Université Mohamed Khider de Biskra
Département Informatique
Module: Analyse de données

Première Année Master
2023/2024

Corrigé type Examen

Solution de l'Exercice 1

- Calculer les moyennes des différents échantillons : $\overline{X}_1 = 24.73$, $\overline{X}_2 = 21.53$ et $\overline{X}_3 = 23.60$.
- Calculer la moyenne globale de toutes les observations : $\overline{X} = \frac{1}{n}(n_1\overline{X}_1 + n_2\overline{X}_2 + n_3\overline{X}_3) = 23.29$.

$$\underbrace{\sum_{j=1}^{p}\sum_{i=1}^{n_j}(X_{ij} - \overline{X})^2}_{SC_{Tot}} = \underbrace{\sum_{j=1}^{p}\sum_{i=1}^{n_j}(X_{ij} - \overline{X}_j)^2}_{SC_{Res}} + \underbrace{\sum_{j=1}^{p} n_j(\overline{X}_j - \overline{X})^2}_{SC_{Fac}}, \quad (1)$$

où, n_j : est la taille du $j^{\text{ième}}$ échantillon (groupe).

SC_{Tot} : est la variation totale qui représente dispersion des données autour de la moyenne générale.

SC_{Fac} : est la variation due au facteur qui représente dispersion des moyennes autour de la moyenne générale.

SC_{Res} : est la variation résiduelle qui représente dispersion des données à l'intérieur de chaque échantillon autour de sa moyenne.

- Compléter le tableau de l'ANOVA à un seul facteur :

source de variation	Somme des carrés SC	Degrés de libertés ddl	Carré moyen CM	ratio F_{obs}	Ficher c
Inter-groupe	31.59	2	15.70	12.02	3.68
Intra-groupe	19.71	15	1.31		
Total	51.30	17			

- ***Décision*** : on constate que $f_{obs} = 12.02 > f_\alpha = 3.68$ (

$$f_\alpha = f(p-1, n-p, 1-\alpha) = f(2, 15, 0.95) = 3.68.$$

), donc les espaces moyens occupés par les informations sont significativement différents d'une bases de données à une autre. Cela signifié que le facteur bases de données influe sur l'espace mémoire occupé par les informations stockées.
Le test à réaliser est:

$$H_0 : \text{ " } \mu_1 = \mu_2 = \mu_3 = \mu \text{ " contre } H_1 : \text{ " } \exists\, i, j \in \{1,2,3\} \text{ tel que } \mu_i \neq \mu_j \text{ ",}$$

Afin de réaliser le test $ANOVA$, principalement trois conditions doit être vérifiées préalablement, à savoir:

- Les p échantillons comparés sont indépendants.
- La variable quantitative étudiée suit une loi normale dans les p populations comparées.
- Les p populations comparées ont la même variance : *Homogénéité* des variances ou *homoscédasticité*.

Solution de l'Exercice 2

							Somme
X $\mu g/\mu l$	0	20	40	60	80	100	300
Y	0	0.21	0.33	0.52	0.58	0.67	2.31
X^2	0	400	1600	3600	6400	10000	22000
Y^2	0	0.042	0.11	0.27	0.34	0.45	1.21
$X * Y$	0	4.10	13.24	30.90	46.72	67.10	162.06

Afin de modéliser ces données, nous avons proposé le modèle linéaire suivant :

$$Y = a\,x + b.$$

1. Calcul des estimateurs des paramètres a et b. On a :

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{6}300 = 50.$$

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{6}3002.3060 = 0.38.$$

$$Cov(x,y) = \frac{1}{n}\sum_{i=n}^{n} x_i y_i - \overline{X}\,\overline{Y} = \frac{1}{6}(162.06) - (50)\,(0.3843) = 7.80$$

$$Var(x) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{X}^2 = \frac{1}{6}(22000) - (50)^2 = 1166.67$$

alors,

$$\hat{a} = \frac{Cov(x,y)}{Var(x)} = \frac{\frac{1}{n}\sum_{i=n}^{n} x_i y_i - \overline{X}\,\overline{Y}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{X}^2} = 0.01.$$

$$\hat{b} = \overline{Y} - \hat{a}\overline{X} = 0.05,$$

de ce fait la droite de régression de l'absorbance (Y) en fonction de la concentration (x) est donnée par :

$$\hat{Y} = 0.01\,x + 0.05.$$

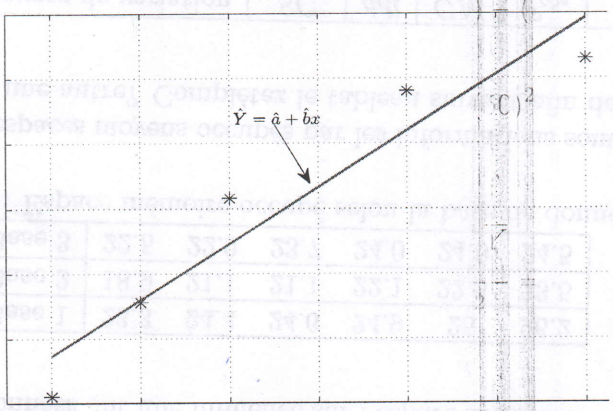

Figure 1: Présentation graphique du nuage des points (X_i, Y_i)

2. Quelle absorbance prévoyez-vous à une concentration 40 $\mu g/\mu l$? Que peut-on conclure?

$$\hat{Y} = 0.0067\ (40) + 0.0503 = 0.3183.$$

On constate que la valeur de régression est très proche de la vraie valeur (0.33), donc à priori le modèle retenu est adéquate pour la représentation des données du tableau.

3. Calcul du coefficient de corrélation linéaire.

$$r = r(x,y) = \frac{Cov(x,y)}{\sigma_x \sigma_y} = 0.99,$$

avec $\sigma_y = \sqrt{var(Y)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \overline{Y}} = \sqrt{0.05} = 0.23$; La valeur du coefficient de corrélation est très proche de 1, i.e. X et Y sont fortement linéairement liés donc le modèle est efficace ce qui confirme les résultats de la question 3).

4. Pour un seuil de risque $\alpha = 5\%$, le modèle proposé est-il pertinent?
Pour répondre à cette question on utilise le test de validation du modèle (Fisher). On d'une part

$$f_c = \frac{\sum_{i=1}^{n}(\hat{y} - \overline{Y})^2/1}{\sum_{i=1}^{n}(y_i - \hat{y})^2/(n-2)} = \frac{0.31/1}{0.01/(6-2)} = 131.54,$$

et d'autre par

$$f_\alpha = f(1, n-2, 1-\alpha) = f(1, 4, 0.95) = 7.71.$$

On constate que $f_c > f_\alpha$, alors on accepte le modèle proposé, c'est-à-dire le modèle est valide (pertinent)