TP1_SEM

2025-09-27

Partie 1 QUIZ

Question 1 : 5. Aucune de ces réponses.

Le fichier est séparé par tabulations (sep = '\t'), a un en-tête (header = TRUE) et le séparateur décimal est la virgule dec = ','. La commande correcte en R est par exemple : read.table("data", header = TRUE, sep = "\t", dec = ","), la réponse 3 est la plus proche mais oublie dec = ","

Question 2 : 2. df_pats[df_pats$BP>17.5, c("gender", "Temp")]

La condition sur les lignes est df_pats$BP > 17.5. En seconde position on fournit les colonnes dans l'ordre voulu : c("gender","Temp").

Question 3. Réponse : 1. Corrélation positive forte.

Dans la figure A, quand X augmente, Y augmente aussi (nuage de points ascendant).

Question 4. Réponse : 3. Transformer un revenu en classes produit une variable ordinale.

Les autres propositions décrivent mal le type d'échelle. Par exemple pour la 1 si les données sont pris en celsius en convertissant en kelvin, on n'a pas le même ratio, pour la réponse 2 elle est mesurée en échelle ordinale. Pour la 4 (sauf exeption) il n'y a aucun sens a faire une moyenne et si on le faisait on obtiendra par exemple 2,3 qui n'a aucun sens par rapport à notre variable.

Question 5. Réponse : 4. 100% des clients sont moyennement satisfaits.

Moyenne = 1 et écart-type = 0 signifie que toutes les réponses valent 1 (exeption de la Question 4).

Question 6. Réponse : 4. 25% des fleurs ont une largeur de sépale > 3,3 et 25% $\leq 2,8$.

Le boxplot montre la médiane autour de 3 (et non la moyenne, aucune information sur la moyenne), la boîte (Q1–Q3) environ [2.8, 3.3]; il y a des points isolés au-dessus et en dessous. L'option 4 décrit correctement les quartiles et car Q1 2.8 et Q3 3.3. Les autres affirmations sont fausses.

Question 7. Réponse : 2. 44,68% des hommes sont fumeurs.

Taux des hommes fumeurs : $21/47 \approx 44,68\%$. Taux d'hommes dans l'échantillion : 47 Taux des femmes fumeurs : $13/40 = 32,5$.

Question 8. Réponse : 4. Skewness positive $\Rightarrow$ mode < moyenne.

Dans une distribution asymétrique à droite : mode < médiane < moyenne.

Question 9. Réponse : 1. La valeur moyenne est 6,10.

La somme est 61, divisée par 10 observations = 6,1. Médiane = 4, mode = 3.

Question 10. Réponse : le code calcule la **médiane**.

Si n est pair, on prend la moyenne des deux valeurs centrales ; sinon, on prend l'élément du milieu.

Partie 2 Initiation à la language R

```
#Exercice 1
a=c(1,2,3)
b=c(4,5,6)
c=c(7,8,9)
a
```

```
## [1] 1 2 3
```

```
b
```

```
## [1] 4 5 6
```

```
c
```

```
## [1] 7 8 9
```

```
mat=cbind(a,b,c)
mat
```

```
##      a b c
## [1,] 1 4 7
## [2,] 2 5 8
## [3,] 3 6 9
```

```
#Exercice 2
table0 = read.table("~/Table0.txt")
table0
```

```
##         V1 V2  V3 V4 V5
## 1     Alex 25 177 57  F
## 2    Lilly 31 163 69  F
## 3     Mark 23 190 83  M
## 4   Oliver 52 179 75  M
## 5   Martha 76 163 70  F
## 6    Lucas 49 183 83  M
## 7 Caroline 26 164 53  F
```

```
#a) Changement du nom des colonnes en Nom, Age, Taille, Poids et Sexe
colnames(table0) <- c("Nom", "Age", "Taille", "Poids", "Sexe")
#b) Changement du nom des lignes par les noms puis supprimer la colonne Nom
rownames(table0) <- table0$Nom
table0$Nom = NULL
table0
```

```
##          Age Taille Poids Sexe
## Alex      25    177    57    F
## Lilly     31    163    69    F
## Mark      23    190    83    M
```

```
## Oliver    52    179    75    M
## Martha    76    163    70    F
## Lucas     49    183    83    M
## Caroline  26    164    53    F
```

```
#Exercice 3
table1 = read.table("~/Table1.txt")
table1
```

```
##         V1  V2     V3     V4  V5
## 1     Name Age Height Weight Sex
## 2     Alex  25    177     57   F
## 3    Lilly  31    163     69   F
## 4     Mark  23    190     83   M
## 5   Oliver  52    179     75   M
## 6   Martha  76    163     70   F
## 7    Lucas  49    183     83   M
## 8 Caroline  26    164     53   F
```

```
#a) Il contient 8 lignes et 5 colonnes
#b) Changement du tableau pour avoir le même tableau que Exercice 2 b) (sauf ici nom des colonnes en an
table1 =read.table("~/Table1.txt", header = TRUE, stringsAsFactors = FALSE)
rownames(table1) <- table1$Name
table1$Name = NULL
table1
```

```
##          Age Height Weight Sex
## Alex      25    177     57   F
## Lilly     31    163     69   F
## Mark      23    190     83   M
## Oliver    52    179     75   M
## Martha    76    163     70   F
## Lucas     49    183     83   M
## Caroline  26    164     53   F
```

```
#Exercice 4
df <- read.csv("~/Cereals.csv", stringsAsFactors = FALSE)
head(df)
```

```
##                 Cereal.name Supplier Cold.or.Hot calories protein fat sodium
## 1                 100%_Bran        N           C       70       4   1    130
## 2          100%_Natural_Bran        Q           C      120       3   5     15
## 3                   All-Bran        K           C       70       4   1    260
## 4 All-Bran_with_Extra_Fiber        K           C       50       4   0    140
## 5             Almond_Delight        R           C      110       2   2    200
## 6   Apple_Cinnamon_Cheerios        G           C      110       2   2    180
##   fiber carbo sugars potass vitamins   rating
## 1  10.0   5.0      6    280       25 68.40297
## 2   2.0   8.0      8    135        0 33.98368
## 3   9.0   7.0      5    320       25 59.42551
## 4  14.0   8.0      0    330       25 93.70491
## 5   1.0  14.0      8     NA       25 34.38484
## 6   1.5  10.5     10     70       25 29.50954
```

```r
str(df)
```

```
## 'data.frame':    77 obs. of  13 variables:
##  $ Cereal.name: chr  "100%_Bran" "100%_Natural_Bran" "All-Bran" "All-Bran_with_Extra_Fiber" ...
##  $ Supplier   : chr  "N" "Q" "K" "K" ...
##  $ Cold.or.Hot: chr  "C" "C" "C" "C" ...
##  $ calories   : int  70 120 70 50 110 110 110 130 90 90 ...
##  $ protein    : int  4 3 4 4 2 2 2 3 2 3 ...
##  $ fat        : int  1 5 1 0 2 2 0 2 1 0 ...
##  $ sodium     : int  130 15 260 140 200 180 125 210 200 210 ...
##  $ fiber      : num  10 2 9 14 1 1.5 1 2 4 5 ...
##  $ carbo      : num  5 8 7 8 14 10.5 11 18 15 13 ...
##  $ sugars     : int  6 8 5 0 8 10 14 8 6 5 ...
##  $ potass     : int  280 135 320 330 NA 70 30 100 125 190 ...
##  $ vitamins   : int  25 0 25 25 25 25 25 25 25 25 ...
##  $ rating     : num  68.4 34 59.4 93.7 34.4 ...
```

```r
#Nous avons un tableau nutritionnel décrivant 77 céréales différents par rapport 13 variables comme le 
#a) Ajouter une nouvelle variable "totalcarb" = carbo + sugars
df$totalcarb <- df$carbo + df$sugars
#b) Nombre de céréales "hot"
sum(df$Cold.or.Hot == "H")
```

```
## [1] 3
```

```r
#c) Nombre de fournisseurs uniques
length(unique(df$Supplier))
```

```
## [1] 7
```

```r
#d) Sous-ensemble uniquement fournisseur "K" (Kellogg's)
df_K <- subset(df, Supplier == "K")

#e) Sous-ensemble : moins de 80 calories ET plus de 20 vitamines
df_calvi <- subset(df, calories < 80 & vitamins > 20)

#f) Sous-ensemble : au moins 1 sucre en ne gardant que "Cereal.name", "calories", "vitamins"
df_sugar <- subset(df, sugars >= 1, select = c(Cereal.name, calories, vitamins))
head(df_sugar)
```

```
##               Cereal.name calories vitamins
## 1               100%_Bran       70       25
## 2       100%_Natural_Bran      120        0
## 3                All-Bran       70       25
## 5          Almond_Delight      110       25
## 6 Apple_Cinnamon_Cheerios      110       25
## 7             Apple_Jacks      110       25
```

```r
#g) Sauvegarder un sous-ensemble en CSV (df_sugar)
write.csv(df_sugar, "Cereals_sugar.csv", row.names = FALSE)

#h) Renommer la colonne "Supplier" en "Producteur"
names(df)[names(df) == "Supplier"] <- "Producteur"
```

```r
#Exercice 5
data("islands")
# Nombre total d'observations
length(islands)
```

```
## [1] 48
```

```r
# Description jeu  de données :
# 'islands' est un vecteur nommé : noms = îles/continents, valeurs = superficie (en milliers de miles²)
head(islands)
```

```
##       Africa   Antarctica         Asia    Australia Axel Heiberg       Baffin
##        11506         5500        16988         2968           16          184
```

```r
# Mesures de tendance centrale
mean(islands)     # Moyenne
```

```
## [1] 1252.729
```

```r
median(islands)   # Médiane
```

```
## [1] 41
```

```r
# Range (minimum et maximum)
range(islands)                          # min et max
```

```
## [1]    12 16988
```

```r
islands[which.max(islands)]             # plus grande île
```

```
##  Asia
## 16988
```

```r
islands[which.min(islands)]             # plus petite île
```

```
## Vancouver
##        12
```

```r
# Mesures de dispersion
sd(islands)              # Écart-type
```

```
## [1] 3371.146
```

```r
quantile(islands, probs = c(0, 0.25, 0.5, 0.75, 1))     # 0%, 25%, 50%, 75%, 100%
```

```
##       0%      25%      50%      75%     100%
##    12.00    20.50    41.00   183.25 16988.00
```

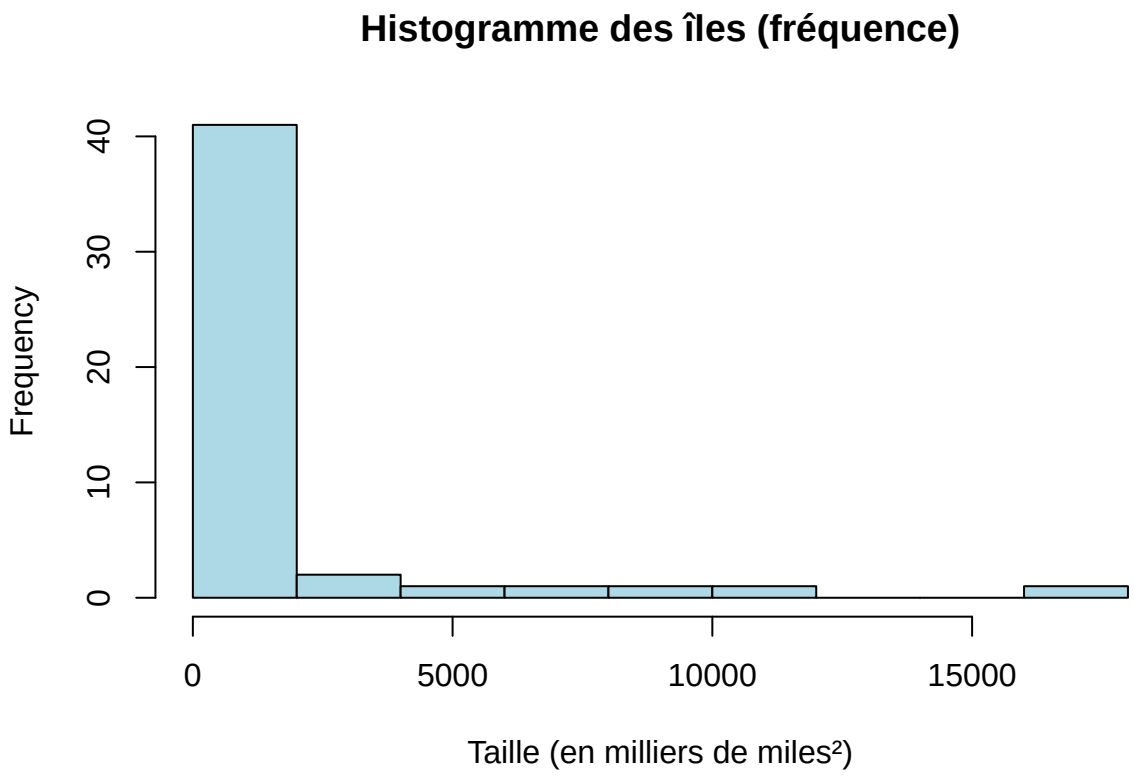
```r
quantile(islands, probs = c(0.005, 0.95))              # 0.5% et 95%
```

```
##     0.5%      95%
##   12.235 8481.750
```

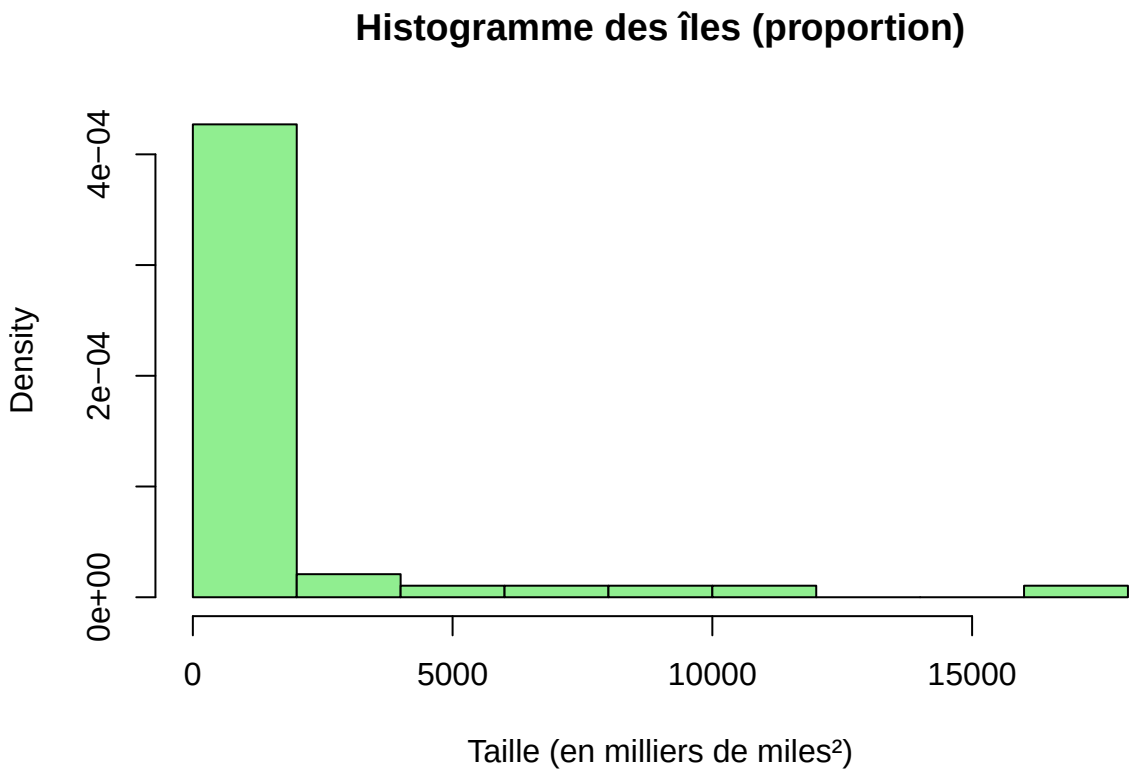
```r
# Intervalle interquartile
IQR(islands)              #Quantile(0,75) - Quantile(0,25)
```

```
## [1] 162.75
```

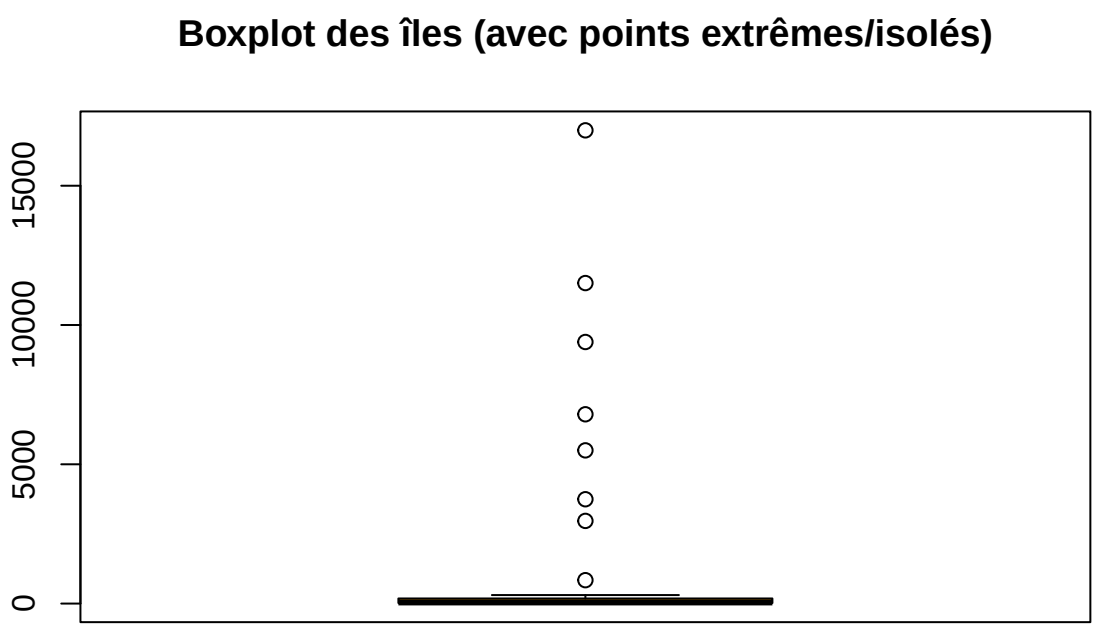
```r
# Histogramme
# A. Fréquence
hist(islands, main="Histogramme des îles (fréquence)", xlab="Taille (en milliers de miles²)", col="ligh
```



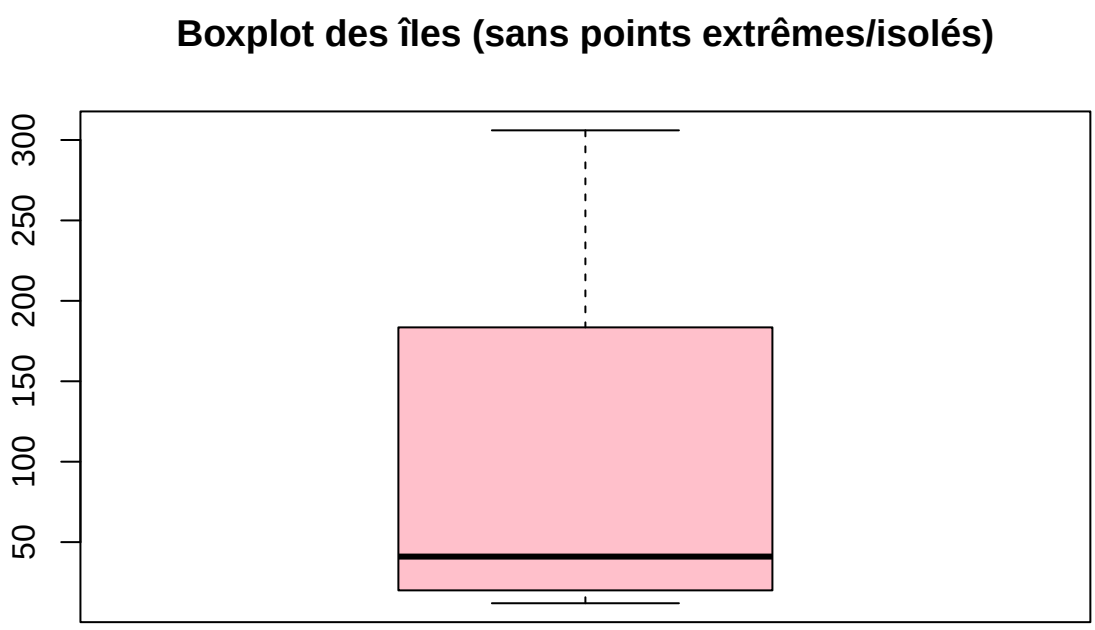

```r
# B. Proportion
hist(islands, main="Histogramme des îles (proportion)", xlab="Taille (en milliers de miles²)", col="ligh
```




```r
# Diagrammes en boîte
# C. Avec valeurs aberrantes
boxplot(islands, main="Boxplot des îles (avec points extrêmes/isolés)", col="orange")
```




```r
# D. Sans valeurs aberrantes
boxplot(islands, outline=FALSE, main="Boxplot des îles (sans points extrêmes/isolés)", col="pink")
```




```
#Exercice 6
sales = read.csv("~/yearly_sales.csv", stringsAsFactors = FALSE)

# Créer la nouvelle variable
sales$spender <- cut(sales$sales_total, breaks = c(-Inf, 100, 500, Inf), labels = c("small", "medium",

# Vérifier le résultat
head(sales)
```

```
##   cust_id sales_total num_of_orders gender spender
## 1  100001      800.64             3      F     big
## 2  100002      217.53             3      F  medium
## 3  100003       74.58             2      M   small
## 4  100004      498.60             3      M  medium
## 5  100005      723.11             4      F     big
## 6  100006       69.43             2      F   small
```

```
str(sales)
```

```
## 'data.frame':    10000 obs. of  5 variables:
##  $ cust_id      : int  100001 100002 100003 100004 100005 100006 100007 100008 100009 100010 ...
##  $ sales_total  : num  800.6 217.5 74.6 498.6 723.1 ...
##  $ num_of_orders: int  3 3 2 3 4 2 2 2 2 2 ...
##  $ gender       : chr  "F" "F" "M" "M" ...
##  $ spender      : Ord.factor w/ 3 levels "small"<"medium"<..: 3 2 1 2 3 1 1 1 2 1 ...
```