Statistiques descriptives

Objectifs Décrire efficacement d'importants jeux de données.
Rechercher l'existence d'une relation (corrélation) affine entre deux variables.
Interpoler et extrapoler des données.

I - Série statistique à une variable

1) Un peu de vocabulaire...

Un **caractère**, ou **variable**, est une propriété commune aux **individus** d'une **population**.
Un **échantillon** est une partie de la population complète.
L'**effectif** d'une population ou d'un échantillon est le nombre d'individus qui la compose.
Un caractère peut-être **quantitatif**, s'il peut s'exprimer par un nombre, ou **qualitatif** (couleur des yeux, nationalité,...) dans le cas contraire.
On peut de plus distinguer les caractères quantitatifs **discrets**, qui ne prennent que des valeurs numériques isolées (ex. nombre d'élèves par classes), des caractères quantitatifs **continus**, lorsque toutes les valeurs peuvent être prises dans un intervalle (ex. taille des élèves, durée de vie d'un composant).

2) Description par la moyenne et l'écart-type

Définition *On considère N valeurs d'un caractère x_1, x_2,...,x_N. La moyenne, notée $\overline{x}$, est :*

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Si la valeurs x_1 est prise n_1 fois par le caractère, la valeur x_2 prise n_2 fois, ..., alors

$$\overline{x} = \frac{n_1x_1 + n_2x_2 + \cdots + n_Nx_N}{N} = \frac{1}{N}\sum_{i=1}^{n} n_ix_i \quad avec \quad N = \sum_{i=1}^{n} n_i$$

On parle alors de moyenne **pondérée**.

La moyenne d'une série permet de situer le niveau global de celle-ci : c'est une **caractéristique de position**, mais ne donne pas d'information sur la répartion, ou **dispersion**, des valeurs autour de cette position centrale.
Par exemple les séries statistiques : 10, 10, 10, 10, 10, 10, 10 et 2, 2, 2, 10, 18, 18 18 ont le même effectif et la même moyenne, alors qu'elles sont nettement différentes.

Définition *La* **variance d'une série** *est la moyenne des carrés des écarts à la moyenne :*

$$V = \frac{n_1\left(x_1 - \overline{x}\right)^2 + n_2\left(x_2 - \overline{x}\right)^2 + \ldots n_N\left(x_N - \overline{x}\right)^2}{N} = \frac{1}{N}\sum_{i=1}^{N} n_i\left(x_i - \overline{x}\right)^2$$

*L'***écart type** σ *de la série est la racine carrée de la variance :* $\sigma = \sqrt{V}$.

Propriété *La variance est égale à la moyenne des carrés moins le carré de la moyenne :* $V = \overline{x^2} - \overline{x}^2$

3) Description par la médiane et les quantiles

Définition *La médiane M_e d'une série statistique* **ordonnée** *est une valeur qui partage la population en deux groupes de même effectif.*

Si l'effectif total de la série est impair : $N = 2p + 1$, la médiane est la $(p+1)^{ème}$ valeur.
Si l'effectif est pair : $N = 2p$, on prend en général pour médiane la moyenne de la $p^{ème}$ et de la $(p+1)^{ème}$ valeur.

Le **mode** *d'une série statistique est la valeur du caractère la plus fréquente.*

Exercice 1 1) Dans une petite société, le patron gagne chaque mois 10 000 euros et ses 9 employés gagnent eux 1500 euros. Quel est le salaire moyen dans l'entreprise ? Le salaire médian ?

2) Rechercher les montants des salaires moyen et médian en France. Commenter.

De même que pour la moyenne, la médiane est une **caractéristique de position** et ne rend pas compte de la **dispersion** des valeurs. Pour décrire une série statistique, on doit donc en plus caractériser la dispersion des valeurs autour de cette position.

Définition *L'***étendue** *d'une série est l'écart entre les valeurs extrêmes de la série.*
Les **quartiles** Q_1, Q_2 *et* Q_3 *d'une série sont trois valeurs de la série ordonnée qui la partagent en quatre séries de même effectif (25% de l'effectif total).*
Le deuxième quartile est la médiane : $Q_2 = M_e$.
L'écart inter-quartile est le nombre $Q_3 - Q_1$.

On définit de la même façon les **déciles** D_1, D_2,...,D_9 *d'une série, en partageant la série en dix séries de même effectif (10% de l'effectif total).*
L'écart inter-décile est le nombre $D_9 - D_1$.

Remarque : Dans le cas d'une série statistique continue, on regroupe les valeurs en classes (ou intervalles). Les indicateurs statistiques sont alors calculés en utilisant le centre des classes.

Exercice 2 Soit la série statistique :

Notes x_i	6	8	10	12	15	18
Nombre d'élèves n_i	1	5	3	4	2	2

La moyenne de cette série est $\overline{x} = \ldots$

La variance est : $V = \ldots$

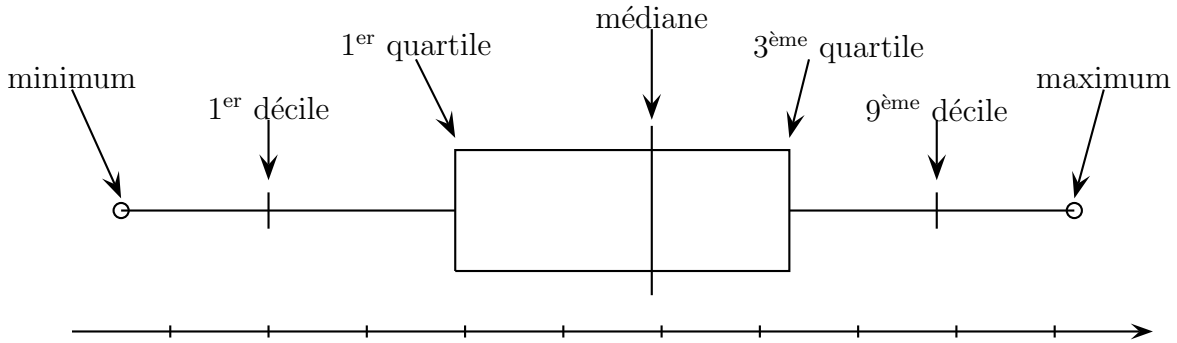
et l'écart-type : $\sigma = \ldots$

L'effectif total est $N = \ldots$.
La médiane est donc la $\ldots^{\text{ème}}$ valeur de la série ordonnée, soit $M_e = \ldots$

Son mode est 8.

4) Diagramme en boîte

La représentation d'une série à l'aide d'un diagramme en boîte (ou diagramme à pattes, ou boîte à moustaches, ou Whiskers plots) repose sur la description de la série par ses quantiles.
Cette représentation a été introduite en 1977 par John Tukey [1].



Exercice 3 Soit la série statistique :

Longueur x_i (mm)	4.7	4.8	4.9	5.0	5.1	5.2	5.3
Effectifs n_i	1	4	23	30	27	9	6

1. Calculer la moyenne et l'écart-type de cette série.
2. Déterminer la médiane, l'étendue, et les écarts inter-quartiles et inter-deciles de cette série. Représenter alors le diagramme en boîte de cette série.

1. John Wilder Tukey (16 juin 1915 - 26 juillet 2000) est un important statisticien américains. Il a créé et développé de nombreuses méthodes statistiques. Il est notamment connu pour son développement en 1965, avec James Cooley, de l'algorithme de la transformée de Fourier rapide (*fft*).

Exercice 4 On mesure, en millimètres, le diamètre de 100 pièces prises au hasard dans la production d'une machine. On obtient les résultats ci-contre.

Soit σ l'écart type de cette série statistique. Un réglage de la machine est nécessaire lorsque $\sigma > 0,013$. Faut-il régler la machine ?

Diamètre x_i (mm)	Effectifs n_i
80,36	8
80,37	19
80,38	55
80,39	36
80,40	10
80,41	11
80,42	5

II - Série statistique à deux variables - Ajustement affine

On s'intéresse à l'étude, sur une population donnée, du lien qui peut exister entre deux caractères.

Valeurs du 1er caractère x_i	x_1	x_2	x_3	...	x_k
Valeurs du 2ème caractère y_i	y_1	y_2	y_3	...	y_k

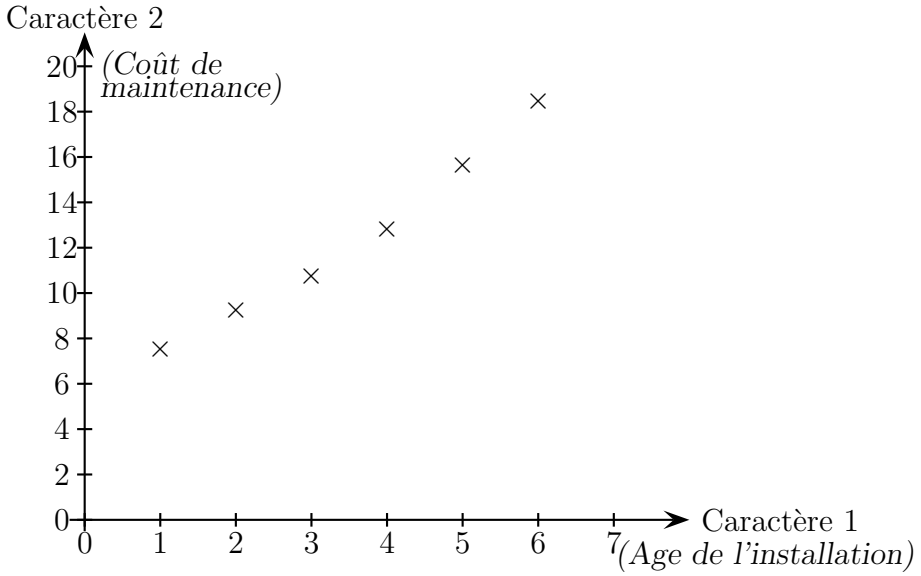
Exemple : L'étude du coût de maintenance annuel d'une installation de chauffage dans un immeuble de bureaux, en fonction de l'âge de l'installation, a donné les résultats suivants :

Age x_i (années)	1	2	3	4	5	6
Coût y_i (k€)	7,55	9,24	10,74	12,84	15,66	18,45

Objectifs Y'a-t-il un lien crédible entre l'âge de l'installation et le coût de maintenance ? Si oui, peut-on le quantifier, et peut-on, par exemple, prévoir le coût de maintenance d'une installation de 7 ans ? 8 ans ? 10 ans ?

1) Représentation graphique - Nuage de points

On appelle **nuage de points**, l'ensemble des points A_i de coordonnées $(x_i; y_i)$.



Définition *Le* **point moyen** *du nuage de points est le point de coordonnées* $(\overline{x}; \overline{y})$.

Exemple : Dans l'exemple précédent, le point moyen G a pour coordonnées $(3,5\,;\,12,41)$.

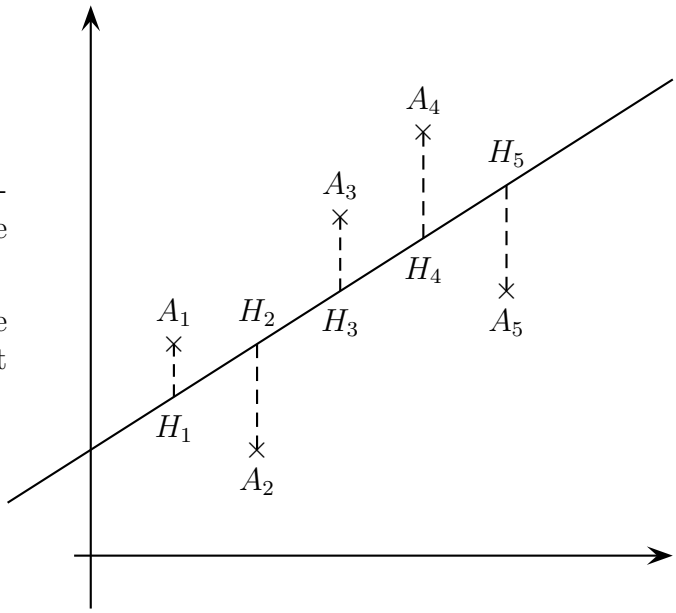
2) Ajustement affine par la méthode des moindres carrés

Les points de l'exemple précédents ne sont pas alignés. Néanmoins, ces points semblent se distribuer approximativement autour d'une droite.

La méthode des moindres carrés permet de déterminer l'équation de la "meilleure" droite passant dans le nuage de points, ainsi que de quantifier la "qualité de l'alignement des points" du nuage.

On considère un nuage de points $A_k(x_k; y_k)$.
Pour une droite quelconque, on peut définir la "distance" de la droite au nuage de points par la somme des distances A_kH_k.

Ainsi, la "meilleure" droite passant dans le nuage de points est celle dont la distance au nuage de points est la plus petite.



Propriété *Il existe une unique droite telle que la somme des distances*

$$d = A_1H_1^2 + A_2H_2^2 + \cdots + A_nH_n^2 = \sum_{k=1}^{n} A_kH_k^2$$

soit minimale. Cette droite est appelée **droite de régression de** y **en** x*, ou encore* **droite des moindres carrés**.

Cette droite de régression passe par le point moyen $G(\overline{x}; \overline{y})$.

La calculatrice, ou un tableur, permet de calculer l'équation de la droite de régression.

Exercice 5 La droite de régression de l'exemple précédent a pour équation $y = 2,17x + 4,83$.

Retrouver cette équation à l'aide de la calculatrice.

Estimer à partir de ce modèle le coût de maintenance pour une installation de 7 ans, de 8 ans, puis de 10 ans.

3) Coefficient de corrélation

La droite de régression est la droite la plus proche de tous les points du nuage. Néanmoins, l'idée d'approcher tous les points du nuage par une droite peut-être plus ou moins pertinent.

Le coefficient de corrélation est un nombre qui quantifie justement ce degré de pertinence.

Propriété

- *Le coefficient de corrélation* r *prend des valeurs entre* -1 *et* 1 *:* $-1 \leqslant r \leqslant 1$
- r *a le même signe que le coefficient directeur de la droite de régression.*
- *La corrélation est d'autant meilleure que* $|r|$ *est proche de* 1 *(si* $r = 1$ *ou* $r = -1$*, les points sont alignés et la corrélation est parfaite).*

Exercice 6 **Temps de chargement et fréquentation d'un site web**

Le temps de chargement d'une page sur internet dépend de nombreux paramètres, entre autre le nombre d'utilisateurs qui y sont connectés simultanément.

Par ailleurs, le temps de chargement influe en retour sur le nombre de visiteurs : plus le temps de chargement est long, plus les utilisateurs sont susceptibles de se diriger vers d'autres ressources.

Le responsable d'un site a relevé le nombre d'internautes sur son site en fonction de sa durée de chargement :

Nombre d'internautes connectés (en millier), x_i	0,5	1	2,5	3	4	5	6
Durée de chargement (en secondes), y_i	0,3	0,4	0,6	0,9	1,3	2	2,8

1. Représenter le nuage de points de coordonnées $(x_i; y_i)$ associés à cette série statistique. (Axes orthogonaux ; unités : 2 cm pour 1000 internautes et 1cm pour 0,2 seconde).
2. À l'aide de la calculatrice, déterminer l'équation $y = ax + b$ de la droite d'ajustement $\mathcal{D}$ obtenue par la méthode des moindres carrés (Arrondir les coefficients au millième).
3. Pour la suite, on prendra $y = 0,44x - 0,19$ pour équation de la droite $\mathcal{D}$.
   a) Tracer la droite $\mathcal{D}$.
   b) Avec ce modèle, estimer la durée de chargement pour 8000 personnes connectées.
   c) Une étude indépendante a montré que 60% des internautes cesse de charger une page pour se diriger vers un autre site dès que le temps de chargement dépasse 3,5 secondes.

      Avec le modèle précédent, estimer le nombre de visiteurs sur ce site lorsque la durée de chargement est de 3,5 secondes.

      Combien de visiteurs perdrait-il alors ?

4) Remarque fondamentale : Corréler n'est pas expliquer

Une erreur (malheureusement) assez répandue consiste à confondre corrélation avec causalité.

Observer que deux variables sont corrélées entre elles ne signifie pas que l'une soit la cause ou la conséquence de l'autre, c'est-à-dire qu'il y ait un lien de cause à effet.

Par exemple en France au 20ème siècle, le nombre de mariages a augmenté ainsi que le nombre de suicides. Ces deux variables sont sûrement corrélées, ce qui ne montre en aucun cas l'existence d'un lien de cause à effet d'un phénomène à l'autre (en fait ces deux augmentations peuvent être directement reliées à une variable commune, ici cachée : l'augmentation de la démographie).

Exercice 7 Le tableau suivant donne les évolutions, de Mai à Septembre, du nombre de climatiseurs

vendus et de noyade par accident dans un secteur littoral.

Mois	Nombre de climatiseurs x_i	Nombre de noyades y_i
Mai	66	1
Juin	88	3
Juillet	90	5
Août	110	8
Septembre	60	0

1. Représenter graphique le nuage de points correspondant.
2. Un ajustement affine semble-t'il pertinent ?
   Donner l'équation de la droite d'ajustement par moindres carrés.
3. Prévoir le nombre de noyades si en Avril de l'année d'après, 88 climatiseurs sont vendus ?
4. Commenter la relation de causalité entre les variables étudiées.

Exercice 8 **Équilibrer offre et demande**

Une étude statistique effectué sur un produit à permis de quantifier l'offre et la demande de ce produit, pour différentes valeurs de son prix unitaire.

Prix unitaire (en euros) x_i	Demande (en milliers d'unités) y_i	Offre (en milliers d'unités) z_i
1,2	8,4	0,75
2,5	6	1,25
3,5	5	1,75
4,5	4,2	2,25
5	3,5	2,5
7	2,1	3,5
8,5	1,2	4,25

1. Représenter graphiquement, sur un même graphique, les nuages de points des séries à deux variables :
   — la demande en fonction du prix unitaire (série à deux variables x et y)
   — l'offre en fonction du prix unitaire (série à deux variables x et z)
2. Tracer, "au jugé" une droite d'ajustement pour chacun des deux nuages.
3. Estimer graphiquement,
   a) la demande et l'offre pour un prix unitaire de 6 euros ;
   b) le prix unitaire pour une demande de 8 milliers ;
   c) le prix unitaire pour une offre de 1 millier.
4. Déterminer, à l'aide de la calulatrice, les deux équations des droites d'ajustement par moindres carrés.
   Tracer alors ces deux droites, et préciser les estimations de la question précédente.
5. Pour quel prix unitaire y a-t'il équilibre entre l'offre et la demande ?

Exercice 9 Le tableau suivant donne le nombre de clients annuel, en millier, d'une nouvelle chaîne de magasins.

Année	2006	2007	2008	2009	2010	2011	2012	2013
Rang de l'année	0	1	2	3	4	5	6	7
Nombre de clients	11,2	20,6	29,7	37,0	39,6	41,7	44,5	48,0

Représenter le nuage de points $(x_i; y_i)$ du nombre de clients en fonction du rang de l'année.

Partie A. Ajustement affine.

1. Déterminer une équation de la droite d'ajustement obtenue par la méthode des moindres carrés.

   Pour la suite, on utilisera l'ajustement affine donné par la droite D d'équation $y = 4,9x + 16,7$.
2. Tracer la droite D sur le nuage de points précédent.
3. Prévoir à l'aide de ce modèle le nombre de clients en 2015 et 2016.

Partie B. Ajustement par une fonction logistique.

Un autre ajustement est obtenu à l'aide de la fonction f définie par $f(x) = \dfrac{52}{1+3e^{-0,6x}}$.

1. Compléter le tableau de valeurs :

Rang de l'année x	0	1	2	3	4	5	6	7
$f(x)$		19,6	27,3					

2. Tracer l'allure de la courbe $\mathcal{C}$ représentative de la fonction f sur le nuage de points précédent.
3. Donner à l'aide de ce modèle le nombre de clients estimé en 2015 et 2016.

Exercice 10 **Durée de vie et maintenance d'équipements.**

Les pourcentages $R(t_i)$ des appareils mécaniques encore en service après un nombre t_i d'heures de fonctionnement ont été relevés et notés dans le tableau suivant :

t_i	100	200	300	400	500	600	750	1000	1500
$R(t_i)$	0,80	0,64	0,52	0,40	0,32	0,28	0,20	0,12	0,04

1. On pose $y_i = \ln R(t_i)$. Représenter graphiquement le nuage de points M_i de coordonnées $(t_i; y_i)$.
2. Peut-on envisager un ajustement affine de ce nuage de points ?
   Donner l'équation de la droite de régression de y en t.
   En déduire une expression de la forme $R(t) = ke^{-\lambda t}$, avec k et λ des constantes.
3. Déterminer à l'aide du modèle précédent, le nombre d'équipements encore en service au bout de 900 heures de fonctionnement.

Exercice 11 Le tableau suivant donne la durée moyenne d'intervention, en minutes, sur les postes de télévision en panne dans un atelier de dépannage, de 1992 à 2000.

Rang x_i	1	2	3	4	5	6	7	8	9
Année	1992	1993	1994	1995	1996	1997	1998	1999	2000
Durré moyenne d_i	83	82	80	75	73	74	71	71	70

Partie A. Ajustement à l'aide de la droite de régression

1. Calculer le coefficient de corrélation linéaire entre x et d (à 10^{-3} près).
   Semble-t-il y avoir une dépendance affine entre l'année et la durée moyenne des interventions ?
2. Donner la droite de régression des moindres carrés de d en x (valeurs arrondies à 10^{-2} près).
3. En supposant que l'évolution se poursuit ainsi pendant les 5 années futures, estimer la durée moyenne d'intervention dans cet atelier en 2002.

Partie B. Ajustement à l'aide de la droite de Mayer

La méthode de Mayer consiste à partager la série en 2. Soit S_1 la série correspondant aux années 1992-1996, et S_2 la série correspondant aux années 1997-2000.

1. Calculer les coordonnées du point moyen G_1 de la série S_1, et du point moyen G_2 de la série S_2.
2. Déterminer l'équation de la droite (G_1G_2) appelée droite de Mayer.
3. Estimer la durée moyenne d'intervention dans cet atelier en 2002 avec la droite de Mayer et comparer avec la droite de régression.

Exercice 12 Après un accident nucléaire, on procède à intervalles de temps réguliers à des mesures de radioactivité sur un site donné. Le tableau suivant donne les résultats de ces mesures.

Rang x_i de la mesure	1	2	3	4	5	6
Valeur y_i mesurée	100	61	37	22	14	7

1. Tracer le nuage de points correspondant. Utiliser une droite de régression linéaire semble-t'il pertinent (justifier en donnant le coefficient de corrélation de cette droite).
2. Pour chaque mesure on pose $z_i = \ln y_i$ et on étudie alors la série statistique $(x_i; z_i)$.
   Compléter le tableau :

Rang x_i de la mesure	1	2	3	4	5	6
$z_i = \ln y_i$						

3. Calculer le coefficient de corrélation de cette série à 0,001 près. Commenter le résultat.
4. Donner une équation de la droite D de régression de z en x (arrondir les coefficients à 0,01 près).
5. En déduire une relation entre x et y du type $y = \alpha e^{\beta x}$, où α et β sont deux constantes à déterminer.
6. En supposant que le modèle reste valable, en déduire pour la prochaine mesure ($x_i = 7$) une estimation de y.
7. En supposant toujours que le modèle reste valable, déteminer à partir de quelle mesure la valeur y mesurée sera inférieure à $0,01$.

Exercice 13 Au cours d'une séance d'essai, un pilote automobile doit, quand il reçoit un signal sonore dans son casque, arrêter le plus rapidement possible son véhicule. Au moment du top sonore, on mesure la vitesse de l'automobile puis la distance nécessaire pour arrêter le véhicule.
Pour six expériences, on a obtenu les résultats suivants :

v_i (km/h)	27	43	62	80	98	115
distance y_i d'arrêt (m)	6,8	20,5	35,9	67,8	101,2	135,8

On pose $x_i = v_i^2$ et on considère la série $(x_i; y_i)$.

1. Compléter le tableau

x_i						
y_i	6,8	20,5	35,9	67,8	101,2	135,8

2. Dans un repère orthogonal représenter le nuage de points associé à cette nouvelle série (unités : 1cm pour 1000 en abscisse, et 1 cm pour 10 en ordonnée).
3. a. Déterminer, à l'aide de la calculatrice, l'équation de la droite de régression de y en x sous la forme $y = mx + p$. Tracer cette droite dans le repère précédent.
   b. A l'aide de cette équation, déterminer la valeur estimée de x correspondant à une distance d'arrêt de 180 m, puis la vitesse correspondante du véhicule.
   c. Quelle est la vitesse d'arrêt estimée correspondant à une vitesse de 150 km/h.
   d. Le manuel du code de la route donne, pour calculer la distance d'arrêt, en mètres, la méthode suivante : "Prendre le carré de la vitesse exprimé en dizaines de kilomètres par heure."
   Comparer le résultat obtenu au c. à celui que l'on obtiendrait par cette méthode.