EISTI

TP DE STATISTIQUE MULTIVARIÉE

17.12.2012

LA BASE DE DONNÉES

Les données "iris" sont des données proposées en 1933 par le statisticien Ronald Aylmer Fisher et elles correspondent à 3 types (classes) de fleurs, à savoir iris setosa, iris virginica et iris versicolor.

Pour chaque classe il y a 50 observations, donc au total il y a 150 observations.

Chaque observation contient les mesures de quatre variables (variables quantitatives) qui sont, dans l'ordre, la longueur et la largeur des sépales (LOSE et LASE), la longueur et la largeur des pétales (LOPE et LAPE) et aussi le numéro de la classe à laquelle appartient la fleur observée (variable ordinale). Toutes les variables quantitatives sont exprimées en millimètres.

Les trois lignes suivantes présentent un échantillon du fichier de données, chaque ligne correspondant à une classe différente.

```
Long. s\'ep.  Larg. S\'ep   Long. p\'et.  Larg. p\'et. Classe
  5.100         3.500         1.400         0.200        1
  7.000         3.200         4.700         1.400        2
  6.300         3.300         6.000         2.500        3
```

On se propose de faire une étude complète de cet ensemble de données et, en particulier, l'influence de la longueur des pétales (3e variable quantitative) à l'ensemble des observations.

LES RÉSULTATS DES CALCULS PRÉLIMINAIRES

Sur l'ensemble de mesures, on a effectué plusieurs calculs dont les résultats sont donnés ci après :

En utilisant l'ensemble des observations (150 observations, 4 variables quantitatives, 1 variable qualitative (numéro de classe)), on a calculé :

1. moyenne et écart-type non biaisé (table 1)

	LOSE	LASE	LOPE	LAPE
Moyenne	5.84	3.06	3.76	1.20
Écart-type non biaisé	0.83	0.43	1.76	0.76

TABLE 1 – Grandeurs statistiques de la base de données totale

2. matrice des variances-covariances (table 2)

	LOSE	LASE	LOPE	LAPE
LOSE	0.6811	-0.0422	1.2658	0.5128
LASE	-0.0422	0.1887	-0.3275	-0.1208
LOPE	1.2658	-0.3275	3.0955	1.2870
LAPE	0.5128	-0.1208	1.2870	0.5771

TABLE 2 – Matrice des variances-covariances de la base de données totale

3. matrice des corrélations (table 3)

La table 4 fournit la moyenne et l'écart-type non biaisé pour la variable "longueur pétales" selon les trois classes des fleurs

On considère pour la suite que les distributions des valeurs de toutes les variables suivent une distribution normale $\mathcal{N}(\mu, \sigma^2)$.

TRAVAIL À FAIRE

1. Nous voulons d'abord étudier l'influence de la variable "longueur des pétales" sur la séparation en classes des iris en comparant les moyennes de cette variable dans chacune des classes à l'aide de l'analyse de variance.

	LOSE	LASE	LOPE	LAPE
LOSE	1.0000	-0.1176	0.8718	0.8179
LASE	-0.1176	1.0000	-0.4284	-0.3661
LOPE	0.8718	-0.4284	1.0000	0.9629
LAPE	0.8179	-0.3661	0.9629	1.0000

TABLE 3 – Matrice des corrélations de la base de données totale

	Classe 1	Classe 2	Classe 3
Moyenne	1.46	4.26	5.55
Écart-type non biaisé	0.18	0.47	0.55

TABLE 4 – Grandeurs statistiques pour la variable "longueur pétales" par classe

   (a) Indiquer, de façon schématique, le tableau de données à utiliser pour faire l'analyse de variance. Par "De façon schématique" on entend qu'il s'agit d'un tableau à double entrée et vous indiquerez ce que représentent ses lignes et ses colonnes, comme suit

   (b) En utilisant le tableau de l'analyse de variance et le programme ANOVA qu'on vous a fourni, vérifier, si les moyennes de la variable "longueur des pétales" dans les trois classes sont équivalentes avec une confiance de 99%.

   (c) Donnez vos commentaires pour le résultat de l'analyse de la variance. Est-il possible de corroborer ce résultat par des analyses supplémentaires et lesquelles ?

2. Nous voulons, en utilisant la régression multilinéaire sur la totalité de la base de données, établir une relation linéaire entre la valeur de la variable "longueur des pétales" et les autres variables.
   En utilisant le programme de régression qui vous a été fourni, établir la meilleure relation linéaire. Justifier votre réponse.
3. En utilisant la totalité du tableau, on procède à une analyse en composantes principales dont les résultats sont donnés à l'annexe 2.
   (a) Expliquer brièvement le principe de l'ACP.
   (b) Commenter les graphiques obtenus.
4. On procède ensuite à une analyse factorielle des correspondances en utilisant le programme fourni.
   (a) Expliquer brièvement le principe de l'AFC.
   (b) Commenter les graphiques obtenus.
   (c) Vérifier si les résultats sont en conformité avec les résultats de l'ACP.
   (d) De façon générale, quelle est la différence entre ACP ET AFC.
   (e) Que faut-il faire pour améliorer avec l'AFC la séparation entre les classe 2 et 3. Tester votre idée et commenter les résultats.
5. Faire une analyse factorielle discriminante avec le programme fourni.
   (a) Expliquer brièvement le principe de l'AFC.
   (b) Commenter le graphique obtenu.
   (c) Faire une nouvelle AFD en utilisant les fleurs de deux dernières classes et en mettant en points supplémentaires les fleurs de la première classe. Commenter le graphique obtenu.
6. Nous allons maintenant examiner si une méthode de classification retrouve les classes d'origine de la population. Utiliser pour cela le programme de k plus proches voisins qui vous a été fourni. Commenter les résultats.
7. On va maintenant évaluer la capacité de cet ensemble de classer correctement des nouvelles fleurs. Utiliser le programme de classement flou qui vous a été fourni et commenter les résultats.
8. En récapitulant tous les résultats, faites une conclusion concernant la séparation entre la 2e et la 3e classes.

	colonne 1 $\cdots$
ligne 1	$\cdots$
$\cdots$	