Corrigé Rattrapage Data Explo 2024

1. Exercice 1 :

1) Population, taille : les 209 pays.
Variables : Espérance de vie et PIB/habitant.
Nature : les 2 variables sont quantitatives continues.

2)

a) Un nuage allongé suivant une ligne droite indiquerait une bonne régression linéaire.

b) Une valeur atypique a tendance à détourner la droite de régression des autres points du nuage, et à induire une régression de mauvaise qualité.

c) L'élimination d'une valeur pourrait révéler d'autres points atypiques qui étaient cachés.

3) $\sum_{i=1}^{n} x_i = 3420$, $\sum_{i=1}^{n} y_i = 14728$, $\sum_{i=1}^{n} x_i^2 = 126677$, $\sum_{i=1}^{n} y_i^2 = 1058461$ et $\sum_{i=1}^{n} x_i y_i = 257436.$

a) $\overline{x} = \dfrac{3420}{209} = 16.36, \quad \overline{y} = \dfrac{14728}{209} = 70.47$

$Var(x) = s_x^2 = \dfrac{126677}{209} - (16.36)^2 = 338.34 \Longrightarrow s_x = \sqrt{338.34} = 18.4$

$Var(y) = s_y^2 = \dfrac{1058461}{209} - (70.47)^2 = 98.54 \Longrightarrow s_y = \sqrt{98.54} = 9.93$

$Cov(x,y) = c_{xy} = \dfrac{257436}{209} - 16.36 \times 70.47 = 78.62$

b) $r_{xy} = \dfrac{c_{xy}}{s_x s_y} = \dfrac{78.62}{\sqrt{338.34 \times 98.54}} = 0.43$

Ce coefficient est positif, indiquant que le PIB/hab. évolue dans le même sens que l'espérance de vie.
Mais il est plutôt faible. L'espérance de vie n'explique qu'une petite partie de la variabilité du PIB/hab.

c) $\widehat{a} = \dfrac{c_{xy}}{s_x^2} = \dfrac{78.62}{338.34} = 0.232$ et $\widehat{b} = \overline{y} - \widehat{a}\,\overline{x} = 70.47 - 0.232 \times 16.36 = 66.67$

L'équation de la droite de régression est : $y = 0.232\,x + 66.67$

d) Pour $x = 15$, on obtient : $y = 0.232 \times 15 + 66.67 = 70.15$

L'epérance de vie serait de 70.15 ans pour un pays au PIB/hab. de 15 milles dollars.

4)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.569e+01  2.577e+00  29.367   <2e-16 ***
PIB_30prem  3.278e-05  4.165e-05   0.787    0.438
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.469 on 28 degrees of freedom
Multiple R-squared:  0.02164,   Adjusted R-squared:  -0.0133
F-statistic: 0.6194 on 1 and 28 DF,  p-value: 0.4379
```

a) $R^2 = 0.02164 \implies 2.164\ \%$ seulement de la variabilité de l'espérance de vie sont expliqués par le PIB/hab.

b) Le coef de corrélation est la racine carrée positive de R^2 car le coef. $\widehat{a} = 3.27 \times 10^{-5} > 0$

$r_{xy} = \sqrt{0.02164} = 0.147$

c) L'équation de la droite de régression donnée par R est la suivante :

$$y = 3.27 \times 10^{-5} x + 75.69$$

d) Le coef. de corrélation est passé de 0.43 à 0.147

La nouvelle droite de régression explique encore moins l'espérance de vie qu'avant.

La régression est encore moins justifiée. Ceci s'explique par le fait que l'on s'est limié aux pays les plus riches où les différences d'espérance de vie peuvent s'expliquer par des choses bien différentes du PIB/hab., par le système de santé par exemple.

2. Exercice 2 :

	Classes d'âge						
nombre d'accidents	18-30	31-40	41-50	51-60	61-70	>70	Total
aucun	758	851	786	723	693	691	4502
un seul	74	62	51	66	50	35	338
deux accidents	31	25	22	18	15	10	121
plus que deux accidents	9	10	6	5	7	2	39
Total	872	948	865	812	765	738	5000

1) $n_{2,1} = 74$ conducteurs sont agés de 18 à 30 ans et ont eu un seul accident.

$n_{\bullet,3} = 865$ conducteurs sont agés de 41 à 50 ans.

$f_{2,\bullet} = \dfrac{338}{5000} = 6.76\%$ des conducteurs ont eu un seul accident.

2)

	Classes d'âge					
nombre d'accidents	18-30	31-40	41-50	51-60	61-70	>70
aucun	0,869	0,898	0,909	0,890	0,906	0,936
un seul	0,085	0,065	0,059	0,081	0,065	0,047
deux accidents	0,036	0,026	0,025	0,022	0,020	0,014
plus que deux accidents	0,010	0,011	0,007	0,006	0,009	0,003

La somme de chaque colonne vaut 1, il s'agit donc de profils-colonnes.

3) Ces profils-colonnes doivent être comparés au profil-colonne moyen obtenu en divisant les effectifs marginaux des nombres d'accidents par l'effectif total. On obtient :

0,9004	0,0676	0,0242	0,0078

4) Le profil-colonne 18-30 est légèrement différent tout comme celui des >70.
On peut s'attendre à un faible lien entre âge et nombre d'accidents.

5)

	Classes d'âge						
nombre d'accidents	18-30	31-40	41-50	51-60	61-70	>70	Total
aucun	785,1	853,6	778,8	731,1	688,8	664,5	4502
un seul	58,9	64,1	58,5	54,9	51,7	49,9	338
deux accidents	21,1	22,9	20,9	**19,7**	18,5	17,9	121
plus que deux accidents	6,8	7,4	6,7	6,3	6,0	5,8	39
Total	872	948	865	812	765	738	5000

6) Ce tableau d'effectifs théoriques est calculé sous l'hypothèse d'indépendance entre les 2 variables.

La formule de calcul est $n_{i,j,theo} = \dfrac{n_{i,\bullet} \times n_{\bullet,j}}{n_{\bullet,\bullet}} = \dfrac{\text{total ligne} \times \text{total colonne}}{\text{effectif total}}$

L'effectif en gras a été obtenu par : $\dfrac{121 \times 812}{5000} = 19.7$

7) La distance du khi2 est donnée par : $\chi^2 = \sum_{i,j} \dfrac{(n_{i,j,theo} - n_{i,j,obs})^2}{n_{i,j,theo}}$.

8) Le nombre de degrés de liberté est : $(4-1)(6-1) = 15$ d.d.l.
Le seuil est donc $C = 25.00$.
$\chi^2 = 27.57 > C \implies$ Les 2 variables sont liées.

3. Exercice 3 :

Catégorie	Effectif	Moyenne	Variance
Cad	25	3.92	1.51
OQ	50	7.34	5.26
ONQ	75	6.74	3.15
Total	150	6.47	4.96

1) Si on note oq_i les durées de chômage des ouvriers qualifiés, la moyenne est :

$$\frac{1}{50}\sum_{i=1}^{50} oq_i.$$

Si on note c_i les durées de chômage des cadres, et $\overline{c}$ la moyenne, la variance est donnée par : $\frac{1}{25}\sum_{i=1}^{n}(c_i-\overline{c})^2 = \frac{1}{25}\sum_{i=1}^{n}c_i^2-(\overline{c})^2$

2) La variance intra-classes est donnée par :

$$V_{intra} = \frac{1}{150}(25\times 1.51 + 50\times 5.26 + 75\times 3.15) = 3.58$$

3) La variance inter-classes est donnée par :

$$V_{inter} = \frac{1}{150}\left[25\times(3.92-6.47)^2 + 50\times(7.34-6.47)^2 + 75\times(6.74-6.47)^2\right]$$

$$V_{inter} = 1.37$$

4) La formule de décomposition de la variance assure que :

$$V_{totale} = V_{intra} + V_{inter}$$

Et on a bien : $3.58 + 1.37 \simeq 4.96$

5) Le rapport de corrélation donne : $R^2 = \frac{V_{inter}}{V_{totale}} = \frac{1.37}{4.96} = 27.6\%$

Conclusion : La catégorie socio-professionnelle explique environ 27.6% de la variabilité de la durée de chômage.