# Covariance linéaire

## Présentation

La covariance **indique ou mesure le degré de dépendance** entre deux variables X et Y.

Plus la dépendance entre X et Y est grande plus la covariance est forte.

**Le fait que deux variables est une grande covariance ne démontre pas qu'il y ait une relation de causalité entre l'une et l'autre.** Dans la littérature on prend souvent l'exemple d'un dataset pour un pays donné des cygognes et du nombre de naissances. On montre alors que la corrélation est forte entre ces deux variables : Nombre de cygognes & Naissances. Il est évident ici qu'il n'y a pas de relation entre corrélation et causalité.

Deux variables X et Y peuvent être liées par une relation mathématiques de type affine, puissance ou même exponentiel.

## Relation affine

La covariance de X et Y deux variables (aléatoires) est définie par :

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{x})}{n}$$

Que l'on peut écrire également sous cette forme :

$$Cov(X,Y) = \frac{\sum_{i=1}^{n} x_i \times y_i}{n} - \bar{x} \times \bar{y}$$

La covariance peut être représentée par un nuage de points, lorsqu'il y a une relation linéaire, la corrélation est forte et une droite passera en moyenne assez proche de tous les points.

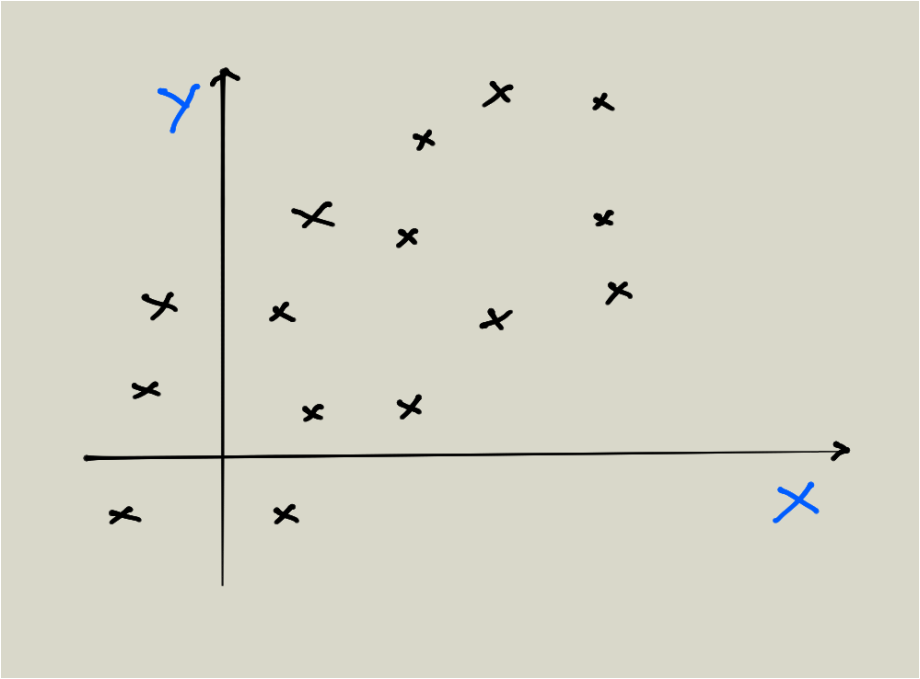


Figure 1: nuage de points

## Centre de gravité

Les axes $\bar{x}$ et $\bar{y}$ représentent le centre de gravité du nuage de point.

Dans les deux situations suivantes les variables X et Y sont dépendantes.

Des produits positifs s'additionnent et donne une quantité non négligeable positive. Les variables X et Y sont donc apparamment dépendantes.

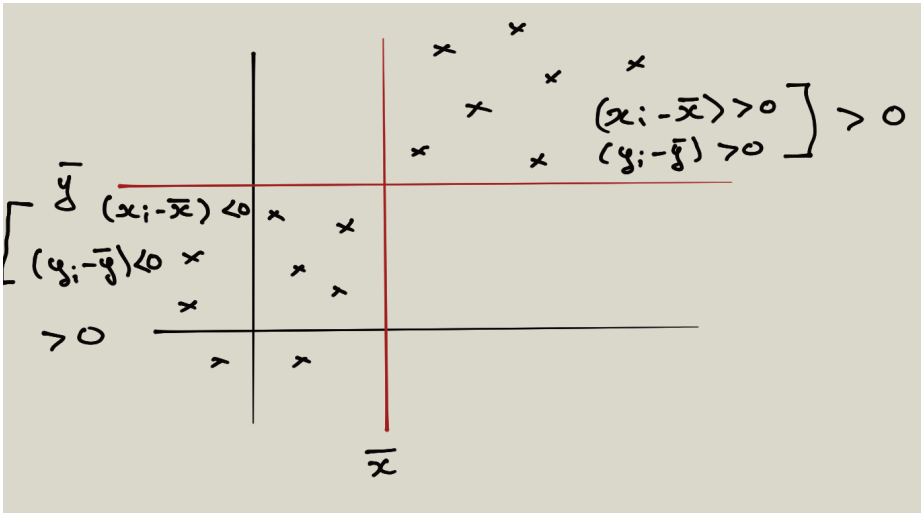


Figure 2: nuage de points

Inversement des quantités négatives peuvent s'additionner. Dans ce cas également on dira que les variables X et Y sont dépendantes négativement.

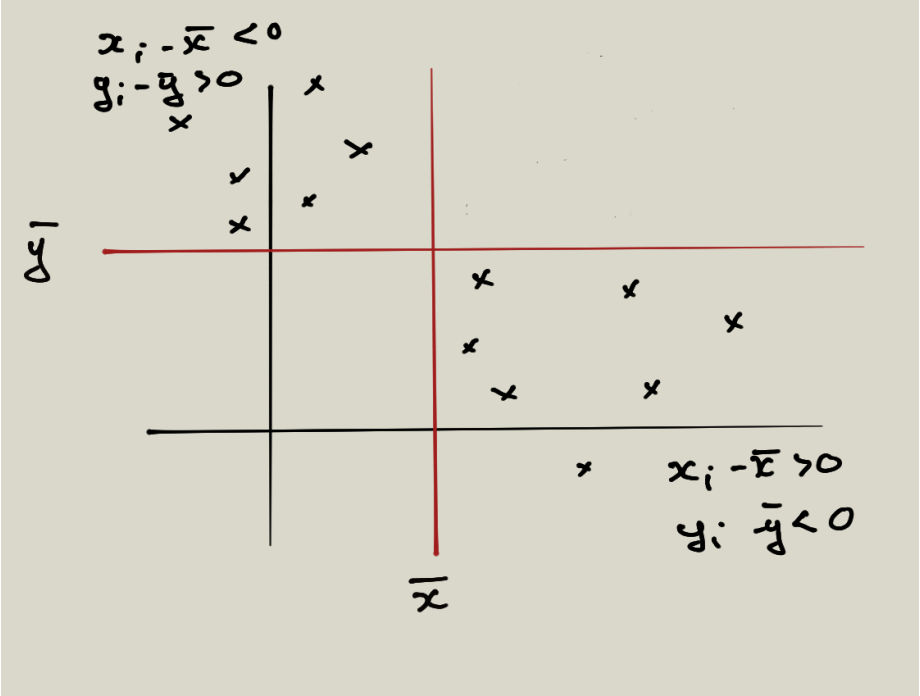


Figure 3: nuage de points

## Exemple de deux variables dépendantes

Ci-dessous le prix des loyers en fonction de la surface en m2 dans une ville de France.

| Surface en m2 | Prix des loyers |
|---|---|
| 19 | 280 |
| 30 | 485 |
| 39 | 615 |
| 52 | 670 |
| 65 | 865 |
| 79 | 1080 |
| 92 | 1260 |
| 108 | 1645 |

## Exercice calculer la covariance

Créez un DataFrame à partir des valeurs suivantes et créez une fonction cov(x,y) permettant de calculer la covariance de deux variables. Attention il existe une méthode **cov** dans Pandas, celle-ci vous donnera la covariance non-biaisée statistique ne l'utilisez pas pour l'instant.

```
import numpy as np
import pandas as pd

area = [19,30,39,52,65,79,92,108]
price = [280 ,485  ,615  ,670  ,865  ,1080 ,1260 ,1645 ]

df = pd.DataFrame({'area' : area, 'price' : price})

def cov(x,y):
    pass
```

Supposons maintenant que l'on change l'unité de mesure de la surface et que l'on prenne des cm pour mesurer la surface des habitations. Que pouvez-vous dire de la covariance que vous avez calculer précédemment ?

Lorsque vous changez d'unité le calcul d'une ou des variables alors la valeur de la covariance change et celle-ci peut-être selon l'unité très différente.

## Normaliser les valeurs

Nous allons introduire la notion de cœfficient corrélation linéaire qui est un réel compris entre -1 et 1. Il permet de mettre nos variables dans le même rapport de valeur, dans ce cas l'unité n'a plus d'effet sur nos calculs.

$r = \frac{Cov(x,y)}{\sigma_x \times \sigma_y}$ Ce cœfficient n'a pas d'unité de mesure et est toujours compris entre -1 et 1. On l'appelle **cœfficient de corrélation de Pearson.**

On admettra que si r est proche de 1 ou -1 il y a une forte corrélation linéaire entre les deux variables X et Y.

Voici un tableau qui vous aidera dans l'interprétation des valeurs de ce cœfficient :

| Cœfficient | Interprétation |
|---|---|
| 0 | pas de corrélation |
| -0.25 | Faible corrélation <0 |
| -0.75 | Forte corrélation <0 |
| -1.0 | Parfaite corrélation <0 |
| 0.25 | Faible corrélation >0 |
| 0.75 | Forte corrélation >0 |

| Cœfficient | Interprétation |
|---|---|
| 1.0 | Parfaite corrélation <0 |

**Exercice fonction de corrélation**

Créez une fonction **corr** en Python permettant de calculer la corrélation.

```
def corr(x,y):
    pass
```

Reprennez l'exercice précédent et déterminez la corrélation entre les deux variables area et price. Que pouvez-vous conclure ?

## Exercice d'application avec Pandas

Soit les données suivantes que l'on représente en nuage de points comme suit :

```
%matplotlib inline
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

df = pd.DataFrame({
    'X' :  np.arange(5)+3,
    'Y' : [1, 3, 4, 8, 12]
})

plt.scatter(
    df['X'], df['Y'],
    s=200,
    c='red',
    marker = '*',
    edgecolors = 'blue'
)

plt.show()
```

Avec Pandas vous pouvez calculer le cœfficient de Pearson de corrélation de deux variables, cette méthode vous retournera la matrice de corrélation, elle est symétrique.

```
df.corr(method='pearson')
```

Donnez le cœfficient de corrélation des variables X et Y.

Que pouvez-vous dire des variables X et Y ?

## Exemples de corrélation forte

```
x = [-5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5]
y = [-5, -4.1, -3, -2, -1.1, 0, 1.3, 2.2, 3, 4.1, 5]

df2 = pd.DataFrame({'x' : x, 'y' : y})

df2.corr(method='pearson')
```

| | X | Y |
|---|---|---|
| X | 1 | 0.999461 |
| Y | 0.999461 | 1 |

## Graphique de présentation de la corrélation

Vous pouvez utiliser un graphique pour présenter vos corrélations. Ci-dessous une représentation de valeurs uniformément entre 0 et 20 distrubuées :

```
%matplotlib inline
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

plt.figure(figsize=(15, 10))
uniformData = np.random.rand(10, 10)
ax = sns.heatmap(uniformData, annot=True)

plt.show()
```

## Exercice d'application Titanic

Prennez le dataset du Titanic dans Seaborn et étudiez les corrélation sur les variables qui vous semble pertinente.

```
%matplotlib inline
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

titanic = sns.load_dataset('titanic')
```

## Droite de régression

Rechercher une droite de régression linéaire, c'est rechercher une relation linéaire entre les variables X et Y. Dis autrement c'est rechercher une relation d'explication linéaire de Y par X. Par exemple on pourrait essayer d'expliquer l'espérance de vie Y en fonction de la natalité X avec un jeu de données correspodant.

On cherche à miniser la somme des distances en bleu ci-dessous. Ce sont les projections des points sur la droite D recherchée. Cette droite passe par un point remarquable qui est le centre de gravité du nuage de points noté G et qui a pour coordonnées la moyenne des X et Y (nos variables) : $G(\bar{X}, \bar{Y})$

On cherche ici une relation affine entre nos deux variables X et Y. On démontre que le cœfficient directeur de cette droite est donnée par la formule suivante :

$y = a \times x + b$ avec $a = \frac{Cov(X,Y)}{Var(X)}$

De plus on sait que cette droite passe par le centre de gravité du nuage de points. On a donc deux relations permettant de calculer les deux paramètres a et b et donc de déterminer l'équation de la droite (D) de régression.

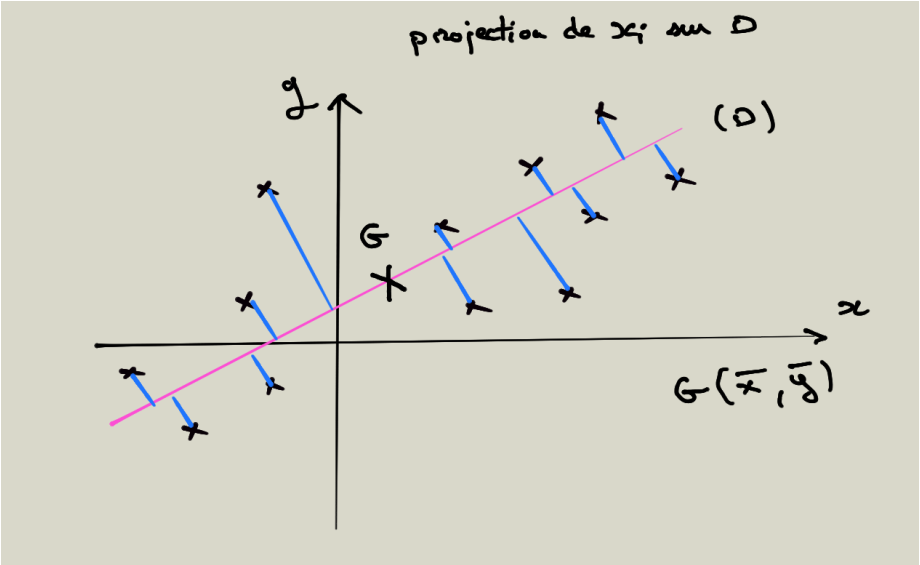


Figure 4: regression linéaire