Statistiques à deux variables

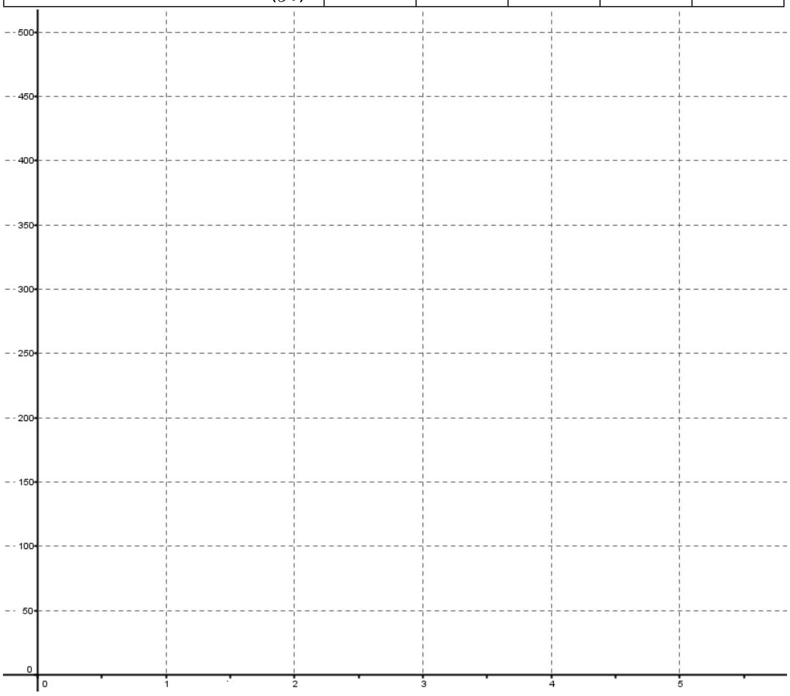
1. Nuage de points et point moyen

On s'intéresse à deux caractères d'une population, notés x et y.
On obtient deux listes de nombres, qui sont deux séries statistiques.
On peut alors tracer un ***nuage de points*** en prenant en abscisse les valeurs x_i de la première série statistique et en ordonnée les valeurs y_i de la deuxième.

A. Nuage de points

Exercice 9.1 Une salle de musculation a ouvert ses portes en 2008.
Le tableau suivant donne le nombre d'adhérents chaque année :

Année	2008	2009	2010	2011	2012
Rang de l'année (x_i)	1	2	3	4	5
Nombre d'adhérents (y_i)	126	201	275	375	438



Pour les plus rapides : Tracer une droite qui passe "à peu près" par les points obtenus, et essayer de donner une équation approximative de cette droite (par lecture graphique).

. . . .

Définition 9.1 Les points obtenus s'appellent un ***nuage de points*** .

B. Point moyen

<u>Exercice</u> 9.2 1. Avec les deux séries de données (x_i et y_i) de l'exercice précédent, calculer :

(a) La moyenne de la série (x_i), la noter $\overline{x}$

....

....

(b) La moyenne de la série (y_i), la noter $\overline{y}$

....

....

2. Placer le point de coordonnées ($\overline{x}$; $\overline{y}$) sur le graphique précédent, noter ce point G.

Pour les plus rapides : Modifier légèrement l'équation de la droite trouvée précédemment pour qu'elle passe par G.

Définition 9.2 Le point G de coordonnées ($\overline{x}$; $\overline{y}$) s'appelle le ***point moyen*** .

C. Exercices bilan

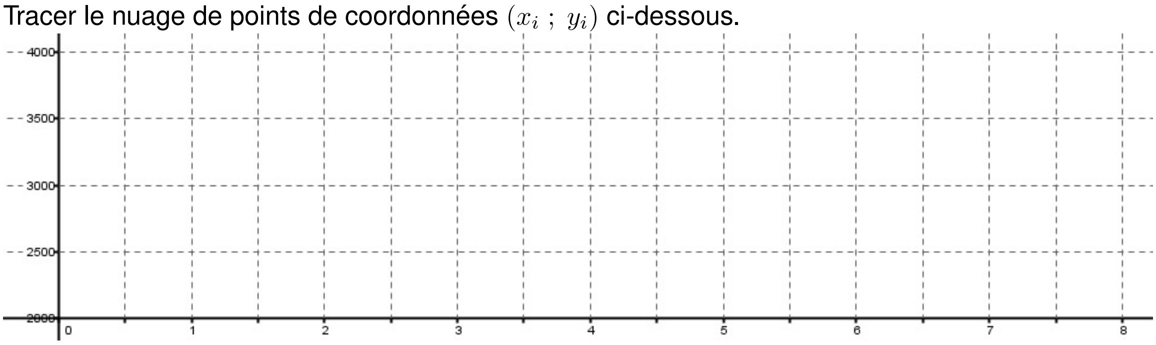
<u>Exercice</u> 9.3 On s'intéresse à l'évolution du prix des appartements neufs en France métropolitaine.

Le tableau ci-dessous indique le prix des appartements neufs en France métropolitaine, en euros par m^2, entre 2004 et 2012.

Année	2004	2005	2006	2007	2008	2009	2010	2011	2012
Rang de l'année : x_i	0	1	2	3	4	5	6	7	8
Prix de l'appartement (e/m^2) : y_i	2563	2852	3071	3276	3344	3368	3571	3773	3861

Source Insee

1. Tracer le nuage de points de coordonnées $(x_i\ ;\ y_i)$ ci-dessous.



2. Calculer les coordonnées du point moyen, et le placer en couleur sur le graphique ci-dessus.

....

....

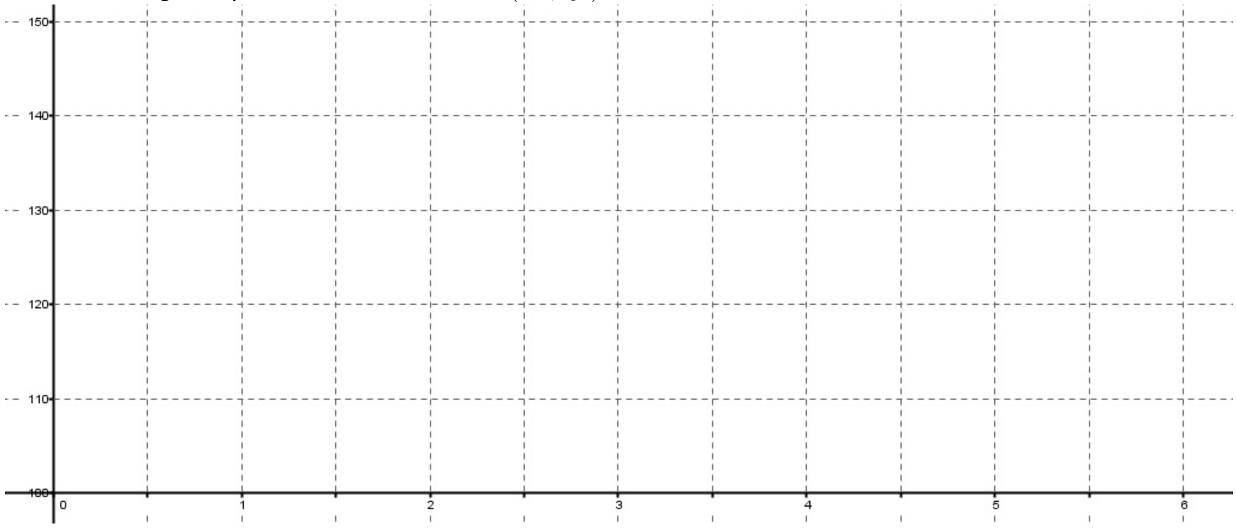
....

....

<u>Exercice</u> 9.4 Le tableau ci-dessous donne le nombre de voitures neuves (en milliers) vendues en France durant les six premiers mois de l'année 2013.

Mois	Janvier	Février	Mars	Avril	Mai	Juin
Rang du mois x_i	1	2	3	4	5	6
Nombre de ventes (en milliers) y_i	149	144	150	140	139	135

1. Tracer le nuage de points de coordonnées $(x_i\ ;\ y_i)$ ci-dessous.

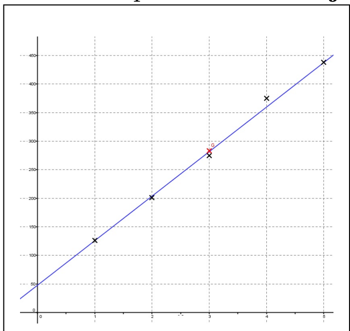
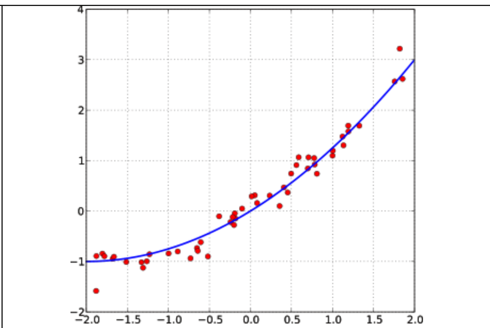
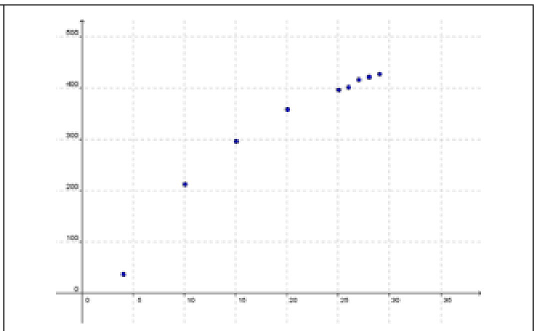


2. Calculer les coordonnées du point moyen, et le placer en couleur sur le graphique ci-dessus.

....
....
....
....

2. Ajustement affine

Lorsque le nuage de point est trop "désorganisé", on ne peut pas forcément trouver de courbe qui le représente. Cependant il arrive que l'on puisse déterminer une courbe qui traduise ce nuage de point, dans ce cas on dit que l'on fait un **ajustement**.

		
Ajustement affine	Ajustement exponentiel ou parabolique	Ajustement logarithmique ou racine carrée

Lorsque les points du nuage de points sont "presque alignés", on peut chercher une droite D d'équation $y = ax + b$ qui passe "le plus près possible" des points du nuage.
On dit alors que l'on fait un **ajustement affine**.

Il existe plusieurs façons d'obtenir cette droite (droite des points extrêmes, droite de Mayer...), dont la **méthode des moindres carrés** détaillée dans ce cours.

Définition 9.3 Le **coefficient de corrélation** est le nombre

$$r = \frac{cov(X;Y)}{\sqrt{V(X) \times V(Y)}}$$

On a toujours $-1 \leqslant r \leqslant 1$

Propriété 9.1
- Si $|r| \simeq 1$, alors les nuages du point sont quasiment alignés, l'ajustement affine par les moindres carrés est pertinent.
- Si $r \simeq 0$, alors les points sont très dispersés, l'ajustement affine n'est pas adapté à la situation.

A. Covariance de deux séries statistiques

On rappelle la formule de la variance : $V(X) = \frac{x_1^2+x_2^2+x_3^2+...}{N} - \overline{x}^2$
La covariance d'une série statistique à deux variables se calcule ainsi :

$$cov(x,y) = \frac{x_1y_1 + x_2y_2 + x_3y_3 + ... + x_ny_n}{n} - \overline{xy}$$

Exemple :

Année	2008	2009	2010	2011	2012
Rang de l'année (x_i)	1	2	3	4	5
Nombre d'adhérents (y_i)	126	201	275	375	438

On a vu précédemment que $\overline{x} = 3$ et que $\overline{y} = 283$

La **covariance** est :
....
....
....
....

Exercice 9.5 Calculer la covariance de la série statistique à deux variables ci-dessous :

Année	2004	2005	2006	2007	2008	2009	2010	2011	2012
Rang de l'année : x_i	0	1	2	3	4	5	6	7	8
Prix de l'appartement (e/m^2) : y_i	2563	2852	3071	3276	3344	3368	3571	3773	3861

....
....
....
....

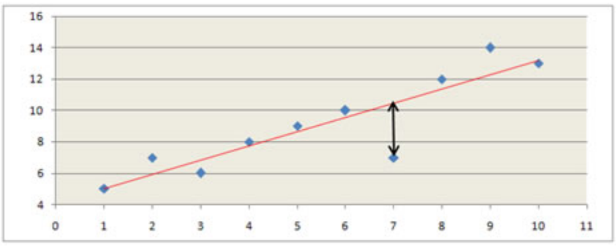
Exercice 9.6 Calculer la covariance de la série statistique à deux variables ci-dessous :

Mois	Janvier	Février	Mars	Avril	Mai	Juin
Rang du mois x_i	1	2	3	4	5	6
Nombre de ventes (en milliers) y_i	149	144	150	140	139	135

....
....
....
....

B. Méthode des moindres carrés

Cette méthode donne l'équation $y = ax + b$ d'une droite D.
C'est la droite pour laquelle les *carrés des distances* (mesurées verticalement) entre les points du nuage et D sont les plus petits.



L'équation de cette droite s'obtient ainsi :

- Le coefficient directeur est $a = \frac{cov(x,y)}{V(x)}$, où $cov(x,y)$ est la covariance, et où $V(x)$ est la variance de la série statistique x_i
- L'ordonnée à l'origine b est choisie de telle manière que **le point moyen** $G(\overline{x};\overline{y})$ **appartienne à la droite**.

Exemple :

Année	2008	2009	2010	2011	2012
Rang de l'année (x_i)	1	2	3	4	5
Nombre d'adhérents (y_i)	126	201	275	375	438

On a vu que la covariance est $cov(x, y) = 159,6$, que $\overline{x} = 3$ et que $\overline{y} = 283$.
On rappelle la formule de la variance : $V(X) = \frac{x_1^2+x_2^2+x_3^2+...}{N} - \overline{x}^2$ (les n_1, n_2, n_3... n'apparaissent pas car dans la série statistique x_i, chaque valeur n'apparaît qu'une fois).
Déterminer $V(X)$:

....

....

Déterminer la covariance $cov(x, y)$:

....

....

Déterminer le coefficient directeur a de la **droite des moindres carrés** :

....

....

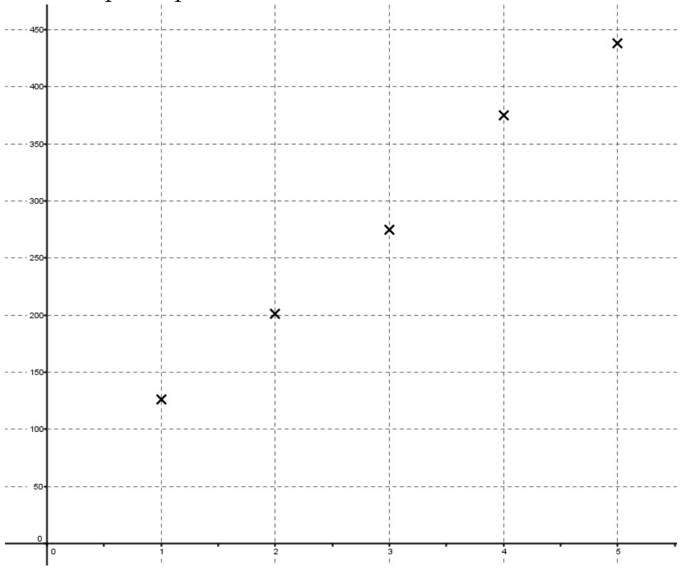
Déterminer l'ordonnée à l'origine b de la **droite des moindres carrés**, en utilisant les coordonnées du point moyen :

....

....

....

Tracer cette droite sur le graphique ci-dessous, placer également le point moyen G et vérifier que la droite obtenue passe par G :



Exercice 9.7 On reprend la série statistique à deux variables :

Année	2004	2005	2006	2007	2008	2009	2010	2011	2012
Rang de l'année : x_i	0	1	2	3	4	5	6	7	8
Prix de l'appartement (e/m^2) : y_i	2563	2852	3071	3276	3344	3368	3571	3773	3861

On a vu que la covariance est $cov(x, y) = 1005,09$, que $\overline{x} = 4$ et que $\overline{y} = 3297,7$.
On rappelle la formule de la variance : $V(X) = \frac{x_1^2+x_2^2+x_3^2+...}{N} - \overline{x}^2$ (les n_1, n_2, n_3... n'apparaissent pas car dans la série statistique x_i, chaque valeur n'apparaît qu'une fois).
Déterminer $V(X)$ et $V(Y)$:

....

....

....

....

....

....

Un ajustement affine est-il envisageable ? :

....

....

....

....

....

....

....

. . . .

Déterminer le coefficient directeur a de la **droite des moindres carrés** :

. . . .

. . . .

. . . .

Déterminer l'ordonnée à l'origine b de la **droite des moindres carrés**, en utilisant les coordonnées du point moyen :

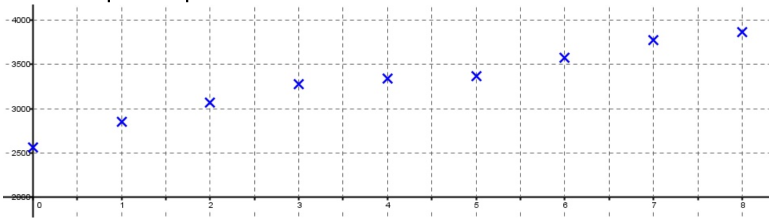
. . . .

. . . .

. . . .

. . . .

Tracer cette droite sur le graphique ci-dessous, placer également le point moyen G et vérifier que la droite obtenue passe par G :



C. A la calculatrice

Heureusement, l'équation de cette droite des moindres carrés peut être obtenue à la calculatrice.
Un exemple de vidéo sur Youtube illustrant cette utilisation :

`https://youtu.be/DQ0Lwl3rk88`

... mais il y a beaucoup d'autres tutoriels à ce sujet.

Pour les calculatrices TI

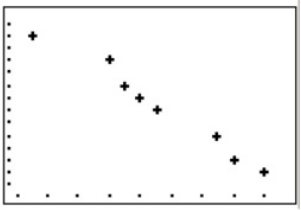
- Après avoir appuyé sur la touche *stats*, sélectionner *EDIT*
- Entrer alors les données x_i dans la colonne *L1*, et les données y_i dans la colonne *L2*
- Appuyer sur la touche *stats* ; sélectionner *CALC*, puis *RégLin(ax+b)*
- Vérifier que la calculatrice a bien sélectionné *XList :L1* et *YList :L2* (ne rien écrire sur les lignes suivantes), puis valider : on obtient les coefficient a et b.

Représenter le nuage de points avec une TI

- Configurer la calculatrice pour les graphiques statistiques :*2nd STAT PLOT* ou *2nd graph stats* puis *Enter* et valider les choix représentés ci-dessous :

- Régler la fenêtre d'affichage : *WINDOW*, puis *Zoom 9 : Zoomstats*

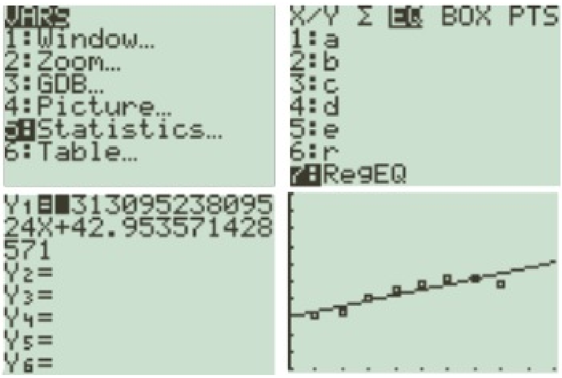
- Afficher le nuage de points : *GRAPH*



Représenter la droite des moindres carrés avec une TI

Avec les mêmes réglages que ci-dessus pour le nuages de points, et après avoir calculé les coefficient a et b qui donnent l'équation de la droite (voir ci-dessus).

- Appuyer sur la touche *f(x)* en haut à gauche
- Se placer sur *Y1=*
- Appuyer sur *var* , choisir *5 :Statistiques*
- Sélectionner l'onglet *EQ*, puis la ligne *EqReg*
- La calculatrice revient automatiquement à l'écran "$Y1 = ...$", et a écrit l'équation de la droite des moindres carrés sur la ligne $Y1$.
- Appuyer sur *graphe*



Pour les calculatrices Casio

- Sélectionner le menu *Stat*
- Entrer les données x_i en *List1*, et les données y_i en *List2*
- Vérifier le réglage de la calculatrice : *CALC*, puis *SET* : *List1* doit être écrit devant *2Var XList*, et *List2* doit être écrit devant *2Var YList*
- Appuyer deux fois sur *EXIT* pour revenir aux listes
- Pour obtenir les coefficient a et b de l'équation de la droite : *CALC*, *REG*, *X* puis $ax + b$.

Pour les calculatrices Numworks

- Sélectionner le menu *Regression*
- Entrer les données x_i en *X1*, et les données y_i en *Y1*
- Aller dans l'onglet *Graph*, vous obtenez le graphique, et en-dessous, les valeurs de $\overline{x}, \overline{y}$ (point moyen), a, et b (coefficients de l'équation $ax + b$ de la droite des moindres carrés).

Exercice 9.8 Le tableau ci-dessous donne le nombre de voitures neuves (en milliers) vendues en France durant les six premiers mois de l'année 2013.

Mois	Janvier	Février	Mars	Avril	Mai	Juin
Rang du mois x_i	1	2	3	4	5	6
Nombre de ventes (en milliers) y_i	149	144	150	140	139	135

1. (a) Représenter le nuage de points de la série $(x_i\ ;\ y_i)$ dans le repère fourni en annexe 1.
   (b) Expliquer pourquoi ce nuage de points permet d'envisager un ajustement affine.
   ....
   ....
   ....
   ....

2. Déterminer une équation de la droite D d'ajustement affine de y en x obtenue par la méthode des moindres carrés. On arrondira au centième les coefficients.

. . . .

. . . .

. . . .

. . . .

3. On décide de modéliser l'évolution du nombre y de ventes de voitures neuves en fonction du rang x du mois par l'expression $y = -2,7x + 152$.

   (a) Représenter graphiquement, dans le repère fourni en annexe 1, la droite traduisant cette évolution.

   (b) Quel nombre de ventes de voitures neuves pouvait-on prévoir pour le mois de décembre 2013 en utilisant ce modèle ?

   . . . .

   . . . .

   . . . .

   . . . .

   (c) À partir de quel mois pouvait-on prévoir que le nombre de voitures neuves en France serait strictement inférieur à 130000 véhicules ?

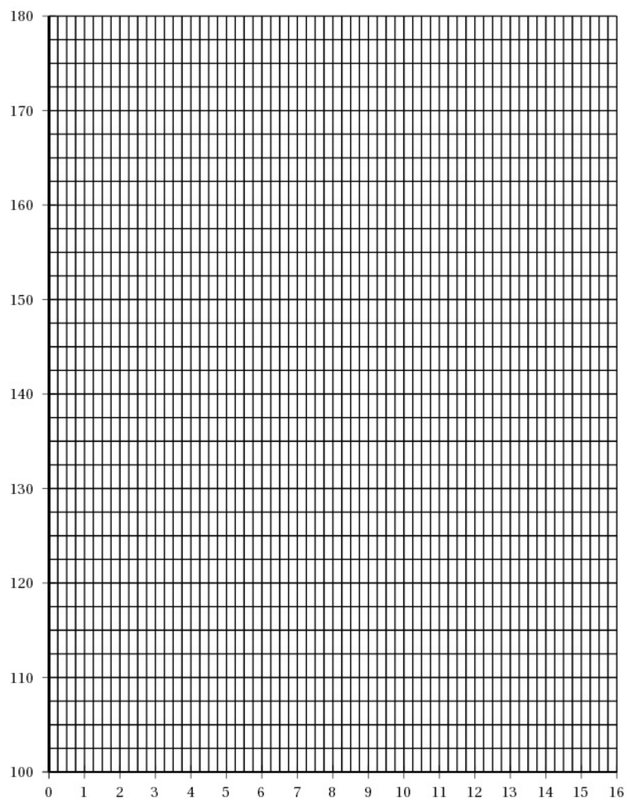
   . . . .

   . . . .

   . . . .

   . . . .

Annexe 1 :



D. Ajustement et changement de variable

Quand le nuage de points d'une série statistique ne suggère pas un ajustement affine, il arrive qu'un "changement de variable" permette tout de même d'avoir recours à un ajustement affine.
En "remontant le changement de variable", on obtient un ajustement du nuage initial (exponentiel, logarithmique, parabolique etc...).

Exercice 9.9 Un restaurateur effectue une étude de marché dans l'intention de d'installer un "grand buffet à volonté", avec des tarifs allant de 15 à 35 euros. Voici le résultat de cette étude, où y_i représente le nombre de couverts potentiel en fonction du tarif x_i (en euros).

x_i	15	17	20	22.5	25	30	35
y_i	500	350	270	190	140	75	50

1. Représenter graphiquement, à la calculatrice si possible, le nuage de points correspondant. Calculer le coefficient de corrélation : un ajustement affine est-il adapté ?
   ....
   ....
   ....
   ....
2. (a) Compléter le tableau suivant, en arrondissant à 10^{-2}.

x_i	15	17	20	22.5	25	30	35
$z_i = \ln(y_i)$	....	....	....	....	....	....	....

   (b) Calculer le nouveau coefficient de corrélation. Un ajustement affine est-il adapté ? Si oui, déterminer à la calculatrice l'équation de la droite des moindres carrés.
   ....
   ....
   ....
   ....

   (c) En déduire que la fonction f définie sur $[15; 35]$ par $f(x) = 2639e^{-0.116x}$ permet de modéliser le nombre de couverts en fonction du tarif.
   ....
   ....
   ....
   ....
3. Le restaurateur estime que si moins de 100 couverts sont servis, son projet ne sera pas viable. Déterminer le prix maximum qu'il doit fixer pour assurer la rentabilité de son établissement.
   ....
   ....
   ....
   ....
   ....
   ....
   ....
   ....