16/09/2018

Analyse de données

TP2 : Statistiques descriptives bivariées

Numa Benamer et Marie Dénès

3. Statistiques Descriptives bivariées sur des données d'Iris

3.1 Etude de la longueur du pétale en fonction de la largeur du pétale

Représentation graphique

La représentation graphique liant deux variables quantitatives est le nuage de points ou l'histogramme 2D.

1)Tracer le nuage de points de la longueur du pétale en fonction de la largeur du pétale pour les 150 iris contenus dans les données (plot, scatter ou scatterhist). Ne pas oublier de mettre des titres sur les axes. Décrire le nuage de points.

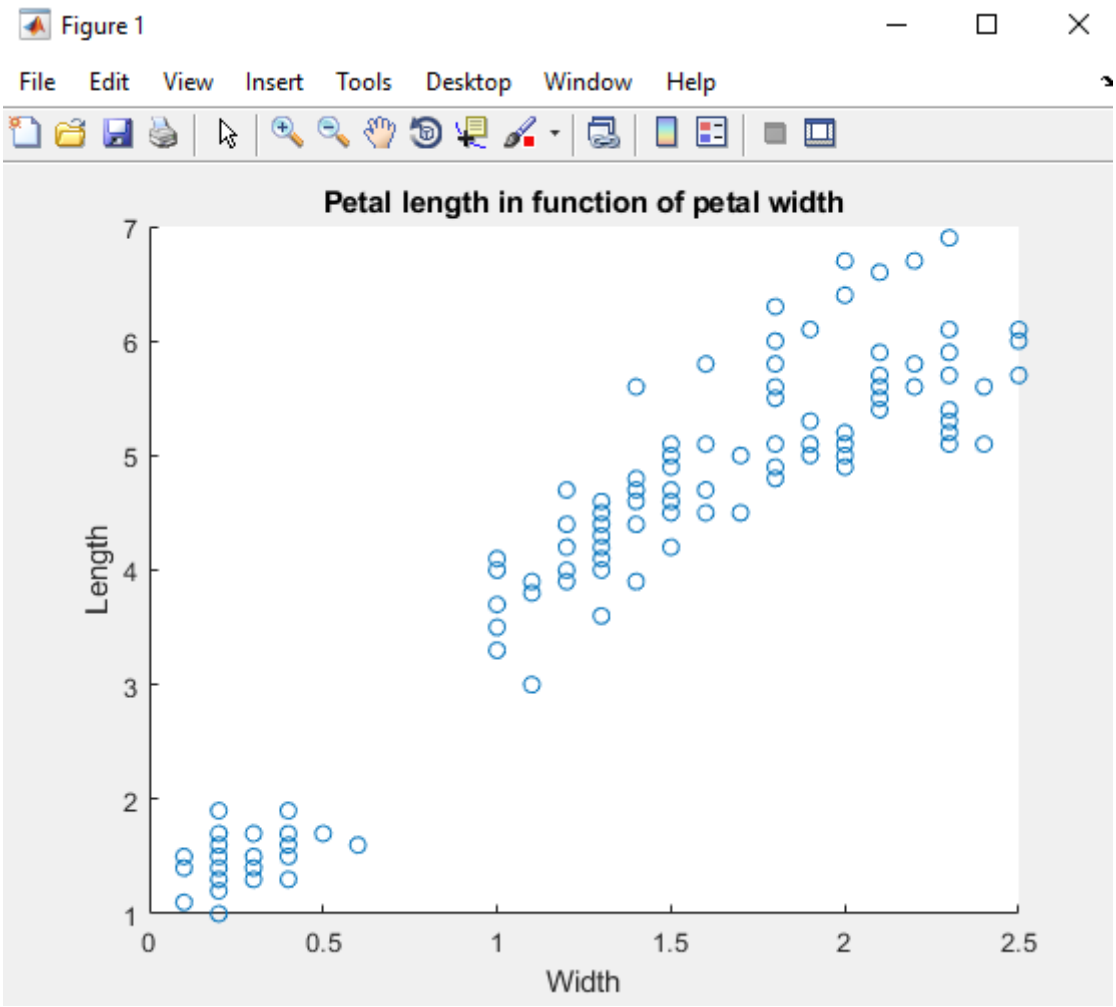

Figure 1:Nuage de points des Iris : longueur en fonction de la largeur

Le nuage de points ou diagramme de dispersion correspond à une représentation graphique dans un repère du plan d'une série statistique à deux variables ici petalWidth et petalLength. Chaque individu est représenté par un point avec en abscisse sa largeur et en ordonnée sa longueur. Le nuage de point semble former une droite (fonction affine : ax+b).
Il semblerait qu'il y ait proportionnalité entre la longueur et la largeur du pétale.
Par ailleurs, il semble qu'il y ait plusieurs groupes de valeurs. Ces valeurs pourraient appartenir aux différentes espèces d'Iris (Setosa, Versicolor, Virginica). Il convient donc de vérifier la corrélation.

Code :

```
load ("fisheriris.mat")

sepalLength = meas(:,1:1);
sepalWidth = meas(:,2:2);
petalLength = meas(:,3:3);
petalWidth = meas(:,4:4);

%Question 1
figure(1)
scatter(petalWidth, petalLength);
title('Petal length in function of petal width');
xlabel('Width');
ylabel('Length');
```

2)Tracer l'histogramme bidimensionnel associé (hist3). A quoi correspond chaque bâton ?

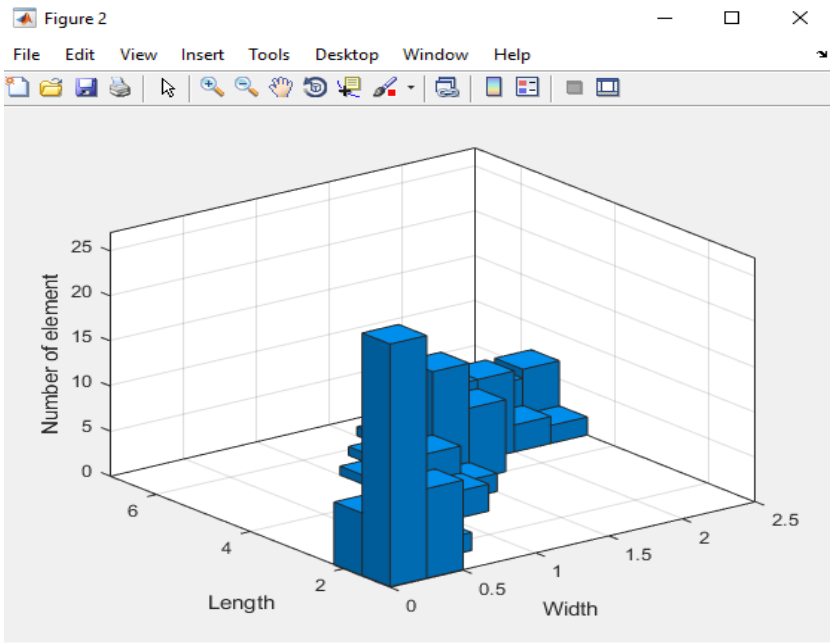

Figure 2 : Histogramme des individus de la série statistique (Length, Width, Number of element)

La position de chaque bâton correspond sa longueur et sa largeur. La hauteur représente le nombre d'effectifs correspondant à cette longueur et cette largeur précise.

Code :

```
%Question 2

figure(2);
histogram2(petalWidth, petalLength, 'NumBins', 10);
xlabel('Width');
ylabel('Length');
zlabel('Number of element');
```

Coefficient de corrélation et droite de régression

3)Tracer le nuage de points de la longueur du pétale en fonction de la largeur du pétale pour les 150 iris contenus dans les données.

Voir question 1.

4)Rappeler la définition de coefficient de corrélation et calculer le par la fonction (corrcoeff).

$$\rho X, Y = \frac{\mathrm{Cov}\{X, Y\}}{SXSY}$$

Le coefficient de corrélation linéaire est obtenu par l'équation ci-dessus. Cov est la covariance des variables quantitatives X et Y. Sx et Sy correspondent aux écarts types des deux variables.

Si $\rho X, Y$ est égale à 1 alors il existe une relation linéaire parfaite entre X et Y. De plus, quand X augmente, Y augmente.

Si $\rho X, Y$ est égale à -1 alors il existe une relation parfaite entre X et Y. De plus, quand X augmente, Y diminue.

Si $\rho X, Y$ est égale à 0, il n'y a pas de relation linéaire entre X et Y.

Quand on utilise la fonction corrcoeff, on obtient une matrice qui donne les coefficients de covariance X avec X, X avec Y, Y avec X et Y avec Y.

Code :

```
% Question 4

[corr]=corrcoef(petalLength,petalWidth);
covXY = corr(2)

% calcul coefficient corrélation linéaire

StdX= std(petalLength;
StdY=std(petalWidth);
r=CovXY/(StdX*StdY)
```

Résultats :

```
CovXY =

    0.9629

r =

    0.7156
```

5)Donner l'équation de la droite de régression linéaire, créer une fonction permettant de la calculer à partir de deux variables X et Y et tracer la sur le même graphique.

Pour déterminer la droite de régression linéaire on utilise la méthode des moindres carrés. On cherche la droite d'équation $y = \hat{a}x + \hat{b}$ avec les coefficients tels que :

$$\hat{a} = \rho X,Y\frac{SY}{SX} \quad \text{et} \quad \hat{b} = \bar{y} - \rho X,Y\frac{SY}{SX}\bar{x}$$

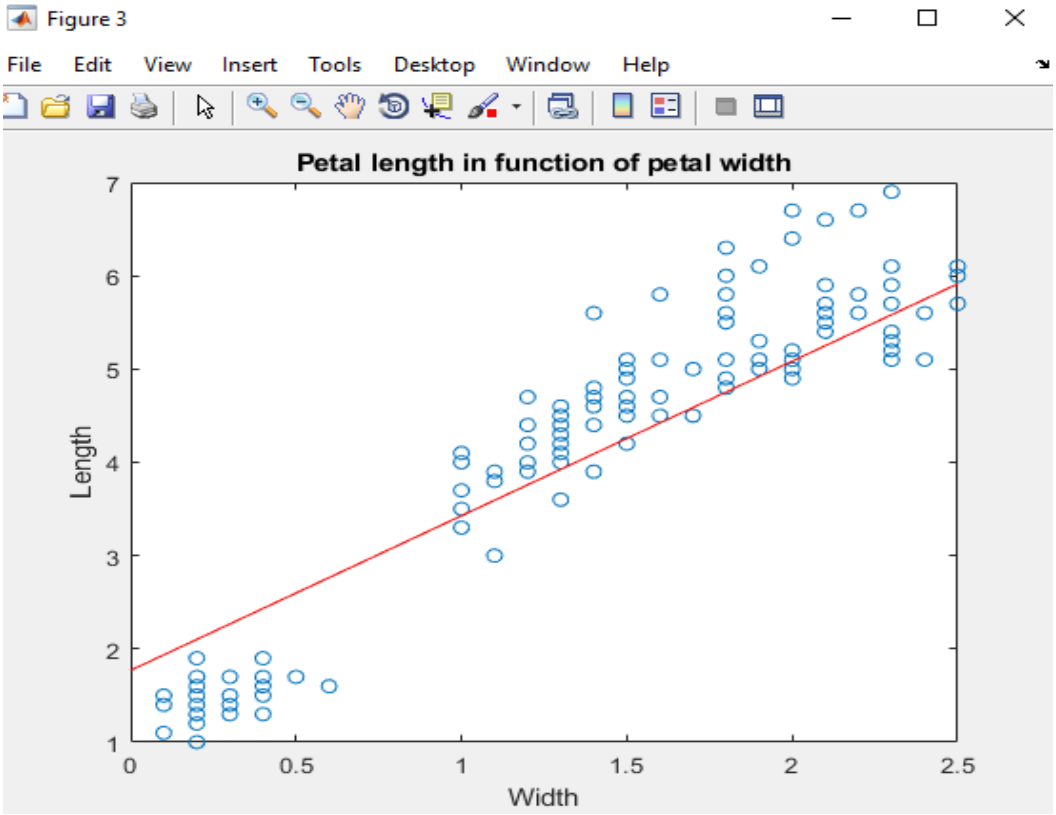

Figure 3 : Nuage de point avec la droite de corrélation

Code :

```
%Question 5
%Calcul des paramètres de l'équation de la droite de régression

a=r*(StdY/StdX)
b=mean(petalWidth)-r*(StdY/StdX)*mean(petalLength)

%Calcul de deux point pour tracer la droite
p1 = [0, 2.5];
p2 = [b, 2.5*a + b];

plot(petalWidth, petalLength, 'o', p1, p2, 'r-');
title('Petal length in function of petal width');
xlabel('Width');
ylabel('Length');
```

6)Analyser le lien entre les deux variables.

Le coefficient de corrélation vaut r = 0.7156 ce qui est relativement élevé. Nous pouvons dire que les deux variables sont donc corrélées positivement (si l'une augmente l'autre augmente également). Comme nous nous rapprochons de 1 mais que nous ne sommes pas à 1 nous aurons un nuage de points qui sera incliné comme notre droite de régression.

3.2 Etude de la longueur de pétale selon les différentes espèces

Représentations graphiques

7)Représenter sur une même figure l'histogramme par espèce (trois histogrammes avec une couleur pour chaque espèce)

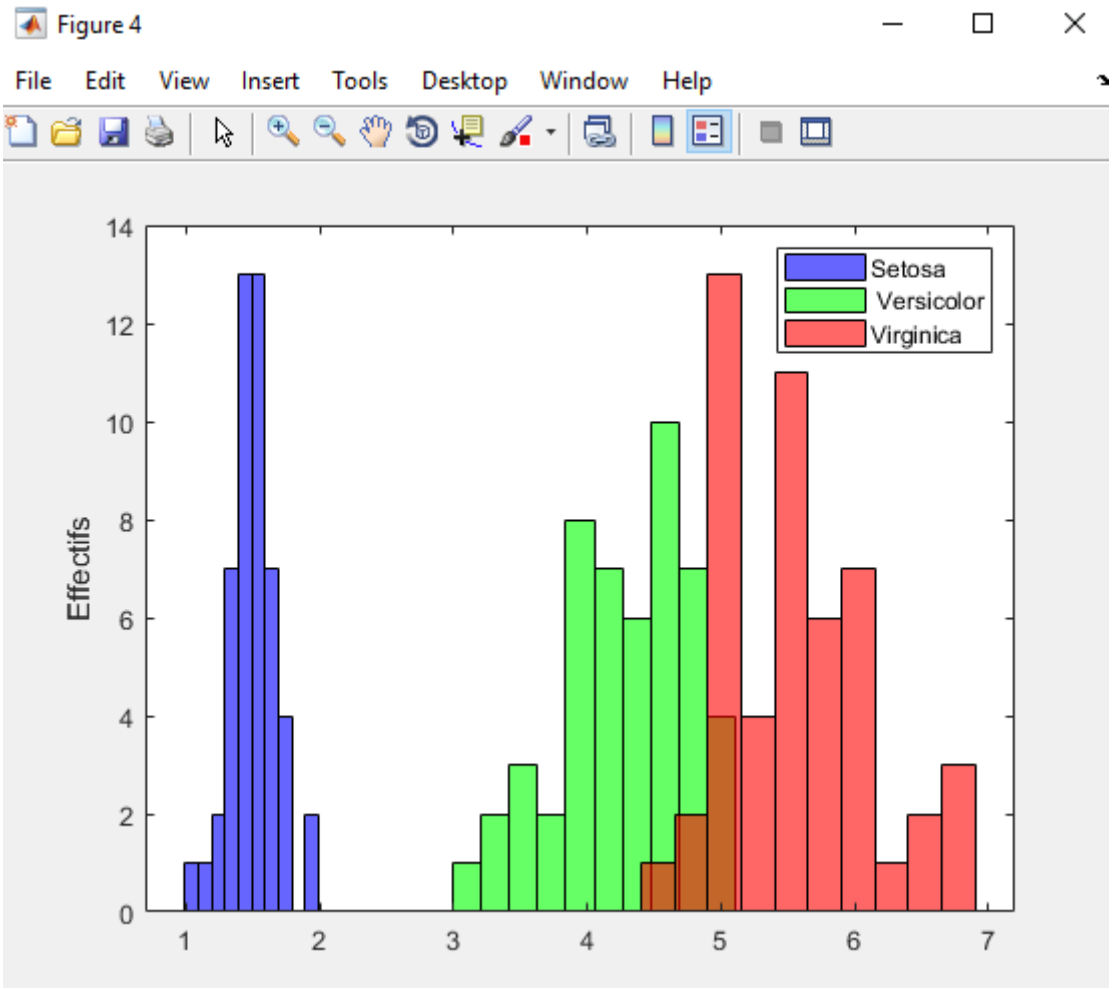

Figure 4 :Histogramme par espèces en fonction des effectifs

Code :

```
%Question 7

setosaPL = petalLength(1:50);
versicolorPL = petalLength(51:100);
virginicaPL = petalLength(101:150);

figure(4)

y = [setosaPL; versicolorPL ; virginicaPL];
histogram(setosaPL, 'FaceColor', 'b', 'NumBins', 10)
hold on
histogram(versicolorPL, 'FaceColor', 'g', 'NumBins', 10 )
hold on
histogram(virginicaPL, 'FaceColor', 'r', 'NumBins', 10 )
ylabel("Effectifs")
legend('Setosa', ' Versicolor','Virginica', 'Location', 'Northeast');
```

8)Représenter une boite à moustache par espèce

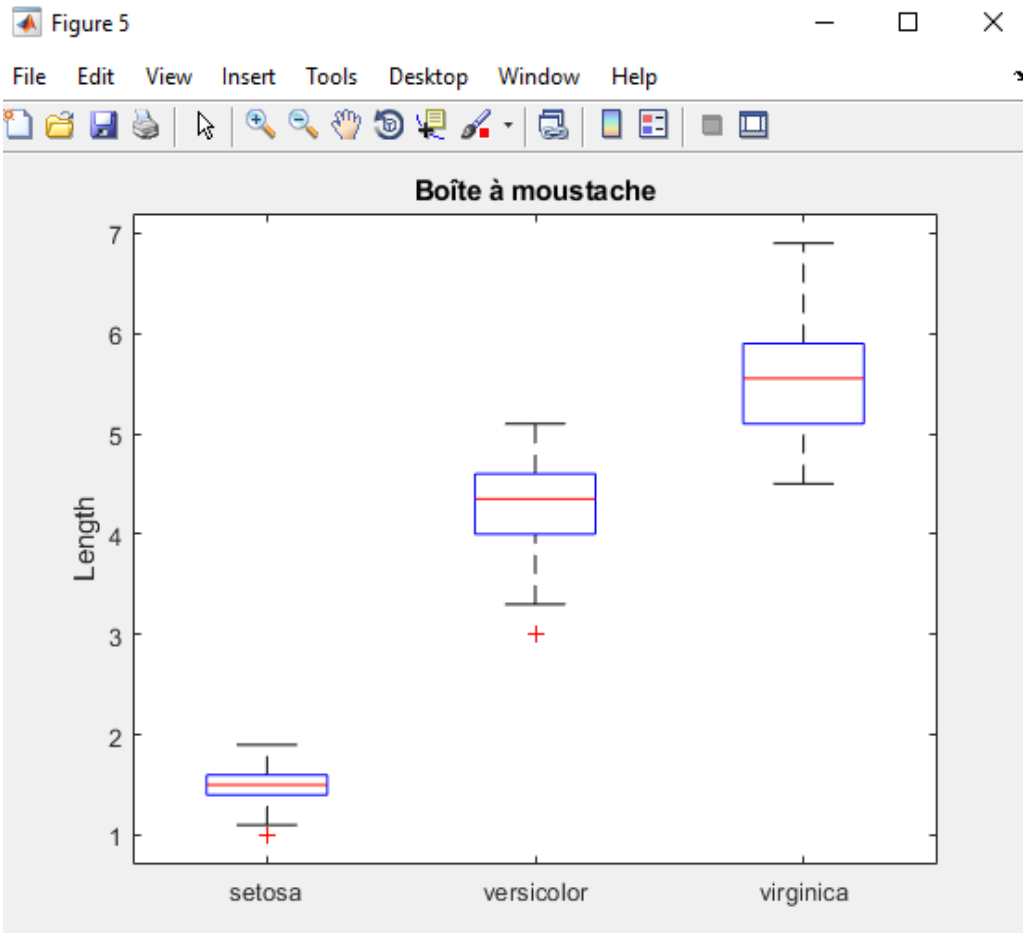

Figure 5 : Boîte à moustache

Code :

```
%Question 8
figure(5)
boxplot(y, species) %ici species correspond à la variable contenant les espèces
                    %d'iris
```

Mesures de corrélation

9)Calculer le rapport de corrélation lié à la décomposition de la variance en variance intra-classe et interclasse. Qu'en concluez-vous?

La variance totale se décompose en somme de variance interclasse et intraclasse telle que :

$$Sy^2 = \frac{1}{n}\sum_{i=1}^{p} n_i.(\overline{y_i} - \bar{y})^2 + \frac{1}{n}\sum_{i=1}^{p} n_i S_{y_i}^2$$
$$= \quad S_B^2 \quad + \quad S_W^2$$
$$= Variance\ interclasse + variance\ intraclasse$$

Le rapport de corrélation lié à la décomposition est définit par la formule ci-dessous :

Code :

```
%Question 9
%Calcul de la variance intra-classe

mean_petal = mean(y);
set_mean = mean(setosaPL);

%Calcul des moyennes
mean_setosa = mean(setosaPL);
mean_versicolor = mean(versicolorPL);
mean_virginica = mean(virginicaPL);

%Calcul des variances
var_inter = 50 .*(mean_setosa - mean_petal) .^ 2 + 50 .*(mean_versicolor -
mean_petal) .^ 2 + 50 .* (mean_virginica - mean_petal) .^ 2;
var_intra = (50 .* var(setosaPL) + 50.*var(versicolorPL) + 50.*var(virgin-
icaPL))./150;
var_inter = var_inter ./ 150;
my_variance = var_intra + var_inter;

var(y) - my_variance

%Calcul du coefficient de correlation entre nos variables qualitatives et
%quantitative
```

```
S_yx = sqrt(var_inter/my_variance)
```

Résultats :

```
S_yx =

    0.9697
```

Ici le rapport est proche de 1 donc $S_B^2 \gg S_w^2$. Il y a une bonne séparabilité des sous-échantillons. Ces données viennent confirmer notre intuition : la longueur du pétale dépend de l'espèce étudiée.