

STATISTIQUES

Baptiste GORIN

I. Séries statistiques : vocabulaire de base

Définition I.1. — *Tout ensemble (au sens mathématique du terme) que l'on soumet à une étude statistique est appelé population. On nomme individu tout élément de l'ensemble en question. L'étude statistique d'une population concerne un aspect, une propriété, un trait commun à tous ces individus; on lui donne le nom de caractère. Enfin, les valeurs prises par le caractère sont appelées modalités.*

Remarque I.2. — Les premiers problèmes auxquels s'est intéressée la statistique étant de nature démographique, les termes « populations », « individu » et « caractère » sont restés.

Exemple I.3. — • population : une classe de seconde; individu : chaque élève de la classe; caractère : moyenne trimestrielle en mathématiques; modalités : valeurs des notes.

• population : population française adulte; individu : toute personne de nationalité française ayant plus de 18 ans révolus; caractère : revenus de l'année 2000; modalités : montant des revenus.

• population : population française; individu : toute personne résidant en France; caractère : région de résidence; modalités : l'Alsace, l'Aquitaine, l'Auvergne...

Définition I.4. — *Un caractère est dit quantitatif si ses modalités sont des nombres, qualitatif sinon.*

Définition I.5. — *Un caractère quantitatif est dit :*

- discret si les valeurs prises par le caractère sont isolées ;
- continu si les valeurs du caractère sont prises dans tout intervalle de $\mathbb{R}$.

Exemple I.6. — • Le nombre de frères et sœurs est un caractère quantitatif discret car les modalités sont 0,1,2,3,...

• Le temps passé quotidiennement à regarder la télévision est un caractère quantitatif continu.

• « Fumeur » est un caractère qualitatif car les modalités sont « oui » et « non ».

Remarque I.7. — Les modalités d'un caractère quantitatif peuvent être regroupées dans des intervalles disjoints appelés classes (statistiques). C'est le cas pour un caractère continu, mais également pour certains caractères discrets dont les modalités sont trop nombreuses.

Exemple I.8. — • population : les marathoniens; caractère (quantitatif continu) : temps mis pour effectuer le marathon de New-York; classes : temps compris entre 2h et 2h30 (strictement), entre 2h30 et 3h (strictement),..., supérieur à 4h.

• population : ensemble des communes françaises; caractère (quantitatif discret) : nombre d'habitants; classes : nombre d'habitants compris entre 1 et 49, entre 50 et 99, entre 100 et 199,..., entre 50000 et 99999, supérieur à 100000.

Définition I.9. — *L'effectif d'une modalité est le nombre d'individus de la modalité. L'effectif total est le nombre d'individus de la population.*

Remarque I.10. — L'effectif total est la somme des effectifs de chaque modalité.

Définition I.11. — *La fréquence d'une modalité est le rapport f de son effectif à l'effectif total. $100f\%$ est l'expression de la fréquence en pourcentage.*

Exemple I.12. — Considérons les notes d'une classe :

5, 7, 8, 8, 9, 9, 10, 10, 10, 10, 11, 11, 11, 12, 12, 12, 13, 13, 13, 14, 14, 15, 16, 17.

La fréquence de la modalité 10 est : $f = \frac{4}{25} = 0,16$, soit $f = 16\%$.

La fréquence de la modalité 7 est : $f = \frac{1}{25} = 0,04$, soit $f = 4\%$.

Théorème I.13. — *La somme des fréquences est 1 ou 100%.*

Définition I.14. — *Une série statistique est la donnée :*

- d'une population déterminée ;
- d'un caractère précis ;
- de l'effectif ou de la fréquence de chaque valeur, classe ou modalité du caractère.

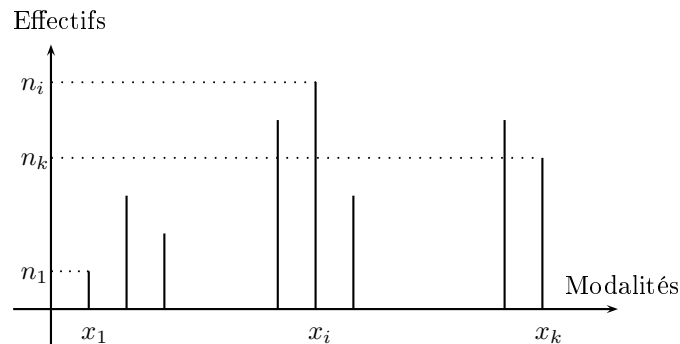
II. Représentation d'une série statistique

Généralement présentée sous la forme d'un tableau de données, une série statistique est également représentée par un diagramme. Nous allons nous intéresser aux principales représentations graphiques.

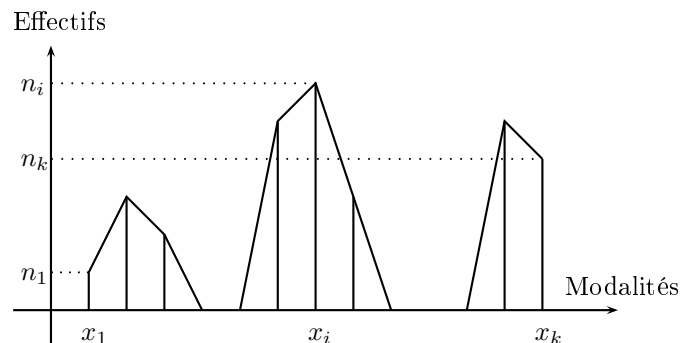
1. Cas d'un cas caractère quantitatif discret

Diagramme en bâtons. Polygones d'effectifs

Désignons par $x_1, x_2, \dots, x_k$ les modalités du caractère avec $x_1 < x_2 < \dots < x_k$ et $n_1, n_2, \dots, n_k$ les effectifs correspondants. On a une représentation de la forme, dite diagramme en bâtons :



Les lignes brisées obtenues en joignant les « sommets des bâtons » s'appellent polygones d'effectifs :



Histogramme

Voir le paragraphe suivant.

2. Cas d'un caractère quantitatif continu

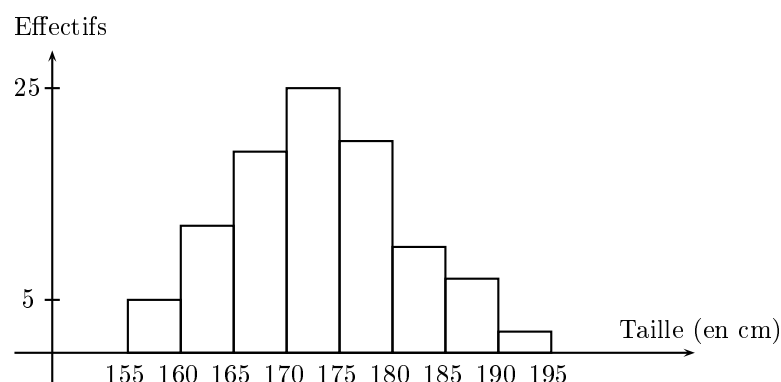
Histogramme

L'histogramme d'une série s'obtient en traçant des rectangles ayant pour bases les intervalles de classes et une aire proportionnelle à l'effectif correspondant.

Exemple II.1. — Considérons la série statistique suivante portant sur une population adulte de 100 personnes :

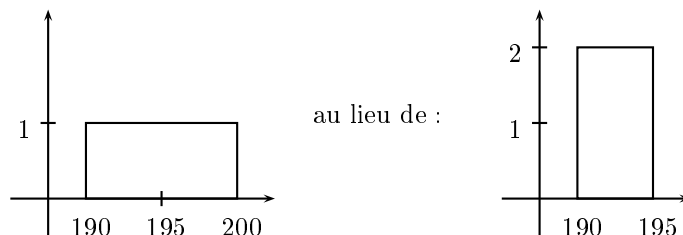
Taille (en cm)	Effectifs
[155, 160[	5
[160, 165[	12
[165, 170[	19
[170, 175[	25
[175, 180[	20
[180, 185[	10
[185, 190[	7
[190 et plus [	2

On a l'histogramme suivant :



Notons que, dans l'histogramme précédent, les intervalles de classes ont même amplitude, à savoir 5, étant entendu que la classe [190 et plus [a été assimilée à la classe [190, 195[.

Si la classe [190 et plus [avait été assimilée à la classe [190, 200[, on aurait eu :



en conservant toujours l'aire du rectangle.

Remarque II.2. — Les histogrammes peuvent être employés pour représenter un caractère quantitatif discret (remarque I.7).

Remarque II.3. — Les classes peuvent être d'amplitudes inégales.

3. Cas d'un caractère qualitatif

Diagramme circulaire ou demi-circulaire

Un disque (ou un demi-disque) est partagé en secteurs d'aires proportionnelles aux effectifs ou aux fréquences de chaque modalité du caractère étudié.

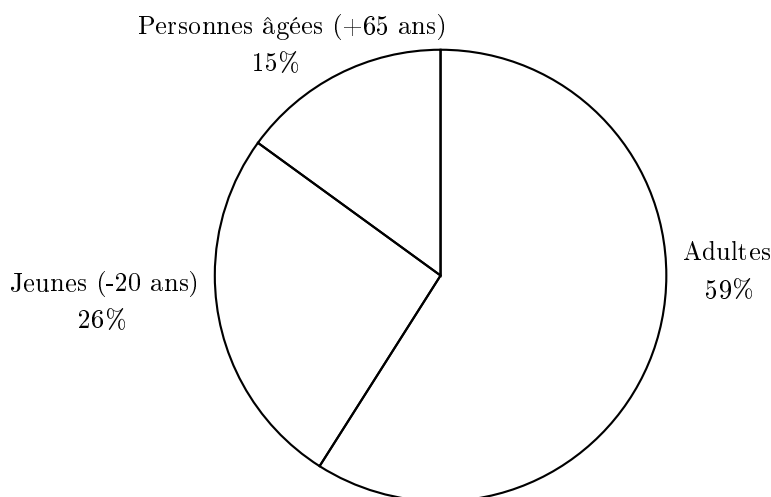
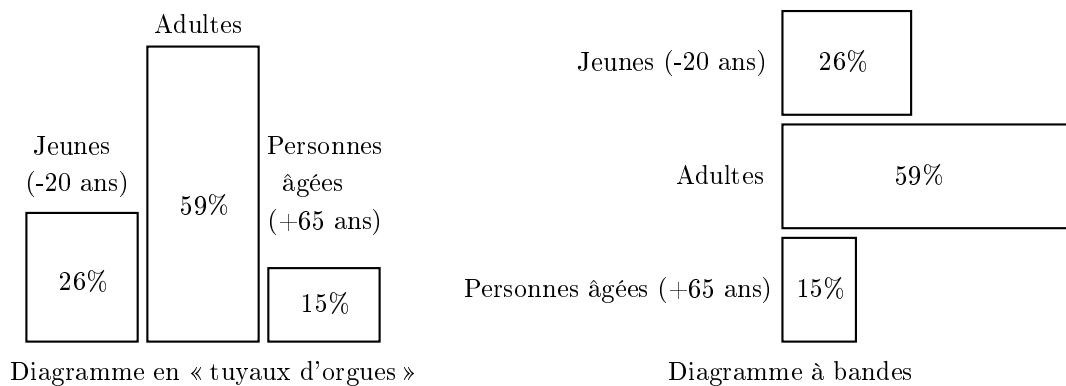


Diagramme à bandes. Diagramme en « tuyaux d'orgues »

Les aires des rectangles sont proportionnelles aux effectifs ou aux fréquences de chaque modalité du caractère étudié.



III. Indicateurs statistiques

1. Indicateurs de position

Définition (moyenne) III.1. — La moyenne, notée $\bar{x}$, d'une série correspondant à un caractère quantitatif est le quotient de la somme de toutes les modalités du caractère par l'effectif total.

Remarque III.2. — Considérons une série statistique correspondant à un caractère quantitatif :

Modalités	x_1	x_2	$\cdots$	x_k
Effectifs	n_1	n_2	$\cdots$	n_k

Alors la moyenne de la série est le nombre : $\bar{x} = \frac{n_1x_1 + n_2x_2 + \cdots + n_kx_k}{n_1 + n_2 + \cdots + n_k}$.

Exemple III.3. — Considérons les notes d'une classe :

5, 7, 8, 8, 9, 9, 10, 10, 10, 10, 11, 11, 11, 12, 12, 12, 13, 13, 14, 14, 15, 16, 17.

La moyenne de la classe est : $\bar{x} = \frac{282}{25} = 11,28$.

Si on représente cette série statistique sous forme de tableau :

Notes	5	7	8	9	10	11	12	13	14	15	16	17
Effectifs	1	1	2	2	4	3	4	3	2	1	1	1

on retrouve : $\bar{x} = \frac{282}{25} = 11,28$

Définition (mode) III.4. — On appelle mode d'une série statistique toute valeur, ou modalité, ou classe d'effectif maximal. Lorsque les données sont regroupées en classes, le mode est appelé classe modale.

Une série statistique peut présenter plusieurs modes ; la série est alors dite multimodale.

Exemple III.5. — La série statistique de l'exemple III.3 est multimodale puisque ces modes sont 10 et 12, modalités pour lesquelles l'effectif, égal à 4, est maximal.

Définition (médiane) III.6. — Soit une série statistique correspondant à un caractère quantitatif aux modalités ordonnées. On appelle médiane de la série tout nombre M ayant la propriété suivante : l'effectif des valeurs strictement inférieur à M et l'effectif des valeurs strictement supérieures à M ne dépassent pas la moitié de l'effectif.

Exemple III.7. — Considérons les notes obtenues en mathématiques par un élève tout au long de l'année :

10, 12, 8, 13, 15, 15, 10, 9, 6, 11, 14.

Ordonnons les notes : 6, 8, 9, 10, 10, 11, 12, 13, 14, 15, 15.

11 est la note médiane puisqu'elle occupe le rang central ; il y a autant de notes supérieures à 11 que de notes inférieures à 11.

Considérons les notes d'un autre élève, absent à l'un des contrôles :

9, 5, 10, 12, 11, 8, 10, 7, 11, 9.

Ordonnons les notes : 5, 7, 8, 9, 9|10, 10, 11, 11, 12.

Les notes obtenues étant en nombre pair, il n'existe pas ici de rang central ; tout nombre compris entre 9 et 10 peut jouer le rôle de note médiane. On convient en général que 9,5 est la note médiane.

2. Indicateurs de dispersion

Définition (étendue) III.8. — L'étendue d'une série statistique correspondant à un caractère quantitatif est la différence entre la plus grande modalité et la plus petite modalité prise par le caractère.

Exemple III.9. — Reprenons la série statistique de l'exemple III.3. L'étendue est : $17 - 5 = 12$.

Définition (écart-moyen) III.10. — L'écart-moyen, noté e , d'une série statistique correspondant à un caractère quantitatif est la moyenne des écarts (à la moyenne).

Remarque III.11. — Considérons une série statistique correspondant à un caractère quantitatif :

Modalités	x_1	x_2	$\cdots$	x_k
Effectifs	n_1	n_2	$\cdots$	n_k

Alors l'écart-moyen de la série est le nombre : $e = \frac{n_1|x_1 - \bar{x}| + n_2|x_2 - \bar{x}| + \cdots + n_k|x_k - \bar{x}|}{n_1 + n_2 + \cdots + n_k}$.

Définition (variance) III.12. — La variance, notée V , d'une série statistique correspondant à un caractère quantitatif est la moyenne des carrés des écarts (à la moyenne).

Remarque III.13. — Considérons une série statistique correspondant à un caractère quantitatif :

Modalités	x_1	x_2	$\cdots$	x_k
Effectifs	n_1	n_2	$\cdots$	n_k

Alors la variance est le nombre : $V = \frac{n_1(x_1 - \bar{x})^2 + n_2(x_2 - \bar{x})^2 + \cdots + n_k(x_k - \bar{x})^2}{n_1 + n_2 + \cdots + n_k}$.

Théorème III.14. — Avec les notations de la remarque III.13, on a :

$$V = \frac{n_1x_1^2 + n_2x_2^2 + \cdots + n_kx_k^2}{n_1 + n_2 + \cdots + n_k} - \bar{x}^2.$$

Définition (écart-type) III.15. — L'écart-type, noté σ , d'une série statistique correspondant à un caractère quantitatif est la racine carrée de la variance : $\sigma = \sqrt{V}$.

Exemple III.16. — Reprenons la série statistique de l'exemple III.3. Après calculs, il vient :

$$\bar{x} = 11,28, \quad e \simeq 2,21, \quad V = 7,96, \quad \sigma \simeq 2,82.$$