ING1 - GI/GM/DU-CPI/MIM

Rattrapage Data Exploration

Département : **Mathématiques**	*Date :* **5 février 2024**
Nombre de pages : **3**	*Durée :* *2h*

2 feuilles manuscrites recto verso et calcultarice autorisées.

Exercice 1. -
On s'intéresse au lien qu'il peut y avoir entre l'espérance de vie (en années) et le PIB par habitant (en milliers de dollars en 2017) dans 209 pays (données du Fonds Monétaires International FMI). Le tableau ci-dessous représente certaines valeurs recueillies par pays.

Pays	Espérance de vie	PIB/habitant
Albanie	77.22	7.7
Algérie	74.26	7.1
Samoa américaines	73.97	8.0
Andorre	82.36	44.9
Angola	38.48	8.3
Anguilla	80.77	12.2
Antigua et Barbuda	75.26	17.2
⋮	⋮	⋮

1. Quelle est la population étudiée ? Donner sa taille. Quelles sont les variables étudiées ? Donner leur type.

2. (a) Quelle forme particulière du nuage de points entre les deux variables étudiées va nous diriger vers une modélisation par la méthode de régression ?
   (b) Quelle peut être l'influence d'une valeur atypique sur la droite de régression ?
   (c) Quels risques peut impliquer la suppression d'une valeur atypique unique, en terme d'apparition de nouvelles valeurs atypiques ?

3. On note X le PIB (par habitant) et Y l'espérance de vie. On dispose des calculs intermédiaires suivants :

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

   (a) Déterminer les moyennes et les écarts-types de X et Y ainsi que la covariance entre X et Y.
   (b) En déduire le coefficient de corrélation entre ces deux variables. Interpréter.
   (c) Déterminer la droite de régression de l'espérance de vie en fonction du PIB.
   (d) Donner une estimation de l'espérance de vie dans un pays ayant un PIB/habitant de 15 milliers de dollars.

4. En utilisant le logiciel R sur ces **données restreintes aux trente pays ayant le plus fort PIB**, on a obtenu le résultat suivant :

```
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) Donner le pourcentage de variance de l'espérance de vie expliquée par le PIB.

(b) Donner une estimation du coefficient de corrélation entre l'espérance de vie et le PIB.

(c) Déterminer la droite de régression de l'espérance de vie en fonction du PIB.

(d) Pouvez-vous comparer les résultats obtenus avant et après la sélection des pays ? Peut-on faire confiance aux résultats obtenus sur les 30 pays ?

Exercice 2. -
On souhaite savoir s'il existe un lien entre l'âge du conducteur ou conductrice et le nombre d'accidents. Pour cela on a fait une sondage et on a regroupé les résultats dans le tableau de contingence suivant :

	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. On note $n_{i,j}$ l'effectif de la i-ème modalité du nombre d'accidents et la j-ème modalité des classes d'âge, et $f_{i,j}$ la fréquence correspondante.

   Déterminer les valeurs de $n_{2,1}$, de $n_{\bullet,3}$ et de $f_{2,\bullet}$ en indiquant ce que ces valeurs représentent.

2. Afin d'avoir une première impression du lien éventuel entre âge et nombre d'accidents, on dresse le tableau suivant :

	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

   Indiquer, en justifiant votre réponse, s'il s'agit d'un tableau de fréquences, de profils-lignes ou de profils-colonnes ?

3. A quel profil moyen doit-on le comparer ? Calculer ce profil moyen.

4. Faites cette comparaison et donner une première réponse à la question du lien ou de l'indépendance entre âge et nombre d'accidents.

5. Afin de donner une réponse plus précise, nous construisons le tableau de contingence théorique suivant :

	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. Quelle est l'hypothèse qui permet de calculer ces effectifs théoriques ? Donner la formule permettant de calculer la valeur d'une cellule, et appliquer la pour obtenir la valeur de la cellule en gras.

7. La distance du khi2 calculée entre les effectifs théoriques et observés a été calculée et vaut : $\chi^2 = 27.57$. Expliquer comment elle est calculée.

8. Déterminer le nombre de degrés de liberté de cette distance et utiliser la table suivante pour conclure le test d'indépendance.

nombre de degrés de liberté	11	12	13	14	15	16
Seuil du χ^2	19.68	21.03	22.36	23.68	25.00	26.30

Exercice 3. -
On s'intéresse ici au lien éventuel entre la durée de chômage et la catégorie socio-professionnelle.
Nous disposons des données concernant la durée de chômage (exprimée en nombre de semaines) de $n_1 = 25$ cadres(Cad), $n_2 = 50$ ouvriers qualifiés(OQ) et $n_3 = 75$ ouvriers non qualifiés(ONQ).
Voici un aperçu du fichier de données nommé "chomage.csv" :

Durée	Catégorie
2	OQ
4	Cad
0	Cad
11	ONQ
1	OQ
⋮	⋮

Le résumé numérique de la durée de chômage global et par catégorie est le suivant :

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. Donner la formule permettant de calculer la moyenne pour la catégorie OQ et la formule de la variance pour la catégrie Cadres.

2. Calculer la variance intra-classes.

3. Calculer la variance inter-classes sans utiliser le résultat précédent.

4. Quelle formule lie les 2 variances précédentes ? Vérifer à l'aide de vos 2 calculs précédents cette formule.

5. Calculer le rapport de corrélation et donner une réponse à la question posée sur le lien entre catégorie et durée de chômage.