TP0 – Probabilités avancées

Mines Saint-Etienne – Majeure Science des Données 2020/2021

La première partie de ce TP est l'occasion de revoir quelques commandes et fonctions de base du logiciel libre R, outil très largement utilisé par les « data scientists » (avec Python). Dans une deuxième partie, on étudie par simulation la loi normale bidimensionnelle associée à un couple de variables aléatoires réelles (X, Y) ou vecteur gaussien bidimensionnel.

Environnement RStudio

Créer un répertoire de travail **…/TP0** dans lequel vous déposerez le script **ScriptTP0** fourni sur Campus. Lancer **RStudio** à partir du menu de votre PC. Se placer ensuite dans le répertoire de travail à partir du menu **Session / Set Working Directory / Choose Directory** ou, directement, en exécutant dans la console la commande R : `setwd("…/TP0")`.

Utilisation du ScriptTP0

On rappelle qu'un script est un fichier texte d'extension .**R** qui va contenir les commandes à faire exécuter dans la console. Ouvrir le fichier **ScriptTP0** avec le menu **File / Open File…** de **RStudio**. Pour exécuter la ligne courante ou une sélection de lignes de commandes R de ce script, utiliser le bouton **Run** ou le raccourci clavier (**Ctrl+Entrée)**. Attention, les calculs sont seulement effectués dans la console et R ne connaît pas les variables déclarées dans un script qui n'a jamais été exécuté.

1. Exécuter le code jusqu'à la ligne 21 en cherchant à anticiper le résultat retourné (le cas échéant) Que représente la bande en pointillés ?

2. Obtenir le graphique montrant l'histogramme normalisé des données et la densité réelle superposée : bloc de lignes 27 à 37. Jouer sur le nombre de tirages et l'option `breaks` de la fonction `hist()`.

3. Retrouver les valeurs de quelques quantiles de la loi normale (lignes 42 à 50).

4. Comparer avec les quantiles estimés à partir de données simulées (lignes 55 à 67)

5. Obtenir les fameux diagrammes quantiles-quantiles ou qq-plots (lignes 74 à 89) permettant de tester ici l'adéquation à la loi normale. Jouer sur les valeurs des différents paramètres.

6. Expliquer la forme du tracé (en $\pm\infty$) dans le cas de données issues d'une loi exponentielle (lignes 92 et 93).

Cette première partie a permis de revoir un certain nombre de fonctions **R** comme `c()`, `seq()`, `plot()`, `lines()`, `abline()`, `rnorm()`, `dnorm()`, `qnorm()`, `quantile()`, `pnorm()`, `hist()`, `qqplot()`, `qqline()` qui sont souvent utilisées en pratique.

Voir le fichier **aide_memoire_R.pdf** sur Campus pour une liste plus complète.

Etude de la loi normale bidimensionnelle

7. Soient X et Y indépendantes de même loi normale N(0, 1). La densité f du vecteur aléatoire (X, Y) est de la forme $(x, y) \rightarrow f(x, y) = \frac{1}{2\pi} \exp(-\frac{x^2+y^2}{2})$. Le graphe ou surface associée est la fameuse cloche de Gauss (bidimensionnelle) et les lignes de niveau ou courbes d'iso-probabilité sont des cercles centrés en O dans le plan Oxy. Exécuter les lignes 100 à 118 pour visualiser un jeu de données simulées selon cette loi ainsi que quelques lignes de niveau.

8. Visualiser ce qui se passe si l'on multiplie X par un scalaire ainsi que Y en considérant par exemple le couple (3X, Y). Que deviennent les cercles ? Comment s'écrit la densité ?

9. On considère en plus de la transformation précédente (multiplication par des scalaires) l'action d'une rotation d'angle θ :

$X = s_X \cos(\theta)\, U - s_Y \sin(\theta)\, V$
$Y = s_X \sin(\theta)\, U + s_Y \cos(\theta)\, V$

où les variables U, V sont **indépendantes de même loi normale N(0, 1)** et s_X , s_Y deux scalaires > 0.

Visualiser à nouveau l'effet d'une telle transformation en visualisant un jeu de données simulées.

10. On considère enfin une transformation linéaire générale de la forme :

X = a U + b V
Y = c U + d V

Faire le même travail. A quelle conclusion arrivez-vous ?