Résumé des méthodes de l’analyse bivariée de variables quantitatives :

Table des matières

1. Introduction

Le datascientist, dans son travail doit réaliser plusieurs types d'étude sur les jeux de données qui l'intéresse. Parmi elles, les analyses bivariés permettent de mettre en lumière les potentielles corrélations entre les différentes variables du jeu de donnée.

Nous verrons donc tout d'abord la réalisation de l'import des données puis la manière dont un data analyste réalise des analyses bivariées et plus particulièrement, des analyses bivariées entres variables quantitatives.

Dans le cadre de ce TP, les packages « Pandas », « Scipy », « Numpy » et « matplotlib » ont été utilisés. Le jeu de donnée lense.csv, représentant les prescriptions de lentilles de contacts en fonctions des recommandations, de l'âge, de l'astigmatie et du taux de larmes, composera la base de notre étude.

2. Import des données

Afin d'apporter les données à étudiant dans python, on utilise le package Pandas qui simplifie la création et la manipulation de dataframe.

Suite à l'import des packages,

```python
# Import des packages
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import chi2_contingency, chi2
```

Il suffit d'utiliser read_csv, une fonction de pandas, pour lire un csv et le récupérer sous forme de dataframe exemple :

```python
# Import du dataframe
df = pd.read_csv("data/lenses.txt", sep="\t")
```

Il en sort notre dataframe sur Python.

	Age	Prescription	Astigmatic	Tears	Recommendation
0	young	myope	no	reduced	none
1	young	myope	no	normal	soft
2	young	myope	yes	reduced	none
3	young	myope	yes	normal	hard
4	young	hypermetrope	no	reduced	none

3. Matrice de contingence

On réalise une matrice de contingence sur les effectifs :

```
# Réalisation de la matrice de contingence pour les effectifs
M_cont = pd.crosstab(df['Age'], df['Recommendation'], margins = True)
```

Recommendation	hard	none	soft	All
Age				
pre-presbyopic	1	5	2	8
presbyopic	1	6	1	8
young	2	4	2	8
All	4	15	5	24

Puis en fréquence :

```
# Réalisation de la matrice de contingence pour les fréquences
M_cont_eff = M_cont/M_cont.loc['All', 'All']
```

Recommendation	hard	none	soft	All
Age				
pre-presbyopic	0.041667	0.208333	0.083333	0.333333
presbyopic	0.041667	0.250000	0.041667	0.333333
young	0.083333	0.166667	0.083333	0.333333
All	0.166667	0.625000	0.208333	1.000000

Les fréquences semblent cohérentes. On retrouve 1 en somme de fréquences.

Nous avons donc les corrélations entre l'âge et les recommandations pour le port des lentilles.

On réalise ensuite l'étude par ligne :

Recommendation	hard	none	soft
Age			
pre-presbyopic	0.125	0.625	0.250
presbyopic	0.125	0.750	0.125
young	0.250	0.500	0.250

Puis part colonne :

Recommandation	hard	none	soft
Age			
pre-presbyopic	0.25	0.333333	0.4
presbyopic	0.25	0.400000	0.2
young	0.50	0.266667	0.4

On peut vérifier nos études axiales en sommant les lignes ou les colonnes (suivant les sens de l'étude). Si la somme vaut 1 pour toute les lignes/colonnes, on peut supposer que les tableaux

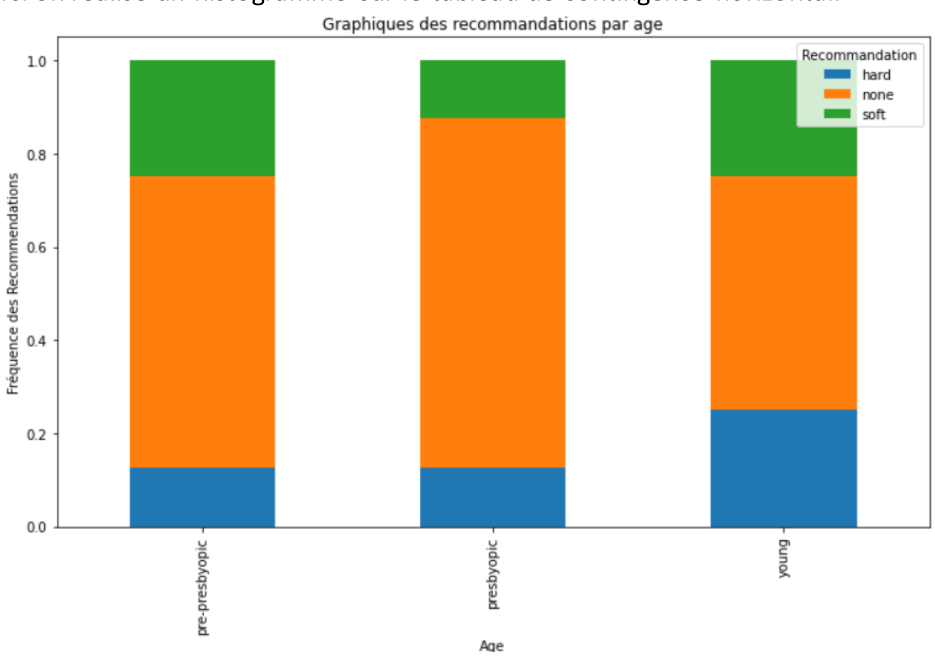
4. Approfondissement

Il existe plusieurs moyens plus ou moins pertinent suivant les cas, de comparer les variables quantitatives entres elles.

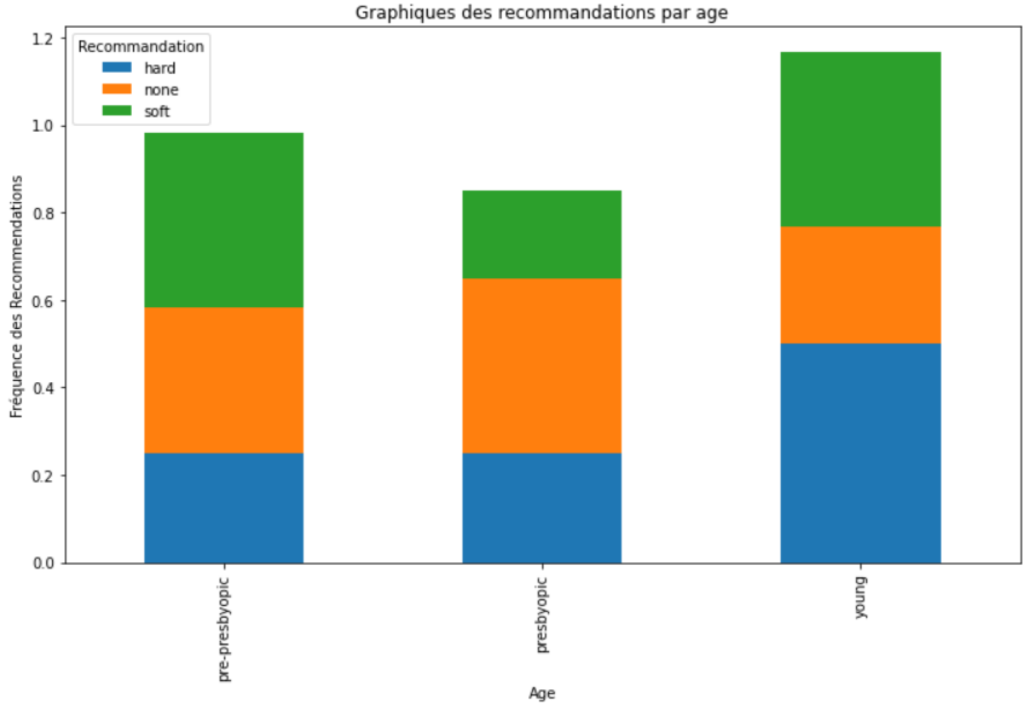
4.1. Graphiques

On peut ainsi réaliser des graphiques pour rendre l'analyse de données plus visuel. L'utilisation d'histogramme est la plus répandue du fait de sa clarté et son efficacité.

Ici on réalise un histogramme sur le tableau de contingence horizontal.



Grâce à ce graphique nous pouvons rapidement poser l'hypothèse que les jeunes ont plus de recommandations pour le port des lentilles que les plus âgés.



En réalisant le même graphique avec le tableau de contingence vertical on réalise encore plus l'écart de recommandation de port de lentille en fonction de l'âge.

Une valeur semble perturber le graph puisqu'un pays semble avoir bien trop d'étudiant. On « nettoie » ce dataset de cette valeur aberrante et réalisons une régression linéaire pour vérifier visuellement l'existence d'une potentielle corrélation entre les variables :

Les points semblent alignés de manière plutôt logique sur la droite de régression, il semble qu'il existe une corrélation entre la variable représentant le nombre d'étudiant et les dépenses du pays dans l'établissement.

4.2. Independence des variables et test du chi-2

On réalise un tableau de fréquence théorique :

Recommendation	hard	none	soft	All
Age				
pre-presbyopic	0.053333	0.2	0.066667	0.32
presbyopic	0.053333	0.2	0.066667	0.32
young	0.053333	0.2	0.066667	0.32
All	0.16	0.6	0.2	0.96

Fréquences théoriques

Recommendation	hard	none	soft	All
Age				
pre-presbyopic	0.041667	0.208333	0.083333	0.333333
presbyopic	0.041667	0.250000	0.041667	0.333333
young	0.083333	0.166667	0.083333	0.333333
All	0.166667	0.625000	0.208333	1.000000

Les valeurs entre le tableau de contingence en fréquence et le tableau de fréquence théoriques sont proches donc cela montre que les variables ont une forte indépendance.

Si on réalise la même chose avec les valeurs et non plus la fréquence, on arrive au même constat :

	hard	none	soft	All
Age				
pre-presbyopic	1.333333	5.0	1.666667	8.0
presbyopic	1.333333	5.0	1.666667	8.0
young	1.333333	5.0	1.666667	8.0
All	4.000000	15.0	5.000000	24.0

Valeurs théoriques

Recommendation	hard	none	soft	All
Age				
pre-presbyopic	1	5	2	8
presbyopic	1	6	1	8
young	2	4	2	8
All	4	15	5	24

Les valeurs théoriques et celles du tableau de contingence sont proches donc les variables semblent indépendantes.

Afin de certifier nos résultats nous pouvons réaliser le test du chi-2 :

```
# Calcul du test chi-2
chi_2, p, deg2lib, tab_freq = chi2_contingency(M_cont)
```

```
la valeur du chi_2 : 1.3
la p-valeur : 0.998376448363871
le degrée de liberté : 9
```

Comme nous avons un degré de liberté vaut 9, en se référant au tableau des seuils de corrélations :

Seuil	3,84	5,99	7,82	9,49	11,07	12,59	14,07	15,51	16,92
d.d.l.	1	2	3	4	5	6	7	8	9

On voit que le chi-2 est bien inférieurs au seuil qu'implique le degré de liberté. On en déduit donc que les variables sont indépendantes. Cette conclusion va aussi dans le sens des études réalisées précédemment lors de l'étude des valeurs prédites.