Chapitre 3 : Statistique

Un peu de culture : En mathématiques, la notion de statistique est utilisée depuis l'Antiquité afin de faire des recensements de la population. Aujourd'hui elle joue un rôle important dans l'interprétation des résultats, et rend notamment lisible les valeurs d'un caractère, grâce à des calculs de paramètres.
Le premier recensement « moderne » au niveau national aurait ainsi été celui ordonné en 1694 par Louis Phélypeaux, comte de Pontchartrain. Celui-ci sera suivi par divers recensements, dénombrements et enquêtes nationales conduits à intervalles irréguliers.
Le recensement de population de 1801 préparé par Lucien Bonaparte et Jean-Antoine Chaptal a été le point de départ d'une série de recensements effectués –avec plus ou moins de régularité - tous les cinq ans jusqu'en 1946. Depuis, les recensements ont été organisés par l'Institut national de la statistique et des études économiques (Insee) jusqu'en 1999, et sous une forme rénovée depuis 2004.
En 2013 en France, près de 350 articles de lois ou de codes se réfèrent au recensement, dont pour l'organisation des élections municipales, la répartition de la dotation globale de fonctionnement, la répartition des services de santé, certaines politique de prévention et gestion des risques. (#Wikipédia)

I. Pour prendre un bon départ : maîtriser le vocabulaire

Les statistiques utilisent un vocabulaire précis qu'il faut connaître. Les définitions suivantes déjà abordées en collège sont donc à connaître.

1. Population / Individu :

On appelle **population** tout ensemble soumis à une étude statistique. Un **individu** est alors un élément de cet ensemble.
Le nombre d'individus constitue la **taille** de la population.

Remarque : *Si la population est de taille très importante, on peut ne faire porter l'étude statistique que sur une partie de la population ; dans ce cas on dit qu'on prend un* ***échantillon*** *de la population et on essaye, autant que faire se peut, que l'échantillon soit* ***représentatif*** *de la population. Dans le cas contraire, l'étude n'aurait pas d'intérêt.*

2. Caractère / Modalité :

L'aspect ou le trait sur lequel porte l'étude statistique est appelé **caractère** ou **variable statistique**.
Les différents traits observés sur ce caractère portent le nom de **modalités**.
Les modalités sont souvent notées x_i où i désigne un entier naturel. Par exemple, x_1 , x_2 , x_3 ,…

3. Quantitatif / Qualitatif :

Un caractère est dit **quantitatif** si les modalités sont des valeurs numériques (mesures physiques, physiologiques, sociologiques, démographiques, économiques…).
Dans le cas contraire, c'est-à-dire lorsque les valeurs ne peuvent être ni ordonnées ni ajoutées, le caractère est dit **qualitatif**.

Exemple 1: la couleur des yeux ou la taille d'un vêtement est un caractère qualitatif, tandis que la taille en cm ou l'âge en années d'une population sont des caractères quantitatifs. Les différentes modalités du caractère « couleur » peuvent être par exemple : marron, bleu, vert ; les différentes modalités de la taille d'un vêtement peuvent être S, M, L, XL.

4. Discret / Continu :

Un caractère **quantitatif** est dit **discret** lorsqu'il ne prend qu'un nombre fini de valeurs numériques, isolées.
Dans le cas contraire, le caractère est quantitatif **continu**, il prend une infinité de valeurs.

Exemple 2: la taille en cm est un caractère quantitatif continu tandis que l'âge en années est un caractère quantitatif discret.

5. Classe / Centre :

Lorsque le caractère est **quantitatif et continu,** les modalités, *c'est-à-dire les différentes valeurs prises par le caractère* peuvent être regroupées en intervalles appelés **classes**. La valeur centrale de la classe est alors appelée **centre de la classe**.
Le centre de la classe est obtenu en effectuant la moyenne des bornes de la classe.

Exemple 3 : Dans un hôpital, on mène une enquête sur la taille des nourrissons à la naissance. Les tailles étant nombreuses, on les regroupe par classes. Les crochets qui indique des intervalles de tailles sont importants afin qu'une taille ne soit pas comptée dans deux classes.

Masses (kg)	[2,5 ; 3[	[3 ; 3,5[	[3,5 ; 4[	[4 ; 4,5[	TOTAL
Effectifs	17	8	5	12	

Dans cette série statistique, le caractère étudié est la masse du nourrisson à la naissance. Les modalités sont les différentes classes de masse.
Les centres de classe sont respectivement : $(2,5+3)/2=2,75$; $(3+3,5)/2=3,25$; $(3,5+4)/2=3,75$; $(4+4,5)/2=4,25$

6. Effectif / Effectif total / Série des effectifs :

L'**effectif** d'une modalité est le nombre d'individus possédant cette valeur du caractère.
L'**effectif total** est le nombre d'individus de la population. C'est la somme des effectifs de chaque modalité.
La **série statistique des effectifs** ou **distribution des effectifs** est la fonction qui à chaque valeur du caractère (modalité) associe l'effectif de cette modalité. Elle est définie le plus souvent à l'aide d'un tableau de valeurs.

Exemple 3: Dans l'exemple ci-dessous (exemple 4), l'effectif de la modalité « 3 enfants » est $n = 26$.
L'effectif total de la série est $N = 48 + 62 + 35 + 26 + 15 + 9 + 5 = 200$

7. Fréquence / Pourcentage

La **fréquence** f_i d'une **modalité** x_i est le **nombre compris entre 0 et 1** obtenu en divisant son effectif n_i par l'effectif total N.

$$f_i = \frac{n_i}{N}$$

Remarques : On peut obtenir le pourcentage correspondant en multipliant la fréquence par 100.

- ✓ **La somme des fréquences est égale à 1.**
- ✓ **La somme des pourcentages est égale à 100.**

Exemple 4 : Enquête sur le nombre d'enfants par famille dans un village comptant 200 famille.

La **population** est constituée des 200 familles, la taille de la population est **N=200**.
Le **caractère étudié** est le nombre d'enfants : c'est un **caractère quantitatif**.

Si au terme de l'étude on obtient des valeurs 0, 1, 2, 3, 4, 5 ou 6 le caractère possède ces **7 modalités**.

On peut résumer ces informations dans le tableau de suivant qui constitue **la série statistique des effectifs.**

Modalités	0	1	2	3	4	5	6	TOTAL
Effectifs	48	62	35	26	15	9	5	
FREQUENCES								

Exemple 5 : Enquête sur la couleur des yeux dans une classe de 40 élèves d'un lycée.

La **population** est constituée par les 40 élèves de la classe, sa taille est **N = 40**.
Le **caractère étudié** est la couleur des yeux, c'est un **caractère qualitatif**. Si au terme de l'étude on obtient des valeurs bleu, marron, noir, ou vert, le caractère possède ces **4 modalités**.

On peut résumer ces informations dans le tableau de suivant qui constitue **la série statistique des effectifs.**

Modalités	BLEU	MARRON	VERT	NOIR	TOTAL
Effectifs	15	12	8	5	
FREQUENCES					
FREQUENCES en %					

II. Effectifs et fréquences cumulé(e)s

1. Définitions

On note x_i la $i_{ème}$ modalité prise par un caractère quantitatif.

L'effectif (resp. fréquence) cumulé(e) croissant(e) ECC, (resp. FCC) de x_i est la somme des effectifs (resp. fréquences) des valeurs inférieures ou égales à x_i.

L'effectif (resp. fréquence) cumulé(e) décroissant(e) ECD, (resp. FCD) de x_i est la somme des effectifs (fréquences) des valeurs supérieures ou égales à x_i.

2. Exemples

Exemple 4 : enquête sur le nombre d'enfants par famille dans un village

Modalités	0	1	2	3	4	5	6	TOTAL
Effectifs	48	62	35	26	15	9	5	
ECC								

Point méthode : On calcule les effectifs cumulés croissants (ECC) de la …… …. vers la ……

On peut lire par exemple que …….. familles ont ……..…..2 enfants (c'est-à-dire soit 0, 1 ou 2 enfants).

Exemple 3 : enquête sur la masse des nouveaux nés dans un hôpital.

Masses (kg)	[2,5 ; 3[	[3 ; 3,5[	[3,5 ; 4[	[4 ; 4,5[	TOTAL
Effectifs	17	8	5	12	
Fréquences					
FCC					

Point méthode : les données sont regroupées en intervalles semi-fermés appelés classes.

On calcule les fréquences cumulées décroissantes de la …… vers la …….

On peut lire par exemple que …… des nouveaux nés ont une masse …….. à ……..

III. Représentations graphiques d'une série statistique

Le choix de la représentation graphique dépend de la série statistique et du type d'interprétation souhaité.

1. Caractère qualitatif

Pour un caractère qualitatif, on peut faire :
- un diagramme à bâtons (ou en barres) ;
- un diagramme circulaire.

Exemple 5 : enquête sur la couleur des yeux dans un groupe de 40 élèves d'un lycée

Modalités	BLEU	MARRON	VERT	NOIR	TOTAL
Effectifs	15	12	8	5	
Angle en degrés					

Point méthode : l'angle au centre est proportionnel à l'effectif de la modalité. On peut donc le calculer par les techniques usuelles sur les tableaux de proportionnalité (addition ou soustraction, multiplication ou division, ou encore par produit en croix).

Diagramme circulaire : *Diagramme en barres :*

Laisser l'espace nécessaire

2. Caractère quantitatif

Pour un caractère quantitatif, on peut utiliser les représentations graphiques précédentes ou encore faire un nuage de points.

Point méthode : polygône des effectifs cumulés

On reprend l'exemple 3 :

Avec les effectifs cumulés croissants des masses des des nouveaux nés, on obtient le polygône suivant :

Avec les fréquences cumulées décroissantes des masses des nouveaux nés, on obtient le diagramme suivant :

Laisser l'espace nécessaire

IV. Médiane ; quartiles ; déciles

On ne s'intéresse à présent qu'à des séries statistiques à caractère quantitatif.

1. Définition : médiane

Dans une série statistique **ordonnée**, la médiane partage l'ensemble des valeurs prises en deux groupes de même effectif.

Point méthode : Si les valeurs sont peu nombreuses, il est simple d'ordonner la série et de séparer les valeurs prises par le caractère en deux groupes de même effectif.

Si l'effectif total est impair : une valeur restera entre les deux demi-groupes.
Cette valeur est la médiane.

Exemple 6: Soit la série suivante :
0 – 11 – 3 – 9 – 3 – 18 – 12 – 3 – 20 – 14 – 4 – 3 – 7 – 14 – 7 – 7 – 9– 13 – 14 - 15 –18

Ordonnons la série :

0 – 3 – 3 – 3 – 3 – 4 – 7 – 7 – 7 – 9 – 9 – 11 – 12 – 13 – 14 – 14 – 14 - 15 – 18 – 18 - 20

N = 21, on peut donc partager la série en 10 valeurs basses et 10 valeurs hausses. De part et d'autre de ces 2 séries, on trouve la $11^{ème}$ valeur qui partage alors la série en 2 séries de même effectif.

Me = x_{11} = 9

Si l'effectif total est pair : aucune valeur de la série ne sépare la série en deux groupes de même effectif. Par convention, on considère que la médiane est la moyenne de la dernière valeur du premier groupe et de la première valeur du deuxième groupe.

Exemple 7 : Soit la série ordonnée suivante

0 – 2 – 3 – 3 – 3 – 4 – 7 – 7 – 8 – 9 – 10 – 11 – 12 – 13 – 14 – 14 – 15 – 18 – 19 - 20

N = 20, on peut donc partager la série en 10 valeurs basses et 10 valeurs hausses. Aucune valeur de la série ne partage donc la série en deux séries de même effectif.
Par convention, Me= $(x_{10} + x_{11}) / 2 = (9 + 10) / 2 = 9{,}5$

2. Définition : quartiles

Le **premier quartile** d'une série statistique, noté **Q_1** est la **plus petite valeur** prise par le caractère telle **qu'au moins un quart** soit **au moins 25%** des valeurs lui sont **inférieures ou égales**.

Le **troisième quartile** d'une série statistique, noté **Q_3** est la **plus petite valeur** prise par le caractère telle **qu'au moins trois quarts** soit **au moins 75%** des valeurs lui sont **inférieures ou égales**.

Vocabulaire :

- la différence **$Q_3 - Q_1$** est appelée **écart interquartiles**.
- l' intervalle **[$Q_3 - Q_1$]** est appelé **intervalle interquartiles** : il rassemble environ 50% de l'effectif. Plus l'écart interquartile est grand, plus les valeurs de la série sont **dispersées** autour de la médiane.

Exemple 8 : Soit la série ordonnée suivante

0 – 3 – 3 – 3 – 3 – 4 – 7 – 7 – 7 – 9 – 9 – 11 – 12 – 13 – 14 – 14 – 14 - 15 – 18 – 18 - 20

L'effectif de la série est N = 21. La médiane est $x_{11} = 9$.

Déterminons Q1 et Q3. Pour cela calculons la taille correspondant à un quart de l'effectif : N/4 = 5,25 donc $Q_1 = x_6 = 4$
De même 3N/4 = 15,75 donc $Q_3 = x_{16} = 14$.

L'écart interquartiles est donc $Q_3 - Q_1 = 14 - 4 = 10$

Interprétation : au moins 25% des valeurs de la série sont inférieures ou égales à 4

au moins 50% des valeurs de la série sont inférieures ou égales à 9

au moins 75% des valeurs de la série sont inférieures ou égales à 14

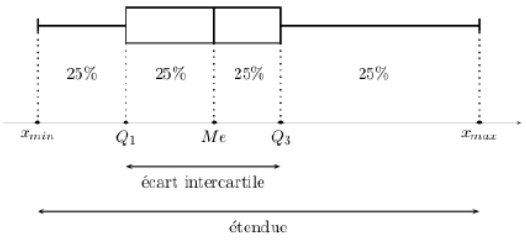
3. Diagramme en boîte :

On peut résumer une série statistique par un diagramme gradué faisant apparaître les indicateurs :

Min, *Max*, Q_1 ; Q_3 et *Me*.

Un tel diagramme est appelé **diagramme en boîte** ou **boîte à moustache**.

Remarque : on peut facilement comparer plusieurs séries statistiques en superposant leurs diagrammes en boîte.



Application : exercice n° 44 page 128

Exercices supplémentaires :

Exemple 9 :

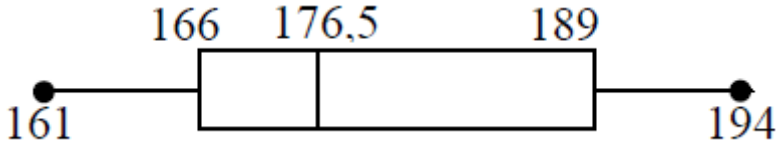
Les élèves de 2nde 2 du Lycée S. Weil ont reçu une invitation pour assister aux enregistrements de l'émission de télévision The Voice.
Les spectateurs sont placés sur deux rangs suivant leur taille.
Aidez les organisateurs en déterminant la taille médiane, puis donnez le diagramme en boîte.

172; 162; 190; 190; 169; 164; 177; 181; 189; 161; 164; 182; 185; 188; 169; 180; 193; 189; 179; 180; 173; 193; 166; 164; 163; 164; 190; 176; 176; 192; 173; 194; 165; 172.

Correction :

Voici la liste ordonnée des tailles des élèves de 2nde 2 :
161 ;162 ;163 ;164 ;164 ;164 ;165 ;166 ;169 ;169 ;172 ;172 ;173 ;173 ; 176 ;176 ;177 ;179 ;180 ;180 ;181 ;182 ;185 ;188 ;189 ;189 ;190 ;190 ; 190 ;192 ;193 ;193 ;194.
L'effectif total de cette série est $N = 34$.
Une fois la série ordonnée, les 17ème et 18ème valeurs partagent cette série en deux groupes de 17 élèves.
Leur moyenne est $\frac{x_{17}+x_{18}}{2} = \frac{176+177}{2} = 176{,}5$
Une valeur possible de la médiane est $Me = 176{,}5$

$\frac{N}{4} = 8{,}5$ donc $Q_1 = x_9 = 166$: au moins 25% des élèves de 2nde 2 ont une taille inférieure ou égale à 1,66m.
$\frac{3N}{4} = 25{,}5$ donc $Q_3 = x_{26} = 189$: au moins 75% des élèves de 2nde2 ont une taille inférieure ou égale à 1,89m.



Point méthode : déterminer les quartiles avec les fréquences

Exemple 10 :

Le tableau ci-dessous indique la répartition du nombre d'enfants de moins de 25 ans dans les familles du secteur est parisien.
Donner la distribution des fréquences et en déduire la médiane et les quartiles. Interpréter par une phrase chacun de ces paramètres statistiques.

Nombre de familles avec	en 2009	en 1999
Aucun enfant	244 918	220 109
1 enfant	131 271	124 597
2 enfants	109 776	102 135
3 enfants	35 907	35 708
4 enfants ou plus	13 311	14 564
Ensemble	535 183	497 113

Source : Insee, RP1999 et RP2009 exploitations complémentaires.

Correction :

Après avoir donné la distribution des fréquences on travaille avec la distribution des fréquences cumulées croissantes.

Fréquences en % des familles avec	en 2009	cumuls
Aucun enfant	45,76	45,76
1 enfant	24,53	70,29
2 enfants	20,51	90,80
3 enfants	6,71	97,51
4 enfants ou plus	2,49	100
Ensemble	100	

- Les 25% de la population sont atteints pour les familles sans enfant. Donc le premier quartile est 0 enfant.
- Les 50% sont atteints pour les familles avec un enfant ou moins. Donc la médiane est 1 enfant.
- Les 75% sont atteints pour les familles avec deux enfants ou moins. Donc le 3ème quartile est 2 enfants.

Remarque : *pour une série dont les données sont regroupées en classe, les valeurs brutes prises par le caractère ne sont pas accessibles. Il est possible d'obtenir une approximation de la médiane et des quartiles par lecture graphique sur le polygone des effectifs (ou fréquences) cumulées.*

Exemple 11 :

Déterminer par lecture graphique la médiane et les quartiles de la série constituée par les tailles des exploitations agricoles professionnelles en Franche-Comté.

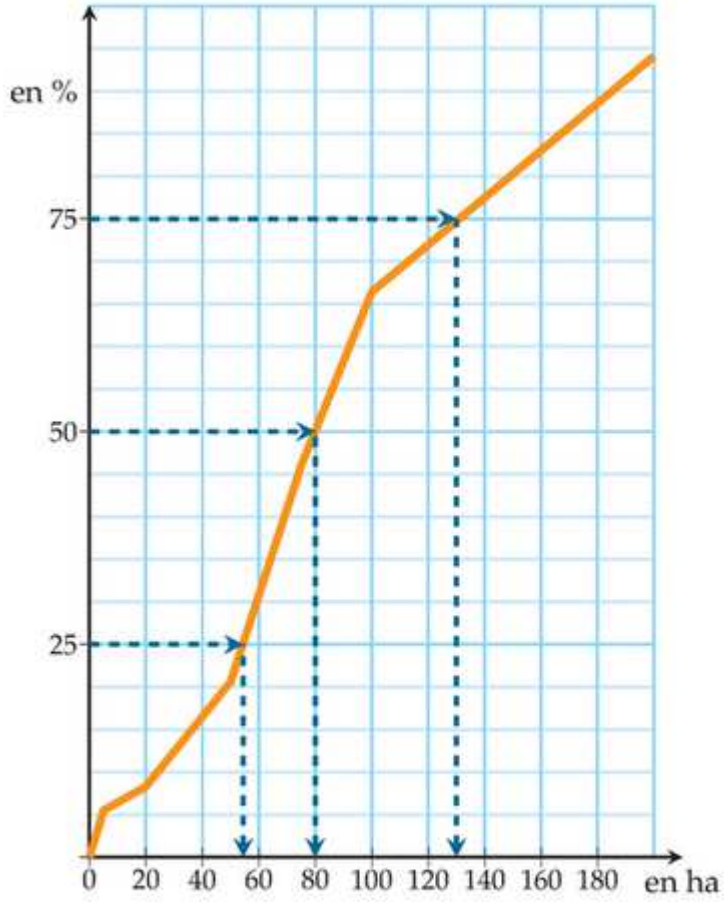
	Effectifs
Moins de 5 ha	370
De 5 à moins de 20 ha	190
De 20 à moins de 50 ha	840
De 50 à moins de 75 ha	1720
De 75 à moins de 100 ha	1380
De 100 à moins de 200 ha	1880
200 ha et plus	400
Ensemble	**6780**

Source Insee : Enquête structure des exploitations 2005

	F.C.C. en %
Moins de 5 ha	5,5
De 5 à moins de 20 ha	8,3
De 20 à moins de 50 ha	20,6
De 50 à moins de 75 ha	46
De 75 à moins de 100 ha	66,4
De 100 à moins de 200 ha	94,1
200 ha et plus	100

Correction :

Polygone des fréquences cumulées croissantes des tailles des exploitations agricoles de la région Franche-Comté en 2006.



Par lecture graphique, on lit que :

- le 1er quartile est 55 ha ;
- une médiane est 80 ha ;
- le 3e quartile est 130 ha.

V. Moyenne

1. Définition :

La moyenne d'une série statistique quantitative est souvent notée $\bar{x}$.

Avec les notations précédemment introduites, on a :

$$\bar{x} = \frac{somme\ totale\ des\ valeurs\ prises\ par\ le\ caractère}{nombre\ de\ valeurs} = \frac{n_1x_1 + n_2x_2 + ... + n_px_p}{N} = f_1x_1 + f_2x_2 + ... + f_px_p$$

avec $N = n_1 + n_2 + ... + n_p$

2. Calculs de moyennes : exemples et méthodes.

MÉTHODE **Calculer une moyenne à partir des fréquences**

Exercice d'application Calculer le salaire net annuel moyen en France en 2005.

Régions	Fréquences (en %)	Salaires (en euros)
Région Parisienne	25,3	29 237
Bassin Parisien	15,7	20 318
Nord	5,8	20 501
Est	8	20 946
Ouest	12,1	19 891
Sud-Ouest	9,3	20 542
Centre-Est	11,9	25 811
Méditerranée	10	20 993
DOM	1,8	20 495

Correction $\overline{x} = f_1 \times x_1 + f_2 \times x_2 + f_3 \times x_3 + \cdots + f_p \times x_p$

Ici, après avoir exprimé les fréquences sous forme décimale :

$$\begin{aligned}\overline{x} &= 0,253 \times 29\,237 + 0,157 \times 20\,318 + 0,058 \times 20\,501 + 0,08 \times 20\,946 + 0,121 \times 19\,891 \\ &\quad + 0,093 \times 20\,542 + 0,119 \times 25\,811 + 0,1 \times 20\,993 + 0,018 \times 20\,495 = 23\,308,6\end{aligned}$$

Le salaire annuel net moyen en France en 2005 était d'environ 23 308 €.

REMARQUE : Pour une **série triée en classes**, la répartition à l'intérieur d'une classe est souvent considérée comme **homogène.** La valeur prise par le caractère est supposée unique et égale au **centre de la classe**. Le centre c de la classe $[a;b[$ vaut : $c = (a+b) \div 2$.

Exemple

Déterminer l'âge moyen d'un demandeur d'emploi dans les Bouches-du-Rhône en 2009.

Tranche d'âge	Nombre de demandeurs
$[15;25[$	24 146
$[25;55[$	107 761
$[55;65]$	29 441

Correction

- Le centre de la classe $[15;25[$ est 20 ;
- celui de la classe $[25;55[$ est 40 ;
- celui de la classe $[55;65[$ est 60.

$$\overline{x} = \frac{24146 \times 20 + 107761 \times 40 + 29441 \times 60}{24146 + 107761 + 29441}$$

soit $\overline{x} \approx 40,66$

L'âge moyen d'un demandeur d'emploi dans les Bouches-du-Rhône en 2009 était d'environ 41 ans.