Statistiques descriptives
Cours

1 Définitions et rappels

L'objectif des statistiques est d'étudier une **population** à partir d'observations faites sur ses **individus**. On choisit des **caractères**, aussi appelés **variables statistiques** que l'on souhaite traiter. Une **statistique** est l'information extraite à partir des observations réalisée et de leur étude.

On peut rarement interroger l'ensemble de la population, on va donc nommer **échantillon** l'ensemble des individus sur lesquels les observations ont été faites.

On distingue deux grandes familles de statistiques par leur finalité : les statistiques **descriptives** dont l'objectif est d'informer sur l'état de la population en observant le maximum d'individus, et les statistiques **inférentielles** qui permettent d'obtenir des informations sur toute la population en questionnant le minimum d'individus.

De plus, on distingue aussi les statistiques de par leur nature : les statistiques **quantitatives** dont les données sont sous forme de valeurs numériques, et les statistiques **qualitatives** dont les données ne sont pas des valeurs numériques.

Ce cours traitera uniquement de l'étude de statistiques descriptives quantitatives.

Ces statistiques peuvent être traités sous forme brutes, on parlera de statistique **discrète**, ou en rassemblant leurs données en classe (intervalle de valeur par exemple), et on parlera de statistique **continue**.

Remarque :

Nous utiliserons le symbole $\sum$ pour décrire une somme d'éléments. Par exemple $\sum_{i=1}^{4} u_i$ représente la somme des termes d'indices 1, 2, 3 et 4 de la suite $(u_n)_n$, c'est-à-dire $u_1 + u_2 + u_3 + u_4$. De même $\sum_{i=0}^{100} 2^i = 2^0 + 2^1 + ... + 2^{99} + 2^{100}$, ce qui permet de comprendre l'intérêt de l'utilisation de cette notation.

1.1 Statistiques discrètes

On présente une statistique comme un ensemble de **classes**. Chaque classe C_i correspond à une **valeur**, notées x_i qui est le résultat donné par les individus au sondage, et un **effectif**, noté n_i, qui correspond au nombre d'individus dans la classe, c'est-à-dire le nombre d'individus ayant répondu la valeur x_i au sondage.

Une statistique est dite **discrète** quand les valeurs des différentes classes sont **séparées**. Cela signifie qu'il n'est pas toujours possible, pour deux valeurs possibles, d'en trouver une troisième entre.

Exemple 1:
Les notes d'étudiants à un devoir est une statistique discrète.

Exercice 1:
Dans l'exemple précédent, combien y a-t-il de classes et quelles sont les valeurs associées ?

Définition :
Soit S une statistique. On suppose qu'il y a k résultats possibles quand on observe S sur une population.

Pour une **classe** C_i de S, on note x_i sa **valeur** et n_i son **effectif** avec $i \in \{1, 2, .., k\}$. On suppose de plus que les valeurs des classes sont rangées par ordre croissant, c'est-à-dire $x_1 < x_2 < ... < x_k$.

- On définit l'**effectif total**, noté n, qui représente le nombre total d'individus dans la population, par :

$$n = \sum_{i=1}^{k} n_i = n_1 + n_2 + ... + n_k$$

- On définit les **fréquences**, notées f_i de chacune des classes, qui représente la proportion d'individus dans cette classe, par :

$$f_i = \frac{n_i}{n}$$

  On remarque de plus que $\sum_{i=1}^{k} f_i = 1$.

- On définit les **effectifs cumulés croissants**, notés N_i, qui représentent le nombre d'individus ayant répondu un résultat inférieur à celui de la classe C_i, par :

$$N_i = \sum_{j=1}^{i} n_j$$

- On définit les **fréquences cumulées croissantes**, notées F_i, qui représentent la proportion d'individus ayant répondu un résultat inférieur à celui de la classe C_i, par :

$$F_i = \sum_{j=1}^{i} f_j = \frac{N_i}{n}$$

Remarque :
On aurait pu ranger les valeurs en ordre décroissants et définir les effectifs et fréquences cumulé.e.s décroissant.e.s.

Exercice 2:
La population étudiée est un ensemble de 30 familles. La variable statistique discrète étudié X est le nombre d'enfants. Les classes et leurs effectifs sont donnés par le tableau suivant :

classes	C_0	C_1	C_2	C_3	C_4	C_5	C_6	C_7
valeurs	0	1	2	3	4	5	6	7
effectifs	5	7	8	4	2	2	1	1
effectifs cumulés								
fréquences								
fréquences cumulées								

1- Complétez ce tableau.

2- Quel est le pourcentage de familles admettant deux enfants ?

3- Quel est le pourcentage de familles admettant au plus un enfant ?

4- Quel est le pourcentage de familles admettant 2 à 5 enfants ?

1.2 Statistiques continues

Par opposition à un statistique discrète, une statistique **continue** est une statistique où entre 2 valeurs possibles, il est toujours possible d'en trouver une troisième. Dans ce cas on regroupera les résultats possibles dans des intervalles et chacun de ces intervalles représentera une classe. On supposera alors que les résultats sont répartis uniformément dans ces intervalles.

Exemple 2:
La taille des étudiants d'une classe est une statistique continue.

Remarque :
On peut décider, quand les effectifs dans chaque classe est trop faible, de traiter une statistique discrète de manière continue. Pour cela on regroupe les valeurs dans des intervalles qui formeront de nouvelles classes.

Dans le cas continue, la **valeur** d'une classe ne sera pas un nombre mais un intervalle. Dans le cas où l'on aurait à faire des calculs (cf partie suivante), nous utiliserons le centre de l'intervalle.

Les définitions précédentes d'**effectifs**, d'**effectif total**, de **fréquences**, d'**effectifs cumulés croissants** et de **fréquences cumulées croissantes** restent inchangés.

Exercice 3:
La distribution Y des tailles d'une population de 100 collégiens est donnée par le tableau :

classes	C_1	C_2	C_3	C_4
valeurs en cm.	[145,155[	[155,160[	[160,165[	[165,175[
effectifs	30	25	23	22
effectifs cumulés				
fréquences				
fréquences cumulées				

1- Complétez le tableau de la distribution de Y.

2- Quelle est la proportion d'élèves ayant une taille inférieur à 165cm ?

3- Quelle est la proportion d'élèves ayant une taille inférieur à 163cm ?

2 Statistiques à une dimension

Nous allons dans un premier temps voir comment décrire et représenter des statistiques en une dimension, c'est-à-dire qui ne considère qu'un seul caractère de la population observée (on peut se dire que le sondage ne consistait qu'en une seule question).

2.1 Représentation graphique

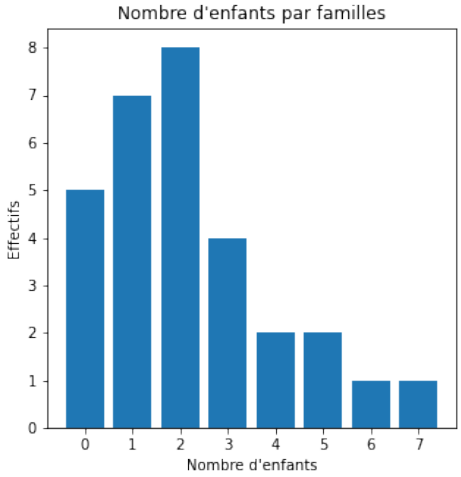
La manière la plus simple pour comprendre une statistique est de faire une représentation graphique. Il faut cependant faire attention à la pertinence du type de représentation choisie suivant les données étudiées, et le sens que l'on veut leur donner.

2.1.1 Diagramme en barres

Les valeurs des différentes classes vont être placés en abscisses de ce graphique. Pour chacune de ces abscisses, on trace une barre dont la hauteur représente l'effectif (ou la fréquence) de la classe.

Remarque :
Les diagrammes en barres sont plutôt utilisés pour la représentation de statistiques quantitatives discrètes (mais aussi de statistiques qualitatives).

Exemple 3:
Voici le diagramme en barre de la statistique X de l'exercice 2.

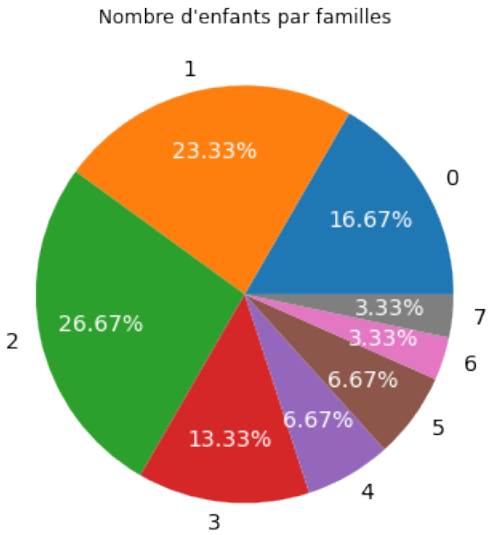


2.1.2 Diagramme circulaire

On utilise les fréquences pour calculer la proportion d'arc de cercle qui correspond à chacune des classes ($2\pi \times f_i$). On découpe alors un disque en colorant de couleurs différentes chaque portion qui correspond aux différentes classes.

Remarque :
Les diagrammes circulaires sont plutôt utilisés pour la représentation de statistiques quantitatives discrètes (mais aussi de statistiques qualitatives).

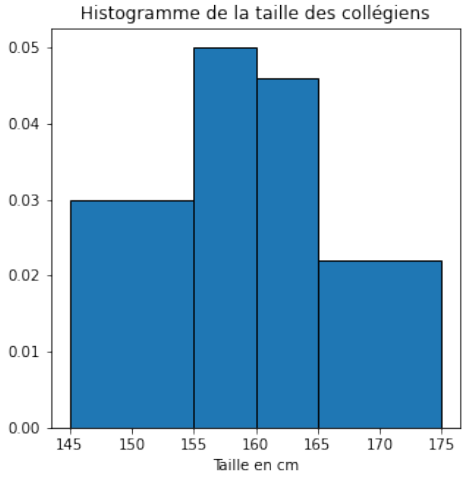
Exemple 4:
Voici le diagramme circulaire de la statistique X de l'exercice 2.



2.1.3 Histogramme

L'histogramme des fréquences représente la proportion d'individus dans des intervalles. Cette représentation est donc réservée au traitement de données continues. Contrairement au diagramme en barres, ce n'est pas la hauteur de la barre qui est proportionnelle à la fréquence mais son aire. Ainsi deux classes qui ont les mêmes effectifs, pourront avoir des barres de hauteurs différentes (si la taille de l'intervalle de l'une est plus grande que celui de l'autre).

Exemple 5:
Voici l'histogramme de la statistique Y de l'exercice 3.

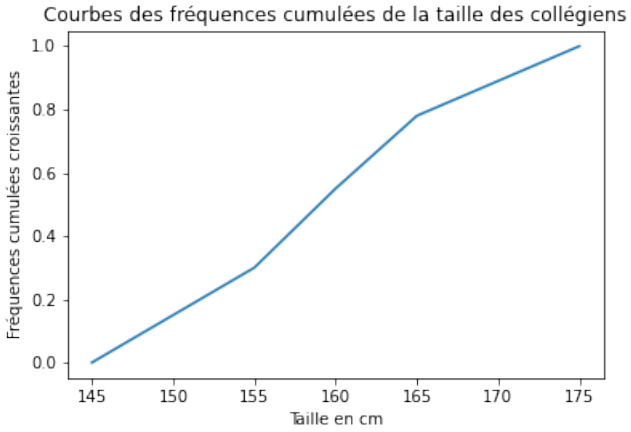


2.1.4 Courbe

La courbe est une représentation souvent bien tentante. Malheureusement, elle n'est pertinente que dans quelques cas bien précis.

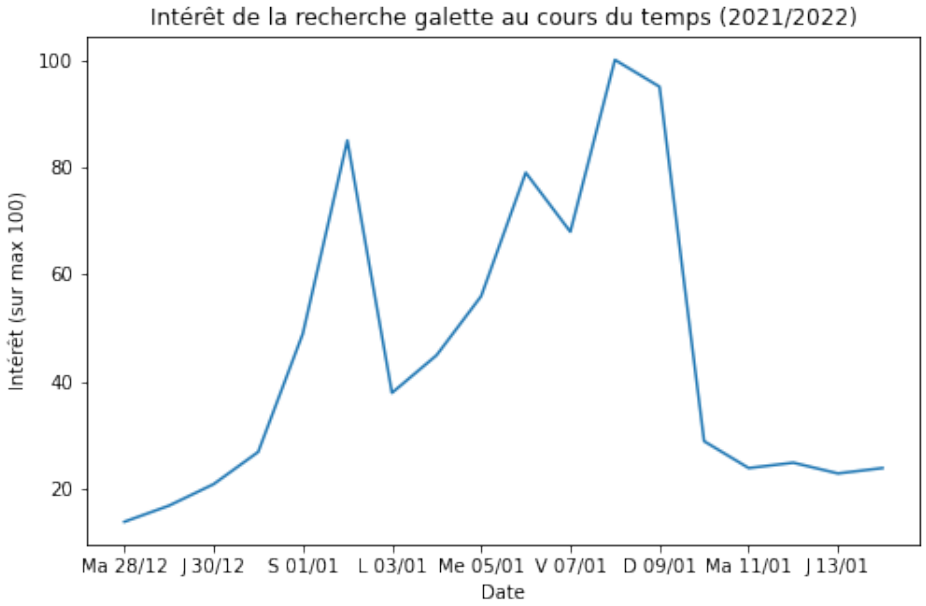
Le premier cas est la représentation cumulée d'une statistique continue. On place la fin de chaque intervalle en abscisse et la fréquence cumulée croissante (ou l'effectif cumulé croissant) en ordonné. On place aussi un point ayant pour abscisse le début du premier intervalle et 0 pour ordonnée. Un relie ces points 1 à 1 pour obtenir la courbe.

Exemple 6:
Voici la courbe des fréquences cumulées de la statistique Y de l'exercice 3.



Le second cas d'utilisation est le traitement de séries temporelles (qui sort de ce que l'on abordera dans ce cours). Ici, nous pourrons nous en servir quand les valeurs d'une statistiques sont des dates ou des durées. Cela reviens à faire un diagramme en barre, en plaçant un point en haut de chaque barre et en reliant ces points entre eux (et sans tracer les barres).

Exemple 7:
On regarde l'intérêt de la recherche du mot galette sur google autour de la fête des rois.



2.2 Paramètres d'une distribution statistique

La représentation graphique d'une statistique permet de l'appréhender et d'appuyer un argument mais ne doit pas être confondu avec réel étude mathématique des données.

Pour donner du sens à des observations, nous allons utiliser des outils mathématiques, aussi appelés **paramètres statistiques**. Nous en étudierons 2 familles : les paramètres de **position** et les paramètres de **dispersion**.

2.2.1 Paramètres de position

Comme leur nom l'indique, ils permettent de positionner une série (par exemple sur un axe, ou une par rapport à une autre).

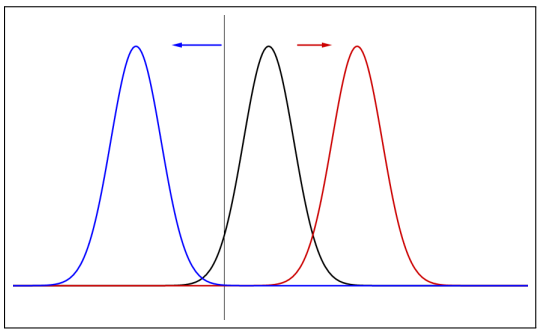


Figure 1: Illustration de la modification d'un paramètre de position

Définition :

La valeur **moyenne** d'une série statistique, aussi appelée tout simplement **moyenne**, correspond au centre (ou barycentre) des résultats observée. Elle se note souvent m_X ou $\bar{x}$ (si la statistique est notée X, sinon on change les lettres respectivement) et se calcul par la formule :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i = \sum_{i=1}^{k} f_i x_i$$

Remarque :

Il s'agit de ce que l'on appelle parfois une moyenne **pondérée**. Dans le cas où les valeurs de la série statistique ne seraient pas regroupées en classe, une simple moyenne suffit.

Remarque :

Dans le cas du statistique continue, on prend le centre des intervalles pour remplacer x_i dans la formule.

Exercice 4:
Calculer les moyennes $\overline{x}$ et $\overline{y}$ des statistiques X et Y des exercices 2 et 3.

Propriété :

La moyenne est linéaire, ce qui signifie que pour X et Y deux séries statistiques et λ un nombre réel, on a :

$$\overline{x+y} = \overline{x} + \overline{y} \qquad \text{et} \qquad \overline{\lambda x} = \lambda \overline{x}$$

Exemple 8:

- *Les élèves d'une classes mangent en moyennes 2 pommes et 3 clémentines par jour. Ils mangent donc en moyennes* $2+3=5$ *fruits (en supposant qu'il n'y en a pas d'autres) par jour.*
- *Si le salaire moyen dans une entreprise est de 2000 euros mensuel, et que le patron décide d'augmenter tout le monde de 10%, le salaire moyen de l'entreprise sera de* $1.1 * 2000 = 2200$ *euros mensuel.*

Définition :

La **médiane**, notée $x_{1/2}$, d'une série statistique est le résultat (on parle ici des réponses au sondage et pas des valeurs possibles) qui se situe au milieu de la série. Elle se calcul différemment dans le cas discret et dans le cas continue.

- Dans le cas discret, il suffit de ranger les résultats par ordre croissant et de trouver celui en position $\frac{n+1}{2}$ (parce que l'on commence à compter à 1). Si n est impaire on tombe sur une valeur de la série, si n est paire on tombe entre 2 valeur dont on fait la moyenne pour déterminer la médiane.

Exercice 5:
Calculer la médiane $x_{1/2}$ de la statistique X de l'exercice 2.

- Dans le cas continue, il faut tout d'abord trouver la classe médiane. Pour cela on utilise les fréquences cumulées croissantes, la classe qui contient la première valeur au-dessus de 0.5 est la classe médiane. Il faut alors calculer la proportion de l'intervalle qui est en-dessous de 0.5 et enfin déterminer la médiane. En pratique cela donne le calcul suivant :

$$x_{1/2} = a_i + (a_{i+1} - a_i)\frac{0.5 - F_{i-1}}{F_i - F_{i-1}}$$

où C_i est la classe médiane, $[a_i; a_{i+1}]$ son intervalle de valeurs et F_i sa fréquence cumulée croissante.

Exercice 6:
Calculer la médiane $y_{1/2}$ de la statistique Y de l'exercice 3.

Définition :
De la même manière on peut définir différents **quantiles** d'une série statistique. Par exemple les quantiles qui se situent au premier, deuxième et troisième quarts d'une série statistique (le deuxième quartile étant la médiane). On peut encore définir les déciles, qui permettent de découper une population en dix parties égales.

2.2.2 Paramètres de dispersion

Les paramètres de dispersion permettent de savoir si les résultats sont proches ou éloignés de la position de la série.

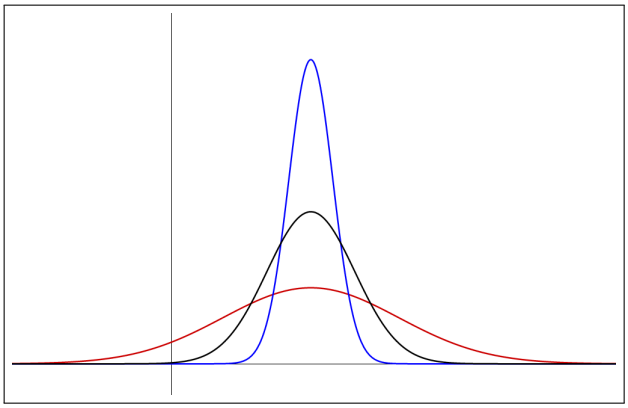


Figure 2: Illustration de la modification d'un paramètre de dispersion

Définition :
Une façon, assez intuitive de se rendre compte de la dispersion d'une série est le calcul de son étendu, c'était dire l'écart entre sa plus grande et sa plus petite valeur.

Exercice 7:
Calculer l'étendue des statistiques X et Y des exercices 2 et 3.

Définition :

L'**écart-type**, noté σ_x, représente la distance moyenne (quadratique) entre les résultats et la moyenne. On le calcul en utilisant la variance, notée V_x, via la formule suivante :

$$\sigma_x^2 = V_x = \frac{1}{n}\sum_{i=1}^{k}\left[n_i(x_i-\overline{x})^2\right] = \sum_{i=1}^{k}\left[f_i(x_i-\overline{x})^2\right]$$

ou, autrement dit

$$\sigma_x = \sqrt{V_x} = \sqrt{\frac{1}{n}\sum_{i=1}^{k} n_i(x_i-\overline{x})^2} = \sqrt{\sum_{i=1}^{k} f_i(x_i-\overline{x})^2}$$

Exercice 8:
Calculer l'écart-type σ_x de la statistique X de l'exercice 2.

Propriété : (Formule de Koenig)

On peut développer la formule de la variance pour obtenir une nouvelle manière de calculer l'écart-type

$$\sigma_x^2 = V_x = \frac{1}{n}\sum_{i=1}^{k}\left[n_i x_i^2\right] - \overline{x}^2 = \sum_{i=1}^{k}\left[f_i x_i^2\right] - \overline{x}^2$$

Remarque :

En pratique, on utilisera la formule de Koenig qui diminue le nombre de calculs à effectuer.

Exercice 9:
Calculer l'écart-type σ_y de la statistique Y de l'exercice 3.

Vocabulaire :

En anglais, moyenne se dit *mean*, médiane s'écrit *median* et écart-type se dit *standard deviation* souvent abrégé en *std*.

Définition :

Le demi-ecart interquartile est le paramètre de dispersion associé à la médiane. Il se calcul via la formule $\dfrac{x_{3/4}-x_{1/4}}{2}$.

Remarque :

En pratique, on représentera la dispersion autour de la médiane en donnant les premier et troisième quartile (et éventuellement par la réalisation d'un boite à moustache).

Exercice 10:
Calculer les demi-ecarts interquartiles des statistiques X et Y des exercices 2 et 3.

3 Statistiques à deux dimensions

Dans le cas de statistique en plusieurs dimensions (nous nous limiterons ici à deux), il intéressant de se demander comment chacune des dimension de la statistique varie par rapport aux autres. On donnera le nom de **variable** à chacune de ces dimensions et nous allons donc introduire la notion de corrélation.

3.1 Échantillon de petite taille

On va tout d'abord s'intéresser à des observations faites sur des échantillons de petite taille (faible nombre de réponses). Dans ce cas là on ne prend pas la peine de regrouper les individus par classe. Chaque individu est seul dans sa classe et 2 valeurs lui sont associé.

Exemple 9:
On possède 6 spécimens fossiles d'un animal disparu et ces spécimens sont de tailles différentes. On estime que si les animaux appartiennent à la même espèce il doit exister une relation linéaire entre la longueur de deux de leurs os, l'humérus et le fémur. Voici les mesures de ces longueurs en cm pour les 5 spécimens possédant ces deux os intacts :

x humérus	*44*	*62*	*71*	*73*	*87*
y fémur	*40*	*57*	*59*	*65*	*77*

3.1.1 Corrélation de deux données

Définition :
La **covariance** représente la manière dont varie deux variables l'une par rapport à l'autre. Elle se calcul par la formule :

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{k}[(x_i - \overline{x})(y_i - \overline{y})]$$

Remarque :
On peut constater une similitude avec la formule de la variance. En fait on peut écrire $V_X = Cov(X,X)$. D'ailleurs la formule de Koenig s'applique elle aussi et sera la formule qui sera appliquée dans la plus part des cas pour calculer la covariance.

Propriété : Formule de Koenig
On peut réécrire la formule de la covariance de la manière suivante

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{k}[x_i y_i] - \overline{xy}$$

Exercice 11:
Calculer la covariance sur l'exemple 9.

Remarque :
Une covariance peut être positive ou négative (voir nulle). Dans le cas où elle est positive, cela signifie que les variables ont une influence positive l'une sur l'autre (par exemple : plus une voiture roule plus elle fait de kilomètres). Dans le cas où elle est négative, ces deux variables ont une

influence négative l'une sur l'autre (par exemple : plus une voiture roule, moins il y aura d'essence dans son réservoir).

Définition :

Le **coefficient de corrélation linéaire** entre deux variables, noté $r(X, Y)$, permet de savoir à quel point deux variables sont *linéairement* liées. Il se calcule via la formule

$$r(X, Y) = \frac{Cov(X, Y)}{\sigma_x \sigma_y}$$

Propriété :

Le coefficient de corrélation linéaire est toujours compris entre -1 et 1 : $-1 \leq r(X, Y) \leq 1$. De plus, si $|r(X, Y)| = 1$ alors X et Y sont linéairement liées.

Convention :

Suivant la précision nécessaire dans le cas traité, on peut accepter que $|r(X, Y)|$ ne soit pas égal à 1 et tout de même conclre à la relation linéaire des deux variables. En général, on demande que $|r(X, Y)| > 0.95$, mais cette valeur, arbitraire, peut variée suivant le contexte (on peut être moins rigoureux pour régler un argument entre amis que quand on prépare le lancement d'une navette).

Remarque :

Les écart-types étant positifs, le coefficient de corrélation linéaire a le même signe que la covariance, on peut donc en faire la même lecture.

Remarque :

Le cas du coefficient de corrélation linéaire nul est un cas particulier. Cela signifie qu'il n'y a pas de corrélation entre les deux variables (à ne pas confondre avec l'indépendance).

Exercice 12:
Calculer le coefficient de corrélation linéaire sur l'exemple 9.

Définition :

Quand la relation linéaire est avérée, on peut déterminer un **ajustement linéaire** des deux variables, c'est-à-dire déterminer l'équation linéaire qui permet de passer de l'une à l'autre. On cherche alors une équation de la forme $y = ax+b$, où x et y sont les variables et a et b les coefficients de l'ajustement (a est souvent appelé la pente et b l'ordonnée à l'origine). Ces deux derniers ce calcule de la manière suivante

$$a = \frac{Cov(X, Y)}{\sigma_X^2} \qquad \text{et} \qquad b = \overline{y} - a\overline{x}$$

Remarque :

Dans ce cas on dit que l'on fait l'ajustement linéaire de Y en fonction de X. On peut faire l'inverse et déterminer les coefficient de l'équation $x = ay + b$ avec les formules

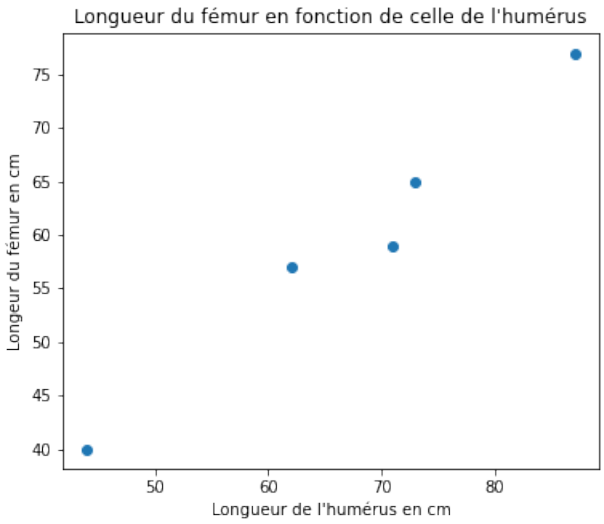
$$a = \frac{Cov(X, Y)}{\sigma_Y^2} \qquad \text{et} \qquad b = \overline{x} - a\overline{y}$$

Exercice 13:
Calculer l'équation de l'ajustement linéaire de Y en fonction de X sur l'exemple 9.

3.1.2 Représentation et ajustement linéaire

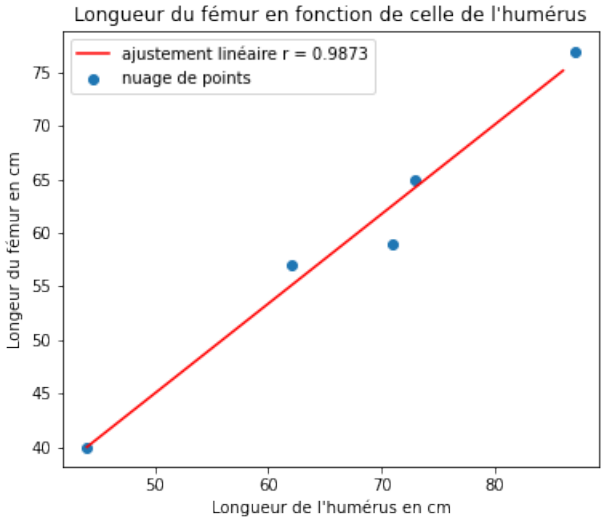
Une bonne manière de représenter une statistique à deux dimensions est via un nuage de points. On place une des variables en abscisses et l'autre en ordonnées. On place alors un point par individu, au point où les coordonnées correspondent à la valeur prise par les variables pour cet individu.

Exemple 10:
Voici le nuage de point des variables X et Y de l'exemple précédent.

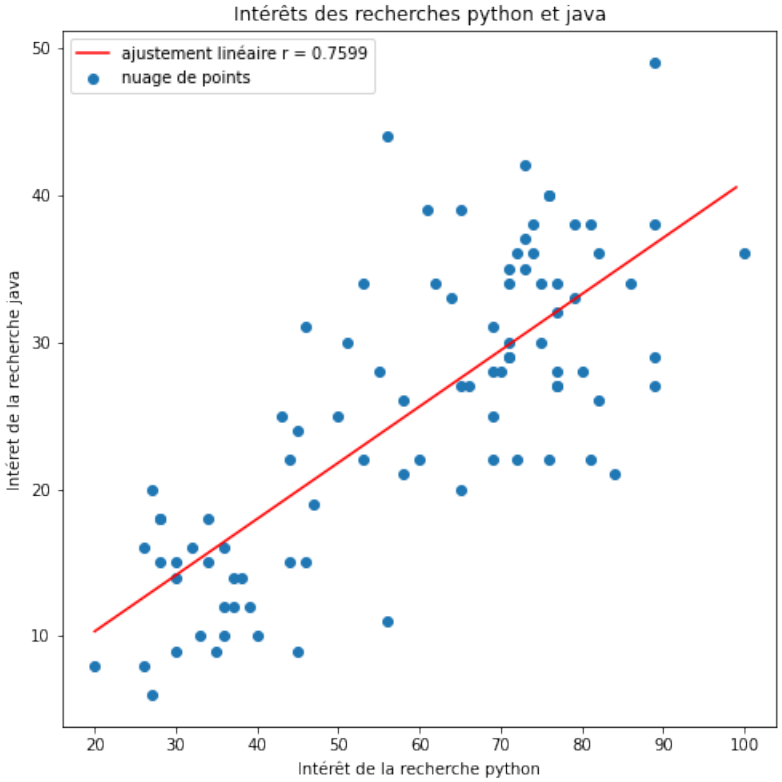


Afin d'apporter de l'information à ce graphe, on peut tracer la droite associée à l'équation d'ajustement linéaire entre les deux variables. On pensera néanmoins à préciser le coefficient de corrélation linéaire, gage de la qualité de cet ajustement.

Exemple 11:
Voici le nuage de point des variables X et Y de l'exemple précédent.



Exemple 12:
On peut comparer l'intérêt que les développeurs portent aux langages de programmation en comparant le nombre de recherche faites sur google(à l'aide de `trends.google.fr` *).*



Remarque :
L'ajustement linéaire permet aussi de faire ce que l'on appelle de le prédiction, c'est-à-dire que l'on est capable d'estimer une variable à partir de l'autre. Nous ne développerons pas plus ce sujet dans ce cours car nous rentrerions dans le cadre des statistiques inférentielles.

3.2 Cas général

Dans le cas où les données sont plus conséquentes, on préfèrera les représenter sous forme d'un tableau à double entrée : les lignes représenterons les valeurs de l'une des variables, et les colonnes celles de l'autre. Chaque case de ce tableau est une classe d'individu ayant les mêmes résultats pour les deux dimensions de la statistique. On indice généralement les lignes par i et les colonnes par j, et on note $n_{i,j}$ l'effectif de la case en position (i, j).

Exemple 13:
On reprend la population des collégiens de l'exercice 3, mais on mesure maintenant les deux variables statistiques (T, P) = (Taille, Poids). Le tableau des effectifs n_{ij} est le suivant :

$P \setminus T$	[145,155[	[155,160[	[160,165[	[165,175[	
[40,45[	20	2	0	0	
[45,50[	9	18	5	1	
[50,55[	1	4	12	7	
[55,60[	0	1	6	14	

Définition :
On peut toujours récupérer les informations sur chacune des variables de la statistiques. Pour cela on calcul les **effectifs marginaux** de chacune des variables, en faisant la somme des effectif dans chaque ligne ou dans chaque colonne.

Exercice 14:
Compléter le tableau de l'exemple 13, en utilisant la dernière ligne pour les effectifs marginaux de la variable T, et la dernière colonne pour les effectifs marginaux de la variable P.

Définition :
La covariance, dans le cadre général, se calcul via la formule

$$\begin{aligned} Cov(X,Y) &= \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{l}[n_{i,j}(x_i-\overline{x})(y_j-\overline{y})] \\ &= \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{l}[n_{i,j}x_iy_j]-\overline{xy} \qquad \text{(Formule de Koenig)} \end{aligned}$$

Les autres paramètres (le coefficient de corrélation linéaire r, et les paramètres de la droite d'ajustement a et b) se calculent de la même manière que précédemment.

Exercice 15:
Calculer la covariance, le coefficient de corrélation linéaire et les paramètres de la droite d'ajustement sur l'exemple 13.

Comme précédemment, on réalisera un nuage de point, auquel on ajoutera la droite d'ajustement linéaire, afin de représenter graphiquement la statistique étudiée.

Exemple 14:
Représentation graphique de la statistique de l'exemple précédent.