TP1 de Probabilités: GIS3

Benjamin Arras *

1 Premier pas avec R et RStudio

Un descriptif détaillé du logiciel R ainsi que son téléchargement sont disponibles sur le site `www.r-project.org`. C'est un logiciel de statistiques libre.

Pour ouvrir une session R, ouvrir un terminal et tapez la commande "R". Vous obtenez alors le prompteur de R ">". Vous venez d'ouvrir une session R à laquelle un environnement est associé. Dans cet environnement, vos variables et objets sont reconnus par vos programmes (scripts R). Dans le répertoire de travail, sont stockés vos sorties, comme des graphiques ou des fichiers texte. Pour connaître le répertoire de travail, tapez la commande "getwd()", et pour en changer tapez "setwd()".

Le logiciel R fonctionne initialement en ligne de commande mais des interfaces permettent désormais une utilisation plus conviviale. Nous proposons ici de travailler avec l'interface RStudio que l'on ouvre en tapant "usr/local/RStudio/bin/RStudio".

Pour obtenir de l'aide sur une commande ou une fonction R, il suffit de taper "help(nomfonction)". Pour quitter une session R, tapez "q()".

Le logiciel R comporte des bibliothèques de fonctions appelées packages qui ont été développées par les scientifiques du monde entier et mises à disposition des utilisateurs dans des entrepôts : choisissez l'un d'eux (par exemple le CRAN de Toulouse depuis le site `www.r-project.org`) pour accéder aux packages disponibles. Pour qu'un package soit utilisable, il doit d'abord être installé et ensuite chargé dans votre session. Pour cela, utilisez les commandes "install.packages" et "require". Pour voir les packages déjà installés sur votre machine, tapez "installed.packages()". Sous RStudio, on peut aussi installer les packages via Tool, Install Packages... Celui qui nous intéresse pour ce TP est le package "stat" : il est installé par défaut et également chargé automatiquement par R dans votre environnement.

2 Les objets de base

Les objets de base du logiciel sont :

1. Les vecteurs et les matrices
2. Les data-frames

*Université de Lille; benjamin.arras@univ-lille.fr

3. Les listes

Le contenu des objets peut être de différents types :

1. numérique
2. chaîne de caractères
3. booléen

Au cours de ces séances de travaux pratiques, nous nous intéresserons essentiellement à des vecteurs et des matrices numériques ainsi qu'aux objets de type booléens.
La fonction concaténant des nombres ou des vecteurs pour obtenir des vecteurs est "c()"

$$x \leftarrow c(-7, 84, 19) \quad y \leftarrow c(x, 1, 3) \quad z \leftarrow c(1:10, x, y)$$

Il existe également des commandes pour créer des matrices avec des structures particulières. Par exemple, la fonction "diag()" permet de construire des matrices diagonales. Testez la sur les exemples suivants

$$M \leftarrow \mathrm{diag}(1:8) \quad N \leftarrow \mathrm{diag}(\mathrm{rep}(1, 10))$$

En général, les matrices seront créés à partir de vecteurs à l'aide de la commande "matrix()" en fixant le nombre de colonnes et/ou le nombre de lignes. Par défaut, la matrice est remplie colonne par colonne. Pour la remplir ligne par ligne, il faut ajouter la commande byrow=TRUE.

$$M \leftarrow \mathrm{matrix}(1:8, \mathrm{ncol} = 4) \quad N \leftarrow \mathrm{matrix}(1:8, \mathrm{nrow} = 4)$$
$$O \leftarrow \mathrm{matrix}(1:8, \mathrm{nrow} = 4, \mathrm{byrow} = \mathrm{TRUE})$$

On peut assembler horizontalement ou verticalement des vecteurs et des matrices à l'aide des commandes "cbind()" et "rbind()".

$$P \leftarrow \mathrm{cbind}(\mathrm{N}, \mathrm{O}) \quad Q \leftarrow \mathrm{rbind}(\mathrm{N}, \mathrm{O})$$

Pour extraire les éléments d'une matrice, on procède comme suit

$$P[1, 3] \quad P[, 3] \quad P[3,] \quad P[1:3, 1:3]$$

Le logiciel R dispose de certaines fonctions pour manipuler des matrices et des vecteurs. Tester les fonctions suivantes :

$$P \,\%*\%\, P \quad \det(P) \quad \mathrm{eigen}(\mathrm{P}) \quad \mathrm{solve}(\mathrm{P}) \quad t(P)$$

Il existe aussi une fonction "chol()" qui à toute matrice symétrique réelle définie positive S associe la matrice triangulaire supérieure T telle que $S = T^t T$ où T^t est la matrice transposée de T. Pour conclure cette section, voici quelques exercices ; pour chaque exercice, écrire le code demandé dans l'éditeur de texte en haut à gauche puis le sauvegarder dans un fichier ".R".

Exercice 1
Créer un vecteur contenant 10 valeurs régulièrement espacées entre 4 et 13 et utilisez-le comme argument des fonctions "sum()", "prod()", "mean()", "median()", "var()", "sd()", "max()", "min()", "length()","pmax()", "which.max()", "which.min()", "cumsum()", "cumprod()".

Exercice 2
Calculer la somme des carrés des inverses des 10, 100 et 1000 premiers entiers non nuls et comparer les valeurs obtenues à $\sum_{k=1}^{+\infty} 1/k^2$. Rappeler la valeur théorique de cette série et conclure. Soit X une variable aléatoire distribuée uniformément sur les entiers de la forme k^2 où k est un entier compris entre 1 et 10. Calculer à l'aide des commandes vues jusqu'ici l'espérance et la variance de X.

Exercice 3
Utiliser R pour résoudre le système suivant

$$\begin{cases} 3x + 2y + z = 5 \\ 2x + 3y + z = 1 \\ x + 2y + 3z = 7 \end{cases}$$

Exercice 4
A l'aide des commandes vues jusqu'ici et de l'approximation d'une intégrale de Riemann par une somme de Riemann, calculer une approximation de $\int_{-\pi}^{+\pi} (\sin(10x))_+ \, dx$ où $(\sin(u))_+ = \max(\sin(u), 0)$. Faire le calcul exact de $\int_{-\pi}^{+\pi} (\sin(10x))_+ \, dx$ et comparer cette quantité à l'approximation obtenue à la question précédente.

3 Les fonctions

Grâce au logiciel R, il est possible de construire ses propres fonctions. La structrure générale d'une fonction est la suivante

```
nom_de_la_fonction ← function(liste_des_variables){
commandes
return(sortie_de_la_fonction)}
```

On a vu plus haut un exemple de fonction dont le paramètre d'entrée est un scalaire et qui renvoie un scalaire. Généralement, quand on entre un vecteur $(x_1, \ldots, x_N)$ comme paramètre à une telle fonction, elle renvoie $(f(x_1), \ldots, f(x_N))$. On voit sur le code générique ci-dessus que l'on peut passer plusieurs arguments à une fonction. Cependant, elle ne doit retourner qu'un objet, lequel peut être par exemple un vecteur. Ainsi

```
premiers_carres ← function(x){
u ← 1 : x
return(u^2)
}
```

retourne les carrés des x premiers entiers non-nuls. Pour pouvoir passer un vecteur comme argument à une telle fonction il faut utiliser la commande Vectorize(). Exécuter

```
vpremiers_carres <− Vectorize(premiers_carres, 'x')
x ← 1 : 5
vpremiers_carres(x)
```

Considérons maintenant le cas d'une fonction de deux arguments scalaires renvoyant un scalaire. Par exemple la fonction suivante

$$\begin{aligned}&\mathrm{norme} \leftarrow \mathrm{function(x, y)\{}\\&\mathrm{return(sqrt}\, x^2 + y^2)\\&\}\end{aligned}$$

Pour passer deux vecteurs $(x_1, \ldots, x_N)$ et $(y_1, \ldots, y_M)$ en argument à une telle fonction et récupérer la matrice de terme général $\mathrm{norme(x_i; y_j)}$ il faut utiliser la commande "outer()"

$$\begin{aligned}&x \leftarrow 1:5\\&y \leftarrow 1:3\\&M \leftarrow \mathrm{outer(x, y, 'norme')}\end{aligned}$$

Finalement, on peut définir des fonctions qui associent des vecteurs ou matrices à plusieurs scalaires, comme dans

$$\begin{aligned}&\mathrm{premieres\_puissances} \leftarrow \mathrm{function(x, y)\{}\\&u \leftarrow 1:x\\&\mathrm{return(u^y)}\\&\}\end{aligned}$$

qui retourne la liste des puissances y-ièmes des x premiers entiers non-nuls.

Exercice 5
À l'aide de la fonction factorial() et des commandes vues jusqu'ici créer une fonction qui à tout entier n associe la probabilité qu'une variable aléatoire de loi de Poisson de paramètre π soit supérieure ou égale à n.

Exercice 6
Créer à l'aide des commandes vues jusqu'ici une fonction qui à tout entier non-nul n associe la probabilité qu'une variable aléatoire de loi géométrique de paramètre $1/\pi$ soit supérieure ou égale à n.

Exercice 7
Calculer à l'aide des commandes vues jusqu'ici la probabilité qu'une variable aléatoire de loi binomiale de paramètres 8 et $1/\pi$ soit supérieure ou égale à 4.

4 Les graphes

Le langage R dispose de fonctions dédiées aux graphes. Pour superposer plusieurs tracés sur un même graphe il faut, au préalable, introduire la commande "par(new=TRUE) "comme dans l'exemple suivant :

$$\begin{aligned}&x \leftarrow \mathrm{seq}(-\pi, \pi, 0.1)\\&y \leftarrow \sin(x^2 - x)\\&z \leftarrow \cos(x^2 - x)\\&\mathrm{plot(x, y, type = 'l', col = 'red')}\end{aligned}$$

observer le résultat puis continuer en exécutant

$$\text{par}(\text{new} = \text{TRUE})$$
$$\text{plot}(\text{x}, \text{z}, \text{type} = \text{'l'}, \text{col} = \text{'cyan'}, \text{main} = \text{'Superposition de deux courbes'})$$

Les tracés ainsi obtenus peuvent être sauvegardés en utilisant les menus de la fenêtre située en bas à droite. On peut aussi créer des figures en 3 dimensions

$$x \leftarrow \text{seq}(-10, 10, \text{length} = 30)$$
$$y \leftarrow x$$
$$f \leftarrow \text{function}(x, y)\{$$
$$r \leftarrow \text{sqrt}(x^2 + y^2)$$
$$\text{return}(10 * \sin(r)/r)$$
$$\}$$
$$z \leftarrow \text{outer}(x, y, f)$$
$$\text{persp}(x, y, z, \text{theta} = 30, \text{phi} = 30, \text{expand} = 0.5, \text{col} = \text{'lightblue'})$$

On utilisera fréquemment des histogrammes pour rendre compte des résultats de simulations. En effet, il s'agit d'un moyen pratique de comparer visuellement des distributions empiriques et des distributions théoriques. La fonction "hist()" appliquée à un vecteur de données $(x_1, \ldots, x_N)$ permet d'obtenir un histogramme. On a plusieurs façons de définir les classes de l'histogramme, ce qui se fait via l'option breaks

1. Si on prend comme classes des intervalles $[a_1, a_2[$, $[a_2, a_3[$,.., $[a_{k-1}, a_k[$ alors il faut passer comme paramètre de l'option breaks le vecteur de coordonnées $a_1, \ldots, a_k$
2. Si on veut un nombre de classes m fixé à l'avance on passe cet entier m comme paramètre de l'option breaks

Par exemple, le code suivant permet d'obtenir le résultat de 100 lancers d'un dé à six faces équilibré (les détails seront vu au paragraphe suivant)

$$x \leftarrow \text{sample}(\text{x} = 1{:}\,6, \text{prob} = \text{rep}(1/6, 6), \text{size} = 100, \text{replace} = \text{TRUE})$$
$$\text{bords} \leftarrow \text{seq}(0.5, 6.5, 1)$$
$$h \leftarrow \text{hist}(\text{x}, \text{breaks} = \text{bords}, \text{plot} = \text{TRUE})$$
$$\text{print}(\text{h})$$

Attention : Par défaut l'histogramme affiche les fréquences. Il faut préciser $\text{freq} = \text{F}$ dans les options de hist pour afficher un histogramme avec une aire totale de 1 (et ainsi pouvoir comparer avec une densité de référence).

5 Les probabilités avec R

Le logiciel R permet de manipuler un certain nombre de lois : "beta", "binom", "cauchy", "chisq", "exp", "f", "gamma", "geom", "norm", "hyper", "pois", "t", "unif", "weibull", en combinant leurs noms avec les préfixes d, p, q ou r. Par exemple,

1. dnorm(x) donne la valeur en x de la densité de la loi $\mathcal{N}(0,1)$;
2. pnorm(x) donne la valeur en x de la fonction de répartition de la loi de densité $\mathcal{N}(0,1)$;
3. qnorm(x) donne le quantile d'ordre $x \in]0;1[$ de la loi de densité $\mathcal{N}(0;1)$;
4. rnorm(n) retourne le résultat de n tirages au sort indépendants effectués suivant la loi $\mathcal{N}(0,1)$.

Exercice 8

1. Simuler le tirage de 1000 variables aléatoires indépendantes de loi $\mathcal{N}(0,1)$.
2. Tracer un histogramme à 50 classes correspondant aux résultats de ces tirages.
3. Sur le même graphique, tracer la densité de la loi $\mathcal{N}(0,1)$ et comparer les deux figures (**indication** : utiliser la fonction "curve" correctement paramétrée).
4. On note $x_1 \dots x_{1000}$ les résultats de ces tirages. Tracer l'extrapolation linéaire de la famille de points $(n; 1/n \sum_{k=1}^{n} x_k)_{1 \leq n \leq 1000}$.

Le logiciel R permet également de traiter les tirages avec ou sans remise dans une population finie à l'aide de la commande sample(). Sa syntaxe est sample(x,n,replace= , prob=) où

1. x est un vecteur listant la population sur laquelle est effectuée le tirage
2. n est le nombre de tirages effectués
3. $\mathrm{replace} = \mathrm{TRUE}$ signifie que le tirage est effectué avec remise tandis que $\mathrm{replace} = \mathrm{FALSE}$ signifie que le tirage est effectué sans remise. Par défaut replace=FALSE.
4. Si $\mathrm{replace} = \mathrm{TRUE}$, dans $prob = p$ l'argument p est un vecteur de même taille que x et $p[i]$ est la probabilité de tirer $x[i]$. Par défaut, le tirage avec remise s'effectue suivant la loi uniforme. Si $\mathrm{replace} = \mathrm{FALSE}$, dans $prob = p$ l'argument p est un vecteur de même taille que x et $p[i]$ est la probabilité de tirer $x[i]$ en premier. Ensuite, le vecteur de probabilités de référence est adapté pour que le résultat corresponde bien à un tirage sans remise.

Exercice 9

1. Simuler 1000 tirages indépendants d'un dé pipé de sorte que $p_4 = 1/3$ et $p_1 = p_2 = p_3 = p_5 = p_6$.
2. Tracer un histogramme à 6 classes correspondant aux résultats de ces tirages.
3. On note $x_1, \dots, x_{1000}$ les résultats de ces tirages. Tracer l'extrapolation linéaire de la famille de points $(n; 1/n \sum_{k=1}^{n} x_k)_{1 \leq n \leq 1000}$.

6 Itérations et structures de contrôle

Les itérations se font suivant un nombre d'exécutions fixé à l'avance, par exemple

$$\mathrm{for(i\,in\,1:10)\{print(i)\}}$$

ou en subordonnant les exécutions à la vérification d'une condition (ce qui peut conduire à une boucle infinie)

$$\begin{array}{l} i \leftarrow 1 \\ \mathrm{while}(i < 11)\{\mathrm{print(i)}; i \leftarrow i+1\} \end{array}$$

On dispose aussi de la construction if (condition) {A} else {B} que l'on peut utiliser par exemple comme dans

$$\begin{aligned}&\text{xlogx} \leftarrow \text{function}(x)\{\\&\text{if}(x > 0)\{\text{return}(x * \log(x))\}\\&\text{else}\{\text{return}(0)\}\\&\}\end{aligned}$$

Exercice 10 : Le paradoxe des anniversaires

1. Créer une fonction qui à tout entier N tel que $2 \leq N \leq 365$ associe la probabilité que dans une classe de N élèves au moins 2 élèves fêtent leurs anniversaires le même jour.
2. Trouver la plus petite valeur de N pour laquelle cette probabilité dépasse $1/2$.
3. Tracer le graphe de la fonction trouvée à la première question.