Chapitre 3 : Statistiques

Exercice 1
Lire les effectifs sur un histogramme : ressource 1111

Dans tout le chapitre, on considère une série statistique à caractère quantitatif.
On note l'effectif total N.

I Médiane, quartiles et déciles

I.1 Médiane

Définition (cas discret)
On considère une série à caractère discret dont les valeurs sont rangées dans l'ordre croissant.
On appelle médiane de la série et on note Me tout nombre tel que :
- au moins 50% des individus ont une valeur inférieure à Me,
- et au moins 50% des individus ont une valeur supérieure à Me.

Méthode de détermination
Si l'effectif total N est impair, la médiane est la valeur centrale de la série.
Si l'effectif total N est pair, on choisit pour médiane la demi-somme des deux valeurs centrales.
Plus précisément,
pour N impair, $N = 2k + 1$, alors $Me = x_{k+1}$, c'est-à-dire la $(k+1)^{\text{e}}$ valeur ;
pour N pair, $N = 2k$, alors on prend $Me = \dfrac{x_k + x_{k+1}}{2}$.

Remarque
Si N est pair et que les deux valeurs centrales sont distinctes, alors tout nombre compris strictement entre ces deux valeurs centrales peut être appelé médiane (il n'y a pas unicité).
Dans les autres cas, la médiane est unique.

Exercice 2
1. Médiane à partir des ECC : ressource 811
2. Médiane à partir des FCC : ressource 812
3. Médiane à partir du diagramme des FCC : ressource814

Remarque (Médiane d'une série à caractère continu)
On trace la courbe des fréquences cumulées croissantes.
La médiane Me est l'abscisse du point de la courbe qui a pour ordonnée $\dfrac{1}{2}$.

I.2 Quartiles

Définition
On considère une série à caractère discret dont les valeurs sont rangées dans l'ordre croissant.
Le premier quartile $Q1$ est la plus petite valeur de la série telle qu'au moins 25% des données soient inférieures ou égales à ce nombre.
Le troisième quartile $Q3$ est la plus petite valeur de la série telle qu'au moins 75% des données soient inférieures ou égales à ce nombre.
$[Q1; Q3]$ est l'intervalle interquartile, $Q3 - Q1$ est l'écart interquartile.

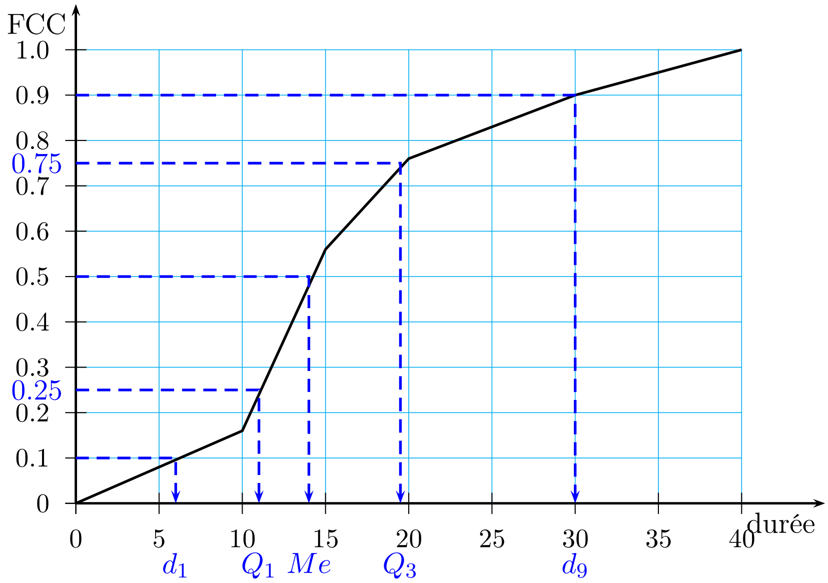
Exercice 3

1. Médiane et quartiles à partir des ECC : ressource 3896
2. Médiane et quartiles à patir des FCC : ressource 3897

I.3 Déciles

Définition
On considère une série à caractère discret dont les valeurs sont rangées dans l'ordre croissant.
Le premier décile d_1 est la plus petite valeur de la série telle qu'au moins 10% des données soient inférieures ou égales à ce nombre.
Le neuvième décile d_9 est la plus petite valeur de la série telle qu'au moins 90% des données soient inférieures ou égales à ce nombre.

Remarque
Dans le cas continu, on procède pour les quartiles et les déciles comme pour la médiane.



Exercice 4

1. Extrema et quantiles à partir des effectifs : ressource 60
2. Extrema et quantiles à partir des effectifs cumulés croissants : ressource 1458
3. Extrema et quantiles à partir des fréquences : ressource 1457

I.4 Diagramme en boîte

Ce type de diagramme permet de résumer les principaux indicateurs de la série.

Exercice 5

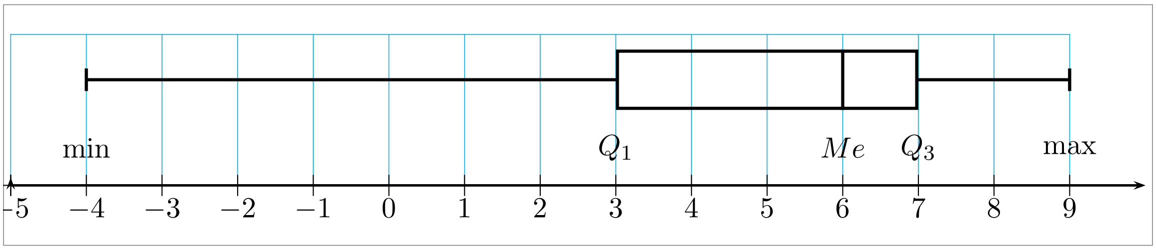
Voici un relevé des températures durant le mois de février :

Température observée	-4	-3	-2	0	2	3	4	5	6	7	8	9
Nombre de jours	1	2	1	1	1	1	2	3	4	5	4	3
Effectifs cumulés croissants												

1. Compléter les effectifs cumulés croissants.
2. L'effectif total est $N = \dots$
3. $\dfrac{N}{4} = \dots$, donc Q_1 est la …$^{\text{e}}$ valeur : $Q_1 = \dots$.
4. $\dfrac{3N}{4} = \dots$, donc Q_3 est la …$^{\text{e}}$ valeur : $Q_3 = \dots$.
5. Détermination de la médiane Me

   ……………………………………………………………………………………
   ……………………………………………………………………………………
   ……………………………………………………………………………………

On résume la série par le diagramme en boîte (ou boîte à moustache) suivant :



Cette série a pour valeur minimale -4, pour maximum 9.
Le premier quartile est 3, la médiane est 6, et le troisième quartile est 7.

Remarque

On fait parfois figurer en plus la valeur du premier décile d_1 et celle du neuvième d_9 sur le diagramme en boîte.
Pour construire le diagramme en boîte, toujours commencer par l'axe gradué !

Exercice 6

1. Lire un diagramme en boîte : ressource 3517
2. Construire un diagramme en boîte à partir des FCC : ressource 3519
3. Construire un diagramme en boîte à partir des ECC : ressource 3520

II Moyenne et écart-type

Notation : On s'intéresse à une série statistique de valeurs x_i $(1 \leqslant i \leqslant r)$ qui apparaissent chacune n_i fois (n_i est l'effectif de x_i).
L'effectif total de la série est $N = n_1 + n_2 + \cdots + n_r$.
La fréquence de la valeur x_i est le nombre $f_i = \dfrac{n_i}{N}$.
On a $f_1 + f_2 + \cdots + f_r = 1$.

II.1 Moyenne

Définition
La moyenne de la série statistique de valeurs et fréquences (x_1, f_1), $\cdots$, (x_r, f_r) est le nombre noté $\bar{x}$ défini par :

$$\bar{x} = f_1x_1 + f_2x_2 + \cdots + f_rx_r = \sum_{i=1}^{r} f_ix_i$$

Remarque

1. Dans le cas où la série statistique est donnée par les valeurs x_i et effectifs n_i, la moyenne $\bar{x}$ s'obtient par :

$$\bar{x} = \frac{1}{N}\left(n_1x_1 + n_2x_2 + \cdots + n_rx_r\right) = \frac{1}{N}\sum_{i=1}^{r} n_ix_i$$

   où bien sûr N désigne l'effectif total, et donc $N = n_1 + n_2 + \cdots + n_r$.

2. Dans le cas où les données de la série statistique sont regroupées en classes, les formules précédentes s'appliquent en prenant pour valeurs x_1, $\ldots$, x_r les centres des classes. On a alors une estimation de la moyenne.

Exercice 7

1. Calcul de moyenne : ressource 819
2. Moyenne d'une série à caractère continu : ressource 816

II.2 Écart-type et variance

Définition
La variance de la série statistique est le nombre positif V défini par :

$$V = \frac{n_1(\bar{x} - x_1)^2 + n_2(\bar{x} - x_2)^2 + \cdots + n_r(\bar{x} - x_r)^2}{N} = \frac{1}{N}\sum_{i=1}^{r} n_i(\bar{x} - x_i)^2$$

Autrement dit, la variance est la moyenne des carrés des écarts la moyenne.

Remarque
La variance d'une série est un indicateur sur la dispersion des valeurs de la série.
Plus précisément, elle informe sur l'éloignement des valeurs par rapport à la moyenne.

Propriété

$$V = \sum_{i=1}^{r} f_i(\bar{x} - x_i)^2$$

Démonstration

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

Comme $f_i = \dfrac{n_i}{N}$, on a bien $V = \displaystyle\sum_{i=1}^{r} f_i(\bar{x}-x_i)^2$. □

Définition
On appelle fonction dispersion associée à la série statistique la fonction d définie par :

$$d(x) = \sum_{i=1}^{r} f_i(x-x_i)^2$$

Théorème
La fonction dispersion admet un minimum lorsque la variable x est en $\bar{x}$, et ce minimum est la variance V de la série statistique.

Démonstration
La fonction d est dérivable de dérivée :
$d'(x) = 2f_1(x-x_1) + \cdots + 2f_r(x-x_r) = 2(f_1+\cdots+f_r)x + -2(f_1x_1+\cdots+f_rx_r)$.
Comme $f_1+\cdots+f_r = 1$, il vient :

$$d'(x) = 2(x-\bar{x})$$

En en déduit facilement les variations de la fonction d.

x	$-\infty$		$\bar{x}$		$+\infty$
d'		$-$	0	$+$	
d	$+\infty$	$\searrow$	V	$\nearrow$	$+\infty$

Par définition, puisque $d(x) = \displaystyle\sum_{i=1}^{r} f_i(x-x_i)^2$, la valeur prise en $\bar{x}$ est

$$d(\bar{x}) = \sum_{i=1}^{r} f_i(\bar{x}-x_i)^2 = V$$

On peut aussi démontrer ce résultat en utilisant seulement le cours sur le second degré : la fonction dispersion d est une fonction du second degré.

En développant, on montre que $d(x) = x^2 - 2x\bar{x} + \displaystyle\sum_{i=1}^{r} f_i{x_i}^2$.

On en déduit que la parabole est tournée vers le haut ($a = 1 > 0$), et que l'abscisse du sommet est $\bar{x}$. □

Remarque
On peut aussi s'intéresser à la fonction $F(x) = \displaystyle\sum_{i=1}^{r} |x-x_i|$ (autre dispersion).

Cette fonction admet un minimum, qui est atteint lorsque $x = Me$, la médiane de la série.
La moyenne ne minimise pas $F(x)$.

Propriété (Formule de Huyghens-König)

$$V = \left(\sum_{i=1}^{r} f_i {x_i}^2\right) \; - \; \bar{x}^2$$

La variance est la moyenne des carrés diminuée du carré de la moyenne.

Démonstration
Pour alléger l'écriture, on supprime les indices. Les sommes vont toutes de $i = 1$ à $i = r$.

$$\begin{aligned} V &= \sum f_i (\bar{x} - x_i)^2 \\ &= \sum f_i \left(\bar{x}^2 - 2\bar{x}x_i + {x_i}^2\right) \\ &= \sum f_i \bar{x}^2 - 2\sum f_i \bar{x} x_i + \sum f_i {x_i}^2 \\ &= \bar{x}^2 \sum f_i - 2\bar{x} \sum f_i x_i + \sum f_i {x_i}^2 \\ &= \bar{x}^2 - 2\bar{x} \times \bar{x} + \sum f_i {x_i}^2 \\ &= \sum f_i {x_i}^2 - \bar{x}^2 \end{aligned}$$

Définition
L'écart-type de la série est le nombre noté s et défini par $s = \sqrt{V}$.

Exercice 8
Calculer un écart-type :
ressource 3066
ressource 3069

Exercice 9
Voici un tableau présentant les salaires mensuels dans une entreprise.

Salaire	1200	1650	2100	2400	6500
Effectif	5	8	4	3	1
Effectifs cumulés croissants					

1. Compléter les effectifs cumulés.
2. Donner l'effectif total.
3. Quel est le mode ?
4. Quel est le salaire moyen ?
5. Calculer l'écart-type de la série.
6. Quel est le salaire médian (la médiane de la série) ?
7. Déterminer l'intervalle interquartiles.
8. Reprendre les questions 4 à 7 lorsqu'on remplace la plus grande valeur par 150 000. Comparer

Remarque
La médiane ne dépend pas des valeurs extrêmes (ni les quartiles).
La moyenne est sensible aux valeurs extrêmes.

III Utilisation de la calculatrice

Pour les calculatrices Texas (82-83).

1. Entrer les données :
   `Stats` puis le menu `EDIT`, et `Edite`.
   Entrer les valeurs x_i dans la liste L_1, les effectifs n_i dans la liste L_2.
   Pour effacer, utiliser la touche `annul`. On peut effacer les valeurs une à une ou toute une liste.
2. Obtenir les indicateurs de la série :
   `Stats`, puis le menu `CALC`, `Stats 1-var`, et L_1 `,` L_2.
   Attention à bien taper la virgule entre L_1 et L_2.

Voir aussi à la fin du livre : page 353 (TI), page 357(Casio).

IV Complément : transformation de données

Théorème
Soient a et b deux réels.
Soient x_1, …, x_r les valeurs d'une série statistique X.
Notons pour cette série X : m sa moyenne, V sa variance, s son écart-type, et I son écart interquartile.
On considère alors la série statistique $aX + b$ dont les valeurs sont $ax_1 + b$, …, $ax_r + b$, et on note m' sa moyenne, V' sa variance, s' son écart-type, et I' son écart interquartile.
Alors : $m' = am + b$, $V' = a^2V$, et $s' = |a|s$.
Si, de plus, $a > 0$, $I' = aI$.

Démonstration
Pour alléger la rédaction, on ne note pas les indices des sommes. Toutes les sommes vont de l'indice $i = 1$ à l'indice $i = r$.
On a $m = \sum f_i x_i$.

$$\begin{aligned} m' &= \sum f_i(ax_i + b) \\ &= \sum af_ix_i + \sum f_ib \\ &= a\sum f_ix_i + b\sum f_i \\ &= am + b \end{aligned}$$

Pour le calcul de la variance, on utilise l'expression $V = \sum f_i(x_i - m)^2$.

$$\begin{aligned} V' &= \sum f_i(ax_i + b - m')^2 \\ &= \sum f_i(ax_i + b - am - b)^2 \\ &= \sum f_i[a(x_i - m)]^2 \\ &= a^2\sum f_i(x_i - m)^2 \\ &= a^2V \end{aligned}$$

Il s'ensuit directement que $s' = \sqrt{V'} = \sqrt{a^2V} = |a|\sqrt{V} = |a|s$. □