Résumé des méthodes de l'analyse bivariée mixte:

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 mixe 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 mixes.

Dans le cadre de ce TP, les packages « Pandas », « Scipy », « Numpy » et « matplotlib » ont été utilisés. Le jeu de donnée SalairesData.csv, représentant les salaires de membres d'entreprise en fonctions de l'âge, le sexe, la catégorie et l'établissement, 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,

```
# Import des packages
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

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

```
# import des datas SalairesData.csv
df = pd.read_csv('Data\SalairesData.csv', sep=';')
df.drop('Unnamed: 0', axis=1, inplace=True)
df.head()
```

Il

Ici on drop la colonne 'Unnamed : 0' car elle ne contient aucune information utile à l'étude. En sort notre dataframe sur Python.

	Sexe	Salaire	Categorie	Age	Etablissement
0	H	140	CS	58	A
1	F	120	CS	55	A
2	H	118	CS	50	C
3	H	117	CS	44	C
4	H	117	CS	45	B

3. Cas particulier : Étude du salaire en fonction de la catégorie

Nous allons étudier le cas particulier de l'étude de corrélation entre le salaire et la catégorie des salariés :

3.1. Création des sous populations

On construit des sous populations pour chaque possible catégorie :

```
# création des sous populations
df_categorie = df_categorie.drop('Age', axis=1)
df_categorie_CS = df_categorie[df_categorie['Categorie']== 'CS']
df_categorie_CM = df_categorie[df_categorie['Categorie']== 'CM']
df_categorie_OE = df_categorie[df_categorie['Categorie']== 'OE']
# retire la colonne Age de df_categorie_CS
df_categorie_OE.head()
```

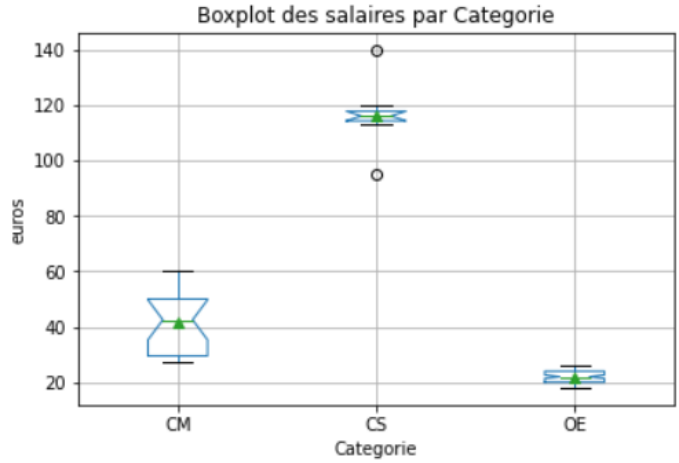
	Categorie	Salaire
30	OE	26
31	OE	26
32	OE	26
33	OE	26
34	OE	26

On étudie ensuite les indicateurs de position et de dispersion :

```
Indicateurs de position:
Moyenne:
 Salaire    32.038462
dtype: float64

Médiane:
 Salaire    23.0
dtype: float64

Quantiles:
      Salaire
0.25     21.0
0.50     23.0
0.75     26.0

Indicateurs de dispersion:
Écart-type:
 Salaire    26.085252
dtype: float64

Variance:
 Salaire    680.44037
dtype: float64

Écart interquartile:
 Salaire    5.0
dtype: float64
```

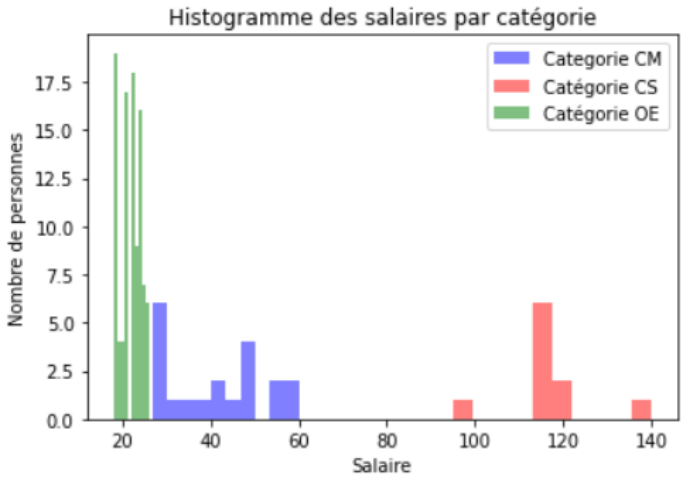
En réalisant une étude avec les boîtes de Tukey, on réalise plus visuellement qu'en fonction des catégories, la médiane, les quantiles et la moyenne varie beaucoup :

```
# boite de Tukey des Salaire par Categorie
df_categorie.boxplot(column='Salaire', by='Categorie', showmeans=True, notch =
True)
plt.ylabel('euros')
plt.suptitle('')
plt.title('Boxplot des salaires par Categorie')
plt.show()
```



On peut également réaliser un histogramme pour accentuer le rendu visuel des écarts de salaires entre les catégories :

```
# histogramme pour la variables Categorie pour le salaire et l'age
df_categorie_CM['Salaire'].plot(kind='hist', bins=10, color='blue', alpha=0.5,
label='Categorie CM')
df_categorie_CS['Salaire'].plot(kind='hist', bins=10, color='red', alpha=0.5,
label='Catégorie CS')
df_categorie_OE['Salaire'].plot(kind='hist', bins=10, color='green', alpha=0.5,
label='Catégorie OE')
plt.xlabel('Salaire')
plt.ylabel('Nombre de personnes')
plt.title('Histogramme des salaires par catégorie')
plt.legend()
plt.show()
```



4. Approfondissement

Il est courant, afin d'expliquer les corrélations entre les données mathématiquement de calculer les rapports de corrélations entre les variables. Pour ce faire, on calcule la variance intra et la variance inter.

4.1. Étude de la variance intra :

La variable intra se calcule à l'aide des variances et effectifs de chaque sous populations ainsi que la population totale.

```
var_intra = (eff_A*var_A +eff_B*var_B+eff_C*var_C)/(eff_C+eff_B+eff_A)
variance : Salaire    116.933333
dtype: float64
effectif 10
variance : Salaire    135.726316
dtype: float64
effectif 20
variance : Salaire    6.037475
dtype: float64
effectif 100


 variation intra :  Salaire    34.520055
dtype: float64
```

Nous trouvons ici une variance intra de 34.520055

4.2. Étude de la variance inter

La variable inter se calcule à l'aide des moyennes et effectifs de chaque sous populations ainsi que la moyenne et la population totale.

```
moy_global = (1/(eff_C+eff_B+eff_A))/(eff_A*moy_A +eff_B*moy_B+eff_C*moy_C)
variation_inter = (eff_A*(moy_A-moy_global)+eff_B*(moy_B-
moy_global)+eff_C*(moy_C-moy_global))/(eff_C+eff_B+eff_A)

 variation inter :  32.03845969156894
```

Nous trouvons ici une variance inter de 32.03845969156894

4.3. Étude de la variance globale et de corrélation

La variance globale permet de calculer la variance de corrélation nécessaire pour se rendre compte numériquement le taux de corrélation entre deux variables. Ici nous étudiant les salaires en fonction des catégories et d'après nos étude graphiques et numérique précédentes, nos deux variables doivent avoir un bon taux de corrélation.

Pour calculer la variance globale, il suffit de sommer la variance intra et la variance inter :

```
# Calcule de la variation globale
var_global = var_intra['Salaire']+variation_inter
```
```
Variance globale : 503.9654737031984
```

On calcule ensuite la variance de corrélation en divisant la variance inter par la variance globale :

```
# Rapport de corrélation
print('Rapport de corrélation : ', variation_inter/var_global)
```
```
Rapport de corrélation :  0.4813577937666804
```

On obtient comme rapport de corrélation 0.481357793766684 ce qui s'interprète comme 0.48% des variations de salaire s'explique par la catégorie. Le rapport de corrélation est élevé, surtout en comparant le salaire avec les autres variables :

Rapports de corrélations :

- En fonction du sexe : `Rapport de corrélation :  0.04528583938387755`
- En fonction de l'établissement : `Rapport de corrélation :  0.04497901124403027`
- En fonction de la catégorie : `Rapport de corrélation :  0.4813577937666804`
- En fonction de l'âge : `Rapport de corrélation :  0.06357272742544547`

Grâce à tout ces rapports de corrélations, on peut déduire que c'est les catégorie des salariés qui impacte largement le plus directement leur salaire

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.