Chapitre 4. Statistiques descriptives : Paramètres de position

Mathématiques et statistiques appliquées

Département TC1-IUT de Sceaux

Damien Thomine

Objectifs

- Savoir calculer et interpréter la moyenne et la médiane.
- Savoir calculer et interpréter quartiles et déciles.
- Faire le lien avec les représentations de données (histogrammes, fréquences cumulées).

Plan du cours

Paramètres statistiques

Les **paramètres statistiques**

sont des nombres qui ont pour but de résumer l'essentiel de l'information relative à une variable statistique quantitative.

- Les **paramètres de position** indiquent la valeur "typique" autour de laquelle les observations sont réparties.
  Les deux paramètres de position les plus importants sont la **moyenne** et la **médiane**.
- Les **fractiles**, notamment les **quartiles** et les **déciles**, donnent des informations plus fines sur la série. Ils sont une généralisation de la médiane.
- Les **paramètres de dispersion** mesurent combien les observations s'éloignent de la valeur centrale. Le plus important est l'**écart-type**.

Section 1

Moyenne

Moyenne

La **moyenne** d'une variable est la somme des valeurs observées x_i divisée par le nombre total n d'individus dans la population :

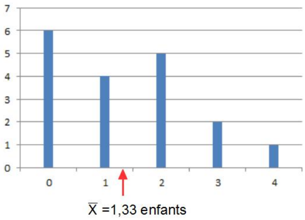
$$\bar{x} = \frac{x_1 + \cdots + x_n}{n}$$

Exemple. À droite, la série brute d'une enquête sur le nombre d'enfants de 18 familles d'un immeuble.

$\overline{x} = \frac{0+0+0+0+0+0+1+1+1+1+2+2+2+2+2+3+3+4}{18} \simeq 1,33$

Il y a en moyenne 1,33 enfants par famille.

Famille N°	Nb Enfants
1	0
2	0
3	0
4	0
5	0
6	0
7	1
8	1
9	1
10	1
11	2
12	2
13	2
14	2
15	2
16	3
17	3
18	4



Calcul de la moyenne à partir des effectifs et fréquences

On peut calculer la moyenne à partir de modalités $y_1, \dots, y_p$ et des effectifs n_i ou fréquences f_1 associés.

$$\overline{x} = \frac{y_1 \times n_1 + \cdots + y_p \times n_p}{n} = y_1 \times f_1 + \cdots + y_p \times f_p$$

Exemple. À partir du tableau des effectifs

Nb. d'enfants y_i	0	1	2	3	4	Total
Effectif n_i	6	4	5	2	1	18

$$\overline{x} = \frac{0\times6+1\times4+2\times5+3\times2+4\times1}{18} \simeq 1,33$$

À partir du tableau des fréquences :

Nb. d'enfants y_i	0	1	2	3	4	Total
Fréquence f_i	33%	22%	28%	11%	6%	100%

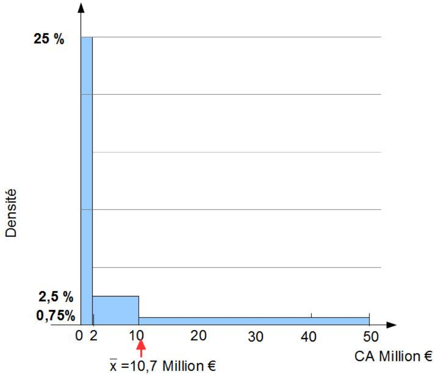
$$\overline{x} = 0\times0,33+1\times0,22+2\times0,28+3\times0,11+4\times0,06 \simeq 1,33$$

Famille N°	Nb Enfants
1	0
2	0
3	0
4	0
5	0
6	0
7	1
8	1
9	1
10	1
11	2
12	2
13	2
14	2
15	2
16	3
17	3
18	4

Calcul de la moyenne avec des classes

Si on dispose seulement des effectifs par classe, on peut calculer une valeur approchée de la moyenne en assignant à chaque classe $[e_i, e_{i+1}[$ son **centre** $c_i = \frac{e_i + e_{i+1}}{2}$.

On utilise ces valeurs à la place des modalités dans la formule.



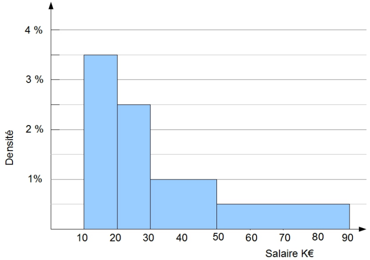
Exemple. Parmi les PME d'une région, on a

CA annuel en Millions d'euros	Micros Entreprises [0,2[	Petites Entreprises [2,10[	Moyennes Entreprises [10,50[
Fréquence	50%	20%	30%
Centre	$\frac{0+2}{2} = 1$	$\frac{2+10}{2} = 6$	$\frac{10+50}{2} = 30$

$\overline{x} = 1 \times 0,5 + 6 \times 0,2 + 30 \times 0,3 = 10,7$
Le CA moyen des entreprises de la région est d'environ 10,7 M€.

Test

Voici la distribution par classes des salaires annuels des habitants d'une région.



Salaire annuel K€	[10,20[	[20,30[	[30,50[	[50,90[
Fréquence	35%	25%	20%	20%
Centre	$\frac{10+20}{2} = 15$	$\frac{20+30}{2} = 25$	$\frac{30+50}{2} = 40$	$\frac{50+90}{2} = 70$

Moyenne $= \overline{x} \simeq$

$$\boxed{15 \times 0,35 + 25 \times 0,25 + 40 \times 0,2 + 70 \times 0,2 = 33,5 \text{ K€}}$$

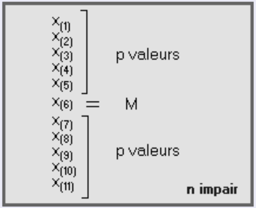
Section 2

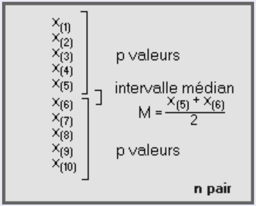
Médiane

Médiane

La **médiane**

M est la valeur du milieu de la série des données, c'est-à-dire la valeur telle qu'il y ait autant d'observations "au-dessous" qu'"au-dessus".





Si la série brute des valeurs observées est triée par ordre croissant $x_1 \leq x_2 \leq \ldots \leq x_n$:

- si n est **impair**, soit $n = 2p + 1$ alors la médiane est $M = x_{p+1}$.
- si n est **pair**, soit $n = 2p$ on choisit en général $M = \frac{x_p + x_{p+1}}{2}$.

Exemple et Test

Immeuble **A**

Famille N°	Nb Enfants
1	0
2	0
3	0
4	0
5	0
6	0
7	1
8	1
9	1
10	1
11	2
12	2
13	2
14	2
15	2
16	3
17	3
18	4

Immeuble **B**

Famille N°	Nb Enfants
1	0
2	0
3	1
4	1
5	1
6	1
7	1
8	1
9	2
10	2
11	2
12	2
13	3
14	3
15	4
16	4

Immeuble **C**

Famille N°	Nb Enfants
1	0
2	0
3	0
4	0
5	0
6	0
7	0
8	0
9	1
10	1
11	1
12	1
13	1
14	1
15	2

Calculer le nombre médian d'enfants par famille dans les trois immeubles.

Immeuble A :

$M_A = \frac{x_9 + x_{10}}{2} = \frac{1+1}{2} = 1$ car

$n = 18 = 2 \times 9$

Immeuble B :

$M_B = \boxed{\frac{x_8 + x_9}{2} = \frac{1+2}{2} = 1,5}$

car $n = 16 = 2 \times 8$

Immeuble C :

$M_C = \boxed{x_8 = 0}$

car $n = 15 = 2 \times 7 + 1$

Médiane à partir des effectifs et fréquences cumulés

La médiane M peut s'obtenir à l'aide des fréquences (ou des effectifs) cumulés :
la médiane est la première modalité dont la fréquence cumulée est supérieure à 50% (ou dont l'effectif cumulé est supérieur à la moitié de l'effectif total).

Exemple. Pour le nombre d'enfants des 18 familles étudiées,

Nombre d'enfants x_i	0	1	2	3	4
Effectifs cumulés N_i	6	10	15	17	18
Fréquences cumulées F_i	33%	55%	83%	94%	100%

On remarque que 55% des familles ont au plus un enfant et que 33% des familles n'ont pas d'enfant, donc la médiane est $M = 1$.
De même "1" est la première valeur dont l'effectif cumulé (10) dépasse la moitié de la moitié effectif total $\frac{18}{2} = 9$

Test

Voici le tableau des effectifs et effectifs cumulés concernant la variable "nombre de salariés" sur une population composée des 107 PME d'une ville.

Nombre de salarié x_i	0	1	2	3	4	5	6	Total
Effectif n_i	13	20	35	19	12	5	3	107
Effectifs cumulées N_i	13	33	68	87	99	104	107	

La médiane est M= $\boxed{2}$
En effet $\frac{107}{2} = 53,5$.
Le premier effectif cumulé $> 53,5$ est 68, et c'est l'effectif cumulé de la modalité 2.

Médiane à partir des classes

Si on dispose des données par classes, on peut obtenir une valeur approchée de la médiane à l'aide du graphique des fréquences cumulées.

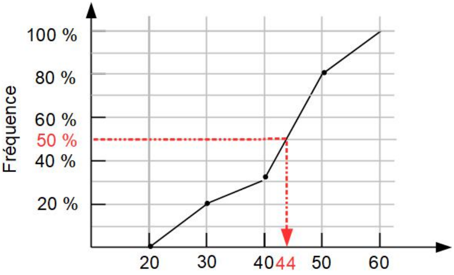
Exemple. Le tableau des fréquences cumulées de l'âge des employés d'une entreprise est :

Âge	[20,30[	[30,40[	[40,50[	[50,60[
Fréquence	20%	10%	50%	20%

Ce qui donne les fréquences cumulées

Âge	30	40	50	60
Fréquences cumulées	20%	30%	80%	100%

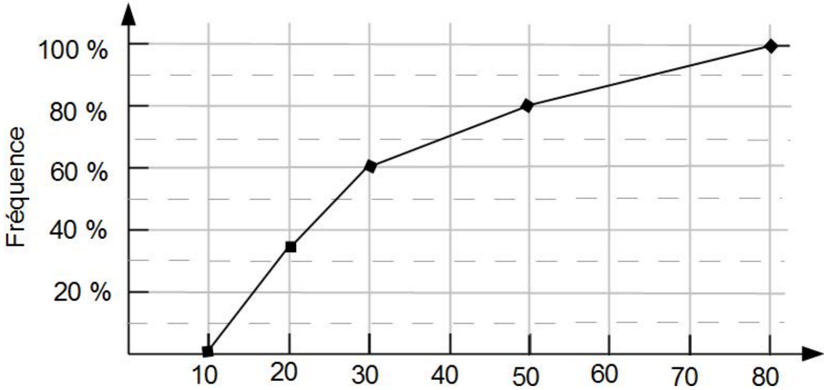
On lit sur le graphique que environs 50% des employés ont moins de 44 ans.
L'âge médian est donc $M \simeq 44$



Test

Voici la distribution par classes des salaires annuels des habitants d'une région et ses fréquences cumulées

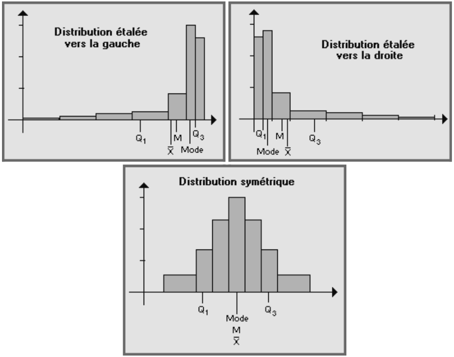
Salaire annuel K€	[10,20[	[20,30[	[30,50[	[50,80[
Fréquences	35%	25%	20%	20%



Le salaire médian est $M \cong$ $\boxed{26 \text{ K€}}$

Médiane vs. Moyenne

Dans le cas d'une distribution symétrique, médiane et moyenne coïncident. Cela n'est pas vrai en général.



Distribution étalée vers la gauche :
$\Rightarrow \overline{x} =$ Moyenne$<M=$médiane

Distribution étalée vers la droite :
$\Rightarrow M=$Médiane $< \overline{x} =$Moyenne

Exemple. La distribution des salaires est étalée vers la droite.
Selon l'INSEE en France en 2013 :
Salaire mensuel médian = 1 772 €< 2 202 €= Salaire mensuel moyen.

Section 3

Fractiles : quartiles, déciles et centiles

Quartiles

Les **quartiles** Q_1, Q_2 et Q_3 sont les trois valeurs qui partagent la série statistique en 4 parties de même effectif.

N°	Note	
1	2	
2	3	
3	4	
4	4	
5	6	Q1
6	6	
7	8	
8	9	
9	10	
10	10	Q2
11	10	
12	11	
13	11	
14	11	
15	11	Q3
16	12	
17	14	
18	14	
19	16	

Q_1 est la donnée de la série qui sépare les 25 % inférieurs des 75% supérieurs

Q_2 est la donnée de la série qui sépare en deux parties de meme effectifs. Q_2 coïncide avec la médiane.

Q_3 est la donnée de la série qui sépare les 75 % inférieurs des 25% supérieurs

Exemple. Pour les notes de la classes de 19 étudiants à gauche, $Q_1 = 6$, $Q_2 = 10$ et $Q_3 = 11$.

Le deuxième quartile Q_2 est la médiane.

Déciles et centiles

Les **déciles** D_1, D_2 .... et D_9 sont les valeurs qui partagent la série statistique en 10 parties de même effectif. Particulièrement significatif sont

D_1 la donnée qui sépare les 10 % inférieurs des 90% supérieurs

D_9 la donnée qui sépare les 90 % inférieurs des 10% supérieurs

Exemple. Selon l'INSEE pour le salaire mensuel en France en 2013 on a $D_1 = 1\ 200$ et $D_9 = 3\ 544$.
C'est-à-dire 10% des français gagnent moins de 1 200 € et 10% plus de 3 544 € par mois.

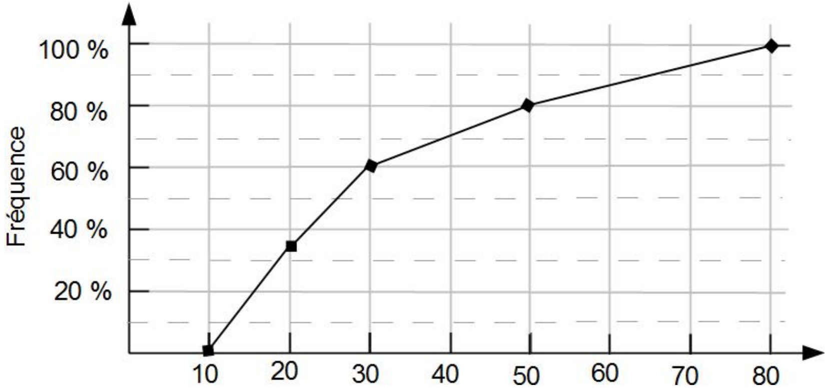
Les **centiles** C_1, ... et C_{99} sont les valeurs qui partagent la série statistique en 100 parties de même effectif.

Exemple. Pour le salaire mensuel en France on a $C_{95} = 4\ 526$ et $C_{99} = 8\ 061$.
C'est-à-dire 5% des français gagnent plus de 4 526 € et 1% gagnent plus de 8 061 € par mois.

Test : à partir du graphique des fréquences cumulés

Voici la distribution par classes des salaires annuel des habitants d'une région et ses fréquences cumulés

Salaire annuel K€	[10,20[	[20,30[	[30,50[	[50,80[
Fréquences	35%	25%	20%	20%



Les quartiles sont $Q_1 \cong$ $\boxed{17 \text{ K€}}$ $Q_2 \cong$ $\boxed{26 \text{ K€}}$ $Q_3 \cong$ $\boxed{45 \text{ K€}}$

Le premier décile est $D_1 \cong$ $\boxed{12 \text{ K€}}$ et le dernier $D_9 \cong$ $\boxed{65 \text{ K€}}$

Test : Fractiles d'une variable discrète

Voici le tableau des effectifs et effectifs cumulés du "Nombre de salariés" sur une population composée des 107 PME d'une ville.

Nb. de salariés x_i	0	1	2	3	4	5	6	Total
Effectif n_i	13	20	35	19	12	5	3	107
Effectifs cumulées N_i	13	33	68	87	99	104	107	

$Q_1 = \boxed{1}$ et $Q_3 = \boxed{3}$

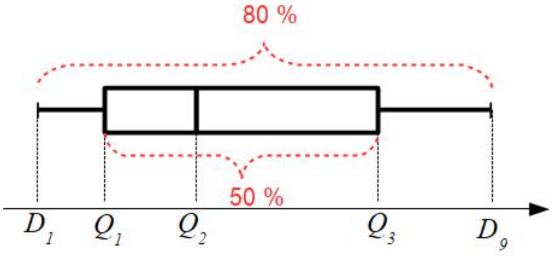
Le premier effectif cumulé$> 107 \times \frac{1}{4} = 26,75$ est celui associé à 1.
Le premier effectif cumulé$> 107 \times \frac{3}{4} = 80,25$ est celui associé à 3.

$D_1 = \boxed{0}$ et $D_9 = \boxed{4}$

Le premier effectif cumulé $> 107 \times \frac{1}{10} = 10,7$ est celui associé à 0.
Le premier effectif cumulé $> 107 \times \frac{9}{10} = 96,3$ est celui associé à 4.

Boite à moustache

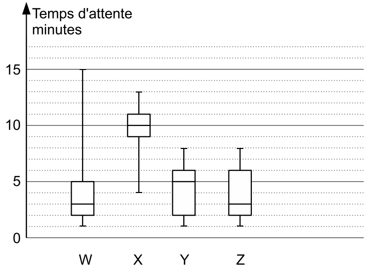
Le **diagramme en boîte** (ou “boîte à moustaches”) permet de représenter les quartiles et déciles d'une distribution.



L'intérêt de ces diagrammes en boîtes est de pouvoir comparer par un moyen visuel plusieurs distributions d'un même caractère sur des populations différentes.

Test

Une association de consommateurs a testé les services d'assistance téléphonique de quatre assurances W, X, Y et Z. Pour chaque appel au service, on a mesuré le temps d'attente du client.



- Pour le temps d'attente au **service clients de W** on a :
  $D_1 = \boxed{1}$ $D_9 = \boxed{15}$
  $Q_1 = \boxed{2}$ $Q_2 = \boxed{3}$ $Q_3 = \boxed{5}$
- 10% des clients de W attendent plus de $\boxed{\text{15 min.}}$ et 50% plus de $\boxed{\text{3 min.}}$
- Le pire des services clients est $\boxed{\text{X}}$.
- Le meilleur est $\boxed{\text{W ou Z}}$.