STATISTIQUES DESCRIPTIVES

Table des matières

★ ★ ★ ★ ★ ★

I Définitions et vocabulaire des statistiques

La **population** est l'ensemble des individus sur lesquels portent l'étude statistique. (Par exemple classe de seconde ISI, hommes, habitants de la France . . .)

Le **caractère** (ou **variable**) d'une série statistique est une propriété étudiée sur chaque individu :

- Lorsque le caractère ne prend que des valeurs (ou **modalités**) numériques, il est **quantitatif** :
  - **discret** s'il ne peut prendre que des valeurs isolées (notes, âge . . .)
  - **continu** dans le cas contraire (poids, taille . . .). Dans ce cas on effectue souvent un regroupement des valeurs par **classes**.
- Sinon, on dit qu'il est **qualitatif** (couleur des yeux, sport pratiqué . . .) : les modalités ne sont pas des nombres.

A chaque valeur (ou classe) est associée un **effectif** n : c'est le nombre d'individus associés à cette valeur.

Faire des **statistiques**, c'est recueillir, organiser, synthétiser, représenter et exploiter des données, numériques ou non, dans un but de comparaison, de prévision, de constat...

Les plus gros "consommateurs" de statistiques sont les **assureurs** (risques d'accidents, de maladie des assurés), les **médecins** (épidémiologie), les **démographes** (populations et leur dynamique), les **économistes** (emploi, conjoncture économique), les **météorologues** . . .

Définition 1
On considère une série statistique à caractère quantitatif, dont les p valeurs sont données par : x_1, x_2, . . ., x_p d'effectifs associés n_1, n_2, . . ., n_p avec $n_1 + n_2 + ... + n_p = N$.

- *A chaque valeur (ou classe) est associée une fréquence f_i : c'est la proportion d'individus associés à cette valeur.*
- *$f_i = \dfrac{n_i}{N}$ est un nombre compris entre 0 et 1, que l'on peut écrire sous forme de pourcentage.*
- *L'ensemble des fréquences de toutes les valeurs du caractère s'appelle la caractère s'appelle la distribution des fréquences de la série statistique.*

Exemple 1
Voici les notes obtenues à un contrôle dans une classe de 30 élèves : (**Série A :**)
$2-3-3-4-5-6-6-7-7-7-8-8-8-8-8-9-9-9-9-9-9-10-10-11-11-11-13-13-15-16$
On peut représenter cette série par un tableau d'effectifs, et le compléter par la distribution des fréquences :

Notes	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Eff.	0	1	2	1	1	2	3	5	6	2	3	0	2	0	1	1	0	0	0
Fréq. en %	0	3	7	3	3	7	10	17	20	7	10	0	7	0	3	3	0	0	0

Remarque 1
On peut vérifier que la somme des fréquences est égale à 1 (ou à 100 si on les exprime en pourcentages).

On peut aussi faire un regroupement par classe, ce qui rend l'étude moins précise, mais qui permet d'avoir une vision plus globale.

Exemple 2
Toujours pour la **série A**, si on regroupe les données par classes d'amplitude 5 points, on obtient :

Notes	$[\,0\,;\,5\,[$	$[\,5\,;\,10\,[$	$[\,10\,;\,15\,[$	$[\,15\,;\,20\,[$	total
Effectif	4	17	7	2	30
Fréquence	$0,13$	$0,57$	$0,23$	$0,07$	1

II Caractéristiques de position

II.1 Moyenne

Définition 2
Soit une série statistique à caractère quantitatif, dont les p valeurs sont données par x_1, x_2, …, x_p d'effectifs associés n_1, n_2, …, n_p avec $n_1+n_2+...+n_p=N$.
La moyenne pondérée de cette série est le nombre noté $\overline{x}$ qui vaut

$$\overline{x}=\frac{n_1x_1+n_2x_2+...+n_px_p}{n_1+n_2+...+n_p}=\frac{1}{N}\sum_{i=1}^{p}n_ix_i.$$

Remarque 2
Lorsque la série est regroupée en classes, on calcule la moyenne en prenant pour valeurs x_i le **centre de chaque classe** ; ce centre est obtenu en faisant la moyenne des deux extrémités de la classe.

Exemple 3
- Dans la **série A**, la moyenne du contrôle est égale à $\overline{m}=\dfrac{2\times 1+3\times 2+\cdots+16\times 1}{30}=\dfrac{254}{30}\approx 8,47$,
- si on regroupe par classe d'amplitude 5 points, une estimation de la moyenne est :
$\overline{m}\dfrac{2,5\times 4+7.5\times 17+\cdots+17,5\times 2}{30}=\dfrac{260}{30}\approx 8,67.$

Remarque 3
On peut aussi calculer une moyenne à partir de la distribution de fréquences :

$$\overline{x} = f_1x_1 + f_2x_2 + \cdots + f_px_p = \sum_{i=1}^{p} f_ix_i.$$

Propriété 1 (Linéarité de la moyenne)

- Si on ajoute (ou soustrait) un même nombre k à toutes les valeurs d'une série, alors la moyenne de cette série se trouve augmentée (resp. diminuée) de k.
- Si on multiplie (ou divise) par un même nombre non nul k toutes les valeurs d'une série, alors la moyenne de cette série se trouve multipliée (resp. divisée) par k.

Exemple 4
On considère la **série A** :

- Si on ajoute $1,5$ points à chaque note du contrôle, alors la moyenne de classe devient $\overline{m} = 8,47 + 1,5 = 9,97$.
- Si on augmente chaque note de 10%, cela revient à multiplier chaque note par $1,1$, ce qui donne $\overline{m} = 8,47 \times 1,1 = 9,32$.

Propriété 2 (Moyenne par sous-groupes)
Soit une série statistique, d'effectif total N, de moyenne $\overline{x}$.
Si on divise cette série en deux sous-groupes **disjoints** d'effectifs respectifs p et q (avec $p + q = N$) de moyennes respectives $\overline{x}_1$ et $\overline{x}_2$, alors on a :

$$\overline{x} = \frac{p}{N} \times \overline{x}_1 + \frac{q}{N} \times \overline{x}_2.$$

Exemple 5
On suppose par exemple que les 12 garçons de la classe de la **série A** ont obtenu une moyenne globale de 8 sur 20.

- La moyenne du groupe formé par les filles de la classe vérifie : $9,47 = \frac{12}{30} \times 8 + \frac{18}{30} \times \overline{m}_f$.
- Soit $\overline{m}_f = \frac{30}{18}\left(9,47 - \frac{12}{30} \times 8\right) = 10,45$.

II.2 Médiane

Définition 3
Soit une série statistique ordonnée dont les n valeurs sont $x_1 \leqslant x_2 \leqslant x_3 \leqslant \cdots \leqslant x_n$.
La médiane est un nombre M qui permet de diviser cette série en deux sous-groupes de même effectif.

- *Si n est impair, n est la valeur de cette série qui est située au milieu, à savoir la valeur dont le rang est $\frac{n+1}{2}$, notée $x_{\frac{n+1}{2}}$.*
- *Si n est pair, n est le centre l'intervalle médian, qui est l'intervalle formé par les deux nombres situés « au milieu » de la série, à savoir $x_{\frac{n}{2}}$ et $x_{\frac{n}{2}+1}$.*

Exemple 6

- La médiane de la série « $2 - 5 - 6 - 8 - 9 - 9 - 10$ » est 8.
- La médiane de la série « $2 - 5 - 6 - 8 - 9 - 9$ » est 7.
- La médiane de la série « $2 - 5 - 6 - 6 - 9 - 10$ » est 6.

II.3 Quartiles

Définition 4
Soit une série statistique, on appelle quartiles de la série un triplet de réels (Q_1 ; Q_2 ; Q_3) qui sépare la série en quatre groupes de même effectif.

Remarque 4
Par définition, si X est une série statistique, $Q_2 = Med(X)$.

La calculatrice donne les valeurs de Q_1, Med et Q_3.
Nous verrons dans un prochain chapitre d'autres façons de trouver la médiane et les quartiles.

Exemple 7
Pour la **série A**, la calculatrice nous donne $Q_1 = 7$, $Méd = 8,5$ et $Q_3 = 10$.

III Caractéristiques de dispersion

Définition 5
On appelle étendue d'une série discrète X le réel défini par $E(X) = \max(X) - \min(X)$.

Il s'agit de la première mesure de la dispersion d'une série statistique. Son principal mérite a longtemps été d'exister, et de fournir une information sur la dispersion très simple à obtenir.

Exemple 8
L'étendue de la **série A** est de $E(A) = 16 - 2 = 14$.

Définition 6
On appelle intervalle inter-quartiles l'intervalle [Q_1 ; Q_3].
L'amplitude de cet intervalle est appelée écart inter-quartiles.

Exemple 9

- Dans la **série A**, l'intervalle inter-quartile est l'intervalle [7 ; 10] dont l'écart vaut $10 - 7 = 3$.
- Cet intervalle comprend donc la moitié des notes de la série située au centre de celle-ci.