

## Quelques définitions

- Le coefficient d'asymétrie (skewness en anglais) : mesure l'asymétrie d'une distribution d'une variable réelle. Un coefficient positif (resp. négatif) indique une queue de distribution étalée vers la droite (resp. gauche).
- Le 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. positivement).
- 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.

## 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 ethnique 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*, *protein2*, *protein3* : la concentration de trois protéines d'intérêt dans le sang
- *n\_visit* : le nombre de visites médicales au cours des 24 derniers mois.

## 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 vue d'ensemble rapide des données. A première vue, les données contiennent-elles des valeurs aberrantes ou atypiques ?

Tracer les boîtes à moustaches des variables quantitatives.

## 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]=histcounts(X,unique(X))
```

```
N=cumsum(n)
```

Etudier ces tableaux pour en déduire la valeur de la médiane, du 1er quartile et du 3e quartile.

## 3 Représentation graphique unidimensionnelle

Représenter graphiquement les effectifs des variables *gender*, *ethnicity*, *weight*.

Tracer les boîtes à moustaches des variables quantitatives.

## 4 Représentation graphique multidimensionnelle

Tracer les nuages de point de la variable *weight* en fonction des variables *protein*, *protein2* et *protein3*

Comment décririez-vous les relations entre ces différentes variables ?

Calculer les coefficients de corrélations entre la variable *weight* et *protein*, *protein2* et *protein3*. Commenter

Tracer les histogrammes des variables *age*, *protein*, *protein2* et *weight*.

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.

## 5 Recodage d'une variable

Ecrire un code pour recoder la variable *age* en 5 classes d'effectifs égaux. Les résultats seront stockés dans une variable de la table *study* sous le nom '*age\_5cat*'.

Tracer l'histogramme de cette nouvelle variable.

Aide : regarder les fonctions *quantile* et *discretize*.