

Iris dataset

Justine Guégan

7 mars 2017

Ce document présente l'analyse du jeu de données `iris`, interne à R.

(Il existe de multiples jeu de données distribués avec R. Pour les découvrir, tapez `data()` dans R. Pour charger un jeu de données, tapez `data(nomDuJeuDeDonnees)`. Pour obtenir des informations sur un jeu de données, tapez `?iris`).

Chargement et découverte du jeu de données `iris`

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa

`iris` donne les mesures en centimètres de la longueur et largeur des sépales et pétales de 150 fleurs provenant de 3 espèces d'iris. Les espèces sont *Iris setosa*, *versicolor*, and *virginica*. La répartition des fleurs par espèce est la suivante :

```
##
##      setosa versicolor  virginica
##         50         50         50
```

Dans *RStudio*, vous pouvez visualisez l'ensemble du jeu de données `iris` grâce à la commande `View(iris)`

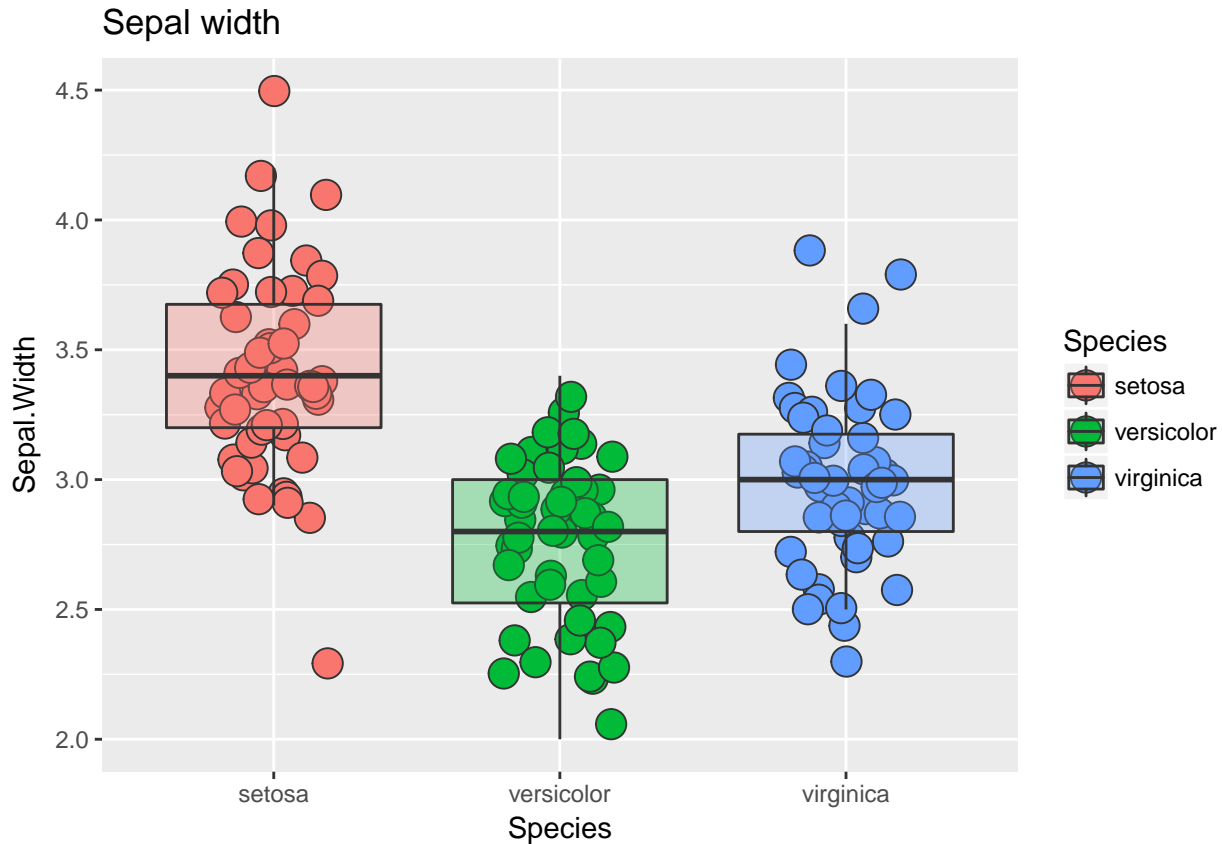
Analyses

Une commande très intéressante afin d'avoir une vue statistique des données est la commande `summary()`

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
Min. :4.300	Min. :2.000	Min. :1.000	Min. :0.100	setosa :50
1st Qu.:5.100	1st Qu.:2.800	1st Qu.:1.600	1st Qu.:0.300	versicolor:50
Median :5.800	Median :3.000	Median :4.350	Median :1.300	virginica :50
Mean :5.843	Mean :3.057	Mean :3.758	Mean :1.199	NA
3rd Qu.:6.400	3rd Qu.:3.300	3rd Qu.:5.100	3rd Qu.:1.800	NA
Max. :7.900	Max. :4.400	Max. :6.900	Max. :2.500	NA

Graphique

On souhaite étudier la largeur des sépales des 3 espèces. Pour cela, une représentation adéquate est le boxplot.



Visuellement, on peut voir que la largeur des sépales diffère entre les 3 classes. Cette différence est-elle significative ?

Statistique

Pour répondre à la question précédente, nous allons faire un test de comparaison de moyennes, appelé t-test (ou test de Student).

```
##
##  Welch Two Sample t-test
##
## data:  iris$Sepal.Width[which(iris$Species == "setosa")] and iris$Sepal.Width[which(iris$Species ==
## t = 9.455, df = 94.698, p-value = 2.484e-15
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.5198348 0.7961652
## sample estimates:
## mean of x mean of y
##      3.428      2.770
```

On peut voir que le test pour les longueurs de sépales entre setosa et versicolor est très significatif : la pvalue est inférieur à 10^{-16} . En est-il de même entre versicolor et virginica ?

```
##
## Welch Two Sample t-test
##
## data: iris$Sepal.Width[which(iris$Species == "versicolor")] and iris$Sepal.Width[which(iris$Species
## t = -3.2058, df = 97.927, p-value = 0.001819
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.33028364 -0.07771636
## sample estimates:
## mean of x mean of y
##      2.770      2.974
```

La largeur des sépales est significative entre versicolor et setosa mais de manière bien moindre puisque la pvalue est cette fois égale à 0.0018195.

