POLYTECH MAM3 - **Introduction à la statistique** **7 décembre 2023**

TP1 - STATISTIQUE DESCRIPTIVE AVEC R, PREMIÈRES NOTIONS

Pour chacun de vos travaux sur machine, choisissez votre éditeur de texte préféré ou plus simplement l'éditeur de Rstudio pour conserver l'ensemble de vos commandes.

Ne craignez pas d'être généreux en commentaires et en rigueur de programmation, c'est un investissement très rentable sur le long terme ! Idem pour le recours à l'aide de R via `?nom.procedure`.

Exercice 1. Un certain nombre de jeux de données issus de contextes divers sont disponibles par défaut sur **R**. Exécuter data() pour les charger.

(1) Exécuter le code suivant, et expliquer ce qu'il fait :

```
data()                          x2 = HairEyeColor
x1 = trees                      str(x2)
str(x1)                         x2[1,2,2]
x1$Girth                        x2[,2,2]
```

(2) Que contiennent les variables `x1` et `x2` ? Comparer et discuter les commandes `str` versus `summary`.

(3) Quels types de graphes vous semblent appropriés pour résumer graphiquement l'information contenue dans le jeu de données `x1` ? Justifiez. Même question pour le jeu de données `x2`.

(4) Tracer la boîte à moustaches (fonction `boxplot`) des valeurs du volume des arbres contenues dans `x1`. Commenter le graphe obtenu.

(5) Tracer l'histogramme (fonction `hist`) des valeurs du diamètre des arbres contenues dans `x1`. Commenter le graphe obtenu.

(6) Tracer la fonction de répartition empirique (fonction `ecdf`) des valeurs du diamètre des arbres contenues dans `x1`. Commenter le graphe obtenu.

(7) Construire un nouveau jeu de données `x3` contenant `x1`, et dans lequel la dernière valeur de diamètre (soit 20.6 pouces) a malencontreusement été modifiée en 206 pouces. Quel est l'effet sur les différents résumés (numériques et graphiques) précédents ?

(8) À l'aide de la fonction `pie`, tracer pour le jeu de données `x2` un diagramme en camembert de la répartition des couleurs des yeux chez les hommes aux cheveux bruns. Commenter le graphe obtenu.

(9) Utiliser la fonction `barplot` pour tracer le diagramme en bâton spécifiant la répartition des couleurs des yeux pour les femmes aux cheveux noirs. Ajouter un titre adéquat. Commenter le graphe obtenu.

(10) Représenter le diagramme en bâton spécifiant la répartition en terme de couleurs des yeux et des cheveux pour les hommes, en ajoutant titre et légende. Commenter le graphe obtenu.

(11) Représenter sur une même figure (commande `par(mfrow = ...)`) les diagrammes de répartition 'couleurs de cheveux' et 'couleurs des yeux' pour les hommes d'une part, et pour les femmes d'autre part. Sauvegarder ce graphe dans un fichier au format pdf.

Exercice 2. On reprend le jeu de données `x1` considéré dans l'exercice précédent, à savoir les données `trees` de **R**.

(1) Exécuter le code suivant, et expliquer ce qu'il fait :

```
x1 = trees
plot(x1$Girth,x1$Volume)
pairs(x1)
cor(x1)
cor(x1,method="kendall")
cor(x1,method="spearman")
```

(2) Effectuer une recherche sur internet pour trouver les définitions nécessaires.

Exercice 3. Prendre connaissance du jeu de données disponible sur **R** intitulé `iris`.

(1) Tracez le boxplot et l'histogramme des longueurs de pétales des 150 observations.
(2) Même question en différenciant les trois échantillons constitués par les trois espèces. On pourra pour cela étudier l'argument `formula` de la commande `boxplot`. Commenter.
(3) Comment illustrer la problématique soulevée dans la question précédente ?
(4) Discuter ce que ce type de phénomène illustre, et proposer des pistes de prise en compte en terme de modélisation.
(5) Avez-vous déjà entendu parler du *paradoxe de Simpson* ? Si oui, l'expliquer puis l'illustrer sur un exemple. Si non, effectuer une recherche internet pour ensuite l'expliquer et l'illustrer sur un exemple.