Université Grenoble Alpes
L2/STA351

Le but de ces TP est que vous soyez capable de mener une étude statistique sur des données expérimentales, qu'elles soient qualitatives, quantitatives discrètes ou continues. Il est important que vous compreniez les commandes et que vous sachiez les reproduire. Elles ne seront pas données lors des évaluations.

TP1 : Prise en main de R - Description d'une ou deux variables qualitatives

Objectifs :
I) apprendre les commandes de base du logiciel R, savoir créer des vecteurs ou des matrices, en extraire certains éléments et effectuer des calculs sur des vecteurs
II) savoir charger un jeu de données, les trier, décrire une ou deux variables qualitatives, calculer des statistiques standards : fréquences, fréquences marginales et conditionnelles

1 Premiers pas avec R et l'environnement R Studio :

RStudio est un logiciel libre et collaboratif de statistique. Nous utiliserons dans les TP l'interface Rstudio de ce logiciel, que l'on peut télécharger gratuitement depuis le site

`www.rstudio.com`,

en ayant au préalable installé le logiciel R lui même, disponible sur le site

`www.r-project.org`.

Attention, dans RStudio, une majuscule et une minuscule n'ont pas le même sens.

RStudio est séparé en 4 fenêtres graphiques :

- la console (en bas à gauche) : permet d'exécuter les instructions avec la touche Entrée ↩
- le script (en haut à gauche) : permet d'écrire l'ensemble des commandes (ou instructions) que l'on veut exécuter et de pouvoir les sauvegarder dans un fichier. Une instruction s'exécute dans la fenêtre en bas à gauche en appuyant sur Ctrl+Entrée. Pour ne pas perdre de temps, vous pouvez copier-coller les commandes qui sont données dans les énoncés de TP dans votre script.
- la fenêtre Files/Plots/Packages/Help (en bas à droite) : permet entre autres de visualiser les graphiques ou l'aide.
- la fenêtre Workspace/History (en haut à droite) : permet de voir l'ensemble des objets en mémoire et historique de toutes les instructions réalisées.

Exercice 1

1. Créer un répertoire STA351 dans votre dossier "documents", un sous-répertoire TPSTA351 et un sous répertoire pour chaque fiche de Tp (il y en aura huit) dans lequel vous rangerez tous les documents relatifs à la fiche étudiée : énoncé de la fiche .pdf, scripts .R, graphiques .pdf, donnees.csv, ...

   Enregistrer l'énoncé du TP1 dans le répertoire TP1.

   Ce répertoire servira pour toutes les séances relatives à la fiche TP1. On créera un nouveau répertoire et un nouveau 'script' au début de chaque nouvelle fiche de TP.

2. Pour se simplifier la vie, nous allons travailler dans R studio depuis le répertoire de travail TP1 de TPSTA351. Pour cela, aller dans l'onglet session de la barre supérieure du menu, choisir Set Working Directory (Répertoire de Travail) et sélectionner le répertoire TP1 nouvellement créé.
3. Créer un nouveau script (File/New File/R script) taper n'importe quoi en première ligne et le sauvegarder.
4. Avant de continuer à remplir le script, s'assurer qu'on sait le sauvegarder correctement et au bon endroit. Fermer RStudio (sans sauvegarder l'environnement de travail). Relancer ensuite Rstudio et vérifier après s'être remis sur le répertoire de travail TP1 que le script précédemment sauvegardé est bien rechargé dans la fenêtre supérieure gauche en l'ouvrant au besoin avec le menu.

Exercice 2

Dans R, une fonction s'écrit toujours avec des parenthèses, dans lesquelles les paramètres (arguments et options) de la fonction sont précisés. Pour connaître l'utilité et les paramètres d'une fonction, vous pouvez vous reporter à l'aide (fonction `help`).
Les fonctions de bases sont listées dans le document "Commandes-R-1.pdf" que vous trouverez sur le site de l'UE. N'hésitez pas à vous y reporter lorsque vous cherchez une fonction. Google est aussi une bonne aide.

1. Dans le script, créer la suite de données $(1, 2, 3, 4, 5)$ avec l'instruction :
   `c(1,2,3,4,5)`
   `c(.)` est la concaténation qui "range" les valeurs ou caractères tapés les uns à la suite des autres dans un vecteur.
   Exécuter la ligne en cliquant sur "Run" ou avec Ctrl+R.
2. Le précédent vecteur n'a pas de nom. Donner lui un nom :

   ```
   x<-c(1,2,3,4,5)
   ```

   La flèche permet d'assigner des valeurs à l'objet x créé. RStudio ne retourne rien. Pour vérifier que le vecteur x contient les valeurs, exécuter l'instruction
   `x`
   Prenez l'habitude de vérifier les objets que vous créez, utilisez.
   A la place de `<-`, on peut aussi utiliser `=`.
3. Créer le vecteur y contenant les valeurs $(2, 4, 6, 8, 10)$.
4. Prenez l'habitude de sauvegarder votre script au fur et à mesure ! (cliquer sur le script et appuyer sur Ctrl+S).
5. Vérifier que les vecteurs x et y ont la même longueur (le même nombre de valeurs)
   `length(x)`
   `length(y)`
6. Tracer sur un graphique les points définis par les deux vecteurs (x, y) :
   `plot(x,y)`
7. Personnaliser votre graphique :

   ```
   plot(x,y, type = "p", pch = 3) # change les symboles
   plot(x,y, type = "b") # ajoute une ligne
   plot(x,y, col = "red") # change la couleur
   ```

   On peut aussi rajouter des titres

   ```
   plot(x,y,main="y selon x", type="p",xlab="abscisse",ylab="ordonnée") # ajoute un titre
   (paramètre main) et des légendes sur chaque axe (parametres xlab et ylab)
   ```

   Créer votre propre graphique en changeant les couleurs, les symboles, les titres.
   Toutes les fonctions sont décrites dans l'aide. Par exemple pour plot, taper :
   `help(plot)`
   ou
   `?plot`

8. Sauver votre graphique comme un fichier pdf en cliquant sur "export"
9. Pour connaitre l'ensemble des objets en mémoire, taper :
   `ls()`
   ou regarder la fenêtre environnement (en haut à droite).

Exercice 3

1. Opérations basiques : comprendre les opérations suivantes
   `x/5`
   `x+5`
   `sum(x)`
   `cumsum(x)`
   `sqrt(x)`
   `x ^ 3`
2. Rajouter des valeurs à la suite du vecteur x
   `c(x,6)`
   Cette commande ne change pas x puisqu'on n'a pas utilisé la flèche ou `x`. Pour changer x, faire :

   ```
   x<-c(x,6)
   x
   ```

   Pour reprendre les valeurs d'origine de x (dont on se sert ensuite)
   ```
   x<-c(1,2,3,4,5)
   z<- c(x,1,1,1,1,1)
   c(x,rep(1,5))
   c(x,seq(from=1, to=10, by=2))
   c(x, 6:15)
   ```
   Créer la suite de valeurs $(1, 4, 7, 10, \ldots, 61, 2, 2, 2, \ldots, 2, 1, 2, 3, \ldots, 20)$ où le nombre 2 a été répété 20 fois au milieu de la suite et où la première suite est celle des entiers de 1 à 61 par pas de 3 et la dernière celle des entiers de 1 à 20.
3. Dans R, on peut faire des tests logiques sur les éléments d'un vecteur. Comprendre les instructions suivantes :
   ```
   (y>4)
   (y!=4)
   y==4)
   (y>4)&(y<=6)
   ```
4. Dans R, les crochets permettent d'aller chercher des éléments d'un vecteur. Par exemple, pour extraire les deuxième et quatrième valeurs de y :
   `y[c(2,4)]`
   Comprendre les instructions suivantes :
   ```
   y[1:4]
   y[(y>4)]
   ```

   Extraire les valeurs de y plus grandes strictement que 2 et plus petites ou égales à 8.
   Extraire les valeurs de z différentes de 1.
   Extraire les valeurs de x égales à 2.
5. Comprendre les opérations de base avec deux vecteurs :
   ```
   x+y
   x*y
   x/y
   ```

6. Créer une table (ou une matrice) avec les deux vecteurs x et y

   ```
   cbind(x,y) # matrice avec 5 lignes et 2 colonnes (cbind permet de coller des colonnes les unes à la suite des autres)
   rbind(x,y) # matrice avec 2 lignes et 5 colonnes (rbind permet de coller des lignes les unes à la suite des autres)
   ```

7. De la même manière que pour les vecteurs, les crochets permettent d'aller chercher des éléments d'une matrice ou d'un tableau. Il y a alors 2 paramètres à définir, le premier pour les lignes et le second pour les colonnes à sélectionner (les 2 séparés par une virgule). Si rien n'est préciser pour le premier paramètre, cela signifie que l'on prend toutes les lignes. Si rien n'est préciser pour le deuxième paramètre, cela signifie que l'on prend toutes les colonnes.

   ```
   M<-rbind(x,y,y,x) # matrice avec 4 lignes et 5 colonnes
   M # toute la matrice
   M[3,2] # élément de la 3ieme ligne 2ieme colonne
   M[ ,1] # élément de la premiere colonne
   ```

   Extraire les élements de la deuxieme et troisieme lignes.
   Extraire les éléments de la première et troisième colonnes.

2 Statistique pour une ou deux variables qualitatives

Exercice 4

Dans l'ensemble des centres de transfusion sanguine de la région Rhône-Alpes on a observé sur un échantillon de n patients choisis au hasard parmi ceux ayant donné leur sang en 2013 les deux variables qualitatives : Groupe sanguin et Rhésus. Les effectifs observés sont donnés dans la table de contingence suivante :

Groupe / Facteur	O	A	B	AB
Rhésus +	40	38	6	1
Rhésus -	7	7	1	0

1. Quelles sont les variables d'intérêt ? Quels sont leurs natures ?
2. Rentrer les effectifs des Rhésus + dans un vecteur

   ```
   Rp <-c(40,38,6,1)
   ```

   Faire de même avec les effectifs des Rhésus -, stockés dans un vecteur `Rm`.
3. Créer une table `S` rassemblant les deux vecteurs. Ajouter le nom des groupes sanguins en faisant

   ```
   colnames(S)<- c("O", "A", "B", "AB")
   ```

4. Quelle est la taille de la population ?
5. Quels sont les effectifs de chaque Rhesus ? Quelles sont les fréquences des deux Rhesus ? Quelle est la proportion de donneurs ayant un Rhesus négatif ?
6. Représenter à l'aide de trois diagrammes en barres (on utilisera la fonction `barplot()`) la distribution marginale en fréquences du groupe (repartition du groupe sur l'échantillon complet indépendamment de la modalité du rhésus), la distribution conditionnelle du groupe sachant le rhesus positif (répartition du groupe observée sur le sous-échantillon des individus de rhésus positif) et la conditionnelle du groupe sachant le rhésus négatif. Comparer les graphes obtenus et conclure.

Exercice 5

On travaille dans cet exercice sur la base de données `titanic.csv`. Cette base contient les informations de 1046 passagers du Titanic :
— `pclass` la classe dans laquelle ils ont voyagé (1ere, 2eme ou 3eme classe)
— `survived` : `yes` ou `no` selon s'ils ont survit au naufrage ou non
— `gender` : sexe (`F` ou `M`)
— `age` : l'age en années
Enregistrer la base dans votre répertoire de travail TP1.

(a) Charger la base de données `titanic.csv`.

```
TI <- read.table(file="titanic.csv",header=TRUE,sep=";")
```

A quoi servent les paramètres "`header`" et "`sep`" ?

(b) Afficher les 6 premieres lignes de la base :

```
head(TI)
```

Ceci permet de vérifier que la base a bien été importée (nom des colonnes et format des variables) et d'en donner un premier apercu.

(c) Afficher le nom des colonnes de la base en utilisant la fonction `names()`.

(d) Affecter la colonne `pclass` à la variable `P`, la colonne `survived` à la variable `S`, la colonne `gender` à la variable `G` et la colonne `age` à la variable `A`. Cela permettra dans la suite d'utiliser des noms raccourcis.

```
P <- TI[,1]
```

On peut aussi utiliser

```
P <- TI[,"pclass"]
```

ou

```
P <- TI$pclass
```

Quel est le type de ces 4 variables ?

(e) Afficher les lignes 250 à 257.

```
TI[250:257, ]
```

Afficher les lignes 3, 45, 73. Afficher les colonnes 1 et 3. Afficher les colonnes 2 et 4 des lignes 67, 83, 101.

(f) Afficher les données des passagers de première classe.

```
TI[P==1,] # donnees des passagers de premiere classe
```

ou

```
TI[which(P==1),] # donnees des passagers de premiere classe
```

Afficher les données des passagers de troisième classe.

Afficher les données des passagers de deuxième et troisième classe.

```
TI[(P==2 | P==3),]
```

ou

```
TI[-(P==3),]
```

(g) Afficher les données des femmes de première classe. `TI[which(P==1 & G=="F"),]`

Il faut toujours mettre des guillemets pour les valeurs de variables "chaines de caractères".

Afficher les données des hommes de deuxieme classe.

(h) Afficher les données des bébés (moins de 1 an).

```
TI[A<1,] # données des bébés
```

Afficher les données des enfants (moins de 18 ans), des adolescents (12-18 ans) des adultes (plus de 18 ans).

(i) Calculer les effectifs et proportions des 3 classes de passagers :

`table(P)`

`prop.table(table(P))`

Quelle est la proportion des passagers en première classe ? On peut représenter cette répartition avec un diagramme en barres :

`barplot(table(P))`

(j) Calculer les effectifs et proportions des survivants et des décédés. Quelle est la proportion de survivants ? Calculer les effectifs et les proportions d'hommes et femmes. Quelle est la proportion de femmes ? Représenter graphiquement ces 2 variables.

(k) Calculer les effectifs des classes de passagers en fonction de la survie :

`table(P,S)`

Calculer les proportions de passager en fonction de la classe et de la survie :

`prop.table(table(P,S))`

Calculer la répartition des 3 classes de passagers selon la survie et la survie selon la classe de voyage. On utilisera la fonction `prop.table()` en rajoutant l'option 1 ou 2.

`prop.table(table(P,S),1)`

`prop.table(table(P,S),2)`

Combien de passagers de première classe ont survécu ? Quelle proportion parmi tous les passagers étaient en première classe et ont survécu ? Quelle proportion, parmi les passagers de première classe, ont survécu ? Quelle proportion de survivants était des passagers de première classe ? Répéter pour la 2ème et 3ème classe.

Proposer des représentations graphiques avec un diagramme en barres :

`barplot(table(P,S))`

`barplot(table(S,P))`

Commenter.

(l) Calculer la répartition des 3 classes de passagers selon le genre et le genre selon la classe de voyage. Combien de passagers de première classe sont des femmes ? Quelle proportion parmi tous les passagers étaient en première classe et sont des femmes ? Quelle proportion, parmi les passagers de première classe, sont des femmes ? Quelle proportion de femmes était des passagers de première classe ? Répéter pour la 2ème et 3ème classe. Proposer une représentation graphique.

Exercice 6 (pour s'entraîner)

On travaille dans cet exercice sur la base de données `bosson.csv` (fournies par le Professeur Jean-Luc Bosson). Cette base contient les informations de 209 patients venant de France ou du Vietnam :

— `country` : `Vietnam` ou `France`
— `gender` : `F` ou `M`
— `aneurysm` : taille de l'anévrisme en mm
— `bmi` : indice de masse corporelle
— `risk` : nombre de facteurs de risque entre 0 et 5

Enregistrer la base dans votre répertoire de travail TP2 de TPSTA351.

(a) Charger la base de données `bosson.csv`.

(b) Afficher les 6 premieres lignes de la base. Afficher le nom des colonnes de la base.

(c) Affecter la colonne `country` à la variable `C`, la colonne `gender` à la variable `G`, la colonne `aneurysm` à la variable `a`, et la colonne `risk` à la variable `R`. Quel est le type de ces variables ?

(d) Afficher les lignes 120 à 123. Afficher les colonnes 1 et 3 des lignes 67, 83, 101. Afficher les données des patients vietnamiens. Afficher le nombre de facteurs de risque des hommes.

(e) Calculer les effectifs et proportions des 2 pays. Quelle est la proportion de vietnamiens ? Calculer les effectifs et les proportions d'hommes et femmes. Quelle est la proportion de femmes ? Représenter graphiquement ces 2 variables.

(f) Calculer la répartition des pays selon le genre et le genre selon le pays. Combien de francais sont des hommes ? Quelle proportion parmi tous les patients sont des hommes francais ? Quelle proportion, parmi les patients vietnamiens, sont des femmes ? Quelle proportion de femmes sont vietnamiennes ? Représenter graphiquement la répartition de ces deux variables.

(g) Calculer la répartition des pays selon le nombre de facteurs de risque et le nombre de facteurs de risque selon le pays. Combien de vietnamiens ont 0 facteur de risque ? Quelle proportion parmi tous les patients sont des vietnamiens sans facteur de risque ? Quelle proportion, parmi les patients vietnamiens, ont 0 facteur de risque ? Quelle proportion de patients sans facteur de risque sont vietnamiens ? Représenter graphiquement la répartition de ces deux variables.

(h) Calculer la répartition des genres selon le nombre de facteurs de risque et le nombre de facteurs de risque selon le genre. Combien de femmes ont 0 facteur de risque ? Quelle proportion parmi tous les patients sont des femmes sans facteur de risque ? Quelle proportion, parmi les patients femmes, ont 0 facteur de risque ? Quelle proportion de patients sans facteur de risque sont des femmes ? Représenter graphiquement la répartition de ces deux variables.