PRATIQUE

- Corrélation de Pearson

Le coefficient de corrélation de Pearson (du nom de Karl Pearson) peut être utilisé pour résumer la force de la relation linéaire entre deux échantillons de données.

Le coefficient de corrélation de Pearson est calculé comme la covariance des deux variables divisée par le produit de l'écart type de chaque échantillon de données. C'est la normalisation de la covariance entre les deux variables pour donner un score interprétable.

L'utilisation de la moyenne et de l'écart type dans le calcul suggère la nécessité pour les deux échantillons de données d'avoir une distribution gaussienne ou de type gaussienne.

Le résultat du calcul, le coefficient de corrélation peut être interprété pour comprendre la relation.

Le coefficient renvoie une valeur comprise entre -1 et 1 qui représente les limites de corrélation d'une corrélation négative complète à une corrélation positive complète. Une valeur de 0 signifie aucune corrélation. La valeur doit être interprétée, où souvent une valeur inférieure à -0,5 ou supérieure à 0,5 indique une corrélation notable, et des valeurs inférieures à ces valeurs suggèrent une corrélation moins notable.

La fonction pearsonr() SciPy peut être utilisée pour calculer le coefficient de corrélation de Pearson entre deux échantillons de données de même longueur.

Nous pouvons calculer la corrélation entre les deux variables dans notre problème test.

```
# calculate the Pearson's correlation between two variables
from numpy.random import randn
from numpy.random import seed
from scipy.stats import pearsonr
# seed random number generator
seed(1)
# prepare data
data1 = 20 * randn(1000) + 100
data2 = data1 + (10 * randn(1000) + 50)
# calculate Pearson's correlation
corr, _ = pearsonr(data1, data2)
print('Pearsons correlation: %.3f' % corr)
```

- Corrélation de Spearman

eux variables peuvent être liées par une relation non linéaire, de sorte que la relation est plus forte ou plus faible dans la distribution des variables.

De plus, les deux variables considérées peuvent avoir une distribution non gaussienne.

Dans ce cas, le coefficient de corrélation de Spearman (du nom de Charles Spearman) peut être utilisé pour résumer la force entre les deux échantillons de données. Ce test de relation peut également être utilisé s'il existe une relation linéaire entre les variables, mais aura un peu moins de puissance (par exemple, peut entraîner des scores de coefficient inférieurs).

Comme pour le coefficient de corrélation de Pearson, les scores sont compris entre -1 et 1 pour les variables parfaitement corrélées négativement et parfaitement corrélées positivement respectivement.

Au lieu de calculer le coefficient en utilisant la covariance et les écarts types sur les échantillons eux-mêmes, ces statistiques sont calculées à partir du rang relatif des valeurs sur chaque échantillon. Il s'agit d'une approche courante utilisée dans les statistiques non paramétriques, par exemple les méthodes statistiques où nous ne supposons pas une distribution des données telle que gaussienne.

```python
# calculate the spearmans's correlation between two variables
from numpy.random import randn
from numpy.random import seed
from scipy.stats import spearmanr
# seed random number generator
seed(1)
# prepare data
data1 = 20 * randn(1000) + 100
data2 = data1 + (10 * randn(1000) + 50)
# calculate spearman's correlation
corr, _ = spearmanr(data1, data2)
print('Spearmans correlation: %.3f' % corr)
```