Statistiques Descriptives – 2nde

A) Généralités en Statistiques

1) Vocabulaire statistique

Définitions : On donne les différentes bases de vocabulaire

- **Population** : Ensemble avec lequel on travaille les données
- **Échantillon** : extraction d'individus de la population
- **Caractère** : type d'élément que l'on étudie
    - **qualitatif** : la variable étudiée n'est pas numérique
    - **quantitatif** : la variable étudiée est numérique
        - **discret** : les variables sont des entiers naturels
        - **continu** : les variables sont des nombres réels
- **Effectif** : valeur associée à une variable donnée
- **Effectif total** : ensemble des valeurs de la population

Exemples :

- Le nombre de frères et sœurs d'un élève de seconde : quantitatif discret
- La taille des élèves de seconde : quantitatif continu
- La couleur des yeux des élèves de 2nde : qualitatif
- Le mois de naissance des élèves de 2nde : quantitatif discret
- Le type de loisir des élèves de 2nde : qualitatif

2) Représentations graphiques

Définition : On donne les différents types de graphiques utilisés en statistique

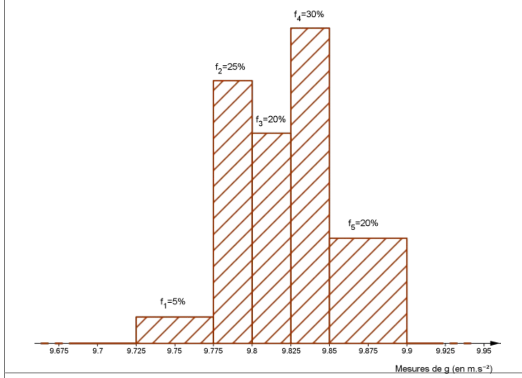
- Le **diagramme circulaire** : utilisé pour comparer des caractères qualitatifs
- L'**histogramme** : utilisé pour comparer des caractères quantitatifs continus
- Le **diagramme en barres** : utilisé pour comparer des caractères quantitatifs discrets (avec une ou plusieurs séries)
- Le **diagramme de Tukey** : utilisé pour comparer des caractères quantitatifs discrets (avec une ou plusieurs séries)

Exemples : Déterminer quels types de diagrammes est le plus adapté

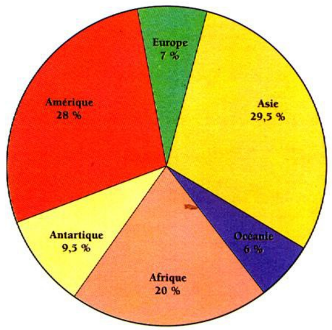
- Le nombre de frères et sœurs d'un élève de seconde : diagramme en barres
- La taille des élèves de seconde : histogramme
- La couleur des yeux des élèves de 2nde : diagramme circulaire
- Le mois de naissance des élèves de 2nde : diagramme en barres
- Le type de loisir des élèves de 2nde : diagramme de Tukey

Les différents types de graphiques sont :

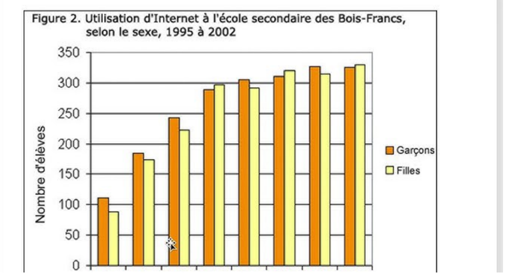
les ***histogrammes***



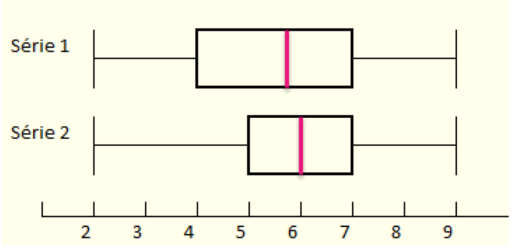
les ***diagrammes circulaire***



les ***diagrammes en barres***



les ***diagrammes de Tukey***
(ou « en boîtes »)



B) Les paramètres de position

1) Les moyennes

Définition : On étudie ici uniquement les moyennes de type arithmétiques

- Si les valeurs $x_1, x_2, x_3, \dots, x_n$ sont dites simples alors la ***moyenne arithmétique simple*** est $\bar{x} = \frac{1}{n} . \sum_{i=1}^{n} x_i$
- Si les valeurs $x_1, x_2, x_3, \dots, x_p$ sont pondérées par les coefficients $n_1, n_2, n, \dots, n_p$ alors la ***moyenne arithmétique pondérée*** est $\bar{x} = \frac{1}{N} . \sum_{i=1}^{p} n_i . x_i$ où $N = \sum_{i=1}^{p} n_i$ correspond à la somme des coefficients

Exemples : Calculer les moyennes de notes obtenues par 3 élèves :

- Jérôme : 4 ; 6 ; 18 ; 7 ; 17 ; 12 ; 12 ; 18
- Bertrand : 13 ; 13 ; 12 ; 10 ; 12 ; 3 ; 14 ; 12 ; 14 ; 15
- Julie : 15 ; 9 ; 14 ; 13 ; 10 ; 12 ; 12 ; 11 ; 10

exemple : Calculer les moyennes des notes coefficientées par 3 élèves :

- Jérôme : (4;1) ; (6;2) ; (18;1) ; (7;1) ; (17;2) ; (12;1) ; (12;2) ; (18;1)
- Bertrand : (13;1) ; (13;2) ; (12;1) ; (10;2) ; (12;1) ; (3;2) ; (14;1) ; (12;1) ; (14;1) ; (15;1)
- Julie : (15;1) ; (9;2) ; (14;1) ; (13;1) ; (10;2) ; (12;1) ; (12;1) ; (11;1) ; (10;2) ; (11;1)

Comparer ces 3 élèves en calculant leur moyennes arithmétiques respectives

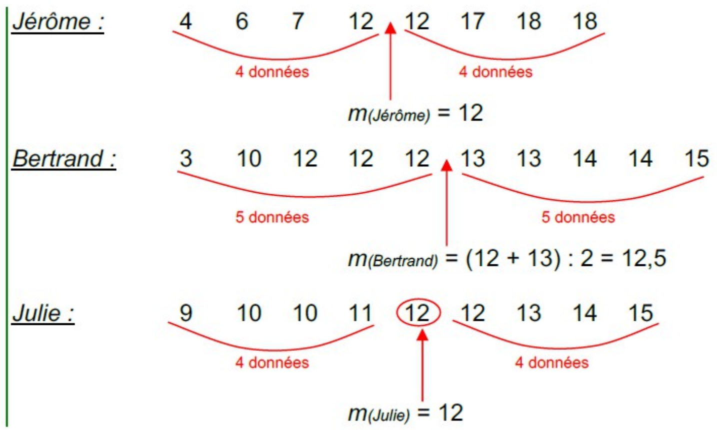
2) Les quartiles

Méthode : *Calculer une médiane*
Vidéo : https://youtu.be/kr90dXv0NFY

Voici les séries de notes obtenues par 3 élèves :

- Jérôme : 4 ; 6 ; 18 ; 7 ; 17 ; 12 ; 12 ; 18
- Bertrand : 13 ; 13 ; 12 ; 10 ; 12 ; 3 ; 14 ; 12 ; 14 ; 15
- Julie : 15 ; 9 ; 14 ; 13 ; 10 ; 12 ; 12 ; 11 ; 10

Calculer la médiane pour Jérôme, Bertrand et Julie.



Définition : Soit $x_1, x_2, x_3, \dots, x_p$ valeurs pondérées par les coefficients $n_1, n_2, n, \dots, n_p$ alors on définit les 3 quartiles de la série :

- le **1er quartile** Q_1 correspond à la valeur de $x_{N/4}$
- le **2eme quartile** Q_2 correspond à la valeur de $x_{N/2}$
- le **3eme quartile** Q_3 correspond à la valeur de $x_{3N/4}$

Rque : On peut par prolongement définir également les **déciles**, les **centiles**, ...
Par ailleurs, il est évident que la médiane est égale au 2eme quartile

exemples : déterminer les quartiles des 4 séries ci-dessous

Série **A**

x_i	7	8	9	10	12	14	15	17
n_i	2	3	1	4	2	1	3	1

Série **B**

x_i	7	8	9	10	12	14	15	17
n_i	1	3	4	5	1	1	2	1

Série **C**

x_i	7	8	9	10	12	14	15	17
n_i	2	4	1	3	1	4	1	3

Série **D**

x_i	7	8	9	10	12	14	15	17
n_i	3	2	2	3	4	3	2	1

Réponses : on obtient les résultats suivants

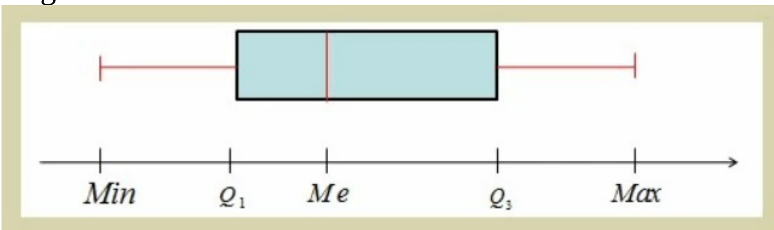
- Série **A** : $Q_1=8$; $Q_2=10$; $Q_3=14{,}5$
- Série **B** : $Q_1=9$; $Q_2=10$; $Q_3=12$
- Série **C** : $Q_1=8$; $Q_2=10$; $Q_3=14$
- Série **D** : $Q_1=8{,}5$; $Q_2=10{,}5$; $Q_3=14$

Rque : le ***2ème quartile*** s'appelle aussi la ***médiane*** puisque $\frac{N}{2}=\frac{2N}{4}$

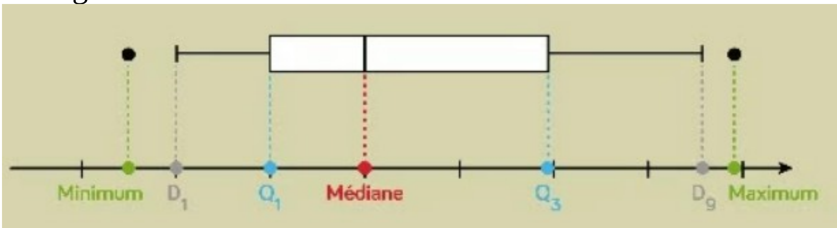
3) Le diagramme de Tukey

Définition : Il existe 2 types de diagrammes « en boîtes »

- Le diagramme en boîtes « sans valeur aberrante »



- Le diagramme en boîtes « avec valeurs aberrantes »

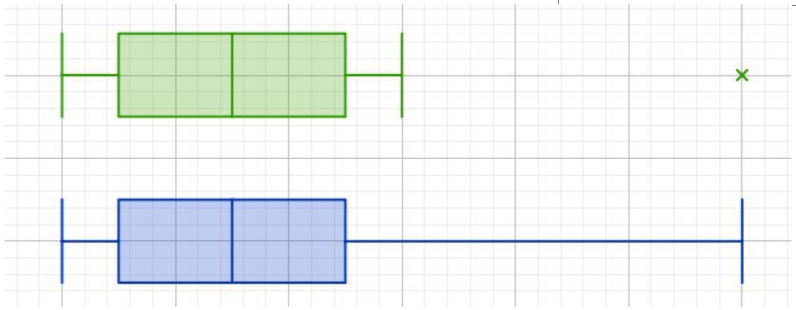


avec $D_1=Q_1-1{,}5(Q_3-Q_1)$ et $D_9=Q_3+1{,}5(Q_3-Q_1)$

Exemple : On donne la série statistique suivante ; construire les 2 types de diagrammes de Tukey

x_i	6	7	9	11	12	13	15	18
n_i	2	3	2	4	1	0	0	1

On obtient :



$Min=6$

$Q_1=7$

$Q_2=9$

$Q_3=11$

$Max=18$

Interprétation : On remarque que le diagramme de Tukey incluant les « *Outliers* » est beaucoup plus précis car il met en évidence des cas particuliers

C) Les paramètres de dispersion

1) Les écarts de bases

Définition : Soit $x_1, x_2, x_3, \dots, x_p$ valeurs pondérées par les coefficients $n_1, n_2, n, \dots, n_p$ alors on a 2 types d'écarts « basiques » :

- l'étendue est : $e=Max(x_i)-Min(x_i)$
- l'écart inter-quartile est $IQR=Q_3-Q_1$

Exemple : On donne la série statistique suivante ; calculer l'étendue et l'écart inter-quartile ; On obtient :
$e=18-6=12$ et $IQR=11-7=4$

x_i	6	7	9	11	12	13	15	18
n_i	2	3	2	4	1	0	0	1

2) La Variance & écart-type

Définition : Soit $x_1, x_2, x_3, \dots, x_p$ valeurs pondérées par les coefficients $n_1, n_2, n, \dots, n_p$ alors on a :

- la Variance est $V=\frac{1}{N}.\sum_{i=1}^{p} n_i.(x_i)^2-(\overline{x})^2$
- l'écart-type est $\sigma=\sqrt{V}$

<u>*exemple*</u> : Soit la série suivante :

$V=\frac{2161}{18}-10{,}5^2\simeq 9{,}8056$

$\sigma=\sqrt{9{,}8056}\simeq 3{,}13$

x_i	6	7	9	11	12	13	15	18
n_i	2	3	2	5	2	2	1	1

<u>*Rque*</u> : l'écart-type correspond alors à la moyenne quadratique des valeurs de la série $(x_i; n_i)$ soit $\sigma=\sqrt{\frac{1}{N}.\sum_{i=1}^{p} n_i.(x_i-\overline{x})^2}$; on trouve $\sigma=\sqrt{9{,}81}\simeq 3{,}13$

<u>*exercice*</u> : On reprend les 4 Séries **A, B, C, D** précédentes

Série **A**

x_i	7	8	9	10	12	14	15	17
n_i	2	3	1	4	2	1	3	1

Série **B**

x_i	7	8	9	10	12	14	15	17
n_i	1	3	4	5	1	1	2	1

Série **C**

x_i	7	8	9	10	12	14	15	17
n_i	2	4	1	3	1	4	1	3

Série **D**

x_i	7	8	9	10	12	14	15	17
n_i	3	2	2	3	4	3	2	1

a) Calculer les moyennes des 4 séries
b) Calculer les variances des 4 séries
c) Calculer les écarts-type des 4 séries
d) Interpréter les résultats

<u>*Réponses*</u> :

Série **A** :
$\overline{x}=11$
$V\simeq 9{,}53$
$\sigma\simeq 3{,}18$

Série **B** :
$\overline{x}\simeq 10{,}56$
$V\simeq 7{,}60$
$\sigma\simeq 2{,}77$

Série **C** :
$\overline{x}\simeq 11{,}53$
$V\simeq 12{,}06$
$\sigma\simeq 3{,}48$

Série **D** :
$\overline{x}=11{,}1$
$V=8{,}79$
$\sigma\simeq 3{,}04$

Interprétations : On peut effectuer le classement suivant des 4 séries

- la meilleure série est la série C
- la série la plus régulière (ou homogène) est la série B
- le classement est donc : C – D – B – A

<u>Rque</u> : En statistique comme dans de nombreux domaines on privilégie la qualité de « régularité » au lieu de la qualité de « performance »