1 Chapitre Cours

A Série statistique à deux variables

1 Étude de deux exemples

Exemple 1

Énoncé

Le tableau suivant donne la moyenne y des maxima de tension artérielle en fonction de l'âge x d'une population donnée.

Âge (x_i)	36	42	48	54	60	66
Tension (y_i)	12	13,5	12,6	14,3	15,4	15

Partie A *Étude de la série statistique à* ***une variable y***

1. Donner l'intervalle médian de cette série statistique y. En déduire une valeur pour la médiane Med.

2. Lire sur la calculatrice la médiane Med ainsi que les quartiles Q_1 et Q_3. Donner l'intervalle interquartile et l'écart interquartile.

Représenter la série y par un diagramme en boîte.

3. Calculer la moyenne $\overline{y}$ de la série y.

4. On veut calculer la variance et l'écart type de la série y.

Méthode 1

▶ On sait que : $V(y) = \frac{1}{6}\left(\sum y_i^2\right) - (\overline{y})^2$.

$S(y) = \sqrt{V(y)}$.

Méthode 2

▶ On utilise la calculatrice.

Vérifier que les deux méthodes donnent les mêmes résultats.

Partie B *Étude de la* ***double série statistique*** $(x ; y)$

1. Représenter graphiquement le nuage des six points $M_i(x_i ; y_i)$ dans un repère orthogonal.

On prendra pour unités graphiques :

▶ 0,5 cm pour 1 cm en abscisse ;

▶ 3 cm pour l'unité de tension artérielle en ordonnée.

On placera l'origine au point $K(34 ; 11)$.

2. Déterminer les coordonnées du point G qui est le point moyen du nuage.

3. Les six points M_i forment un nuage ayant une forme « allongée rectiligne ». La droite (M_1M_6) semble passer « assez près » des six points du nuage. Déterminer l'équation de cette droite sous la forme $y = ax + b$ et la tracer.

Quelle tension artérielle peut-on prévoir pour une personne de 78 ans ?

4. On partage le nuage en deux sous-nuages de trois points (les 3 premiers et les 3 derniers). Déterminer le point moyen G_1 du premier sous-nuage et le point moyen G_2 du second sous-nuage. Déterminer une équation de la droite (G_1G_2) sous la forme $y = mx + p$ (on prendra pour m et p des valeurs arrondies à 2 décimales).

Tracer (G_1G_2) et vérifier qu'elle passe par le point G.
Quelle tension artérielle peut-on prévoir pour une personne de 78 ans ?

Solution

Partie A

❶ Quand on détermine un intervalle médian, ainsi qu'une médiane, il faut classer les valeurs dans l'ordre croissant.

Cela donne :

y_i ↗	12	12,6	13,5	14,3	15	15,4

[13,5 ; 14,3]

Comme la série comporte 6 valeurs l'intervalle médian est l'intervalle dont les extrémités sont les troisième et quatrième valeurs, celles-ci étant classées dans l'ordre croissant.

L'intervalle médian de la série y est [13,5 ; 14,3].

Par convention on choisit comme médiane le centre de l'intervalle médian, c'est-à-dire $\frac{1}{2}(13{,}5 + 14{,}3)$.

La médiane Med de la série y est Med = 13,9.

❷ ▶ Sur une TI 82 on peut obtenir les quartiles et la médiane de deux manières différentes :

- faire afficher successivement les 3 valeurs ;
- utiliser le diagramme en boîte.

On donne d'abord les résultats, on montrera ensuite comment les obtenir.

La calculatrice donne $Q_1 = 12{,}6$; Med = 13,9 ; $Q_3 = 15$.

L'intervalle interquartile est $[Q_1 ; Q_3] = [12{,}6 ; 15]$.

L'écart interquartile est $Q_3 - Q_1 = 2{,}4$.

Représentons la série y par un diagramme en boîte (voir **figure 1**).

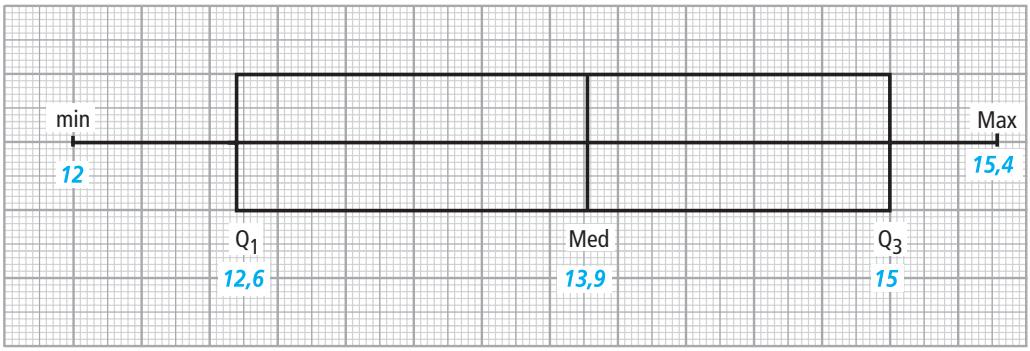


Fig. 1

▶ On peut aussi trouver les valeurs Q_1, Med, Q_3 sans la calculatrice.

Pour Q_1 : on divise le nombre n de valeurs par 4.

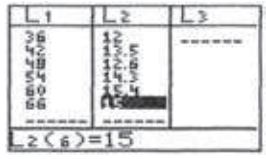
- comme $\frac{6}{4} = 1{,}5$ on prend pour Q_1 la seconde valeur.

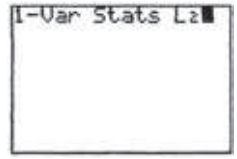
Pour Q_3 : on divise n par 4 et on multiplie par 3.

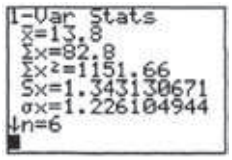
- comme $\frac{6}{4} \times 3 = 4{,}5$ on prend pour Q_3 la cinquième valeur.

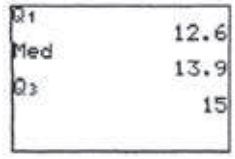
Pour Med : on prend le centre de l'intervalle médian.

Calcul des quartiles Q_1 et Q_3 et calcul de la médiane Med









▶ On peut commencer par vider les listes (éventuellement les 6 listes) à l'aide de la touche STAT :

Faire : STAT 4 2nd 1 , 2nd 2 , ... 2nd 6 ENTER

▶ On rentre la liste des années en L_1 et la liste des tensions en L_2 :

Faire : STAT ENTER 36 ENTER 42 ENTER 48 ENTER 54 ENTER 60 ENTER
66 ENTER ▶ 12 ENTER 13.5 ENTER 12.6 ENTER 14.3 ENTER 15.4 ENTER 15 ENTER

▶ On va calculer le premier quartile de la liste L_2, noté Q_1.

Faire : STAT ▶ 1 2nd 2 ENTER VARS 5 ▶ ▶ ▶ 1 ENTER pour obtenir $Q_1 = 12{,}6$

▶ On peut de même calculer la médiane Med et le troisième quartile Q_3.

Faire : VARS 5 ▶ ▶ ▶ 2 ENTER pour obtenir Med = 13,9

Faire : VARS 5 ▶ ▶ ▶ 3 ENTER pour obtenir $Q_3 = 15$.

Le diagramme en boîte (ou encore boîte à pattes, **B à P** en abrégé)

Faire : 2nd Y= ENTER ENTER ▼ ▶ ▶ ENTER ▼ ▶ ENTER ENTER

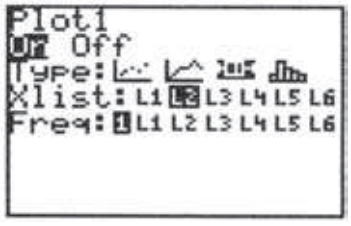
Il faut bien sûr se placer sur On, choisir le logo de la boîte dans Type, se placer sur L_2 dans Xlist et ensuite sur 1 dans Freq.

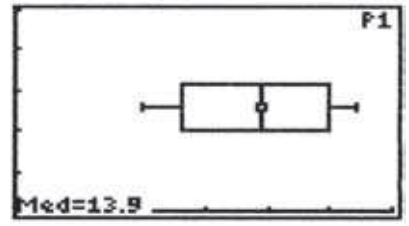
▶ Choix de la fenêtre : choisir une fenêtre convenable

Attention

Ici les valeurs X min et X max sont en réalité les valeurs y des tensions artérielles car c'est le diagramme en boîte de la série y que l'on veut.

▶ Tracé de la boîte : on fait TRACE. En se déplaçant à l'aide des flèches ▶ et ◀ on peut lire min X = 12 ; $Q_1 = 12{,}6$; Med = 13,9 ; $Q_3 = 15$ et max X = 15,4.





❸ Le calcul de la moyenne $\bar{y}$ peut se faire « à la main » ou directement sur la calculatrice.

▶ « à la main » : $\bar{y} = \frac{1}{6}\sum y_i = \frac{1}{6}(12 + 12{,}6 + \ldots + 15, 4) = \frac{1}{6}(82{,}8)$.

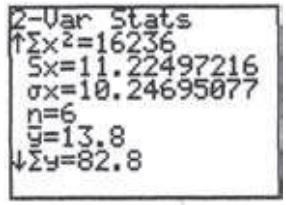
$\bar{y} = 13{,}8$.

▶ sur la TI 82 :

Faire : STAT ▶ 2 2nd 1 , 2nd 2 ENTER ▼ ▼

On peut lire la moyenne $\bar{y} = 13{,}8$ des valeurs de y ainsi que la somme $\sum y = 82{,}8$ des 6 valeurs de y.

À l'aide des flèches ▲ et ▼ on peut monter ou descendre dans ce tableau et lire aussi la moyenne $\bar{x} = 51$ des « x » ainsi que leur somme $\sum x = 306$.



❹ Calcul de la variance et de l'écart type de la série y.

Méthode 1

▶ $\sum y^2 = 12^2 + 13{,}5^2 + \ldots + 15^2 = 1\,151{,}66$.

D'où $$V(y) = \frac{1}{6}\left(\sum y_i^2\right) - (\bar{y})^2$$

$$V(y) = \frac{1}{6}(1\,151{,}66) - 13{,}8^2$$

$$V(y) = 1{,}503\,3\ldots.$$

et $$s(y) = \sqrt{V(y)} = 1{,}226\,1\ldots .$$

En prenant des valeurs arrondies on trouve :

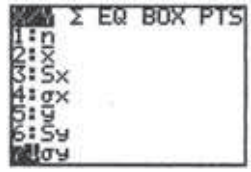
$$\boxed{V(y) = 1{,}503 \text{ et } s(y) = 1{,}226}$$

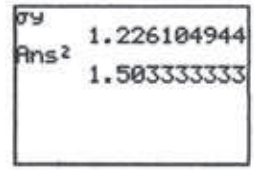
Méthode 2

▶ On utilise la calculatrice.

On calcule d'abord l'écart type : VARS 5 7 ENTER et on trouve $s(y) = 1{,}226$.

On calcule ensuite le carré de l'écart type pour trouver la variance : 2nd (–) x^2 ENTER soit $V(y) = 1{,}503$.





Attention

Maintenant **l'écart type** d'une **série statistique** se note **s** et non plus σ.

Mais sur les calculatrices il y a deux écarts types, l'un noté s_y et l'autre σ_y . On prend la valeur notée σ_y mais on l'appelle s_y . Ne pas prendre la valeur s_y de la calculatrice.

Partie B

❶ Le nuage des 6 points est représenté sur la **figure 2**.

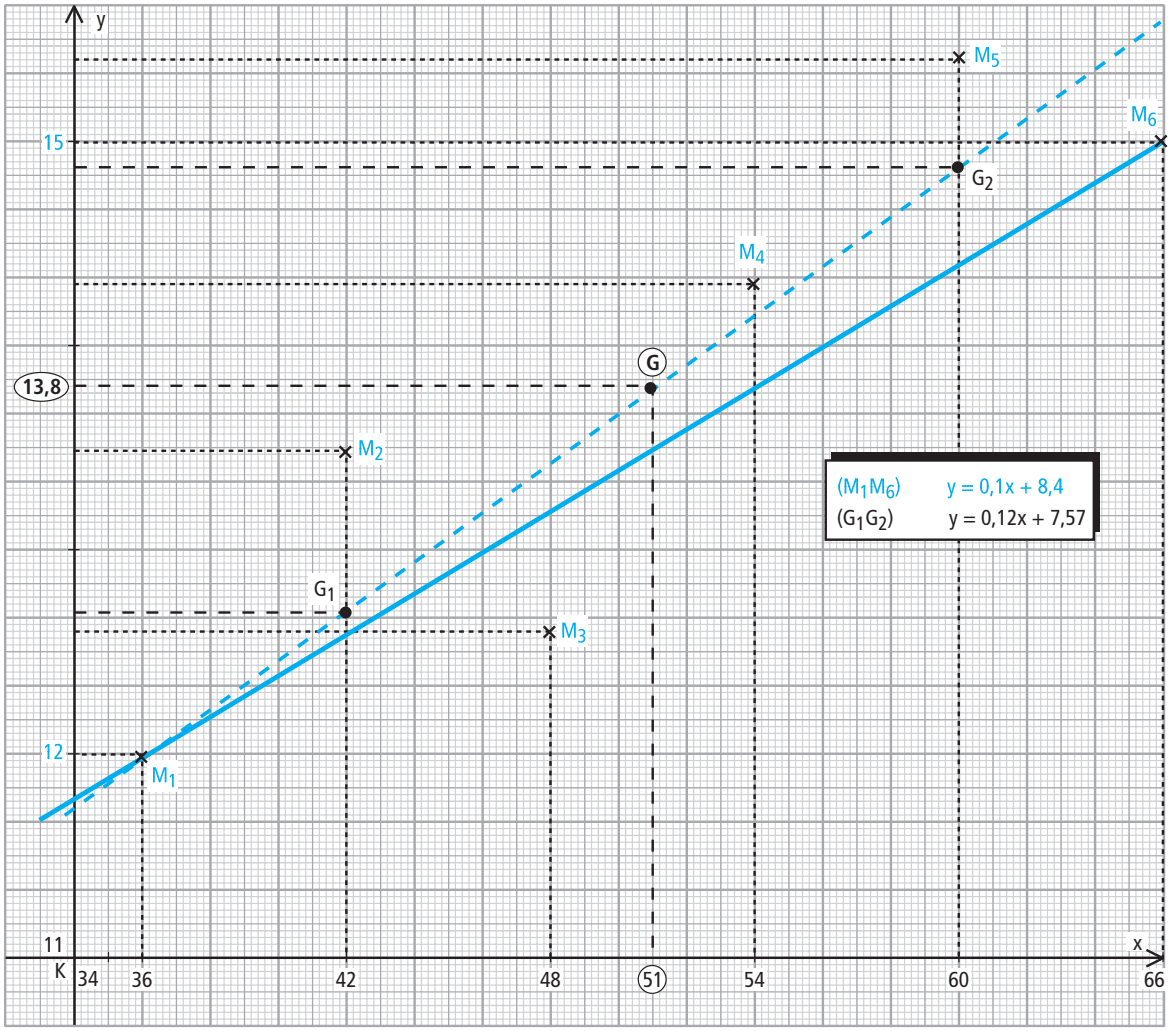


Fig. 2

❷ Le point moyen G a pour coordonnées $(\overline{x} ; \overline{y})$.

D'où $\boxed{G(51 ; 13{,}8)}$.

❸ La droite (M_1M_6) a une équation de la forme $y = ax + b$.

On a : $a = \dfrac{y_6 - y_1}{x_6 - x_1} = \dfrac{15 - 12}{66 - 36} = \dfrac{3}{30} = \dfrac{1}{10} = 0{,}1$.

Cela donne : $y = 0{,}1x + b$.

En $M_1(36 ; 12)$ on peut écrire $12 = 0{,}1 \times 36 + b$ soit $b = 8{,}4$.

La droite (M_1M_6) a pour équation $y = 0{,}1x + 8{,}4$.

On va utiliser l'équation de la droite (M_1M_6) pour prévoir quelle peut être la tension artérielle d'une personne de 78 ans.

Pour $x = 78$ on obtient $y = 7{,}8 + 8{,}4 = 16{,}2$.

On peut estimer la tension artérielle d'une personne de 78 ans à 16,2.

❹ Le point G_1 a pour abscisse $\dfrac{1}{3}(36 + 42 + 48) = 42$.

Le point G_1 a pour ordonnée $\frac{1}{3}(12 + 13{,}5 + 12{,}6) = 12{,}7$.

Le point G_2 a pour abscisse $\frac{1}{3}(54 + 60 + 66) = 60$.

Le point G_2 a pour ordonnée $\frac{1}{3}(14{,}3 + 15{,}4 + 15) = 14{,}9$.

On a donc $G_1(42 ; 12{,}7)$ et $G_2(60 ; 14{,}9)$.

La droite (G_1G_2) a pour équation $y = mx + p$.

On a : $m = \frac{14{,}9 - 12{,}7}{60 - 42} = \frac{2{,}2}{18} = \frac{11}{90} = 0{,}122\,2...$

Cela donne : $y = \frac{11}{90}x + p$.

En $G_2(60 ; 14{,}9)$ on peut écrire : $14{,}9 = \frac{11 \times 60}{90} + p$ soit $p = \frac{22{,}7}{3}$.

En prenant pour m et p des valeurs arrondies à 2 décimales on obtient $y = 0{,}12x + 7{,}57$.

Une équation de la droite (G_1G_2) est $y = 0{,}12x + 7{,}57$.

Pour $x = 78$ on obtient $y = 0{,}12 \times 78 + 7{,}57 = 16{,}93...$

On peut estimer la tension artérielle d'une personne de 78 ans à 16,9.

Montrons que les coordonnées de G vérifient l'équation de (G_1G_2).

Pour $x = 51$ on a $y = \frac{11}{90} \times 51 + \frac{22{,}7}{3} = 13{,}8$.

Ceci prouve que la droite (G_1G_2) passe par le point moyen G.

Remarque *On note une différence assez sensible égale à environ 0,75 unité. Il faut avoir bien présent à l'esprit que ce ne sont que des estimations.*

Exemple ❷

Énoncé

Lors d'une période de sécheresse, un agriculteur relève la quantité totale d'eau, exprimée en m^3, utilisée dans son exploitation depuis le premier jour. On obtient les résultats suivants :

nombre de jours écoulés : x_i	1	3	5	8	10
volume utilisé (en m^3) : y_i	2,25	4,3	8	17,5	27

Le plan est muni d'un repère orthogonal. On prendra pour unités graphiques : sur l'axe des abscisses 1 cm pour un jour et sur l'axe des ordonnées 0,5 cm pour 1 m^3.

❶ Représenter graphiquement la série statistique $(x_i ; y_i)$.

❷ La calculatrice donne l'équation d'une droite Δ qui est la droite de régression de y en x. Cette droite Δ a pour équation $y = 2{,}75x - 3{,}04$.

Vérifier que Δ passe par le point moyen du nuage et tracer Δ.

❸ Le nuage de points permet d'envisager un ajustement par la parabole $\mathscr{P}$ passant par les points $A(1 ; 2{,}25)$; $B(10 ; 27)$ et d'équation $y = ax^2 + b$.

a) Déterminer les deux réels a et b et donner l'équation de $\mathscr{P}$.

b) Représenter la parabole $\mathscr{P}$ sur le graphique.

❹ Estimer le volume d'eau utilisé le 12^e jour de sécheresse en utilisant l'équation de la droite Δ puis l'équation de la parabole $\mathscr{P}$.

Lequel des deux résultats paraît le plus vraisemblable ? Pourquoi ?

Solution

❶ Le nuage des 5 points représentant la série statistique $(x_i ; y_i)$ est sur la **figure 3**.

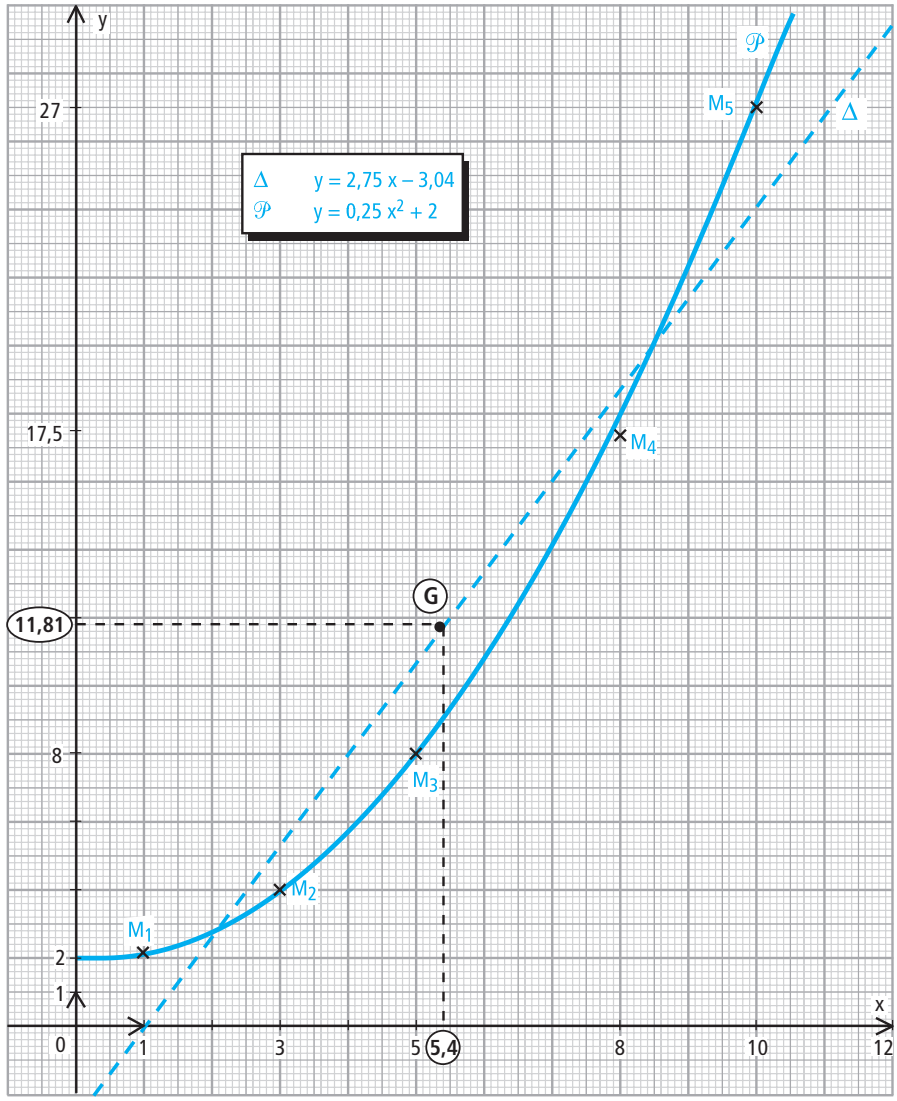


Fig. 3

❷ La droite Δ a pour équation $y = 2{,}75x - 3{,}04$.

Le point moyen G du nuage a pour coordonnées $(\overline{x} ; \overline{y})$.

La calculatrice donne $\overline{x} = 5{,}4$ et $\overline{y} = 11{,}81$.

Pour $x = 5{,}4$ on a $y = 2{,}75 \times 5{,}4 - 3{,}04 = 11{,}81$.

Ceci montre que Δ passe par le point moyen du nuage.

Pour tracer Δ on peut choisir deux autres points.

La droite Δ passe par les points de coordonnées $(3 ; 5{,}21)$ et $(10 ; 24{,}46)$.

Le tracé de la droite Δ est sur la **figure 3**.

❸ **a)** Au point $A(1 ; 2{,}25)$ on obtient $2{,}25 = a + b$.

Au point $B(10 ; 27)$ on obtient $27 = 100a + b$.

Résolvons le système $\begin{cases} 2{,}25 = a + b & [L_1] \\ 27 = 100a + b & [L_2] \end{cases}$.

En faisant $[L_2] - [L_1]$ on obtient : $24{,}75 = 99a$ soit $\mathbf{a = 0{,}25}$.

En reportant la valeur de a dans $[L_1]$ on trouve $\mathbf{b = 2}$.

La parabole $\mathscr{P}$ a pour équation $y = 0{,}25x^2 + 2$.

b) Faisons un tableau de valeurs pour le tracé de $\mathscr{P}$.

x	0	1	2	3	4	5	6	7	8	9	10
$0{,}25x^2 + 2$	2	2,25	3	4,25	6	8,25	11	14,25	18	22,25	27

❹ Estimons le volume d'eau utilisé le 12e jour.

▶ avec la droite Δ.

Pour $x = 12$ on trouve $y = 2{,}75 \times 12 - 3{,}04 = 29{,}962$.

▶ avec la parabole $\mathscr{P}$.

Pour $x = 12$ on trouve $y = 0{,}25 \times 12^2 + 2 = 38$.

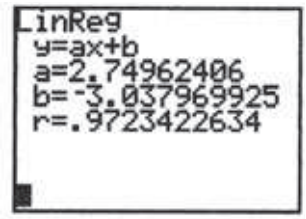
Les deux résultats obtenus sont bien différents.

On voit sur le graphique que la parabole $\mathscr{P}$ passe très près des cinq points du nuage. Cette parabole réalise donc un meilleur ajustement du nuage que la droite Δ. L'estimation $y = 38$ est la plus vraisemblable.

On peut estimer la consommation d'eau à 38 m^3 environ au 12e jour de sécheresse.

Remarque

Détermination, par la calculatrice, de l'équation de la droite de régression Δ :
On met les valeurs de la série x en L_1 *et celles de la série y en* L_2.
Faire : STAT ▶ 5 2nd 1 , 2nd 2 ENTER
On obtient : $a = 2{,}749\,6\ldots$ *et* $b = -3{,}037\,9\ldots$
On peut choisir comme valeurs arrondies $a = 2{,}75$ *et* $b = -3{,}04$.



❷ Nuage des points

Soit $(x_i ; y_i)$ une série statistique à deux variables.

On suppose $1 \leq i \leq n$ ce qui signifie que x_i et y_i peuvent prendre n valeurs.

On munit le plan d'un repère orthogonal dont les unités seront judicieusement choisies.

À chaque couple $(x_i ; y_i)$ on associe le point $M_i(x_i ; y_i)$. C'est l'ensemble de ces n points qui forme un nuage de points.

Définition ❶

Soit $(x_i ; y_i)$ une série statistique à deux variables $(1 \leq i \leq n)$.

L'ensemble des points $M_i(x_i ; y_i)$ constitue **le nuage de points** associé à la série statistique double $(x_i ; y_i)$.

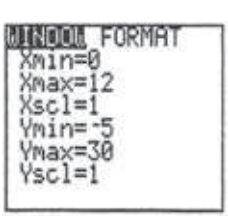
Le plus souvent on représente le nuage de points par des petites croix (x).

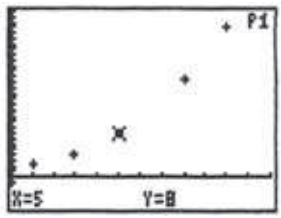
Pour représenter un nuage de points sur la calculatrice

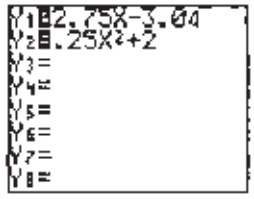
Considérons le nuage formé par les 5 points de l'**exemple ❷**.

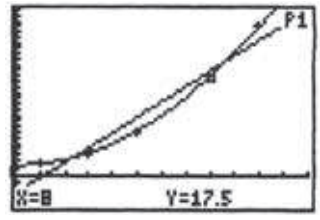
On choisit une fenêtre convenable : WINDOW ▼

Faire ensuite : [2nd] [Y=] [ENTER] [ENTER] [▼] [ENTER] [▼] [ENTER] [▼] [▶] [ENTER] [▼] [▶] [ENTER] [TRACE]









On peut aussi tracer la droite Δ ainsi que la parabole $\mathcal{P}$ dans le même repère que le nuage.

On écrit l'équation de Δ en Y_1 et l'équation de $\mathcal{P}$ en Y_2.

Il suffit alors de faire [TRACE] pour voir les deux courbes ainsi que le nuage de points dans un même repère.

❸ Point moyen

Définition ❷ On appelle **point moyen** d'un nuage de n points $M_i(x_i ; y_i)$ le point G de coordonnées $(\overline{x} ; \overline{y})$ où $\overline{x}$ et $\overline{y}$ sont les moyennes respectives des variables x et y.

$$G(\overline{x} ; \overline{y}) \quad \text{avec} \quad \overline{x} = \frac{1}{n}\sum x_i \quad \text{et} \quad \overline{y} = \frac{1}{n}\sum y_i .$$

B Ajustement affine par moindres carrés

❶ Somme des carrés des résidus

On considère un nuage formé de n points : $M_1(x_1 ; y_1)$, $M_2(x_2 ; y_2)$, ..., $M_n(x_n ; y_n)$.

Soit Δ une droite d'ajustement de ce nuage, d'équation $y = ax + b$.

Définition ❸ On appelle *somme des carrés des résidus*, associée à la droite d'ajustement d'équation $y = ax + b$, le nombre réel positif S défini par :

$$S = [y_1 - (ax_1 + b)]^2 + [y_2 - (ax_2 + b)]^2 + ... + [y_n - (ax_n + b)]^2$$

Notation

On peut encore écrire : $\boxed{S = \sum_{i=1}^{n} [y_i - (ax_i + b)]^2}$.

Exemple ❸ *Énoncé*

On reprend la série statistique double $(x_i ; y_i)$ de l'**exemple ❶**.

❶ Calculer la somme des carrés des résidus correspondant à chacune des deux droites d'ajustement.

❷ En utilisant la fonction « Lin Reg(ax + b) » dans le menu [STAT], option [CALC] d'une TI 82 (voir la remarque de l'**exemple ❷**) on obtient l'équation d'une droite d'ajustement Δ dont l'équation est : $y = 0{,}107x + 8{,}360$.

Calculer, pour la droite Δ, la somme des carrés des résidus.

Comparer le résultat trouvé aux deux précédents.

Solution

Rappel

- l'équation de (M_1M_6) est $y = 0{,}1x + 8{,}4$.
- l'équation de (G_1G_2) est $y = 0{,}12x + 7{,}57$.

▶ On peut présenter les résultats dans un tableau pour la droite (M_1M_6).

x_i	36	42	48	54	60	66	
y_i	12	13,5	12,6	14,3	15,4	15	
$e_i = y_i - (0{,}1x_i + 8{,}4)$	0	0,9	−0,6	0,5	1	0	
$e_i^2 = [y_i - (0{,}1x_i + 8{,}4)]^2$	0	0,81	0,36	0,25	1	0	$\sum = 2{,}42$

La somme des carrés des résidus pour la droite (M_1M_6) est le nombre, noté S_1, tel que $S_1 = 2{,}42$.

▶ Pour la droite (G_1G_2) on va tout faire dans les listes de la calculatrice.

On place les valeurs de la série x_i dans la liste L_1.

On place les valeurs de la série y_i dans la liste L_2.

On écrit l'équation de la droite (G_1G_2) en Y_1 soit $Y_1 = 0{,}12x + 7{,}57$.

On va placer les valeurs de $y_i - (ax_i + b)$ en L_3 et les valeurs de $[y_i - (ax_i + b)]^2$ en L_4.

On aura donc : $L_3 = L_2 - Y_1(L_1)$ et $L_4 = (L_3)^2$.

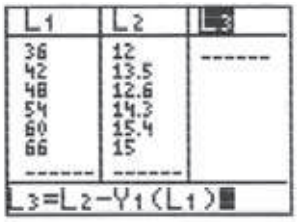
Pour rentrer la liste $L_3 = L_2 - Y_1(L_1)$ on se place avec la touche ▲ sur (L_3).

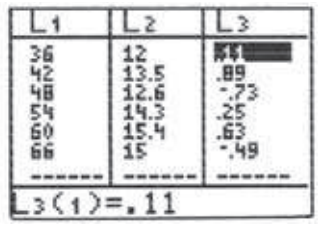
Une fois que l'on est sur (L_3) comme indiqué sur le premier tableau on fait :

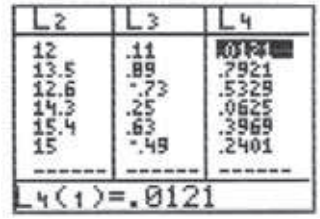
2nd 2 − 2nd VARS 1 1 (2nd 1) ENTER pour obtenir le second tableau.

Pour rentrer la liste $L_4 = (L_3)^2$ on se place avec la touche ▲ sur (L_4).

Une fois que l'on est sur (L_4) on fait : 2nd 3 x^2 ENTER pour obtenir le dernier tableau.

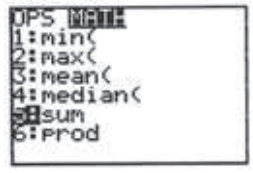


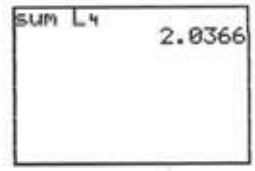




On peut maintenant calculer la somme de tous les nombres de la liste L_4, c'est-à-dire la somme des carrés des résidus correspondant à la droite (G_1G_2).

Faire : 2nd MODE 2nd STAT ▶ 5 2nd 4 ENTER ce qui donne $S_2 = 2{,}036\,6\ldots$





Pour calculer la somme des carrés des résidus correspondant à la droite Δ d'équation $y = 0{,}107x + 8{,}360$ on procède comme pour la droite (G_1G_2).

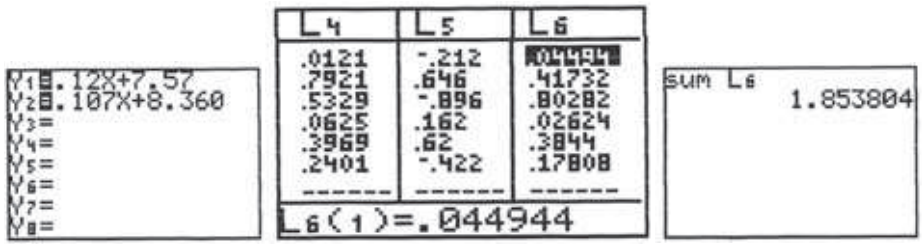
On écrit l'équation de la droite Δ en Y_2 soit $Y_2 = 0{,}107x + 8{,}360$.

Pour rentrer la liste $L_5 = L_2 - Y_2(L_1)$ on se place avec la touche ▲ sur L_5.

Une fois que l'on est sur L_5 on fait : 2nd 2 − 2nd VARS 1 2 (2nd 1) ENTER.

Pour rentrer la liste $L_6 = (L_5)^2$ on se place avec la touche ▲ sur L_6.

Une fois que l'on est sur L_6 on fait : 2nd 5 x^2 ENTER pour obtenir le dernier tableau.



On peut maintenant calculer la somme de tous les nombres de la liste L_6, c'est-à-dire la somme des carrés des résidus correspondant à la droite Δ ce qui donne $S_3 = 1{,}853\ 804$.

Faire : 2nd | MODE | 2nd | STAT | ▶ | 5 | 2nd | 6 | ENTER

Récapitulons les résultats trouvés.

	Pour (M_1M_6)	Pour (G_1G_2)	pour Δ
somme des carrés des résidus	$S_1 = 2{,}42$	$S_2 = 2{,}036\ 6$	$S_3 = 1{,}853\ 8$

La somme des carrés des résidus est minimale pour la droite Δ.

On admet que c'est cette droite Δ qui minimise la somme des carrés des résidus : on l'appelle parfois droite des « moindres carrés ».

❷ Ajustement affine par moindres carrés

Définition de la droite des moindres carrés

Définition ❹

On appelle **droite de régression de y en x** (ou droite d'ajustement), par la méthode des moindres carrés, la droite Δ d'équation $y = ax + b$ qui rend **minimale** la somme $S = \sum_i [y_i - (ax_i + b)]^2$.

Propriété ❶

Parmi toutes les droites d'ajustement possibles d'un nuage, **la droite de régression de y en x** par moindres carrés est la seule à rendre **la somme des carrés des résidus minimale**.

Relation entre les coefficients a et b

Soit $y = ax + b$ l'équation de la droite de régression de y en x, notée Δ.

On admet la propriété suivante.

Propriété ❷

La droite de régression de y en x passe toujours par le point moyen $G(\bar{x} ; \bar{y})$ du nuage.

On peut écrire : $y = ax + b$

d'où $\bar{y} = a\bar{x} + b$ car Δ passe par G.

En soustrayant la deuxième équation de la première on obtient : $y - \bar{y} = ax - a\bar{x} = a(x - \bar{x})$.

Propriété ❸

Soit $y = ax + b$ l'équation de la droite Δ de régression de y en x.

On a alors : $b = \bar{y} - a\bar{x}$.

L'équation de Δ s'écrit : $\mathbf{y - \bar{y} = a(x - \bar{x})}$.

Il ne nous reste plus qu'à trouver une formule donnant a.

Détermination du coefficient directeur a

On admet la formule donnant l'expression de a.

Propriété 4

L'équation de la droite Δ de régression de y en x est de la forme $y - \bar{y} = a(x - \bar{x})$ avec $a = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$.

Autre expression donnant a

On peut écrire $a = \dfrac{\frac{1}{n}\sum (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum (x_i - \bar{x})^2}$.

On reconnaît au dénominateur la variance de x.

Par analogie avec la variance, le numérateur est appelé covariance de x et de y.

Définition 5 La **covariance** de la série double $(x\ ;\ y)$ est le nombre réel défini par :

$$\text{cov}(x\ ;\ y) = \frac{1}{n}\sum (x_i - \bar{x})(y_i - \bar{y}).$$

On sait qu'il existe deux formules donnant la variance d'une série statistique.

De même il existe deux formules donnant la covariance d'une série double $(x\ ;\ y)$.

Propriété 5

La covariance d'une série double $(x\ ;\ y)$ est égale à :

$$\text{cov}(x\ ;\ y) = \frac{1}{n}\sum (x_i - \bar{x})(y_i - \bar{y}) = \left(\frac{1}{n}\sum x_i y_i\right) - \bar{x}\,\bar{y}.$$

Conséquence On a donc $a = \dfrac{\text{cov}(x\ ;\ y)}{V(x)} = \dfrac{\text{cov}(x\ ;\ y)}{s^2(x)}$.

Important

Dans la pratique on ne vous demandera pas de calculer « a » en utilisant l'une ou l'autre de ces formules.

La calculatrice nous donne les coefficients a et b de la droite de régression de y en x obtenue par la méthode des moindres carrés.

L'équation de cette droite est obtenue, sur une TI 82, par la ligne 5 du menu STAT, option CALC.

3 Étude d'un exemple à la calculatrice

Exemple 4

Énoncé

Le tableau suivant donne l'évolution du nombre de spectateurs dans les salles de cinéma en France sur une période de 7 ans.

année	1989	1993	1994	1995	1996
rang (x_i)	0	4	5	6	7
nombre de spectateurs (y_i)	120,9	132,7	124,5	130,2	136,3

Le nombre de spectateurs est exprimé en millions.

1 On peut associer un nuage de points $M_i(x_i ; y_i)$ à la série statistique double $(x ; y)$. Soit G le point moyen de ce nuage.

Déterminer, à la calculatrice, les coordonnées du point G.

2 Donner une équation de la droite de régression Δ de y en x par la méthode des moindres carrés. On donnera cette équation sous la forme $y = ax + b$, les réels a et b étant arrondis à 10^{-3} près.

3 Si l'évolution constatée s'est poursuivie jusqu'à la fin du XX^e siècle, donner une estimation du nombre de spectateurs dans les salles de cinéma en France en l'an 2000.

Solution

1 La calculatrice donne : $\bar{x} = 4{,}4$ et $\bar{y} = 128{,}92$.

Le point moyen G a donc pour coordonnées $(4{,}4 ; 128{,}92)$.

Ainsi $G(4{,}4 ; 128{,}92)$.

2 On cherche l'équation de la droite de régression Δ de y en x sous la forme $y = ax + b$.

La calculatrice donne : $a = 1{,}793\,1...$ et $b = 121{,}030\,1$.

En prenant pour a et b des valeurs arrondies à 10^{-3} près on obtient comme équation de Δ :

$$y = 1{,}793x + 121{,}030$$

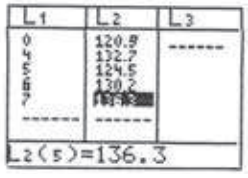
3 Si l'évolution s'est poursuivie on peut estimer le nombre de spectateurs de cinéma en l'an 2000 en utilisant la droite d'ajustement Δ.

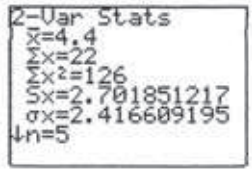
L'an 2000 correspond au rang $x = 11$.

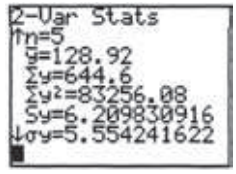
Pour $x = 11$ on a $y = 1{,}793 \times 11 + 121{,}030 = 140{,}753$.

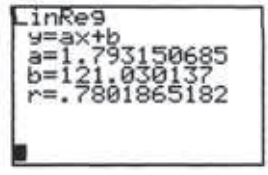
On peut estimer à 140,753 millions le nombre de spectateurs de cinéma en France en l'an 2000 (valeur arrondie 140,8 millions).

Voici les résultats demandés obtenus sur l'écran d'une TI 82.









C Transformation affine des données

Exemple 5

Énoncé

Le tableau suivant donne le taux d'équipement en lave-linge des ménages français de 1955 à 1985.

Année : x_i	1955	1960	1965	1970	1975	1980	1985
Taux en % : y_i	10	25	41	60	69	80	86

Dans tout l'exercice, le détail des calculs n'est pas demandé. Les résultats pourront être obtenus à la calculatrice et seront arrondis à 10^{-3} près.

Partie A

1 Donner les coordonnées $(\bar{x} ; \bar{y})$ du point moyen G.

2 Le plan est muni d'un repère orthogonal $(O ; \vec{i}, \vec{j})$.

Soit Δ la droite de régression de y en x, obtenue par la méthode des moindres carrés, sous la forme $y = ax + b$. Déterminer une équation de la droite Δ.

Partie B Changement d'origine

1. On pose : $X = x - \overline{x}$ et $Y = y - \overline{y}$.

Donner l'équation de la droite Δ dans le nouveau repère $(G ; \vec{i}, \vec{j})$.

2. *Généralisation*

▶ On donne : $y = ax + b$; $X = x - x_0$ et $Y = y - y_0$.

Trouver une relation entre Y et X.

▶ Cas particulier : $y = ax + b$; $X = x - \overline{x}$ et $Y = y - \overline{y}$.

Trouver une relation entre Y et X.

Partie C Changement d'échelle

On pose : $X = x$ et $Y = \dfrac{y}{100}$.

Donner l'équation de la droite Δ dans le nouveau repère dont l'origine est O.

Partie D

1. On pose $X = \dfrac{x - 1955}{5}$ et $Y = y$.

Donner l'équation de la droite Δ dans le nouveau repère.

2. On pose $X = \dfrac{x - 1955}{5}$ et $Y = \dfrac{y}{100}$.

Donner l'équation de la droite Δ dans le nouveau repère.

Solution

Partie A

1. La calculatrice donne : $\boxed{G(1970 ; 53)}$

2. La droite Δ a une équation de la forme $y = ax + b$ dans le repère $(O ; \vec{i}, \vec{j})$.

La calculatrice donne : $a = 2{,}614\,2\ldots$ et $b = -5\,097{,}1\,428\ldots$

D'où $\boxed{y = 2{,}614x - 5\,097{,}143}$.

Partie B

1. On pose : $X = x - 1970$ et $Y = y - 53$.

La calculatrice donne : $\boxed{Y = 2{,}614X}$

2. *Généralisation*

▶ $y = ax + b$; $X = x - x_0$ et $Y = y - y_0$.

On a donc : $x = X + x_0$ et $y = Y + y_0$.

L'équation de Δ s'écrit alors : $Y + y_0 = a(X + x_0) + b$

$$\boxed{Y = aX + ax_0 - y_0 + b}$$

On remarque que le coefficient directeur a est inchangé.

▶ cas particulier : $y = ax + b$; $X = x - \overline{x}$ et $Y = y - \overline{y}$.

Comme la droite de régression passe alors par le point moyen $G(\overline{x} ; \overline{y})$ on a : $\overline{y} = a\overline{x} + b$.

Ainsi $\begin{cases} y = ax + b \\ \overline{y} = a\overline{x} + b \end{cases}$ d'où $y - \overline{y} = a(x - \overline{x})$

$$\boxed{Y = aX}$$

Partie C

On pose $X = x$ et $Y = \dfrac{y}{100}$.

La calculatrice donne : $\boxed{Y = 0{,}026X - 50{,}971}$

On note que le coefficient directeur a été divisé par 100.

Partie D

❶ On pose $X = \dfrac{x - 1955}{5}$ et $Y = y$.

La calculatrice donne : $Y = 13{,}071X + 13{,}786$

❷ On pose $X = \dfrac{x - 1955}{5}$ et $Y = \dfrac{y}{100}$.

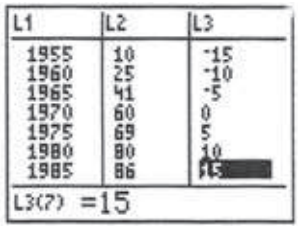
La calculatrice donne : $Y = 0{,}131X + 0{,}138$

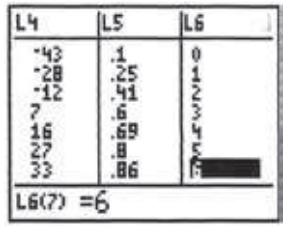
Montrons comment tous ces résultats peuvent être obtenus rapidement à la calculatrice.

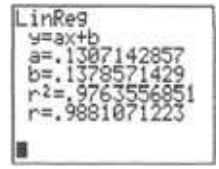
On va utiliser les 6 listes de la calculatrice de la manière suivante :

L_1	L_2	L_3	L_4	L_5	L_6
x	y	$x - 1970$	$y - 53$	$\frac{y}{100}$	$\frac{x - 1955}{5}$
		$L_1 - 1970$	$L_2 - 53$	$\frac{L_2}{100}$	$\frac{L_1 - 1955}{5}$

Lin Reg (ax + b) L_1, L_2	$y = 2{,}614x - 5\ 097{,}143$
Lin Reg (ax + b) L_3, L_4	$Y = 2{,}614X$
Lin Reg (ax + b) L_1, L_5	$Y = 0{,}026X - 50{,}971$
Lin Reg (ax + b) L_6, L_2	$Y = 13{,}071X + 13{,}786$
Lin Reg (ax + b) L_6, L_5	$Y = 0{,}131X + 0{,}138$







Pour obtenir, par exemple, la liste L_6, on fait d'abord ▲ pour se placer sur L_6.

Ensuite : (2nd 1 − 1 9 5 5) ÷ 5 ENTER

Pour obtenir l'équation de la première droite de régression : STAT ▶ 5 2nd 1 , 2nd 2 ENTER

Pour obtenir l'équation de la dernière droite de régression : STAT ▶ 5 2nd 6 , 2nd 5 ENTER

D Adéquation à une loi équirépartie

❶ Diagramme en boîte et déciles

Exemple ❻ *Énoncé*

Le tableau suivant donne le temps de parcours des 32 élèves d'une classe de seconde, correspondant au trajet quotidien : domicile – lycée – domicile (c'est donc un aller-retour).

t : temps de parcours des 32 élèves, exprimés en minutes															
0	0	15	15	15	20	20	25	25	25	30	30	30	30	35	35
40	40	40	40	40	45	45	50	50	60	70	80	80	90	90	110

❶ Calculer le temps total de parcours quotidien pour les 32 élèves. Calculer le temps de parcours moyen.

❷ Déterminer la médiane, le premier et le dernier quartiles de cette série statistique. En déduire l'intervalle interquartile et l'écart interquartile.

❸ Calculer le premier décile et le neuvième décile de cette série statistique. Calculer l'écart interdécile.

Représenter la série par un diagramme en boîte où figurent les premier et neuvième déciles, ainsi que les valeurs extrêmes.

Solution

❶ Désignons la série par t.

La calculatrice donne comme temps total 1 320 min.

Le temps moyen de parcours est $\bar{t} = \frac{1\,320}{32} = 41{,}25$.

Temps total de parcours : 1 320 min soit 22 h exactement.

Temps moyen de parcours : 41,25 min.

❷ Le nombre d'élèves est pair.

Les temps sont rangés dans l'ordre croissant.

On a : $t_{16} = 35$ et $t_{17} = 40$.

La médiane est alors la demi-somme des nombres 35 et 40.

$$\frac{35 + 40}{2} = 37{,}5 .$$

Le premier quartile est le plus petit élément Q_1 des valeurs de t, ordonnées par ordre croissant, tel qu'au moins 25 % des données soient inférieures ou égales à Q_1 .

$$\frac{32}{4} = 8 .$$

Le premier quartile est donc égal à $t_8 = 25 = Q_1$.

Le troisième quartile est le plus petit élément Q_3 des valeurs de t, ordonnées par ordre croissant, tel qu'au moins 75 % des données soient inférieures ou égales à Q_3 .

$$\frac{32 \times 3}{4} = 24 .$$

Le troisième quartile est donc égal à $t_{24} = 50 = Q_3$.

Ainsi $\boxed{\text{Med} = 37{,}5 \ ; \ Q_1 = 25 \ ; \ Q_3 = 50}$.

L'intervalle interquartile est l'intervalle **[25 ; 50]** .

L'écart interquartile est égal à $\mathbf{Q_3 - Q_1 = 25}$.

❸ Pour les déciles on a des définitions analogues à celles des quartiles.

$\frac{32}{10} = 3{,}2$. On prend donc la $4^{\text{ème}}$ valeur de t.

Ainsi $t_4 = 15 = d_1$.

$\frac{32 \times 9}{10} = 28{,}8$. On prend donc la $29^{\text{ème}}$ valeur de t.

Ainsi $t_{29} = 80 = d_9$.

On a donc $\boxed{d_1 = 15 \text{ et } d_9 = 80}$.

Le diagramme en boîte est sur la **figure 4**.

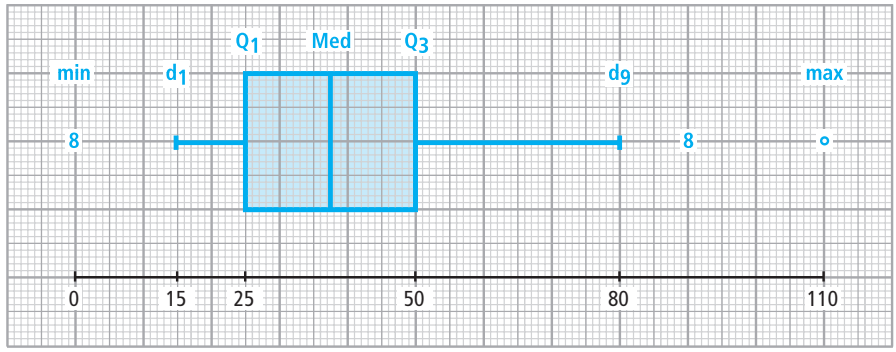


Fig. 4

L'écart interdécile est égal à $d_9 - d_1 = 80 - 15 = 65$.

L'écart interdécile est 65.

Remarque *Les deux valeurs nulles correspondent à deux élèves internes.*

❷ Le jeu de pile ou face

Exemple ❼

Énoncé

Un joueur A lance une pièce de 1 euro : il obtient 450 fois « Pile » et 550 fois « Face ».

Un joueur B lance une pièce de 2 euros : il obtient 520 fois « Pile » et 480 fois « Face ».

Les pièces sont-elles équilibrées ?

Solution

On ne peut pas répondre directement à cette question car on n'a pas de définition précise de ce qu'est une pièce équilibrée.

▶ On va vérifier si les résultats obtenus sont « compatibles » avec un modèle d'équiprobabilité sur $\{P ; F\}$.

La probabilité théorique d'obtenir Pile quand on lance une pièce équilibrée est égale à $\frac{1}{2}$.

Les probabilités théoriques sont donc : $p(\text{Pile}) = p(\text{Face}) = 0{,}5$.

▶ Pour le joueur A la probabilité observée pour Pile est égale à 0,45.

Pour le joueur B la probabilité observée pour Pile est égale à 0,52.

Comment peut-on faire pour savoir si ces probabilités peuvent être considérées « proches » de 0,5 ?

▶ On peut, par exemple, s'intéresser à la « distance » entre la distribution des fréquences (f_1, f_2) obtenues en lançant **n** fois une pièce et la loi de probabilité équirépartie $\left\{\frac{1}{2} ; \frac{1}{2}\right\}$, et regarder si cette distance est « petite ».

Remarque *On prend* f_1 *pour la fréquence des « Pile » et* f_2 *pour la fréquence des « Face ».*

En prenant une définition classique de la distance, on peut essayer de baser la notion de compatibilité sur l'étude du carré de cette distance, à savoir :

$$d^2 = (f_1 - 0{,}5)^2 + (f_2 - 0{,}5)^2 .$$

Dans le cas du jet d'une pièce n fois on a $f_1 + f_2 = 1$.

Ainsi
$$f_2 - 0{,}5 = 1 - f_1 - 0{,}5 = 0{,}5 - f_1 .$$
$$d^2 = (f_1 - 0{,}5)^2 + (0{,}5 - f_1)^2 .$$
$$d^2 = 2(f_1 - 0{,}5)^2 .$$

La quantité d^2 devrait être « petite », mais elle est soumise à la fluctuation d'échantillonnage, c'est-à-dire que sa valeur varie d'une série de lancers à l'autre.

C'est précisément l'étude de la fluctuation d'échantillonnage qui va nous permettre de convenir d'un seuil entre valeur « petite » et valeur « non petite » de d^2.

▶ Imaginons qu'un joueur lance 1 000 fois une pièce de monnaie. Il obtient une distribution de fréquences (f_1, f_2) conduisant à une valeur d^2 observée qui sera notée d^2_{obs}.

Pour déterminer le seuil on va simuler des séries de 1 000 chiffres choisis au hasard dans $\{0 ; 1\}$ (on peut attribuer 0 à « Face » et 1 à « Pile »).

▶ On décide d'effectuer 100 simulations de séries de 1 000 tirages de chiffres au hasard dans $\{0 ; 1\}$.

Dans le tableau suivant on a indiqué le nombre de « Pile » obtenus lors de chaque série de 1 000.

Les cent valeurs obtenues ont ensuite été classées dans l'ordre croissant (le tirage a été fait à l'aide du tableur Excel).

	100 simulations de séries de 1 000 tirages de chiffres au hasard dans $\{0 ; 1\}$										
nb P	fréq	$(f_1 - 0,5)^2$	nb P	fréq	$(f_1 - 0,5)^2$	nb P	fréq	$(f_1 - 0,5)^2$	nb P	fréq	$(f_1 - 0,5)^2$
467	0,467	0,001089	487	0,487	0,000169	496	0,496	0,000016	508	0,508	6,4E-05
468	0,468	0,001024	487	0,487	0,000169	496	0,496	0,000016	509	0,509	8,1E-05
469	0,469	0,000961	488	0,488	0,000144	496	0,496	0,000016	509	0,509	8,1E-05
476	0,476	0,000576	488	0,488	0,000144	497	0,497	9E-06	510	0,51	0,0001
477	0,477	0,000529	488	0,488	0,000144	497	0,497	9E-06	510	0,51	0,0001
479	0,479	0,000441	489	0,489	0,000121	498	0,498	4E-06	510	0,51	0,0001
479	0,479	0,000441	489	0,489	0,000121	499	0,499	0,000001	510	0,51	0,0001
481	0,481	0,000361	490	0,49	0,0001	499	0,499	0,000001	513	0,513	0,000169
481	0,481	0,000361	490	0,49	0,0001	499	0,499	0,000001	514	0,514	0,000196
481	0,481	0,000361	490	0,49	0,0001	500	0,5	0	515	0,515	0,000225
481	0,481	0,000361	490	0,49	0,0001	501	0,501	0,000001	516	0,516	0,000256
481	0,481	0,000361	491	0,491	8,1E-05	501	0,501	0,000001	518	0,518	0,000324
481	0,481	0,000361	492	0,492	6,4E-05	501	0,501	0,000001	521	0,521	0,000441
482	0,482	0,000324	492	0,492	6,4E-05	502	0,502	4E-06	523	0,523	0,000529
482	0,482	0,000324	493	0,493	4,9E-05	502	0,502	4E-06	523	0,523	0,000529
483	0,483	0,000289	494	0,494	3,6E-05	504	0,504	0,000016	524	0,524	0,000576
483	0,483	0,000289	494	0,494	3,6E-05	504	0,504	0,000016	524	0,524	0,000576
484	0,484	0,000256	495	0,495	0,000025	505	0,505	0,000025	524	0,524	0,000576
484	0,484	0,000256	495	0,495	0,000025	505	0,505	0,000025	526	0,526	0,000676
484	0,484	0,000256	495	0,495	0,000025	507	0,507	4,9E-05	527	0,527	0,000729
486	0,486	0,000196	495	0,495	0,000025	507	0,507	4,9E-05	531	0,531	0,000961
486	0,486	0,000196	495	0,495	0,000025	507	0,507	4,9E-05	532	0,532	0,001024
486	0,486	0,000196	496	0,496	0,000016	507	0,507	4,9E-05	536	0,536	0,001296
487	0,487	0,000169	496	0,496	0,000016	508	0,508	6,4E-05	537	0,537	0,001369
487	0,487	0,000169	496	0,496	0,000016	508	0,508	6,4E-05	539	0,539	0,001521

(nb P désigne le nombre de « Pile » obtenu quand on lance 1 000 fois une pièce).

▶ On peut représenter le diagramme en boîte des 100 valeurs de d^2 obtenues lors des 100 simulations de séries de 1 000 tirages de chiffres au hasard dans $\{0 ; 1\}$.

Pour cela on a rangé les 100 valeurs de $d^2 = 2(f_1 - 0{,}5)^2$ dans l'ordre croissant afin de déterminer d_1, d_9, Q_1, Q_3 et Med.

n°	d^2	n°	d^2	n°	d^2	n°	d^2
1	0	26	0,00005	51	0,000242	76	0,000722
2	0,000002	27	0,00005	52	0,000242	77	0,000722
3	0,000002	28	0,000072	53	0,000288	78	0,000722
4	0,000002	29	0,000072	54	0,000288	79	0,000722
5	0,000002	30	0,000098	55	0,000288	80	0,000722
6	0,000002	31	0,000098	56	0,000338	81	0,000882
7	0,000002	32	0,000098	57	0,000338	82	0,000882
8	0,000008	33	0,000098	58	0,000338	83	0,000882
9	0,000008	34	0,000098	59	0,000338	84	0,001058
10	**0,000008**	35	0,000128	60	0,000338	85	0,001058
11	0,000018	36	0,000128	61	0,000392	86	0,001058
12	0,000018	37	0,000128	62	0,000392	87	0,001152
13	0,000032	38	0,000128	63	0,000392	88	0,001152
14	0,000032	39	0,000128	64	0,000392	89	0,001152
15	0,000032	40	0,000162	65	0,00045	**90**	**0,001152**
16	0,000032	41	0,000162	66	0,000512	91	0,001352
17	0,000032	42	0,000162	67	0,000512	92	0,001458
18	0,000032	43	0,0002	68	0,000512	93	0,001922
19	0,000032	44	0,0002	69	0,000512	94	0,001922
20	0,000032	45	0,0002	70	0,000578	95	0,002048
21	0,00005	46	0,0002	71	0,000578	96	0,002048
22	0,00005	47	0,0002	72	0,000648	97	0,002178
23	0,00005	48	0,0002	73	0,000648	98	0,002592
24	0,00005	49	0,0002	74	0,000648	99	0,002738
25	**0,00005**	**50**	**0,0002**	**75**	**0,000722**	100	0,003042

Les valeurs de ce tableau sont le double de celles du tableau précédent.

Le tableau nous permet aussi d'obtenir le diagramme en boîte des valeurs simulées de 1 000 d^2 représenté sur la **figure 5**.

Cela donne :

	pour d^2	pour 1000 d^2
d_1	0,000 008	0,008
Q_1	0,000 05	0,05
Med	0,000 2	0,2
Q_3	0,000 722	0,722
d_9	0,001 152	1,152

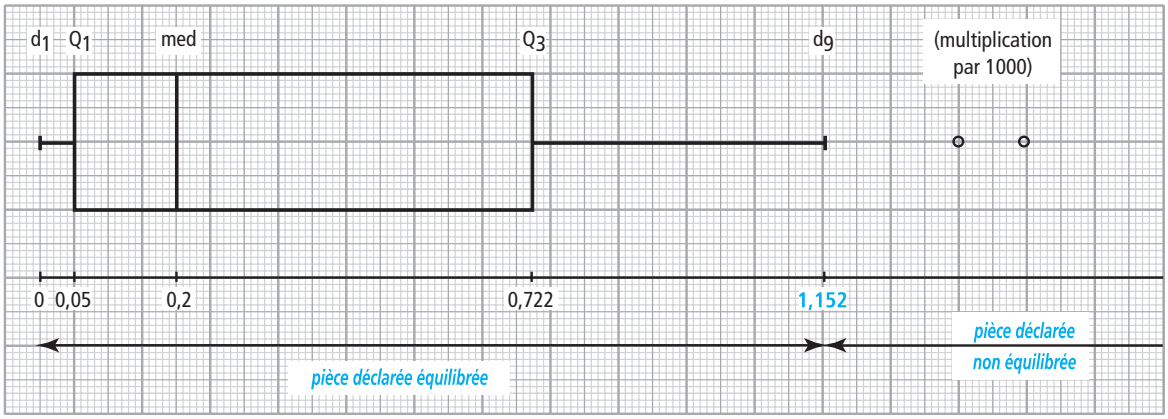


Fig. 5

▶ le premier décile est presque confondu avec le minimum.

▶ les huit plus grandes valeurs ne sont pas représentées (problème d'échelle si on veut rendre la boîte lisible).

Le 9ème décile de la série des valeurs simulées de d^2 est 0,001 152. Cela signifie que 90 % des valeurs simulées de d^2 sont dans l'intervalle $[0 ; 0{,}001\ 152]$.

Conclusion

On convient que :

▶ si $d^2_{obs} \leq 0{,}001\ 152$ alors la pièce sera déclarée équilibrée.

▶ si $d^2_{obs} > 0{,}001\ 152$ alors la pièce sera déclarée non équilibrée.

On associe à cette conclusion un risque d'erreur de 10 % : en effet, en utilisant cette méthode sur les données simulées, on se serait « trompé » dans 10 % des cas.

▶ Revenons maintenant à la question posée dans **l'exemple ❼** concernant les pièces des joueurs A et B.

On peut présenter les calculs dans un tableau.

	Joueur A					Joueur B				
		f_{obs}	p_T	$f_{obs} - p_T$	$(f_{obs} - p_T)^2$		f_{obs}	p_T	$f_{obs} - p_T$	$(f_{obs} - p_T)^2$
Pile	450	0,45	0,5	– 0,05	0,002 5	520	0,52	0,5	0,02	0,000 4
Face	550	0,55	0,5	0,05	0,002 5	480	0,48	0,5	– 0,02	0,000 4
					0,005 0					0,000 8

f_{obs}	fréquence observée.
p_T	probabilité théorique.

Pour le joueur A on obtient $d^2_{obs} = 0{,}005\ 0$.

Pour le joueur B on obtient $d^2_{obs} = 0{,}000\ 8$.

On a : $0{,}005\ 0 > 0{,}001\ 152$ et $0{,}000\ 8 \leq 0{,}001\ 152$.

D'après ce que l'on a convenu :

▶ on déclare la pièce du joueur A non équilibrée ;

▶ on déclare la pièce du joueur B équilibrée.

On peut aussi se poser la question suivante :

« On effectue 1 000 lancers d'une pièce. À quel intervalle doit appartenir le nombre de « Pile » obtenu pour pouvoir déclarer la pièce équilibrée ? »

On garde le même seuil et on élimine les dix valeurs correspondant aux 10 plus grandes valeurs de d^2.

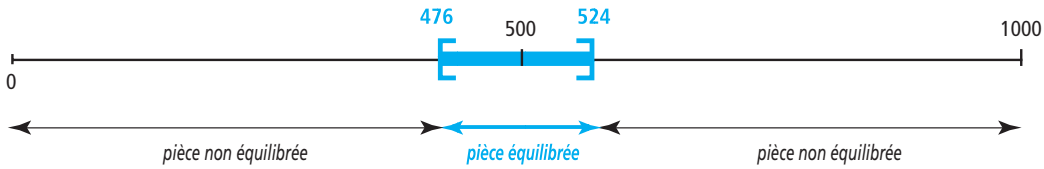
Écrivons ces 10 valeurs dans un tableau.

nombre de « Pile »	467	468	469	...	526	527	531	532	536	537	539
10 plus grandes valeurs de d^2	0,002	0,002	0,001 9	...	0,001 3	0,001 4	0,001 9	0,002 0	0,002 6	0,002 7	0,003 0

Il reste l'intervalle $[476 ; 524]$ qui est un intervalle centré en 500.

Au vu des simulations faites, on peut déclarer que si le nombre P de « Pile » est tel que $476 \leq P \leq 524$ la pièce est équilibrée (le nombre F de « Face » est aussi dans le même intervalle).

On peut présenter ce résultat sur un axe où l'on a indiqué le nombre de « Pile » (ou de « Face ») pour 1 000 lancers.



❸ La loi du Khi-deux

Loi de probabilité de nd^2

On peut se demander si le seuil obtenu dans l'exemple précédent est satisfaisant. En effet le nombre de tirages (1 000) et le nombre de séries (100) ont été fixés arbitrairement.

Qu'obtiendrait-on si on fixait le nombre de tirages à 2 000 ?

Qu'obtiendrait-on si on fixait le nombre de séries à 1 000 ?

Nous n'allons pas simuler ces expériences mais plutôt considérer la loi des grands nombres vue en première.

Pour une expérience donnée, plus le nombre de tirages est grand, plus la distribution des fréquences est proche de la loi de probabilité.

Ainsi plus le nombre de lancers augmente et plus les valeurs de d^2 auront tendance à être faibles.

Des résultats théoriques expliquent la ressemblance entre les histogrammes des fréquences lorsque le nombre n de tirages est assez grand (ainsi $n = 1\ 000$ dans **l'exemple ❼**).

On peut démontrer que la loi de probabilité de nd^2 ne dépend pas de n pourvu que n soit assez grand : **dans la pratique on peut considérer que la loi de probabilité de nd^2 ne dépend pas de n pour tout $n > 100$.**

Loi du Khi-deux

Considérons une distribution expérimentale sur un échantillon de taille n, les « individus » de cet échantillon étant répartis en un certain nombre k de classes.

Il existe alors une loi de probabilité, dite loi du Khi-deux, qui ne dépend pas de n (pourvu que n soit assez grand) mais seulement de k.

Définition 6 On considère un univers théorique défini par une loi $P = (p_1, p_2, ..., p_k)$.

La loi de probabilité de la quantité :

$$\chi^2 = n \sum_{i=1}^{k} \frac{(f_i - p_i)^2}{p_i}$$

est, pour n assez grand, distribuée selon une loi qui dépend uniquement de k.

Cette loi s'appelle **loi du Khi-deux** à $k - 1$ degrés de liberté (d.d.l).

p_i = probabilité d'obtenir une observation de la loi théorique P dans la classe i.

f_i = fréquence obtenue par chaque observation de la classe i.

n = effectif total observé.

k = nombre de classes.

Remarques ▶ *La lettre χ est la lettre grecque « khi » (prononcer ki).*

▶ *Si n_i est l'effectif observé dans la classe i alors* $f_i = \frac{n_i}{n}$.

▶ *La loi du Khi-deux est aussi appelée loi de Pearson, du nom du mathématicien anglais Karl PEARSON (1857-1936).*

On trouve des tables numériques donnant, entre autres, la liste des $9^{èmes}$ déciles de ces lois : voir le tableau suivant où figure un extrait de ces tables.

d.d.l $k - 1$	1	2	3	4	5	6	7	10	20	30
$9^{ème}$ décile d_9	2,71	4,61	6,25	7,78	9,24	10,64	12,02	15,99	28,41	40,26

lecture	la probabilité pour que χ^2 dépasse la valeur 2,71 est égale à 0,10 pour 1 d.d.l.

Application

Exemple 8 ***Énoncé***

On reprend les deux joueurs de **l'exemple 7**.

	A	B
nombre de « Pile »	450	520
nombre de « Face »	550	480

Tester, en utilisant la loi du Khi-deux, si les pièces sont équilibrées (au risque de 10 %).

Solution

On a deux classes ▶ la classe 1 pour les « Pile » ;

▶ la classe 2 pour les « Face ».

L'effectif total observé est $n = 1\,000$.

Le nombre de classes est $k = 2$, donc le nombre de degrés de liberté est $k - 1 = 1$.

Les probabilités théoriques sont $p_1 = \frac{1}{2}$ et $p_2 = \frac{1}{2}$.

Présentons les calculs dans un tableau.

	classe	résultat observé	fréquence observée f_i	probabilité théorique p_i	$(f_i - p_i)^2$	$\frac{(f_i - p_i)^2}{p_i} \times n$
A	1	450	0,45	0,5	0,002 5	5
	2	550	0,55	0,5	0,002 5	5
						$\chi^2 = 10$
B	1	520	0,52	0,5	0,000 4	0,8
	2	480	0,48	0,5	0,000 4	0,8
						$\chi^2 = 1,6$

Pour le joueur A la valeur du χ^2 est 10. Comme $10 > 2,71$ on considère sa pièce comme non équilibrée.

Pour le joueur B la valeur du χ^2 est 1,6. Comme $1,6 < 2,71$ on considère sa pièce comme équilibrée.

Le test du Khi-deux donne ici les mêmes résultats qu'avec la simulation.

Exemple 9

Énoncé

Dans le système AB0 on distingue 4 groupes sanguins : A – B – AB – 0.

On peut considérer que la répartition théorique en France est proche de :

groupe	A	B	AB	0
pourcentage	40 %	10 %	5 %	45 %

On prélève un échantillon de $n = 800$ personnes parmi la population française.

Dans cet échantillon on trouve 342 personnes du groupe A, 88 personnes du groupe B, 32 personnes du groupe AB et 338 personnes du groupe 0.

Peut-on dire, au seuil de risque de 10 %, que l'échantillon choisi est représentatif de la population française ?

Solution

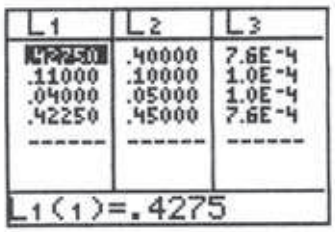
Comme le nombre n est assez grand on peut faire le test du Khi-deux.

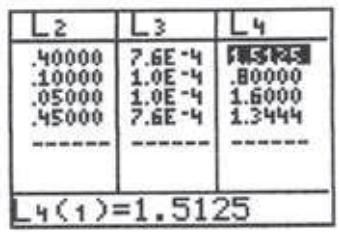
On peut, soit faire un tableau comme précédemment, soit utiliser les listes de la calculatrice.

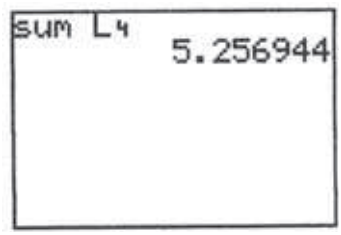
Utilisons une calculatrice.

L_1	L_2	L_3	L_4
fréquence observée f_i	probabilité théorique p_i	$(f_i - p_i)^2$	$\frac{(f_i - p_i)^2}{p_i} \times n$
		$L_3 = (L_1 - L_2)^2$	$L_4 = L_3 \div L_2 \times 800$

On obtient les écrans suivants :







La valeur arrondie du χ^2 est égale à 5,26.