PARTIE 3
CORRIGES DES EXERCICES

Avertissement !!!!!!!!Les calculs proposés ont été réalisé à l'aide d'Excel avec une précision nettement supérieure à ce que vous pouvez obtenir sur une calculatrice, il se peut que vos résultats diffèrent un peu sans pour autant être faux, donc vérifier votre méthode de calcul en cas de divergence de résultats.

EXERCICE 1

Un sondage a été réalisé sur l'adaptation à l'outil informatique en fonction de l'âge (voir tableau empirique). Peut-on conclure que le niveau d'adaptation est lié à l'âge des répondants?

Cette étude porte sur 1184 individus sur lesquels on étudie 2 caractères : le niveau d'adaptation (variable qualitative) et la tranche d'âge (caractère quantitatif mais qui sera considéré comme du qualitatif). Pour «étudier l'existence d'un lien entre ces 2 variables on réalise un **test d'indépendance**, test du X^2. Le modèle dont nous disposons est le modèle de l'indépendance. On va vérifier si la distribution des effectifs observés peut être « assimilée » à ce que serait cette distribution si les variables étaient indépendantes.

H0: le niveau d'adaptation est indépendant de la tranche d'âge de l'apprenant
H1: le niveau d'adaptation est lié à la de la tranche d'âge de l'apprenant

Tableau empirique (nij: effectifs observés)

effectifs nij (O)	niveau d'adaptation			
âge	très difficile	difficile	facile	total
[18 , 30[	98	140	150	388
[30 , 50[	126	150	90	366
[50 , 70[	230	120	80	430
total	454	410	320	1184

Tableau théorique (n'ij: effectifs théoriques) construit à partir du modèle probabiliste de l'indépendance.

effectifs n'ij (C)	niveau d'adaptation			
âge	très difficile	difficile	facile	total
[18 , 30[	148,78	134,36	104,86	388
[30 , 50[	140,34	126,74	98,92	366
[50 , 70[	164,88	148,90	116,22	430
total	454	410	320	1184

Exemples de calcul
(388*454)/1184 = 148.78
(366*454)/1184 = 140.34
(430*410)/1184 = 148.90……..etc

pour chaque case on calcule la quantité $(O-C)^2 / C$, on fait la somme de ces quantité pour obtenir le X^2 calculé.

(O - C)² / C	niveau d'adaptation		
âge	très difficile	difficile	facile
[18 , 30[	17,33	0,24	19,43
[30 , 50[	1,47	4,27	0,80
[50 , 70[	25,72	5,61	11,29

Exemples de calculs :
$(98 – 148.78)^2 / 148.78 = 17.33$
$(150 – 126.74)^2 / 126.74 = 4.27$…….etc

X^2 calculé = 17.33 + 0.24 + 19.43 + 1.47 + 4.27 + 0.8 + 25.72 + 5.61 + 11.29 = **86.15**
ddl = (3-1)*(3-1) = **4**

A l'aide de la table on positionne le X^2 calculé :
$X^2_{0,95}(4) = 9.49$
$X^2_{0,999}(4) = 18.5$
X^2 calculé $> X^2_{0,999}(4)$
α calculé $< 0,001$ (calculé à l'aide d'Excel = α calculé $= 9.\ 10^{-18}$

Si on décide de rejeter H0 on le fait avec un risque inférieur à 0.1% ($9.\ 10^{-16}$%)de chance de se tromper. Ce risque est inférieur aux 5% classiquement toléré, donc on peut prendre ce risque et on accepte donc H1.
Conclusion: on a pu mettre en évidence l'existence d'un lien hautement significatif entre le niveau d'adaptation de l'apprenant et sa tranche d'âge, avec moins de 0.1% de chance de se tromper en affirmant cela.

EXERCICE 2

En 1959 une étude a été faite sur un échantillon de 350 personnes qui comportait 10 % de fumeurs et 20% de gens atteints du " cancer du fumeur " . Dans cet échantillon il y avait 23 individus fumeurs et atteints d'un cancer. Que concluez-vous?

Cette étude porte sur 350 personnes, 10% (350*0.1 = 35) sont des fumeurs, 20% (0.2*350 = 70) sont atteintes d'un cancer et 23 individus sont à la fois fumeurs et sont atteints du cancer.

Tableau empirique (nij: effectifs observés)

effectifs nij (O)	Fumeur	Non fumeur	total
cancer	23	47	70
non cancer	12	268	280
total	35	315	350

Dans cette enquête on étudie 2 caractères : le fait d'être fumeur ou non (variable qualitative) et le fait d'être atteint du cancer ou non (variable qualitative). Pour «étudier l'existence d'un lien entre ces 2 variables on réalise un **test d'indépendance**, test du X^2. Le modèle dont nous disposons est le modèle de l'indépendance. On va vérifier si la distribution des effectifs observés peut être « assimilée » à ce que serait cette distribution si les variables étaient indépendantes.

H0: le fait d'être atteint du cancer ou non est indépendant du fait d'être fumeur ou non
H1: le fait d'être atteint du cancer ou non est lié au fait d'être fumeur ou non

Tableau théorique (n'ij: effectifs théoriques) construit à partir du modèle probabiliste de l'indépendance.

Effectifs n'ij (C)	Fumeur	Non fumeur	total
cancer	7,00	63,00	70
non cancer	28,00	252,00	280
total	35	315	350

Exemples de calcul :
35*70/350 = 7
280 * 315 / 350 = 252……..etc

pour chaque case on calcule la quantité $(O-C)^2/C$, on fait la somme de ces quantité pour obtenir le X^2 calculé.

$(O - C)^2 / C$	Fumeur	Non fumeur
cancer	36,57	4,06
non cancer	9,14	1,02

Exemples de calcul :
$(23 – 7)^2 / 7 = 36.57$
$(268 – 252)^2 / 252 = 1.02$

X^2 calculé = 36.57 + 4.06 + 9.14 + 1.02 = **50.79**
ddl = (2 – 1)*(2 – 1) = **1**

A l'aide de la table on positionne le X^2 calculé :
$X^2_{0,95}(1) = 3.84$
$X^2_{0,999}(1) = 10.8$
X^2 calculé $> X^2_{0,999}(1)$
α calculé $< 0,001$ (calculé à l'aide d'Excel = α calculé $= 1.\ 10^{-12}$

Si on décide de rejeter H0 on le fait avec un risque inférieur à 0.1% (10^{-10}%) de chance de se tromper. Ce risque est inférieur aux 5% classiquement toléré, donc on peut prendre ce risque et on accepte donc H1.
Conclusion : on a pu mettre en évidence l'existence d'un lien très significatif entre le fait d'être atteint du cancer ou non et le fait d'être fumeur ou non, avec moins de 0.1% de chance de se tromper en affirmant cela.

EXERCICE 3

La division recherche d'une entreprise chimique a expérimenté 4 produits anti-mildiou sur 300 pieds de vigne situés en des lieux aléatoires. Quelques semaines plus tard, la proportion des pieds contaminés était réparties selon le tableau ci-dessous. Que concluez-vous?

	traitement 1	traitement 2	traitement 3	traitement 4
contaminés	6	6	8	7
non contaminés	12	31	16	14

!!! Attention!!! ce tableau est un tableau de fréquences et non d'effectifs!!!!!

Cette étude porte sur 300 unités statistiques sur lesquelles on étudie 2 caractères qualitatifs: le type de traitement (4 modalités) et le type de contamination (2 modalités). Pour «étudier l'existence d'un lien entre ces 2 variables on réalise un **test d'indépendance**, test du X^2. Le modèle dont nous disposons est le modèle de l'indépendance. On va vérifier si la distribution des effectifs observés peut être « assimilée » à ce que serait cette distribution si les variables étaient indépendantes.

Tableau empirique (nij: effectifs observés)

nij (O)	traitement 1	traitement 2	traitement 3	traitement 4	total
contaminés	18	18	24	21	81
non contaminés	36	93	48	42	219
total	54	111	72	63	300

H0: il n'existe pas de lien entre le type de traitement et le type de contamination.
H1: il existe un lien entre le type de traitement et le type de contamination.

Tableau théorique (n'ij: effectifs théoriques) construit à partir du modèle probabiliste de l'indépendance.

n'ij (C)	traitement 1	traitement 2	traitement 3	traitement 4	total
contaminés	14,58	29,97	19,44	17,01	81
non contaminés	39,42	81,03	52,56	45,99	219
total	54	111	72	63	300

Exemples de calculs :
$(81 * 54)/ 300 = 14.58$
$(219 * 111) / 300 = 81.03$

pour chaque case on calcule la quantité $(O-C)^2/C$, on fait la somme de ces quantité pour obtenir le X^2 calculé.

$(O - C)^2 / C$	traitement 1	traitement 2	traitement 3	traitement 4
contaminés	0,802	4,781	1,070	0,936
non contaminés	0,297	1,768	0,396	0,346

Exemples de calcul :
$(18 – 14.58)^2 / 14.58 = 0.802$
$(42 – 45.99)^2 / 45.99 = 0.346$

X^2 calculé $= 0.802 + 4.781 + ……. + 0.346 =$ **10.40**
ddl $= (4 – 1) *(2 – 1) =$ **3**

A l'aide de la table on positionne le X^2 calculé :

$X^2_{0,975}(3) = 9.35$
$X^2_{0,99}(3) = 11.3$
$X^2_{0,975}(3) < X^2$ calculé $< X^2_{0,99}(3)$
$0,01 < \alpha$ calculé $< 0,025$

α calculé à l'aide d'Excel = 0.0155

Si on décide de rejeter H0 on le fait avec un risque compris entre 1% et 2.5% (1.55%)de chance de se tromper. Ce risque est inférieur aux 5% classiquement toléré, donc on peut le prendre et on accepte donc H1.
Conclusion: on a pu mettre en évidence l'existence d'un lien significatif entre le type de traitement et le type de contamination.

EXERCICE 4

Le nombre de pannes d'un équipement est supposé suivre une loi de Poisson de moyenne égale à 0.4panne par jour. Le service de contrôle qualité à vérifier l'équipement pendant 100 jours et a obtenu :

Nombre de pannes	Nombre de jours	1°) peut-on considérer comme vraisemblable l'hypothèse selon laquelle le nombre de pannes suit une loi de Poisson de paramètre 0.4? 2°) peut-on considérer comme vraisemblable l'hypothèse selon laquelle le nombre de pannes suit une loi de Poisson ?
0	45	
1	36	
2	14	
3	4	
4	1	

Dans ce cas, il s'agit de comparer la distribution observée d'une variable à une distribution théorique; il s'agit d'un **test du X^2 d'adéquation (appelé aussi test de conformité).** Ici on dispose du modèle théorique de la loi (paramètre connu dans la 1ère question, paramètre de la loi à estimer dans la 2ème question)

Dans ce type de test le ddl = nombre de terme qui rentrent dans le calcul du X^2 - 1 – nombre de paramètres estimés.

!!Attention !! un test du X^2 ne peut s 'appliquer uniquement si les effectifs théoriques sont supérieurs ou égaux à 5, si ce n'est pas le cas, il faut procéder à un regroupement en sachant que l'on tolère un effectif théorique inférieur à 5, dans chacune des classes des extrémités de la distribution.

1°)
H0: L(X) est conforme à une loi P(0.4)
H1: L(X) n'est pas conforme à une loi P(0.4)

Pj se lit dans la table P(0.4)
nj = pj * 100

L(X) = P(0,4)

xj	nj observés (O)	pj = P(X=xj)	nj calculés (C)	(O - C)² / C
0	45	0,6703	67,03	7,24
1	36	0,2681	26,81	3,15
2	14	0,0536	5,36	13,91
3 et plus de 3	5	0,0079	0,79	22,35
total	100	1,00	100,00	**46,65**

X^2 calculé = **46.65**

ddl = 4 – 1 = **3**
$X^2_{0.999}(3) = 16.3$
X^2 calculé > $X^2_{0.999}(3)$
α calculé à l'aide d'Excel = $4\ 10^{-10}$

Si on décide de rejeter H0 on le fait avec un risque inférieur à 0.1% ($4\ 10^{-8}$ %) de chance de se tromper. Ce risque est inférieur aux 5% classiquement toléré, donc on peut le prendre et on accepte donc H1.
Conclusion: on a pu mettre en évidence que la loi suivie par X n'était pas une loi de Poisson de paramètre 0.4.

2°)
H0: L(X) est conforme à une loi de Poisson
H1: L(X) n'est pas conforme à une loi de Poisson

On va estimer le paramètre de la loi:

xj	nj observés	xj*nj
0	45	0
1	36	36
2	14	28
3	4	12
4	1	4
5	0	0
6	0	0
total	100	80
moyenne		0,8

E(X) = 0.8

Pj se lit dans la table P(0.8)
nj = pj * 100

L(X) = P(0,8)

xj	nj observés (O)	pj = P(X=xj)	nj calculés (C)	(O - C)² / C
0	45	0,44933	44,93	0,0001
1	36	0,35946	35,95	0,0001
2	14	0,14379	14,38	0,0100
3 et plus de 3	5	0,0474	4,74	0,0142
total	100	1,00	100,00	**0,0244**

X^2 calculé = **0.024**
ddl = 4 – 1 – 1 = **2**
$X^2_{0.010}(2) = 0.020$
$X^2_{0.025}(2) = 0.051$
$X^2_{0.010}(2) < X^2$ calculé < $X^2_{0.025}(2)$
$0.975 < \alpha$ calculé < 0.99
α calculé à l'aide d'Excel = 0.9879

Si on décide de rejeter H0 on le fait avec un risque compris entre 97.5% et 99% (98.8 %) de chance de se tromper. Ce risque est supérieur aux 5% classiquement toléré, donc on ne peut pas le prendre et on conserve donc H0.
Conclusion: on n'a pas pu mettre en évidence que la loi suivie par X n'était pas une loi de Poisson.

EXERCICE 5

Une enquête menée sur la lecture de 4 revues scientifiques a été menée auprès des laboratoires de recherche d'une ville universitaire. On voulait vérifier si elles répondaient toujours dans les mêmes proportions aux attentes des chercheurs. Que concluez-vous des résultats suivants :

type de revues	pourcentage théorique des préférence	nombre de chercheurs selon leur préférence
A	10	32
B	50	181
C	35	125
D	5	22

Cette étude porte sur 360 chercheurs, pour chaque individu on ne mesure qu'une seule variable, à savoir: leur type de revue préférée. Dans ce cas, il va s'agir de comparer la distribution observée des attentes actuelles en matière de type de revues préférées par rapport aux attentes antérieures (à une distribution théorique connue); on doit mettre en œuvre un **test du X^2 d'adéquation (appelé aussi test de conformité).**
On voudrait savoir si l'échantillon constitué actuellement est conforme à la répartition connue antérieurement. On dispose donc du modèle théorique (les proportions des préférences pour chaque type de revue) et on est donc capable de dire comment devraient se répartir les 360 chercheurs si ces attentes étaient inchangées.

Ho:les attentes des chercheurs sont encore conformes à celles mesurées antérieurement.
H1: les attentes des chercheurs ne sont plus conformes à celles mesurées antérieurement.

type de revues	pj	nj observés (O)	nj calculés (C)	$(O - C)^2 / C$
A	0,1	32	36	0,444
B	0,5	181	180	0,006
C	0,35	125	126	0,008
D	0,05	22	18	0,889
somme	1	360	360	**1,35**

nj = pj * 360

X^2 calculé = **1.35**
ddl = 4 – 1 = **3**
$X^2_{0.10}(3) = 0.584$
$X^2_{0.50}(3) = 2.37$
$X^2_{0.10}(3) < X^2$ calculé $< X^2_{0.50}(3)$
$0.50 < \alpha$ calculé < 0.90
α calculé à l'aide d'Excel = 0.718

Si on décide de rejeter H0 on le fait avec un risque compris entre 50% et 90% (72 %) de chance de se tromper. Ce risque est supérieur aux 5% classiquement toléré, donc on ne peut pas le prendre et on conserve donc H0.
Conclusion: on n'a pas pu mettre en évidence que les attentes des chercheurs n'étaient plus conformes à celles mesurées antérieurement.

EXERCICE 6

Soient le nombre d'heures de révision à un examen et la note obtenue à cet examen pour 6 élèves :

heures	16	13	19	17	10	6
note	17	12	18	15	9	5

Pour cet exercice vous établirez un tableau de détails des calculs dans lequel la dernière ligne sera constituée de :

Σxi Σxi² $\overline{SCEx(\bar{x})}$

Σyi Σyi² $\overline{SCEy(\bar{y})}$

Σxiyi Σ[$(xi - \bar{x})$ (yi - $\bar{y}$)],

$\sum \hat{y}i$ $\sum \hat{y}i^2$ $\overline{SCE\hat{y}(\bar{y})}$

Σei Σei² SCE ei ($\bar{e}$)

et la liste des résidus standardisés.

1°) Déterminer la variable explicative et la variable à expliquer et calculer la moyenne et l'écart-type de X et de Y
2°) Calculer à l'aide de la calculatrice et interpréter les coefficients de corrélation et de détermination
3°) Déterminer les coefficients de la régression à l'aide de la calculatrice
4°) Déterminer l'équation de la droite de régression de Y pour X fixé
5°) Tracer le nuage de points et la droite de régression, où se trouve le centre de gravité du nuage ?
6°) calculer à l'aide de la calculatrice sans saisir d'autres données que celles de X et Y:
SCE x SCE y SPE xy cov xy SCE résiduelle
la variance résiduelle la variance expliquée par la régression
7°) Si on suppose qu'il existe un lien entre ces 2 variables, quel devrait être la note d'un élève qui révise 13 heures ? calculer le résidu correspondant.

Cette étude qui porte sur 6 individus, pour chaque individu on relève 2 variables quantitatives (note et nombre d'heures). On veut vérifier si on peut expliquer les variations entre les notes obtenues par les variations des heures consacrées aux révisions. Pour cela on va utiliser la **méthode de régression linéaire simple et la corrélation**.

Notez bien !!!!!!! $\bar{x}$ pourra être noté m_x et $\bar{y}$ pourra être noté m_y

On réalise donc l'ensemble des calculs nécessaire à l'interprétation de l'information donnée par les couples (xi , yi)

	xi	yi	xi²	yi²	(xi - m_x)²	(yi - m_y)²	xiyi	(xi - m_x) * (yi - m_y
	16	17	256	289	6,25	18,78	272	10,83
	13	12	169	144	0,25	0,44	156	0,33
	19	18	361	324	30,25	28,44	342	29,33
	17	15	289	225	12,25	5,44	255	8,17
	10	9	100	81	12,25	13,44	90	12,83
	6	5	36	25	56,25	58,78	30	57,50
somme	**81**	**76**	**1211,00**	**1088,00**	**117,50**	**125,33**	**1145,00**	**119,00**
moyenne	**13,5**	**12,67**	**201,83**	**181,33**	**19,58**	**20,89**	**190,83**	**19,83**

	xi	yi	$yi_{estimé}$	$yi^2_{estimé}$	$(yi_{estimé} - m_{yestimé})^2$	ei	ei²	$(ei - m_e)^2$
	16	17	15,20	231	6,41	1,80	3,245	3,245
	13	12	12,16	147,87	0,26	-0,16	0,026	0,026
	19	18	18,24	332,58	31,03	-0,24	0,056	0,056
	17	15	16,21	262,81	12,56	-1,21	1,467	1,467
	10	9	9,12	83,211	12,56	-0,12	0,015	0,015
	6	5	5,07	25,714	57,70	-0,07	0,005	0,005
somme	**81**	**76**	**76,00**	**1083,19**	**120,52**	**0,00**	**4,814**	**4,814**
moyenne	**13,5**	**12,67**	**12,67**	**180,53**	**20,09**	**0,00**	**0,802**	**0,802**

On remarque que la moyenne des Y observés est égale à la moyenne des Y estimés.
On remarque que la SCE des résidus est égale à la SC des résidus.

1°)

Xi: heures (explicative)
Yi: note (à expliquer)
m_x =13,50
m_y = 12,67
s^2_x = 117.50 / 6 = 19,58 ou s^2_x = 1211 / 6 – (81 / 6)² = 19,58
s^2_y = 125.33 / 6 = 20,89 ou s^2_y = 1088 / 6 – (76 / 6)² = 20,89
s_x = 4,43
s_y = 4,57

2°)

SPE = $\Sigma xy - n*m_x*m_y$ = 1145 – 6*13.50*12.67 = 119
ou SPE = Σ (xi - m_x) * (yi - m_y) = 119

Covariance = SPE / n = 119 / 6 = 19,83
ou Cov = $\Sigma xy/n – m_x*m_y$ = 1145 / 6 – 13.50 * 12.67 = 19.83

$r^2 = cov_{x,y}^2 / (s^2_x * s^2_y) = 0,9616$
r = 0,9806

Coefficient de détermination = r^2: 96,16% des variations des notes entre les individus sont expliquées par les variations entre les heures qu'ils ont consacré aux révisions.
Donc 1 - r^2: 3,84% des variations des notes entre les individus ne sont pas expliquées par les variations entre les heures qu'ils ont consacrées aux révisions, ces variations sont liées aux fluctuations naturelles entre les individus.

Coefficient de corrélation: r: il est positif, les variables varient donc dans le même sens. Ce coefficient est du même signe que la SPE et que la covariance.

3°)

$b = SPExy / SCEx = 1.013$
$a = m_y - b*m_x = -1.006$

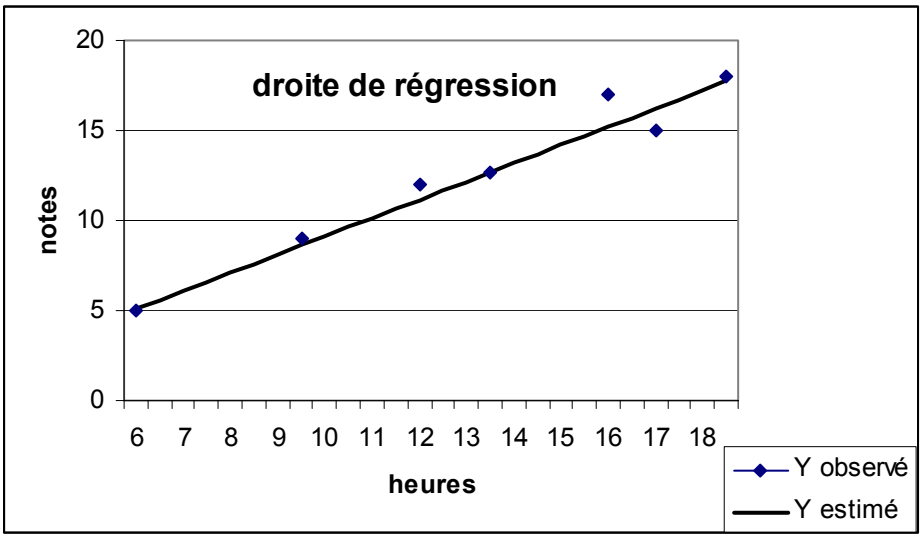
Interprétation des coefficients de la régression:
Ordonnée à l'origine: **a**: si un élève n'a pas révisé (x = 0) alors sa note estimée devrait être égale à –1.006 (ceci a un sens mathématiquement mais pas concrètement dans cette étude)

Pente: **b**: c'est la valeur de la variation de Y quand X varie de 1 unité; quand un élève révise1 heure de plus, sa note est augmentée de 1.013 points.

4°) Equation de la droite de régression de Y pour X fixé: Dy/x
$\hat{Yi}$ = b Xi + a = 1.013 Xi – 1.006
ou Yi = b Xi + a + ei = 1.013 Xi – 1.006 + ei

5°)

Le point (m_x, m_y) = (13.50, 12.67) correspond au centre de gravité du nuage, il se trouve sur la droite de régression.



6°)
On saisit la liste des xi et des yi et on trouve

- en faisant l'écart-type au carré: **s^2 y = 20,89**
- en faisant n fois l'écart-type au carré: SCEx **= 117,50 et** SCE y **= 125,33**
- en appliquant la formule de calcul manuel: somme des produits – n fois le produit des moyennes**:** SPE x,y **= 119** puis SPE / n = cov x,y **=19,83**

On saisit la liste des ei et on trouve

- en faisant l'écart-type au carré: **s^2 e = 0,80**
- en faisant n fois la variance ou la somme des carrés: **SCE e = 4,81**

On saisit la liste des des yi estimés et on trouve en faisant l'écart-type au carré: **s^2 ŷ=20,09**

On remarque que s^2 y = s^2 ŷ + s^2 e
La variance totale des Y est égale à la somme entre la variance des Y expliquée par la régression et la variance des Y non expliquée (résiduelle)

Important !!: **r^2 est aussi égal à SCE $\hat{Y}i$ / SCE yi = 120.52 / 125.33 = 0.9616**

7°)

$\hat{Y}i$ = 1.013 *13 – 1.006 = **12.2**

EXERCICE 7

Les données suivantes représentent la mesure de la concentration en chlore d'échantillons prélevés à des semaines différentes (vous devez compléter la colonne vide* de l'énoncé):

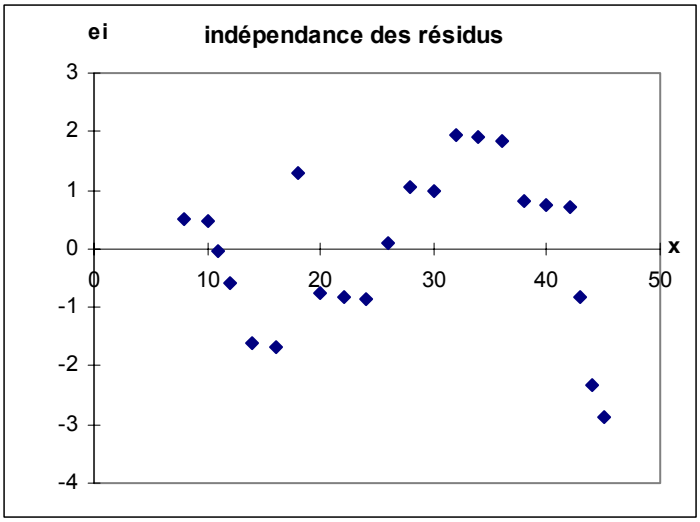
X	Y	Y estimé*	ei	ei standardisé**
8	49	48,48	0,52	0,39
10	48	47,53	0,47	0,35
11	47	47,05	-0,05	-0,04
12	46	46,57	-0,57	-0,43
14	44	45,62	-1,62	-1,22
16	43	44,67	-1,67	-1,26
18	45	43,72	1,28	0,96
20	42	42,77	-0,77	-0,58
22	41	41,81	-0,81	-0,61
24	40	40,86	-0,86	-0,65
26	40	39,91	0,09	0,07
28	40	38,96	1,04	0,78
30	39	38,00	1,00	0,75
32	39	37,05	1,95	1,47
34	38	36,10	1,90	1,43
36	37	35,15	1,85	1,39
38	35	34,20	0,80	0,60
40	34	33,24	0,76	0,57
42	33	32,29	0,71	0,53
43	31	31,82	-0,82	-0,62
44	29	31,34	-2,34	-1,76
45	28	30,86	-2,86	-2,15

Dans cet exercice vous analyserez les résidus. Vous utiliserez la variance résiduelle observée = SCEe / n pour calculer l'écart type résiduel (logiquement on utilise une variance estimée = SCEe / (n-2) mais nous verrons en 2ème année cette utilisation).

n = 22
SCEe = 38.83
s^2e = 1.77 (variance observée = SCEe / 22)
se = 1.33

a) D'après le tableau ci dessus, tous les résidus standardisés (colonne ** ei / s_e) sont compris entre –2 et +2, sauf pour le dernier individu qui peut être considéré comme une donnée « out ». Dans ce cas il faudrait refaire les calculs des coefficients de la régression sans cette donnée.

b) Représentation des résidus ei en fonction des xi:



Il n' existe pas de lien linéaire entre les ei et les xi, mais l'allure de la courbe est cyclique, le lien peut être sinusoïdal, on peut penser que l'hypothèse d'indépendance entre les ei et les xi n'est pas respectée.

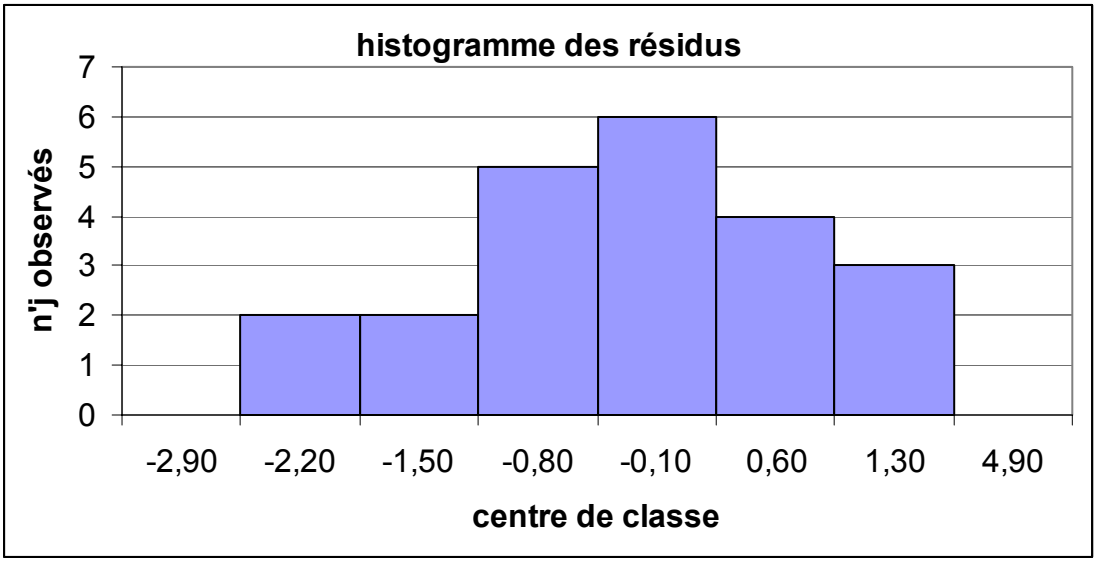
c) Il faut également s'assurer que les résidus suivent une loi Normale autour de la valeur 0, on peut grouper les résidus en classes pour faire un test d'adéquation à une loi Normale (test d'adéquation du X^2).

H0: les résidus suivent une loi N(0; 1.33)
H1: les résidus ne suivent pas une loi N(0; 1.33)

pj = P(lim inf< X < lim sup) = P(t inf< T < t sup)
n'j calculés = pj * 22
ddl = nombre de termes qui rentrent dans le calcul du X^2 - 1 – nombre de paramètres estimés (ddl = 8 – 1 – 1 = 6)

lim inf	lim sup	n'j observés	pj= P(lim inf< X < lim sup)	n'j calculés	(O - C)² / C
infini	-2,2	0	0,0490	1,08	1,079
-2,2	-1,5	2	0,0806	1,77	0,029
-1,5	-0,8	2	0,1441	3,17	0,431
-0,8	-0,1	5	0,1963	4,32	0,108
-0,1	0,6	6	0,2040	4,49	0,509
0,6	1,3	4	0,1618	3,56	0,055
1,3	2	3	0,0979	2,15	0,333
2	infini	0	0,0663	1,46	1,459
	total	**22**	**1,0000**	**22,00**	

X^2 calculé = 1.079 + 0.029 + … + 1.459 = **4.00**
$X^2_{0,10}$ (6) < X^2 calculé < $X^2_{0,50}$ (6)
0,50 < α calculé < 0,90
α calculé = 0.6763 (à l'aide d'Excel)
On a entre 50% et 90% (68%) de risque de se tromper en rejetant H0, donc on conserve H0 et on n'a pas pu mettre en évidence que les résidus ne suivaient pas une loi N(0; 1.33)



Dans cet exemple R^2 = SCE $y_{estimé}$ / SCE y = 710.60 / 749.45 = 0.9481 !! on aurait pu conclure que le modèle linéaire expliquait bien les variations de Y, en réalité l'étude des résidus montre qu'il n'en ait rien!! Les 2 premières conditions sur les résidus ne sont pas validées. Il faut émettre l'hypothèse d'un autre modèle mathématique qu'une droite!

EXERCICE 8 (EXERCICE SUPPLEMENTAIRE DE SYNTHESE SUR LA REGRESSION LINEAIRE)

Une étude porte sur 20 individus. Sur chaque individu on relève 2 variables quantitatives (note QCM et moyenne de l'individu).
Les auteurs cherchent à expliquer des variations dans la réussite scolaire (évaluée par une moyenne générale) par des variations dans les taux d'acquisition de vocabulaire. Pour cela il faut utiliser la **méthode de régression linéaire simple et la corrélation**.

Les calculs sont réalisés selon le même principe que l'exercice 6 de cette partie3. Ils sont représentés dans le tableau ci-dessous.

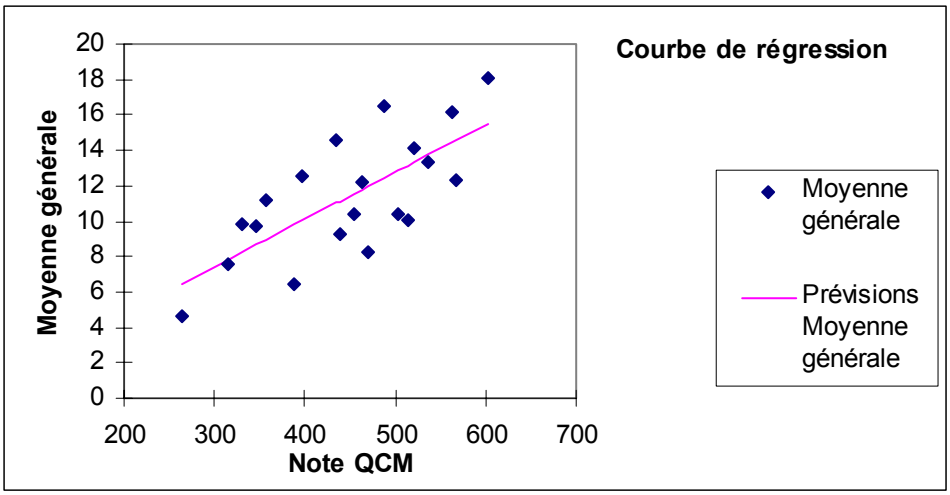
Pour analyser et interpréter ces résultats nous nous intéresserons successivement :

- à l'allure du nuage de points,
- à l'équation de la droite de régression de Dy/x,
- aux coefficients de détermination et de corrélation,
- aux coefficients de la régression,
- à l'analyse des résidus.

	X (explicative)	Y (à expliquer)		yi estimé	résidus
Elève	Note QCM	Moyenne générale	XiYi	$\hat{Y}i$	ei
1	603	18,05	10884,15	15,533	2,517
2	264	4,58	1209,12	6,459	-1,879
3	537	13,33	7158,21	13,766	-0,436
4	347	9,67	3355,49	8,680	0,990
5	463	12,24	5667,12	11,785	0,455
6	562	16,18	9093,16	14,435	1,745
7	520	14,11	7337,2	13,311	0,799
8	314	7,54	2367,56	7,797	-0,257
9	397	12,56	4986,32	10,019	2,541
10	504	10,37	5226,48	12,883	-2,513
11	331	9,85	3260,35	8,252	1,598
12	357	11,21	4001,97	8,948	2,262
13	568	12,34	7009,12	14,596	-2,256
14	454	10,43	4735,22	11,544	-1,114
15	471	8,21	3866,91	11,999	-3,789
16	438	9,23	4042,74	11,116	-1,886
17	389	6,45	2509,05	9,805	-3,355
18	514	10,1	5191,4	13,150	-3,050
19	488	16,54	8071,52	12,455	4,085
20	435	14,58	6342,3	11,036	3,544
somme	8956	227,57	106315,39	227,57	0,0000
n	20	20	20	20	20
somme / n	**447,80**	**11,38**	**5315,77**	**11,38**	**0,00**
SC	4175238	2816,92		2707,43	109,49
SCE	164741,20	227,52		118,03	109,49
Variance	8237,06	11,38		5,90	5,47
écart type	**90,76**	**3,37**		2,43	2,34

SPE	4409,544
covariance	220,48
r^2	**0,5188**
r	**0,7202**
b	**0,0268**
a	**-0,6075**

- allure du nuage de points : le nuage représente une forme d'ellipse ascendante qui laisse penser qu' il existe une relation de type linéaire entre X et Y, X et Y variant dans le même sens. Il ne semble pas y avoir de données « out »



- équation de la droite de régression de Dy/x, les calculs donnent :

$$\hat{Y}i = b\ Xi + a = 0.0268\ Xi - 0.6075$$

- interprétation des coefficients de détermination et de corrélation :

r^2	**0,5188**
r	**0,7202**

Interprétation du coefficient de détermination **r^2: 51.88%** des variations de la moyenne générale sont expliquées par les variations des résultats au QCM. Il reste donc 48.12% ($1 - r^2$) des variations de Y qui ne sont pas expliquées, il s'agit de variations liées aux fluctuations naturelles entre les individus mais aussi on peut penser qu'il existe un autre régresseur qu'il aurait faire rentrer dans le modèle (sous la forme $\hat{y}i = b_0 + b_1Xi_1 + b_2Xi_2$ par exemple), ce régresseur pourrait être lié à l'environnement culturel de l'élève ou à l'intérêt qu'il porte aux études par exemple.

Interprétation du coefficient de corrélation **r: il est positif**, du même signe que b, que la SPE, et que la covariance, X et Y varient dans le même sens, ils sont proportionnels.

- interprétation des coefficients de la régression:

Interprétation de la valeur **a** (ordonnée à l'origine) : Un élève qui aurait une note x = 0 au QCM devrait avoir une moyenne générale Y = -0.6075, cette valeur est non sens dans cette étude, d'autant plus qu'elle résulte d'une extrapolation du domaine des X, or, l' équation Y = f(X) proposée n'est calculée que pour un domaine des X compris entre 264 et 603, à l'extérieur de ce domaine on ne sait pas comment Y varie en fonction de X.

Interprétation du coefficient **b** (pente) : quand le score d'un élève au QCM augmente de 1 point, sa moyenne générale doit augmenter de 0.0268 points (ou si X augmente de 100 points Y augmente de 2.68).

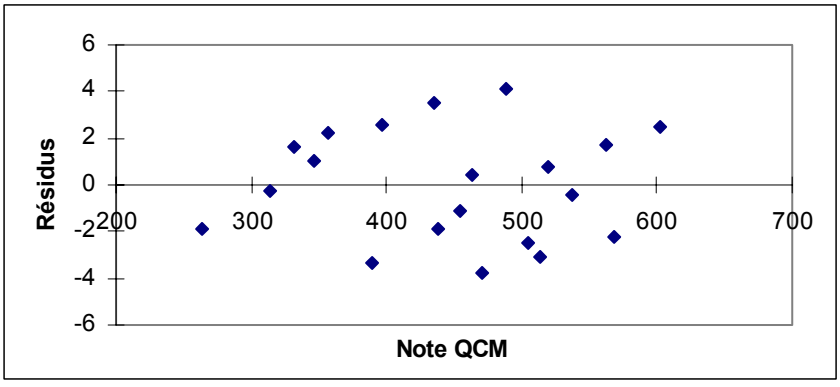
- analyse des résidus (résidus standardisé, indépendance entre ei et xi, normalité des ei)

a) D'après le tableau ci dessous, tous les résidus standardisés (ei / s_e) sont tous compris entre –2 et +2, donc il n'y a pas de données « out »

yi estimé	ei	ei/se
15,53	2,52	1,08
6,46	-1,88	-0,80
13,77	-0,44	-0,19
8,68	0,99	0,42
11,79	0,45	0,19
14,44	1,74	0,75
13,31	0,80	0,34
7,80	-0,26	-0,11
10,02	2,54	1,09
12,88	-2,51	-1,07
8,25	1,60	0,68
8,95	2,26	0,97
14,60	-2,26	-0,96
11,54	-1,11	-0,48
12,00	-3,79	-1,62
11,12	-1,89	-0,81
9,80	-3,35	-1,43
13,15	-3,05	-1,30
12,45	4,09	1,75
11,04	3,54	1,51

SCEe = 109.49
N = 20
s^2e = 5.47
se = 2.34

b) Représentation des résidus ei en fonction des xi:



Il ne semble pas exister de lien entre les ei et les xi, le nuage à une forme de « patate », on peut penser que l'hypothèse d'indépendance entre les ei et les xi est respectée.

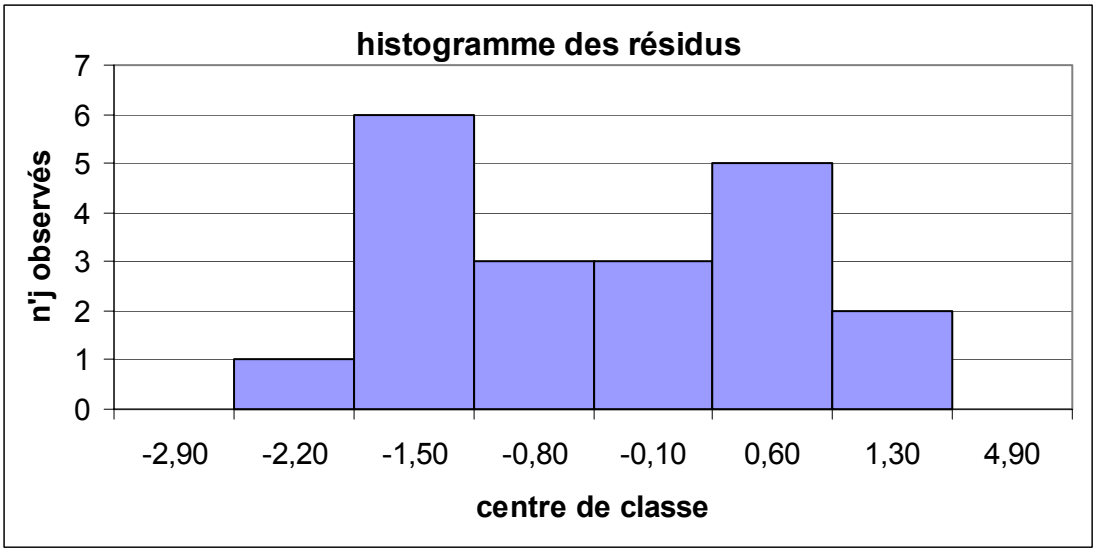
c) Il faut également s'assurer que les résidus suivent une loi Normale autour de la valeur 0, on peut grouper les résidus en classes pour faire un test d'adéquation à une loi Normale (test d'adéquation du X^2).

H0: les résidus suivent une loi N(0; 2.34)
H1: les résidus ne suivent pas une loi N(0; 2.34)

pj= P(x'j inf< X < x'j sup) = P(t inf< T < t sup)
n'j calculés = pj * 20
ddl = nombre de termes qui rentrent dans le calcul du X^2 - 1 – nombre de paramètres estimés
(ddl = 8 – 1 – 1 = 6)

lim inf	lim sup	n'j observés	pj= P(lim inf< X < lim sup)	n'j calculés	(O - C)² / C
infini	-2,2	0	0,0490	0,98	0,981
-2,2	-1,5	1	0,0806	1,61	0,233
-1,5	-0,8	6	0,1441	2,88	3,376
-0,8	-0,1	3	0,1963	3,93	0,218
-0,1	0,6	3	0,2040	4,08	0,286
0,6	1,3	5	0,1618	3,24	0,962
1,3	2	2	0,0979	1,96	0,001
2	infini	0	0,0663	1,33	1,326
	total	**20**	**1,0000**	**20,00**	

X^2 calculé = 0.981 + 0.233 + … + 1.326 = **7.38**
$X^2_{0,50}$ (6) < X^2 calculé < $X^2_{0,90}$ (6)
0,10 < α calculé < 0,50
α calculé = 0.39 (à l'aide d'Excel)
On a entre 10% et 50% (39%) de risque de se tromper en rejetant H0, donc on conserve H0 et on n'a pas pu mettre en évidence que les résidus ne suivaient pas une loi N(0; 2.34)



L'analyse des résidus est satisfaisante, par contre $r^2 = 58\%$ ce qui signifie aussi que 42% des variations de Y ne sont pas expliquées par les variations de X. On peut penser que si l'on disposait d'un autre facteur explicatif (résultats à un autre type de test), on pourrait avoir un modèle linéaire à 2 régresseurs additifs avec un meilleur R^2 (nous verrons cela en $2^{ème}$ année)

EXERCICE 9 (EXERCICE SUPPLEMENTAIRE D'ADEQUATION)

Un cabinet de communication est chargé de mettre en place un dispositif d'information juridique portant sur le harcèlement sexuel des femmes au travail. Avant de lancer une campagne nationale, ce cabinet décide de tester différents formats de communication et souhaite évaluer les éléments retenus par les femmes à qui ils sont présentés. Comme il s'agit d'un dispositif préliminaire nécessitant une disponibilité importante de la part des participantes, seules 250 personnes acceptent d'y prendre part jusqu'au bout. Comme il est important de bien contrôler le niveau d'instruction des participantes à ce pré-test, le cabinet souhaite vérifier si l'échantillon des 250 femmes possède les mêmes caractéristiques de formation scolaire que celles que l'on retrouve dans la population féminine générale. A l'aide d'une classification en 7 catégories de diplômes, on observe les répartitions suivantes (tableau ci-dessous). L'échantillon est-il représentatif de la population L'échantillon est-il représentatif de la population ?

Diplômes	effectifs observés	pourcentage national %
aucun	72	34
BEPC	24	8
CAP, BEP	65	25
Bac, BP	25	11
BTS, DUT	29	10
2ème, 3ème cycle	19	7
non déclarés	16	5
total	250	100

Cette étude porte sur 250 femmes, pour chaque individu on ne mesure qu'une seule variable, à savoir: leur niveau d'instruction. Dans ce cas, il va s'agir de comparer la distribution observée des femmes par rapport à leur niveau d'instruction à une distribution théorique connue; on doit mettre en œuvre un **test du X^2 d'adéquation (appelé aussi test de conformité).**
On voudrait savoir si l'échantillon constitué est conforme à la répartition connue des femmes d'une population selon ce critère d'instruction. On dispose donc du modèle théorique (les proportions de femmes pour chaque niveau) et on est donc capable de dire comment devraient se répartir les 250 femmes si elles étaient représentatives de la population des femmes actives.

H0: la répartition du niveau d'instruction dans l'échantillon des 250 femmes est conforme à ce que l'on retrouve dans la population des femmes actives
H1: la répartition du niveau d'instruction dans l'échantillon des 250 femmes n'est pas conforme à ce que l'on retrouve dans la population des femmes actives

Diplômes	effectifs observés: O *	pourcentage national % **	effectifs calculés: C	$(O - C)^2 / C$
aucun	72	34	85,00	1,99
BEPC	24	8	20,00	0,80
CAP, BEP	65	25	62,50	0,10
Bac, BP	25	11	27,50	0,23
BTS, DUT	29	10	25,00	0,64
2ème, 3ème cycle	19	7	17,50	0,13
non déclarés	16	5	12,50	0,98
total	250	100	250	**4,86**

Exemples de calculs pour C :
250 * 0.34 = 85
250 * 0.08 = 20
250 * 0.25 = 62.50……...etc

Dans ce type de test le ddl = nombre de terme qui rentrent dans le calcul du X^2 - 1 – nombre de paramètres estimés
Ici, le X^2 résulte de la somme de 7 termes, il n'y a pas de paramètres de loi à estimer, donc ddl = 7 – 1 – 0 = **6**
X^2 calculé = **4.86**
ddl = 6

A l'aide de la table on positionne le X^2 calculé:
$X^2_{0,95}(6) = 12.6$
$X^2_{0,10}(6) = 2.2$
$X^2_{0,50}(6) = 5.35$
$X^2_{0,10}(6) < X^2$ calculé $< X^2_{0,50}(6)$
$0,50 < \alpha$ calculé $< 0,90$
α calculé à l'aide d'Excel = 0.5619

Si on décide de rejeter H0 on le fait avec un risque compris entre 50% et 90% (56%)de chance de se tromper. Ce risque est supérieur aux 5% classiquement toléré, donc on ne peut pas prendre ce risque et on conserve donc H0.
Conclusion: on n'a pas pu mettre en évidence que l'échantillon des 250 femmes n'était pas représentatif de la population des femmes actives en ce qui concerne leur niveau d'instruction..