Exercice 1. *Le nombre de décès cumulés (en milieu hospitalier) dus au covid-19 en France jour par jour pendant 15 jours consécutifs du 14/03/2020 au 28/03/2020 (numérotés de 1 à 15) est résumé dans le tableau ci-dessous.*

jour	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Nombre de décès	91	127	148	175	264	372	450	562	674	860	1100	1331	1696	1995	2314

(1) *Indiquer: le caractère, les individus, le nombre d'individus, la nature du caractère (qualitatif, quantitatif discret, quantitatif continu).*
(2) *Calculer le nombre moyen de décès par jour dans la période.*
(3) *Quel est le nombre médian de décès par jour dans la période.*
(4) *Calculer la variance et l'écart-type.*
(5) *Calculer les premier et troisième quartiles (Q_1 et Q_3), puis l'écart interquartile ($Q_3 - Q_1$);*

(1) Caractère observé : Nb de décès par jour

Individus : jours

15 individus

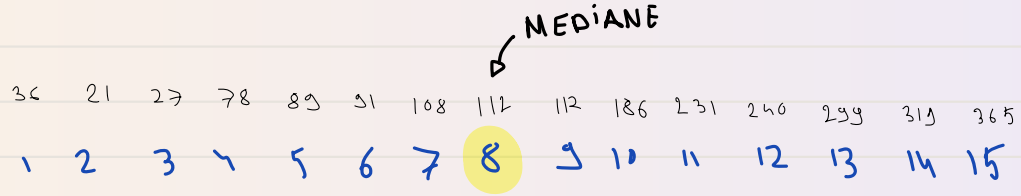
Nature du caractère : quantitatif discret

jour	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Décès cumulés	91	127	148	175	264	372	450	562	674	860	1100	1331	1696	1995	2314
Décès par jour	91	36	21	27	89	108	78	112	112	186	240	231	365	299	319

(2) $\sum X_i = 2314$

$E(X) = \frac{\sum X_i}{15} = 154$

(3)



Définition *La* **médiane** *M de la série statistique simple $x = (x_1, ..., x_n)$ est la valeur du caractère telle qu'au moins 50% des individus ont une valeur inférieure ou égale à M et au moins 50% des individus ont une valeur supérieure à M.*

- *Si $F_j = 0,5$, on dit que l'intervalle $[a_j, a_{j+1}[$ est médian et, par convention, on pose : $M = \frac{a_j + a_{j+1}}{2}$ (dans ce cas, il y a d'autres conventions possibles, tout élément de l'intervalle $[a_j; a_{j+1}[$ peut convenir, on peut choisir de prendre a_j ou a_{j+1}).*
- **Pour un caractère discret**, *si $F_{j-1} < 0,5 < F_j$, on pose $M = a_j$.*
  Avec cette convention, Si la donnée statistique $(x_1, ..., x_n)$ est rangée dans $(x'_1, ..., x'_n)$ par ordre croissant des valeurs du caractère, alors $M := x'_{m+1}$ si $n = 2m+1$ est impair; et $M = \frac{x'_m + x'_{m+1}}{2}$ si $n = 2m$ est pair.
- **Pour un caractère continu**, *la convention usuelle si $F_{j-1} = \bar{F}(a_{j-1}) < 0,5 < F_j = \bar{F}(a_j)$ est de prendre pour M la valeur qui la fréquence 0,5 sur le graphe de la fréquence empirique, c'est-à-dire prendre la moyenne de a_{j-1} (avec le poids $(F_j - 0,5)/f_j$) et de a_j (avec le poids $(0,5 - F_{j-1})/f_j$)*

$$M = a_{j-1}\frac{F_j - 0,5}{f_j} + a_j\frac{0,5 - F_{j-1}}{f_j},$$

en notant f_j la fréquence de l'intervalle $]a_{j-1}, a_j]$.

ici $15 = 2m+1$

$\Rightarrow m = 7$

Donc

$M = x_{m+1} = x_8 = 112$

(4)

jour	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Décès cumulés	91	127	148	175	264	372	450	562	674	860	1100	1331	1696	1995	2314
Décès par jour	91	36	21	27	89	108	78	112	112	186	240	231	365	299	319
(Décès par jour)^2	8281	1296	441	729	7921	11664	6084	12544	12544	34596	57600	53361	133225	89401	101761

$$Var(X) = E(X^2) - \left[E(X)\right]^2$$

$$E(X_i)^2 = 531\,448$$

$$E(X^2) = 35\,430 \qquad \left[E(X)\right]^2 = 23\,716$$

$$Var(X) = 35\,430 - 23\,716$$

$$\boxed{Var(X) = 11\,714}$$

$$\boxed{\sigma(X) = 108}$$

(5)

Le **quantile** d'ordre $p \in]0;1[$ noté q_p de la série statistique simple $x = (x_1, ..., x_n)$ est la valeur du caractère telle qu'il y ait une proportion au moins égale à p de valeurs inférieures ou égales à q_p et telle qu'il y ait au moins une proportion au moins égale à $1 - p$ de valeurs supérieures ou égales à q_p :

Définition *Les* **quartiles** Q_1, Q_2 *et* Q_3 *de la série statistique simple* $x = (x_1, ..., x_n)$ *sont les quantiles d'ordre respectif* $0,25$, $0,50$ *et* $0,75$.

On remarque que Q_2 est la médiane.
On peut définir de manière analogue les déciles et les centiles.

36	21	27	78	89	91	108	112	112	186	231	240	299	319	365
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

p- Quantile :

$$Q_p = \begin{cases} X_{\lceil np \rceil} & \text{si } np \notin \mathbb{N} \\ \dfrac{X_{np} + X_{np+1}}{2} & \text{si } np \in \mathbb{N} \end{cases}$$

$\frac{15}{4} = 3,75$ donc $Q_{0,25} = X_4 = 78$

$\frac{3 \cdot 15}{4} = 11{,}25$ donc $Q_{0{,}75} = X_{12} = 240$

Exercice 2. *On s'intéresse au gain en euro d'un jeu de hasard (on appelle gain la différence entre la somme gagnée et la somme misée). Sur 10 000 000 de cartes à gratter d'un euro, on a obtenu les gains suivants:*

Gain	*-1*	*0*	*1*	*4*	*9*	*199*	*4 999*
Nombre de cartes	*7 387 198*	*1 176 471*	*986 193*	*250 000*	*200 000*	*133*	*5*

(1) *Indiquer: le caractère, les individus, le nombre d'individus, la nature du caractère (qualitatif, quantitatif discret, quantitatif continu).*
(2) *Déterminer la fréquence des cartes perdantes (correspondant à un gain est < 0).*
(3) *Déterminer la fréquence des cartes gagnantes (correspondant à un > 0).*
(4) *Calculer le gain moyen. Interpréter ce résultat.*
(5) *Calculer les fréquences et fréquences cumulées.*
(6) *Quel est le gain médian.*
(7) *Calculer les premier et troisième quartiles (Q_1 et Q_3), puis l'écart interquartile ($Q_3 - Q_1$);*
(8) *calculer la moyenne, la variance et l'écart-type.*

(1) Caractère : Gains

Individus : Cartes

$n = 10 \cdot 10^6$

Nature caractère : Quantitatif discret

(2) $f_p = \frac{7\,387\,198}{10^7} = 0{,}74$

(3) $f_g = 0{,}14$

(4) $E(G) = \frac{\sum N_i \times G_i}{n} = -0{,}35$: On perd en moyenne...

(5)

Gain	*-1*	*0*	*1*	*4*	*9*	*199*	*4 999*
Nombre de cartes	*7 387 198*	*1 176 471*	*986 193*	*250 000*	*200 000*	*133*	*5*
Fréquences	0,738720	0,117647	0,098619	0,025000	0,020000	0,000013	0,000000
Fréquences cum.	0,738720	0,856367	0,954986	0,979986	0,999986	1,000000	1,000000

(6) Méd : −1

(7) $Q_1 = -1$ $\quad Q_3 = 0$

(8) $E(G) : -0{,}35$

$$Var(G) = E(G^2) - (E(G))^2 = \frac{\sum N_i G_i^2}{n} - (E(G))^2$$

$$= 15{,}75$$

$$\sigma(G) = 3{,}97$$

Exercice 3. *Un tour automatique produit des axes cylindriques. Les diamètres (en dixièmes de mm), mesurés sur un lot de 1000 pièces ont donné les résultats suivants:*

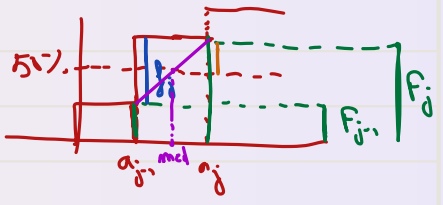
classes	*[244;248[*	*[248;249[*	*[249;250[*	*[250;251[*	*[251;252[*	*[252;258]*
effectifs	*143*	*152*	*200*	*194*	*158*	*153*
f_i	0,143	0,152	0,200	0,194	0,158	0,153
F_i	0,143	0,295	0,495	0,689	0,847	1,000

(1) *Indiquer: le caractère, les individus, le nombre d'individus, la nature du caractère (qualitatif, quantitatif discret, quantitatif continu).*
(2) *Calculer les fréquences et fréquences cumulées.*
(3) *Donner une représentation graphique des fréquences cumulées et en déduire la médiane.*

(1) Caractère : diamètre des tours

Individus : Tours

Nature du caractère : quantitatif continu



$a_j = 251$

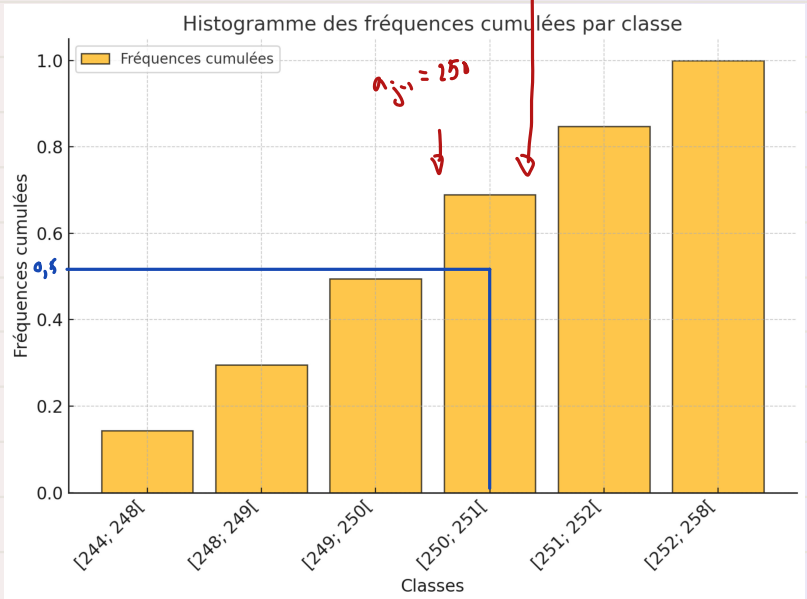
(2) cf. énoncé

(3) ~~Méd = $\frac{250+251}{2} = 250{,}5$~~

NON : il faut interpoler...

La cfq la dif de la médiane est dans :

$$\text{méd} = 250 \times \frac{0{,}689 - 0{,}5}{0{,}194} + 251 \times \frac{0{,}5 - 0{,}495}{0{,}194}$$



= 250,026

Exercice 5. *Une personne possède des actions d'une entreprise. Dans l'espoir de gagner un maximum d'argent, elle note les informations suivantes sur ses actions :*

jour	1	2	3	4	5	6	7	8	9	10	11
nombre d'actions vendues	105	92	60	80	91	51	70	92	80	50	99
prix de l'action	88	74	70	61	67	69	58	63	73	61	55

(1) *Le coefficient de corrélation linéaire entre les caractères x (prix de l'action au jour j) et y (nombre d'actions vendues le même jour) est $\rho_{x,y} \approx 0.28$. Est-il raisonnable de proposer un modèle du type $Y = aX + b$? Justifier.*

(2) *L'actionnaire décide de chercher une relation entre le prix de l'action au jour j (variable Z) et le nombre d'actions vendues le jour $j-1$ (variable T).*

(a) *Dessiner le nuage de points correspondant. Quel modèle proposez-vous ?*

(b) *Calculer le coefficient de corrélation linéaire entre les variables Z et T, puis la droite de régression de Z en fonction de T.*

(c) *D'après ce modèle, lui conseillez-vous de vendre ses actions le $11^{ème}$ jour ou plutôt d'attendre le $12^{ème}$ jour pour les vendre?*

(1) $\rho = 0,28$ éloigné de 1 $\Rightarrow$ MAUVAISE approximation linéaire...

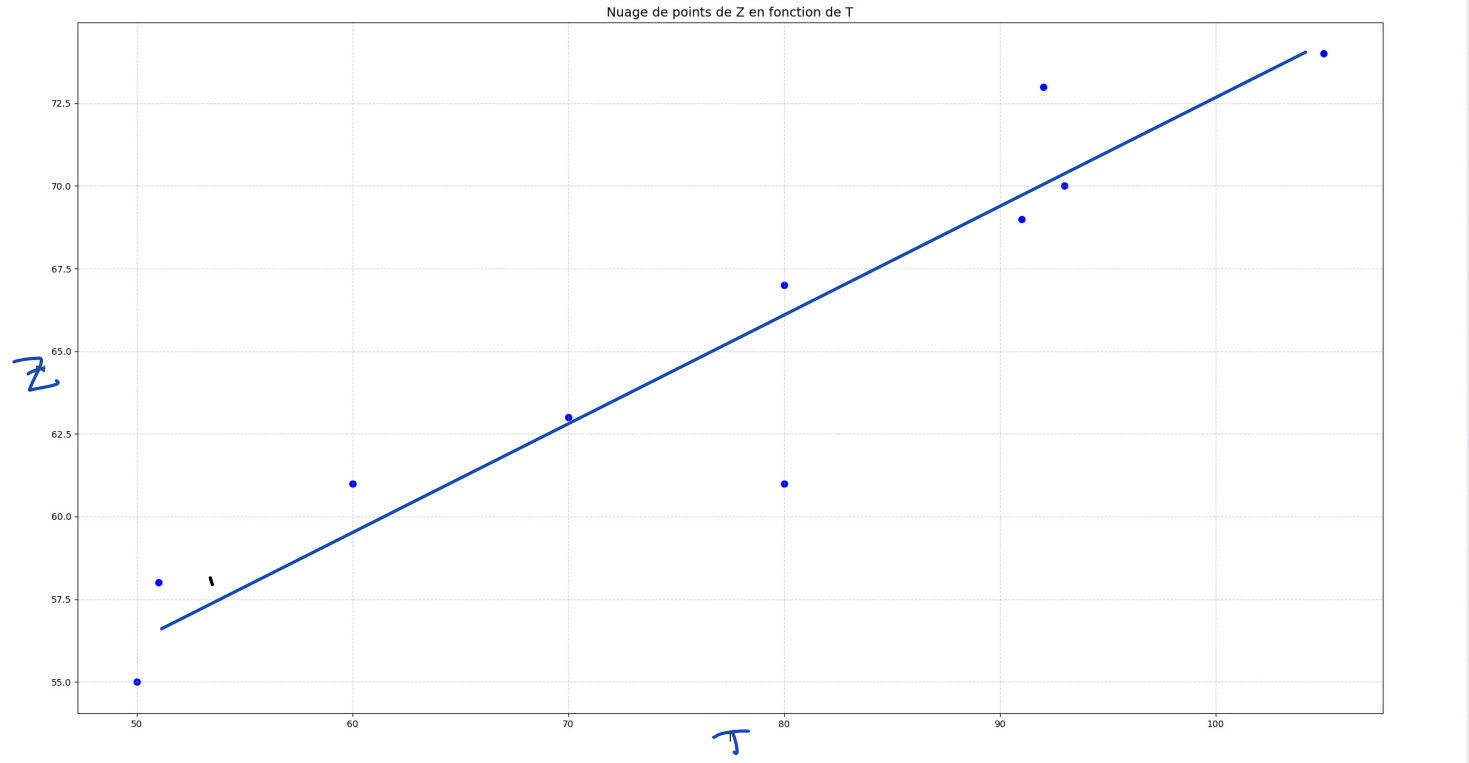
(2)

j	2	3	4	5	6	7	8	9	10	11
Z	74	70	61	67	69	58	63	73	61	55
T	105	93	60	80	91	51	70	92	80	50

(a) Approximation linéaire semble correcte...

(b) $$\rho = \frac{\mathrm{cov}(T,Z)}{\sigma(T)\sigma(Z)} = \frac{\mathbb{E}(TZ) - \mathbb{E}(T)\mathbb{E}(Z)}{\sqrt{(\mathbb{E}(T^2) - \mathbb{E}(T)^2)(\mathbb{E}(Z^2) - \mathbb{E}(Z)^2)}} = \frac{103.58}{109.90} = 0,9425$$

Proche de 1 !



$Z = aT + b$

$a = \dfrac{\text{cov}(T,Z)}{\text{Var}(T)} = \dfrac{103,58}{322.16} = 0,3215$

$b = E(Z) - a\,E(T)$

$= 65,1 - 0,3215 \times 77,2$

$= 40,28$

$\boxed{Z = 0,3215\,T + 40,28}$

(c) Prix de l'action au jour 11 : 55

Prédiction du prix de l'action au jour 12 : $Z = 0,3215 \times 55 + 40,28$

$= 57,96$

Donc c'est mieux d'attendre le jour 12 !

Exercice 4. *Nous reprenons les données de l'Exercice 1 de cette feuille. Le nombre de décès cumulés (en milieu hospitalier) dus au covid-19 en France jour par jour pendant 15 jours consécutifs du 14/03/2020 au 28/03/2020 est résumé dans le tableau ci-dessous. Les jours sont numérotés de 1 à 15, le nombre de décès jusqu'au jour numéro i est noté $N(i)$.*

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$N(i)$	91	127	148	175	264	372	450	562	674	860	1100	1331	1696	1995	2314

Etudiez et comparez les deux modèles suivants de régression linéaire:

(1) $N(i) = a \times i + b$.

(2) $N(i) = a \times i^{2,5} + b$ *(ce qui correspond à* $N(i) = a \times (i^2 \times \sqrt{i}) + b$*, dans un tableur vous pouvez taper* ^2.5 *pour mettre un nombre à la puissance 2,5)*

Pour chacun de ces modèles, identifier le X et le Y, calculer moyennes, variances, covariance, le $\widehat{R}^2$ (représentant la part de la variance de la variable N expliquée par le modèle), les coefficients a et b de la régression linéaire (la meilleure au sens des moindres carrés ordinaires). Quel modèle vous paraît le meilleur? Justifier.

Utiliser les 2 modèles pour estimer les nombres $N(16)$, $N(17)$ et $N(26)$ de décès cumulés jusqu'au dates respectives 29/03/2020, 30/03/2020 et 8/04/2020 respectivement. Comparez avec les vraies valeurs: $N(16) = 2606$ et $N(17) = 3024$ et $N(26) = 7632$.

↳ Modèle (1) : $X = i$; $Y = N(i)$

Modèle (2) : $X = i^{2,5}$; $Y = N(i)$

↳ Régression linéaire : $Y = aX + b$

$$a = \frac{cov(X,Y)}{Var(X)} = \frac{\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)}{\mathbb{E}(X^2) - \mathbb{E}(X)^2}$$

$$b = \mathbb{E}(Y) - a\,\mathbb{E}(X)$$

```
=== Détails pour Modèle 1 (N = a*i + b) ===
X_mean = 8.0000, N_mean = 810.6000
X2_mean = 82.6667, XN_mean = 9340.8667, NN_mean = 1145989.1333
Var(X) = 18.6667, Var(N) = 488916.7733
Cov(X, N) = 2856.0667
a = 153.0036
b = -413.4286
R^2 = 0.8938
```

↳ Modèle 1 :

$$a_1 = \frac{cov(X,Y)}{Var(X)} = \frac{\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)}{\mathbb{E}(X^2) - \mathbb{E}(X)^2}$$

$$\mathbb{E}(X) = \frac{1}{15}\sum_{i=1}^{15}(i) = 8$$

$$\mathbb{E}(X^2) = \frac{1}{15}\sum (i)^2 = 83$$

$$\mathbb{E}(Y) = \frac{1}{15}\sum N(i) = 811$$

$$\mathbb{E}(XY) = \frac{1}{15}\sum (i \cdot N(i)) = 9341 \qquad \mathbb{E}(Y^2) = 1145989$$

$$a_1 = \frac{9341 - 8 \times 811}{83 - 8^2} = 150$$

$b_1 = \mathbb{E}(Y) - a_1 \mathbb{E}(X)$

$= 811 - 150 \times 8$

$= -389$ **Donc $Y_1 = 150X - 389$**

↳ $R^2 = \frac{(Cov(X,Y))^2}{Var(X)Var(Y)} = a \times \frac{cov(X,Y)}{Var(Y)}$

$= 150 \times \frac{3341 - 8 \times 811}{1145989 - 811^2} = 0,88 = R^2$

↳ Modèle 2 : Il faut calculer les $i^{2,5}$

```
i =
[ 1.  2.  3.  4.  5.  6.  7.  8.  9. 10. 11. 12. 13. 14. 15.]

i**2.5 =
[  1.           5.65685425  15.58845727  32.          55.90169944
  88.18163074 129.64181424 181.01933598 243.        316.22776602
 401.31159963 498.83063258 609.33816555 733.36484781 871.4212529 ]

N =
[  91.  127.  148.  175.  264.  372.  450.  562.  674.  860. 1100. 1331.
 1696. 1995. 2314.]
```

```
=== Détails pour Modèle 2 (N = a*i^2.5 + b) ===
X_mean = 278.8323, N_mean = 810.6000
X2_mean = 153280.0000, XN_mean = 418029.5467, NN_mean = 1145989.1333
Var(X) = 75532.5650, Var(N) = 488916.7733
Cov(X, N) = 192008.1083
a = 2.5421
b = 101.7924
R^2 = 0.9983
```

```
=== Estimation des valeurs pour i=16, i=17 et i=26 ===

--- Modèle 1 (N = a*i + b) ---
N(16) prédit = 2034.6  | valeur réelle = 2606
N(17) prédit = 2187.6  | valeur réelle = 3024
N(26) prédit = 3564.7  | valeur réelle = 7632

--- Modèle 2 (N = a*(i^2.5) + b) ---
N(16) prédit = 2704.9  | valeur réelle = 2606
N(17) prédit = 3130.9  | valeur réelle = 3024
N(26) prédit = 8864.1  | valeur réelle = 7632
```