Analyse et Traitement des Données (ATD) — Examen théorique
Master 1 DSC — Saint-Étienne, France

Remarques préliminaires

Examen sur feuille d'Analyse et Traitement des Données (ATD) de deux heures. Aucun document n'est autorisé. L'examen est noté sur 20 points mais il est possible d'obtenir jusqu'à 22 points, donc choisissez de répondre aux questions qui vous semblent les plus faciles. Les 3 exercices sont indépendants les uns des autres. L'examen théorique correspond à 70 % de la note finale.

Exercice 1 — Tests d'hypothèses (7 points)

Dans les années 1970, les athlètes féminines de RDA (la République Démocratique d'Allemagne, c'est-à-dire l'Allemagne de l'Est avant la chute du mur de Berlin et la réunification avec la RFA) étaient réputées pour leur forte corpulence. Le comité éthique olympique soupçonnait ces athlètes de dopage via la prise de substances hormonales virilisantes (dites hormones *androgènes*). Des mesures ont été effectuées sur la quantité d'androgènes (exprimée en nano mol) par litre de sang chez 36 athlètes, et on a obtenu les résultats suivants :

Concentration d'hormones androgènes (en nmol/l)	2,9	3,0	3,1	3,2	3,3	3,4	3,5	3,6
Nombre d'athlètes	1	3	4	7	8	6	5	2

3 pts **1.1.** Réaliser un graphique adapté permettant de voir la distribution des concentrations d'hormones androgènes des athlètes féminines est-allemandes. Est-ce que cette distribution semble suivre une loi normale ? Pour quelle raison ?

4 pts **1.2.** On veut tester l'hypothèse nulle « les athlètes de RDA ne sont pas dopées » sachant que, de manière générale, chez les femmes de 10 à 45 ans, la quantité moyenne d'androgènes est de 3,1 nmol/l, avec un écart-type de 0,3 nmol/l. Peut-on rejeter l'hypothèse nulle à un seuil α de 5 % ?

Rappel : quand on cherche à tester une moyenne μ, que la population suit une distribution normale et que la variance est connue, la statistique $\sqrt{n}\left(\frac{\overline{X}-\mu}{\sigma}\right)$ suit une loi normale réduite $\mathcal{N}(0,1)$.

Exercice 2 — Coefficient de corrélation (7 points)

Lors d'une étude sur des tas de variables enregistrées tous les ans aux États-Unis, un analyste américain du Bureau Fédéral des Statistiques s'est arrêté sur les deux phénomènes suivants : les importations américaines de pétrole brut en provenance de Norvège, d'une part, et le nombre de décès de conducteurs tués dans une collision avec un train d'autre part. Ces valeurs, enregistrées entre les années 1999 et 2009 par le Département Américain de l'Énergie et le CDC, sont données dans le tableau suivant, avec les importations américaines de pétrole brut en provenance de Norvège (en millions de barils) issu et le nombre de décès de conducteurs tués dans une collision avec un train par an :

	1999	**2000**	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**
Quantité de pétrole	96	110	103	127	60	54	43	36	20	11	
Nombre de décès	76	74	76	87	66	59	63	60	55	52	46

Questions

3 pts **2.1.** Tracez sur le même graphique, avec des échelles adaptées, l'évolution de ces deux variables (avec deux couleurs différentes) en fonction du temps.

3 pts **2.2.** Calculez le coefficient de corrélation entre ces deux variables. Est-il significatif ?

1 pt **2.3.** Que peut-on en conclure de la valeur de ce coefficient de corrélation ? Est-il possible de réaliser un modèle prédictif telle qu'une régression linéaire à partir de ces données ? Quelle pourrait être son application ?

Rappel : le coefficient de corrélation entre deux variables X et Y est $\rho_{XY} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$, avec :

- la covariance $\text{cov}(X,Y) = E(XY) - E(X)E(Y)$,
- σ_X l'écart-type de X, avec $\sigma_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$
- et $E(X)$ l'espérance (ou la moyenne $\overline{X}$) de X.

Exercice 3 — Classification ascendante hiérarchique (8 points)

Soit un jeu de données constitué par 4 points dans l'espace de représentation en 2 dimensions avec les coordonnées suivantes : $a(0, 0)$, $b(4, 3)$, $c(10, 0)$ et $d(10, 7)$.

1 pt **3.1.** Tracez un graphique avec l'ensemble des 4 points dans un repère à 2 dimensions.

2 pts **3.2.** Calculez la matrice de distances entre les différents points. On utilisera pour cela la distance euclidienne. Pour ne pas avoir trop de chiffres avec virgules, vous pouvez indiquer les distances entre les points élevées au carré.

2 pts **3.3.** Réalisez une classification hiérarchique ascendante avec ces 4 points en utilisant comme méthode d'agglomération le lien simple (saut minimal). Sur votre dendrogramme, vous indiquerez les valeurs de distance correspondant aux fusions des clusters.

2 pts **3.4.** Faites de même en utilisant cette fois-ci comme méthode d'agglomération le lien multiple (saut maximal). Sur votre dendrogramme, vous indiquerez les valeurs de distance correspondant aux fusions des clusters.

1 pt **3.5.** Comparez les deux classifications issues des réponses aux questions **3.3** et **3.4**. Que pouvez-vous en déduire sur la forme des classifications obtenues par ces deux méthodes différentes de classification hiérarchique ?