L. Noiret,
A. Rogozan

TP
Statistiques Descriptives
Analyse multi-dimensionnelle

3e année

Quelques définitions

- coefficient d'asymétrie (skewness en anglais) : mesure l'asymétrie d'une distribution d'une variable. Un coefficient positif (resp. négatif) indique une queue de distribution étalée vers la droite (resp. gauche).
- coefficient de corrélation linéaire (r) : mesure l'intensité de la liaison linéaire entre deux variables quantitative. r varie entre -1 et 1. Une valeur proche de 0, indique que les données ne sont pas liées linéairement, une valeur proche de -1 (resp. 1) indique que les données sont fortement liées négativement (resp. postivement).
- nuage de points (scatter plot en anglais) est une représentation de données (x versus y) qui permet de mettre en évidence le type de liaisons entre deux variables.

1 Description d'un base de données cliniques

On s'intéresse aux résultats d'une étude clinique réalisée aux Etats-Unis. Le jeu de données contient :

- *gender* : le sexe du patient ;
- *ethnicity* : l'origine éthnique déclarée (cette question est couramment posée aux Etats Unis) ;
- *age* : l'âge du patient au moment de l'étude ;
- *weight* : le poids du patient ;
- *protein*, *protein*2, *protein*3 : la concentration de trois protéines d'intérêt dans le sang
- *n_visit* : le nombre de visites médicales au cours des 24 derniers mois.

1.1 Chargement et pré-traitements des données

Importer le fichier *study.csv* et enregistrer les données au format table.
Quels sont les types des différentes variables ?

- Les variables *gender* et *ethnicity* sont pour l'instant enregistrées en tant que chaînes de caractères. Nous allons les transformer en variables catégorielles (ce qui nous facilitera leur analyse). Exécuter les lignes suivantes :
  study.ethnicity=categorical(study.ethnicity) ;
  study.gender=categorical(study.gender) ;
- Supprimer les individus dont l'âge est inférieur à 15 ans :
  study(study.age≤15, :)=[] ;

Utiliser la fonction *summary* afin d'obtenir une vie d'ensemble rapide des données. A premiè vue, les données contiennent-elles des valeurs abérrantes ou atypiques ?

1.2 Tableau des fréquences

Donner le tableau des fréquences f et des fréquences cumulées F de la variable n_visit en adaptant le code suivant :
[n,edges]=hist(X,unique(X))
N=cumsum(hist(X),unique(X))
Etudier ces tableaux pour en déduire la valeur de la médiane, du 1er quartile et du 3e quartile.

1.3 Représentation graphiques

Représenter graphiquement les effectifs des variables *gender*, *ethnicity*, *weight*.
Quelle indicateur de tendance central pouvez-vous utiliser pour chaque variable ?

1.4 Représentation graphique multidimensionelle

1.4.1 Matrice des nuages de points

Adapter le code suivant pour obtenir les nuages de points et histogrammes de toutes les variables quantitatives :
dataLabels = {*'label 1','label 2','label 3'*}
$[H, AX] = plotmatrix(data)$;
for i = 1 : length(AX)
ylabel(AX(i,1),dataLabels{i}) ;
xlabel(AX(end,i),dataLabels{i}) ;
end
Notes : la fonction *plotmatrix* n'accepte que des données de type array. Pour transformer une table en array, vous devez utiliser $table2array$.
Pour récuper le noms des variables dans une table, utiliser : *study.Properties.VariableNames*.

1.4.2 Coefficient d'asymétrie

A votre avis, les coefficients d'asymétrie associés aux variables *age* et $protein2$ sont-ils positifs ou négatifs ?
Vérifier votre intuition en calculant le coefficient d'asymétrie sur toutes les variables quantitatives.
Aide : la fonction *'skewness'* de matlab prend en entrée un array et non une table.

1.4.3 Analyse des nuages de points

Comment décririez-vous les relations entre les différentes variables quantitatives continues ?
Les variables sont-elles liées linéairement, non linéairement, les variables ne sont pas liées ?
Calculer les coefficients de corrélations entre chaque paire de variables. Commenter.

Aide : regarder l'aide sur les fonctions : *corrcoef*, *corr* et *corrplot*. Attention, les deux premières fonctions n'acceptent pas de *table* en entrée.

1.5 Recodage d'une variable

Ecrire un code pour recoder la variable *age* en 5 classes d'effectifs égaux. Les résulats seront stockés dans une variable de la table *study* sous le nom *'age_5cat'*.
Aide : regarder les fonctions *quantile* et *discretize*.