

Biométrie

Mm. Mebrek N,

Rappels sur certaines notions statistiques

L'analyse des données recueillies est l'un des objectifs de la statistique.

Le vocabulaire de la statistique vient des premières études qui portaient sur la démographie

On effectue généralement des mesures sur les individus qui composent une population.

Ces mesures sont rattachées à des variables ou caractères.

L'ensemble sur lequel porte l'étude s'appelle Population

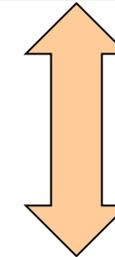
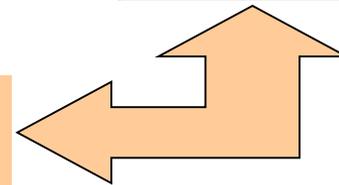
Un élément de cet ensemble est un individu

Répartition des
étudiants selon leur note



[0;4]	20
[4;6]	60
[6;8]	90
[8;10]	100
[10;12]	70
[12;14]	70
[14;16]	40
[16;20]	20

Ici, la variable d'observation (la note)
est découpée en classes



À chaque modalité
correspond un
effectif observé

Deux types de caractères ou variables sont observées

- Les variables continues

Variables de type quantitatif,
On peut calculer une moyenne (loyer, taille, âge...)

- Les variables nominales

Variables de type qualitatif, (profession, statuts matrimoniaux, catégories...)

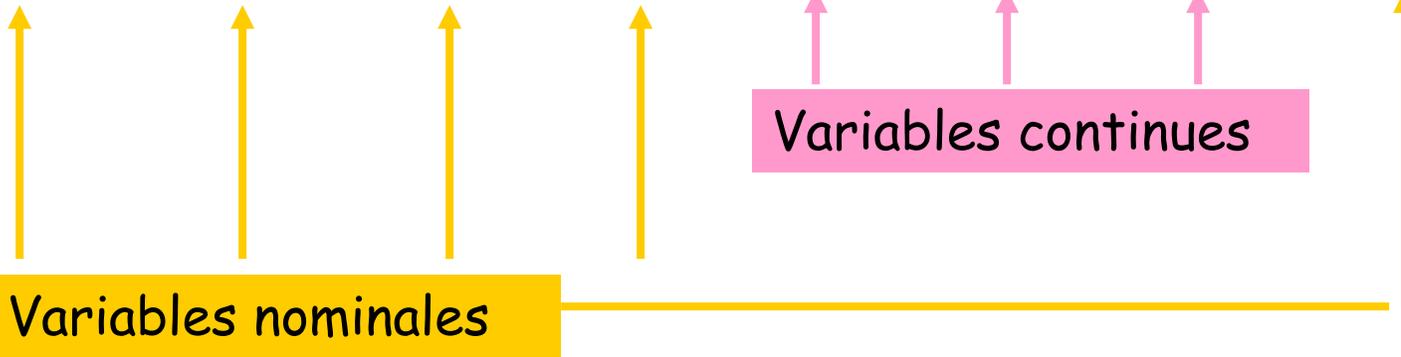
variables

individus

arrdt	pièces	parking	étage	surface	loyer	charges	bailleur
7	1	0	1	37	4100	390	AGF
9	5	0	1	150	13500	1200	AGF
10	3	1	4	78	6120	590	AGF
12	5	1	3	87	7305	958	LOC
14	3	1	1	69	5865	822	SGI
14	3	1	1	68	6644	500	LOC
14	2	1	15	56	5345	788	SOLVEG

Variables nominales

Variables continues



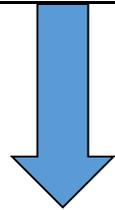
Une série statistique

Variable

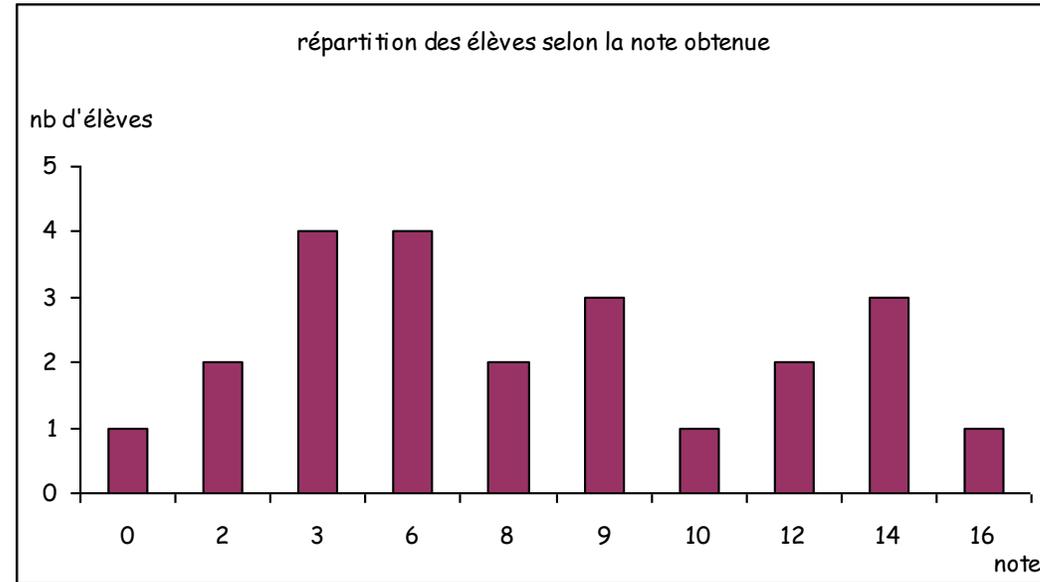
note	effectif
0	1
2	2
3	4
6	4
8	2
9	3
10	1
12	2
14	3
16	1



fréquence $3/23 = 0,131$



Effectif total : 23



Indicateurs statistiques

Nous considérons une population de n individus pour lesquels une variable x est observée.

On notera (x_1) la valeur de x pour l'individu 1,

(x_i) la valeur de x pour l'individu i et

(x_n) la valeur de x pour l'individu n .

x_1	0,
	2,2,
	3,3,3,
	6,6,6,6
32 élèves, 32 notes	8,8,
	9,9,9,9,9,9,
	10,10,10,10,10,
	12,12,12,12,
	14,14,14,
x_{32}	16

note	effectif
0	1
2	2
3	3
6	4
8	2
9	7
10	5
12	4
14	3
16	1

Autre exemple, la population de communes x :

Code	NOM	PSDC99
95 203	EAUBONNE	22870
95 306	HERBLAY	23081
95 277	GONESSE	24721
95 582	SANNOIS	25331
95 607	TAVERNY	25905
95 063	BEZONS	26087
95 680	VILLIERS-LE-BEL	26089
95 280	GOUSSAINVILLE	27224
95 500	PONTOISE	27418
95 219	ERMONT	27542
95 252	FRANCONVILLE	33494
95 268	GARGES-LES-GONESSE	39963
95 127	CERGY	53995
95 585	SARCELLES	57940
95 018	ARGENTEUIL	94019

La médiane :

La médiane est la valeur de la série qui partage la distribution en deux sous-ensembles d'égal effectif.

50% des valeurs sont supérieures à la médiane.

50% des valeurs sont inférieures à la médiane.

Si on range les valeurs selon un ordre croissant, la médiane se trouve au rang

$$i = \frac{n + 1}{2}$$

Si n est impair i est un entier et la valeur médiane existe dans la série statistique.

Par contre si n est pair le rang i tombe entre deux valeurs, la médiane correspondra à la moyenne des 2 valeurs qui encadrent ce rang.

Les valeurs extrêmes et l'étendue d'une série statistique

Code	NOM	PSDC99
95203	EAUBONNE	22870
95306	HERBLAY	23081
95277	GONESSE	24721
95582	SANNOIS	25331
95607	TAVERNY	25905
95063	BEZONS	26087
95680	VILLIERS-LE-BEL	26089
95280	GOUSSAINVILLE	27224
95500	PONTOISE	27418
95219	ERMONT	27542
95252	FRANCONVILLE	33494
95268	GARGES-LES-GONESSE	39963
95127	CERGY	53995
95585	SARCELLES	57940
95018	ARGENTEUIL	94019

minimum

médiane

étendue
(94019-22870)

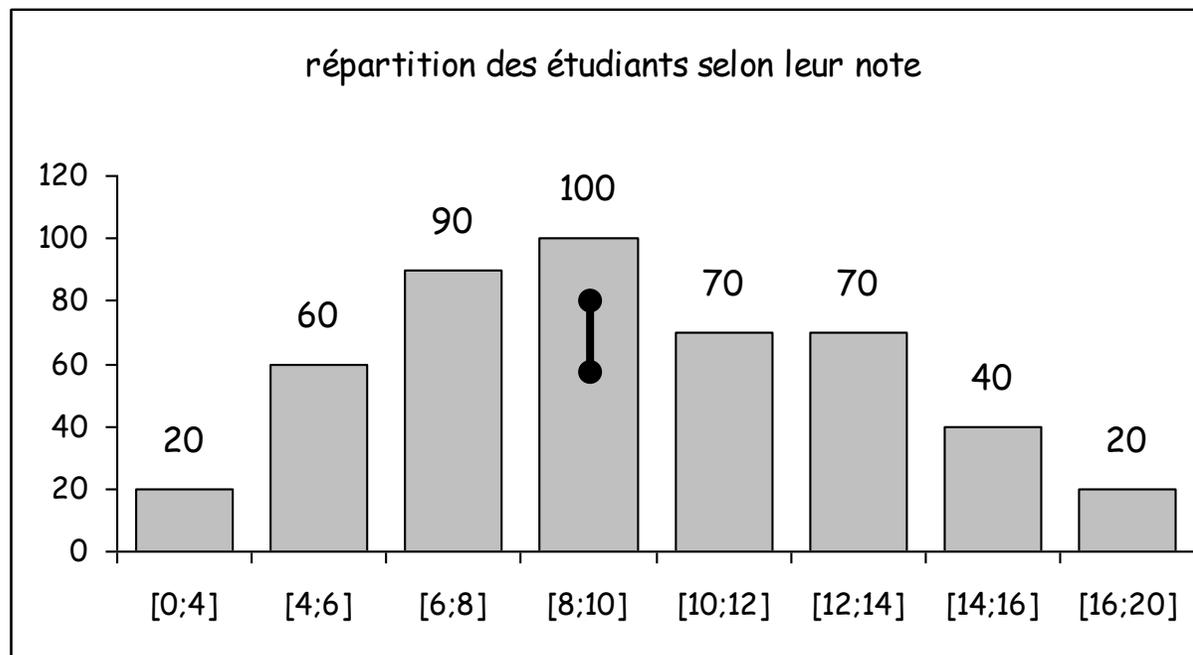
maximum

Les valeurs sont rangées par ordre croissant

Le mode :

C'est la valeur la plus fréquente de la série statistique, c'est à dire celle qui a l'effectif le plus important.

Dans une série en classes, on détermine d'abord la classe modale,



Moyenne :

La moyenne de X est égale à la somme des Xi divisée par l'effectif N.

$$\bar{X} = \frac{1}{N} \sum_{i=1}^N x_i$$

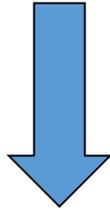


$$\frac{x_1 + x_2 + \dots + x_n}{N}$$

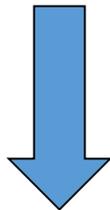
$$\bar{X} = \frac{1}{N} \sum_{i=1}^k n_i x_i$$

Propriété de la moyenne : $\sum_{i=1}^n (x_i - \bar{X}) = 0$

$$\frac{1}{n} \sum_{i=1}^n x_i = \bar{X}$$



$$\sum_{i=1}^n x_i = n\bar{X}$$



$$\sum_{i=1}^n x_i - n\bar{X} = 0$$



4	-7
7	-4
10	-1
13	2
14	3
18	7
11	

66

Propriété de la moyenne :

$$\sum_{i=1}^n (x_i - \bar{X}) = 0$$

La variance de x est égale à la somme des carrés des écarts à la moyenne divisée par l'effectif N.

$$\frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2 = S^2$$

Code	NOM	PSDC99	écart à la moyenne	(écart à la moyenne) ²
95203	EAUBONNE	22 870		
95306	HERBLAY	23 081		
95277	GONESSE	24 721		
95582	SANNOIS	25 331		
95607	TAVERNY	25 905		
95063	BEZONS	26 087		
95680	VILLIERS-LE-BEL	26 089		
95280	GOUSSAINVILLE	27 224		
95500	PONTOISE	27 418		
95219	ERMONT	27 542		
95252	FRANCONVILLE	33 494		
95268	GARGES-LES-GONESSE	39 963		
95127	CERGY	53 995		
95585	SARCELLES	57 940		
95018	ARGENTEUIL	94 019		
	somme	535 679	0,0	0,0
	effectif	15		0
	moyenne	35 711,9		

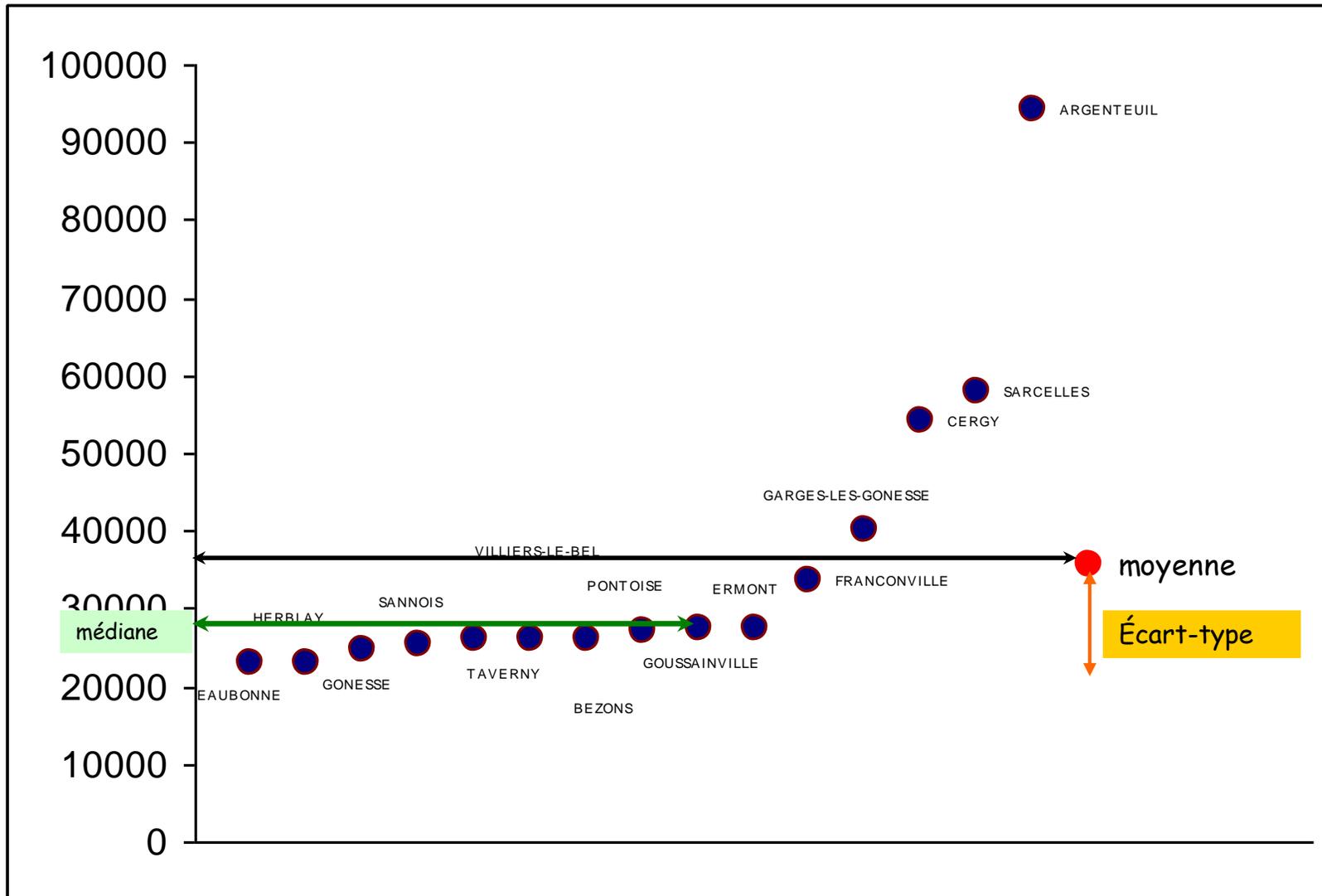
L'écart type d'une variable est égal à la racine carrée de sa variance.

$$\sqrt{\frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2} = S$$

L'écart type s'exprime dans la même unité que les observations

Code	NOM	PSDC99	écart à la moyenne	(écart à la moyenne) ²
95203	EAUBONNE	22 870		
95306	HERBLAY	23 081		
95277	GONESSE	24 721		
95582	SANNOIS	25 331		
95607	TAVERNY	25 905		
95063	BEZONS	26 087		
95680	VILLIERS-LE-BEL	26 089		
95280	GOUSSAINVILLE	27 224		
95500	PONTOISE	27 418		
95219	ERMONT	27 542		
95252	FRANCONVILLE	33 494		
95268	GARGES-LES-GONESSE	39 963		
95127	CERGY	53 995		
95585	SARCELLES	57 940		
95018	ARGENTEUIL	94 019		
	somme	535 679		
	effectif	15		
	moyenne	35 711,9		
				écart-type
				0,0

Moyenne et écart-type d'une série statistique



Code	NOM	PSDC99
95 203	EAUBONNE	22870
95 306	HERBLAY	23081
95 277	GONESSE	24721
95 582	SANNOIS	25331
95 607	TAVERNY	25905
95 063	BEZONS	26087
95 680	VILLIERS-LE-BEL	26089
95 280	GOUSSAINVILLE	27224
95 500	PONTOISE	27418
95 219	ERMONT	27542
95 252	FRANCONVILLE	33494
95 268	GARGES-LES-GONESSE	39963
95 127	CERGY	53995
95 585	SARCELLES	57940
95 018	ARGENTEUIL	94019

Médiane

27 224

Moyenne

35 711,9

Écart-type

18 784,9

coefficient de variation 0,53

écart-type/moyenne

Exercice : Afin de tester le taux de mortalité des nouveaux nés, N femmes ont été questionnées,

pour chacune d'elles on a noté le nombre d'enfants vivants, Les résultats sont dans le tableau suivant:

1	3	0	0	3	2	1	2	3	2
1	1	0	2	1	2	0	3	1	1
2	2	3	2	1	3	2	3	2	0
1	2	4	1	4	4	0	4	2	4
0	2	1	1	0	2	0	1	2	3

1) dresser le tableau des fréquences absolues, fréquences relatives et fréquences cumulées,

2) donner les représentations graphiques des deux fréquences,

3) Calculer la moyenne, la médiane, la variance et l'écart type.

Les quartiles :

Le terme vient de quart. On distingue 3 quartiles Q1, Q2 et Q3. Ce sont des quantiles d'ordre 25%, 50% et 75%. Partageant la série statistique en 4 parties de même taille.

25 % des observations sont inférieures à Q1

50% des observations sont inférieures à Q2

75% des observations sont inférieures à Q3



La médiane est le deuxième quartile ou le quantile d'ordre 50%.

Les déciles

Le terme vient de dix. Les déciles notés D_1, D_2, \dots, D_9 partagent la série en 10 parties d'égal effectif.

-

-

-

Les centiles

Le terme vient de cent. Les paramètres C_1, C_2, \dots, C_{99} (99 centiles) d'ordre 1, 2, ... 99% partagent la série en 100 parties de même taille.

Intervalle interquartile :

C'est un indicateur de dispersion.

Il mesure l'écart entre le 3^{ème} et 1^{er} quartile.

Cet intervalle correspond à 50% des observations.

Cette mesure n'est pas sensible aux valeurs éloignées.

On calcule la longueur de l'intervalle : $Q3-Q1$.

Série en classes

Classes	Effectifs
M_1	n_1
M_2	n_2
M_2	n_2
.	.
.	.
M_i	n_i
.	.
.	.
M_k	n_k

Ce tableau renseigne sur l'effectif ou fréquence absolue pour chacune des classes M_i .

*N est l'effectif total
 k étant le nombre de classes.*

Les n_i représentent les effectifs des différentes classes.

$$N = \sum_{i=1}^k n_i$$

sous Excel deux méthodes sont possibles pour découper une série en classes :

- Utiliser la fonction fréquence*
- Le tableau croisé dynamique et la fonction grouper.*