Probabilités discrètes et Statistiques descriptives **L2 Université Lyon 1 2018**

TP : Initiation à R

R est un logiciel libre qui peut être téléchargé gratuitement sur le site du CRAN (Comprehensive R archive project, https ://cran.r-project.org/), qui fonctionne sous Microsoft Windows, Mac OS et Linux.

1 Interface Rstudio

Pour travailler avec R on dispose de l'interface Rstudio, qui est installée sur les machines, et dont une version open source est téléchargeable à l'adresse https ://www.rstudio.com/.

Cette interface comporte

- une console, permettant d'exécuter des commandes R (en bas à gauche),
- un éditeur, permettant d'écrire des scripts R, que l'on peut ensuite lancer dans la console (en haut à gauche),
- un espace en haut à droite permettant d'afficher au choix l'historique des commandes lancées ou la liste des variables en mémoire,
- un espace en bas à droite, permettant d'afficher en particulier une fenêtre d'aide (à utiliser sans modération), une fenêtre d'affichage de graphiques ou un navigateur de fichiers.

2 Prise en main

Ouvrir un nouveau script et l'enregistrer en utilisant un nom se terminant par .R . Taper une commande dans le script, par exemple 1+2. Pour la lancer dans la console, sélectionner la ligne correspondant et taper Ctrl + Entrée.

Remarque : si besoin plusieurs lignes peuvent être sélectionnées et lancées en même temps. Pour mettre une ligne en commentaire (cette ligne ne sera alors pas lancée dans la console, même si elle est sélectionnée), la commencer par #.

Pour obtenir de l'aide sur une fonction, par exemple plot, on peut utiliser directement le moteur de recherche situé dans la fenêtre d'aide, ou taper dans la console ?plot ou help(plot). Pour lancer une recherche plus générale, par exemple sur le mot 'matrix', taper help.search('matrix') dans la console. Écrire dans un script et lancer les commandes suivantes. Interpréter les résultats.

Commandes de base :

```
1+1
0.1*4
a <- exp(2)
a
b <- cos(1) ; b
```

Création de vecteurs :

```
v <-c(1,9,-4,0.5) ; v
v[3]
v[1]
1 :10
seq(from=0, to=1, by=0.1)
seq(from=0, to=2, length=11)
rep(x=1, times=10)
rep(x=1 :2, times=3)
rep(x=1 :2, each=4)
```

Opérations sur les vecteurs :

```
A<-c(1,2,3,6) ; B<-c(0,-4,9,4) ; exp(A)
A+B
A*B
```

Opérations sur les vecteurs (suite) :

```
A%*%B
A/B
2*A
A+1
A+c(1,2,3)
A==2
T<-A>1 ; A[T]
u <- (1 :10) ; sum(u)
cumsum(u)
```

Création et opération sur les matrices :

```
C<-matrix(A,ncol=2,nrow=2) ;
D<-matrix(B,ncol=2,nrow=2,byrow=TRUE)
t(C)
class(C)
C%*%D
C[1,]
D/C[1,]
eigen(C) #Trouvez les valeurs propres
et les vecteurs propres de C
```

Tracés de graphiques :

u<- seq(from=0, to=2*pi, by=0.01)
v<-sin(u)
plot(x=u, y=v, type='l', ylim=c(-1,1), xlim=c(0,2*pi))

3 Lois usuelles discrètes

3.1 Diagrammes en bâtons

Pour tracer le diagramme en bâtons décrivant de la loi binomiale $\mathcal{B}(10, 0.25)$, on peut utiliser les commandes suivantes :

x <- 0 :10
y <- dbinom(x, size=10, prob=0.25)
plot(x, y, type='h')

Tracer de même un diagramme en bâton des lois $\mathcal{B}(10, 0.5)$, $\mathcal{B}(100, 0.25)$, $\mathcal{P}(2)$, $\mathcal{P}(10)$, $\mathcal{G}(0.75)$, $\mathcal{G}(0.25)$ et de la loi uniforme sur $\{1, 2, 3, 4, 5, 6\}$. Il peut être utile de faire la recherche distribution dans l'aide (faire attention à la définition de la loi géométrique $\mathcal{G}(p)$, qui est à support sur $\mathbb{N}^*$ selon le cours).

3.2 Fonctions de répartition

Tracer la fonction de répartition de la loi $\mathcal{B}(10, 0.25)$ (on pourra utiliser la fonction pbinom et l'option type='s' de la fonction plot). Tracer de même la fonction de répartition des lois $\mathcal{B}(100, 0.5)$ et $\mathcal{P}(3)$.

Que valent $P(X \leq 3)$ et $P(X > 30)$ si X suit la loi $\mathcal{B}(50, 0.2)$?

3.3 Lien binomiale/Poisson

Rappeler le théorème de convergence de la loi binomiale vers la loi de Poisson.

Illustrer cette convergence en représentant sur un même graphique les lois $\mathcal{B}(500, 0.02)$ et $\mathcal{P}(10)$.

On pourra pour cela par exemple représenter la loi sous forme de diagrammes à bâtons, et la loi $\mathcal{P}(10)$ à l'aide de la fonction points.

4 Lois usuelles continues

4.1 Densités

Pour tracer la densité de la loi $\mathcal{N}(0, 1)$, on peut utiliser les commandes suivantes :

x <- seq(from=-4, to=4, by=0.01)
y <- dnorm(x, mean=0, sd=1)
plot(x, y, type='l')

Tracer de même la densité des lois $\mathcal{N}(1, 2)$ (de moyenne 1 et variance 2), $\mathcal{N}(1, 0.5)$, $\mathcal{E}(1)$ et de la loi de Cauchy de paramètres 0 et 1.

4.2 Fonctions de Répartition

Tracer la fonction de répartition des lois $\mathcal{N}(0, 1)$, $\mathcal{N}(-2, 1)$ et $\mathcal{E}(2)$.

Pour X de loi $\mathcal{N}(0, 1)$, calculer $P(X \leq 2)$, $P(-1 \leq X \leq 0.5)$ et trouver u tel que $P(X \leq u) = 0.9$.

5 Statistiques descriptives

5.1 Import de données

Pour illustrer les outils basiques de statistiques descriptives, on va charger un jeu de données, en utilisant la commande (attention au caractère ∼, obtenu avec AltGr+2)

Don<-read.delim("http ://math.univ-lyon1.fr/∼dabrowski/enseignement/ProbaL2/Donnees.csv")

Les données correspondent à l'âge, au poids, à la taille, à la consommation hebdomadaire d'alcool (en nombre de verres bus), au sexe, au ronflement et au tabagisme, d'un échantillon de 100 personnes.

On récupère (par exemple) les données concernant l'âge, l'alcool, et le ronflement, de la façon suivante :

age<-Don[1 :100,1] ;alcool<-Don[1 :100,4] ;ronfle<-Don[1 :100,6]

ou utiliser attach(Don), pour le faire pour toutes les variables.

Commencez par identifier les variables qualitatives nominales, ordinales et quantitative discrètes et continues. Typez les variables avec les classes factor, ordered, integer ou double.

5.2 Résumés numériques

En utilisant les fonctions mean, var, quantile, déterminer la moyenne empirique, la variance empirique, la variance empirique non-biaisée et l'écart type empirique pour les variables quantitatives. Déterminer la médiane pour chaque caractéristique ordinale ou quantitative de l'échantillon.

Représenter également les données à l'aide d'un diagramme en boîte (boxplot). Rappeler les éléments représentés dans ce diagramme.

5.3 Résumés graphiques

Représenter les variables quantitatives discrètes et continues à l'aide respectivement d'un diagramme en bâton et d'un histogramme (fonction hist). Représenter leurs fonctions de répartitions empiriques. (On peut utiliser les fonctions ecdf, plot et hist si on veut représenter la fonction de répartition de l'histogramme pour les variables continues)

Représenter les variables qualitatives par un diagramme en tuyaux d'orgue (fonction barplot).

Donner les tables de contingences pour les variables sexes et tabagisme. Représenter par un nuage de point les variables poids et taille.

6 Analyse en composantes principales

On continue avec le jeu de données de la section précédente. Évaluer les corrélations entre les 4 différentes caractéristiques quantitatives (en utilisant la fonction cor).

Évaluer l'échantillon des variables centrées réduites associées aux 4 différentes caractéristiques quantitatives (on pourra convertir en matrice A le data.frame et utiliser $apply(A, MARGIN = 2, FUN = sd)$) puis la covariance de ces variables (en utilisant la fonction cov).

Effectuez l'analyse en composantes principales. Représentez par un nuage de points les 2 composantes principales de l'ACP puis le cercle des corrélations (on pourra utiliser install.packages("plotrix"); library("plotrix") et la commande draw.circle$(0, 0, 1)$ pour représenter le cercle). Quelles sont les variables de départ bien représentées par ces 2 composantes ? Quelle est la variance empirique de ces deux composantes (obtenir cette variance de deux façons) ?

7 Simulations de lois usuelles

7.1 Lois discrètes

Simuler N réalisations indépendantes (N grand) de la loi $\mathcal{B}(10, 0.5)$ (on pourra utiliser la fonction rbinom). Tracer le diagramme en bâton correspondant et représenter sur le même graphique les probabilités théoriques correspondantes (on pourra utiliser la fonction points).

Procéder de même avec les loi $\mathcal{P}(3)$, $\mathcal{G}(0.4)$.

7.2 Lois continues

Simuler N réalisations indépendantes de la loi $\mathcal{N}(0, 1)$. Tracer l'histogramme correspondant (on pourra définir les bords des intervalles de l'histogramme à l'aide de l'option breacks et demander une graduation en densité), et la densité de la loi $\mathcal{N}(0, 1)$ (pour tracer la densité on pourra utiliser la fonction lines).

Procéder de même avec les lois $\mathcal{U}([0, 3])$ et $\mathcal{E}(1)$.

8 Test d'indépendance

8.1 Rappel du test vu en cours

.

Si on a les variables statistiques X, Y à valeurs I, J finies respectivement et qu'on a les effectifs empiriques $n = (n_{(i,j)})$ ($n_{i,j}$ nombre d'individus de l'échantillon avec $X = i$ et $Y = j$) et l'effectif total $N = \sum_{i\in I}\sum_{j\in J} n_{i,j}$. On peut calculer les effectifs marginaux $N_{(i,.)} = \sum_{j\in J} n_{i,j}$ et $N_{(.,j)} = \sum_{i\in I} n_{i,j}$. En pratique, comme le test est asymptotique, il faut au moins que $N \geq 30$ et tous les $N^{th}_{i,j} \geq 5$ pour pouvoir appliquer le test (sinon on doit regrouper les classes et R affiche un avertissement). Sous l'hypothèse d'indépendance les effectifs théoriques sont : $N^{th}_{i,j} = \frac{N_{(i,.)}N_{(.,j)}}{N}$. Le test va donc quantifier l'écart à la validité de cette équation. On considère la statistique de Pearson :

$$D^2(n) = \sum_{i\in I}\sum_{j\in J} \frac{(n_{i,j} - N^{th}_{i,j})^2}{N^{th}_{i,j}}.$$

On considère une loi $\chi^2(d)$ avec $d = (Card(I) - 1)(Card(J) - 1)$ est appelé nombre de degrés de libertés. Le test dit donc qu'au niveau α (en général $\alpha = 0.05$) si $\chi^2_\alpha(d)$ est le fractile tel que $P(\chi^2(d) \leq \chi^2_\alpha(d)) = 1 - \alpha$ et :

1. si $D^2(n) > \chi^2_\alpha(d)$ on rejette l'hypothèse d'indépendance.
2. si $D^2(n) \leq \chi^2_\alpha(d)$ on ne peut pas rejeter l'hypothèse d'indépendance au niveau α, et **par ABUS** on dit parfois qu'on l'accepte.

8.2 Application

En utilisant les commandes chisq.test, qchisq, appliquer le test au niveau $\alpha = 0.05$ pour savoir si on peut considérer le ronflement indépendant du sexe sur notre échantillon.

En regroupant si nécessaire en classes, faire le même test avec les variables alcool et sexe.

9 Loi des grands nombres

Simuler un échantillon $(X_1, X_2, ..., X_N)$ de variables i.i.d. de loi $\mathcal{B}(1, 0.5)$ pour N grand. Illustrer la loi des grands nombres en traçant $\overline{X}_n = \frac{X_1+X_2+\cdots X_n}{n}$. en fonction de n (compris entre 1 et N).

On pourra pour cela utiliser la fonction cumsum.

Procéder de même avec des échantillons de loi $\mathcal{N}(1, 4)$ et de Cauchy de paramètre 0 et 1. Que remarque-t-on ?

10 Théorème centrale limite

Pour un échantillon $(X_1, X_2, ..., X_N)$ de variables i.i.d. de loi $\mathcal{B}(1, p)$, on définit la variable aléatoire V_N par

$$V_N = \sqrt{N}\frac{\overline{X}_N - p}{\sqrt{p(1-p)}}$$

avec $\overline{X}_N = \frac{X_1+X_2+\cdots X_N}{N}$.

Illustrer le Théorème centrale limite (théorème de Moivre-Laplace) en simulant k réalisations indépendantes de V_N (on pourra utiliser une boucle for), et en traçant dans le même graphique l'histogramme correspondant et la densité de la loi $\mathcal{N}(0, 1)$.

En utilisant le théorème, obtenir un intervalle de confiance de niveau $1-\alpha = 0.95$ pour la proportion de femme dans l'échantillon de données de la section 5. Faire de même pour la proportion de ronfleurs et fumeurs.