Statistiques

1. Données brutes

Le résultat d'une étude statistique quantitative est une liste de nombres appelée série statistique.

Exemple :

Les densités de population en habitants par km² de 25 pays de l'union européenne (source wikipédia, 2007). (Remarque : actuellement, il y a 27 pays dans l'union européenne).

Pays	Densité
Allemagne	231
Autriche	98
Belgique	340
Chypre	84
Danemark	126
Espagne	80
Estonie	29
Finlande	15
France	111
Grèce	81
Hongrie	108
Irlande	57
Italie	193
Lettonie	35
Lituanie	55
Luxembourg	181
Malte	1261
Pays-Bas	395
Pologne	124
Portugal	114
Rép. Tchèque	130
Royaume-Uni	243
Slovaquie	111
Slovénie	99
Suède	20

Les valeurs doivent être ordonnées (de préférence dans l'ordre croissant).

Pour l'exemple, il faut écrire 2 fois 111 : une fois pour la France et une fois pour la Slovaquie.

1	15
2	20
3	29
4	35
5	55
6	57
7	80
8	81
9	84
10	98
11	99
12	108
13	111
14	111
15	114
16	124
17	126
18	130
19	181
20	193
21	231
22	243
23	340
24	395
25	1261

1.1. Médiane. Quartiles. Déciles.

a) **Définitions**

La médiane d'une série est le nombre Me qui partage la série en deux sous-séries de même effectif.

Le premier quartile d'une série, noté Q_1 , est la plus petite valeur de la série telle que 25% de valeurs de la série soient inférieures ou égales à Q_1 .

Le troisième quartile d'une série, noté Q_3 , est la plus petite valeur de la série telle que 75% de valeurs de la série soient inférieures ou égales à Q_3 .

L'écart interquartile est le nombre égal à $Q_3 - Q_1$.

Le premier décile d'une série, noté D_1 , est la plus petite valeur de la série telle que 10% de valeurs de la série soient inférieures ou égales à D_1 .

Le neuvième quartile d'une série, noté D_9 , est la plus petite valeur de la série telle que 90% de valeurs de la série soient inférieures ou égales à D_9 .

b) **Remarques**

Dans la pratique :

- Si l'effectif total est n , impair, alors Me est la valeur centrale de la série, c'est à dire celle de rang $\frac{n+1}{2}$
- Si l'effectif total est n , pair, alors Me est la demi-somme des deux valeurs centrales de la série, c'est à dire celles de rang $\frac{n}{2}$ et $\frac{n}{2}+1$.
- Q_1 et Q_3 se déterminent de la même façon quel que soit l'effectif total : le rang de Q_1 est l'entier le plus proche supérieur ou égal à $\frac{1}{4}\times n$; le rang de Q_3 est l'entier le plus proche supérieur ou égal à $\frac{3}{4}\times n$.
- D_1 et D_9 se déterminent de la même façon quel que soit l'effectif total : le rang de D_1 est l'entier le plus proche supérieur ou égal à $\frac{1}{10}\times n$; le rang de D_9 est l'entier le plus proche supérieur ou égal à $\frac{9}{10}\times n$.

c) **Pour l'exemple**

L'effectif total est 25.

$\frac{25}{2}=12,5$ donc la médiane est au $13^{\text{ième}}$ rang, soit $Me=111$.

$\frac{1}{4}\times 25=6,25$ donc le premier quartile est au $7^{\text{ième}}$ rang, soit $Q_1=80$.

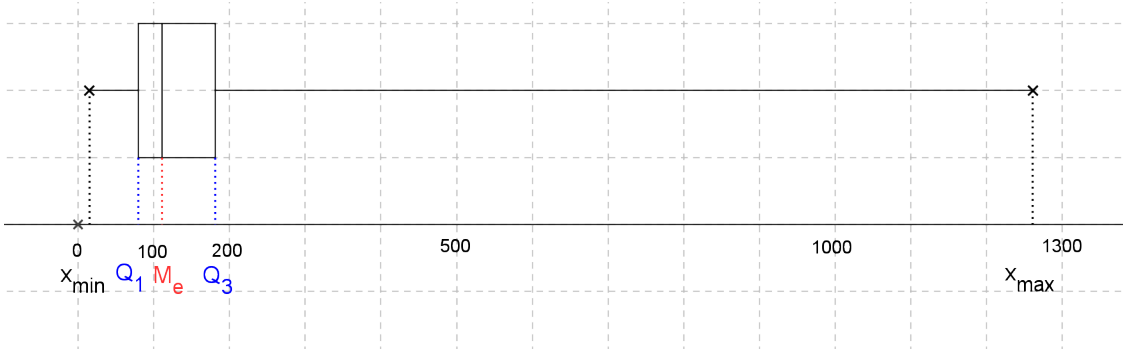
$\frac{3}{4}\times 25=18,25$ donc le troisième quartile est au $19^{\text{ième}}$ rang, soit $Q_3=181$.

$\frac{1}{10}\times 25=2,5$ donc le premier décile est au $3^{\text{ième}}$ rang, soit $D_1=29$.

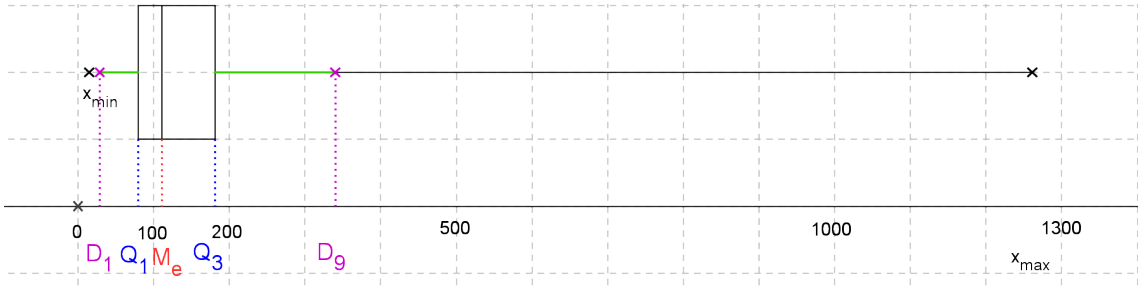
$\frac{9}{10}\times 25=22,5$ donc le neuvième décile est au $23^{\text{ième}}$ rang, soit $D_3=340$.

d) **Diagramme en boîte**

A l'aide de la médiane et des quartiles, ainsi que de la valeur minimale (notée x_{min})et de la valeur maximale de la série (notée x_{max}), on construit un diagramme en boîte (ou boîte à moustaches).



Il existe aussi un diagramme en boîte élagué dont les extrémités sont le premier décile et le neuvième décile.



1.2. Moyenne. Écart type.

a) **Définitions**

La moyenne d'une série statistique donnée sous forme de liste est égale à :

$$\text{moyenne} = \frac{\text{somme des valeurs}}{\text{effectif total}}$$

soit encore

$$\bar{x} = \frac{\sum x_i}{n}$$

Remarque :

En règle générale, la moyenne et la médiane sont différentes.

La variance d'une série statistique est le nombre :

$$V = \frac{\sum \left(x_i - \bar{x}\right)^2}{n}$$

L'écart type d'une série statistique est le nombre :

$$\sigma = \sqrt{V}$$

b) **Exemple**

La moyenne est :

$$\bar{x} = \frac{15 + 20 + 29 + \ldots + 1261}{25} \approx 172{,}84$$

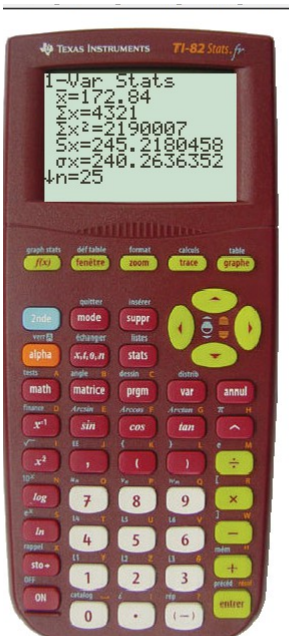
La variance est :

$$V = \frac{(15-172{,}84)^2 + (20-172{,}84)^2 + \ldots + (1261-172{,}84)^2}{25} \approx 57726{,}61$$

L'écart-type est :

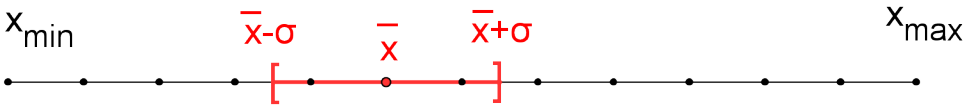
$$\sigma = \sqrt{V} \approx 240{,}26$$

A la calculatrice, on obtient l'écran suivant :



c) **Remarque**

Plus l'écart-type est petit, plus les valeurs de la série sont concentrées autour de la moyenne.



On retiendra que l'intervalle $[\bar{x}-\sigma;\bar{x}+\sigma]$ contient au moins 68% des valeurs de la série. On appelle cet intervalle **plage de normalité au seuil de 68%**.

On retiendra que l'intervalle $[\bar{x}-2\sigma;\bar{x}+2\sigma]$ contient au moins 95% des valeurs de la série. On appelle cet intervalle **plage de normalité au seuil de 95%**.

2. Données regroupées

Les données peuvent être regroupées par effectifs, ou par intervalles (appelés classes).

Données regroupées par effectifs:

Notes x_i	10	12	13	18
Nombre d'élèves n_i	5	8	10	1

Données regroupées en classes :

Taille des élèves en cm x_i	[150;160[	[160;165[	[165;180[	[180;200[
Nombre d'élèves n_i	4	12	15	2

2.1. Médiane ; Intervalle interquartile.

On est amené à calculer **les effectifs cumulés croissants** de la série étudiée.

Exemple :

Notes x_i	10	12	13	18
Nombre d'élèves n_i	5	8	10	1
Eff.cum.croissants	5	13	23	24

La médiane est la demi-somme des 12ième et 13ième valeurs, soit $Me=\frac{12+12}{2}=12$.

Le premier quartile est au rang 6 $\left(\frac{1}{4}\times 24=6\right)$, soit $Q_1=12$.

Le troisième quartile est au rang 18 $\left(\frac{3}{4}\times 24=18\right)$, soit $Q_3=13$.

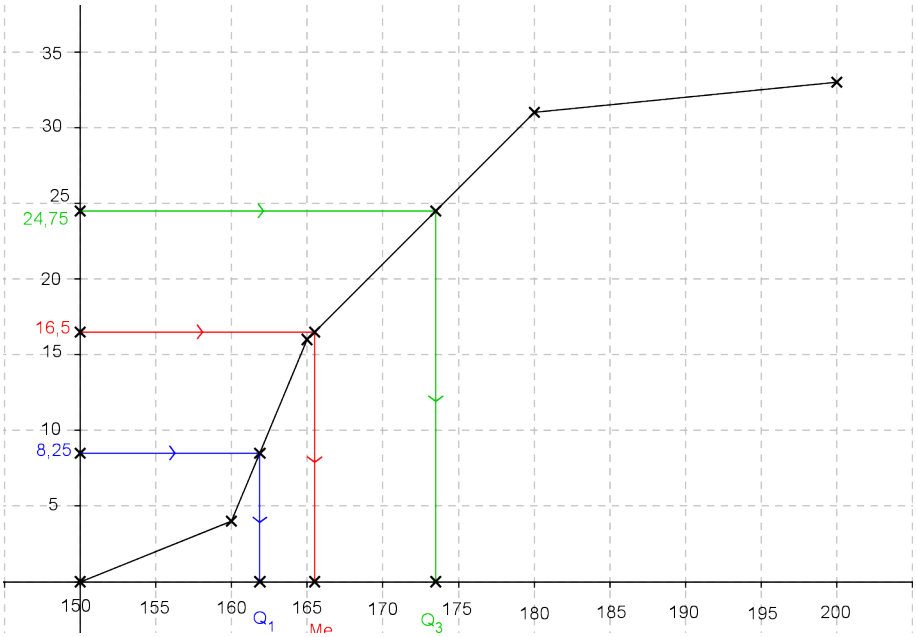
L'écart interquartile est donc $Q_3-Q_1=13-12=1$.

Exemple :

Taille des élèves en cm x_i	[150;160[	[160;165[	[165;180[	[180;200[
Nombre d'élèves n_i	4	12	15	2
Eff.cum.croissants	4	16	31	33

La médiane est au rang 17, elle se situe donc dans l'intervalle [165;180[.
Cet intervalle se nomme **intervalle médian**.

Pour connaître une valeur plus précise de la médiane, on construit le **polygone des effectifs cumulés croissants**.
Il passe par les points de coordonnées (150, 0), (160, 4), (165, 16), (180, 31) et (200, 33).



Par lecture graphique, $Me \approx 165{,}5$, $Q_1 \approx 162$, $Q_3 \approx 173{,}5$.

2.2. Moyenne-Écart type.

a) **Définitions**

La moyenne d'une série dont les valeurs sont regroupées est :

$$\text{Moyenne} = \frac{\text{somme des (valeurs} \times \text{effectifs)}}{\text{effectif total}}$$

ou

$$\text{Moyenne} = \frac{\text{somme des (centres des classes} \times \text{effectifs)}}{\text{effectif total}}$$

ou encore

$$\overline{x} = \frac{\sum x_i \times n_i}{n}$$

La variance d'une série regroupée est :

$$V = \frac{\sum \left(\left(x_i - \overline{x} \right)^2 \times n_i \right)}{n}$$

L'écart-type est :

$$\sigma = \sqrt{V}$$

b) **Exemple**

On donne la répartition des auditeurs (en milliers) d'une radio suivant leur âge :

Age	[8;13[	[13;16[	[16;18[	[18;22[	[22;27[
Centre	$\frac{3+13}{2}=10,5$	$\frac{13+16}{2}=14,5$	$\frac{16+18}{2}=17$	$\frac{18+22}{2}=20$	$\frac{22+27}{2}=24,5$
Effectif	15	27	24	24	10

La moyenne est $\bar{x}=\frac{10,5\times 15+14,5\times 27+\ldots+24,5\times 10}{100}=16,82$.

La variance est $V=\frac{(10,5-16,82)^2\times 15+(14,5-16,82)^2\times 27+\ldots+(24,5-16,82)^2\times 10}{100}=15,7776$

L'écart-type est $\sigma=\sqrt{15,7776}\approx 3,97$

2.3. Complément:histogrammes.

On considère une série de donnée regroupée en classe. **Un histogramme** est un diagramme constitué de rectangles dont les bases sont les classes de la série et dont les aires sont proportionnelles à l'effectif (ou à la fréquence) de la classe.

Exemple : Avec la série des auditeurs de radio.