Licence Informatique S4
Probabilités-Statistiques TP 1
2019-2020

1er février 2020

Les TP's se font avec le logiciel R. On apprend R comme tout autre logiciel : en essayant, testant, tâtant, expérimentant. Si vous ne savez pas quoi faire, demander à "help", ou à Google. À défaut, à un humain. N'oubliez néanmoins pas que, dans ce cours, l'apprentissage de R n'est pas un but en soi. Le but est d'utiliser quelques fonctionnalités de R pour apprendre, illustrer et rendre plus concrètes les notions de probabilités et de statistique introduites et étudiées dans le cours et dans les TD's. Il s'agit donc de se familiariser suffisamment avec l'environnement fourni par l'ordinateur et RStudio (voir ci-dessous) pour faciliter cet apprentissage. A contrario, une maîtrise insuffisante de l'outil informatique sera une entrave à la compréhension du cours.

1 Lancer R ; quitter R ; premières fonctions.

Une fois connecté à votre espace personnel, lancer le le logiciel RStudio, qui fournit un environnement de développement pour R. Son utilisation est intuitif, mais un tutoriel efficace de 11 minutes que je vous recommande vivement se trouve ici : `https://www.youtube.com/watch?v=1mSzskE1TWs` .

Pour commencer, créer d'abord un premier project R. Nommez-le et sauvegardez-le dans votre espace personnel. Tout est expliqué dans les 3 premières minutes du tutoriel.

Si vous souhaitez travailler sur votre propre ordinateur, il est facile d'installer R et RStudio. Utiliser par exemple les liens suivants :
Pour le logiciel R : `https://cran.r-project.org` .
Pour RStudio : `https://www.rstudio.com/products/rstudio/download/`
.

Quelques remarques générales qui peuvent être utiles pour la suite :

1. Les fonctions à exécuter s'écrivent toujours avec des parenthèses () qui peuvent contenir des arguments optionnels par exemple `ls()` qui affiche la liste des noms des objets en mémoire ou `ls(pat="^m")` affiche la liste des noms des objets en mémoire commençant par m.
2. Pour avoir de l'aide spécifique sur l'utilisation d'une fonction par exemple `ls`, tapez `help("ls")`. Voir aussi `help.start()` pour une aide générale.
3. Pour obtenir des exemples sur certaines fonctions (`plot,hist`) tapez `example(plot)`.

2 Données : les charger, les créer, les manipuler

1. Générer les données suivantes :
   (a) une suite d'entiers de 1 à 20 à mettre dans la variable `x`,
   (b) une suite de 1 à 10 par pas de 0.5 à mettre dans la variable `y`,
   (c) une suite constitué de 2 répété 12 fois à mettre dans la variable `z`.
   (d) La suite de caractères A...A B...B C...C où A,B,C sont répétés 20 fois.

   On pourra utiliser les commandes `c(), seq(), rep(), gl()`. Pour savoir ce que font ces commandes et quelles arguments elles acceptent, utiliser `help()`. Par exemple `help(gl)` vous informera sur la command `gl()`.
2. Pour obtenir la liste des bases de données disponibles tapez `data()`. Taper ensuite par exemple `data(morley)` ou `morley`. Pour comprendre la signification des données montrées, taper `help(morley)`
3. Extraire des données
   (a) Utiliser la commance `matrix()` pour créer une matrice de 12 lignes et 12 colonnes contenant les données de AirPassengers. Attention, que s'est-il passé avec les lignes et les colonnes de AirPassengers ? Corriger cela éventuellement.
   (b) Extraire de cette matrice les données correspondant aux mois Février-Juin, des années 1958-1959, en forme de tableau.
4. Essayer de stocker les données suivantes de 4 individus :
   taille : 168 ; 175 ; 172 ; 183
   poids : 67 ; 75 ; 69 ; 81
   Il y a plusieurs méthodes : voir les fonctions `c(),scan(),data.frame(),matrix()`.
   Donner des noms aux lignes et aux colonnes : par exemple "Jean", "Fatima", "Kevin", "Eduardo", pour les colonnes et "taille", "poids", pour les lignes.
   Calculer l'Indice de Masse Corporelle pour les quatre personnes, dans une colonne ajoutée au tableau ci-dessus.

3 Échantillon et histogramme

1. La base de données `morley` donne des valeurs expérimentales de la vitesse de la lumière, obtenues au 19ème siècle par le physicien E. W. Morley. Utiliser `help(morley)` pour comprendre le contenu précis de la base de données.

   Calculer, à l'aide de R, la moyenne empirique (`mean()`), la médiane, et l'écart type (`sd()`) de ces données, pour chacune des 5 expériences, et globalement. Tracer aussi les 5 histogrammes et l'histogramme global.

   Est-ce que vous constatez des différences notables entre les résultats de ces 5 expériences ?

   Quelle est, selon vous, la meilleure estimation qu'on peut donner de la valeur de la vitesse de la lumière sur la base de ces données. Avec quelle erreur ? Comparer avec la valeur de cette quantité utilisée aujourd'hui.

2. On considère maintenant les données correspondant aux années 1956-1960 de la base de données AirPassengers : 60 valeurs donc. Établir la moyenne empirique, la médiane et l'écart type. Tracer des histogrammes de ces 60 données, en choisissant différentes tailles de classe. Expliquer et commenter les résultats.

3. Créer un vecteur `x` contenant les valeurs de `sin(t)` pour `t` appartenant à $[0, 2\pi]$. *Attention :* à vous de décider du nombre d'éléments dans le vecteur `t` et donc du vecteur `x`. Il faudra les varier pour répondre aux questions suivantes.

   (a) Tracer `plot(t,x)`.

   (b) Tracer un histogramme du vecteur `x` avec des classes bien choisies. Comment expliquer l'allure de cet histogramme ? Notamment son lien avec le graphe précédent.

   (c) Un point qui se déplace sur un cercle du rayon 1m dans le plan, à vitesse angulaire constante et égale à 1, a comme coordonnées cartésiennes $x(t) = \cos(t), y(t) = \sin(t)$. Faire (à la main), un dessin pour se représenter la situation. Le mobile fait donc le tour du cercle en 2π secondes. Expliquer comment on peut utiliser l'histogramme sous (b) pour répondre à la question suivante. Soit $-1 \leq a < b \leq 1$. Pendant quelle fraction de temps $t \in [0, 2\pi]$ a-t-on alors $a < x(t) < b$? Notamment pour $a = -1, b = -0.9$ et $a = 0, b = 0.1$. Comment expliquer la (grande) différence entre ces deux valeurs ?
   Indication. Il ne s'agit pas ici de coder en R, mais de réfléchir à la signification des histogrammes obtenus obtenus sous (b).

Des données aux graphiques

1. Représenter graphiquement la fonction $x \mapsto \dfrac{\mathrm{e}^{-x^2/2}}{\sqrt{2\pi}}$. On utilisera `curve()`.
2. On considère maintenant $x \mapsto \dfrac{\exp(-\frac{(x-\mu)^2}{2\sigma^2})}{\sigma\sqrt{2\pi}}$ où $\mu \in \mathbb{R}$ et $\sigma > 0$. Quelle est l'influence de ces paramètres μ et σ sur le graphe de la fonction ? On pourra superposer les graphes en utilisant l'option `add=TRUE` dans `curve()`.
3. Même question pour la densité de probabilité de la loi exponentielle $x \mapsto \alpha\exp(-\alpha x)$ sur $[0,+\infty[$.