Chapitre 1 : Séries statistiques à une variable

2024/2025

Sommaire

1. Séries à caractère qualitatif
2. Séries à caractère quantitatif
3. Caractéristiques de position
4. Caractéristiques de dispersion

1. Séries à caractère qualitatif

Une série statistique à une variable qualitative peut être représentée sous forme d'un tableau :

(i)

Nom et prénom	date de naissance	lieu de naissance	profession du père	Nombre de frères et soeurs
ALIBERT Jacques	25/02/1988	Paris-15^e	épicier	0
ANDRE Jean	7/11/1988	Colombes (Seine)	ingénieur	2
BERNARD Louis	14/8/1988	Orléans (Loiret)	instituteur	1
....	....	....	....	.....

(ii) Nombre de clients préférant un produit parmi L et N :

Modalités	Nombre de clients
Préférer L	45
Préférer N	35
Indifférence	20
Total	100

1. Séries à caractère qualitatif

Définition 1.1

De façon générale une série statistique à une variable qualitative peut être résumée par un tableau de la forme suivante :

Modalités du caractère X	X_1	X_2	$\cdots$	X_i	$\cdots$	X_k	Total
Effectif	n_1	n_2	$\cdots$	n_i	$\cdots$	n_k	N

où le nombre total des observations est

$$N = n_1 + n_2 + \cdots + n_i + \cdots n_k \text{ ou encore } N = \sum_{i=1}^{k} n_i.$$

Les effectifs peuvent être remplacés par les **fréquences** f_i, où

$$f_i = \frac{n_i}{N} \qquad \text{avec } 1 \leq i \leq k.$$

1. Séries à caractère qualitatif

Remarque 1.1

$$\sum_{i=1}^{k} f_i = \frac{n_1}{N} + \frac{n_2}{N} + \cdots + \frac{n_k}{N} = \frac{n_1 + n_2 + \cdots + n_k}{N} = \frac{N}{N} = 1$$

Les fréquences sont souvent exprimées en pourcentages.

Définition 1.2

Le mode, ou la dominante, est la modalité du caractère qui correspond au plus grand effectif (ou à la plus grande fréquence). Son symbole est M_0.

1. Séries à caractère qualitatif

Exemple 1.1

Modalités	Nombre de clients	Fréquences
Préférer L	45	0,45
Préférer N	35	0,35
Indifférence	20	0,20
Total	100	1

Le mode de la série est la modalité *Préférer L*.

2. Séries à caractère quantitatif

Une série statistique à une variable quantitative peut être également représentée sous forme d'un tableau :

(i) Nombre d'ouvriers ayant eu 0, ou 1, $\cdots$ 5 ou plus de 5 accidents dans l'année écoulée :

Classes Nombre d'accidents	Nombre d'ouvriers
0	450
1	231
2	80
4	9
5 ou plus de 5	5

(ii)

Kilométrage annuel (en milliers de km)	moins de 4	$[4, 8[$	$[8, 12[$	$[12, 16[$	$[16, 20[$	20 et plus	Total
Effectifs (nombres de voitures)	228	634	821	475	233	87	2478

2. Séries à caractère quantitatif

(iii) Une série à caractère quantitatif peut être donnée aussi sous forme d'une liste de nombres :

Nombre de pièces fabriquées par chacun des 70 ouvriers d'une entreprise

87	80	107	91	83	70	91	93	71	86
98	104	107	89	109	126	88	107	85	115
103	89	74	70	97	118	102	122	97	98
91	108	100	80	93	90	90	79	93	102
83	105	62	87	92	107	89	90	118	91
115	88	70	87	99	105	95	98	88	90
89	73	92	94	115	105	88	99	92	76

Il convient alors de dépouiller les observations et de les représenter par un tableau comme dans les exemples ci-dessus.

2. Séries à caractère quantitatif

Définition 2.1

De façon générale une série statistique à une variable quantitative peut être résumée par un tableau de la forme suivante, en indiquant soit les effectifs, soit les fréquences, comme pour une série à variable qualitative :

Valeurs du caractère X	X_1	X_2	$\cdots$	X_i	$\cdots$	X_k	Total
Effectifs	n_1	n_2	$\cdots$	n_i	$\cdots$	n_k	N
Fréquences	f_1	f_2	$\cdots$	f_i	$\cdots$	f_k	1

où N est le nombre total des observations : $N = \sum_{i=1}^{k} n_i$,

$f_i = \frac{n_i}{N} \quad \text{avec } 1 \leq i \leq k$, et $\sum_{i=1}^{k} f_i = 1$.

Définition 2.2

Lorsque les valeurs du caractère sont rangées dans l'ordre croissant on définit l'effectif cumulé croissant (respectivement la fréquence cumulée croissante) de la $i^{\text{ème}}$ valeur par $N_i = n_1 + n_2 + \cdots + n_i$, respectivement par $F_i = f_1 + f_2 + \cdots + f_i$.

Valeurs de X	Effectifs	Effectifs cumulés croissants
X_1	n_1	$N_1 = n_1$
X_2	n_2	$N_2 = n_1 + n_2 = N_1 + n_2$
X_3	n_3	$N_3 = n_1 + n_2 + n_3 = N_2 + n_3$
$\ldots$	$\ldots$	....
X_i	n_i	$N_i = n_1 + ... + n_i = N_{i-1} + n_i$
...	$\ldots$	....
X_k	n_k	$N_k = n_1 + ... + n_k = N_{k-1} + n_k = N$
Total	N	$\downarrow$

.

Exemple 2.1

Kilométrage annuel (en milliers de km)	Effectifs (nombres de voitures)	Effectifs cumulés	Fréquences en %	Fréquences cumulées
moins de 4	228	228	$9,2$	$9,2$
$[4,8[$	634	862	$25,6$	$34,8$
$[8,12[$	821	1 683	$33,1$	$67,9$
$[12,16[$	475	2 158	$19,2$	$87,1$
$[16,20[$	233	2 391	$9,4$	$96,5$
20 et plus	87	2 478	$3,5$	100
Total	2478		100	

Combien de véhicules ont parcouru moins de 12 000 km par an ?
C'est l'effectif cumulé de la classe $[8,12[$ qui nous permet de répondre : 1 683 véhicules .
La fréquence cumulée de cette même classe nous permet d'affirmer que 67,9% de véhicules ont parcouru moins de 12 000 km par an.

3. Caractéristiques de position

Cette section concerne les séries statistiques de variable quantitative.

Définition 3.1

Le mode, ou la dominante, est la valeur de la variable qui correspond au plus grand effectif (ou à la plus grande fréquence).
Son symbole est M_0.

Exemple 3.1

Tailles	Effectifs	Tailles	Effectifs
160	2	165	4
161	3	166	6
162	5	167	5
163	3	168	0
164	0	169	2

Le mode de cette série est $M_0 = 166$.

Définition 3.2

La médiane est la valeur de la variable telle que, la distribution étant ordonnée, il y ait autant d'observations rangées avant elle que d'observations rangées après elle. Son symbole est M_e.

Exemple 3.2

$$28 \quad 33 \quad 34 \quad 35 \quad 36 \quad 37 \quad 37 \quad 39 \quad 40$$

$M_e = 36$ car **4 valeurs** sont classées **avant** 36 et **4 valeurs après** .

Exemple 3.3

Voici les notes obtenues par un candidat à l'issue de 10 épreuves : 6 7 7 8 10 11 12 12 12 14 . Les notes étant en nombre pair, l'intervalle médian est [10,11], et toute valeur de cet intervalle est médiane. Par convention on prendra $M_e = \frac{10+11}{2} = 10,5$.

Définition 3.3

La moyenne arithmétique de plusieurs valeurs est le quotient de leur somme par le nombre de ces valeurs.
Dans la suite nous utiliserons, comme dans le langage courant, le mot **moyenne** pour "**moyenne arithmétique**".
Si les observations sont connues sous forme d'effectifs (ou de fréquences) par valeur :

Caractère quantitatif X	x_1	x_2	...	x_i	...	x_k	Total
Effectifs	n_1	n_2	...	n_i	...	n_k	N

la moyenne des N valeurs, que l'on désigne par $\bar{x}$, est :

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{k} n_i x_i \quad \text{ou} \quad \bar{x} = \sum_{i=1}^{k} f_i x_i$$

Exemple 3.4

Distribution des 35 élèves d'une classe selon leur nombre de frères et soeurs		
nombre x_i de frères et soeurs	nombre d'élèves n_i	$n_i * x_i$
0	1	0
1	12	12
2	9	18
3	6	18
4	6	24
5	0	0
6	1	6
Total	35	78

La moyenne est $\bar{x} = \frac{1}{N} \sum_{i=1}^{7} n_i x_i = \frac{78}{35} = 2,2.$

Exemple 3.5

Les observations sont connues individuellement :

On a mesuré les tailles de $N = 6$ élèves et obtenu, en arrondissant au centimètre le plus proche : $156 \quad 152 \quad 163 \quad 155 \quad 158 \quad 154$.

Alors $\bar{x} = \dfrac{1}{6}\displaystyle\sum_{i=1}^{6} x_i = \dfrac{938}{6} = 156,3$

Remarque 3.1

Lorsque les observations sont connues individuellement la formule de la moyenne devient

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{7} x_i.$$

3. Caractéristiques de position

Exemple 3.6

Distribution de 1 526 manoeuvres agricoles, selon la durée hebdomadaire du travail			
classe de durée (en heures)	effectif n_i	centre x_i	produit $n_i * x_i$
moins de 40	10	37	370
40 à moins de 45	87	42	3 654
45 à moins de 50	473	47	22 231
50 à moins de 55	457	52	23 764
55 à moins de 60	111	57	6 327
60 à moins de 65	304	62	18 848
65 à moins de 70	54	67	3 618
70 et plus	30	72	2 160
Total	1526		80 972

Le centre de la classe "45 à moins de 50'' a été choisi égal à 47, de même pour les autres classes, et par analogie, les centres des classes extrêmes sont 37 et 72.

$$\bar{x} = \frac{80\ 972}{1\ 526} = 53,1\ h$$

Remarque 3.2

Si les valeurs sont regroupées par classes (ou intervalles), les x_i représentent les centres de ces classes.

Remarque 3.3

La moyenne arithmétique est de même nature et s'exprime avec la même unité que la variable observée. On dit qu'elle est homogène aux observations.
C'est la caractéristique de position la plus représentative d'une distribution statistique car sa définition ne laisse aucune place à l'interprétation du statisticien. Sa valeur dépend de toutes les observations, elle est toujours calculable et est unique.

Définition 4.1

L'étendue d'une distribution statistique est la différence entre la plus grande et la plus petite des valeurs observées. On la note W.
Ainsi W=$x_{max} - x_{min}$.

Exemple 4.1

Soit la distribution 12 7 5 -1 4 -7 9.
L'étendue de la distribution est W= 12-(-7)=19.

Remarque 4.1

L'étendue est une caractéristique qui ne dépend que de deux observations extrêmes et risque d'être gravement affectée par une valeur exceptionnelle ou erronée. Ainsi la distribution suivante :
12 7 5 -1 4 -2 9, a une étendue W= 12-(-2)= 14.

4. Caractéristiques de dispersion

Définition 4.2

On appelle quartiles les trois valeurs de la variable qui partagent les observations d'une série statistique, rangées par valeurs croissantes, en quatre groupes de même effectif. On les note Q_1, Q_2, Q_3.
Plus précisément :
le 1^{er} quartile Q_1 est la plus petite valeur de la série telle qu'au moins 25% des valeurs sont inférieures ou égales à Q_1.
Le $3^{\text{ème}}$ quartile Q_3 est la plus petite valeur de la série telle qu'au moins 75% des valeurs sont inférieures ou égales à Q_3.
Q_2 est égal à la médiane.

Définition 4.3

On appelle écart interquartile la différence entre les quartiles extrêmes : $Q_3 - Q_1$.
L'intervalle $[Q_1, Q_3]$ est appelé intervalle interquartile.

4. Caractéristiques de dispersion

Exemple 4.2

Supposons qu'ayant observé 2 séries (A) et (B) de 9 lampes et étudié leur "durée de vie", en heures, nous ayons obtenu les résultats suivants :
type (A) : 780, 790, 790, 800, 800, 800, 810, 810, 820
type (B) : 400, 600, 600, 800, 800, 800, 1 000, 1 000, 1 200

Chaque série a un effectif total $N = 9$ donc $\dfrac{N}{4} = 2,25$ et $\dfrac{3}{4}N = 6,75$.

Alors, pour chacune d'elles, Q_1 est la $3^{\text{ème}}$ valeur et Q_3 la $7^{\text{ème}}$ valeur.

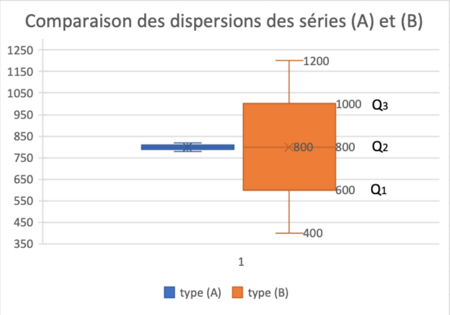
Ainsi pour la série de type (A) : Q_1=790, Q_2 =800 et Q_3 =810.

Pour la série de type (B) : Q_1=600, Q_2 =800 et Q_3 =1 000.

Définition 4.4

Une boîte à moustaches est une représentation graphique permettant de visualiser les quartiles d'une série statistique. Les extrémités de la boîte correspondent aux quartiles extrêmes. Chaque moustache mesure au plus 1,5 fois l'écart interquartile.

Exemple 4.3



Les deux séries ont même médiane 800, même mode 800 et même moyenne arithmétique 800.
Et pourtant la série de type (B) est plus dispersée que la série de type (A) : son étendue est plus importante, son écart inter-quartile également, comme le montrent leurs boîtes à moustaches respectives.

Définition 4.5

La variance V d'une distribution statistique d'effectif $N = \sum_{i=1}^{k} n_i$ et de moyenne $\bar{x}$, est égale à

$$V = \frac{1}{N} \sum_{i=1}^{k} n_i (x_i - \bar{x})^2$$

C'est la moyenne des carrés des écarts des observations à leur moyenne.
L'écart-type est la racine carrée de la variance V : $\sigma = \sqrt{V}$.
Il s'exprime dans la même unité que la variable.

4. Caractéristiques de dispersion

Définition 4.6

Le coefficient de variation CV est le rapport de l'écart-type à la moyenne : $CV = \dfrac{\sigma}{\bar{x}}$, lorsque $\bar{x} \neq 0$.
Le coefficient de variation n'a pas d'unité, ce qui permet de comparer les variations de différentes variables.

Théorème 4.1 (de Koenig-Huygens)

On a aussi : $V = \dfrac{1}{N}\displaystyle\sum_{i=1}^{k} n_i x_i^2 \ - \ \bar{x}^2$.

Remarque 4.2

Lorsque les valeurs sont connues individuellement, on a $V = \dfrac{1}{N}\displaystyle\sum^{k}(x_i - \bar{x})^2 = \dfrac{1}{N}\sum^{k} x_i^2 \ - \ \bar{x}^2$.

Remarque 4.3

Une grande variance indique une distribution dispersée, une petite variance indique une distribution statistique "ramassée" autour de la moyenne. Il en est de même pour le coefficient de variation.

Exemple 4.4

Dans l'exemple précédent, la série de type (A) a pour variance $V(A) = 133,33$, pour écart-type $\sigma(A) = 11,55$ et pour coefficient de variation $CV(A) = 0,014$.
La série de type (B) a pour variance $V(B) = 53\ 333,33$, pour écart-type $\sigma(B) = 230,94$ et pour coefficient de variation $CV(B) = 0,29$.

On retrouve que la série de type (B) est plus dispersée que la série de type (A).

Exemple 4.5 : Disposition pratique

Age x_i	Effectif n_i	$n_i * x_i$	$n_i * x_i^2$
18	4	72	1 296
25	2	50	1 250
31	9	279	8 649
35	7	245	8 575
39	10	390	15 210
42	3	126	5 292
48	1	48	2 304
59	2	118	6 962
72	2	144	10 368
Total	40	1 472	59 906

Age moyen : $\bar{x} = \frac{1}{N}\sum_{i=1}^{9} n_i x_i = \frac{1\ 472}{40} = 36,8$

Variance : $V = \frac{1}{N}\sum_{i=1}^{9} n_i x_i^2 \ - \ \bar{x}^2 = \frac{59\ 906}{40} - 36,8^2 = 143,41.$

Ecart-type : $\sigma = \sqrt{143,41} \approx 11,98$ (en années)