

CHAP
01

Statistiques à une variable



► Compétences

- ✓ Maîtriser le vocabulaire des statistiques
- ✓ Comprendre l'intérêt des indicateurs
- ✓ Savoir utiliser la calculatrice pour déterminer les indicateurs classiques
- ✓ Connaître quelques représentations classiques

► Fil rouge *coloration drone*

Après un vol de reconnaissance, le pilote récupère le log d'altitude du drone, qui a enregistré sa position verticale toutes les 10 secondes.

Comment caractériser statistiquement ce profil de vol et évaluer la stabilité de la phase de croisière ?



I. Approche historique

► Point histoire

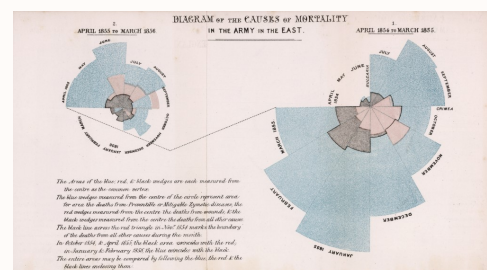
On attribue à l'histoire de la statistique ou des statistiques la date de commencement de 1749, bien que l'interprétation du terme « statistique » a changé au cours du temps.

Aux temps plus anciens, cette science ne consistait qu'à la collection d'informations des États, d'où l'étymologie du nom, de l'allemand Statistik, dérivé de l'italien statista (« Homme d'État »).

Plus tard, cette définition est étendue à tout type d'information collectée et, encore plus tard, les sciences statistiques incluent l'analyse et l'interprétation de ces données. En termes modernes, les statistiques incluent les ensembles de données, telles celles de la comptabilité nationale et les registres de températures, ainsi que le travail d'analyse, lequel requiert les méthodes de l'inférence statistique.

Florence Nightingale est une pionnière de la présentation visuelle de l'information. Elle utilise entre autres les diagrammes circulaires, les « Pie Chart », développés par William Playfair en 1801.

Après la guerre de Crimée, elle se met à utiliser une version améliorée de ces diagrammes afin d'illustrer les causes saisonnières de mortalité des patients de l'hôpital militaire qu'elle gère.



Source : Wikipedia

II. Définitions et notations

II.1. Notations

► Intro

Dans ce chapitre, on étudie des séries à caractères **quantitatifs** discrètes (à valeurs séparées) ou **continues** (dont les valeurs sont regroupées en *classes*, ou *intervalles*).

Dans le cas d'une série continue, on fait toujours l'hypothèse que la *répartition* des valeurs est *uniforme* à l'intérieur de chaque classe.

► Notations 2.1

- X est le caractère étudié et $x_1, x_2, \dots, x_p$ les *valeurs* du caractère ou les centres des classes dans le cas où les valeurs sont regroupées en classes. Ces valeurs sont rangées dans l'ordre croissant.
- $n_1, n_2, \dots, n_p$ sont les *effectifs* respectifs des valeurs $x_1, x_2, \dots, x_p$.
- $f_1, f_2, \dots, f_p$ sont les *fréquences* respectives des valeurs $x_1, x_2, \dots, x_p$.
- N est l'effectif total, et $N = n_1 + n_2 + \dots + n_p$

II.2. Fréquences

► Propriétés 2.1

La fréquence de la valeur x_i est donnée par $f_i = \frac{n_i}{N}$.

Ce nombre est compris dans l'intervalle $[0 ; 1]$ et est souvent écrit sous la forme d'un *pourcentage*.

La somme de toutes les fréquences est égale à 1 : $\sum_{i=1}^p f_i = f_1 + f_2 + \dots + f_p = 1$.

II.3. Effectif cumulé, fréquence cumulée

► Définitions 2.1 (Cas d'une série discrète)

C'est le nombre d'individus pour lesquels la valeur du caractère est *inférieure ou égale* à x_i .

On a $S_i = n_1 + n_2 + \dots + n_i$ et de ce fait on a $S_p = N$ (dernier effectif cumulé).

La fréquence cumulée en x_i est donnée par $F_i = f_1 + f_2 + \dots + f_i = \frac{S_i}{N}$.

► Définition 2.2 (Cas d'une série continue)

Pour l'effectif cumulé d'une classe $]a ; b]$, il s'agit de l'effectif cumulé en b (borne droite de l'intervalle), c'est-à-dire le nombre d'individus pour lesquels la valeur du caractère est inférieure ou égale à b .

Ainsi, lorsqu'on calcule des effectifs cumulés pour une série statistique groupée en classes, ces effectifs cumulés doivent être donc **impérativement** affectés aux *bornes de droite* de ces classes.

► Remarque (Hypothèse de répartition uniforme)

Pour des raisons pratique, on fait l'hypothèse de *répartition uniforme* à l'intérieur de chaque classe, c'est-à-dire qu'il y a « proportionnalité entre les effectifs et les largeurs ».

► Remarque

Il est souvent très pratique et utile de présenter des données statistiques sous forme d'un tableau d'effectifs (ou de fréquences), et d'y rajouter une ligne pour les ECC (ou les FCC).

III. Paramètres d'une série

III.1. Médiane et quartiles

► Définition 3.1

La valeur médiane d'une série statistique est la valeur M séparant la population en deux moitiés : les 50 % ayant une valeur inférieure ou égale à M et les 50 % ayant une valeur supérieure ou égale à M .

► Méthode 3.1 (Cas d'une série discrète)

Si l'effectif est impair, la médiane est la valeur de rang $(N+1)/2$; si l'effectif est pair, la médiane est la moyenne des valeurs de rang $N/2$ et $N/2 + 1$.

► Méthode 3.2 (Cas d'une série continue)

La médiane est la valeur correspondant à la fréquence cumulée égale à 0,5.

► Définitions 3.2

De même, le premier quartile est la valeur Q_1 telle que 25 % de la population a une valeur du caractère inférieure ou égale à Q_1 , les 75 % restants ayant une valeur supérieur ou égale à Q_1 .

Le troisième quartile est la valeur Q_3 telle que 75 % de la population a une valeur du caractère inférieure ou égale à Q_3 , les 25 % restants ayant une valeur supérieur ou égale à Q_3 .

► Méthode 3.3 (Cas d'une série discrète)

Si $N/4$ est un entier, Q_1 est la valeur de rang $N/4$ et Q_3 est la valeur de rang $3N/4$.

Si $N/4$ n'est pas un entier, Q_1 est la valeur de rang immédiatement supérieur à $N/4$ et Q_3 est la valeur de rang immédiatement supérieur à $3N/4$.

► Méthode 3.4 (Cas d'une série continue)

Q_1 est la valeur correspondant à la fréquence cumulée égale à 0,25 et Q_3 est la valeur correspondant à la fréquence cumulée égale à 0,75.

► Définitions 3.3

Intervalle inter-quartile : c'est l'intervalle $[Q_1 ; Q_3]$.

Écart inter-quartile : c'est la valeur de $Q_3 - Q_1$.

► Exemple 3.1 (cas discret)

On étudie les notes obtenues par une classe de $N = 30$ étudiants :

Notes x_i	02	03	04	05	06	07	08	09	10	11	13	15	16
Effectif n_i	1	2	1	1	2	3	5	6	2	3	2	1	1
Cumulé S_i	1	3	4	5	7	10	15	21	23	26	28	29	30

Médiane : $N = 30$ est pair, donc M est la moyenne des valeurs de rang 15 et 16.

On en déduit $M = \frac{8+9}{2} = 8,5$.

Premier quartile : $N/4 = 7,5$ donc Q_1 est la valeur de rang immédiatement supérieur, c'est-à-dire le rang 8.

On en déduit $Q_1 = 7$.

Troisième quartile : $3N/4 = 22,5$ donc Q_3 est la valeur de rang immédiatement supérieur, c'est-à-dire le rang 23.

On en déduit $Q_3 = 10$.

On obtient ainsi l'intervalle inter-quartile $[7 ; 10]$ et un écart inter-quartile de 3.

III.2. Moyenne

► Définition 3.4

Pour une série **discrete**, la moyenne des valeurs $x_1, x_2, \dots, x_p$ d'effectifs $n_1, n_2, \dots, n_p$ (effectif total N) est le réel :

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

Pour une série **continue**, on applique la même formule en prenant pour les valeurs x_i les centres des classes.

III.3. Variance et écart-type

► Intro

La moyenne d'un caractère statistique ne suffit pas pour caractériser le comportement de ses valeurs.

Par exemple, les étudiants d'un premier groupe peuvent avoir obtenu des résultats très homogènes (c'est-à-dire des notes comprises dans un petit intervalle autour de 10) alors que ceux d'un deuxième groupe peuvent avoir des résultats beaucoup plus dispersés autour de 10. L'**écart-type** est un nombre qui mesure de cette **dispersion**, au sens où plus l'écart-type est élevé plus les valeurs sont dispersées autour de la moyenne.

► Définitions 3.5

La **variance** d'une série statistique de caractère X prenant les valeurs (ou centres des classes) $x_1, x_2, \dots, x_p$ d'effectifs respectifs $n_1, n_2, \dots, n_p$ et de moyenne $\bar{x}$ est le nombre

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

On en déduit l'**écart-type** de la série $\sigma(X) = \sqrt{V(X)}$ qui est toujours un nombre positif.

► Point histoire



L'écart type est une grandeur dont l'invention remonte au XIX^e siècle, qui vit la statistique se développer au Royaume-Uni. C'est à **Abraham de Moivre** (1667 – 1754, ) qu'est attribuée la découverte du concept de mesure de la dispersion qui apparaît dans son ouvrage *The Doctrine of Chances* en 1718.



Mais le terme d'écart type (« standard deviation ») a été employé pour la première fois par **Karl Pearson** (1857 – 1936, ) en 1893 devant la Royal Society. C'est aussi lui qui utilise pour la première fois le symbole σ pour représenter l'écart type.

III.4. Utilisation de la calculatrice

► Calculatrice 3.1

Les calculatrices scientifiques permettent de déterminer les principaux paramètres statistiques d'une série.

La compétence en BTS SIO est principalement d'interpréter les paramètres obtenus à l'aide d'un outil numérique.

	List 1	List 2	List 3	List 4
SUB				
1	7.5	15		
2	20	20		
3	30	50		
4	37.5	30		
				7.5

	List 1	List 2
1Var XList	List1	
1Var Freq	List2	
2Var XList	List1	
2Var YList	List2	
2Var Freq	1	

1 variable	
$\bar{x}$	=37.375
Σx	=7475
Σx^2	=324750
σx	=15.0618516
sx	=15.099648
n	=200

IV. Représentations graphiques

► Intro

Il existe, comme présenté en cours, différentes représentations graphiques pour une série statistique.

Pour une série **discrète**, on utilise le diagramme en bâtons (ou en barres).

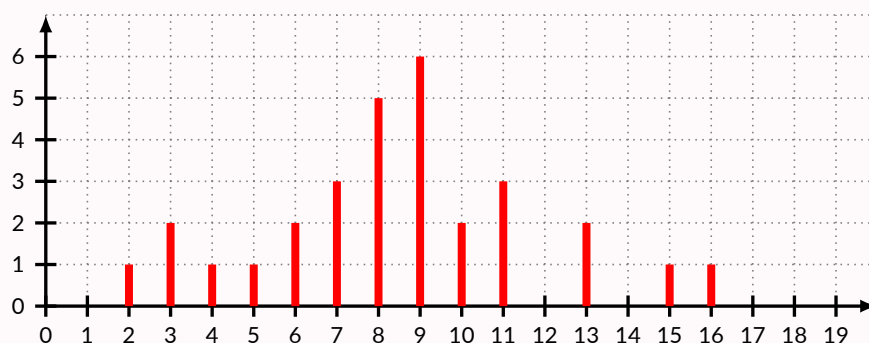
Pour aller plus loin dans l'exploitation d'une série – recherche de médiane, de quartiles... – qu'elle soit discrète ou continue, on dispose de deux représentations « outils », présentées un peu plus loin : la **boîte à moustaches** et la **courbe des effectifs (ou fréquences) cumulés**.

► Exemple 4.1

On reprend la série (discrète) des notes de la classe de 30 étudiants étudiée précédemment :

Notes	02	03	04	05	06	07	08	09	10	11	13	15	16
Eff.	1	2	1	1	2	3	5	6	2	3	2	1	1

► Illustration 4.1



► Logiciel

Le logiciel gratuit SineQuaNon, fonctionnant sous Windows , et disponible sur le site internet de Patrice Rabiller <http://patrice-rabiller.fr/SineQuaNon/menusqn.htm>, permet de traiter et d'exploiter des données statistiques, qu'elles soient quantitatives, qualitatives, discrètes ou continues...

L'interface est un peu *austère* mais les fonctionnalités sont bien présentes, et très pertinentes!

Statistiques à une variable

Variable non numérique | Valeurs isolées | Valeurs regroupées en classes | Boîtes à moustaches multiples

Intervalles

n°	Borne inférieure	Borne supérieure	Effectifs (entiers)
1	0	15	15
2	15	25	20
3	25	35	50
4	35	40	30
5	40	45	35
6	45	55	25
7	55	65	15
8	65	75	10
9			
10			
11			
12			
13			
...			
	Total		200

carreau(x) = 1,0 %

☒ Dessin réalisé avec les pourcentages
☐ Dessin réalisé avec les effectifs

Style du graphique

☒ Histogramme ☒ Afficher les effectifs
☐ Diagramme en boîte (ou boîte à moustaches)
☐ Diagramme en boîte avec déciles
☐ Fréquences cumulées ☒ Croissantes ☐ Décroissantes
☐ Droite de Henry

Calculs :

Moyenne	37,375	1er décile	17,5
Écart type	15,0619	1er quartile	28
Effectif total	200	Médiane	37,5
Minimum	0	3ème quartile	45
Maximum	75	9ème décile	58,3333

☒ Visualiser les paramètres ☐ Médiane seulement
☒ Médiane et quartiles ☐ Tous les paramètres

Titre du graphique : Police (Times New Roman) Effacer tout

☒ Dessiner la série statistique Aide OK

