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

Première Année Master
2023/2024

Corrigé type Examen

Solution de l'Exercice 1 Les différentes moyennes de chaque échantillon sont données par :

$$\overline{X}_1 = 27.40,\ \overline{X}_2 = 23.00,\ \overline{X}_3 = 34.5.(0.75p)$$

$$\underbrace{\sum_{j=1}^{p}\sum_{i=1}^{n_j}(\overline{X}_j - \overline{X})^2}_{SC_{Fac}} = \sum_{j=1}^{p} n_j(\overline{X}_j - \overline{X})^2 = 240.46.(1p)$$

En exploitant ces dernières quantités pour le calcul des différentes variations on obtient :(2p)

	SC	ddl	CM	f	f_α
Inter-groupes	240.46	2	120.23	1.69	4.26
Intra-groupes	640.22	9	71.13		
Total	880.68	11			

On constate que $f < f_\alpha$ cela signifie qu'on doit pas rejeter H_0. C'est-à-dire le facteur type de capteurs n'a pas une influence significative leurs durée moyenne de vie. 0.25p

1. Le test à réaliser dans ce cas est le test d'homogénéité de moyenne des deux échantillons cas petits échantillons, dont la formulation du test est donné par :

$$H_0 :\ ''\ \mu_1 = \mu_2\ ''\ \text{contre } H_1 :\ ''\ \mu_1 < \mu_2\ ''(0.25p), \tag{1}$$

mais on doit vérifier d'abord l'homogénéité de leurs variances, et cela en réalisant le test suivant :

$$H_0 :\ ''\ \sigma_1^2 = \sigma_2^2\ ''\ \text{contre } H_1 :\ ''\ \sigma_1^2 \neq \sigma_2^2\ '', (0.25p)$$

Notons que les moyennes des deux échantillons sont données par :
$\overline{X_1} = \frac{1}{n_1}\sum\limits_{i=1}^{7} X_i = \frac{1}{7}\sum\limits_{i=1}^{7} X_i = 26, (0.5p)$ $\overline{X_2} = \frac{1}{n_2}\sum\limits_{i=1}^{9} X_i = \frac{1}{9}\sum\limits_{i=1}^{9} X_i = 32.(0.5p)$

a) La statistique du test d'homogénéité de variance des deux échantillons est donnée par : $F = \frac{\hat{\sigma}_{c2}^2}{\hat{\sigma}_{c1}^2}(0.5p)$ (le fait que $\hat{\sigma}_{c2}^2 > \hat{\sigma}_{c1}^2(0.25p)$) est sa réalisation est $f = 1.15(0.5p)$.

b) La valeur critique du test, pour $\alpha = 5\%$, est : $f_\alpha = f_{(n_2-1, n_1-1, 1-\alpha/2)} = f_{(8,6,0.975)} = 5.60.(0.5p)$

c) On constate que : $f \in [1, f_\alpha[$, cela signifier que les deux échantillons ont la même variance. (0.5p)

d) Le fait que les deux échantillons ont la même variance donc on calcule la variance commune définie par :

$$\hat{\sigma}_c^2 = \frac{(n_1-1)\hat{\sigma}_{c1}^2 + (n_2-1)\hat{\sigma}_{c2}^2}{n_1+n_2-2} = 51.43.(0.5p)$$

e) Ainsi, la statistique du test d'homogénéité de moyenne (1) est donnée par :

$$T = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\hat{\sigma}_c^2(\frac{1}{n_1} + \frac{1}{n_2})}} \quad (0.5p) \text{ et sa réalisation est } \quad t = -1.66.(0.25p)$$

f) La valeur critique du test est : $t_\alpha = t(n_1+n_2-2, 1-\alpha/2) = t(14, 1-0.05) = 1.76.(0.5p)$.

g) On constate que, $t > -t_\alpha$ on ne rejette pas H_0, on peut dire que la durée de vie moyenne des capteurs de type X_1 est égale à celle des capteurs de type X_2 pour un risque $\alpha = 5\%$.(0.5p)

Solution de l'Exercice 2

1. À partir de la présentation graphique (voir figure 1), on constate que le nuage des points est distribué sous une forme linéaire (0.25p), à priori le modèle proposé est adéquat pour l'explication de Y en fonction de X.

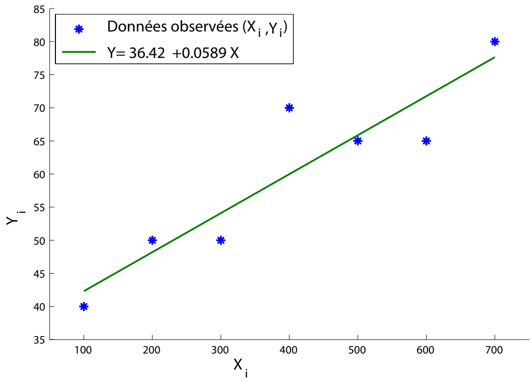

Figure 1: Présentation graphique du nuage des points (X_i, Y_i). (2p)

2. On a d'une part :

$$\hat{b} = \frac{\frac{1}{n}\sum_{i=n}^{n} x_i y_i - \overline{X}\ \overline{Y}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{X}^2} \quad (0.5\text{p}) \text{ et } \quad \hat{a} = \overline{Y} - \hat{b}\overline{X}.(0.5p) \tag{2}$$

et d'autre part :

$$\begin{aligned} \overline{X} &= \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{7}2800 = 400, \quad (0.25\text{p}) \ \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{7}420 = 60, (0.25p) \\ Cov(x,y) &= \frac{1}{n}\sum_{i=n}^{n} x_i y_i - \overline{X}\ \overline{Y} = \frac{1}{7}(184500) - (400)\ (60) = 2357.14, (0.5p) \\ Var(x) &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{X}^2 = \frac{1}{7}(1400000) - (400)^2 = 40000, (0.25p) \end{aligned}$$

ainsi,

$$\hat{b} = 0.0589 \simeq 0.06, \quad (0.5\text{p}) \text{ et } \quad \hat{a} = 36, (0.5p)$$

de ce fait, la droite de régression de la quantité consommées (Y) en fonction de prix (X) est :

$$\hat{Y} = 36 \ + \ 0.06\ x.(0.5p)$$

3. On a d'une part,

$$\rho = r(x,y) = \frac{Cov(x,y)}{\sigma_x \sigma_y}, (0.5p) \tag{3}$$

et d'autre part : $Cov(x,y) = 2357.14$, Écart-type$(x) = \sqrt{40000} = 200$(0.25p) et

Écart-type$(Y) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \overline{Y}^2} = \sqrt{\frac{1}{7}(26350) - (60)^2} = \sqrt{164.28} = 12.82$(0.5$p$), alors

$$\rho = \frac{Cov(x,y)}{\sigma_x \sigma_y} = 92\%, (0.5p) \tag{4}$$

Le fait que la valeur de $\rho \approx 1$, on déduit qu'il y a une forte liaison linéaire entre x et Y.(0.25p)

4. Afin de valider le modèle nous aurons besoin des

$$\hat{Y} = 36 \ + \ 0.06 \ x.$$

dont leurs valeurs sont rangées dans le tableau suivant :

X	100	200	300	400	500	600	700
Y	40	50	50	70	65	65	80
$\hat{Y}$	42	48	54	60	66	72	78
$e_i = y_i - \hat{y}_i$	-2	2	-6	10	-1	-7	2

On a d'une part

$$f_c = \frac{\sum_{i=1}^{n}(\hat{y} - \overline{Y})^2/1}{\sum_{i=1}^{n}(y_i - \hat{y})^2/(n-2)} = \frac{\sum_{i=1}^{n}(\hat{y} - \overline{Y})^2/1}{\sum_{i=1}^{n}e_i^2/(n-2)} = \frac{1008}{198/(7-2)} = 25.45.(1p)$$

et d'autre par

$$f_\alpha = f(1, n-2, 1-\alpha) = f(1, 5, 0.95) = 6.61, (0.25p)$$

On constate que $f_c > f_\alpha$, alors le modèle est valide (pertinent), c'est-à-dire on admet qu'on peut expliquer la quantité consommées (Y) en fonction de prix (X).(0.5p)

$$\hat{Y} = 36 \ + \ 0.06 \ x.$$

5. On a :

$$\hat{Y} = 36 \ + \ 0.06 \ x.$$

alors la quantité consommées qu'on peut prévoir au prix 800 da est: $\hat{Y} = 36 \ + \ 0.06 * 800 = 84.(0.25p)$