Méthodes statistiques
Réponses – Chapitre 2

Exercice 2.1

a) Voici la distribution conjointe des deux variables (exprimées en pourcentages plutôt qu'en fréquences pour faciliter la lecture):

		État matrimonial				
		Célibataires	Marié(s	Veuf(ve)s	Divorcés	Total
Sexe	Hommes	19,420	24,278	3,824	2,948	50,470
	Femmes	22,431	23,990	0,946	2,162	49,530
	Total	41,852	48,268	4,770	5,110	100

b) Les distributions marginales se trouvent dans les marges du tableau ci-dessus.

c) Distributions conditionnelles de la variable « État matrimonial »

		État matrimonial				
		Célibataires	Marié(s	Veuf(ve)s	Divorcés	Total
Sexe	Hommes	38,479	48,104	7,577	5,840	100
	Femmes	45,289	48,436	1,910	4,365	100

d) Distributions conditionnelles de la variable « Sexe »

		État matrimonial			
		Célibataires	Marié(s	Veuf(ve)s	Divorcés
Sexe	Hommes	46,402	50,298	80,165	57,685
	Femmes	53,598	49,702	19,835	42,315
	Total	100,000	100	100,000	100,000

e) Non

Exercice 2.2

i) 47,15 % ; ii) 49,62 % ; iii) 46,56 % ; iv) 23,46 % ; v) 54,73 % ; vi) 23,39 % ; vii) 5,18 % ; viii) 49,75 %

Exercice 2.3

Ceux qui sont disposés à épouser quelqu'un d'une autre race — et seuls ceux-là — sont en majorité disposés à épouser quelqu'un d'une autre religion.

Exercice 2.4

Ceux qui croient qu'on peut communiquer avec les morts ont moins tendance à mépriser l'astrologie.
Si on groupe les catégories 1 avec 2 (plutôt favorables) et 3 avec 4 (plutôt défavorables), nous obtenons les résultats suivants, qui confirment la conclusion précédente :

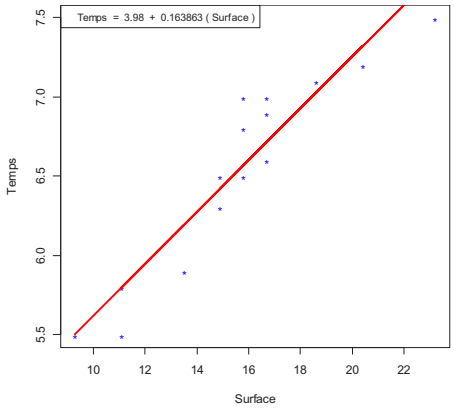
		« L'astrologie est une ânerie »		
		1	2	Tous
« On peut communiquer	1	31,8 %	68,2 %	100 %
avec les morts »	2	70,3 %	29,7 %	100 %
	Tous	55,9 %	44,1 %	100 %

Exercice 2.5

$\overline{x} = 9; \overline{y} = 12$; $\sigma_x = 4$; $\sigma_y = 5{,}621388$; $\sigma_{xy} = 19{,}6$; $r = 0{,}8716709$. Droite de régression : $y = 0{,}975 + 1{,}225x$.

Exercice 2.6

a) Nuage de points :



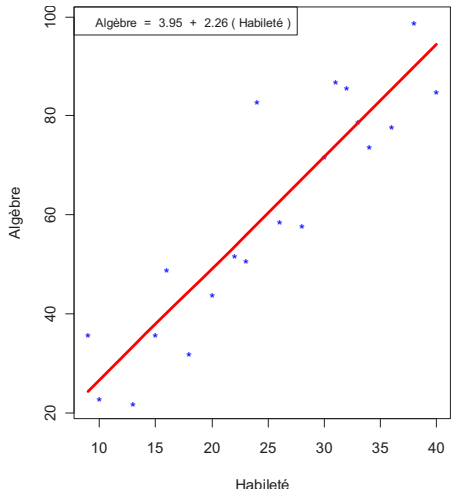
b) « Surface » joue le rôle de variable indépendante (x).

c) Droite de régression : $T = 3{,}9783 + 0{,}1639S$; $r = 0{,}9371572$.

d) $T = 3{,}9783 + 0{,}1639(20) = 7{,}255534$.

Exercice 2.7

a) Voici le graphique :



Imparfaitement linéaire.

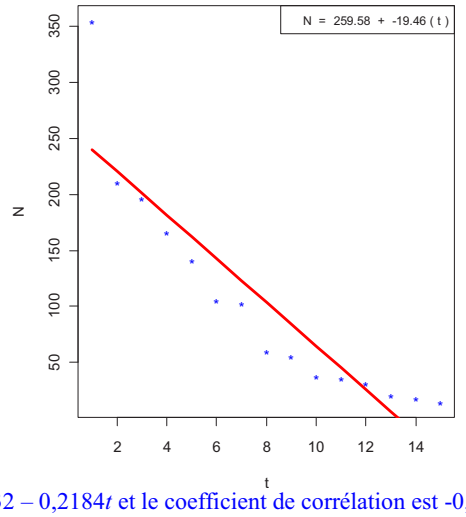
b) H, le test d'aptitude.

c) Droite de régression : $F = 3{,}95 + 2{,}26\,H$; $r = 0{,}908$.

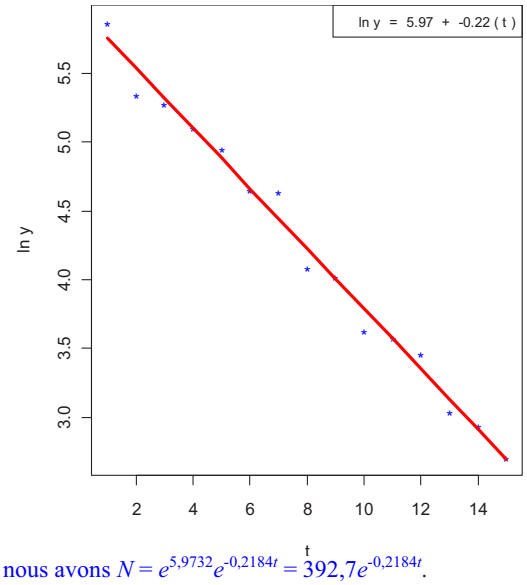
d) 60,47.

Exercice 2.8

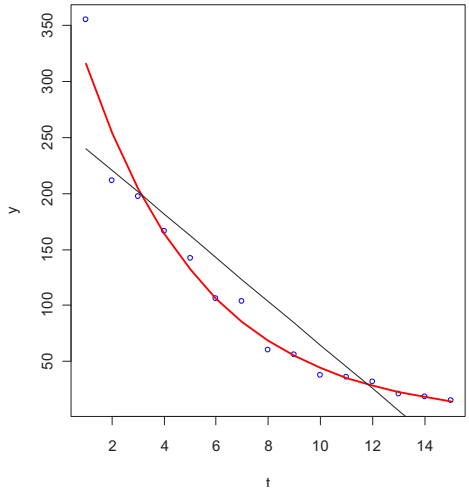
a) La droite des moindres carrés est $N = 259{,}58\text{-}19{,}46t$. Coefficient de corrélation : -0,907. Mais la relation n'est vraiment pas linéaire :



b) Droite des moindres carrés : $y = 5{,}9732 - 0{,}2184t$ et le coefficient de corrélation est -0,99416. La relation semble bien être linéaire :

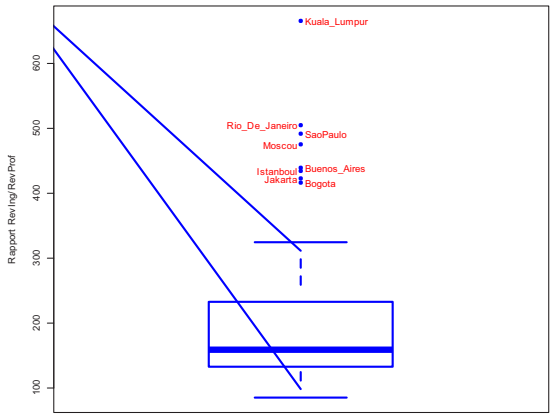


En fonction des données initiales, nous avons $N = e^{5,9732}e^{-0,2184t} = 392,7e^{-0,2184t}$.
Superposée au premier graphique, cette courbe devient



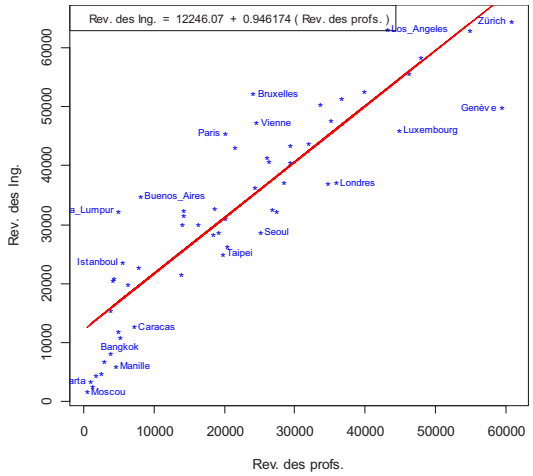
Exercice 2.9

a) Les villes dont les cotes Z sont particulièrement élevées sont marquées dans le graphique ci-dessous :



Les ingénieurs sont beaucoup plus valorisés que les professeurs en Malaisie, au Brésil, en Turquie, en Russie, en Argentine, en Indonésie et en Colombie. Mais le cas de Kuala Lumpur, cependant, avec une cote Z de 6, est suspect.

b)



Exercice 2.10

La catégorie 1 est constituée des pays les plus développés. Voici les moyennes:

	Catégorie 1	Catégorie 2
Ingénieurs	22,63333	20,40741
Professeurs	51,76667	49,44444

Les vacances sont plus longues dans les pays de la catégorie 1, mais l'écart entre ingénieurs et professeurs est à peu près le même dans les deux catégories.

Écarts-types

	Catégorie 1	Catégorie 2
Ingénieurs	5,06	6,15
Professeurs	20,67	19,04

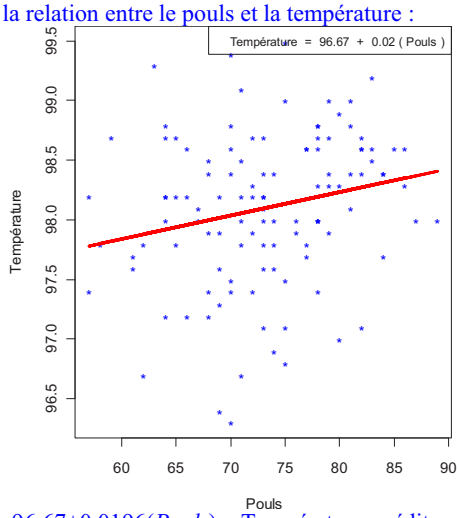
Coefficients de variation

	Catégorie 1	Catégorie 2
Ingénieurs	22,54 %	30,43 %
Professeurs	40,27 %	38,84 %

Exercice 2.11

c) Il y a 7 valeurs dont la valeur absolue est supérieure à 2, soit environ 6 %. Bien qu'elles soient extrêmes, elles n'ont rien de surprenant.

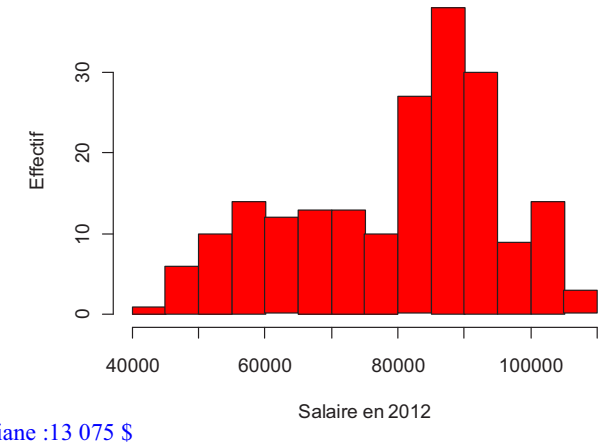
d) Voici un nuage de points représentant la relation entre le pouls et la température :



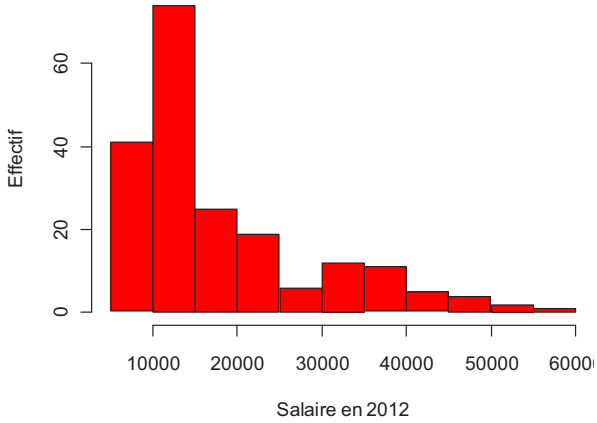
Droite de régression : *Température* = 96,67+0,0196(*Pouls*). Température prédite pour un pouls de 80 : 98,23 . Prédiction peu fiable étant donnée la faible corrélation (r = 0,218).

Exercice 2.12

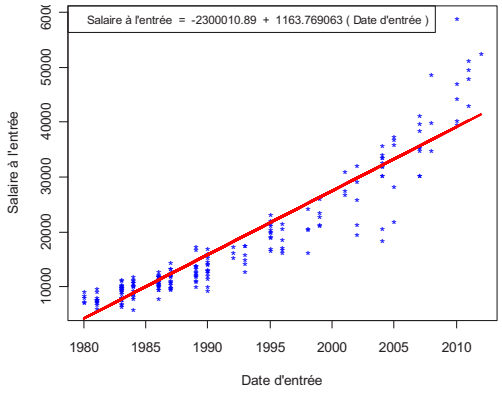
a) i) Moyenne : 79 783 $; Médiane : 84 126 $. Voici l'histogramme :



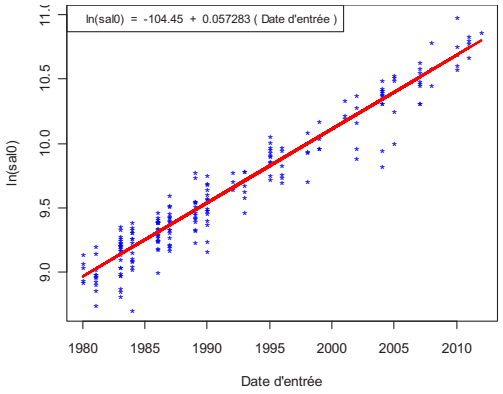
ii) Moyenne : 17 990 $; Médiane :13 075 $



(i)



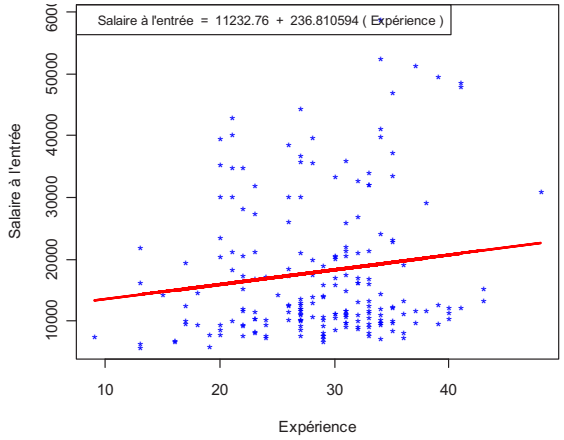
(ii)



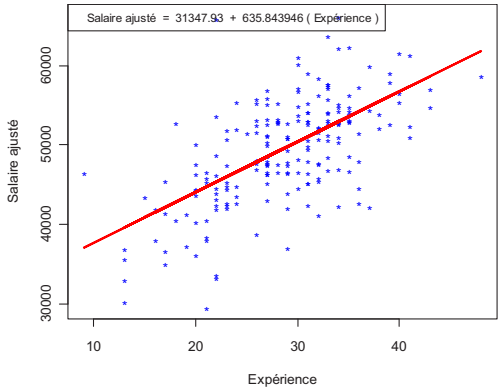
(iii) La droite des moindres carrés est ln *sal*0 = 9,25311 ⇒sal0 = $e^{9,25311}$ = 10 437.
(iv) Accroissement relatif : 0,058955.
(v)

sal0	moyenne : 17 990	écart-type : 11 121	coefficient de variation : 0,620
salajust	moyenne : 49 492	écart-type : 6723	coefficient de variation : 0,136

c) Voici le graphique

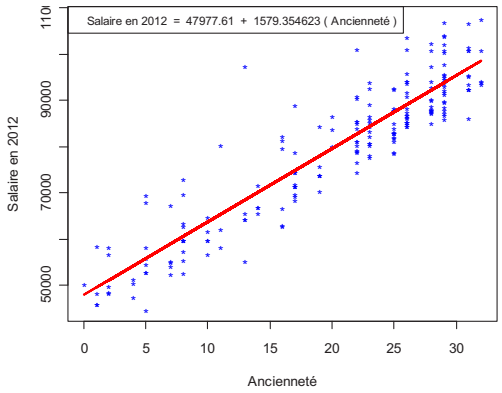


Droite de régression sal0 = 11233 + 236,8(exp). La relation est faible (r = 0,141).



d) salajust = 31348 + 635,8(exp). r = 0,625.

e)



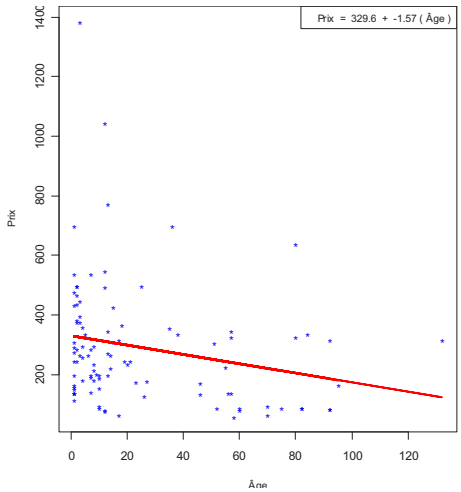
$r = 0{,}918$

f) $r = -0{,}855$: le salaire à l'entrée diminue de 62 ¢ pour chaque dollar de salaire actuel.

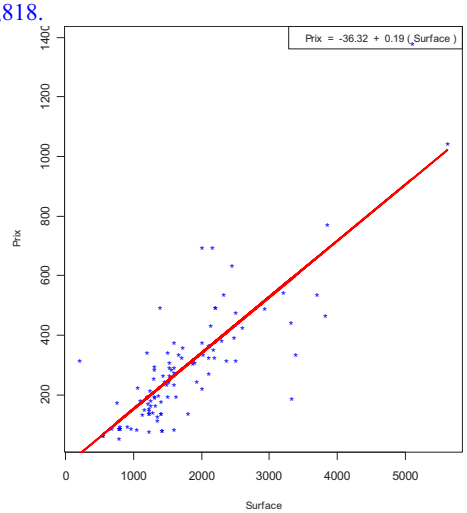
Exercice 2.13

a) Oui, les vieilles maisons coûtent moins cher. Le prix moyen d'une vieille maison est de 271 196 $, alors que pour les maisons jeunes, le prix moyen est 311 667 $.

b) La relation est faible, et elle est non-linéaire. Le coefficient de corrélation est de -0,23. Une maison vieille peut coûter moins cher parce qu'elle est plus vieille, mais une maison vieille est souvent aussi une maison luxueuse qui coûte très cher. La relation entre l'âge et le prix aurait été plus forte si on s'était limité à une certaine classe de maisons, disons, les maisons construites entre 1980 et aujourd'hui, dans une même banlieue.



c) Prix = -36,32 + 0,188×Surface. $r = 0{,}818$.



d) Non, il y a bien plus de 100 000 $ de différence.

e) *Moyennes initiales* :

Moins de deux salles de bains : 193 315 ; Deux salles de bains ou plus : 401 813 ; Différence : 208 498

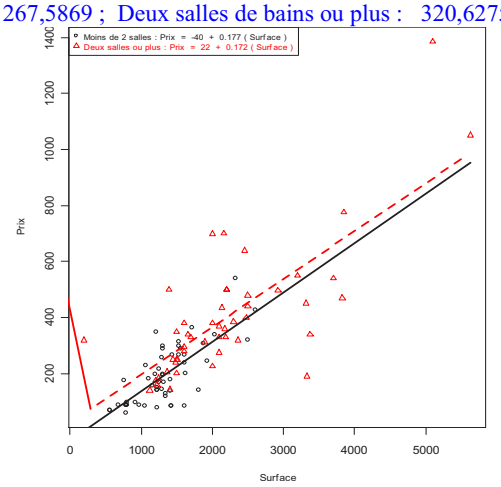
Droites de régression

Moins de deux salles de bains : Prix = -40,1494 + 0,1766 (Surface);

Deux salles de bains ou plus : Prix = 21,6236 + 0,172 (Surface)

Moyennes ajustées:

Moins de deux salles de bains : 267,5869 ; Deux salles de bains ou plus : 320,6275 ; Différence : 53 041

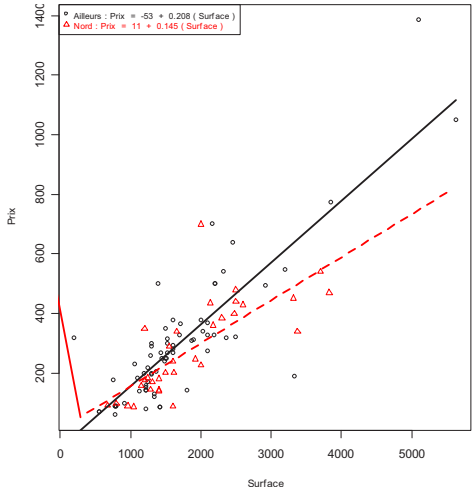


f) *Moyennes ajustees:*

Ailleurs : -53,2586 + 0,2077(1742,4706) = 308,5869;

Nord : 10,7535 + 0,1448(1742,4706) = 263,0686

Différence : 45 518 $ de moins dans le secteur Nord



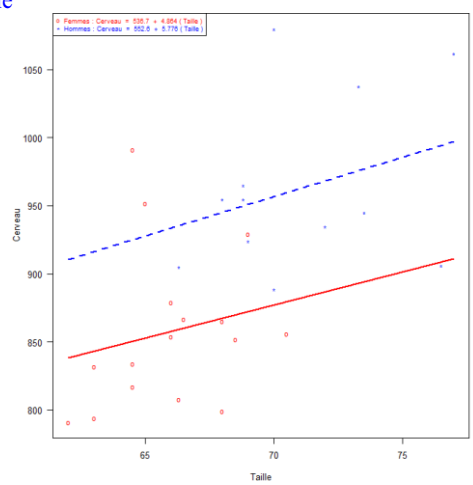
Exercice 2.14

a) Voici les moyennes, médianes et écarts-types de la différence DIFF = (B1+B2)/2-(A1+A2)/2.
Moyennes : Traitement 1 : -1,773 ; Traitement 2 : 0,591

b) Moyennes de B3 : Traitement 1 : 41,05 ; Traitement 2 : 46,73

Exercice 2.15

a) $r = 0{,}424$; pas fort mais pas négligeable



b) $r = 0{,}596$.

c) $r = 0{,}479$

Une meilleure façon d'ajuster les moyennes (des tailles de cerveau) consiste à examiner la relation entre la taille du cerveau et la taille du corps pour les femmes et pour les hommes

Moyennes initiales :

Femmes : 857.53 ; Hommes : 963.25 ;

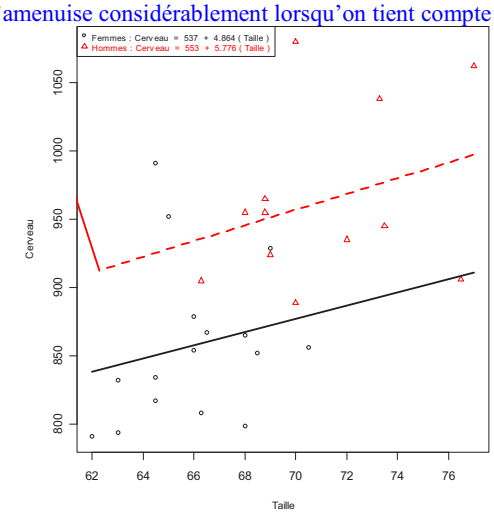
Différence : 105.72

Moyennes ajustées:

Femmes : 575,96 + 4,28 (67,97) = 867; Hommes : 552,60 + 5,78(67,97) = 945,15 ;

Différence : 78,14

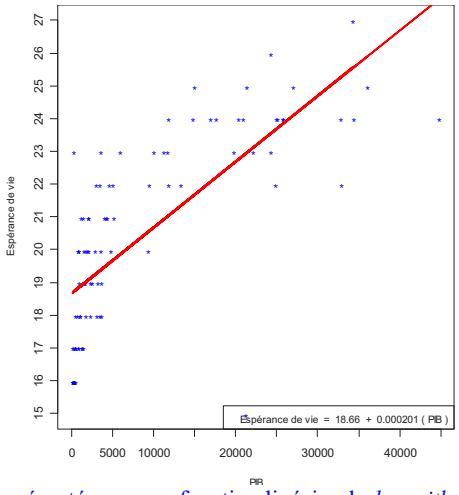
L'importante différence de 105,72 s'amenuise considérablement lorsqu'on tient compte de la taille des femmes.



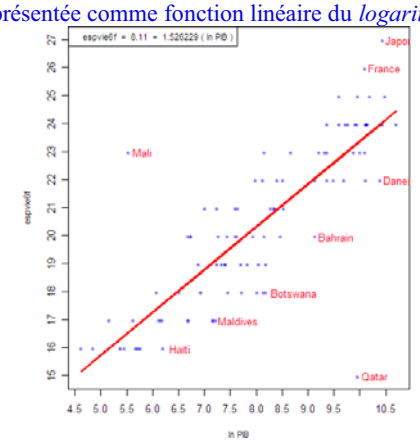
Exercice 2.16

a) L'espérance de vie est particulièrement faible en Afrique, excepté pour Cap Vert, l'Algérie et la Tunisie, dont l'espérance de vie est comparable à celle des pays d'Amérique latine (à l'exception d'Haïti, classée là bien que normalement elle ne fasse pas partie de l'Amérique latine), d'Asie, et d'Océanie. Les pays développés ont une espérance de vie uniformément supérieure à celle de reste du monde, bien que certains pays de l'ancien bloc communiste demeurent au-dessous de la moyenne, comparables aux pays d'autres régions non-africaines.

c) En effet, la relation entre le PIB et l'espérance de vie des femmes à 60 ans est loin d'être linéaire.

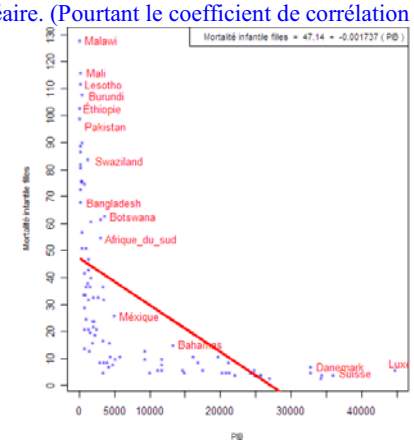


d) L'espérance de vie pourrait bien être représentée comme fonction linéaire du *logarithme* du PIB.

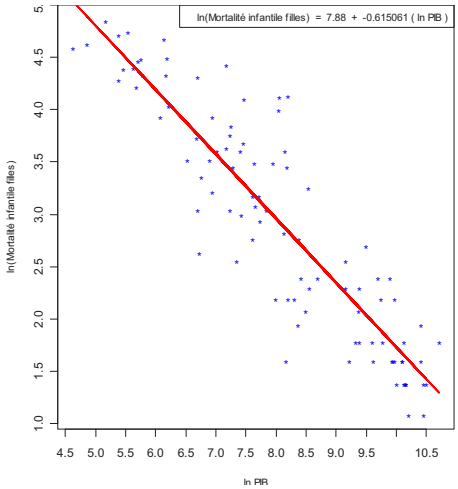


Deux données suspectes : le Mali, dont l'espérance de vie est beaucoup plus élevée qu'on aurait pu le prévoir à partir du PIB; et le Qatar, au contraire, dont l'espérance de vie est beaucoup trop basse. S'agit-il d'une erreur? ou d'un fait sociologique qui mérite qu'on s'y attarde? Une étude plus approfondie serait nécessaire.

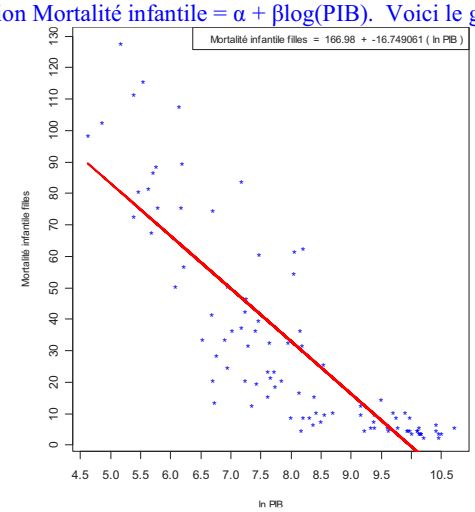
e) En effet la relation à l'air loin d'être linéaire. (Pourtant le coefficient de corrélation est de –0,58, pas négligeable).



f) Le modèle semble se confirmer par le fait que la relation semble être maintenant linéaire :



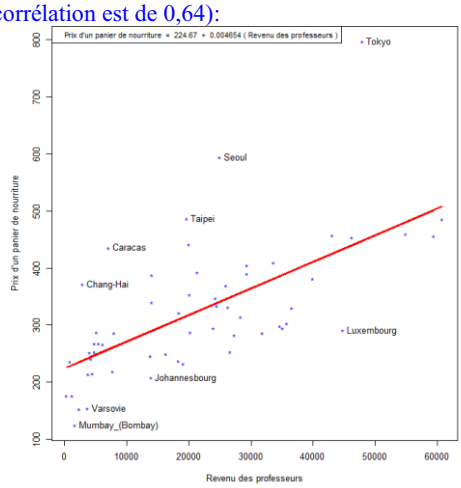
Le coefficient de corrélation est de –0,90. Pour un pays dont PIB = 1000, on estime la mortalité infantile à 37,74462. On aurait pu aussi considérer la relation Mortalité infantile = $\alpha + \beta\log(PIB)$. Voici le graphique :



Le coefficient de corrélation est -0,83. Le modèle Log(Mortalité infantile) = $\alpha+\beta\log(PIB)$ smble légèrement meilleur.

Exercice 2.17

Voici le nuage de points (le coefficient de corrélation est de 0,64):



À Tokyo, les salaires sont assez élevés, mais les prix sont disproportionnés — les profs de Tokyo sont donc moins riches qu'ailleurs. À Genève, à Zürich, à Los Angeles et à Chicago les salaires sont élevés, les prix aussi, mais à peu près conformes aux salaires. C'est au Luxembourg que les profs sont particulièrement avantagés : leurs salaires sont au-dessus de la moyenne mais le coût de la nourriture est inférieur à la moyenne. Les cotes Z confirment ces affirmations :

	Genève	Luxembourg	Tokyo	Zürich

Coût de la nourriture	1,15	-0,28	4,09	1,41
Salaires	2,43	1,5	1,70	2,51

Exercice 2.18

a) Nous recourons à un groupement des classes : les réponses 1 et 2 forment une classe (favorable à la proposition) et les réponses 3 et 4 sont défavorables.
Distributions conditionnelles :

		Croit à l'astrologie?		
		Non	Oui	Total
Croit à la théorie de l'évolution?	Oui	52,73	47,27	100
	Non	55,00	45,00	100
	Total	53,33	46,67	100

Le résultat est surprenant : ceux qui croient à la théorie de l'Évolution sont plus souvent portés à croire à l'astrologie. La différence, cependant, est plutôt faible, et pourrait être fortuite.

b) Dans le cas des deux variables, nous avons réuni les réponses 1, 2, et 3, étant donné les effectifs plutôt faibles dans ces catégories :
Distributions conditionnelles :

		Épouserait une personne d'une autre religion?		
		Non	Oui	Total
Fréquente l'église	Souvent	40,91	59,09	100
	Pas souvent	43,64	56,36	100
	Total	42,86	57,14	100

Selon le tableau, les gens très pratiquants sont plus ouverts à l'idée d'épouser quelqu'un d'une autre religion. Mais cette conclusion serait hâtive : La dépendance et faible et pourrait facilement être attribuée au hasard.

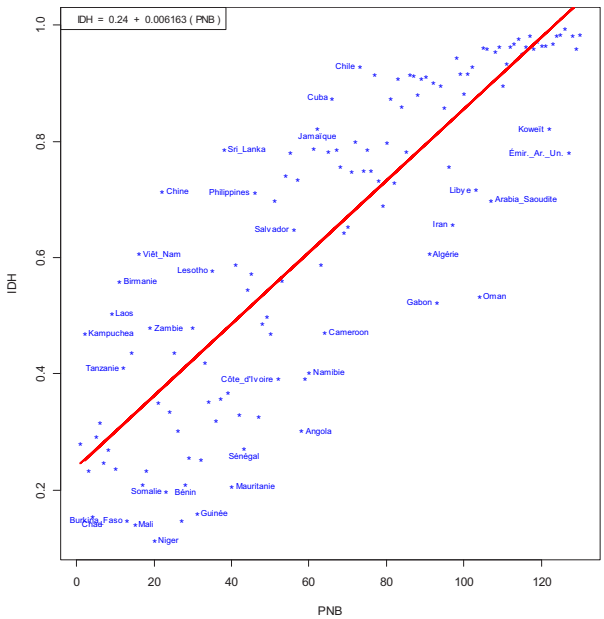
c) Nous réunissons les classes, comme au dernier numéro. Pour la variable « Paume », les valeurs 1, 2 et 3 sont groupées en une classe dont on peut dire qu'ils y croient, la valeur 4 comprenant la classe de ceux qui n'y croient pas. Pour la variable « Astro », nous réunissons les valeurs 1 et 2 en une classe la valeur 4 comprenant la classe dont on dirait qu'ils expriment, à différents degrés, des doutes. Distributions conditionnelles :

		Croit à la signification de la ligne de vie?		
		Oui	Non	Total
Croit à l'astrologie	Non	34,4	65,6	100
	Oui	57,1	42,9	100
	Total	45	55	100

La réponse ici est plus claire : ceux qui croient en l'astrologie sont plus nombreux à croire à la ligne de vie.

Exercice 2.19

Voici le nuage de points :



On perçoit clairement une dépendance entre le PNB et l'IDH. On identifie les pays dont l'indice de développement humain est inférieur à ce qu'il devrait être compte tenu de son PNB (par exemple, la Mauritanie, Oman, la Namibie); ainsi que ceux dont l'indice de

développement humain est supérieur au niveau attendu compte tenu du PNB (par exemple, la Chine, le Sri Lanka, le Chili). L'explication relève de facteurs culturels au sujet desquels le lecteur peut spéculer.

Exercice 2.20

Voici les moyennes brutes et ajustées :

	Femmes	Hommes	Différence (Femmes – Hommes)
Moyennes brutes	20713	16137	4576
Moyennes ajustées	20482	16633	3849

Les femmes ont une moyenne de 4576 $ de plus que les hommes. Mais c'est en partie parce qu'elles travaillent dans des départements où la paie est élevée. Si la distribution (selon le département) des hommes et des femmes avait été la même, la différence serait réduite à 3849 $.

Exercice 2.21

	Femmes	Hommes	Différence Hommes-femmes
Moyennes brutes	73941	83910	9969
Moyennes ajustées	75767	82626	6858

Dans tous les départements, les hommes sont mieux payés que les femmes. C'est en partie ce qui explique l'écart de 9969 $. Mais un autre facteur qui contribue à l'écart, c'est le fait qu'ils sont particulièrement nombreux dans certains départements (dont Management et Sciences économiques) dans lesquels les écarts sont particulièrement importants. Si on élimine cet effet, il reste un écart de 6858 $ qui ne peut être attribué aux départements d'appartenance.

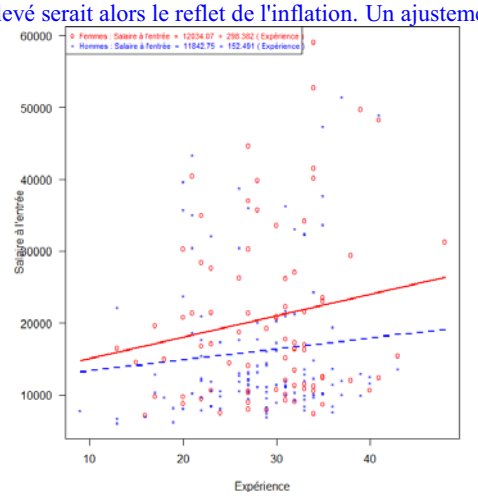
Exercice 2.22

Droites de régression

Femmes : Salaire ajusté = 32600+577(Expérience) Hommes : Salaire ajusté = 30267+688(Expérience)

	Femmes	Hommes	Femmes-Hommes
Salaires réels	20713	16137	4576
Salaires ajustés	20548	16194	4354

Les femmes gagnent plus que les hommes (4576 $ de plus), ce qui pourrait en partie s'expliquer par une expérience légèrement supérieure. Mais une différence importante persiste: l'avantage des femmes se voir réduit à 4354 $. La différence d'expérience est minime et ne saurait expliquer la différence de salaire. Un autre facteur pourrait être en jeu: la date d'engagement, en moyenne plus tardive pour les femmes; leur salaire plus élevé serait alors le reflet de l'inflation. Un ajustement pour l'inflation s'impose.

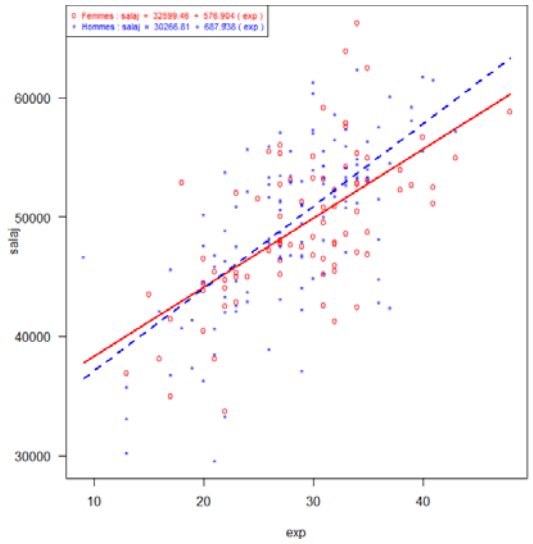


Droites de régression

Femmes : Salaire ajusté = 32600+577(Expérience) Hommes : Salaire ajusté = 30267+688(Expérience)

	Femmes	Hommes	Hommes-Femmes
Salaires réels	49380	49639	259
Salaires ajustés	49064	49899	835

Les femmes gagnent moins (259 $ de moins), malgré leur niveau d'expérience légèrement supérieur. Si les hommes et les femmes avaient eu le même nombre d'années d'expérience, elles auraient gagné encore moins: 835 $ de moins. Le graphique suivant montre qu'il faut interpréter les résultats avec soins. Les droites ne sont pas parallèles, ce qui veut dire que si nous avions supposé un différent niveau d'expérience, les moyennes ajustées auraient été différentes.



Exercice 2.23

Résumé des résultats:

	1995	2007	Augmentation
Moyennes brutes	22673	25137	2464
Moyennes ajustées pour la scolarité	22776	24821	2046

Globalement, les revenus ont augmenté de 2464 $ entre 1995 et 2007. Cette augmentation pourrait-elle être expliquée par une augmentation du niveau de scolarité (qui a augmenté aussi : remarquez l'accroissement des fréquences aux deux plus haut niveaux de scolarité? À peine, si on élimine l'effet de scolarité, l'augmentation passe à 2046, une augmentation qu'on ne peut pas attribuer à un effet de scolarité.

Exercice 2.24

	1995	2007	Augmentation
Moyennes brutes	22618	24994	2376
Moyennes ajustées pour l'âge	22658	24994	2336

Le calcul des moyennes brutes montre une augmentation de 2376 $, mais l'âge n'y contribue presque pas puisque, une fois éliminée, l'augmentation demeure presque inchangée, soit 2336 $.

Exercice 2.25

Un calcul analogue à ceux effectués aux numéros précédents donne les résultats suivants

	2000	2006	Augmentation
Moyennes brutes	37,68	43,27	5,58
Moyennes ajustées	37,99	43,19	5,20

Le taux de fécondité a augmenté, en partie à cause d'un changement dans la distribution de l'âge. La proportion de femmes dans la tranche de 25 à 29 ans a augmenté, ce qui explique en partie l'augmentation globale du taux de fécondité. Si la distribution d'âge n'avait pas changé, l'augmentation serait réduite de 5,58 à 5,20.

Exercice 2.26

Le revenu moyen est de 2187 $ de plus à Halifax. Avec la pondération habituelle, cet écart est réduit légèrement à 1929 $. Pour certains types de familles, les revenus sont supérieurs à St John's alors que pour d'autres types c'est le contraire. Ce que ceci a pour effet, c'est que la différence de moyennes ajustées entre les deux villes peut non seulement varier selon la pondération utilisée, mais elle peut changer de sens: on peut avec un choix approprié de pondération montrer soit que St. John's est plus riche, soit que Halifax est plus riche.

Exercice 2.27

	Winnipeg	Montréal	Différence
Moyenne brutes (ce que sont les moyennes en réalité)	93 295	89 895	3401
Pondération commune (ce qu'auraient été les moyennes si la distribution des types de ménage avait été celle des deux populations réunies)	92 662	90 156	2506
Pondération Winnipeg (ce qu'auraient été les moyennes si les types de ménage étaient distribués à Montréal comme ils le sont à Winnipeg)	93 295	91 401	1895
Pondération Montréal (ce qu'auraient été les moyennes si les types de ménage étaient distribués à Winnipeg comme ils le sont à Montréal)	92 529	89 895	2634
Pondération artificielle (ce qu'auraient été les moyennes si les types de ménage étaient distribués selon les fréquences arbitraire 0,11; 0,41; 0,10; 0,3 ; 0,08)	84 271	85 696	-1425

Exercice 2.28

	1990	1995	*D*ifférence (1995-1990)
Moyennes brutes	37652	37556	-96
Moyennes ajustées pour la scolarité	38105	37099	-1006
Moyennes ajustées pour l'âge	38084	37074	-1010
Moyennes ajustées pour l'âge et la scolarité	38545	36656	-1889

Commandes Excel
Distributions conjointes

Comment déterminer la distribution conjointe de deux variables discrètes

Dans un échantillon de 90 ménages, on obtient les valeurs de deux variables :

`Enfants` : Le nombre d'enfants dans le ménage; et

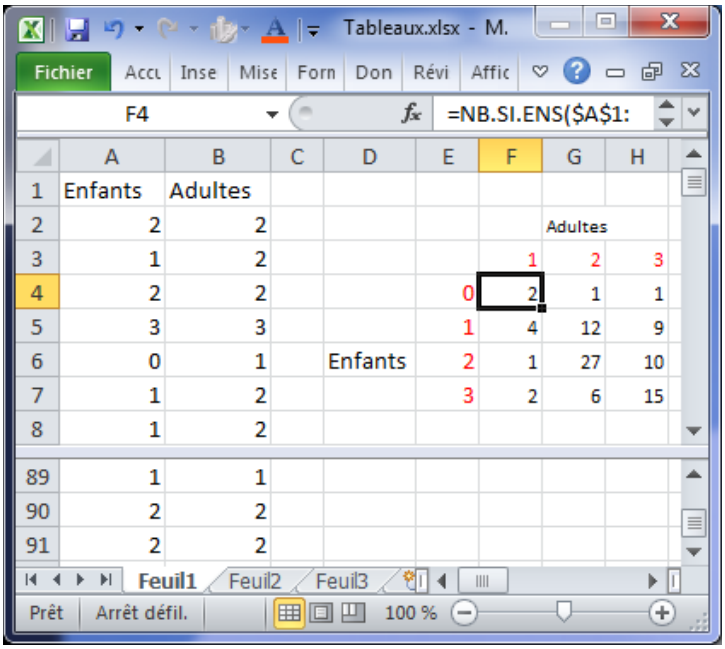
`Adultes` : Le nombre d'adultes dans le ménage.

Les observations sont inscrites dans la plage `A2-B91`.

Les valeurs distinctes de la variable `Enfants` sont inscrites dans la plage `E4-E7` et celles de la variable `Adultes` dans la plage `F4-H4`. Ces deux tables constituent les deux marges du tableau. La commande placée en `F4` (qui sera la première cellule du tableau) est celle-ci :

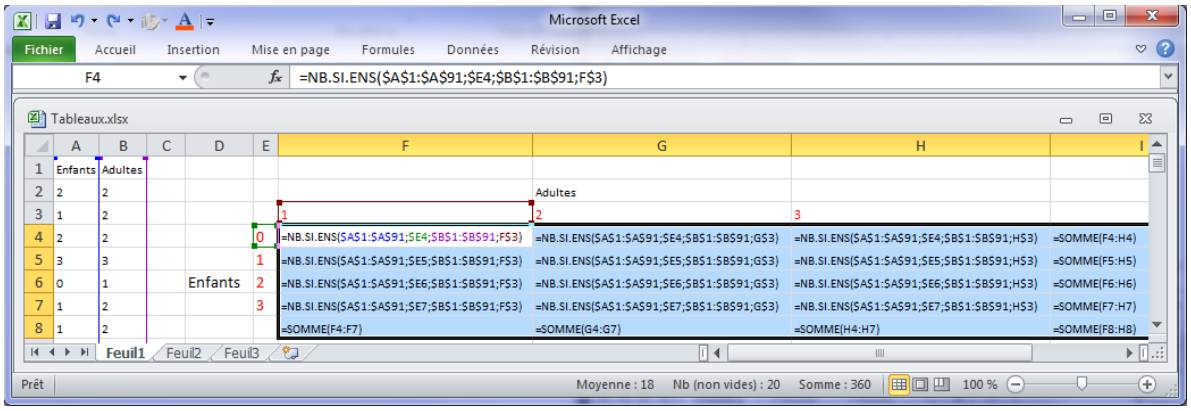
```
=NB.SI.ENS($A$1:$A$91;$E4;$B$1:$B$91;F$3)
```

Cette commande est ensuite étalée sur les lignes et les colonnes du tableau. Voici le résultat :



Voici l'ensemble des commandes (incluant celles qui donnent les sommes des lignes et les sommes des colonnes.

Voici l'ensemble des commandes (incluant celles qui donnent les sommes des lignes et les sommes des colonnes.


Microsoft Excel
Fichier
Accueil
Insertion
Mise en page
Formules
Données
Révision
Affichage
F4
=NB.SI.ENS($A$1:$A$91;$E4;$B$1:$B$91;F$3)
Tableaux.xlsx
Enfants
Adultes
Feuil1
Feuil2
Feuil3
Prêt
Moyenne : 18
Nb (non vides) : 20
Somme : 360
100 %