Rapport SY09 TP01 - Statistique descriptive, analyse en composantes principales

Jean-Baptiste Audibert & Yueqing Qin

8 avril 2011

1 Statistique descriptive

1.1 Données babies

Dans cette première partie du TP, on étudie des données sur des bébés (*1236 échantillons et 23 variables, cependant nous ne réaliserons l'étude qu'avec 9 variables*). Le but est d'étudier les différentes influences du fait que la mère soit fumeuse ou non sur des caractéristiques du bébé et de la grossesse, mais également sur la mère. Nous étudions tout d'abord la différence de poids entre les bébés nés de mères fumeuses et de mères non fumeuses.
En étudiant le poids des bébés des échantillons, en distinguant les mères fumeuses et non fumeuses , on observe que la moyenne du poids est plus élevé pour une mère fumeuse que non fumeuse :

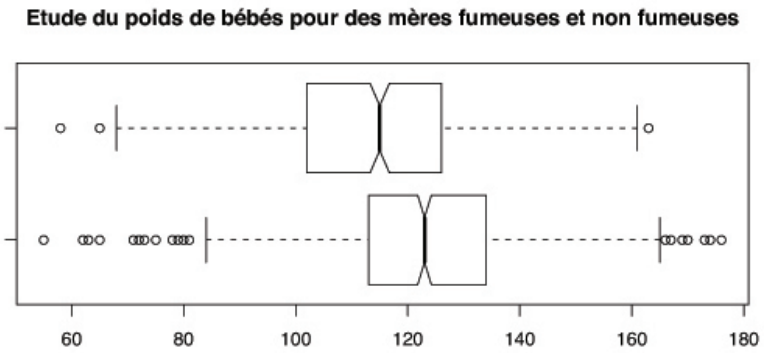
FIGURE 1 – Résumé des données concernant le poids des bébés à la naissance - mère fumeuse

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
58.0	102.0	115.0	114.1	126.0	163.0	10.0

FIGURE 2 – Résumé des données concernant le poids des bébés à la naissance - mère non fumeuse

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
55	113	123	123	134	176	10

De plus, nous utilisons la représentation en *boite à moustaches* afin d'observer la répartition des valeurs, la dispersion des données mais également la matérialisation de de l'intervalle de confiance de la médiane :



On retrouve en haut la représentation pour une mère non fumeuse, et en bas celle pour une mère non fumeuse. On observe d'emblée que la répartition des valeurs est très différente selon que la mère soit fumeuse ou non. De plus, on observe que les intervalles de confiance de la médiane

sont disjoints, ce que nous permet de confirmer R : soit m_f la médiane pour une mère fumeuse et m_{nf} la médiane pour une mère non fumeuse, on a $m_f \in [121.8; 124.2]$ et $m_nf \in [113.3; 116.7]$, soit $IC_{m_f} \cap IC_{m_nf} = \emptyset$. D'après cela, on peut dire que le fait que la mère soit fumeuse a bien une influence sur le poids du bébé à la naissance.

On s'intéresse maintenant au temps de gestation de la mère, selon qu'elle soit fumeuse ou non.

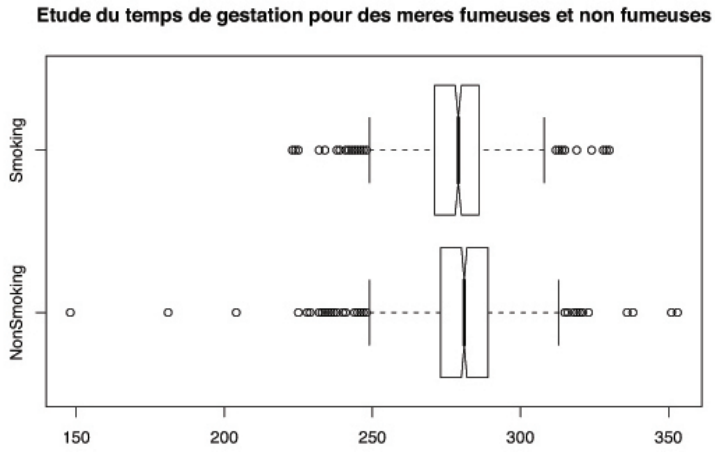
FIGURE 3 – Résumé des données concernant le temps de gestation - mère fumeuse

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
223	271	279	278	286	330	14

FIGURE 4 – Résumé des données concernant le temps de gestation - mère non fumeuse

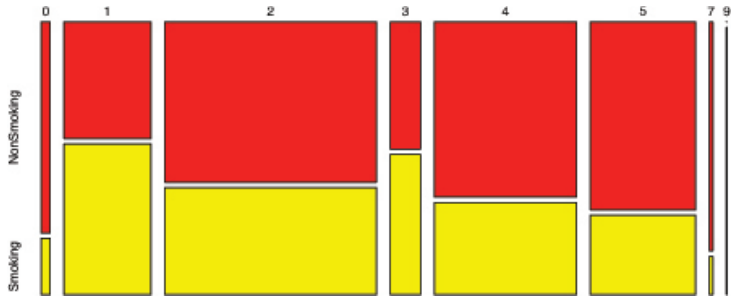
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
148.0	273.0	281.0	280.2	289.0	353.0	19.0

On observe également une différence dans la moyenne, cependant elle parait très peu significative puisque égale à 0.79% Etudions l'affichage avec les boites à moustache :



Ce que l'on pensait est ici confirmé, les intervalles de confiance de la médiane ne sont pas disjoints ($m_{nf} \in [280.05; 281.93]$ et $m_f \in [277.92; 280.08]$, d'où $IC_{m_f} \cap IC_{m_nf} \neq 0$). Ainsi, on peut dire que le fait que la mère soit fumeuse n'influe pas sur le temps de gestation.

On s'intéresse enfin à la corrélation entre le fait que la mère soit fumeuse et le niveau d'étude. Pour cela, on utilise un tableau de contingence, ainsi que son histogramme afin étudier la proportion de mère fumeuse dans chaque *niveau d'étude* :



Les échantillons pour des niveaux d'étude 7 et 9 ne sont pas utilisés car les populations sont trop faibles pour être utiles dans l'étude. On observe ici que plus le niveau d'étude est élevé, plus la proportion de mères *non fumeuses* augmente. Ainsi, on peut bien dire que le niveau d'étude a une influence sur le fait que la mère soit fumeuse ou non.

1.2 Données crabes

Nous allons désormais étudier des données morphologiques sur des crabes. L'échantillon est composé de 200 crabes, étudiés avec huit variables. Cependant, nous n'utiliserons que les 5 variables *quantitatives* afin de mener l'étude.

On souhaite tout d'abord s'il est possible de distinguer les caractéristiques morphologiques des crabes selon leur espèce et/ou leur sexe, puis de les identifier en utilisant ces caractéristiques morphologiques. Pour cela, on étudie tout d'abord deux graphiques de corrélation entre les variables, selon l'espèce, et le sexe.

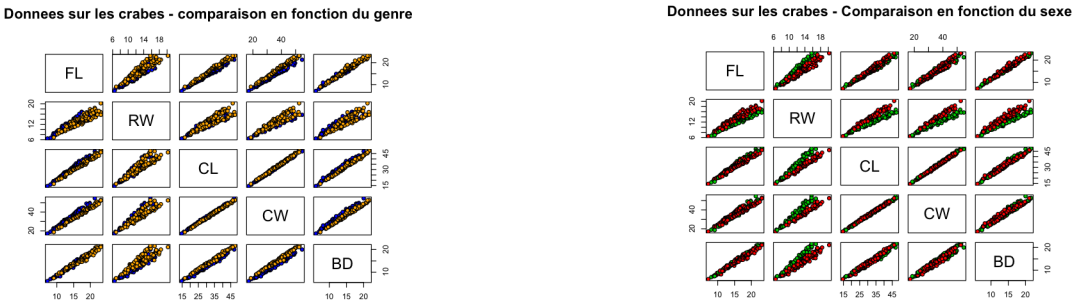


FIGURE 5 – Etude des corrélations avec la fonction pairs

Les variables sont très corrélées et on ne peut pour l'instant affirmer qu'il y a des différences significatives selon le sexe ou l'espèce. On étudie donc l'influence des variables quantitatives sur les variables qualitatives en utilisant des représentations en boite à moustache. Nous n'avons pu inclure tous les graphiques dans le rapport, faute de place.
Toutefois, grâce à ces représentations, et en utilisant les intervalles de confiance sur la médiane, on peut affirmer qu'il y a des différences morphologiques selon l'espèce et le sexe.
Nous allons maintenant étudier la corrélation entre les variables. Les calculs nous donnent :

```
          FL        RW        CL        CW        BD
FL 1.0000000 0.9069876 0.9788418 0.9649558 0.9876272
RW 0.9069876 1.0000000 0.8927430 0.9004021 0.8892054
CL 0.9788418 0.8927430 1.0000000 0.9950225 0.9832038
CW 0.9649558 0.9004021 0.9950225 1.0000000 0.9678117
BD 0.9876272 0.8892054 0.9832038 0.9678117 1.0000000
```

On observe qu'effectivement, les variables sont très corrélées. Ceci est dû à un phénomène physique simple : quand on est en présence d'un *petit crabe*, l'ensemble de ses attributs est *petit*, et inversement. Pour supprimer la corrélation, on pourrait diviser par une des variable, à priori la plus corrélée avec les autres, et recommencer l'étude.

2 Analyse en composante principale

2.1 Exercice théorique

2.1.1 Axes factoriels et pourcentages d'inertie expliquée

Matrice des valeurs propres / axes factoriels :

$$\begin{pmatrix} 0 & 0 & 1 \\ -0.707 & 0.707 & 0 \\ 0.707 & 0.707 & 0 \end{pmatrix}$$

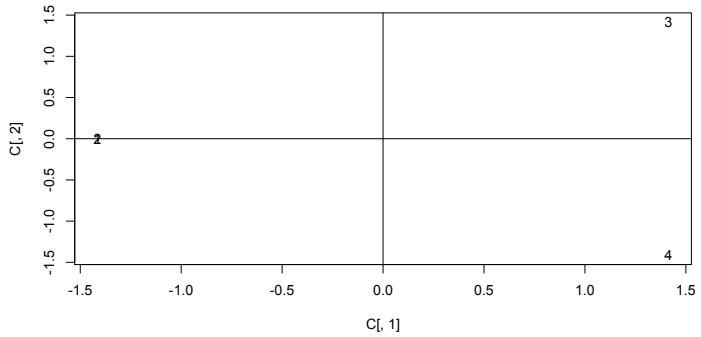
Pourcentage d'inertie expliqué par chacun des axes : 57.14%, 28.57%, 14.59%. Le troisième axe est donc clairement le moins intéressant pour l'étude.

2.1.2 Composantes principales et représentation dans le premier plan factoriel

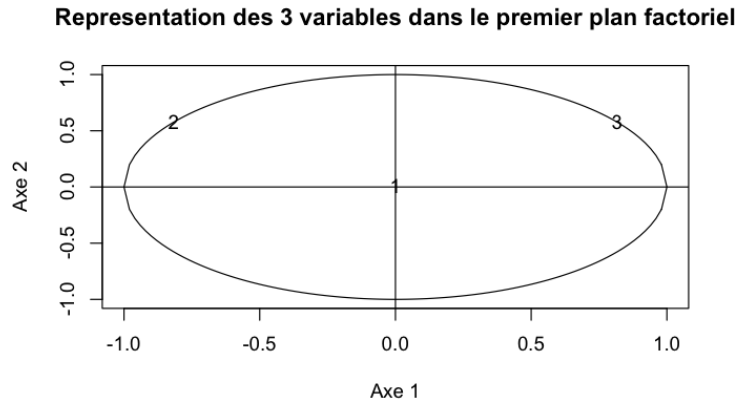
Matrice des composantes principales :

$$\begin{pmatrix} -1.414 & 2.220 * 10^{-16} & 1 \\ -1.414 & 2.220 * 10^{-16} & -1 \\ 1.414 & 1.414 & 0 \\ 1.414 & -1.414 & 0 \end{pmatrix}$$

Et on représente les individus dans le premier plan factoriel :



2.1.3 Représentation des 3 variables dans le premier plan factoriel



2.1.4 Calcul de reconstitution

$k = 1$

$$\sum_{\alpha=1}^{1} c_\alpha u_\alpha^t = \begin{pmatrix} 0 & 1 & -1 \\ 0 & 1 & -1 \\ 0 & -1 & 1 \\ 0 & -1 & 1 \end{pmatrix}$$

$k = 2$

$$\sum_{\alpha=1}^{2} c_\alpha u_\alpha^t = \begin{pmatrix} 0 & 1 & -1 \\ 0 & 1 & -1 \\ 0 & 0 & 2 \\ 0 & -2 & 0 \end{pmatrix}$$

$k = 3$

$$\sum_{\alpha=1}^{3} c_\alpha u_\alpha^t = \begin{pmatrix} 1 & 1 & -1 \\ -1 & 1 & -1 \\ 0 & 0 & 2 \\ 0 & -2 & 0 \end{pmatrix}$$

2.2 Utilisation des outils R

2.2.1 Reprise des résultats

On reprend l'exemple du cours, avec les différentes fonctions fournies par **R**.
Avec l'appel *acp <- princomp(A)*, où A est la matrice de départ, on obtient l'ensemble des résultats dans l'objet *acp*. On obtient ainsi par exemple la matrice des vecteurs propres, ou la mtrice des composantes principales :

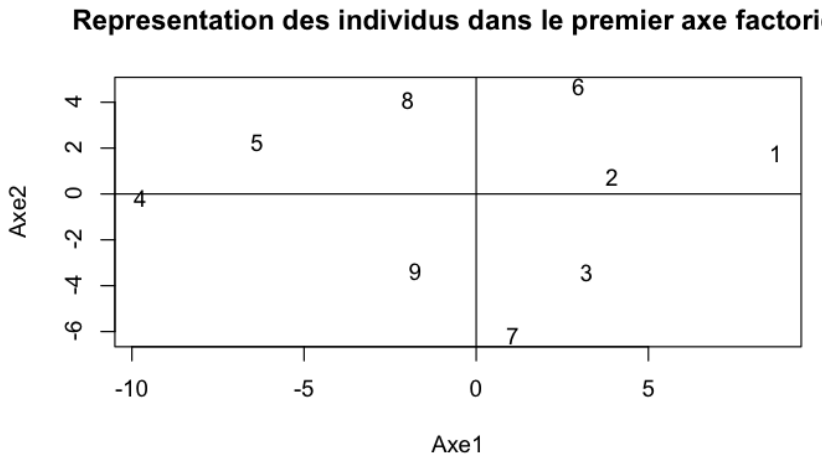
```
> acp$loadings

Loadings:
     Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
[1,] -0.515  0.567         -0.289  0.573
[2,] -0.507  0.372          0.553 -0.546
[3,] -0.492 -0.650  0.108  0.394  0.410
[4,] -0.485 -0.323         -0.674 -0.453
[5,]        -0.113 -0.992
```

```
> acp$scores
          Comp.1     Comp.2     Comp.3      Comp.4      Comp.5
[1,]  8.700907  1.7027046  2.5539182 -0.14945398 -0.11731596
[2,]  3.938596  0.7085441  1.8104644 -0.09068389  0.04349922
[3,]  3.209392 -3.4590552  0.3006617  0.17254286  0.01928215
[4,] -9.755741 -0.2157421  3.3436726 -0.17347137  0.10041455
[5,] -6.371422  2.1733326  0.9570588  0.07066256 -0.18799232
[6,]  2.974017  4.6509322 -2.6349457 -0.02321315  0.14809545
[7,]  1.050967 -6.2271742  1.6880636  0.11529582  0.04281219
[8,] -1.980533  4.0685562 -1.4007122  0.24321198  0.01039742
[9,] -1.766183 -3.4020982 -6.6181814 -0.16489082 -0.05919270
```

FIGURE 6 – Affichage de résultats de la fonction princomp : matrice des vecteurs propres et matrice des composantes principales

Avec la fonction plot classique, on obtient la projection des individus dans le premier axe factoriel :



2.2.2 Fonction princomp, plot et biplot

La fonction **plot** affiche la variance des différents composants. Une variance élevée est intéressante pour effectuer l'analyse. En effet, la variance est égale à la valeur de propre de la composante. La fonction **biplot** affiche la représentation des individus dans les deux principaux axes d'inertie. Elle affiche également les vecteurs de corrélation entre les variables.

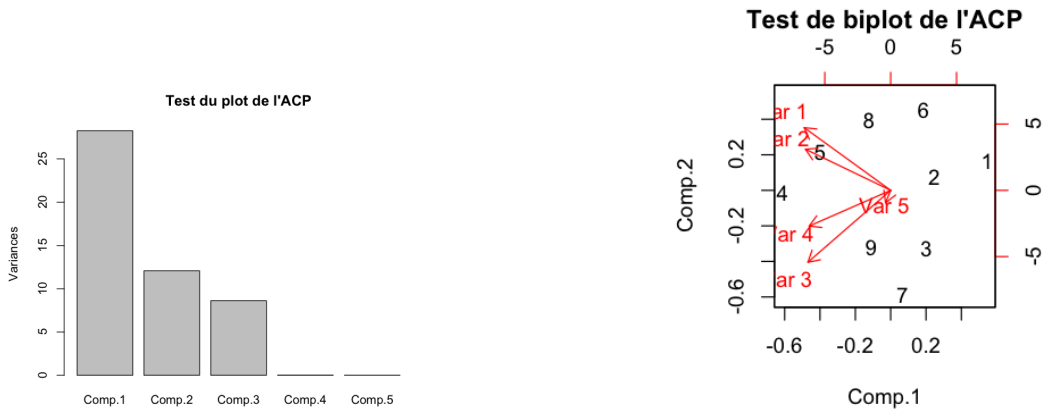
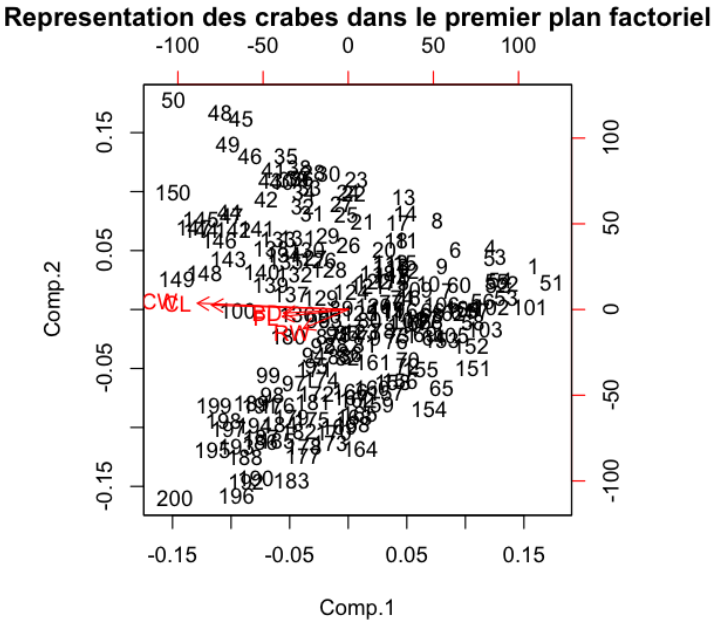


FIGURE 7 – Affichage de résultats des fonctions plot et biplot pour l'ACP

2.3 Traitement des données Crabs par l'ACP

2.3.1 Première analyse sans traitement préalable

Nous allons désormais appliquer l'ACP sur les données **Crabs**, afin d'étudier les données et d'essayer d'identifier des groupes. On effectue tout d'abord l'ACP sans traitement préalable. Avec le *biplot*, on obtient le nuage de points suivants :



On constate qu'il est impossible de distinguer des groupes, selon le sexe ou l'espèce.

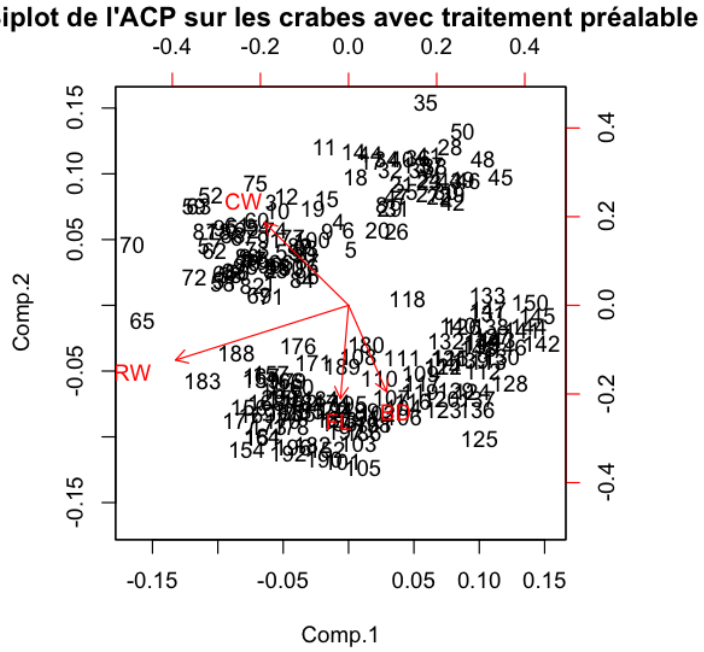
2.4 Traitement des données Crabs avec traitement préalable

On a observé précédemment qu'il était impossible de distinguer des groupes. Ceci est du à un effet de taille, et au fait que les variables sont toutes liés. On va donc chercher à supprimer une variable, afin d'obtenir des données plus promptes à l'analyse. En étudiant les corrélations, on se rend compte que la variable **CL** est la plus corrélée avec les autres. De plus, avec l'histogramme obtenu par le *plot* de la fonction *princomp*, on se rend compte qu'elle est clairement dominante.

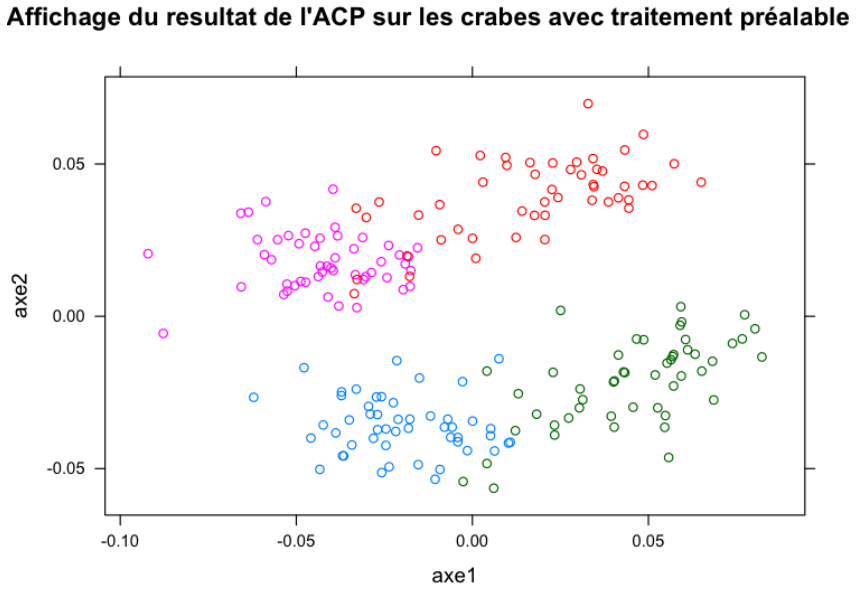
```
          FL        RW        CL        CW        BD
FL 1.0000000 0.9069876 0.9788418 0.9649558 0.9876272
RW 0.9069876 1.0000000 0.8927430 0.9004021 0.8892054
CL 0.9788418 0.8927430 1.0000000 0.9950225 0.9832038
CW 0.9649558 0.9004021 0.9950225 1.0000000 0.9678117
BD 0.9876272 0.8892054 0.9832038 0.9678117 1.0000000
> colSums(cor(crabsquant))
      FL       RW       CL       CW       BD
4.838412 4.589338 4.849811 4.828192 4.827848
```

Nous allons donc *supprimer* cette variable en divisant chaque ligne i par la valeur CL_i correspondante, et étudier le jeu de données ainsi obtenus.

On observe désormais qu'il n'y a plus de variables dominantes comme précédemment. Nous allons projeter dans le premier axe factoriel afin de savoir s'il est possible d'observer distinctement un ou plusieurs groupes :



On observe bien des groupes, cependant, il est difficile de les identifier. Nous allons donc construire quatre groupes avec l'outil *class* de R : MB/MO/FM/FB, correspondant aux différentes associations sexe/espèce. Afin d'afficher ces groupes, on utilise la fonction *xyplot*, du package *lattice*. On obtient le résultat suivant :



On peut désormais observer quatre groupes distincts, et donc identifier les crabes.