Statistique à une variable

I. Vocabulaire et exemples

En statistiques, on étudie un **caractère** d'un échantillon d'une population.
Ce caractère peut être :

- qualitatif (exemple : couleur, opinion politique...)
- quantitatif discret (exemple : nombre d'enfants d'une famille)
- quantitatif continu (taille, durée, masse ...)

Nous étudierons deux exemples.

Exemple 1 :
On considère la série de notes suivante relative à un devoir de mathématiques dans une classe :

$$7 ; 8 ; 12 ; 7 ; 12 ; 10 ; 8 ; 7 ; 10 ; 10 ; 7 ; 8 ; 7.$$

Valeurs prises par le caractère « note » : $x_1 = 7$, $x_2 = 8$, $x_3 = 10$, $x_4 = 12$.

Tableau d'effectifs, d'effectifs cumulés, de fréquences et de fréquences cumulées :

Notes x_i	7	8	10	12
Effectifs n_i	5	3	3	2
Effectifs cumulés croissants	5	8	11	13
Effectifs cumulés décroissants	13	8	5	2
Fréquences f_i	$\frac{5}{13}$	$\frac{3}{13}$	$\frac{3}{13}$	$\frac{2}{13}$
Fréquences cumulées croissantes	$\frac{5}{13}$	$\frac{8}{13}$	$\frac{11}{13}$	1
Fréquences cumulées décroissantes	1	$\frac{8}{13}$	$\frac{5}{13}$	$\frac{2}{13}$

L'effectif total est $N = \sum n_i = 13$.
La **fréquence** d'une valeur x_i est $\boxed{f_i = \frac{n_i}{N}}$.

Le **mode** est la valeur du caractère ayant le plus grand effectif : ici, c'est la note 7.
D'après les effectifs cumulés, on peut dire qu'il y a 11 notes inférieures ou égales à 10 et 8 notes supérieures ou égales à 8.

Exemple 2 :
Voici un relevé de la durée des d'appels d'une certaine personne au cours d'un mois.

Durée de l'appel (en minutes)	[0 ; 20[	[20 ; 40[	[40 ; 60[	[60 ; 90]
Effectif	16	28	17	9

Les valeurs du caractère « durée des appels » sont regroupées en intervalles appelés **classes**.
L'effectif total (nombre total d'appels) est de 70.
La **classe modale** est ici la classe [20 ; 40[.

II. Représentations graphiques

1) Diagramme en bâtons et en barres

Pour le premier exemple :

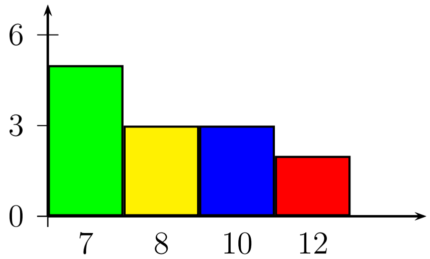


Diagramme en barres.

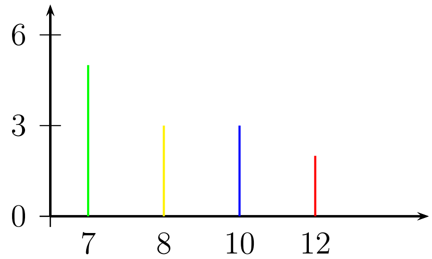
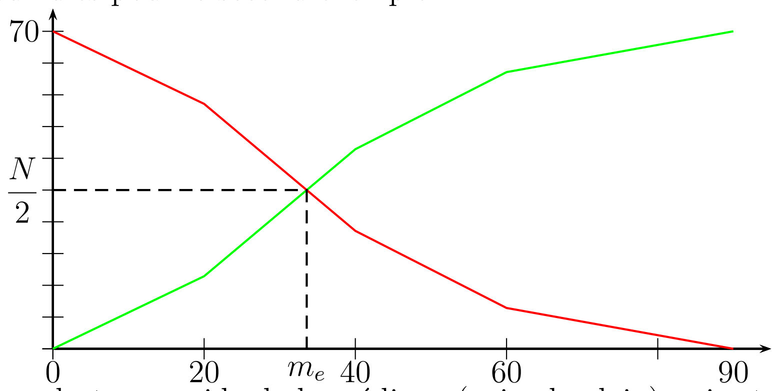


Diagramme en bâtons.

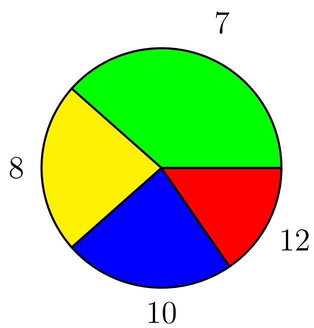
2) Polygônes

Polygônes des effectifs cumulés pour le second exemple.



Ce diagramme permet une lecture rapide de la médiane (voir plus loin) qui est l'abscisse du point d'intersection des polygones des effectifs cumulés croissants et décroissants.

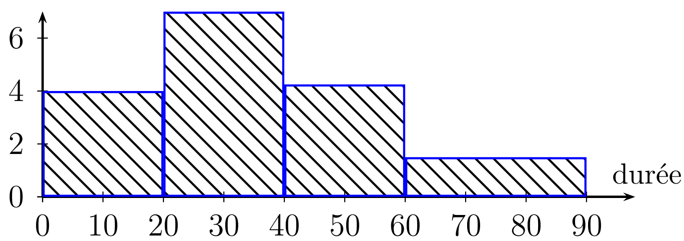
3) Diagramme circulaire



Le diagramme circulaire s'obtient en convertissant les fréquences (ou les effectifs) en angles en utilisant un tableau de proportionalité (100 % correspondent à 360°).

4) Histogramme

Exemple 2 représenté sous forme d'*histogramme*.



Les aires des rectangles sont proportionnelles aux effectifs (ou aux fréquences). On choisit un coefficient de proportionalité entre les effectifs et les aires, par exemple, 5. Ainsi, pour la première classe, l'effectif est de 16, l'aire est donc $16 \times 4 = 64$ et, comme la base est de 20, la hauteur est de 3,2.

Durée de l'appel (en minutes)	[0 ; 20[	[20 ; 40[	[40 ; 60[	[60 ; 90]
Effectif	16	28	17	9
Aire	80	140	85	45
Hauteur du rectangle	4	7	4,25	1,5

Remarques :
- Quand les classes ont toutes la même amplitude alors l'histogramme revient à un diagramme en barres.
- Ne pas mettre de légende sur l'axe des ordonnées.

III. Caractéristiques d'une série

1) Caractéristiques de position

a) Moyenne

La **moyenne arithmétique**, notée $\bar{x}$ se calcule de la façon suivante :

$$\boxed{\bar{x} = \frac{1}{N}\sum n_i x_i = \frac{n_1x_1 + n_2x_2 + n_3x_3 + \dots}{N}}$$

Exemple 1 :
Dans le premier exemple : $\bar{x} = \dfrac{7 \times 5 + 8 \times 3 + 10 \times 3 + 12 \times 2}{13} = \dfrac{113}{13} \simeq 8,69.$

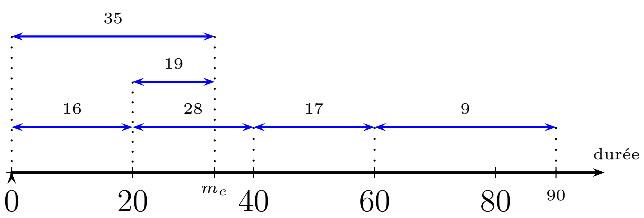
Exemple 2 :
Dans le second exemple, on calcule la moyenne (qui est forcément approximative) en prenant comme valeurs les centres des classes.
On trouve alors : $\bar{x} = \dfrac{10 \times 16 + 30 \times 28 + 50 \times 17 + 75 \times 9}{70} \simeq 36,1$ minutes.

b) Médiane

La **médiane** est la valeur du caractère qui sépare la série en deux séries d'effectifs égaux.

Exemple 1 :
Dans le premier exemple $\dfrac{N}{2} = 6,5$. Il y a 5 ($< 6,5$) notes inférieures ou égales à 7 et 8 ($> 6,5$) notes inférieures ou égales à 8 donc la médiane est égale à 8.

Exemple 2 : Dans le second exemple, la médiane s'obtient par interpolation linéaire.
La moitié de l'effectif est de $\dfrac{N}{2} = 35$. La médiane se trouve donc dans la deuxième classe (car $16 < 35 < 16 + 28$).



$m_e = 20 + \dfrac{19}{28} \times 20 \simeq 33,571$ minutes.

Remarques :
- On peut aussi utiliser un polygone des effectifs cumulés et appliquer le théorème de Thalès.
- On ne prend pas le centre des classes pour le calcul de la médiane !

c) Quartiles

Les trois **quartiles** Q_1, Q_2 et Q_3 sont les valeurs du caractère qui séparent la série en quatre parties de même effectif. Q_2 est la médiane.

> Exemple 1 : Dans le premier exemple, $\dfrac{N}{4} = 3,25$. Le tableau des effectifs cumulés croissants montre alors que $Q_1 = 7$, $Q_2 = m_e = 8$ et $Q_3 = 10$.

> Exemple 2 : Dans le second exemple, on doit procéder à une interpolation linéaire. $\dfrac{N}{4} = 17,5$ donc $Q_1 \in [\,20\,;\,40\,[$. On trouve alors $\dfrac{Q_1 - 20}{20} = \dfrac{17,5 - 16}{28}$ d'où $Q_1 \simeq 21,07$. De même, $\dfrac{3N}{4} = 52,5$ donc $Q_3 \in [\,40\,;\,60\,[$. On trouve alors $\dfrac{Q_3 - 40}{20} = \dfrac{61 - 52,5}{17}$ d'où $Q_3 = 50$.
> On peut aussi faire une lecture graphique des quartiles à partir d'un polygone des effectifs cumulés.

2) Caractéristiques de dispersion

Il s'agit de mesurer la façon dont les valeurs s'éloignent autour de la moyenne.

L'**étendue** est la différence entre la plus grande valeur et la plus petite valeur. Dans l'exemple 1, elle vaut $12 - 7 = 5$ et pour l'exemple 2 elle vaut 90.

L'**intervalle interquartile** est le nombre $Q_3 - Q_1$, qui indique la dispersion de la série autour de sa médiane.

La **variance**, notée V ou V_x est la moyenne des carrés des écarts par rapport à la moyenne :

$$\boxed{V_x = \frac{1}{N}\sum\left(n_i(x_i - \bar{x})^2\right)}$$

et on prouve que :

$$\boxed{V_x = \frac{1}{N}\sum\left(n_i x_i^2\right) - \bar{x}^2}\,.$$

(la variance est aussi la moyenne des carrés moins le carré de la moyenne).
L'**écart type**, noté σ_x est la racine carrée de la variance :

$$\boxed{\sigma_x = \sqrt{V_x}}\,.$$

> Exemple 1 :
> Dans le premier exemple, $V_x \simeq (5 \times 7^2 + 3 \times 8^2 + 3 \times 10^2 + 2 \times 12^2)/13 - 8,69^2 \simeq 3,33$.
> Donc $\sigma_x \simeq 1,825$ (les notes s'écartent en moyenne de 1,825 points de la moyenne 8,69).

> Exemple 2 : Dans le second exemple, on prend les centres des classes pour effectuer le calcul.
> $V_x \simeq (16 \times 10^2 + 28 \times 30^2 + 17 \times 50^2 + 9 \times 75^2)/70 - 36,1^2 \simeq 389,43$. Donc $\sigma_x \simeq 19,73$ minutes.

IV. Utilisation des calculatrices

Entrer les valeurs : Menu Lists (STAT Edit sur Graph25) sur Casio et Menu Stat sur TI
Configurer la machine : Menu Stat Calc Set sur Casio ou STAT Calc Setup sur TI
Calculer la moyenne, etc. : Menu Stat Calc 1var sur Casio ou STAT Calc 1varstats sur TI
Attention à la médiane et aux quartiles des caractères quantitatifs continus (intervalles).