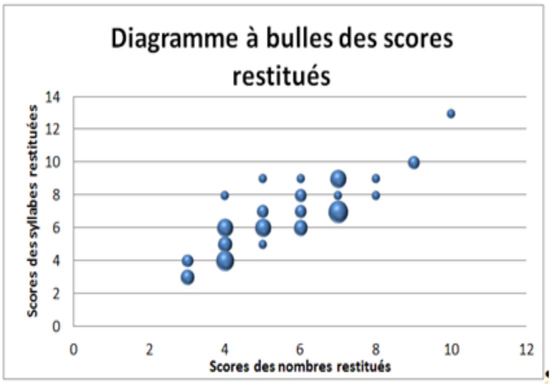
Fiche de TD N° 07
Statistique descriptive bi-variée

Exercice 1

On propose 2 tests de mémorisation à des étudiants. Pour le premier, on présente une série de 30 nombres qu'ils doivent apprendre en un temps limité. On leur demande de rappeler le maximum de nombres une minute après. Pour le second, on leur présente une série de 30 syllabes qu'ils devront également apprendre et restituer. Pour chacun de ces tests, on recueille dans un tableau le nombre d'éléments rappelés après la minute de pause. Chaque sujet est caractérisé par deux scores: un score X au test des nombres et un score Y au test de syllabes.

Les résultats portant sur 50 sujets sont reportés dans le tableau ci-dessous, et représentés par le diagramme à bulles.

S	X	Y	S	X	Y	S	X	Y	S	X	Y	S	X	Y
1	5	7	11	3	3	21	8	9	31	4	4	41	7	7
2	4	5	12	4	6	22	6	7	32	5	6	42	6	6
3	4	6	13	7	9	23	5	6	33	10	13	43	8	8
4	5	9	14	6	6	24	7	8	34	4	4	44	7	7
5	4	8	15	7	7	25	4	6	35	6	8	45	3	3
6	5	7	16	9	10	26	7	7	36	4	4	46	7	9
7	6	9	17	4	4	27	3	4	37	4	5	47	4	5
8	7	9	18	9	10	28	3	3	38	4	4	48	4	6
9	6	7	19	5	6	29	6	6	39	5	6	49	3	4
10	7	7	20	6	8	30	7	7	40	5	5	50	7	9



1. Décrire le graphe.
2. Dresser la table de contingence correspondante à la série double (X, Y) en explicitant les marges.
3. Calculer et interpréter les fréquences suivantes : f_{23}, f_{55}, $f_{3\bullet}$ et $f_{\bullet 4}$
4. Calculer les scores moyens des nombres et des syllabes restituées pour l'ensemble des sujets testés.
5. Calculer le score moyen des nombres des étudiants ayant un score des syllabes égal à 5.
6. Calculer le coefficient de corrélation linéaire entre X et Y. Que peut-on conclure ?
7. Déterminer les équations des droites de régression de Y en X et de X en Y, et les représenter sur le graphe.
8. Peut-on prévoir le score des nombres restitués d'un étudiant avait un score des syllabes égal à 14?

Exercice 2

Un technicien de laboratoire désire étudier l'influence d'un antibiotique sur une culture bactérienne. Il répartit dans 10 tubes des volumes égaux de culture additionnées d'une quantité X d'antibiotique, et il mesure, après incubation, la densité optique D. La densité optique permet de déterminer la concentration en bactérie du milieu de culture.

Antibiotique X	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
Densité optique D	19	21	35	38	64	66	115	130	200	210

1. a. Construire le nuage de points représentant la densité optique en fonction de la concentration d'antibiotique. Que conclure.
   b. Calculer le coefficient de corrélation linéaire entre X et D.
   c. Trouver l'expression de la droite de régression de D en X. Tracer le sur le graphique.
2. On reprend l'analyse en posant $Y = \ln(D)$
   a. Calculer les valeurs de Y on arrondira à 10^{-2} près.
   b. Un ajustement linéaire de Y en X est-il plus légitime ? Pourquoi ?
   c. En déduire le modèle ajustant le mieux les données initiales.
3. Un onzième tube est additionné d'une quantité d'antibiotique égale à 1,2. Donner une estimation de la densité optique de ce tube qu'on mesurera après incubation
4. En réalité, la densité optique a augmenté de 80% entre les deux mesures de la quantité d'antibiotique : 1 et 1,2. Déterminer l'erreur commise sur la densité optique en prenant l'estimation obtenue à la troisième question au lieu de la valeur réelle.

Exercice 3

Ajuster, par la méthode des moindres carrés, chacun des nuages de points suivants par la fonction indiquée :

1.

x_i	0	2	4	6	8	10
y_i	1,1	0,19	0,08	0,06	0,04	0,03

$$y = \frac{1}{ax + b}$$

2.

x_i	1	3	5	7	9
y_i	1,1	4,4	5,8	6,8	7,7

$$y = a \ln x + b$$

3.

x_i	1	2	3	4	5
y_i	1,6	1,4	1,7	2,3	2,7

$$y = \frac{a}{x} + bx$$

Corrigé TD N° 07

Statistique Descriptive bi variée.

Exo1 :

1/ Description du graphe :

Lorsqu'on décrit un nuage de points associé à une série statistique à deux variables, on doit en préciser :

- La forme : affine ou non.
- Sens : positive ou négative.
- Intensité : forte, modérée ou faible.
- Il faut aussi détecter la présence de valeurs aberrantes (points Très éloignés des autres points du nuage).

* Le graphe donné dans l'exercice est un Diagramme à bulles, c'est une Variante d'un graphique en nuage de points dans lequel les données sont représentées par des bulles de tailles proportionnelles aux effectifs. Il représente la distribution des 50 étudiants selon leurs scores de nombres et de syllabes.

* Ce nuage de points fait apparaître une corrélation linéaire, bonne et positive entre les deux scores. il ne semble pas y avoir des points aberrants.

2/ Tableau de contingence :

$x_i \backslash y_j$	3	4	5	6	7	8	9	10	13	$n_{i.}$	$n_{i.}x_i$	$n_{i.}x_i^2$	$\sum n_{ij}y_j$	$x_i\sum n_{ij}y_j$
3	3	2	0	0	0	0	0	0	0	5	15	45	17	51
4	0	5	3	4	0	1	0	0	0	13	52	208	67	268
5	0	0	1	4	2	0	1	0	0	8	40	200	52	260
6	0	0	0	3	2	2	1	0	0	8	48	288	57	342
7	0	0	0	0	6	1	4	0	0	11	77	539	86	602
8	0	0	0	0	0	1	1	0	0	2	16	128	17	136
9	0	0	0	0	0	0	0	2	0	2	18	162	20	180
10	0	0	0	0	0	0	0	0	1	1	10	100	13	130
$n_{.j}$	3	7	4	11	10	5	7	2	1	50	276	1670		1969
$n_{.j}y_j$	9	28	20	66	70	40	63	20	11	329				
$n_{.j}y_j^2$	27	112	100	396	490	320	567	200	169	2321				
$\sum n_{ij}x_i$	9	26	17	54	64	31	47	18	10					
$y_j\sum n_{ij}x_i$	27	104	85	324	448	248	423	180	130	1969				

3/ Calcul de fréquences :

- $f_{23} = \frac{n_{23}}{N} = \frac{3}{50} = 0{,}06$.

6 % des étudiants testés avaient un score de nombres = 4 et un score de syllabes égal à 5.

- $f_{55} = \frac{n_{55}}{N} = \frac{6}{50} = 0{,}12$.

12 % des étudiants avaient $X = 7$ et

- $f_{3.} = \frac{n_{3.}}{N} = \frac{8}{50} = 0{,}16$.

16 % des étudiants avaient un score de nombres $= 5$.

- $f_{.4} = \frac{n_{.4}}{N} = \frac{11}{50} = 0{,}22$.

22 % des étudiants avaient un score syllabes $= 6$.

4/ Calcul de $\bar{X}$, $\bar{Y}$:

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{p} n_{i.}x_i = \frac{276}{50} = \boxed{5{,}52}$$

$$\bar{Y} = \frac{1}{N}\sum_{j=1}^{q} n_{.j}y_j = \frac{329}{50} = \boxed{6{,}58}$$

5/ Calcul de : $\bar{X}_{Y=5}$

$$\bar{X}_{Y=5} = \frac{1}{n_{.3}}\sum_{i=1}^{p} n_{i3}x_i = \frac{17}{4} = \boxed{4{,}25}$$

6/ Calcul de r :

$$r = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y}$$

$$Cov(X,Y) = \frac{1}{N}\sum_{j=1}^{q}\sum_{i=1}^{p} n_{ij}x_iy_j - \bar{X}\cdot\bar{Y}$$

$$= \frac{1}{N}\sum_{j=1}^{q} y_j \sum_{i=1}^{p} n_{ij}x_i - \bar{X}\bar{Y}$$

ou bien égal aussi

$$= \frac{1}{N}\sum_{i=1}^{p} x_i \sum_{j=1}^{q} n_{ij}y_j - \bar{X}\bar{Y}$$

$$= \frac{1969}{50} - 5{,}52 \times 6{,}58$$

$$= \boxed{3{,}0584}$$

$$V(X) = \frac{1}{N}\sum_{i=1}^{p} n_{i.}x_i^2 - \bar{X}^2 = \frac{1670}{50} - 5{,}52^2 = \boxed{2{,}9296}$$

$$\Rightarrow \sigma_X = \sqrt{V(X)} = \boxed{1{,}71}$$

$$V(Y) = \frac{1}{N}\sum_{j=1}^{q} n_{.j}y_j^2 - \bar{Y}^2 = \frac{2321}{50} - 6{,}58^2 = \boxed{4{,}3236}$$

$$\Rightarrow \sigma_Y = \sqrt{V(Y)} = \boxed{2{,}07}$$

$$r = \frac{3{,}0584}{1{,}71 \times 2{,}07} = \boxed{0{,}86}$$

$0{,}7 < r < 0{,}95 \Rightarrow$ bonne corrélation linéaire entre les deux scores.

1. Droite des moindres carrés :

$D_{Y/X}$: $y = ax + b$ avec :

$$a = \frac{Cov(X,Y)}{V(X)} = \frac{3,0584}{2,9236} = \boxed{1,04}$$

$$b = \bar{Y} - a\bar{X} = 6,58 - a \cdot 5,52 = \boxed{0,81}$$

D'où :

$$D_{Y/X} : y = 1,04x + 0,81$$

$D_{X/Y}$: $x = a'y + b'$ avec

$$a' = \frac{Cov(X,Y)}{V(Y)} = \frac{3,0584}{4,3236} = \boxed{0,70}$$

$$b' = \bar{X} - a' \cdot \bar{Y} = 5,52 - a' \cdot 6,58 = \boxed{0,86}$$

$D_{X/Y}$: $x = 0,70y + 0,86 \iff y = 1,42x - 1,22$

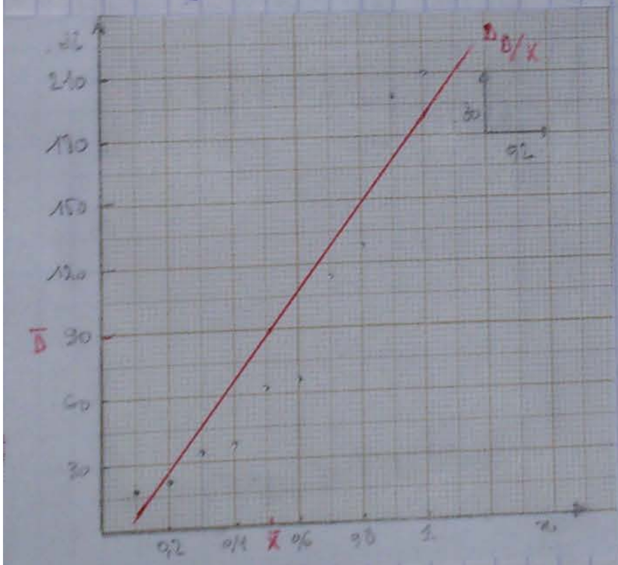
$$D_{X/Y} : y = 1,42x - 1,22$$

4. On peut prévoir le score des nombres restitués d'un étudiant ayant un score de syllabes égal à 14 en utilisant l'équation de la droite de régression de X en Y :

$$x = 0,7(14) + 0,86 = 10,66$$

02 .

a. Nuage de points :



La densité optique en fonction de la quantité d'antibiotique.

Conclusion : le nuage de points semble presque aligné, ce qui justifie une corrélation forte, linéaire et positive entre X et D.

b. Calcul : r(X,D)

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{5,5}{10} = \boxed{0,55}$$

x_i	x_i^2	d_i	d_i^2	$x_i d_i$	y_i	y_i^2	$x_i y_i$
0,1	0,01	19	361	1,9	2,94		
0,2	0,04	21	441	4,2	3,04		
0,3	0,09	35	1225	10,5	3,56		
0,4	0,16	38	1444	15,2	3,64		
0,5	0,25	64	4096	32	4,16		
0,6	0,36	66	4356	39,6	4,19		
0,7	0,49	115	13225	80,5	4,74		
0,8	0,64	130	16900	104	4,87		
0,9	0,81	200	40000	180	5,30		
1	1	210	44100	210	5,35		
5,5	3,85	898	126143	677,9	41,73		

$$V(X) = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \bar{X}^2 = \frac{3,85}{10} - 0,55^2 = \boxed{0,0825}$$

$$\sigma_X = \sqrt{V(X)} = \boxed{0,28}$$

$$\bar{D} = \frac{1}{N}\sum_{i=1}^{N} d_i = \frac{898}{10} = \boxed{89,8}$$

$$V(D) = \frac{1}{N}\sum_{i=1}^{N} d_i^2 - \bar{D}^2 = \frac{126143}{10} - 89,8^2 = \boxed{4550,76}$$

$$\sigma_D = \sqrt{V(D)} = \boxed{67,45}$$

$$Cov(X,D) = \frac{1}{N}\sum_{i=1}^{N} x_i d_i - \bar{X} \cdot \bar{D}$$

$$= \frac{677,9}{10} - 0,55 \times 89,8 = \boxed{18,4}$$

$$r(X,D) = \frac{Cov(X,D)}{\sigma_X \cdot \sigma_D} = \frac{18,4}{0,28 \times 67,45} = \boxed{0,97}$$

Interprétation : $r_{X,D} > 0,95 \Rightarrow$ il existe une forte corrélation linéaire entre X et D.

c. Droite de régression de D en X.

$D_{D/X}$: $d = ax + b$ avec :

$$a = \frac{Cov(X,D)}{V(X)} = \frac{18,4}{0,0825} = \boxed{223,03}$$

$$b = \bar{D} - a\bar{X} = 89,8 - a \times 0,55 = \boxed{-32,86}$$

$$D_{D/X} : d = 223,03x - 32,86$$

2

2. On pose $Y = \ln(D)$

a. Valeurs de Y (Tableau)

b. Pour répondre à cette question, on calcule $r_{X,Y}$ et le compare avec $r_{X,D}$.

$\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} y_i = \frac{41,79}{10} = \boxed{4,179}$

$V(Y) = \frac{1}{N} \sum_{i=1}^{N} y_i^2 - \bar{Y}^2 = \frac{181,5871}{10} - 4,179^2$

$\boxed{V(Y) = 0,69} \Rightarrow \boxed{\sigma_Y = 0,83}$

$Cov(X.Y) = \frac{1}{N} \sum_{i=1}^{N} x_i y_i - \bar{X} . \bar{Y}$

$= \frac{25,354}{10} - 0,55 \times 4,179$

$= \boxed{0,23}$

$r_{X,Y} = \frac{Cov(X.Y)}{\sigma_X \sigma_Y} = \frac{0,23}{0,28 \times 0,83} \simeq \underline{0,99}$

Or : $r_{X,Y} > r_{X,D}$

$\Rightarrow$ l'ajustement de Y en X est plus légitime.

c. Modèle ajustant le mieux les données initiales (x_i, d_i) :

La droite de régression de Y en X :

$D_{Y/X} : y = \alpha x + \beta$ /

$\alpha = \frac{Cov(X.Y)}{V(X)} = \frac{0,23}{0,0825} = \boxed{2,78}$

$\beta = \bar{Y} - \alpha \bar{X} = 4,179 - \alpha \times 0,55 = \boxed{2,64}$

$D_{Y/X} : y = 2,78 x + 2,64$.

$\Leftrightarrow \ln d = 2,78 x + 2,64$

$\Leftrightarrow d = e^{2,64} \times e^{2,78 x}$

$\Leftrightarrow \boxed{d = 14,01 \ (16,11)^x}$

c'est un modèle exponentiel qui ajuste le mieux les données initiales

3. Estimation de la densité optique pour le 11ème tube : $x = 1,2$

$d_E(1,2) = 14,01 \ (16,11)^{1,2} = \boxed{393,50}$.

4. L'erreur commise sur la densité optique :

On note d_R : la densité optique réelle

$d_R(1,2) = d_R(1) + 0,8 . d_R(1)$

$= 1,8 . d_R(1) = 1,8 \times 210 = 378$.

donc l'erreur commise est :

$e = |d_R(1,2) - d_E(1,2)| = |378 - 393,50|$

$\Rightarrow \boxed{e = 15,5}$

Exo3 Ajustement non linéaire :

1/ $y = \frac{1}{ax+b} \Leftrightarrow \frac{1}{y} = ax + b$.

donc X et $\frac{1}{Y}$ sont liées linéairement.

D'où : $a = \frac{Cov(X, 1/Y)}{V(X)} = \frac{37,63}{11,66} = \boxed{3,22}$

$b = \overline{\frac{1}{Y}} - a \bar{X} = 15,6 - a \times 5 = \underline{-0,5}$

x_i	y_i	$1/y_i$	x_i^2	x_i/y_i
0	11	9,90	0	0
2	0,19	5,26	4	10,52
4	0,08	12,50	16	50
6	0,06	16,66	36	99,96
8	0,04	25	64	200
10	0,03	33,33	100	333,3
30		93,65	220	693,78

$\bar{X} = \frac{30}{6} = 5$, $V_X = \frac{220}{6} - 5^2 = 11,66$

$\overline{1/Y} = \frac{93,65}{6} = 15,6$ $Cov(X, 1/Y) = \frac{693,78}{6} - 5 \times 1$
$= 37,6$

Donc : le modèle d'ajustement est :

$$y = \frac{1}{3,22 x - 0,53}$$

2/ $y = a \ln x + b$

$a = \frac{Cov(\ln X, Y)}{V(\ln X)}$,

$b = \bar{Y} - a \overline{\ln X}$

x_i	y_i	$\ln x_i$	$(\ln x_i)^2$	$y_i \cdot \ln$
1	1,1	0	0	0
3	4,4	1,09	1,1881	4,79
5	5,8	1,6	2,56	9,28
7	6,8	1,94	3,7636	13,19
9	7,7	2,19	4,7961	16,86
	25,8	6,82	12,3078	44,1

$\overline{\ln X} = \frac{6,82}{5} = 1,364$ $V(\ln X) = \frac{12,3078}{5} - 1,36$

$= 0,60$

3

$\bar{Y} = \frac{25,8}{5} = 5,16$

$cov(\ln X, Y) = \frac{44,131}{5} - 1,364 \times 5,16$

$= 1,78$

donc : $a = \frac{1,78}{0,60} = \boxed{2,96}$

$b = 5,16 - a \times 1,364 = \boxed{1,11}$

$\Rightarrow \boxed{y = 2,96 \ln x + 1,11}$

3/ $y = \frac{a}{x} + bx \Leftrightarrow y/x = \frac{a}{x^2} + b$

D'où, $a = \frac{Cov(1/x^2 \,;\, Y/X)}{V(1/x^2)}$

$b = \overline{Y/X} - a\,\overline{1/x^2}$

x_i	y_i	$1/x_i^2$	$1/x_i^4$	y_i/x_i	y_i/x_i^3
1	1,6	1	1	1,6	1,6
2	1,4	0,25	0,0625	0,7	0,175
3	1,7	0,11	0,0121	0,56	0,0616
4	2,3	0,06	0,0036	0,57	0,0342
5	2,7	0,04	0,0016	0,54	0,0216
		1,46	1,0798	3,97	1,8924

$\overline{1/x^2} = \frac{1,46}{5} = 0,292$

$V(1/x^2) = \frac{1,0798}{5} - 0,292^2 = 0,13$

$\overline{Y/X} = \frac{3,97}{5} = 0,794$

$Cov(1/x^2, Y/X) = \frac{1,8924}{5} - 0,292 \times 0,794$

$= 0,14$

Alors : $a = \frac{0,14}{0,13} = \boxed{1,07}$

$b = 0,794 - a \times 0,292 = \boxed{0,47}$

$\Rightarrow \boxed{y = \frac{1,07}{x} + 0,47x}$

2ème méthode :

En appliquant la méthode des moindres carrés.

par exemple le 3ème modèle :

Pour déterminer les paramètres du modèle : $y = \frac{a}{x} + bx$ on minimise la fonction : $S(a,b) = \sum_{i=1}^{N} \left(\frac{a}{x_i} + bx_i - y_i\right)^2$

$$\Rightarrow \begin{cases} \frac{\partial S}{\partial a} = 0 \\ \frac{\partial S}{\partial b} = 0 \end{cases} \Rightarrow \begin{cases} 2 \sum_{i=1}^{N} \frac{1}{x_i}\left(\frac{a}{x_i} + bx_i - y_i\right) = 0 \\ 2 \sum x_i \left(\frac{a}{x_i} + bx_i - y_i\right) = 0 \end{cases}$$

$$\Rightarrow \begin{cases} a \sum_{i=1}^{N} \frac{1}{x_i^2} + Nb = \sum_{i=1}^{N} y_i/x_i \\ Na + b \sum_{i=1}^{N} x_i^2 = \sum_{i=1}^{N} x_i y_i \end{cases}$$

Il reste à calculer les sommation du système.

$1/x_i^2$	x_i^2	$x_i y_i$	y_i/x_i
	1	1,6	
	4	2,8	
	9	5,1	
	16	9,2	
	25	13,5	
1,46	55	32,2	3,97

Donc on obtient le système linéaire :

$$\begin{cases} 1,46a + 5b = 3,97 \\ 5a + 55b = 32,2 \end{cases}$$

$\Delta = \begin{vmatrix} 1,46 & 5 \\ 5 & 55 \end{vmatrix} = 55,3 \neq 0$

$a = \frac{1}{55,3} \begin{vmatrix} 3,97 & 5 \\ 32,2 & 55 \end{vmatrix} = \boxed{1,03}$

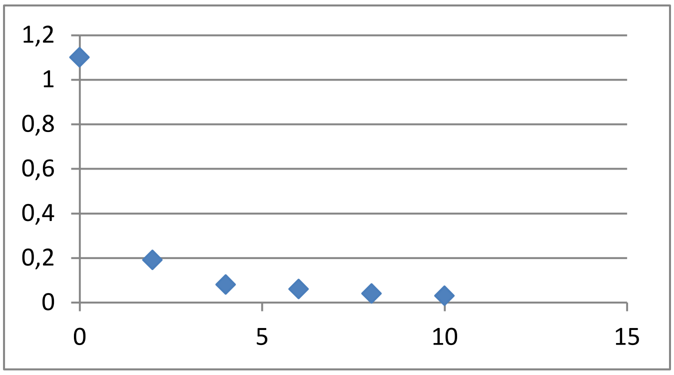
$b = \frac{1}{55,3} \begin{vmatrix} 1,46 & 3,97 \\ 5 & 32,2 \end{vmatrix} = \boxed{0,49}$

4

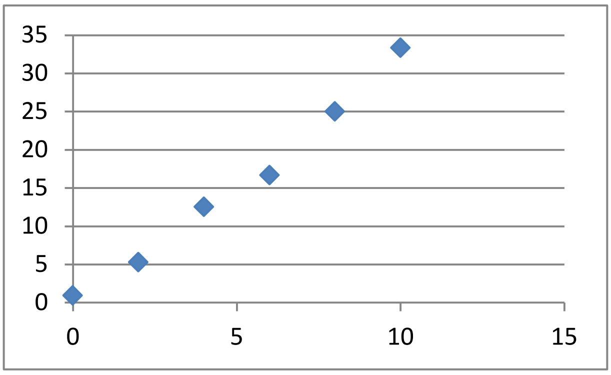
Voici les nuages de points correspondants

1. $y = \frac{1}{ax+b}$

y en fonction de x

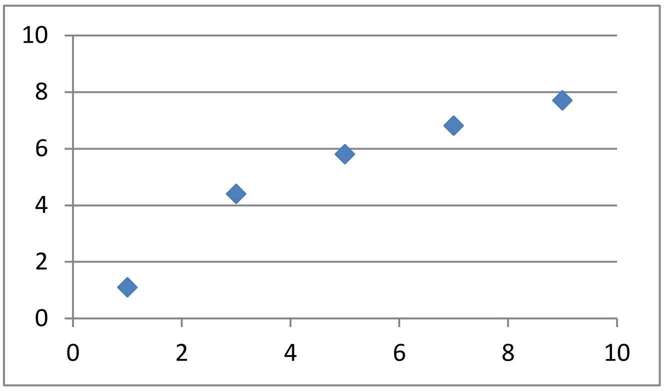


1/y en fonction de x

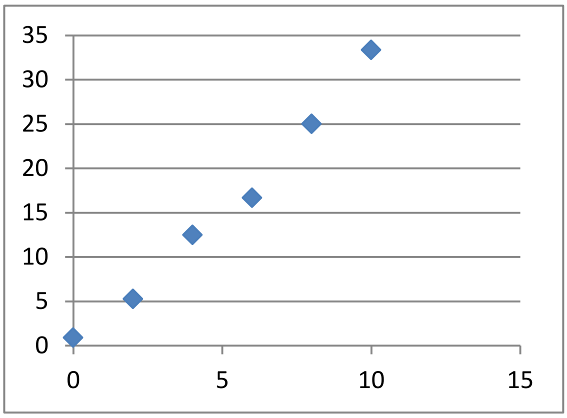


2. $y = a\ln x + b$

y en fonction de x

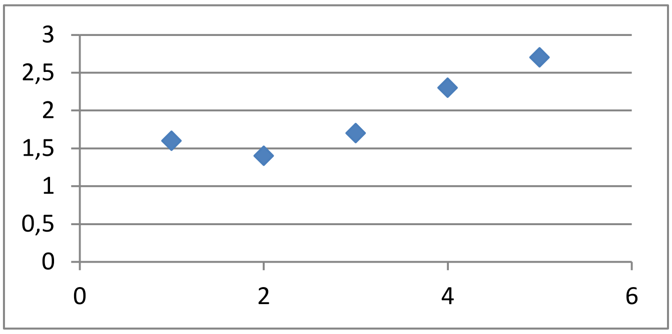


y en fonction de lnx



3. $y = \frac{a}{x} + bx$

y en fonction de x



y/x en fonction de 1/x²