

Université de Abbes Laghrour-Khenchela
Département d'Informatique

EXAMEN FINAL : ANALYSE DE DONNÉES

Niveau : M1-IA

Durée suggérée : 1 h 30

Documents autorisés : Aucun

Calculatrice : Autorisée

PARTIE 1 : QCM (8 points)

Cochez la bonne réponse. Une seule réponse correcte par question. (1 point par question)

1. Quelle méthode est généralement déconseillée pour traiter les valeurs manquantes si la perte d'information est importante ?

- a) L'imputation par la moyenne.
- b) L'imputation par la médiane.
- c) La suppression pure et simple des observations incomplètes.
- d) L'utilisation d'un algorithme k-NN.

2. Dans une analyse de la variance (ANOVA), que compare-t-on pour tester l'effet d'un facteur ?

- a) Les médianes des groupes.
- b) La variance inter-groupes par rapport à la variance intra-groupe.
- c) Les écarts-types des échantillons seulement.
- d) La corrélation entre les groupes.

3. Pour étudier la liaison entre deux variables qualitatives (ex. : couleur des yeux et Couleur des cheveux), quel test ou méthode privilégie-t-on ?

- a) Le coefficient de Pearson.
- b) Le test du Khi-deux (χ^2) ou l'AFC.
- c) L'ANOVA.
- d) La régression linéaire simple.

4. Concernant les valeurs aberrantes (outliers), quelle affirmation est vraie selon le cours ?

- a) Elles doivent toujours être supprimées sans analyse.
- b) Elles ne peuvent pas être détectées par des méthodes graphiques.
- c) Elles peuvent provenir d'erreurs de saisie ou refléter une réalité rare mais correcte.
- d) Elles n'influencent jamais la moyenne.

5. Quel est l'objectif principal de l'analyse en composantes principales (ACP) ?

- a) Classer les individus en groupes homogènes.
- b) Résumer l'information d'un grand tableau de données quantitatives en un nombre réduit de variables.
- c) Tester l'égalité des moyennes.
- d) Prédire une variable qualitative.

6. Le coefficient de corrélation de Pearson (r) mesure :

- a) Une relation monotone quelconque (linéaire ou non)
- b) La force et la direction d'une relation linéaire entre deux variables quantitatives
- c) La dépendance entre deux variables qualitatives

d) Uniquement la causalité entre deux variables

7. Si le coefficient de corrélation de Spearman est proche de +1, cela signifie que :

- a) Les données suivent une loi normale parfaite.
- b) Les classements (rangs) des deux variables sont presque identiques.
- c) Il n'y a aucune relation entre les variables.
- d) Les classements sont inversés.

8. Quelle est l'hypothèse nulle (H0) testée lors d'une ANOVA à un facteur ?

- a) Toutes les variances des groupes sont différentes.
- b) Toutes les moyennes des groupes sont égales ($\mu_1 = \mu_2 = \dots = \mu_k$).
- c) Au moins deux moyennes sont différentes.
- d) Les données ne suivent pas une loi normale.

EXERCICE 1 (6 points) : Une étude vise à analyser la relation entre le budget publicitaire (X) et le chiffre d'affaires (Y) de 10 entreprises. Voici les sommes calculées à partir des données observées (n=10) :

$$\sum X = 100, \sum Y = 200, \sum X^2 = 1100, \sum Y^2 = 4400, \sum XY = 2100.$$

Questions :

- 1. Calculer les moyennes : \bar{X} et \bar{Y} . (1 pt)
- 2. Calculer la covariance $\text{Cov}(X, Y)$. (1.5 pts)
- 3. Calculer les variances et les écarts-types σ_X et σ_Y . (1.5 pts)
- 4. Calculer le coefficient de corrélation linéaire de Pearson (r). (1 pt)
- 5. Interprétez le résultat : la relation est-elle forte ou faible ? Dans quel sens ? (1 pt)

EXERCICE 2 (6 points) : Une entreprise de téléphonie mobile souhaite savoir s'il existe un lien significatif entre le type d'abonnement choisi par les clients (Standard ou Premium) et leur niveau de satisfaction (Satisfait ou Non Satisfait). Pour cela, elle interroge un échantillon de 100 clients. Voici le tableau de contingence (effectifs observés) :

	Non Satisfait	Satisfait	Total
Abonnement Standard	30	20	50
Abonnement Premium	10	40	50
Total	40	60	100

Questions :

- 1. Formulez l'hypothèse nulle (H0) et l'hypothèse alternative (H1) de ce test. (1 pt)
- 2. Calculez la valeur du χ^2 . (4 pts)
- 3. On donne la valeur critique du χ^2 pour 1 degré de liberté au seuil de 5% : 3,84. Comparez votre résultat à cette valeur critique et concluez : le type d'abonnement influence-t-il la satisfaction ?

Corrige type

PARTIE 1 : QCM (8 points)

1. Méthode déconseillée si la perte d'information est importante :
c) **La suppression pure et simple des observations incomplètes.**

2. En ANOVA, on compare :
b) **La variance inter-groupes par rapport à la variance intra-groupe.**

3. Liaison entre deux variables qualitatives :
b) **Le test du Khi-deux (χ^2) ou l'AFC.**

4. Concernant les valeurs aberrantes (outliers) :
c) **Elles peuvent provenir d'erreurs de saisie ou refléter une réalité rare mais correcte.**

5. Objectif principal de l'ACP :
b) **Résumer l'information d'un grand tableau de données quantitatives en un nombre réduit de variables.**

6. Le coefficient de corrélation de Pearson (r) mesure :
b) **La force et la direction d'une relation linéaire entre deux variables quantitatives.**

7. Si le coefficient de Spearman est proche de +1 :
b) **Les classements (rangs) des deux variables sont presque identiques.**

8. Hypothèse nulle (H_0) d'une ANOVA à un facteur :
b) **Toutes les moyennes des groupes sont égales ($\mu_1 = \mu_2 = \dots = \mu_k$).**

Correction – Exercice 1 (6 points)

Données :

$$n = 10$$

$$\sum X = 100, \sum Y = 200, \sum X^2 = 1100, \sum Y^2 = 4400, \sum XY = 2100$$

1. Calcul des moyennes (1 pt)

$$\bar{X} = \sum X / n = 100 / 10 = 10 \quad 0.5$$

$$\bar{Y} = \sum Y / n = 200 / 10 = 20 \quad 0.5$$

2. Calcul de la covariance Cov(X,Y) (1,5 pts)

$$\text{Cov}(X,Y) = (1/n)\sum XY - \bar{X}\bar{Y} \quad 0.5$$

$$\text{Cov}(X,Y) = 2100/10 - (10 \times 20) = 210 - 200 = 10 \quad 0.5$$

3. Variances et écarts-types (1,5 pts)

$$\text{Variance de X} : \sigma X^2 = (1/n)\sum X^2 - \bar{X}^2 = 110 - 100 = 10 \quad 0.5$$

$$\text{Écart-type de X} : \sigma X = \sqrt{10} \approx 3,16 \quad 0.5$$

$$\text{Variance de Y} : \sigma Y^2 = (1/n)\sum Y^2 - \bar{Y}^2 = 440 - 400 = 40 \quad 0.5$$

$$\text{Écart-type de Y} : \sigma Y = \sqrt{40} \approx 6,32 \quad 0.5$$

4. Coefficient de corrélation de Pearson (1 pt)

$$r = \text{Cov}(X,Y) / (\sigma X \sigma Y) = 10 / (3,16 \times 6,32) \approx 0,50 \quad 1$$

5. Interprétation (1 pt)

Le coefficient r est positif, ce qui indique une relation croissante. **0.5**

Sa valeur ($\approx 0,5$) montre une corrélation linéaire modérée entre le budget publicitaire et le chiffre d'affaires. **0.5**

Correction – Exercice 2 (6 points)

Données :

Tableau de contingence (effectifs observés)

Non Satisfait | Satisfait | Total

Abonnement Standard : 30 | 20 | 50

Abonnement Premium : 10 | 40 | 50

Total : 40 | 60 | 100

1. Hypothèses du test (1 pt)

H_0 : Il n'existe aucun lien entre le type d'abonnement et le niveau de satisfaction (variables indépendantes). 0.5

H_1 : Il existe un lien entre le type d'abonnement et le niveau de satisfaction (variables dépendantes). 0.5

2. Calcul de la statistique du χ^2 (4 pts)

Effectifs théoriques :

$$E(\text{Standard, Non satisfait}) = (50 \times 40) / 100 = 20 \quad 0.5$$

$$E(\text{Standard, Satisfait}) = (50 \times 60) / 100 = 30 \quad 0.5$$

$$E(\text{Premium, Non satisfait}) = (50 \times 40) / 100 = 20 \quad 0.5$$

$$E(\text{Premium, Satisfait}) = (50 \times 60) / 100 = 30 \quad 0.5$$

Calcul du χ^2 :

$$\chi^2 = \sum (O - E)^2 / E \quad 0.5$$

$$\chi^2 = (30-20)^2/20 + (20-30)^2/30 + (10-20)^2/20 + (40-30)^2/30 \quad 0.5$$

$$\chi^2 = 5 + 3,33 + 5 + 3,33 = 16,66 \quad 0.5$$

3. Décision et conclusion (1 pt)

Valeur critique au seuil de 5 % (ddl = 1) : $\chi^2 = 3,84$

χ^2 calculé = 16,66 > 3,84 0.5

Décision : rejet de H_0 . 0.5

Conclusion : le type d'abonnement influence significativement la satisfaction des clients. 0.5