_n)$ for $n = 10$ from a normal distribution $\mathcal{N}(5, 2)$. Suppose μ and σ^2 unknown.
2. Compute the 95%-level confidence interval for μ using Student's t distribution.
3. Compute the 95%-level asymptotic confidence interval for μ using the Central limit theorem.
4. Compute the 95%-level asymptotic confidence interval for μ using the Bootstrap Percentiles procedure.
5. Compare the three methods
6. Take a larger n and repeat the exercise.

---

3 Hypothesis testing with the Bootstrap

Problems

In a number of practical problems, we modelize the relation between two quantities Y and X. Suppose that we dispose of n values of X fixed, denoted x_i, and that for each x_i we observe a realization of a random variable Y, denoted y_i. The simple linear regression model consists in :

$$Y_i = \alpha + \beta x_i + E_i$$

where E_i are random variables i.i.d. with null expected value and variance σ^2. We can then estimate the unknown parameters α and β thanks to the least squares criterion. This consist in finding the quantities $\hat{\alpha}$ and $\hat{\beta}$ minimizing :

$$\sum_{i=1}^{n} (Y_i - \alpha - \beta x_i)^2.$$

Let $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$, $\overline{Y} = \sum_{i=1}^{n} Y_i$, $S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2$ and $S_{YY} = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$, $S_{xY} = \sum_{i=1}^{n} (x_i - \overline{x})(Y_i - \overline{Y})$. The least square estimator of α and β can be written as :

$$\hat{\beta} = \frac{S_{xY}}{S_{xx}} \quad \hat{\alpha} = \overline{Y} - \hat{\beta}\overline{x}$$

1. For $n = 30$, $\sigma^2 = 0.5$, $\alpha = 2$, $\beta = 1$ and `x = seq(0, 10, length = n)` , generate some realization of Y_i following the simple linear model in the case where the residuals E_i follow the distribution $\gamma\mathcal{T}(5)$ where γ is a constant that we need to compute.
2. Determine a Boostrap estimation of α and β
3. Determine, with a Boostrap procedure, a 95%-level confidence interval for the parameters α and β.
4. Suppose

$$T = \sqrt{(n-2)S_{xx}}\,\frac{\hat{\beta}-1}{\sum_{i=1}^{n}(Y_I - \hat{\alpha} - \hat{\beta}x_i)2}$$

On the previously generated sample, we want to test the null hypothesis H_0 that $\beta = 1$ versus the alternative hypothesis H_1 that $\beta = 1.5$. We propose to use the decision rule where we reject H_0 if $T > F^{-1}_{St(28)}(0.95)$, optimal strategy when the residuals are Gaussian. Determine, with the Bootstrap method, the error ratio of the previous test (i.e. probability of rejecting H_0 while it's true instead, equal to 5% in the Gaussian case).

Université Paris Dauphine
UFR de mathématiques de la Décision / MD3
Mat 6-3- NOISE

Année 2012-2013

Exercise Sheet #6

Kolmogorov-Smirnov's Test.

In this sheet, we will make use of the faithful data, included in R.

Preliminary study of the data

1. Download the dataset and make a summary analysis (type of data, sample size, quantities under study, means, variances ... etc)

   Use `data(faithful)`, `help(faithful)`, `summary(faithful)`

2. Design a graphic rapresentation describing roughly the distribution of the data in faithful

3. Estimate with an uniform Kernel the density of the faithful data.

4. Estimate with a Gaussian Kernel the density of the faithful data.

5. Study the influence of the window's width on the Gaussian Kenrnel density estimation.

1 Introduction to the Kolmogorov-Smirnov's test

1.1 Test of adequacy to a given law

Let $(X_1, \ldots, X_n)$ be an n-sample from an unknown distribution P. Let P_0 be a known distribution, that is fixed. We try to test the hypothesis:

$\mathcal{H}_0$: "the data $X_1, \ldots, X_n$ are distributed according to the distribution P_0"

versus

$\mathcal{H}_1$: "the data $X_1, \ldots, X_n$ are not distributed according to the distribution P_0"

Principle of the test : The Kolmogorov-Smirnov's test can answer this problem. The idea is that if the hypothesis H0 is correct, then the empirical distribution function $\widehat{F}_n$ of the sample should be close to F_0, the distribution function corresponding to P_0.

Test Statistic: We measure the adequacy of the empirical distribution function of the function F_0 with the Kolmogorov-Smirnov's distance, which is the uniform norm between the distribution functions. To compute that, simply evaluate the difference between $\widehat{F}_n$ and F_0 at the points $X_{(i)}$.

$$D_{KS}(F_0, \widehat{F}_n) = \max_{i=1,\ldots,n} \left\{ \left| F_0(X_{(i)}) - \frac{i}{n} \right| , \left| F_0(X_{(i)}) - \frac{i-1}{n} \right| \right\} .$$

Construction of the test: We're going to reject $\mathcal{H}_0$ if the distance between $\widehat{F}_n$ and F_0 is big, i.e. if $D_{KS}(F_0, \widehat{F}_n)$ exceeds a certain threshold q_α still to be defined.

About the threshold: We'll choose the threshold q_α such that, if the hypothesis $\mathcal{H}_0$ is true, the probability of rejection for $\mathcal{H}_0$ is small (typically $\alpha = 5\%$)

$$\mathbb{P}_{X_i \sim F_0}\left(D_{KS}(F_0, \widehat{F}_n) > q_\alpha\right) = \alpha$$

To obtain this threshold, we need to know the distribution of $D_{KS}(F_0, \widehat{F}_n)$ in the case where the X_is are distributed according to F_0. Now we can show that under the assumption $\mathcal{H}_0$, the distribution of the statistic $D_{KS}(F_0, \widehat{F})$ does not depend on F_0. Thus the distribution of $D_{KS}(F_0, \widehat{F})$ has no simple and explicit expression and has to be computed numerically. This distribution has been tabulated.

$\hookrightarrow$ **In R**,

- This adequacy test has been implemented in `ks.test`
- the output of this function is a liste of objects, comprehending the `p-value`. The p-value is the minimum α at which We would have rejected $\mathcal{H}_0$. **If the p-value is inferior to 5% We will reject the hypothesis $\mathcal{H}_0$ at the 5% level**

> **Exercise. Test of adequacy to a given law**
>
> We are interested in the eruption's times exceeding 3 minutes.
>
> 1. Create a vector `long` containing the eruption's times exceeding 3 minutes.
> 2. Test the hypothesis that the observed times of eruption exceeding 3 minutes follow a $\mathcal{N}(4, 0.1)$ distribution.

Remark: The Kolmogorov-Smirnov's test can be extended to a comparison between two empirical distribution functions, and allows to test the hypothesis that two samples come from the same distribution. For this we use a function similar to `ks.test` but with corrected thresholds.

1.2 Test of adequacy to a family of distributions

Let $X_1 \dots X_n$ be an n-sample from an unknow distribution. Let $\mathcal{F}_\theta$ be a parametric family of distributions. For example, $\mathcal{F}_\theta = \left\{\mathcal{N}(\mu, \sigma^2), \theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^{*+}\right\}$. We'll try now to test the hypothesis that: $\mathcal{H}_0$: "The distribution of $X_1, \dots, X_n$ belongs to the family of distributions $\mathcal{F}_\theta$" contre $\mathcal{H}_1$: "The distribution of $X_1, \dots, X_n$ does NOT belong to the family of distributions $\mathcal{F}_\theta$"

Method: Let $\widehat{\theta}$ be the maximum likelihood estimator of the parameter θ. As before, we'll reject $\mathcal{H}_0$ if

$$D_{KS}(F_{\widehat{\theta}}, \widehat{F}_n) > q'_\alpha$$

About the threshold: As before, we compute the threshold in order to minimize the probability to reject $\mathcal{H}_0$ while it's true. So we need again the distribution of the statistic $D_{KS}(F_{\widehat{\theta}}, \widehat{F}_n) > q'_\alpha$.

- *Attention*: As $F_{\widehat{\theta}}$ depends on the sample, the distribution of the statistic is not the same as in the test of adequcy to a given distribution. The function `ks.test` cannot be used in this case.
- The R function which allow us to realize the Kolmogorov-Smirnov's test adjustement to a gaussian family is `lillie.test(x)`, in the package `nortest`.

> **Exercise. Test of adequacy to a family of distributions**
>
> Test the potential normality of the probability distribution of the observed eruption's times exceeding 3 minutes using the function `lillie.test(x)`.