

Les méthodes Chi2, la corrélation de Pearson et l'ANOVA

Méthode Chi2 (Chi carré) :

- Utilisation : La méthode Chi2 est utilisée pour évaluer la relation entre deux variables qualitatives (catégorielles) dans un tableau de contingence.
- Exemple avec des données qualitatives et qualitatives : Supposons que nous ayons un dataframe contenant des données sur des étudiants, avec les colonnes "Genre" (masculin/féminin) et "Préférence de musique" (rock/pop/jazz). Nous pouvons créer un tableau de contingence pour analyser la relation entre ces deux variables.

```
import pandas as pd
from scipy.stats import chi2_contingency

# Création du dataframe
data = {
    'Genre': ['F', 'M', 'M', 'F', 'F', 'M', 'F', 'M', 'F', 'F'],
    'Préférence de musique': ['Rock', 'Pop', 'Jazz', 'Rock', 'Rock',
                              'Pop', 'Pop', 'Jazz', 'Rock', 'Jazz']
}
df = pd.DataFrame(data)

# Création du tableau de contingence
contingency_table = pd.crosstab(df['Genre'], df['Préférence de
musique'])

# Application de la méthode Chi2
chi2, p_value, _, _ = chi2_contingency(contingency_table)

# Affichage des résultats
print("Statistique Chi2 :", chi2)
print("P-value :", p_value)
```

- Les résultats nous indiqueront la statistique Chi2 et la p-value. Si la p-value est inférieure à un certain seuil (par exemple 0,05), cela suggère qu'il y a une relation significative entre les variables "Genre" et "Préférence de musique".

Corrélation de Pearson :

- Utilisation : La corrélation de Pearson mesure la relation linéaire entre deux variables quantitatives continues.
- Exemple avec des données quantitatives et quantitatives : Supposons que nous ayons un dataframe contenant des données sur des étudiants, avec les colonnes "Âge" et "Score d'examen". Nous pouvons calculer la corrélation de Pearson pour évaluer la relation entre ces deux variables.

```

import pandas as pd
from scipy.stats import pearsonr

# Création du dataframe
data = {
    'Âge': [20, 22, 21, 19, 25, 23, 18, 20, 24, 22],
    'Score d\'examen': [85, 80, 90, 70, 95, 92, 65, 75, 98, 88]
}
df = pd.DataFrame(data)

# Calcul de la corrélation de Pearson
correlation, p_value = pearsonr(df['Âge'], df['Score d\'examen'])

# Affichage des résultats
print("Corrélation de Pearson :", correlation)
print("P-value :", p_value)

```

- Les résultats nous donneront le coefficient de corrélation de Pearson et la p-value associée. Une valeur proche de 1 indique une corrélation positive forte, une valeur proche de -1 indique une corrélation négative forte, et une valeur proche de 0 indique une absence de corrélation linéaire.

ANOVA (Analyse de la variance) :

- Utilisation : L'ANOVA est utilisée pour comparer les moyennes de plusieurs groupes et déterminer s'ils sont statistiquement différents.
- Exemple avec des données quantitative et qualitative : Supposons que nous ayons un dataframe contenant des données sur des étudiants, avec les colonnes "Groupe" (A/B/C) et "Score d'examen". Nous pouvons utiliser l'ANOVA pour déterminer s'il y a une différence significative des scores d'examen entre les groupes.

```

import pandas as pd
from scipy.stats import f_oneway

# Création du dataframe
data = {
    'Groupe': ['A', 'A', 'B', 'B', 'C', 'C'],
    'Score d\'examen': [80, 85, 90, 88, 75, 80]
}
df = pd.DataFrame(data)

# Application de l'ANOVA
groups = []
for group, scores in df.groupby('Groupe')['Score d\'examen']:

```

```
groups.append(scores)
f_value, p_value = f_oneway(*groups)
```

```
# Affichage des résultats
print("Valeur F :", f_value)
print("P-value :", p_value)
```

- Les résultats nous donneront la valeur F et la p-value. Si la p-value est inférieure à un certain seuil (par exemple 0,05), cela suggère qu'il y a une différence significative des scores d'examen entre les groupes.