TP1: introduction et statistiques descriptives
Statistiques pour Sciences Humaines II

Baptiste Perez et Anna Kiriliouk

Année 2021–2022

Pour chaque TP, vous aurez un récapitulatif des nouvelles commandes à utiliser (voir page suivante). Les commandes de tous les TPs seront combinées dans un seul document "Formulaire R" qui sera à votre disposition lors de la partie pratique de l'examen. N'oubliez pas que pour obtenir de l'aide sur n'importe quelle fonction, vous pouvez taper `?nomdelafonction` dans la console même.

Les graphiques La commande de base, `plot()`, permet de créer toutes sortes de graphiques (nuage de points, linéaire, ...). Si nous avons une ou plusieurs variables `var1`, `var2`, ..., il suffit de taper `plot(var1)` ou `plot(var1,var2)`. Pour améliorer la lisibilité d'une graphique, il y a plusieurs arguments que l'on pourrait ajouter à la commande `plot()` :

- Pour avoir un graphique avec des points ou avec une ligne, ajoutez respectivement `type = "p"` ou `type = "l"`
- Pour ajouter/modifier le titre du graphique, ajoutez `main = "Nom du Graphique"`
- Pour ajouter/modifier le nom des axes, ajoutez `xlab = "nom axe horizontal"` et `ylab = "Nom axe vertical"`
- Pour modifier la graduation des axes, ajoutez `xlim = c(valeurmin, valeurmax)` et `ylim = c(valeurmin, valeurmax)`
- Pour mettre en couleur, ajoutez `col= "red"` ou `col= "green"`, ...
- Pour ajouter une légende sur le graphique, ajoutez la ligne de code suivante (après exécution de la commande `plot()`) :
  `legend(xcoord, ycoord, "légende du graphique")`
  où `xcoord` et `ycoord` représentent les coordonnées x et y de l'emplacement de la légende. Ajoutez une légende est fortement conseillé lorsqu'on utilise plusieurs lignes sur le même graphique !

Par exemple :

```
plot(var1,  var2,  type = "l",  main = "mon premier graphique",
     xlab = "variable 1",  ylab = "variable 2",
     xlim = c(0, 10),  ylim = c(-100,  100),  col = "blue")
legend(5, 50, "in dollars", col = "blue")
```

Commandes utiles pour le TP1 :

Commande	Description de la commande
`getwd()`	donne le dossier dans lequel on est en train de travailler
`setwd()`	permet de changer de dossier dans lequel on est en train de travailler
`readRDS()`	lit un fichier RDS placé dans le bon dossier
`head()`	montre les six premières lignes d'un jeu de données
`tail()`	montre les six dernières lignes d'un jeu de données
`str()`	donne la structure d'un jeu de données
`attach()`	permet d'utiliser les noms de variables d'un jeu de données directement
`plot()`	crée un graphique (linéaire ou en nuage de points)
`length()`	donne la longueur d'un vecteur
`mean()`	donne la moyenne d'un vecteur
`median()`	donne la médiane d'un vecteur
`var()`	donne la variance d'un vecteur
`summary()`	donne un résumé (min, Q_1, médiane, moyenne, Q_3, max) d'un objet
`boxplot()`	crée une boîte à moustaches

Données "boysgirls" : garçons et filles nés entre 1629 et 1710 à Londres

Récupérez le fichier `boysgirls.RDS`, disponible sur webcampus, et enregistrez-le sur votre ordinateur dans un dossier "TP1". Créez un nouveau script `R` et enregistrez-le dans le même dossier. Le jeu de données fait référence au Dr John Arbuthnot, médecin, écrivain et mathématicien. Il s'intéressait à la proportion de garçons nouveau-nés par rapport aux filles nouveau-nées. Il a donc rassemblé les registres de baptême des enfants nés à Londres pour chaque année de 1629 à 1710.

Ouvrez le fichier sur Rstudio en utilisant la commande suivante :

```
data <- readRDS("boysgirls.RDS")
```

Rstudio va créer un nouvel objet appelé `data` qui contiendra les données présentes dans le fichier `boysgirls.RDS`. Attention, votre fichier doit se trouver dans le dossier correspondant au projet actif. Si vous avez un doute concernant la localisation de ce dossier, vous pouvez retrouver le chemin en tapant la commande `getwd()` pour voir où RStudio va chercher les fichiers.

Le jeu de données contient 82 observations sur 3 variables :

- `year` : l'année (entre 1629 et 1710)
- `boys` : le nombre de garçons baptisés chaque année
- `girls` : le nombre de filles baptisées chaque année

Si vous tapez `data` dans la console, `R` va afficher le jeu de données complet. On remarque que la première colonne de chiffres est simplement le numéro de la ligne. Les numéros de lignes ne font pas partie des données ; `R` les ajoute dans le cadre de son impression pour vous aider à faire des comparaisons visuelles.

Exercices :

1. Si le jeu de données est grand, il n'est pas toujours pratique d'afficher le jeu de données complet. Afficher les premières six données et les dernières six données avec `head(data)` et `tail(data)`. Examinez les données avez `str(data)`.
2. Que représente le chiffre obtenu en tapant `data[10,3]` ?
3. Montrez à présent uniquement le nombre de garçons baptisés et observez la manière dont les données sont présentées dans la console. *NB : voir la section "Comment sélectionner certaines données sur R ?" dans le document introductif "Installation de R et premiers pas".*
4. Pour utiliser directement les variables `boys` et `girls` sans passer par `data`, on peut taper `attach(data)`. Il suffit de le faire une seule fois par session. Utilisez cette commande et affichez de nouveau le nombre de garçons baptisés.
5. Créez maintenant un nuage de points entre les années et le nombre de baptêmes de filles. Y a-t-il une tendance apparente dans le nombre de filles baptisées au fil des ans ? Pour obtenir un graphique linéaire, répétez la même opération en reliant les points entre eux. Modifiez les noms des axes et ajoutez un titre au graphique. *NB : les graphiques apparaissent sous l'onglet "Plots" du panneau inférieur droit de Rstudio. Il est possible de passer d'un graphique à l'autre en cliquant sur les flèches gauches et droites apparaissant dans cette fenêtre. Il est également possible d'exporter un graphique en cliquant sur "Export".*
6. Créez une nouvelle variable `total` qui est égale à la somme du nombre total d'enfants baptisés (filles + garçons) pour chaque année. Ajoutez cette variable comme quatrième colonne au jeu de données.
7. Créez une nouvelle variable `prop` qui est la proportion de nouveau-nés qui sont des garçons puis faites un graphique de la proportion de garçons dans le temps. Changer la graduation de l'axe des ordonnées pour qu'il aille de 0 à 1. Qu'observez-vous ?
8. Sélectionnez les années pour lesquelles la proportion de garçons baptisés dépasse le 52%. Combien d'années trouve-t-on ?
9. Donnez la moyenne, la médiane et la variance du nombre de garçons et du nombre de filles baptisés. Affichez deux boîtes à moustaches (une pour les garçons et une pour les filles) et interprétez-les.