STATISTIQUE: ÉTUDE DE CAS
durée : 2 heures
Aucun document n'est autorisé
Les téléphones portables doivent être éteints

1. Problème 1

Le responsable du comité de sécurité d'une entreprise a effectué une compilation du nombre d'accidents de travail qui se sont produits dans l'usine depuis 2 ans. Cette compilation correspond à 500 jours ouvrables et est présentée dans le tableau ci-dessous. Il indique la répartition du nombre de jours sans accident, avec 1 accident, ..., avec 4 et plus accidents par jour.

Nombre d'accidents en une journée	0	1	2	3	4 et plus
Nombre de jours (fréquences)	194	138	80	52	36

(a) Calculer le nombre moyen d'accidents par jour.

(b) En supposant que la loi de Poisson est convenable pour décrire ce phénomène, déterminer la probabilité pour qu'arrive 0 accident, 1 accident, ..., 4 et plus accidents par jour (à 10^{-3} près).

(c) Peut-on considérer comme vraisemblable l'hypothèse selon laquelle cette variable statistique se comporte selon une loi de Poisson au seuil de signification $\alpha = 0,01$? Indiquer clairement l'hypothèse nulle, le nombre de degré de liberté, la statistique du test, la valeur p et votre conclusion.

2. Problème 2

Ces données sont celles du sondage Gallup sur le niveau d'adaptation au système métrique en fonction de l'âge des répondants.

		Niveau d'adaptation		
		Très difficile	Passablement difficile	Pas difficile
Age	18 à 29 ans	81	138	132
	30 à 49 ans	126	131	94
	50 ans et plus	203	78	69

(a) Pour l'ensemble des répondants, estimer la proportion associé à chaque modalité du niveau d'adaptation.

(b) Estimer la probabilité pour qu'un répondant appartienne à la modalité "30 à 49 ans" du caractère "Age".

(c) Sous l'hypothèse d'indépendance de ces deux caractères, quelle aurait dû être la répartition des 1052 répondants selon les deux caractères analysés?

(d) Peut-on conclure, au seuil $\alpha = 0,05$, que le niveau d'adaptation est lié à l'âge des répondants?

(e) Estimer la distribution des probabilités conditionnelles du niveau d'adaptation sachant que les répondants appartiennent à la classe des 50 ans et plus.

3. Problème 3

L'entreprise SINTRON fabrique un matériau en matière plastique qui est utilisé dans la fabrication de jouets. Le département de contrôle de la qualité de l'entreprise a effectué une étude qui avait pour but d'établir dans quelle mesure la résistance à la rupture de cette matière plastique pouvait être affectée par l'épaisseur du matériau ainsi que la densité de ce matériau. Douze essais ont été effectués et les résultats sont présentés dans le tableau ci-dessous.

Essai numéro	Résistance à la rupture Y	Épaisseur du matériau X_1	Densité X_2
1	37,8	4	4
2	22,5	4	3,6
3	17,1	3	3,1
4	10,8	2	3,2
5	7,2	1	3,0
6	42,3	6	3,8
7	30,2	4	3,8
8	19,4	4	2,9
9	14,8	1	3,8
10	9,5	1	2,8
11	32,4	3	3,4
12	21,6	4	2,8

(a) Pour chaque régression, compléter le tableau suivant:

	Carré moyen résiduel	Écart-type des résidus
Régression avec X_1		
Régression avec X_2		
Régression avec X_1, X_2		

(b) Compléter le tableau d'analyse de variance suivant pour la régression comportant les deux variables explicatives.

Source de variation	Somme de carrés	Degrés de liberté	Carrés moyens	F
Régression X_1, X_2				
Résiduelle				
Totale				

(c) Tester, au seuil de signification $\alpha = 0,05$, l'hypothèse $H_0 : \beta_1 = \beta_2 = 0$ contre l'hypothèse H_1 : au moins un des $\beta_j \neq 0$. Quelle est votre conclusion?

(d) Dans le cas du modèle de régression ne comportant que l'épaisseur du matériau comme variable explicative, déterminer un intervalle de confiance à 95% pour le paramètre β_1.

(e) Est-ce que la contribution marginale de la variable explicative "densité du matériau", lorsqu'elle est introduite à la suite de la variable "épaisseur du matériau" est significative au seuil $\alpha = 0,05$? Connaissez-vous une autre façon d'effectuer ce test? Expliquer.

4. Problème 4

Weisberg donne un jeu de données sur la consommation d'essence des 48 états continentaux des États-Unis. Le jeu de données **fuel2001** se trouve dans la librairie **alr3** (que vous devez installer) de **R**. Pour chaque état on a: l'abréviation du nom de l'état, la population de l'état (Pop), la valeur de la taxe de vente (Tax), le nombre d'individus ayant un permis de conduire (Drivers), le revenu annuel per capita en milliers de dollars (Income), la longueur totale des routes fédérales en milliers de miles (Miles), miles parcouru par habitant (MPC) et la consommation d'essence(FuelC).

(a) Ajuster un modèle de régression multiple de la consommation d'essence (FuelC) avec toutes les variables explicatives. Identifier le R_a^2 et l'écart type des résidus pour ce modèle.

(b) Identifier le meilleur modèle selon le critère suivant: Choisir parmi les variables explicatives, celle pour lequel le test de Student est le moins significatif (à 5%), c'est à dire avec la plus grande p-value. La retirer du modèle et recalculer l'estimation. Arrêter le processus lorsque tous les coefficients sont considérés comme significativement (à 5%) différents de 0. Attention, la variable INTERCEPT ne peut être considérée au même titre que les autres variables, la conserver dans le modèle. Indiquer clairement le modèle retenu et les estimations de paramètres correspondantes.

(c) Identifier le meilleur modèle selon le critère AIC en indiquant le modèle retenu et les estimations de paramètres correspondants.

(d) Effectuer un test de Shapiro-Wilk sur les résidus studentisés du modèle choisi en (c). Indiquer clairement l'hypothèse nulle, la statistique du test, la valeur p et votre conclusion.