26.1_STATISTIQUE BIVARIÉE
LIEN ENTRE DEUX VARIABLES.

SOMMAIRE

- Variance
- Covariance
- Coefficient de corrélation linéaire
- Test d'indépendance du Khi-

COVARIANCE ET COEFFICIENT DE CORRÉLATION LINÉAIRE :

La variance :

- Elle permet de mesurer la dispersion d'une variables autour de ça moyenne.
- La variance est toujours positive, et ne s'annule que si les valeurs sont toutes égales.
- Elle se calcule comme suit :

$$V(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Avec $\bar{x}$ la moyenne et x_1, x_2, x_3, …, les valeur de la série statistique.

Exemple :
Résultats d'une course avec le poids des participants.

```
print(filles)
```

	temps	poids
0	25.416667	48
1	25.666667	50
2	25.783333	55
3	25.916667	55
4	26.250000	57
5	26.333333	62
6	26.583333	60
7	26.750000	62
8	27.000000	63
9	27.333333	68

```
print(garçons)
```

	temps	poids
0	19.333333	68
1	19.500000	85
2	20.833333	77
3	19.200000	72
4	20.500000	85
5	19.716667	80
6	19.850000	77
7	21.000000	85
8	20.666667	69
9	21.500000	70

```
print(filles.var())
```

```
temps     0.382148
poids    38.222222
dtype: float64
```

```
print(garçons.var())
```

```
temps     0.626123
poids    46.622222
dtype: float64
```

La Covariance

- La covariance entre deux variables est un nombre permettant de quantifier leurs écarts conjoints par rapport à leurs espérances respectives.

Elle est donnée par la formule suivante :

$$\mathrm{Cov}(X, Y) = \sum_{i,j}(x_i - \bar{x})(y_i - \bar{y})f_{i,j}$$

Avec fi,j la fréquence d'apparition du couple $(xi\ ,\ yi)$. (Dans notre exemple toujours = à 1).

Intuitivement, la covariance caractérise les variations simultanées de deux variables aléatoires :

- Elle sera positive lorsque les écarts entre les variables et leurs moyennes ont tendance à être de même signe.
- Négative dans le cas contraire.
- Nulle si les deux variables sont indépendantes

```
print(filles.cov())
```

	temps	poids
temps	0.382148	3.664815
poids	3.664815	38.222222

```
print(garçons.cov())
```

	temps	poids
temps	0.626123	0.192963
poids	0.192963	46.622222

COEFFICIENT DE CORRÉLATION LINÉAIRE DE BRAVAISPEARSON

La covariance est une grandeur non bornée.
Alors il est difficile à un lien entre variables.
Pour avoir une meilleur idées du lien entre variables on va normaliser la covariance en la divisant par le produit des deux écart-type.

On obtient le coefficient de corrélation linéaire de Bravais-Pearson :

$$r = \frac{Cov(X,Y)}{\sqrt{V(X)V(Y)}}$$

La valeur obtenue est bornée par -1 et 1 :

- Si proche de 0 les variables sont non-corrélées.
- Si proche de 1 les variables sont corrélées.
- Si proche de -1 les variables sont anti-corrélées.

```
print(filles.corr())
```

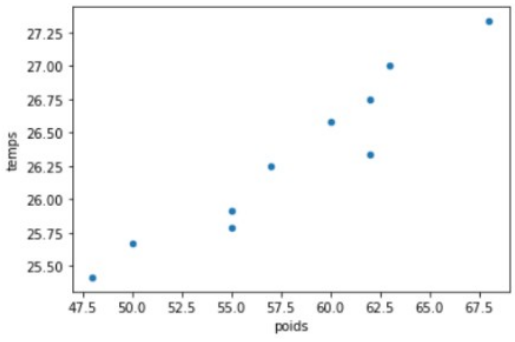
	temps	poids
temps	1.000000	0.958911
poids	0.958911	1.000000

```
print(garçons.corr())
```

	temps	poids
temps	1.000000	0.035715
poids	0.035715	1.000000

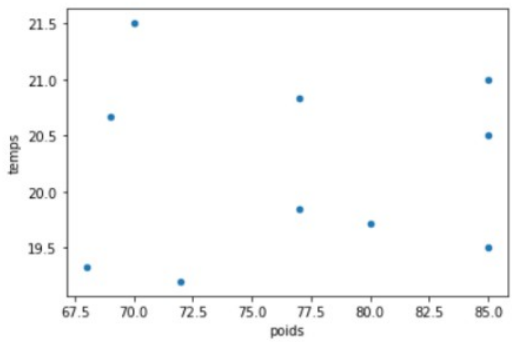
```
import matplotlib.pyplot as plt

filles.plot(kind='scatter',x='poids',y='temps')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x14a68b35508>
```




```
garçons.plot(kind='scatter',x='poids',y='temps')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x14a68b3dc88>
```



TEST D'INDÉPENDANCE DU KHI-2

Le Test d'indépendance du Khi-2 sert à étudier la liaison entre caractères qualitatifs X et Y en comparant leurs effectifs respectifs.

Procédure de test :

1) On définit les hypothèses testées :
   a) H 0 : Les varibles X et Y sont indépendantes
   b) H 1 : Les varibles X et Y ne sont pas indépendante

2) On vérifie les condition d'application du test:
   a) Effectif total de l'échantillon >= 50.
   b) Le produit de l'effectif des modalités pris 2 à 2 pour chacune des variables doit être 5x plusélevés que l'effectif total.

3) Statistique de test : sert à calculer la P-value.

4) Règles de décision On accepte ou on rejette $H0$ en fonction du seuil de risque choisi (généralement 5%).

Exemple :
Le tableau suivant représente la répartition des étudiants français ayant bénéficié du programme Erasmus de 2002/2003 à 2005/2006.

```
print(erasmus)
```

	_2002_2003	_2003_2004	_2004_2005	_2005_2006
Espagne	4470	5115	5167	5481
Royaume-Uni	4705	4652	4564	4499
Allemagne	2808	2804	2863	2884
Italie	1415	1549	1571	1642

On définit les hypothèses testées :

1) On définit les hypothèses testées :
   a) Hypothèse nulle H 0 : L ′ année et le pays d ′ accueil sont indépendantes
   b) Hypothèse alt H 1 : L′année et le pays d ′ accueil ne sont pas indépendantes

2) On vérifie les condition d'application du test:

```
somme = erasmus.sum().sum()
print(somme)
```

```
56189
```

```
somme_annee = erasmus.sum()
print(somme_annee)
```

```
_2002_2003    13398
_2003_2004    14120
_2004_2005    14165
_2005_2006    14506
dtype: int64
```

```
somme_pays = erasmus.T.sum()
print(somme_pays)
```

```
Espagne        20233
Royaume-Uni    18420
Allemagne      11359
Italie          6177
dtype: int64
```

```
Espagne = []
Royaume_Uni = []
Allemagne = []
Italie = []

for i in range(len(somme_annee)):
    Espagne.append(somme_pays[0]*somme_annee[i]/5)
    Royaume_Uni.append(somme_pays[1]*somme_annee[i]/5)
    Allemagne.append(somme_pays[2]*somme_annee[i]/5)
    Italie.append(somme_pays[3]*somme_annee[i]/5)
```

Rafik

```
erasmus = pd.DataFrame([Espagne, Royaume_Uni, Allemagne, Italie],
                       index=pays, columns=['_2002_2003', '_2003_2004',
                       '_2004_2005', '_2005_2006'])

print(erasmus)
```

	_2002_2003	_2003_2004	_2004_2005	_2005_2006
Espagne	54216346.8	57137992.0	57320089.0	58699979.6
Royaume-Uni	49358232.0	52018080.0	52183860.0	53440104.0
Allemagne	30437576.4	32077816.0	32180047.0	32954730.8
Italie	16551889.2	17443848.0	17499441.0	17920712.4

```
import scipy.stats as st

st_chi2, st_p, st_dof, st_exp = st.chi2_contingency(erasmus)
print(st_p)
```

```
8.060723213207895e-15
```

Ce qui nous intéresse ici, c'est la variable st_p, qui contient la P-value. Cette valeur nous donne la probabilité de rejeter $H0$ en ayant tort. Généralement si elle est inférieur à 5% on rejette $H0$ au profit de $H1$.