TP3 : les iris de Fisher

Chargement des données

Importez le fichier `N:l2ie/ad/iris.txt` dans votre dossier personnel AD (disque H).
Si vous travaillez sur votre machine personnelle vous trouverez ce fichier dans le dossier : `https://share-etu.istic.univ-rennes1.fr/l2ie/ad`.

Ouvrez ce fichier : nous avons ici la description de 150 iris appartenant à trois espèces : 50 setosa, 50 virginica, et 50 versicolor. Sur chaque iris on a mesuré en centimètres la longueur et la largeur des sépales, ainsi que la longueur et la largeur des pétales. L'objectif du TP est de savoir quelle variable discrimine le mieux les espèces.

Lancez R et placez-vous dans votre dossier AD (avec Fichier →Changer le répertoire courant).
Rappel : tout ce que vous sauvegardez sur le disque C ou sur le "bureau" des machines en salle de TP est perdu dès que vous vous déconnectez.
Pour vérifier le nom du répertoire courant : `getwd()`.

Ouvrez un nouveau script, inscrivez `#<votre nom>` sur la première ligne du script, puis tapez `iris<-read.table("iris.txt",header=T,dec=",")` dans ce script pour charger le data frame dans un objet "iris" de R (l'argument `header` signifie que la première ligne contient les noms des variables, et `dec` signale que la décimale est notée par une virgule dans le fichier de données), puis `iris` pour l'afficher :

```
> # TP3 de Jean Moulin
> iris<-read.table("iris.txt",header=T,dec=",")
> iris
    SeLo SeLa PeLo PeLa    Espece
1    5.1  3.5  1.4  0.2    setosa
2    4.9  3.0  1.4  0.2    setosa
3    4.7  3.2  1.3  0.2    setosa
...
```

Vous taperez toutes vos commandes dans ce script et déposerez ce script sur Moodle à la fin du TP.

Quelques rappels

- Pour ouvrir une nouvelle fenêtre graphique : `x11()`
- Pour avoir plusieurs graphes dans une même fenêtre graphique, on utilise la fonction "par" : par exemple avec `par(mfrow=c(2,3))`, le prochain tracé donnera alors 6 graphiques dans la fenêtre (2 en ligne et 3 en colonne).
- Pour accéder à une variable, utiliser $ ou le repérage matriciel (par exemple `iris$SeLo` ou `iris[,1]` pour la longueur de sépale).
- Sauvegardez votre script régulièrement.
  On peut exécuter tout le script d'un coup, ou seulement la ligne courante par Ctrl+R.

Analyse des données

- Calculez les corrélations entre les variables quantitatives par la commande `cor`.
  Par exemple :
  `cor(iris$SeLo,iris$SeLa)` pour longueur et largeur de sépales.

  On peut aussi afficher toute la matrice de corrélations : `cor(iris[,-5])` ou `cor(iris[,1:4])`
  Quelles sont les variables les mieux corrélées ? Qu'est-ce que cela signifie ?
- Représentez la variable qualitative *Espece* sur un diagramme en camembert.
- Faites un histogramme de chacune des variables quantitatives. Mettez les 4 graphiques dans la même fenêtre, mettez des titres et des couleurs. Commentez les résultats.
- Faites afficher les statistiques sommaires des variables, globalement puis par espèce (fonction `summary`).
- Pour chacune des variables quantitatives, faites un histogramme par espèce et mettez l'ensemble sur un même graphique (16 histogrammes). Que remarquez-vous ?
  Quelle variable permet de caractériser au mieux les différentes espèces ? Faites afficher tous les iris classés selon cette variable.
- Représentez les boîtes à moustaches de chaque variable quantitative par espèce (commande `boxplot`).
  Par exemple pour la variable longueur de sépale : `boxplot(iris$SeLo~iris$Espece)`
  Mettez les 4 graphiques dans la même fenêtre, mettez des couleurs et des titres.
  Commentez les résultats. Si vous devez discriminer entre espèces, quelle(s) variable(s) suggérez-vous d'utiliser ?
- On peut également s'intéresser aux relations entre deux variables : représentez les nuages de points obtenus en prenant les variables quantitatives deux par deux : commande `plot(iris[,-5])`.
  On les appelle diagrammes de dispersion.
  Pour visualiser les espèces (c'est-à-dire les modalités de la variable qualitative), vous ajouterez `col=as.numeric(iris$Espece)` dans le `plot`. Commentez.

  Réalisez le diagramme de dispersion avec seulement les deux variables les plus pertinentes.
  Par exemple, la commande pour représenter le diagramme de dispersion pour les deux variables longueur et largeur de sépale, avec des couleurs différentes pour les trois espèces, sera :
  `plot(iris$SeLo,iris$SeLa,col=as.numeric(iris$Espece))`

  Ajoutez un titre et une légende. On peut aussi modifier l'aspect des points avec le paramètre `pch`.

==========

En guise de conclusion, soulignons le fait que les représentations graphiques sont une étape fondamentale dans la connaissance des données et que le logiciel R est un excellent outil. Les représentations graphiques sont là pour éclairer la nature des données et guider notre réflexion.