

Coquillages & Poincaré

Statistiques descriptives univariées

NASSIRI Mohamed



Statistiques descriptives univariées

Mohamed NASSIRI

Objectifs :

- Analyser un ou plusieurs caractères sur une population.
- Tracer et interpréter des représentations graphiques : tableaux, diagrammes et graphiques.

Mots – clefs :

Série statistique - Tableau à simples entrées - Diagramme à barres - Histogramme - Diagramme circulaire - Boîte à moustache - Polygone des effectifs ou des fréquences cumulés - Diagramme de Kiviat - Nuage de points

Prérequis :

Nombres réels - Fractions - Notion d'ensemble et de quantité - Notions élémentaires de pourcentage



*La musique du chapitre : [MEUTE - REJ](#). Album : *Tumult* - Date de sortie : 2017.*

Dans ce cours sur les statistiques, nous allons apprendre à analyser des données à travers plusieurs outils et représentations. Une série statistique est une collection de données que l'on peut organiser dans un tableau à simples entrées pour les visualiser plus facilement.

Ensuite, nous verrons comment représenter ces données à l'aide de différents diagrammes, comme le diagramme à barres, l'histogramme, ou encore le diagramme circulaire. Pour une analyse plus poussée, nous utiliserons des outils comme la boîte à moustache pour résumer la dispersion des données, et le polygone des effectifs ou des fréquences cumulés pour suivre l'évolution de ces données.

Nous découvrirons aussi des représentations spécifiques comme le diagramme de Kiviat, qui permet de comparer plusieurs variables, et le nuage de points pour étudier la relation entre deux variables. Ces outils vous permettront d'analyser et de comprendre des données de manière claire et précise.

« La recherche a montré que nous finissons pratiquement tous par ressembler à la moyenne des cinq personnes avec lesquelles nous passons le plus de temps. Les individus que vous côtoyez le plus peuvent être le principal facteur conditionnant votre qualité de vie et la personne que vous devenez. Si vous êtes entouré de personnes paresseuses, faibles desprit et qui se cherchent sans cesse des excuses, vous finirez sans doute par leur ressembler. Passez du temps avec des personnes brillantes et positives et leurs attitudes et habitudes pertinentes déteindront sur vous. Vous leur ressemblerez de plus en plus. »

Hal Elrod, *Miracle Morning*.



Sources & liens :

- Manuel scolaire lelivresolaire.fr
- Manuel scolaire *Math'x 2^{de} Nouveau programme*, Editions didier, 2019.
- Site internet INSEE (Institut National de la Statistique et des Etudes Economiques)
- Site internet [Wikipédia](https://fr.wikipedia.org/)
- Chaîne Youtube [Les Bons Profs](#)
- Chaîne Youtube [m@ths et tiques](#)

1 Le tableau à simple entrée

Définition 1 Un tableau statistique est un outil utilisé afin de rassembler des données (le plus souvent chiffrées), sous forme de lignes et de colonnes, et de les rendre ainsi plus faciles à utiliser et à interpréter. Le tableau à simple entrée possède au moins une colonne, mais une seule ligne.

 Remarque


On a très souvent des séries statistiques de valeurs x_1, x_2, \dots, x_p et d'effectifs respectifs n_1, n_2, \dots, n_p que l'on transcrit sous la forme d'un tableau à simple entrée comme ci-dessous :

Valeur	x_1	x_2	...	x_p
Effectif	n_1	n_2	...	n_p

 Exemple

Le tableau fournit les températures (en °C) moyennes mensuelles à Brest.

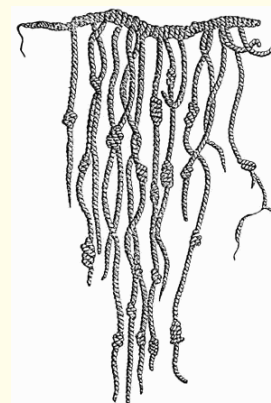
Mois	J	F	M	A	M	J	J	A	S	O	N	D
Brest	9,1	9,4	11	12,5	15,6	18,1	20,4	20,6	18,7	15,3	11,9	10

 Un peu d'histoire

Quipu, quipou, khipu ou quipo, signifie *ú* noeud *z* et *ú* compte *z* en quechua. Le terme désigne aujourd'hui les objets qu'utilisait l'administration inca pour le recensement des données statistiques concernant l'économie et la société de l'empire. En l'absence d'écriture, l'administration figurait les entiers naturels à l'aide de successions de nuds le long de cordelettes de diverses couleurs fixées à une corde : l'ensemble constituait un quipu.

Il est toutefois possible qu'une partie des quipus ait véhiculé une information d'un autre type, notamment des mots-clefs comme payé ou vendu, voire de véritables textes.

Les quipus constituent un système original de consignation de données qui a été développé très tôt dans le Pérou ancien. En effet, certains pourraient dater de quelques milliers d'années comme ceux découverts par Ruth Shady sur le site de la civilisation de Caral remontant à 4 500 ans.



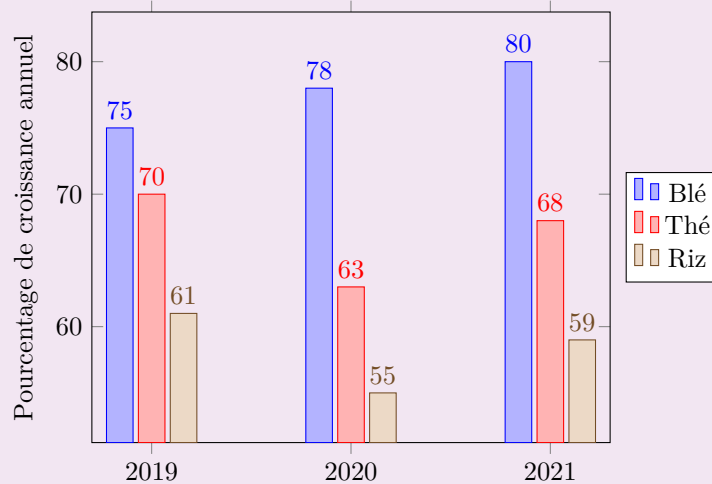
2 Le diagramme à barres

Définition 2 Un diagramme à barres (ou en barres), aussi appelé **diagramme à bâtons** (ou en bâtons), est un graphique représentant des variables avec des barres rectangulaires (verticalement ou horizontalement) avec des hauteurs proportionnelles aux valeurs qu'elles représentent.

 Exemple

Voici des données statistiques et le diagramme à barres correspondant :

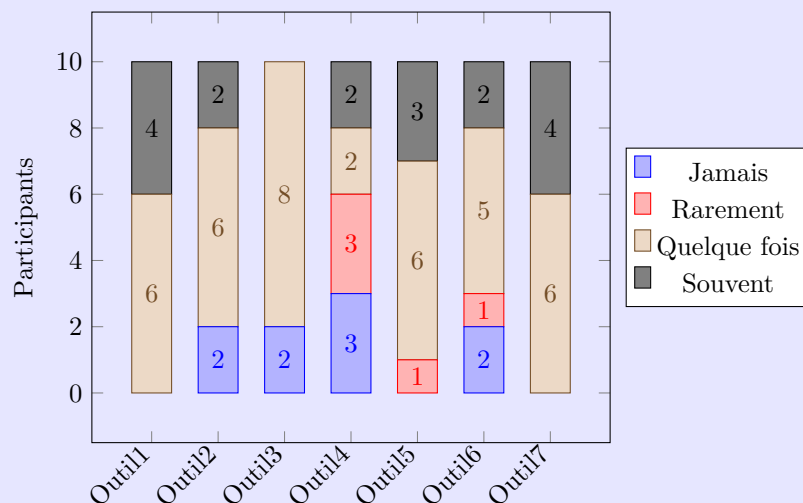
Matières	Blé	Thé	Riz
Quantité en 2019 (en %)	75	70	6
Quantité en 2020 (en %)	70	63	6
Quantité en 2021 (en %)	61	55	59



 Remarque

Il existe une variante que l'on appelle **diagramme à barres empilées**.

Dans l'exemple ci-dessous, on a demandé à 10 personnes leur habitudes de consommation concernant 10 outils numériques.



3 L'histogramme

Définition 3 Un histogramme est un moyen de représenter une série statistique dont le caractère est quantitatif continu. Si la série statistique est donnée par les classes $[a_i, a_{i+1}[$, il est constitué par des rectangles dont la base est le segment $[a_i, a_{i+1}[$ (sur l'axe des réels) et l'aire est proportionnelle à l'effectif de la classe.

 Remarque

Il faut bien noter que c'est l'aire qui doit être proportionnelle à l'effectif de la classe et non la hauteur elle-même. Si toutes les classes ont la même étendue, cela n'a pas d'importance.

Sinon, il faut procéder de la façon suivante : on note n_i l'effectif de la classe $[a_i, a_{i+1}[$. On choisit un rapport de proportionnalité k . La hauteur du rectangle de base $[a_i, a_{i+1}[$ sera $\frac{k \times n_i}{a_{i+1} - a_i}$.

 Exemple

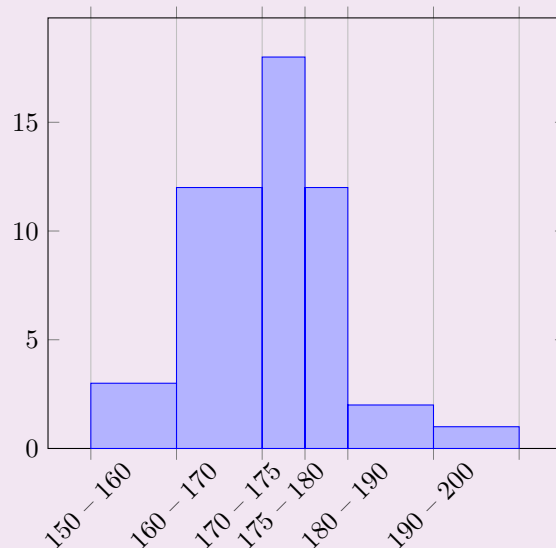
On a demandé la taille des élèves dans une classe de 33 élèves. On obtient les résultats suivants :

Taille (en cm) :	150 – 160	160 – 170	170 – 175	175 – 180	180 – 190	190 – 200
Effectif :	3	12	9	6	2	1

En choisissant un rapport de proportionnalité $k = 10$.

- La hauteur du rectangle de base $[150, 160[$ sera $\frac{10 \times 3}{160 - 150} = 3$.
- La hauteur du rectangle de base $[160, 170[$ sera $\frac{10 \times 12}{170 - 160} = 12$.
- La hauteur du rectangle de base $[170, 175[$ sera $\frac{9 \times 12}{175 - 170} = 18$.
- On réitère l'opération pour tous les intervalles...

L'histogramme correspondant est donc :



4 Le diagramme circulaire

Définition 4 Un diagramme circulaire ou diagramme en secteurs est un type de graphique utilisé en statistiques. Il permet de représenter un petit nombre de valeurs (ou de classes) par des angles proportionnels à la fréquence (ou l'effectif) de ces valeurs.

Exemple

Fatima est une candidate qui prépare sérieusement un concours. Elle décide de lister, par thème, dans un tableau la quantité de notions qu'elle ne maîtrise pas. Voici son tableau :

Matières	Arithmétique	Géométrie	Algèbre	Algorithmique
Nombre de notions	36	72	144	108

Pour mieux se rendre compte de la répartition des notions, elle décide de faire un diagramme circulaire. Elle doit « convertir » ces données en secteurs d'angle.

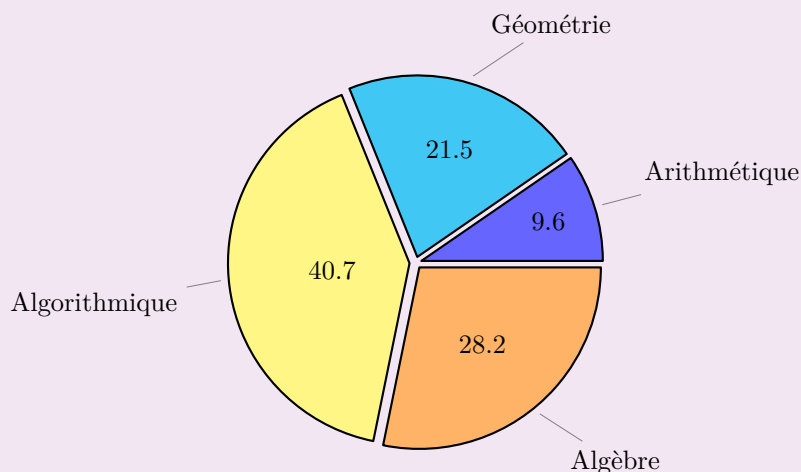
Matières	Arithmétique	Géométrie	Algèbre	Algorithmique	Total
Nombre de notions	17	38	72	50	177
Pourcentage	9,6	21,5	40,7	28,2	100
Secteur d'angle	17	38	72	50	177

Pour déterminer les secteurs d'angles, Aude a tout simplement réalisé un produit en croix. Le nombre total de notions, à savoir 177, correspond à l'angle total, à savoir 360°.

Par exemple, pour l'angle correspondant à la géométrie, on a le calcul suivant :

$$\text{Angle recherché} = \frac{38 \times 360}{177} \simeq 21,5$$

Nombre de notions	Angle correspondant
38	Angle recherché
177	360

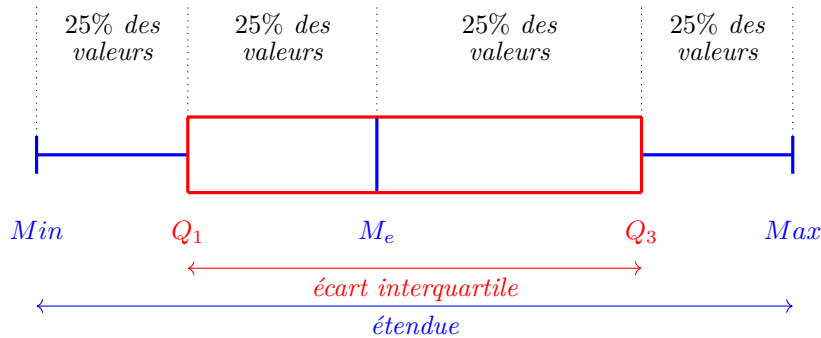


Remarque

À cause d'erreurs d'arrondis, il peut arriver que la somme des pourcentages ne donne pas 100, tout comme la somme des angles au centre ne donne pas 360°.

5 La boîte à moustache

Définition 5 Il est possible de résumer, sous la forme d'un graphique, l'information fournie par l'étendue, ainsi que par la médiane, les deux quartiles et les intervalles qui les séparent. Ce graphique porte le nom de **boîte à moustaches**, ou encore de **boîte à pattes** ou **diagramme en boîte** (boxplot en anglais).



Remarque

On parle aussi de **boîte de Tukey** (de son inventeur, John Tukey, en 1977).



Exemple

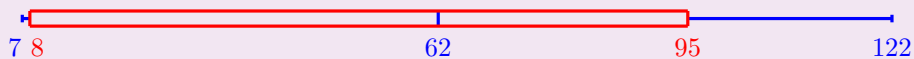
Réalisons la boîte à moustache de la série suivante :

73; 122; 73; 8; 95; 43; 64; 44; 32; 111; 115; 7; 7; 8; 111; 60

On trouve :

$Min = 7$; $Q_1 = 8$; $Me = 62$; $Q_3 = 95$; $Max = 122$

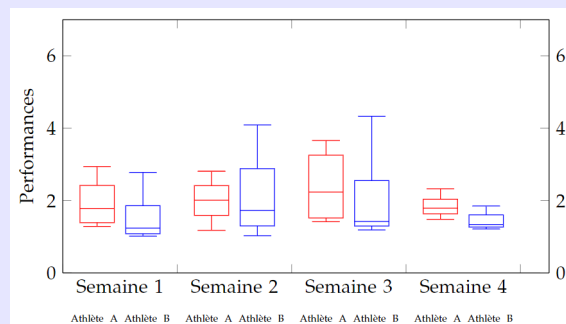
On a donc



Remarque

À noter qu'une boîte à moustache seule n'est pas vraiment très intéressante... On les utilise souvent pour comparer plusieurs séries statistiques car c'est un moyen rapide de figurer le profil essentiel de séries statistiques quantitatives.

Par exemple, dans le graphique ci-dessous, nous pouvons comparer rapidement les performances deux athlètes sur 4 semaines.



6 Le polygone des effectifs ou des fréquences cumulés

Définition 6 L'effectif d'une valeur du caractère étudié est le nombre d'individus de la population ayant cette valeur.

La **fréquence** d'une valeur est le quotient de l'effectif de cette valeur par l'effectif total de la population. (la fréquence peut être exprimée en pourcentage)

$$\text{fréquence} = \frac{\text{effectif de la valeur}}{\text{effectif total}}$$



Remarque

On étudie un caractère quantitatif dans une population.

- Les différentes valeurs x_i du caractère quantitatif constituent une série statistique notée (x_i) .
- On note n_i l'effectif de la valeur x_i .

Définition 7 L'effectif cumulé croissant de la valeur x_i est la somme des effectifs de toutes les valeurs inférieures ou égales à x_i .

La **fréquence cumulée croissante** de la valeur x_i est la somme des fréquences de toutes les valeurs inférieures ou égales à x_i .



Exemple

Dans un service de maintenance, on a répertorié le nombre d'interventions par jour sur un mois. On a obtenu la distribution suivante :

Nombre d'interventions x_i	3	5	6	7	8	9
Nombre de jours n_i	2	4	9	6	3	1

Le nombre total de journées d'intervention est $2 + 4 + 9 + 6 + 3 + 1 = 25$. Les fréquences des différentes valeurs du nombre d'intervention sont :

Nombre d'interventions x_i	3	5	6	7	8	9
Nombre de jours n_i	2	4	9	6	3	1
Fréquence $f_i = \frac{n_i}{25}$	0,08	0,16	0,36	0,24	0,12	0,04

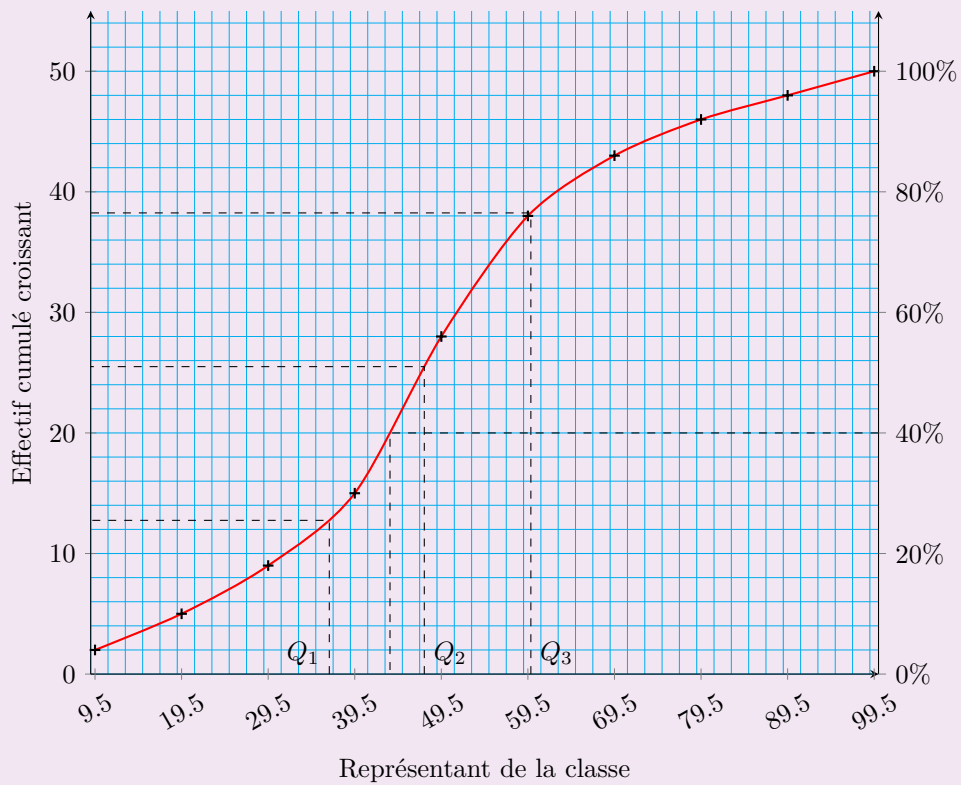
Le tableau suivant donne les effectifs cumulés croissants ainsi que les fréquences cumulées croissantes :

Nombre d'interventions x_i	3	5	6	7	8	9
Nombre de jours n_i	2	4	9	6	3	1
Effectif cumulé	2	6	15	21	24	25
Fréquence cumulée	0,08	0,24	0,6	0,84	0,96	1

💡 Exemple

Réalisons le polygone des effectifs et des fréquences cumulés croissantes

Classe (%)	Effectif	Représentant de la classe	Effectif cumulé croissant
0 – 9	2	9.5	2
10 – 19	3	19.5	5
20 – 29	4	29.5	9
30 – 39	6	39.5	15
40 – 49	13	49.5	28
50 – 59	10	59.5	38
60 – 69	5	69.5	43
70 – 79	3	79.5	46
80 – 89	2	89.5	48
90 – 99	2	99.5	50
Total	50		



7 La diagramme de Kiviat

Définition 8 Le diagramme de Kiviat, ou diagramme en toile d'araignée sert à représenter sur un plan en deux dimensions au moins trois ensembles de données multivariées. Chaque axe, qui part d'un même point, représente une caractéristique quantifiée. Est ainsi facilitée une analyse détaillée de plusieurs objets, ainsi que leur comparaison générale (comparaison des surfaces) ou point par point. Ce type de diagramme n'est utile que si les axes sont correctement normés selon l'importance donnée à chaque caractéristique.



Remarque

- Il est aussi appelé **diagramme en radar** ou encore **diagramme en étoile**.
- Il fut créé en 1877 par le statisticien allemand Georg von Mayr.

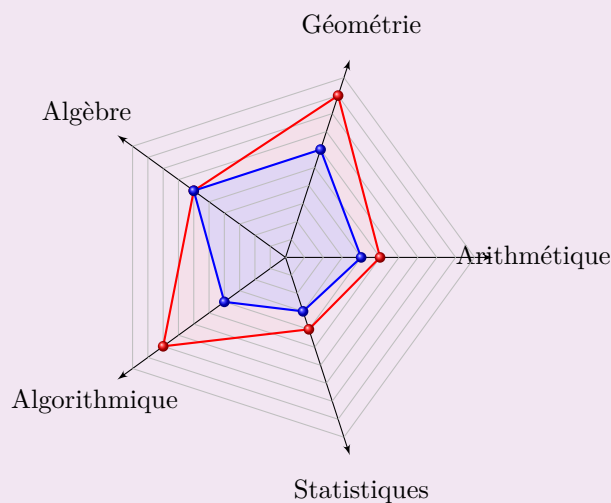


Exemple

Considérons les données suivantes

Matières	Arithmétique	Géométrie	Algèbre	Algorithmique	Statistiques
Notes de l'élève A	5	9	6	8	4
Notes de l'élève B	4	6	6	4	3

Alors le diagramme de Kiviat est



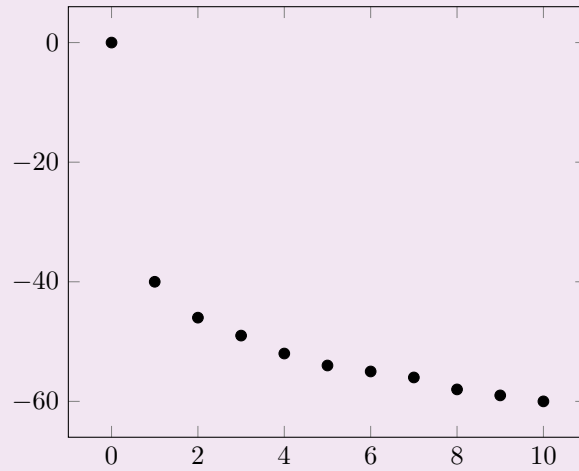
8 Nuage de points

Définition 9 Étant donné deux caractères d'une série statistique, le nuage de points associé est, dans un repère, l'ensemble des points ayant pour abscisses les valeurs du premier caractère et pour ordonnées les valeurs du second.

 Exemple

Considérons le tableau de données suivants (volontairement sans contexte) et traçons son nuage de points

X	0	1	2	3	4	5	6	7	8	9	10
Y	0	-40	-46	-49	-52	-54	-55	-56	-58	-59	-60

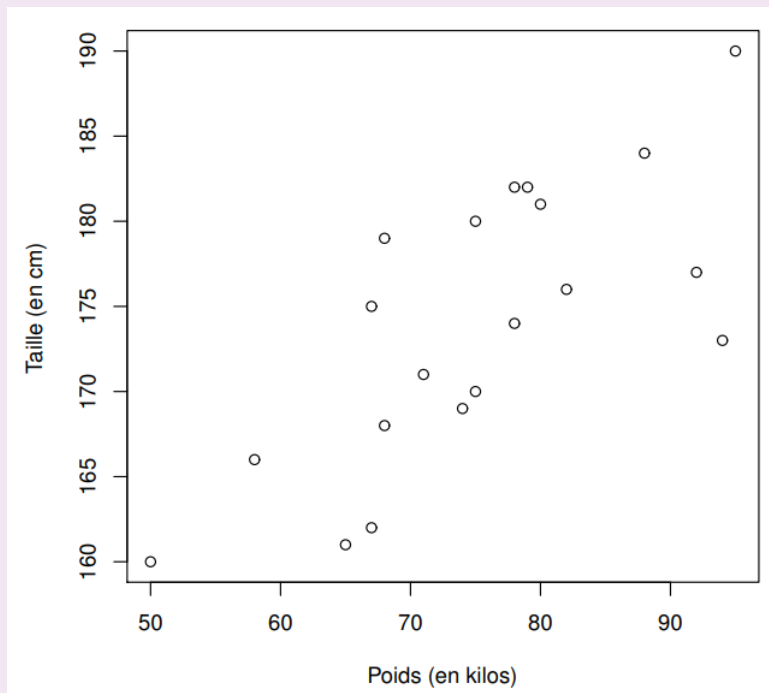


 Exemple

On a relevé le poids X (en kilos) et la taille Y (en centimètres) de 20 individus dans le tableau suivant :

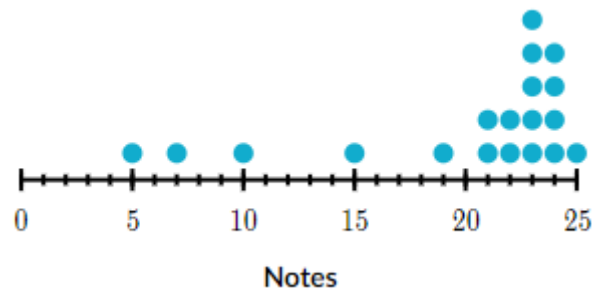
Poids	67	71	92	74	75	94	79	58	65	68
Taille	162	171	177	169	170	173	182	166	161	179
Poids	88	78	78	75	67	95	80	50	82	68
Taille	184	174	182	180	175	190	181	160	176	168

Voici le nuage de points (aussi dit diagramme de dispersion) correspondant à cette double distribution



9 Ouverture

Question ouverte : Le graphique à points suivant donne la distribution des notes obtenues par 19 candidats au permis de conduire à un test sur le code de la route.



Comment pourrait-on définir mathématiquement les *valeurs aberrantes* ?

Culture scientifique : Les statistiques à grande échelle et Python

Sur Internet, il est possible de trouver le fichier *titanic.csv* (et plein d'autres fichiers de ce type!). Il s'agit d'un fichier recensant plusieurs informations sur les survivants et morts du Titanic : s'ils ont survécu ou non, leur nom, prénoms, âge, etc.

J'ai pu trouver un fichier recensant 887 personnes, et j'ai voulu tracer une boîte à moustache me donnant la répartition des âges des personnes étant sur le Titanic.

A l'aide du module **pandas** qui permet de lire des fichiers *.csv* (plus précisément, n'importe quelles informations se trouvant dedans!) et du module **matplotlib**, il est possible de réaliser un tel diagramme sans faire le relevé soi-même à la main! Voici le script et son résultat :

```
import matplotlib.pyplot as plt
import pandas

data=pandas.read_csv('titanic.csv',sep=',')

agetitanic=data.Age

plt.boxplot(agetitanic)

plt.title("Boîte à moustaches pour la répartition des âges sur le Titanic")
plt.show()
```

