

Statistique descriptive

Analyse multi-dimensionnelle

Dans ce TP on s'intéresse aux résultats d'une étude clinique réalisée aux États-Unis. Le jeu de données contient les variables suivantes :



- gender : le sexe du patient ;
- ethnicity : l'origine ethnique déclarée (cette question est couramment posée aux États-Unis) ;
- age : l'âge du patient au moment de l'étude ;
- weight : le poids du patient ;
- protein, protein2 et protein3 : la concentration de trois protéines d'intérêt dans le sang ;
- n_visit : le nombre de visites médicales au cours des 24 derniers mois.
- X : une variable aléatoire gaussienne que nous avons rajouter.

Partie 1

1. Charger les données, sous forme de DataFrame, dans la variable data. Donner le type de chaque variable et vérifier si il y'a des valeurs manquantes.
2. Stocker le nombre de variables dans la variable n et celui des d'observations dans la variable p

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 :



```
import pandas as pd
data.ethnicity = pd.Categorical(data.ethnicity);
data.gender = pd.Categorical(data.gender);
data.info()
```

3. Utiliser la fonction data.describe() afin d'obtenir une vue d'ensemble sur la répartition des données. Commenter les résultats obtenus
4. Donner le tableau des fréquences et des fréquences cumulées (notées respectivement par f_i et F_i où $i = 1, \dots, n$) de la variable n_visit.
5. Étudier ce tableau pour en déduire le mode de cette variable.

Partie 2

Nous nous intéressons maintenant à la liaison entre les deux variables weight et X. Nous cherchons à expliquer la variable weight en fonction de X.

1. Tracer le nuage de points de ces deux variables.
2. Pensez-vous qu'il existe une relation entre ces deux variables? Justifiez votre réponse.
3. Calculer et commenter le coefficient de corrélation entre ces deux variables. Cela est-il cohérent avec les résultats précédents?

Nous pouvons modéliser la relation, qui est linéaire, entre *weight* et *X* par une droite de la forme

$$\text{weight} = aX + b + \text{erreur},$$

où *erreur* est la variable qui mesure l'erreur d'observation (souvent supposée gaussienne). Les coefficients *a* et *b* sont inconnus et peuvent être estimés selon le critère des moindres carrés ordinaire.

4. Trouver les estimateurs des moindres carrés ordinaire de *a* et *b* qu'on notera respectivement par *a_o* et *b_o*.
5. Tracer sur le nuage de points (*X*, *weight*) la droite des moindres carrés (notée DMC)

$$\text{DMC} = a_oX + b_o.$$

6. Commenter les résultats obtenus.

Partie 3

Intéressons-nous maintenant aux variables *age*, *protein* et *protein2*.

1. Tracer le nuage de points entre *age* et *protein*. Quel est le type de la relation entre ces deux variables?
2. Tracer les histogrammes des variables *age* et *protein2*.
3. À votre avis, les coefficients d'asymétrie associés à ces deux variables sont-ils positifs ou négatifs? Vérifiez votre intuition en calculant le coefficient d'asymétrie de chaque variable.