Statistiques avec le logiciel R - TD4

Objectifs

— Etre capable de réaliser une analyse descriptive d'une variable quantitative continue avec **R**.
— Etre capable d'étudier le lien entre deux variables.
— Utiliser la fonction `plot` pour produire un diagramme de dispersion ou tracer une courbe.

1 Statistiques descriptives

Pour illustrer ces propos nous utiliserons les données 'iris', directement téléchargeable sur **R** à partir du package `datasets`.

```
> #install.packages('datasets') #Pour telecharger le package depuis un site miroir du CRAN
> library(datasets) #pour charger le package
> iris=iris #pour charger la table iris
```

1.1 Table des effectifs

On peut obtenir la table des effectifs d'une variable qualitative sous forme de `factor`, avec la fonction `table`.

```
> table(iris$Species)
    setosa versicolor  virginica
        50         50         50
```

Lorsqu'on donne en entrée de la fonction plusieurs facteurs, on obtient la table des effectifs croisés.

1.2 Résumés numériques

Dans cette partie, on présente des résumés numériques d'un échantillon d'observations $(x_1, x_2, \cdots, x_n)$ de taille n. Le variable considérée ici est la taille de sépales d'iris. Le vecteur d'observations peut facilement être extrait de la table `iris` avec l'instruction `sepal=iris$Sepal`, ou

```
> sepal=iris[,1]
```

1.2.1 Résumés de positions de la distribution

La médiane

La médiane d'un échantillon statistique est la valeur m_e de la variable X qui sépare l'échantillon ordonné dans l'ordre croissant en deux parties de même effectif.

— Lorsque n est impair, la médiane est la valeur située à la position $\frac{n+1}{2}$
— Lorsque n est pair, la médiane est la moyenne entre la valeur située à la position $\frac{n}{2}$ et $\frac{n}{2}+1$.

La médiane est obtenue avec la fonction `median(x)`.

```
> median(sepal)
[1] 5.8
```

La moyenne

Elle est obtenue avec la la fonction `mean`

```
> mean(sepal)
[1] 5.843333
```

Les fractiles

Le fractil d'ordre p $(0 < p < 1)$ est la valeur q_p de la variable X qui coupe l'échantillon en deux groupes. L'un composé des observations inférieures à q_p et qui représente $p\%$ des observations de l'échantillon, et l'autre composé des observations supérieures à q_p, et représentant $(1-p)\%$ des observations. Les fractiles s'obtiennent avec la fonction *quantile*(), on ajoutera comme argument un vecteur spécifiant l'ordre des quantiles voulus.

```
> quantile(sepal,probs=c(0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95))
```

```
   5%   10%   25%   50%   75%   90%   95%
4.600 4.800 5.100 5.800 6.400 6.900 7.255
```

1.2.2 Résumés de dispersion d'une distribution

Il en existe plusieurs sortes, tous n'ont pas de fonctions prédéfinies sur **R**, mais peuvent être obtenus facilement à partir d'autres fonctions. Le tableau ci-dessous en donne quelques-uns.

Nom	Défintion	Appel	Résultat
Etendue	$\max_i x_i - \min_i x_i$	`max(sepal)-min(sepal)`	3.6
Intervalle inter-quartiles	$q_{3/4} - q_{1/4}$	`IQR(sepal)`	1.3
Variance	$\sigma^2_{pop} = \frac{1}{n-1}\sum_{i=1}^n (x_i - \overline{x})^2$	`var(sepal)`	0.69
Ecart type	$\sigma_{pop} = \sqrt{\frac{1}{n-1}\sum_{i=1}^n (x_i - \overline{x})^2}$	`sd(sepal)`	0.83

1.2.3 Résumés de forme d'une distribution

Ces résumés ne peuvent être calculés que pour une variable continue. Le coefficient d'asymétrie (*skewness*) donne une information sur la symétrie ou l'asymétrie de la distribution, il est égal au moment d'ordre 3 de la variable X centrée et réduite

$$\gamma_1 = \mathbf{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right].$$

— Lorsque $\gamma_1 < 0$, la queue de distribution à gauche est plus longue qu'à droite,
— au contraire, si $\gamma_1 > 0$, la queue de distribution à droite est plus longue qu'à gauche.
— Enfin, si $\gamma_1 = 0$, la distribution est symétrique.

Le coefficient d'aplatissement (*kurtosis*) informe sur le côté piqué ou aplati de la distribution, il est égal au moment d'ordre 4 de la variable X centrée et réduite :

$$\beta_2 = \mathbf{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right].$$

Celui-ci est égal à 3 pour la distribution gaussienne, il est supérieur à 3 si la distribution est plus "pointue" que la distribution gaussienne et inférieure à 3 la distribution est plus plate que la distribution gaussienne. Vous pourrez aussi rencontrer le coefficient d'aplatissement normalisé, égal à $\beta_2 - 3$. Ce dernier à l'avantage de valoir 0 pour une distribution gaussienne.

Des fonctions *kurtosis* et *skewness* n'existent pas dans la base de **R**, mais il est aisé facile de les créer en utilisant des estimateurs empiriques.

```
> skewness=function(x){return(mean((x-mean(x)/sd(x))^3))}
> kurtosis=function(x){return(mean((x-mean(x)/sd(x))^4))}
> skewness(sepal)
[1] -4.089865
> kurtosis(sepal)
[1] 8.457813
```

1.3 Mesure d'association entre deux variables

On peut mesurer l'association entre deux variables continues $(x_i)_{i=1,\cdots,n}$ et $(y_i)_{i=1,\cdots,n}$ avec la covariance ou le coefficient de corrélation de Pearson. On étudiera ici, la relation entre la longueur des sépales, et la longueur des pétales, nommée `petal`.

```
> petal=iris[,3]
```

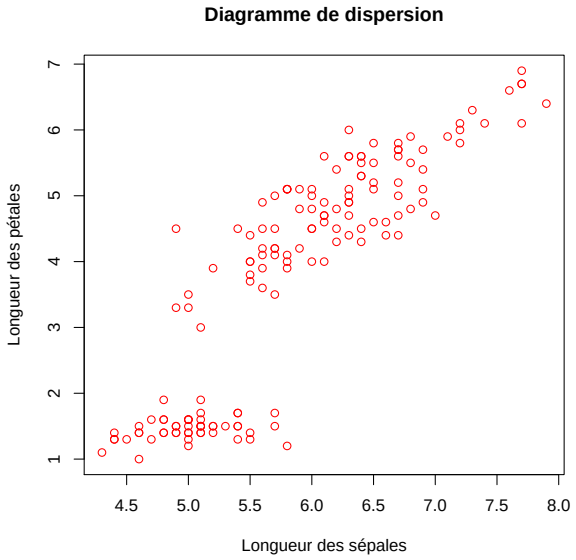
Covariance	Coefficient de corrélation
$cov_{pop} = \frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})$	$\rho_{pop} = \frac{\frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^n (y_i - \overline{y})^2}}$
`> cov(sepal,petal)`	`> cor(sepal, petal)`
`[1] 1.274315`	`[1] 0.8717538`

2 Représentation graphique de données quantitatives

2.1 Diagramme de dispersion

Le diagramme de dispersion d'une variable relativement à une autre s'obtient avec la fonction `plot`. La fonction s'utilise toujours de la façon suivante `plot(x,y)`, ou `x` est la variable figurant en abscisse et `y` la variable en ordonnée. On notera que `x` et `y` doivent être deux vecteurs contenant exactement le même nombre de données, et de type numriques pour obtenir un graphe de dispersion.

```
> plot(sepal,petal, xlab='Longueur des sépales', ylab='Longueur des pétales',
+ main='Diagramme de dispersion', col='red')
```



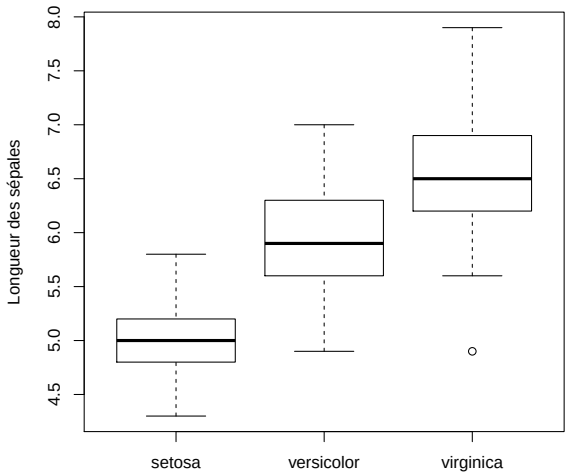
Remarque : La fonction `par` permet de gérer l'ensemble des paramètres graphiques : nombre de graphiques par page, positionnement des axes, taille des marges, taille des polices, ...

2.2 Boîte à moustaches (*boxplot*)

Avec une boîte à moustaches, on pourra représenter les principaux fractiles de la distribution sur un même diagramme. La ligne au centre de la boite représente la médiane, alors que les bords de la boîte sont formés par les 1^{er} et 3^{e} quartiles. Les valeurs situées à l'extérieur de la boîte, mais en deçà de 1.5 fois l'intervalle interquartile sont dîtes adjacentes. L'extrémité de la moustache supérieure correspond à la plus grande valeur adjacente, alors que l'extrémité inférieure correspond à la plus petite valeur adjacente. Les cercles à l'extérieur de la boîte à moustaches figurent les valeurs extrêmes qui se situent au-delà de 1.5 fois l'intervalle interquartile, et qui sont susceptibles d'être aberrantes.

En disposant côte à côte plusieurs boîtes à moustache, il devient aisé de comparer des distributions selon un attribut. Dans l'exemple ci-dessous, on produit pour chacune des trois espèces d'Iris (indiqué dans la colonne `species`) une boîte à moustache de la longueur des sépales.

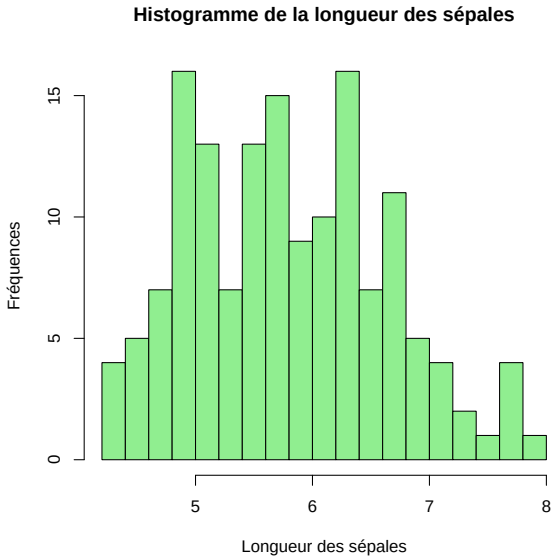
```
> boxplot(Sepal.Length~Species,data=iris, ylab='Longueur des sépales')
```



2.3 Histogramme

Un histogramme est un diagramme composé de rectangles contigus dont les aires sont proportionnelles aux effectifs (ou aux fréquences) et dont les bases sont déterminées par les intervalles de classes. Sur **R** il s'obtient avec la fonction `hist`, on pourra spécifier les intervalles de classes, ou seulement le nombre de classes à l'aide de l'option `breaks`.

```
> hist(sepal,breaks=20,main='Histogramme de la longueur des sépales',
+ xlab='Longueur des sépales', ylab='Fréquences', col='lightgreen')
```



3 Représentation graphique de données qualitatives

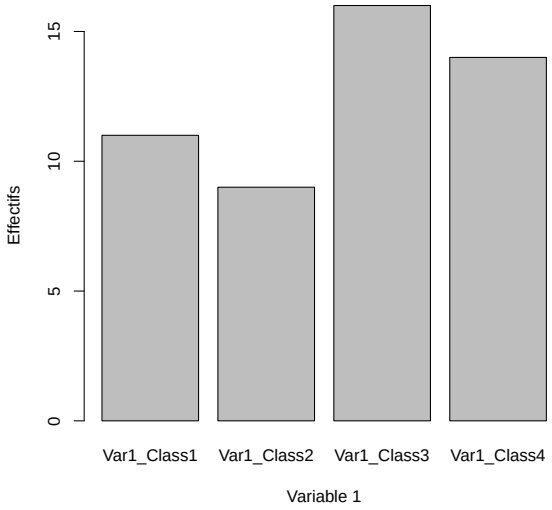
3.1 Représentation sous forme de diagrammes en barres

En utilisant directement le tableau individus $\times$ variables

Lorsqu'une variable est définie comme **factor** la fonction **plot** produit un *diagramme en barres* des effectifs de chaque niveau. Les arguments `ylab` et `xlab` permettent de nommer les axes.

```
> data=data.frame(var1=sample(paste("Var1_Class", c(1,2,3,4), sep=""),50,replace=TRUE),
+ var2=sample(paste("Var2_Class",c(1,2,3),sep=""),50,replace=TRUE))
> head(data)
         var1        var2
1 Var1_Class2 Var2_Class3
2 Var1_Class4 Var2_Class1
```

```
3 Var1_Class1 Var2_Class3
4 Var1_Class1 Var2_Class2
5 Var1_Class4 Var2_Class2
6 Var1_Class1 Var2_Class2
> data$var1=as.factor(data$var1)
> data$var2=as.factor(data$var2)
> plot(data$var1,xlab="Variable 1", ylab="Effectifs")
```



En utilisant la table des effectifs

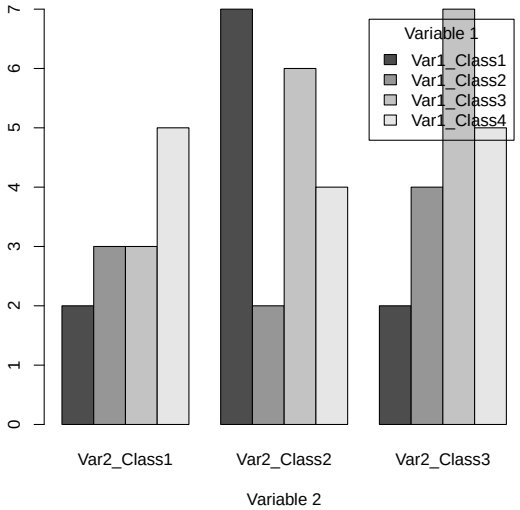
Dans l'exemple ci-dessus la table ne contient pas directement les effectifs mais donne pour chaque individu sa classe. La fonction `plot` se charge donc de compter les effectifs dans chaque classe avant des les afficher. La fonction `barplot` permet quant à elle d'obtenir directement un diagramme en barres à partir des effectifs.

A partir de la table `data`, nous pouvons obtenir la table des effectifs avec la fonction `table(vec1,vec2,...)`. Lorsqu'on spécifie deux vecteurs ou plus, `table` renvoie une matrice des effectifs croisés.

```
> barplot(table(data$var1))
```

On peut aussi entrer une matrice, dans ce cas nous obtenons les effectifs croisés pour chacune des deux variables.

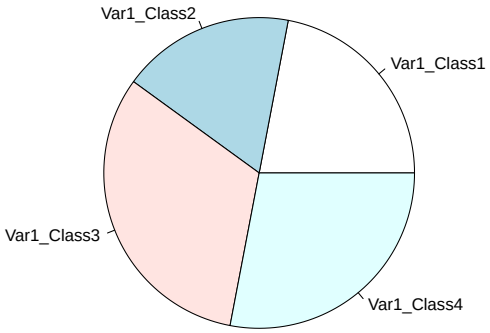
```
> tab=table(data$var1,data$var2)
> barplot(tab,beside=TRUE,legend.text=row.names(tab),
+ xlab='Variable 2',args.legend=list(title='Variable 1'))
```



3.2 Représentation des effectifs sous forme de diagrammes circulaires

Enfin avec la fonction `pie`, on peut obtenir un diagramme circulaire sur les proportions de chacune des classes d'une variable quantitative. Comme pour la fonction `barplot` il faut donner la **table des effectifs**.

```
> pie(table(data$var1))
```



4 Exercices

Exercice 1 : Lecture de la table de données

1. Télécharger et enregistrer la table de données `Olympics_data.csv` dans un nouveau répertoire dédié à ce TD. Créer un nouveau script R, et enregistrer le, vous répondrez aux questions de ce TD dessus.
2. Ouvrir la table de données avec un éditeur de texte, repérer sa structure (séparateur de colonnes, de décimales, nombres de lignes à sauter en début de fichier...).
3. Sur **R**, lire la table de données avec la fonction `read.table`, on enregistrera la table dans une variable nommée `olymp`
4. Donner le nombre de lignes et de colonnes de `olymp`.
5. A l'aide du fichier `Olympic_doc.txt` donner la signification de chaque colonne. En déduire le type de donnée de chaque colonne. Vérifier qu'il correspond bien au type enregistré, obtenu avec la fonction `typeof`. Si ce n'est pas le cas, convertissez le type de donnée.
6. Sur les lignes sont représentés les individus. Qu'est ce qu'un individu dans cet exemple ?

Exercice 2 : Etude descriptive de deux variables quantitatives

Dans cet exercice vous ferez une étude descriptive des variables `Income` et `BordaPoints`.

Cette dernière variable donne un classement des nations selon le nombre de médailles d'or, d'argent et de bronze obtenues, mais tiens aussi compte des disciplines dans lesquelles ces médailles ont été obtenues, puisque certaines disciplines comme la gymnastique ou la natation délivrent d'avantage de médailles que d'autres, comme le football.

Pour chacune des deux variables étudiées, vous donnerez les statistiques numériques standard de position, dispersion et forme de la distribution. Vous noterez aussi les données extrêmes. Finalement, vous résumerez les caractéristiques de chaque distribution en quelques phrases, et vous illustrerez vos propos par des graphiques appropriés.

Exercice 3 : Etude descriptive du lien entre deux variables

Dans cette exercice on cherchera à identifier le lien entre la variable `BordaPoints` représentant la réussite de la nation au J.O. et les autres variables quantitatives socio-économiques.

1. Pour cela, vous calculerez le coefficient de corrélation entre la variable `BordaPoints` et les variables `Income`, `PopnSize` et `GDP`. Qu'en concluez-vous ?
2. Vous représenterez aussi le lien entre deux variables, en réalisant le nuage de points dont l'abscisse est la variable socio-économique et l'ordonnée le score Borda.
3. L'association est-elle meilleure lorsqu'on opère une transformation de la variable socio-économique ? Vous testerez les transformations carré, cube et logarithme. Qu'en concluez-vous ?

Exercice 4 : Représentation graphique de la courbe gaussienne

1. Quelles sont les quantiles à l'ordre 0.025 et 0.0975 de la distribution normale centrée réduite. On cherche à représenter graphique la fonction de distribution de la densité normale. Proposer un intervalles de valeurs pour `x` (absisse).
2. Créez un vecteur `x` en échantillonnant de manière régulière dans cet intervalle.
3. La fonction `dnorm(x,mean=mu,sd=sigma)` donne la densité de probabilité en `x` de la loi normale d'espérance `mu` et d'écart-type `sigma`. Déterminez la densité de probabilité pour chacun des `x` échantillonnés.
4. Tracer la densité de probabilité de la loi normale. Vous soignerez la présentation du graphique.