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 rajouté.

Partie 1

1. Charger les données, sous forme de `DataFrame`, dans la variable `data`. Donner le type de chaque variable et vérifier s'il y'a des valeurs manquantes.
2. Stocker le nombre de variables dans la variable `n` et celui des 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 f_i et F_i où $i = 1, \ldots, n$) 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. Penser vous qu'il existe une relation entre ces deux variables ? Justifier 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 linaire, entre `weight` et `X` par une droite de la forme

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

où `erreur` est la variable qui mesure l'erreur d'observation (souvent supposée gaussienne). Les coefficient `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 `a_o` et `b_o`.
5. Tracer sur le nuage de points (`X`, `weight`) la droite des moindre carrée (notée `DMC`)

$$\texttt{DMC = a\_oX + b\_o.}$$

6. Tracer le nuage des points des erreurs résiduelles.
7. Commenter les résultats obtenues.

Partie 3

Intéressons nous maintenant aux variables `age`, `protein` et `protein2`.

1. Tracer le nuage de point 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 ? Vérifier votre intuition en calculant le coefficient d'asymétrie de chaque variable.