CORRECTION Exercices Chapitre 2 - Statistiques à deux variables.

Exercice 27

1. On complète le tableau auquel on rajoute une ligne et une colonne :

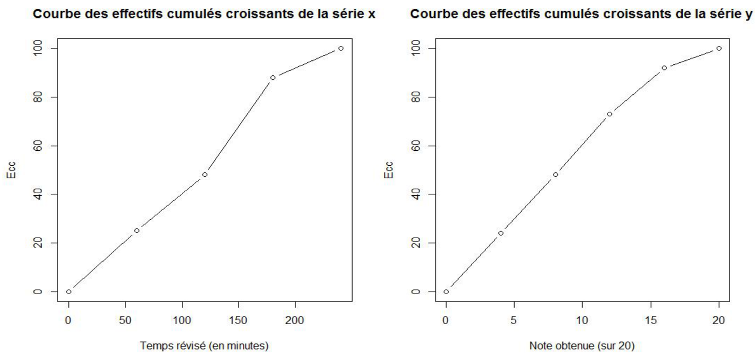
x \ y	$[0;4[$	$[4;8[$	$[8;12[$	$[12;16[$	$[16;20[$	$n_{i.}$	$n_{i.}\nearrow$
$[0;60[$	16	6	2	1	0	25	25
$[60;120[$	5	10	5	3	0	23	48
$[120;180[$	2	7	17	10	4	40	88
$[180;240[$	1	1	1	5	4	12	100
$n_{.j}$	24	24	25	19	8	100	–
$n_{.j}\nearrow$	24	48	73	92	100	–	–

2. Les étudiants ayant eu entre 8 et 20 à leur examen sont au nombre de $n_{.3}+n_{.4}+n_{.5}=25+19+8=52$.
3. Le pourcentage d'étudiants ayant eu une note inférieure à 8 et ayant révisé moins de 2 heures est égal à $\frac{n_{11}+n_{12}+n_{21}+n_{22}}{n}\times 100=\frac{16+6+5+10}{100}\times 100=37\%$
4. Le pourcentage d'étudiants ayant eu une note supérieure à 8 et ayant révisé plus de 2 heures est égal à $\frac{n_{33}+n_{34}+n_{35}+n_{43}+n_{44}+n_{45}}{n}\times 100=\frac{17+10+4+1+5+4}{100}\times 100=41\%$.
5. Les histogrammes sont donnés par les graphiques :

Le code R est le suivant :

```
> ni.<-c(25,23,40,12)
> n.j<-c(24,24,25,19,8)
> classes.X<-c(0,60,120,180,240)
> classes.Y<-c(0,4,8,12,16,20)
> centres.classes.X<-c(30,90,150,210)
> centres.classes.Y<-c(2,6,10,14,18)
> serie.X<-c(rep(centres.classes.X,ni.))
> serie.Y<-c(rep(centres.classes.Y,n.j))
> par(mfrow=c(1,2))
> hist(serie.X,classes.X,xlab='Temps révisé (minutes)',ylab='Effectifs',
+ main='Histogramme de la série x')
> hist(serie.Y,classes.Y,xlab='Note obtenue (sur 20)',ylab='Effectifs',
+ main='Histogramme de la série y')
```

6. Les courbes des effectifs cumulés croissants sont données par les graphiques :

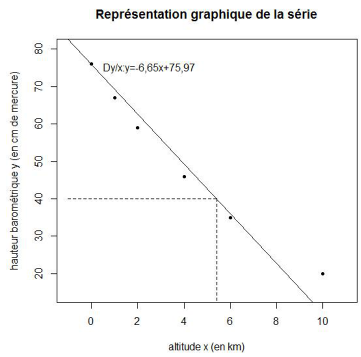


Le code R est le suivant :

```
> diagX<-plot(classes.X,c(0,cumsum(ni.)),type='b',xlab='Temps révisé (en minutes)',ylab='Ecc',
+ main='Courbe des effectifs cumulés croissants de la série x')
> diagY<-plot(classes.Y,c(0,cumsum(n.j)),type='b',xlab='Note obtenue (sur 20)',ylab='Ecc',
+ main='Courbe des effectifs cumulés croissants de la série y')
```

Exercice 28

1. Le nuage de points est donné ci-dessous :



Le code R est le suivant :

```
> x<-c(0,1,2,4,6,10)
> y<-c(76,67,59,46,35,20)
> plot(x,y,xlab='altitude x (en km)',ylab='hauteur barométrique y (en cm de mercure)',
+ main='Représentation graphique de la série',pch=20,xlim=c(-1,11),ylim=c(15,80))
```

2. L'équation de régression de y en x est donnée par $D_{y/x}$: $y = ax + b$ où $a = \frac{Cov(x,y)}{V(x)}$ et $b = \overline{y} - a\overline{x}$. On complète le tableau de l'énoncé :

i	1	2	3	4	5	6	TOTAL
x_i	0	1	2	4	6	10	23
y_i	76	67	59	46	35	20	303
x_i^2	0	1	4	16	36	100	157
x_iy_i	0	67	118	184	210	200	779

On en déduit $\overline{x} = \frac{23}{6} = 3,83$, $\overline{y} = \frac{303}{6} = 50,5$, $V(x) = \frac{157}{6} - (3,83)^2 = 11,50$, $\mathrm{Cov}(x,y) = \frac{779}{6} - 3,83 \times 50,5 = -63,58$ à 10^{-2} près. Par conséquent, $a = \frac{-63,58}{11,47} = -5,54$ et $b = 50,5 - (-5,54) \times 3,83 = 71,72$. Ainsi la droite recherchée admet pour équation $y = -6,65x + 75,97$, qu'on peut tracer sur le graphique précédent à l'aide des commandes sous R(à mettre à la suite du code précédent) :

```
> abline(71.72,-5.54)
> text(2.5,75,'Dy/x :y=-5,54x+71,72')
```

3. Sachant que la hauteur barométrique en un lieu est de 40 cm, l'altitude est estimée en résolvant $40 = -5,54 \times x + 71,72 \Leftrightarrow x = 5,725632$ à 10^{-6} près soit approximativement 5726m. On peut représenter graphiquement ces données sur le graphe précédent grâce aux lignes de commande qui suivent :

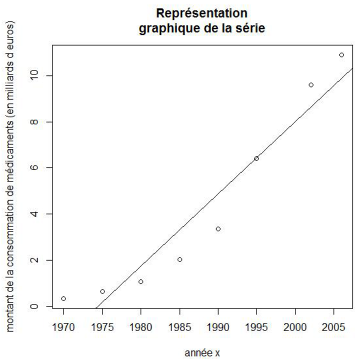
```
> segments(-1,40,5.73,40,lty='dashed')
> segments(5.73,40,5.73,0,lty='dashed')
```

Exercice 29 On complète le tableau de l'énoncé :

i	1	2	3	4	5	6	7	8	TOTAL
x_i	1970	1975	1980	1985	1990	1995	2002	2006	15903
y_i	0,352	0,6660	1,073	2,026	3,369	6,420	9,592	10,89	34,39
x_i^2	3880900	3900625	3920400	3940225	3960100	3980025	4008004	4024036	31614315
y_i^2	0,12	0,44	1,15	4,10	11,35	41,22	92,01	118,59	268,99
x_iy_i	693,44	1315,35	2124,54	4021,61	6704,31	12807,90	19203,18	21845,34	68715,67

1. On déduit du tableau les moyennes $\overline{x} = \frac{15903}{8} = 1987,88$ et $\overline{y} = \frac{34,39}{8} = 4,30$. Par conséquent, $G(1987,88; 4,30)$.
2. Déterminons les variances de x et de y. On a $V(x) = \frac{31614315}{8} - (1987,88)^2 = 122,48$ et $V(y) = \frac{268,99}{8} - (4,30)^2 = 15,13$. On a ensuite $\text{Cov}(x,y) = \frac{68715,67}{8} - 1987,88 \times 4,30 = 41,58$. Finalement, $r = \frac{41,58}{\sqrt{122,48} \times \sqrt{15,13}} = 0,97$. Il y a donc une forte corrélation linéaire entre x et y, ou de manière équivalente un lien de dépendance linéaire très significatif entre ces deux variables.
3. L'équation de régression de y en x est donnée par $D_{y/x} : y = ax + b$ où $a = \frac{Cov(x,y)}{V(x)}$ et $b = \overline{y} - a\overline{x}$. On a $a = \frac{41,58}{122,48} = 0,34$ et $b = 4,30 - 0,34 \times 1987,88 = -671,58$. Ainsi la droite recherchée admet pour équation $y = 0,34x - 671,58$, qu'on peut représenter graphiquement à l'aide du code ci-dessous :

```
> x<-c(1970,1975,1980,1985,1990,1995,2002,2006)
> y<-c(0.352,0.666,1.073,2.026,3.369,6.420,9.592,10.89)
> plot(x,y,xlab='année x',ylab='montant de la consommation de médicaments (en milliards d euros)',
+ main='Représentation graphique de la série')
> abline(-671.5792,0.3395)
```



4. Il suffit de remplacer x par 2010 dans l'équation précédente : on obtient $y = 0,34 \times 2010 - 671,58 = 11,82$ qui est donc une estimation (en milliards d'euros) de la consommation de médicaments des ménages en France en 2010.

Exercice 30

1. On utilise le tableau de la page suivante pour cette question et la suivante. On déduit de ces valeurs que $\overline{x} = \frac{16}{4} = 4$, $\overline{y} = \frac{2,4}{4} = 0,6$, $V(x) = \frac{84}{4} - 4^2 = 5$, $V(y) = \frac{2,18}{4} - 0,6^2 = 0,185$ et $\text{Cov}(x,y) = \frac{13,4}{4} - 4 \times 0,6 = 0,95$. Par conséquent, $r_{x,y} = \frac{0,95}{\sqrt{5} \times \sqrt{0,185}} = 0,988$ à 10^{-3} près.
2. On a, toujours d'après le tableau, $\overline{z} = 2,25$, $V(z) = \frac{23,3}{4} - (2,25)^2 = 0,76$ et $\text{Cov}(x,z) = \frac{41,2}{4} - 4 \times 2,25 = 1,3$. Ainsi, $r_{x,z} = \frac{1,3}{\sqrt{5}\sqrt{0,76}} = 0,667$ à 10^{-3} près.
3. Étant donné que $r_{x,y} > r_{x,z}$, c'est l'élément y_i qui est le plus susceptible d'avoir influencé l'augmentation des prêts de livres x_i.
4. Comme $a = \frac{cov(x,y)}{V(x)} = \frac{0,95}{5} = 0,19$ et $b = 0,6 - 0,19 \times 4 = -0,16$, on trouve aisément que $D_{y/x} : y = 0,19x - 0,16$.

5. Si on pose $x = 9$ dans l'équation précédente, on trouve $y = 0,19 \times 9 - 0,16 = 1,55$. Conclusion, une estimation du nombre de nouveaux lecteurs inscrits y_i, si l'on pense que l'augmentation du nombre des prêts sera de 9000 en 2007, est égale à 1550.

	2003	2004	2005	2006	TOTAL
x_i	3	7	1	5	16
y_i	0,3	1,2	0,1	0,8	2,4
z_i	0,9	3,2	2,1	2,8	9
x_i^2	9	49	1	25	84
y_i^2	0,09	1,44	0,01	0,64	2,18
z_i^2	0,81	10,24	4,41	7,84	23,3
x_iy_i	0,9	8,4	0,1	4,0	13,4
x_iz_i	2,7	22,4	2,1	14	41,2

Exercice 31

1. On a les données suivantes (les chiffres entre crochets correspondent à $n_{ji}y_jx_i$) :

y_j \ x_i	2	8	12	18	$n_{j.}$	$n_{j.}y_j$	$n_{j.}y_j^2$
6	8 [96]	1 [48]	1 [72]	0 [0]	10	60	360
9	1 [18]	10 [720]	2 [216]	0 [0]	13	117	1053
11	1 [22]	2 [176]	14 [1848]	1 [198]	18	198	2178
14	0 [0]	0 [0]	2 [336]	7 [1764]	9	126	1764
$n_{.i}$	10	13	19	8	50	501	5355
$n_{.i}x_i$	20	104	228	144	496	[5514]	
$n_{.i}x_i^2$	40	832	2736	2592	6200		

   On a $\overline{x} = \frac{496}{50} = 9,92$, $\overline{y} = \frac{501}{50} = 10,02$, $V(x) = \frac{6200}{50} - (9,92)^2 = 25,39$, $V(y) = \frac{5355}{50} - (10,02)^2 = 6,70$, $\text{Cov}(x,y) = \frac{5514}{50} - 9,92 \times 10,02 = 10,88$. On en déduit que $a = \frac{\text{Cov}(x,y)}{V(x)} = \frac{10,88}{25,39} = 0,43$ et $b = 10,02 - 0,43 \times 9,92 = 5,75$. Finalement, une équation de régression de Y en X est donnée par $D_{Y/X} : y = 0,43x + 5,75$.
2. L'équation de la droite de régression de X en Y est $D_{X/Y} : x = a'y + b'$ avec $a' = \frac{10,88}{6,70} = 1,62$ et $b' = 9,92 - 1,62 \times 10,02 = -6,31$ soit $D_{X/Y} : x = 1,62y - 6,31$.
3. On a $r = \frac{10,88}{\sqrt{25,39} \times \sqrt{6,70}} = 0,83$.
4. On pose $y = 15$ dans l'équation de $D_{X/Y}$. On obtient $x = 1,62 \times 15 - 6,31 = 17,99$. La note prévue en méthodes quantitatives est approximativement 18/20.
5. On pose $x = 4$ dans l'équation de $D_{Y/X}$. On obtient $y = 0,43 \times 4 + 5,75 = 7,47$. La note prévue en marketing est approximativement 07,5/20.

Exercice 32

1. Le pourcentage de personnes pesant moins de 80 kilos et mesurant plus de 180 centimètres est de $\frac{2}{60} \times 100 = 3,33\%$ à 10^{-2} près.
2. La proportion de personnes pesant entre 60 et 90 kilos est de $\frac{15+19+10}{60} \times 100 = 73,33\%$ à 10^{-2} près.
3. • $2975 = 175 \times 17$.
   • $106875 = 19 \times 75^2 = 106875$.
   • $94350 = 6 \times 185 \times 85$.
4. On a $\overline{x} = \frac{4290}{60} = 71,5$ et $\overline{y} = \frac{10240}{60} = 170,67$ à 10^{-2} près.
5. On a $V(x) = \frac{314900}{60} - (71,5)^2 = 136,08$ et $V(y) = \frac{1755900}{60} - (170,67)^2 = 136,75$ à 10^{-2} près. On en déduit $\sigma(x) = \sqrt{136,08} = 11,67$ et $\sigma(y) = \sqrt{(}136,75) = 11,69$.
6. On a $\text{Cov}(x,y) = \frac{739150}{60} - 71,5 \times 170,67 = 116,26$ et on en déduit $r = \frac{116,26}{\sqrt{136,08} \times \sqrt{136,75}} = 0,85$.
7. Un ajustement linéaire n'est pas justifié car $|r| < 0,87$. On ne peut donc prévoir à l'aide de l'échantillon donné la taille d'une personne à partir de son poids.