Table des matières

Statistiques

Dans ce chapitre, et uniquement ce chapitre, l'usage de la calculatrice/de l'ordinateur est autorisée.

I Série statistique

Définition 1. La statistique descriptive étudie certaines caractéristiques d'une population.
Ces caractéristiques sont appelées **caractères** ou **variables statistiques** .
Les éléments de cette population sont appelés **individus** ou **unités statistiques** .
Une série statistique de taille N est donc un N-uplets de caractères : $(x_1, \dots, x_N)$, où x_i est l'observation effectuée sur l'individu i.

Exemple 1. Pour étudier les terres agricoles d'une région, on peut faire l'inventaire de N exploitations agricoles qui s'y trouvent, et noter pour chacune d'elles sa taille (en hectares).

- Chacune des exploitations est un individu, ou une unité statistique.

- La taille (en ha) est la variable statistique (ou le caractère) étudiée.

Le résultat de la mesure du caractère "taille" sur les individus "exploitations" est un ensemble de N nombres, appelé **série statistique brute**.
Sur la même population, on peut s'intéresser à d'autres variables. On présente alors souvent les résultats sous forme d'un tableau "individus / caractères" :

$n°$ de l'exploitation	Taille (ha)	Culture dominante	Nombre de personnes employées	Qualité du produit (inspections)
1	50	blé	2	excellente
2	50.5	vigne	4	mauvaise
3	35	orge	3	normale
4	62.1	blé	6	très mauvaise
5	20	vigne	1	bonne
6	10	vigne	1	normale
7	20	maïs	2	bonne
8	33	blé	6	excellente
9	56	blé	2	bonne
10	39	orge	4	normale

Dans le tableau représenté ci-dessus, il y a 10 individus et 4 variables.

Définition 2. Type de variable :

- Une variable **qualitative** est une variable qui ne prend pas de valeur numérique. On distingue deux types de variables qualitatives :
  - ⋆ les valeurs ne peuvent pas être ordonnée : variable qualitative nominale
  - ⋆ les valeurs peuvent être ordonnées : variable qualitative ordinale
- Une variable **quantitative** est une variable qui prend uniquement des valeurs numériques. On distingue deux types de variables quantitatives :
  - ⋆ le nombre de valeurs est au plus dénombrable : variable quantitative discrète
  - ⋆ la variable peut prendre toutes les valeurs d'un intervalle : variable quantitative continue

Exemple 2. Dans l'exemple précédent,

- variable qualitative nominale : Type de produit
- variable qualitative ordinale : Qualité du produit
- variable quantitative discrète : Nombre de personnes employées
- variable quantitative continue : Taille en ha.

II Effectifs, fréquences

Définition 3. Effectifs, fréquences.
Le nombre d'individus pour lequel la variable prend la valeur A est appelé effectif
La somme de tous les effectifs est appelé effectif total
Le rapport de l'effectif de la valeur A sur l'effectif total est appelé fréquence

Exemple 3. Reprenons la population de l'exemple précédent, et considérons la variable "culture dominante".

- L'effectif de la valeur "blé" est : 4
- L'effectif total est :10
- La fréquence de la valeur "blé" est : $\frac{4}{10}$

Répondre aux mêmes questions pour les autres valeurs prises par la variable "culture dominante" :

Culture dominante	blé	vigne	orge	maïs
Effectif	4			
Fréquence	$\frac{4}{10}$			

Définition 4. Effectifs cumulés, fréquences cumulées.
Lorsque les valeurs peuvent être ordonnées, on peut définir les fréquences cumulées en ajoutant à chaque fréquences les fréquences des valeurs précédentes.

Le polygône des effectifs cumulés est la ligne brisée qui relie les points ayant pour abscisse la valeur du caractère, et en ordonnée son effectif.

Les fréquences cumulées s'obtiennent de la même façon en sommant les fréquences successives (ou alors en divisant les effectifs cumulés par l'effectif total).

La fonction qui à chaque modalité associe la fréquence cumulée est appelée **fonction de répartition**

Exemple 4. Reprenons l'exemple précédent, et considérons la variable "nombre de personnes employées". Compléter le tableau suivant, puis tracer le polygône des effectifs cumulés et la fonction de répartition.

Nombre de personnes employées	1	2	3	4	5	6
Effectif	2	3				
Fréquence	$\frac{2}{10}$	$\frac{3}{10}$				
Effectif cumulé	2	5				
Fréquence cumulée	$\frac{2}{10}$	$\frac{5}{10}$				

Regroupement par classes

Les variables quantitatives (en particulier pour les variables continues) peuvent être regroupées en **classes**, c'est-à-dire qu'on regroupe les valeurs selon l'intervalle auquel elles appartiennent. L'amplitude de la classe est alors la taille de l'intervalle de la classe.

Exemple 5. Reprenons la population de l'exemple des exploitations agricoles, et considérons la variable "taille de l'exploitation". Compléter le tableau ci-dessous, où les effectifs sont rangés par classes :

Taille de l'exploitation (ha)	$[0;20[$	$[20;40[$	$[40;60[$	$[60;80[$
Effectif	1	5		
Effectif cumulé	1	6		

On peut aussi choisir de séparer en deux l'intervalle $[20,40[$ (qui contient plus de la moitié de l'effectif) pour plus de précision :

Taille (ha)	$[0;20[$	$[20;30[$	$[30;40[$	$[40;60[$	$[60;80[$
Effectif					
Effectif cumulé					

On est alors confronté au problème du choix du nombre de classes : prendre un grand nombre de classe rend malaisé la lecture des classes (tableau ou représentation graphique), mais préserve un maximum d'informations (éventuellement trop, car certains phénomènes marginaux ou secondaires peuvent masquer la présence d'autres plus essentiels) ; en revanche, ne retenir que très peu de classes favorise la simplicité de la lecture des résultats mais peut donner une vision simpliste voire caricaturale de la population étudiée

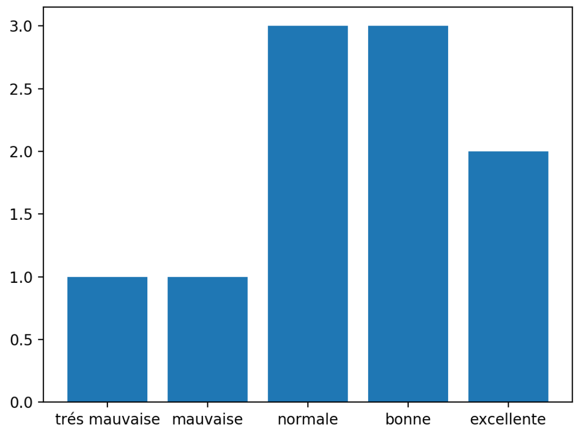
II. 1 Représentations graphiques

Variables qualitatives On utilise principalement deux représentations.

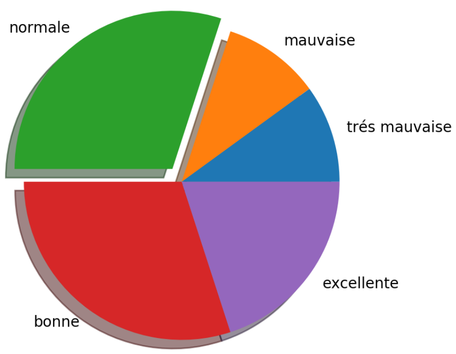
- Le **diagramme en barres** : chaque variable est représentée par un rectangle de base constante, et de hauteur (et donc de surface) proportionnelle à son effectif.
- Le **diagramme circulaire** : chaque variable est représentée par un secteur angulaire dont l'angle (et donc la surface) est proportionnel à son effectif.

Exemple 6. Dans l'exemple des exploitations agricoles, représenter la variable "culture dominante" sous forme de diagramme circulaire, et la variable "qualité du produit" sous forme d'un diagramme en barres.

```
import matplotlib.pyplot as plt
qualite=['tres mauvaise','mauvaise','normale','bonne','excellente']
quantite=[1,1,3,3,2]
plt.bar(qualite,quantite)
plt.show()
```



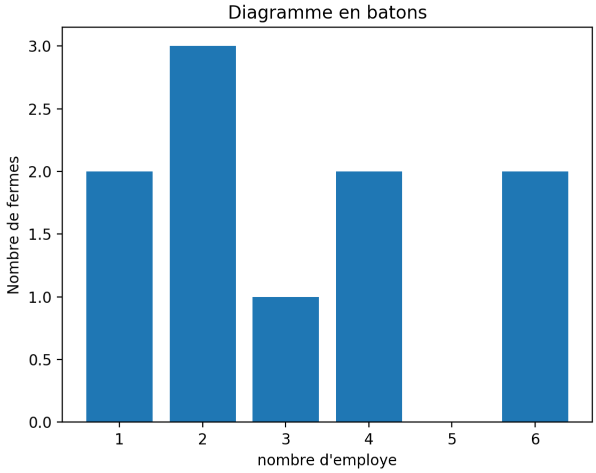

```
import matplotlib.pyplot as plt
qualite=['tres mauvaise','mauvaise','normale','bonne','excellente']
quantite=[1,1,3,3,2]
explode=(0,0,0.1,0,0)
plt.pie(quantite, explode=explode, labels=qualite, shadow=True)
plt.show()
```



Variables quantitatives La représentation des variables quantitatives discrètes se fait à l'aide de **diagrammes en bâtons**, qu'il ne faut pas confondre avec les diagrammes en barres utilisés pour les variables qualitatives. Les valeurs prises par les variables statistiques sont placées sur l'axe des abscisses, et leurs effectifs sont representés par un segment vertical de longueur proportionnelle à l'effectif.

Exemple 7. Tracer le diagramme en bâtons correspondant à la variable “nombre de personnes employées”.

```
import matplotlib.pyplot as plt
nbr_employe=[1,2,3,4,6]
quantite=[2,3,1,2,2]
plt.bar(nbr_employe,quantite)
plt.title('Diagramme en batons')
plt.xlabel('nombre d\'employe')
plt.ylabel('Nombre de fermes')
plt.show()
```



Pour les variables quantitatives continues, ou lorsque l'on regroupe les effectifs par classes, on utilise un **histogramme**. Chaque classe est représentée par un rectangle, dont la base est l'étendue de la classe et la surface proportionnelle à l'effectif.

Exemple 8. Tracer deux histogrammes correspondant à la variable “taille de l'exploitation”, en utilisant successivement les deux tableaux suivants :

Taille (ha)	$[0;20[$	$[20;40[$	$[40;60[$	$[60;80[$
Effectif	1	5	3	1

Taille (ha)	$[0;20[$	$[20;30[$	$[30;40[$	$[40;60[$	$[60;80[$
Effectif	1	2	3	3	1

III Paramètres caractéristiques d'une série statistique

III. 1 Paramètres de position

III. 1. a Mode

Définition 5. Le mode d'une série statistique est le caractère ayant le plus grand effectif.

Remarques. • On peut calculer le mode de n'importe quelle variable statistique (quantitative ou qualitative).

- Le mode n'est pas unique ! On peut en effet avoir plusieurs caractères de plus grand effectif. On distingue ainsi les séries unimodales (un seul mode) des séries plurimodales (plusieurs modes).
- Pour les variables continues, on parle de classe modale : c'est la classe ayant le plus grand rapport effectif/amplitude.

Exemple 9. Donner le mode de la variable statistique "nombre de personnes employées", ainsi que la classe modale de la variable "taille de l'exploitation" dans chacun des tableaux ci-dessus.

III. 1. b Moyenne

Définition 6. La moyenne d'une série statistique $(x_1, \dots, x_N)$ dont les effectifs sont donnés par $(n_1, \dots, n_N)$ est la moyenne des x_i, pondérée par les coefficients n_i. On la note $\overline{x}$, elle est définie par

$$\overline{x} = \frac{1}{\sum_{i=1}^{N} n_i} \sum_{i=1}^{N} x_i n_i = \frac{1}{T} \sum_{i=1}^{N} x_i n_i,$$

où T est la population totale.

Remarques.
- La moyenne n'est définie que pour des variables quantitatives.
- Si l'on connaît les fréquences associées, on a directement $\overline{x} = \sum_{i=1}^{N} x_i f_i$ (C'est la même chose que pour les variables aléatoires...)
- Pour des variables continues, on prend pour les x_i les centres des classes.

Exemple 10.
- Calculer la moyenne de la variable "nombre de personnes employées."
- Calculer la moyenne de la variable "taille de l'exploitation", en utilisant la série brute, puis les classes du premier tableau, puis les classes du deuxième tableau. Que remarque-t-on ?

Proposition 7. Calcul de moyennes.
- Linéarité de la moyenne : soit $(x_1, \dots, x_N)$ une série statistique d'effectifs $(n_1, \dots, n_N)$ ayant pour moyenne $\overline{x}$. Alors, la série statistique $y = (ax_1 + b, \dots, ax_N + b)$, de mêmes effectifs, a pour moyenne : $a\overline{x} + b$
- Soient S_1 et S_2 deux séries statistiques d'effectifs totaux respectifs N_1 et N_2 et de moyennes respectives $\overline{x}_1$ et $\overline{x}_2$. Alors la moyenne de la série S regroupant S_1 et S_2 est donnée par :
$$\frac{1}{N_1 + N_2}(N_1\overline{x}_1 + N_2\overline{x}_2)$$

Exercice 8.
- Les résultats du dernier DS sont catastrophiques, la moyenne est de 6,5. Dans un élan de générosité, M. G. hésite entre rajouter 1 point à tout le monde ou à multipler toutes les notes par 1,2. Quelle est la solution la plus avantageuse pour les élèves ?
- Au concours blanc, les élèves de 1BioA (38 élèves) ont obtenu 10,8 de moyenne, et ceux de la 1BioB (30 élèves) 10,2. Quelle est la moyenne générale ?

III. 1. c Médiane

Définition 9. La médiane d'une série statistique permet de couper la population étudiée en deux groupes de même taille de la façon suivante : 50% de la population a un caractère inférieur à la médiane, et 50% de la population a un caractère supérieur à la médiane.

Selon le type de données, la méthode sera différente :

⋆ Lorsque la série est sous forme brute, il faut commencer par ordonner les valeurs.

- Si l'effectif total est impair, par exemple pour la série $(0, 1, 3, 3, 4, 7, 10)$, il n'y a pas d'ambiguité dans la définition de la médiane, celle-ci vaut :
- En revanche, si l'effectif total est pair, par exemple pour la série $(0, 1, 3, 3, 4, 5, 7, 10)$, on parle d'intervalle médian :
  La médiane est alors définie comme la moyenne des bornes de l'intervalle médian :
  On peut aussi choisir de prendre la borne de gauche, pour être en adéquation avec la définition de quartile.

Remarque.
- On peut calculer la médiane d'une variable quantitative, ou d'une variable qualitative ordinale.
- La médiane permet de limiter l'impact de données extrêmes, contrairement à la moyenne.

Exemple 11. Voici la série des notes obtenues par un très bon élève (qui a été malade au dernier devoir) : $17, 16, 17, 15, 17, 5$. Calculer la moyenne et la médiane de cet élève, et commenter l'impact de la dernière note.

⋆ Lorsque l'on connaît la série à l'aide de ses effectifs, la médiane est la première valeur pour laquelle l'effectif cumulé est superieur ou égale à $\dfrac{n_{\text{tot}}}{2}$.
On peut aussi choisir de prendre les deux valeurs dont les effectifs cumulés encadrent $\dfrac{n_{\text{tot}}}{2}$, de calculer l'équation de la droite $y = ax + b$ reliant ces valeurs sur le polygône des effectifs cumulés, puis de résoudre $ax + b = \dfrac{n_{tot}}{2}$. En pratique, n'importe quel nombre entre les deux valeurs est une médiane de la série.

Exemple 12. Reprendre l'exemple de la variable "nombre de personnes employées", et calculer sa médiane.

⋆ Lorsque les variables sont regroupées par classe, on commence par calculer la classe médiane : c'est la première dont l'effectif cumulé est supérieur ou égal à $\dfrac{n_{\text{tot}}}{2}$.
Pour calculer la médiane, on calcule l'équation de la droite $y = ax + b$ reliant les bornes de la classe médiane sur le polygône des effectifs cumulés, puis on résout $ax + b = \dfrac{n_{tot}}{2}$.

Exemple 13. Reprendre l'exemple de la variable "taille de l'exploitation", et calculer sa médiane à l'aide des deux tableaux.

III. 1. d Quartiles - Déciles

Définition 10. Quartiles

- Le **premier quartile** d'une série, noté Q_1, est la plus petite valeur pour laquelle au moins 25 % des données sont inférieures ou égales à Q_1.
- Le **troisième quartile** d'une série, noté Q_3, est la plus petite valeur pour laquelle au moins 75 % des données sont inférieures ou égales à Q_3.

Remarque. Le deuxième quartile est la médiane

Exercice 11. Le tableau suivant donne les résultats des relevés pluviométriques à Brest en 2008 :

Pluviométrie en mm	0	1	2	3	4	5	6	7	8	9	10	15	20	25	30
Nombre de jours	207	14	8	13	14	21	19	17	9	10	14	8	4	5	3

Calculer les effectifs cumulés associés à cette série statistique. Les dictons suivants : « il pleut plus de la moitié de l'année en Bretagne » et « il pleut au moins 1 cm d'eau par jour le quart de l'année » se vérifient-ils sur l'année 2008 ?

Définition 12. Déciles

- Le **premier décile** d'une série, noté D_1, est la plus petite valeur pour laquelle au moins 10 % des données sont inférieures ou égales à D_1.
- Le **neuvième décile** d'une série, noté D_9, est la plus petite valeur pour laquelle au moins 90 % des données sont inférieures ou égales à D_9.

Remarques.
- Les déciles sont largement utilisés en géologie minière, en hydrologie (niveau de crue décennale), et dans le milieu médical.
- On définit de la même manière les centiles (ou percentiles), quintiles, terciles, ...

III. 2 Paramètres de dispersion

Définition 13.

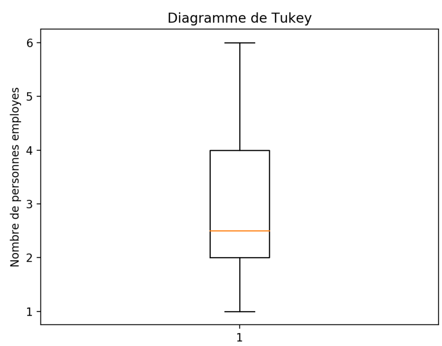
- L'**étendue** d'une série statistique est la différence entre la plus grande et la plus petite valeur observée.
- L'**intervalle interquartile** est l'intervalle $[Q_1, Q_3]$.
  La **distance interquartile** est le nombre $Q_3 - Q_1$.
- L'**intervalle interdécile** est l'intervalle $[D_1, D_9]$.
  La **distance interdécile** est le nombre $D_9 - D_1$.

On peut à l'aide de ces données construire le diagramme de Tukey (ou diagramme en boite, ou encore « boite à moustache ») associé à la série statistique. Il s'agit de tracer un rectangle allant du premier quartile au troisième quartile et coupé par la médiane.

Il sépare la série statistique en 4 blocs de taille environ égale.

Exemple 14. Tracer de diagramme de Tukey du nombre de personne employées pour la série statistique du début du cours.

```
import matplotlib.pyplot as plt
plt.title('Diagramme de Tukey')
plt.ylabel('Nombre de personnes employes')
plt.boxplot([2,4,3,6,1,1,2,6,2,4] )
plt.show()
```



III. 2. a Variance

Définition 14. Soit $x = (x_1, \ldots, x_N)$ une série statistique d'effectifs $(n_1, \ldots, n_N)$. La **variance** de x est donnée par
$$var(x) = \overline{(x - \overline{x})^2}$$
c'est-à-dire
$$var(x) = \frac{1}{T}\sum_{i=1}^{N}(x_i n_i - \overline{x})^2$$

Remarque. Si l'on connaît les fréquences f_i, on a : $V(x) =$ ....
Si l'on a une série brute, dont tous les effectifs valent 1, on a : $V(x) = 0$

Proposition 15. On a $V(x) = \overline{x^2} - \overline{x}^2$

Remarque. C'est la même formule que celle des variables aléatoires.

Proposition 16. Soit y la série statistique définie par $y = (ax_1 + b, \ldots, ax_N + b)$, et ayant les mêmes effectifs que x. Alors on a
$$V(y) = a^2 V(x)$$

Remarque. Contrairement à la moyenne, la variance n'est pas linéaire !

Définition 17. Soit $x = (x_1, \ldots, x_N)$ une série statistique d'effectifs $(n_1, \ldots, n_N)$. L'**écart type** de x est donné par
$$\sigma_x = \sqrt{V(x)}$$

Remarque. L'écart type donne une indication de la dispersion des valeurs d'une série statistique autour de sa moyenne. Plus l'écart type est petit, plus les valeurs de la série sont resserrées autour de la moyenne. Au contraire, plus l'écart type est grand, plus les valeurs sont dispersées.
On en particulier la propriété suivante : $V(x) = 0 \Leftrightarrow \sigma_x = 0 \Leftrightarrow x$ est une série constante

Exercice 18. Un entrepreneur souhaite fabriquer des chaussures pour hommes. Il fait une étude statistique pour connaître la répartition des pointures :

Taille	37	38	39	40	41	42	43	44	45	46	47
Effectifs	2	4	7	17	31	56	85	50	33	12	5

1. Calculer la moyenne et l'écart type de cette série statistique à 10^{-1} près.
2. L'entrepreneur décide de ne produire que les tailles de l'intervalle $[\overline{x} - 2\sigma_x, \overline{x} + 2\sigma_x]$. Quelles sont les tailles qu'il va fabriquer ? Quel pourcentage de l'étude cela représente-t-il ?

IV Statistique bivariée

On observe que dans certains cas, il semble exister un lien entre deux caractères d'une population, par exemple entre le poids et la taille d'une vache, entre l'épaisseur d'un mur et sa résistance thermique, entre la consommation et la vitesse d'une voiture... L'objectif principal de ce paragraphe est de mettre en évidence un lien entre deux caractères d'une même population.

IV. 1 Définitions

Définition 19. Une **série statistique double** est une série statistique portant sur deux variables d'une même population.
On étudie donc des N-uplets d'éléments de $\mathbb{R}^2$: $((x_1, y_1), (x_2, y_2), \dots, (x_N, y_n))$.

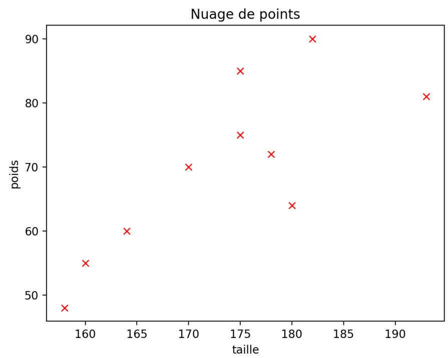
Exemple 15. On étudie la taille et le poids de 10 individus :

Taille (cm)	158	180	170	193	175	160	164	178	182	175
Poids (kg)	48	64	70	81	75	55	60	72	90	85

La série statistique double est la série des couples (x_i, y_i) ou x_i est la taille de l'individu i et y_i est son poids.

Chaque couple de nombre peut ainsi être représenté comme un point dans le plan $\mathbb{R}^2$. On représente graphiquement une série double en traçant l'ensemble des points qu'elle contient : c'est ce qu'on appelle le **nuage de points** associé. On obtient pour l'exemple précédent

```
import matplotlib.pyplot as plt
taille=[158,180,170,193,175,160,164,
             178,182,175]
poids=[48,64,70,81,75,55,60,72,90,85]
plt.plot(taille,poids, 'rx')
plt.title('Nuage de points')
plt.xlabel('taille')
plt.ylabel('poids')
plt.plot()
```



Définition 20. On appelle point moyen d'une série statistique le point $(\overline{x}, \overline{y})$, où $\overline{x}$ est la moyenne de la série $(x_1, \dots, x_N)$ et $\overline{y}$ la moyenne de la série $(y_1, \dots, y_N)$.

Exemple 16. Calculer le point moyen de la série de l'exemple précédent.

IV. 2 Caractéristiques d'une série statistique double

Pour une série double, l'un des objectifs est de chercher une dépendance entre les deux variables étudiées.

Définition 21. La covariance d'une série double est donnée par

$$\mathrm{Cov}(x, y) = \frac{1}{N} \sum_{i=1}^{N} (x_i - \overline{x})(y_i - \overline{y})$$

Proposition 22. Formule de Koenig-Huygens.

$$\mathrm{Cov}(x, y) = \overline{xy} - \overline{x} \cdot \overline{y}$$

Exemple 17. Calculer la covariance pour l'exemple précédent.

```python
 taille=np.array([158,180,170,193,175,160,164,178,182,175])
poids=np.array([48,64,70,81,75,55,60,72,90,85])
xbar=np.sum(taille)/10
ybar=np.sum(poids)/10
xybar=np.sum(taille*poids)/10
sigmax=sqrt(xbar)
sigmay=sqrt(ybar)
print((xybar-xbar*ybar))
```

Remarque. Si les deux variables ont tendance à varier dans le même sens, la covariance est positive. Au contraire, si les deux variables varient en sens contraire, la covariance est négative.

Définition 23. On définit le coefficient de corrélation linéaire à l'aide de la covariance par

$$\rho(x, y) = \frac{\mathrm{Cov}(x, y)}{\sigma_x \sigma_y}$$

Exemple 18. Calculer le coefficient de corrélation pour l'exemple précédent.

Remarques.
- On a toujours $-1 \leq \rho(x, y) \leq 1$.
- Le coefficient de corrélation indique s'il y a un lien linéaire entre x et y. Si $\rho(x, y) = 1$ (resp -1), y dépend linéairement de x, et le nuage de points est aligné dur une droite croissante (resp. décroissante).
- S'il n'y a aucun lien linéaire entre x et y, le coefficient de corrélation est nul. Mais il peut y avoir un lien non linéaire entre x et y.

IV. 3 Droite d'ajustement affine

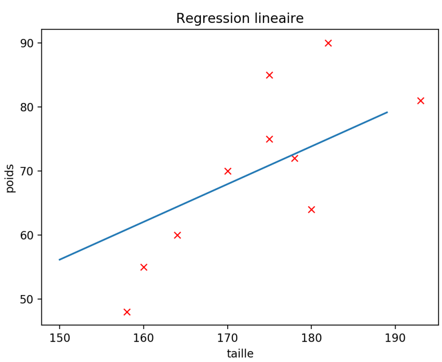
Lorsque deux variables semblent corrélées linéairement, il est naturel de chercher quelle relation linéaire les lie l'une à l'autre. On cherche donc des réels a et b tels que les valeurs de y se rapprochent "le plus" (en un sens que nous définiront) des valeurs de $ax+b$. C'est ce qu'on appelle l'**ajustement affine**.

Définition 24. Soient $((x_1, y_1), \ldots, (x_N, y_N))$ une série statistique double. La **droite d'ajustement affine** (ou **droite de régression linéaire**) de cette série est la droite ajustant le mieux un nuage de point au sens des moindres carrés. Elle a pour équation

$$y = ax + b, \quad \text{avec } a = \frac{\mathrm{Cov}(x, y)}{V(x)} = \frac{\sigma_y}{\sigma_x} \times \rho(x, y), \text{ et } b = \overline{y} - a\overline{x}.$$

Exemple 19. Calculer l'équation de la droite d'ajustement linéaire sur l'exemple précédent.

```python
a=(xybar-xbar*ybar)/(sigmax**2)
b=ybar-a*xbar
X=np.array([150+i for i in range(40)])
Y=a*X+b
plt.title('Regression lineaire')
plt.xlabel('taille')
plt.ylabel('poids')
plt.plot(X,Y)
plt.plot()
plt.show()
```



Remarque. L'ajustement se fait au sens des moindres carrés : ceci veut dire que l'on minimise la somme des carrés des erreurs que l'on commet avec cette approximation linéaire.