Programmation et Logiciels Statistiques
Exercices TD 4 - R

L3 Statistique & LP SIS
Université de Bretagne-Sud

Exercice 1

La fonction de densité d'un vecteur aléatoire réel X de dimension d, de loi $\mathcal{N}(\mu, V)$ tel que $\det(V) \neq 0$ s'écrit

$$f(x) = \frac{1}{(2\pi)^{d/2}\sqrt{\det(V)}} \exp\left(-\frac{1}{2}(x-\mu)'V^{-1}(x-\mu)\right)$$

Ecrire une fonction R permettant de calculer la valeur de cette densité en n points de $\mathbb{R}^d$.

Exercice 2

Utiliser la fonction mise au point à l'exercice précédent pour représenter graphiquement la densité d'un vecteur aléatoire réel de loi $\mathcal{N}(\mu, V)$ en 3D pour $d = 2$, ainsi que ses courbes de niveau.

Le faire d'abord pour une matrice de variance-covariance diagonale, puis pour une matrice de variance-covariance quelconque.

Exercice 3

Soit (X_n) une suite de variables aléatoires réelles indépendantes et identiquement distribuées (i.i.d.) d'espérance μ et de variance $\sigma^2 > 0$. Définissons, pour tout n, la variable centrée réduite

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

où

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

Alors, pour tout x réel,

$$\mathbb{P}(Z_n \leq x) \longrightarrow \Phi(x)$$

où Φ désigne la fonction de répartition associée à la loi normale standard (Théorème Central Limite).

Illustrer le TCL en étudiant la loi de $\bar{X}_{30}$ lorsque $X_1, \ldots, X_n$ sont des v.a.r. i.i.d. de loi uniforme sur $[0, 1]$.
– Pour cela, simuler indépendamment n échantillons i.i.d. de taille 30 associés à la loi $\mathcal{U}([0, 1])$.
– Calculer la moyenne empirique de chacun des échantillons et les valeurs centrées réduites associées ($\mu = 0.5$ et $\sigma^2 = 1/12$).
– Représenter l'histogramme des valeurs ainsi obtenues.
– Comparer graphiquement cet histogramme à la fonction de densité de la loi normale standard.
– On pourra faire en sorte que l'utilisateur puisse spécifier le nombre d'échantillons à la console. Pour cela, faire en sorte que la simulation et la représentation graphique soient réalisées par une fonction.

Exercice 4

Soit (X_n) une suite de variables aléatoires réelles i.i.d. d'espérance finie μ. Alors,

$$\mathbb{P}(\bar{X}_n \to \mu) = 1$$

(Loi Forte des Grands Nombres).

Illustrer la Loi Forte des Grands Nombres en étudiant la convergence de $\bar{X}_n$ lorsque $X_1, X_2, \ldots, X_{10\,000}$ est un échantillon i.i.d. associé à la loi uniforme sur $[0, 1]$.
– Pour cela, simuler en une fois un échantillon i.i.d. de taille 10 000 associé à la loi $\mathcal{U}([0, 1])$.
– Calculer la moyenne empirique de chacun des sous-échantillons de taille 30 à 10 000 et les comparer graphiquement à la valeur de l'espérance $\mu = 0.5$.

Exercice 5

Soit (X_n) une suite de variables aléatoires réelles i.i.d. d'espérance finie μ. Alors, pour tout $\varepsilon > 0$,

$$\mathbb{P}(|\bar{X}_n - \mu| \leq \varepsilon) \longrightarrow 1$$

(Loi Faible des Grands Nombres).

Illustrer la Loi Faible des Grands Nombres en étudiant la variabilité de $\bar{X}_{n,n}$ lorsque $X_{1,n}, \ldots, X_{n,n}$ sont des échantillons i.i.d. indépendants associés à la loi uniforme sur $[0, 1]$, pour n variant de 30 à 10 000.
– Pour cela, simuler indépendamment pour n variant de 30 à 10 000, des échantillons i.i.d. de taille n associés à la loi $\mathcal{U}([0, 1])$.
– Calculer la moyenne empirique de chacun des échantillons et comparer les valeurs obtenues à la valeur de l'espérance $\mu = 0.5$.

Exercice 6

Une fonction *récursive* est une fonction qui s'appelle elle-même. Afin de résoudre un problème de type X à l'aide d'une fonction récursive $f()$, l'idée est en gros la suivante :
1 – Ramener le problème initial de type X à un ou plusieurs problèmes plus petits de type X.
2 – Dans $f()$, appeler $f()$ pour chacun de ces problèmes plus petits.
3 – Dans $f()$, rassembler les résultats du 2 pour résoudre le problème initial.

L'exemple classique est la fonction factorielle, qui peut être implémentée récursivement de la façon suivante :

```
fact <- function(x) {
if (x ≤ 1) 1 else x * fact(x − 1)
}
```

L'exemple auquel nous allons nous intéresser dans cet exercice est Quicksort, un algorithme permettant de trier un vecteur de nombres du plus petit au plus grand. Par exemple, supposons que nous souhaitions trier le vecteur $(5, 4, 12, 13, 3, 8, 88)$. On commence par comparer tous les élements au premier élément 5, ce qui donne deux sous-vecteurs : celui composé des éléments strictement inférieurs à 5 et celui composé des éléments supérieurs ou égaux à 5. Plus précisément, on obtient les sous-vecteurs $(4, 3)$ et

$(12, 13, 8, 88)$. On peut alors appeler la fonction sur les sous-vecteurs, ce qui donne, si tout fonctionne bien, $(3, 4)$ et $(8, 12, 13, 88)$. On les recolle ensuite avec 5, ce qui donne $(3, 4, 5, 8, 12, 13, 88)$, comme souhaité.

Ecrire une fonction récursive sous R qui implémente Quicksort.

Exercice 7

On observe une suite de données de la forme 0 ou 1. Il peut s'agir par exemple du temps qu'il fait : on note 1 s'il a plu au cours de la journée et 0 sinon. Supposons qu'on souhaite prévoir le temps qu'il fera demain : pluvieux ou non-pluvieux à partir des données observées jusque là. De façon plus précise, pour un entier $k \geq 1$ donné, on va prévoir le temps qu'il fera demain à partir des données des k derniers jours. On utilisera un vote à la majorité : si le nombre de 1 au cours des k derniers jours est supérieur ou égal à $k/2$, on prévoit qu'il pleuvra demain, sinon qu'il ne pleuvra pas. Par exemple, si $k = 3$ et qu'on a observé au cours des trois derniers jours 1, 0, 1, notre prévision sera 1 (pluie).

Le problème qui se pose alors est le suivant : comment choisir k ? Si k est trop petit, on aura trop peu de données pour prévoir correctement; s'il est trop grand, on prendra en compte des données trop éloignées dans le temps pour qu'elles aient un quelconque pouvoir prédictif.

La solution usuelle à ce problème consiste à se baser sur des données existantes, appelées *échantillon d'apprentissage*, et à voir ce que donnent diverses valeurs de k sur ces données.

Dans le cas de la météo, supposons qu'on dispose de 500 jours de données et qu'on envisage de prendre $k = 3$. Afin d'évaluer le pouvoir prédictif de cette valeur de k, on "prévoit" chaque jour des données à partir des trois jours précédents puis on compare ces prévisions avec les valeurs effectives. Après avoir fait cela pour toutes les données, on dispose d'un taux d'erreur pour $k = 3$. On répète l'opération pour $k = 1$, $k = 2$, $k = 4$, etc, jusqu'à une valeur maximale de k qu'on se fixe a priori. On utilise alors la valeur de k pour laquelle le taux d'erreur trouvé est le plus bas pour faire de nouvelles prévisions.

Ecrire un script R permettant d'implémenter cette méthode de prévision.