1 Statistiques à une variable

1. Rappels de vocabulaire et notations

L'ensemble sur lequel porte l'étude est la **population**, chaque élément de cet ensemble est un **individu statistique** ou une **unité statistique**.

Ce qu'on étudie est un **caractère**.
Un caractère est dit **qualitatif** lorsqu'il n'est pas mesurable (couleur, lieu de naissance,...).
Un caractère est dit **quantitatif** lorsqu'il s'agit d'une grandeur mesurable (prix, poids, longueurs, etc.).
Nous ne nous intéresserons dans ce chapitre qu'à des caractères quantitatifs.
Un caractère quantitatif peut être **discret** (discontinu) ou **continu**.
Il est discret lorsque le caractère prend un nombre fini de valeurs, il est continu dans le cas contraire, on regroupe alors les valeurs du caractère en classes (il s'agit en général d'intervalles fermés à gauche et ouverts à droite).
En pratique la plupart des séries statistiques, même à caractère discret, auront les valeurs de leur caractère regroupées en classes.
Conséquences du regroupement en classes.
Ce regroupement permettra de faciliter les calculs des paramètres de la série statistique, mais ces paramètres perdront une grande part de leur précision.

On notera généralement X le caractère étudié et $\boldsymbol{x_i}$ les valeurs ou **modalités** du caractère (il s'agira des centres de classes dans le cas d'un caractère continu) et $\boldsymbol{n_i}$ les nombres d'individus correspondants. Les entiers $\boldsymbol{n_i}$ sont appelés **effectifs**.

Les réels $f_i = \dfrac{n_i}{\sum n_i}$ sont les **fréquences**, celles-ci peuvent être exprimées en pourcentages.

Il est fréquent de noter N l'effectif total de la population étudiée, on a alors $f_i = \dfrac{n_i}{N}$.

Exemple

Une unité de production fabrique des axes de moteurs électriques, une étude statistique portant sur leurs longueurs en millimètres a donné les résultats suivants :

Longueurs des axes	[89,7 ; 89,8[	[89,8 ; 89,9[	[89,9 ; 90[	[90 ; 90,1[	[90,1 ; 90,2[	[90,2 ; 90,3[
Nombres d'axe (n_i)	3	14	36	33	13	1
Centres de classe (x_i)	89,75	89,85	89,95	90,05	90,15	90,25
Fréquences (f_i)	3/100 ou 3 %	14 %	36 %	33 %	13 %	1 %

Dans cet exemple $N = 100$.
Pour chacune des classes $[\alpha_i, \alpha_{i+1}[$ on peut calculer :

- $n_1 + n_2 + \ldots + n_i$, il s'agit des **effectifs cumulés croissants** ou,
- $n_i + n_{i+1} + \ldots + n_p$ (s'il y a p classes), il s'agit des **effectifs cumulés décroissants**.

On peut faire le même travail avec les fréquences, il s'agit alors des **fréquences cumulées croissantes** et **décroissantes**. Nous allons voir leur utilisation dans l'exemple suivant.

Exemple

Reprenons l'exemple précédent :

Longueurs des axes	[89,7 ; 89,8[	[89,8 ; 89,9[	[89,9 ; 90[	[90 ; 90,1[	[90,1 ; 90,2[	[90,2 ; 90,3[
x_i	89,75	89,85	89,95	90,05	90,15	90,25
n_i	3	14	36	33	13	1
f_i	0,03	0,14	0,36	0,33	0,13	0,01
Effectifs cumulés croissants	3	17	53	86	99	100
Effectifs cumulés décroissants	100	97	83	47	14	1

Les effectifs cumulés permettent de répondre à des questions du type :

Quel est le nombre d'axes de longueur au plus égale à 90 mm ?
Réponse 53.

Quel est le nombre d'axes de longueur supérieure à 89,9 mm ?
Réponse 83.

Représentations graphiques des séries statistiques à caractères continus

• Histogramme

On représente pour chaque classe un rectangle dont l'**aire est proportionnelle à son effectif**.
Si les classes ont toutes la même amplitude comme dans l'exemple précédent, il suffit de représenter des rectangles dont les hauteurs sont proportionnelles aux effectifs.

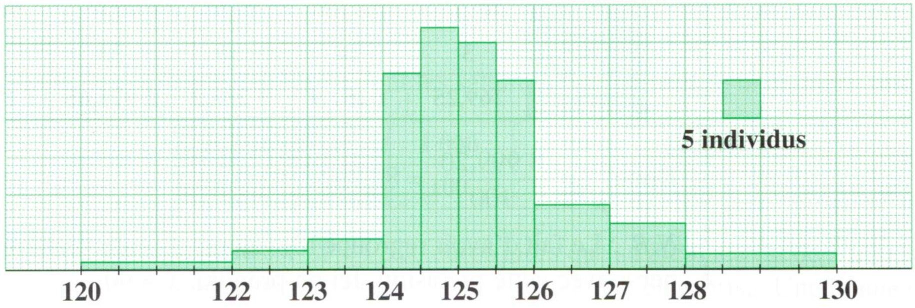
Si les classes n'ont pas la même amplitude, il faut revenir à une amplitude de « base » comme dans l'exemple ci-dessous.

Exemple

On considère la série statistique :

Classes	[120, 122[	[122, 123[	[123, 124[	[124 ; 124,5[	[124,5 ; 125[	[125 ; 125,5[	[125,5 ; 126[	[126, 127[	[127, 128[	[128, 130[
Effectifs	4	5	8	26	32	30	25	17	12	8

On constate que les amplitudes des classes sont 0,5 ; 1 et 2. On peut considérer une amplitude de base égale à 0,5 , les classes d'amplitudes 1 étant considérées comme deux classes d'amplitude 0,5 et celles d'amplitude 2 comme quatre classes de base. On peut alors représenter l'histogramme.



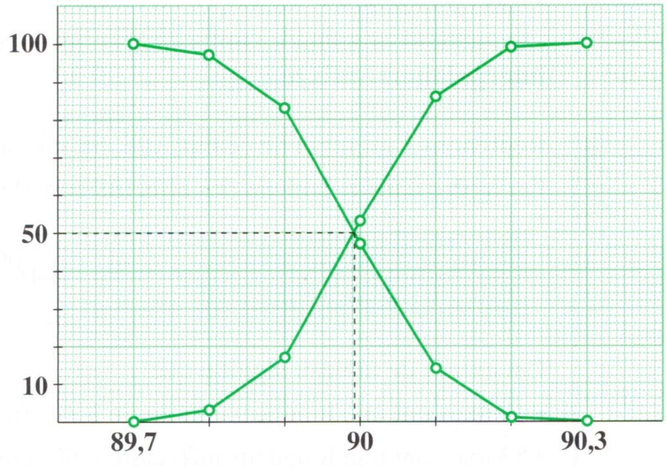
• **Polygones des effectifs cumulés**

On représente le polygone des effectifs cumulés croissants en prenant les bornes droites des intervalles et le polygone des effectifs cumulés décroissants en considérant les bornes de gauche.

(Le polygone des effectifs cumulés croissants « part » toujours de 0 et « arrive » à N pour les ordonnées alors que le polygone des effectifs cumulés décroissants est une ligne brisée allant de N à 0 pour les ordonnées.)

Exemple

On reprend l'exemple sur les longueurs d'axes des moteurs.



L'ordonnée du point de concours des deux polygones est égale à la moitié de l'effectif. Son abscisse est appelée médiane de la série statistique (on retrouvera cette notion dans le paragraphe suivant).

2. Caractéristiques de position

a) La moyenne

C'est de loin la caractéristique la plus utilisée, on la calcule grâce à la formule suivante :

À savoir

$$\overline{x} = \frac{\sum_{i=1}^{p} n_i x_i}{N} = \frac{n_1 x_1 + n_2 x_2 + \dots + n_p x_p}{N}$$

(les x_i sont les valeurs du caractère, chaque x_i correspondant à un effectif n_i).

Exemple

Pour la série concernant les axes, on peut faire le tableau ci-dessous.

x_i	89,75	89,85	89,95	90,05	90,15	90,25	Totaux
n_i	3	14	36	33	13	1	100
$n_i x_i$	269,25	1 257,9	3 238,2	2 971,65	1 171,95	90,25	8 999,20

Donc $\overline{x} = \dfrac{8\,999{,}20}{100} = 89{,}992$.

N.B. : Le fait d'avoir regroupé les valeurs du caractère en classes ne permet pas de donner $\overline{x}$ avec cette précision. Ici on prendra $\overline{x} \approx 90$.

b) La médiane

À savoir

La médiane est la valeur du caractère qui sépare l'effectif en deux parties égales.

Mode de calcul de la médiane

- Cas où la série est discrète

- La série a un effectif impair : $2p + 1$: la médiane est la $(p + 1)^e$ valeur de la variable ;

- La série a un effectif pair : $2p$: la médiane est la moyenne de la p^e et de la $(p + 1)^e$ valeur du caractère.

- Cas où la série est regroupée en classes

- On obtient sa valeur en utilisant les effectifs cumulés croissants et en procédant à une **interpolation linéaire**.

Exemple

Reprenons l'exemple de ce chapitre, on sait que la moitié de l'effectif est 50.

On repère dans le tableau des effectifs cumulés croissants la partie qui nous intéresse.

Classes	…	[89,8 ; 89,9[	[89,9 ; 90[	…
Effectifs cumulés croissants	…	17	53	…

Cet extrait de tableau signifie que :

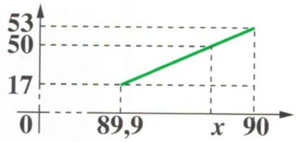
- 17 axes ont une longueur inférieure à 89,9 mm et
- 53 axes ont une longueur inférieure à 90 mm.

Il s'agit de trouver à quelle longueur théorique correspondrait le 50^e axe.

Si l'on fait l'hypothèse que les longueurs des axes sont uniformément reparties dans la classe [89,9 ; 90[, on peut alors procéder à une **interpolation linéaire**.

On peut faire le tableau suivant :

36	33	17	89,9	$x - 89{,}9$	0,1
		50	x		
		53	90		



On obtient $\dfrac{x - 89{,}9}{0{,}1} = \dfrac{33}{36}$, d'où $x \approx 89{,}99$ mm.

On note généralement Me la médiane et on prendra $Me \approx 90$ mm.

N.B. : La valeur de la médiane est l'abscisse du point d'intersection des polygones des effectifs cumulés croissants et décroissants lorsque la série est regroupée en classes.

Utilisation de la médiane

On ne peut se contenter dans un calcul de caractéristiques d'une série statistique de sa moyenne.

Supposons le cas d'une entreprise dont les salaires sont repartis de la manière suivante.

Salaires en euros	1 000	2 000	10 000
Effectifs	100	20	5

La moyenne de cette série est $\overline{x} = 1\,520$ euros alors que sa médiane est de 1 000 euros, la moyenne seule donne une idée fausse des salaires dans l'entreprise. Une bonne étude statistique doit donc comporter ces deux caractéristiques.

c) Les quartiles

Les quartiles ont le même rôle que la médiane, mais au lieu de séparer la série statistique en deux parties égales, ils la séparent en quatre groupes de même effectif.

On note Q_1 le premier quartile (premier quart de l'effectif).

On note Q_3 le dernier quartile (troisième quart de l'effectif)

(le deuxième quartile est en fait la médiane).

d) Les déciles

Les déciles se calculent comme la médiane et les quartiles mais séparent la série en dix parties d'effectifs égaux.

On note : D_1 le premier décile correspondant à un dixième de l'effectif, D_2 le deuxième décile correspondant à deux dixième de l'effectif, etc. (D_5 est la médiane).

3. Caractéristiques de dispersion

a) Intervalle interquartile

À savoir

On appelle intervalle interquartile la différence $Q_3 - Q_1$, cet intervalle permet de connaître la dispersion de la série statistique autour de sa médiane.

Exercice résolu

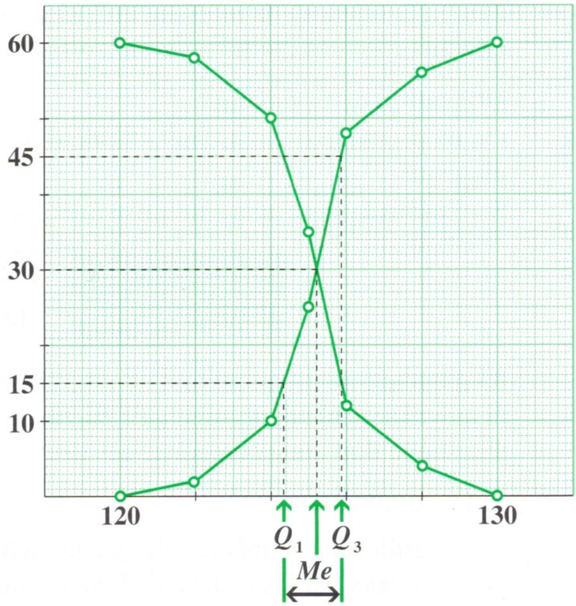
On donne la série statistique mesurant la masse en g des pièces grossièrement usinées.

Classes	[120, 122[	[122, 124[	[124, 125[	[125, 126[	[126, 128[	[128, 130[
Effectifs	2	8	15	23	8	4

Représenter les effectifs cumulés. Calculer la médiane et les quartiles. Calculer l'intervalle interquartile.

On obtient le tableau ci-dessous pour les effectifs cumulés

Classes	[120, 122[	[122, 124[	[124, 125[	[125, 126[	[126, 128[	[128, 130[
Effectifs	**2**	**8**	**15**	**23**	**8**	**4**
Effectifs cumulés croissants	**2**	**10**	**25**	**48**	**56**	**60**
Effectifs cumulés décroissants	**60**	**58**	**50**	**35**	**12**	**4**



Intervalle interquartile

La médiane correspond à un effectif cumulé de 30, elle se situe donc dans l'intervalle [125, 126[.

	moins de 125	**25**		
1	Me	**30**	**5**	**23**
	moins de 126	**48**		

On a $\dfrac{Me - 125}{1} = \dfrac{5}{23}$ donc $Me \approx 125{,}2$.

Le premier quartile correspond à un effectif cumulé de 15, il se situe donc dans l'intervalle [124, 125[.

	moins de 124	**10**		
1	Q_1	**15**	**5**	**15**
	moins de 125	**25**		

donc $\dfrac{Q_1 - 124}{1} = \dfrac{5}{15}$ d'où $Q_1 \approx 124{,}3$.

On procède de même pour Q_3.

	moins de 125	**25**		
1	Q_3	**45**	**20**	**23**
	moins de 126	**48**		

donc $\dfrac{Q_3 - 125}{1} = \dfrac{20}{23}$ d'où $Q_3 \approx 125{,}9$.

L'intervalle interquartile est $125{,}9 - 124{,}3 = 1{,}6$.
Ce nombre signifie que du premier quart au troisième quart (soit la moitié) de l'effectif, il y a un écart de 1,6 g.
Ceci montre que la série statistique est concentrée autour de sa médiane.

b) Variance et écart type

Considérons une série statistique de moyenne $\overline{x}$, on appelle **variance** de cette série le nombre positif V défini par :

Définition

$$V = \frac{\sum_{i=1}^{p} n_i (x_i - \overline{x})^2}{N} \quad \text{où } N = \sum_{i=1}^{n} n_i.$$

À savoir

On montre que $$V = \frac{\sum_{i=1}^{p} n_i x_i^2}{N} - \overline{x}^2.$$

À noter que la deuxième formule est plus pratique à utiliser, notamment avec l'outil informatique.
L'**écart type** est la racine carrée de la variance, on le note σ_x.

À savoir

$$\sigma_x = \sqrt{V}.$$

Si l'écart type est petit par rapport à l'étendue de la série statistique, cela siginifie que la série est très concentrée autour de sa moyenne.

Exercice résolu

On donne de nouveau la série statistique de l'exercice résolu précédent, calculer son écart type.

On fait un tableau dans lequel figurent les lignes pour x_i, n_i, $n_i x_i$ et $n_i x_i^2$. On obtient :

x_i	121	123	124,5	125,5	127	129	Totaux
n_i	2	8	15	23	8	4	60
$n_i x_i$	242	984	1 867,5	2 886,5	1 016	516	7 512
$n_i x_i^2$	29 282	121 032	232 503,75	362 255,75	129 032	66 564	940 669,5

$$\overline{x} = \frac{\sum_{i=1}^{6} n_i x_i}{N} = \frac{7\,512}{60} = 125{,}2.$$

$$V = \frac{\sum_{i=1}^{6} n_i x_i^2}{N} - \overline{x}^2 = \frac{940\,669{,}5}{60} - 125{,}2^2 = 2{,}785.$$

$\sigma_x = \sqrt{2{,}785} \approx 1{,}7$.
En réalité l'écart type de cette série est supérieur à 1,7 car le fait de regrouper les valeurs en classes nous a fait négliger la dispersion à l'intérieur de chacune des classes.