CORRIGÉ DÉTAILLÉ — TP 1 : Analyse des Données

Module : Analyse des données | M1 IA & SD | Février 2026

Question 1 — Déclaration de la matrice X et sa transposée X^t

La matrice des données X est de dimension 20 × 6 : 20 individus (villes) en lignes et 6 variables (sports) en colonnes.

Code R :

```
X <- matrix(c(
  1881.9, 96.8, 14.2, 25.2, 1135.5, 278.3,
  3369.8, 96.8, 10.8, 51.6, 1331.7, 284.0,
  # ... (20 lignes en tout)
), nrow=20, byrow=TRUE)
rownames(X) <- paste0("V", 1:20)
colnames(X) <- c("H.Ball","B.Ball","Tennis","Gym","Natation","F.Ball")
Xt <- t(X)      # Transposée
print(X)
print(Xt)
```

Résultat — Matrice X (20×6) :

Ville	H.Ball	B.Ball	Tennis	Gym	Natation	F.Ball
V1	1881.9	96.8	14.2	25.2	1135.5	278.3
V2	3369.8	96.8	10.8	51.6	1331.7	284.0
V3	4467.4	138.2	9.5	34.2	2346.1	312.3
V4	1862.1	83.2	8.8	27.6	972.6	203.4
V5	3499.8	287.0	11.5	49.4	2139.4	358.0
V6	3903.2	170.7	6.3	42.0	1935.2	292.9
V7	2620.7	129.5	4.2	16.8	1346.0	131.8
V8	3678.4	157.0	6.0	24.9	1682.6	194.2
V9	3840.5	187.9	10.2	39.6	1859.9	449.1
V10	2170.2	140.5	11.7	31.1	1351.1	256.5
V11	3920.4	128.0	7.2	25.5	1911.5	64.1
V12	2599.6	39.6	5.5	19.4	1050.8	172.5
V13	2828.5	211.3	9.9	21.8	1085.0	209.0
V14	2498.7	123.2	7.4	26.5	1086.2	153.5
V15	2685.1	41.2	2.3	10.6	812.5	89.8
V16	2739.3	100.7	6.6	22.0	1270.4	180.5
V17	1662.1	81.1	10.1	19.1	872.2	123.3
V18	2469.9	142.9	15.5	30.9	1165.5	335.5
V19	2350.7	38.7	2.4	13.5	1253.1	170.0
V20	3177.7	292.1	8.0	34.8	1400.0	358.9

La transposée X^t est de dimension 6×20 : on inverse lignes et colonnes (obtenue via t(X) en R).

Question 2 — Liste des individus

Les individus sont les 20 villes. On y accède avec :

```
rownames(X)
```

Résultat : V1, V2, V3, ..., V20

→ Chaque ligne de la matrice représente un individu statistique (une ville).

Question 3 — Extraction des variables

Les variables sont les 6 sports. On y accède avec :

```
colnames(X)
```

Résultat : "H.Ball" "B.Ball" "Tennis" "Gym" "Natation" "F.Ball"

→ Chaque colonne est une variable quantitative (taux de pratique du sport pour 100 000 jeunes).

Question 4 — Accès aux individus V3, V11, V15, V19

```
X[c(3, 11, 15, 19), ]
```

Résultat :

Ville	H.Ball	B.Ball	Tennis	Gym	Natation	F.Ball
V3	4467.4	138.2	9.5	34.2	2346.1	312.3
V11	3920.4	128.0	7.2	25.5	1911.5	64.1
V15	2685.1	41.2	2.3	10.6	812.5	89.8
V19	2350.7	38.7	2.4	13.5	1253.1	170.0

Question 5 — Proximité (Distance euclidienne) entre V3, V11, V15, V19

La distance euclidienne entre deux individus i et j est :

$d(i,j) = \sqrt{[\ \Sigma(xik - xjk)^2\ ]}$ pour k = 1,...,6

Code R :

```
sel <- X[c(3,11,15,19),]
dist(sel)    # distance euclidienne par défaut
```

Paire	Distance euclidienne	Interprétation
V3 — V11	741.5352	Assez proches (même profil sportif élevé)
V3 — V15	2363.9049	Très éloignées (profils très différents)
V3 — V19	2388.6599	Très éloignées (profils très différents)
V11 — V15	1655.9613	Très éloignées
V11 — V19	1707.8656	Très éloignées
V15 — V19	558.9261	Les plus proches de ce sous-groupe

Commentaire :

• V3 et V11 sont les plus proches (d ≈ 741.5) : elles ont des taux similaires pour plusieurs sports.

• V15 et V19 sont aussi relativement proches (d ≈ 559.0) : toutes deux ont des taux modestes.

• V3 est très éloignée de V15 et V19 (d > 2300) car V3 a des taux extrêmement élevés (H.Ball = 4467, Natation = 2346).

• Les grandes distances sont principalement dues à l'échelle dominante du Handball et de la Natation.

Question 6 — Tableau statistique X(j) : Moyenne, Variance, Écart-type

IMPORTANT : On utilise la variance biaisée (division par m=20, car ce sont des données de population) :

$var(j) = (1/m) \times \Sigma(xij - \bar{x}j)^2$

```
var_pop <- function(v) sum((v - mean(v))^2) / length(v)
sd_pop  <- function(v) sqrt(var_pop(v))
tab <- data.frame(
  Moyenne    = round(apply(X, 2, mean), 4),
  Variance   = round(apply(X, 2, var_pop), 4),
  Ecart_type = round(apply(X, 2, sd_pop), 4)
)
print(tab)
```

Résultat :

Variable	Moyenne	Variance	Écart-type
H.Ball	2911.3000	586087.2800	765.5634
B.Ball	134.3200	4766.5926	69.0405
Tennis	8.4050	11.7665	3.4302
Gym	28.3250	116.9119	10.8126
Natation	1400.3650	179029.5633	423.1189
F.Ball	230.8800	9732.2776	98.6523

Interprétation :

• Le H.Ball a la moyenne la plus élevée (2911.3) avec une variance très grande (586087.3) → forte dispersion entre villes.

• Le Tennis a la moyenne la plus faible (8.4) et la variance la plus faible (11.77) → sport peu pratiqué et homogène.

• La Natation présente aussi une grande variance (179029.6), reflétant des inégalités importantes entre villes.

Question 7 — Individu moyen

L'individu moyen (ou profil moyen) est le vecteur des moyennes de chaque variable :

$\bar{x} = (1/m) \times \Sigma\ xij$ pour chaque variable j

```
ind_moy <- round(apply(X, 2, mean), 4)
print(ind_moy)
```

Résultat :

H.Ball	B.Ball	Tennis	Gym	Natation	F.Ball
2911.3000	134.3200	8.4050	28.3250	1400.3650	230.8800

→ L'individu moyen représente une « ville fictive » ayant le profil moyen de toutes les villes. Il sert de référence pour l'analyse.

Question 8 — Matrice centrée Y

La matrice centrée Y est obtenue en soustrayant la moyenne de chaque variable :

$y_{ij} = x_{ij} - \bar{x}_j$

```
Y <- sweep(X, 2, apply(X, 2, mean))
# ou : Y <- scale(X, scale=FALSE)
print(round(Y, 4))
```

Résultat (5 premières lignes) :

Ville	H.Ball	B.Ball	Tennis	Gym	Natation	F.Ball
V1	-1029.4000	-37.5200	5.7950	-3.1250	-264.8650	47.4200
V2	458.5000	-37.5200	2.3950	23.2750	-68.6650	53.1200
V3	1556.1000	3.8800	1.0950	5.8750	945.7350	81.4200
V4	-1049.2000	-51.1200	0.3950	-0.7250	-427.7650	-27.4800
V5	588.5000	152.6800	3.0950	21.0750	739.0350	127.1200

→ Une valeur positive indique que la ville est au-dessus de la moyenne pour ce sport. Une valeur négative, en dessous.

→ La somme de chaque colonne de Y est nulle (propriété fondamentale).

Question 9 — Fonction de calcul des variances

On écrit une fonction en R qui calcule la variance biaisée des 6 variables :

```
calc_variance <- function(mat) {
  m <- nrow(mat)
  variances <- apply(mat, 2, function(v) {
    round(sum((v - mean(v))^2) / m, 4)
  })
  return(variances)
}
# Appel de la fonction :
calc_variance(X)
```

Résultat :

H.Ball	B.Ball	Tennis	Gym	Natation	F.Ball
586087.2800	4766.5926	11.7665	116.9119	179029.5633	9732.2776

Question 10 — Matrice de covariance V

La matrice de covariance V est calculée selon la formule :

$V = (1/m) \times Y^t \cdot Y$

Où Y est la matrice centrée (Question 8). La diagonale de V contient les variances des variables.

```
m <- nrow(X)
Y <- sweep(X, 2, apply(X, 2, mean))
V <- (1/m) * t(Y) %*% Y
print(round(V, 4))
```

Résultat — Matrice V (6×6) :

	H.Ball	B.Ball	Tennis	Gym	Natation	F.Ball
H.Ball	**586087**	**23879**	**-278.00**	**4282**	**280230**	**26625**
B.Ball	23879	4767	83.71	440.74	15483	4217
Tennis	-278.00	83.71	11.77	19.82	146.57	198.28
Gym	4282	440.74	19.82	116.91	2734	800.07
Natation	280230	15483	146.57	2734	179030	19537
F.Ball	26625	4217	198.28	800.07	19537	9732

Question 11 — Commentaire de la matrice V

La matrice V est symétrique, semi-définie positive. Ses éléments sont les covariances Cov(Xi, Xj).

Éléments diagonaux — Variances :

• H.Ball : 586 087 → très grande variance (forte dispersion entre villes)

• Natation : 179 030 → également très dispersée

• Tennis : 11.77 → très faible variance (homogène entre villes)

Éléments hors-diagonale — Covariances :

• Cov(H.Ball, Natation) ≈ 280 230 → très grande covariance positive : les villes qui pratiquent beaucoup le handball pratiquent aussi beaucoup la natation.

• Cov(H.Ball, Tennis) ≈ −278 → légère covariance négative : faible relation inverse.

• Attention : les covariances sont difficiles à interpréter directement à cause des unités différentes. La matrice de corrélation R est plus lisible.

Question 12 — Matrice de corrélation R

La matrice de corrélation R normalise les données. On divise la matrice centrée Y par les écarts-types :

$Z = Y / \sigma$ puis $R = (1/m) \times Z^t \cdot Z$

Les coefficients $r_{ij} \in [-1, 1]$ mesurent l'intensité et le sens de la relation linéaire entre deux variables.

```
sd_vec <- apply(X, 2, sd_pop)    # écarts-types
Z <- sweep(Y, 2, sd_vec, "/")    # standardisation
R <- (1/m) * t(Z) %*% Z
print(round(R, 4))
```

Résultat — Matrice R (6×6) :

	H.Ball	B.Ball	Tennis	Gym	Natation	F.Ball
H.Ball	1.0000	0.4518	-0.1059	0.5173	0.8651	0.3525
B.Ball	0.4518	1.0000	0.3535	0.5904	0.5300	0.6192
Tennis	-0.1059	0.3535	1.0000	0.5344	0.1010	0.5859
Gym	0.5173	0.5904	0.5344	1.0000	0.5977	0.7501
Natation	0.8651	0.5300	0.1010	0.5977	1.0000	0.4680
F.Ball	0.3525	0.6192	0.5859	0.7501	0.4680	1.0000

Question 13 — Commentaire de la matrice R

La diagonale est toujours 1 (un sport est parfaitement corrélé avec lui-même).

Corrélations fortes positives (r > 0.7) :

• r(H.Ball, Natation) = 0.8651 → Forte corrélation : les villes sportives en handball le sont aussi en natation. Ces deux sports partagent un contexte socio-économique similaire.

• r(Gym, F.Ball) = 0.7501 → Bonne corrélation entre la gymnastique et le football.

Corrélations modérées (0.4 < r < 0.7) :

• r(B.Ball, F.Ball) = 0.6192 ; r(H.Ball, Gym) = 0.5173 ; r(Natation, Gym) = 0.5977

Corrélations faibles ou négatives :

• r(H.Ball, Tennis) = −0.1059 → Très faible corrélation négative : pas de lien significatif entre handball et tennis.

• r(Tennis, Natation) = 0.1010 → Quasi-indépendance.

Conclusion : Le handball et la natation sont les variables les plus liées entre elles et les plus structurantes. Le tennis est la variable la plus indépendante des autres.

Question 14 — Représentation graphique : Nuages de points

On représente les 20 individus (villes) dans le plan de 3 couples de variables.

Code R :

```
par(mfrow = c(1, 3))  # 3 sous-fenêtres côte à côte

# Graphe 1 : X1 (H.Ball) vs X4 (Gym)
plot(X[,1], X[,4], xlab="H.Ball", ylab="Gym",
     main="H.Ball vs Gym", pch=16, col="blue")
text(X[,1], X[,4], rownames(X), pos=3, cex=0.7)

# Graphe 2 : X2 (B.Ball) vs X5 (Natation)
plot(X[,2], X[,5], xlab="B.Ball", ylab="Natation",
     main="B.Ball vs Natation", pch=16, col="red")
text(X[,2], X[,5], rownames(X), pos=3, cex=0.7)

# Graphe 3 : X3 (Tennis) vs X6 (F.Ball)
plot(X[,3], X[,6], xlab="Tennis", ylab="F.Ball",
     main="Tennis vs F.Ball", pch=16, col="darkgreen")
text(X[,3], X[,6], rownames(X), pos=3, cex=0.7)
par(mfrow = c(1, 1))  # Réinitialiser
```

Commentaire des graphes :

Graphe 1 — (H.Ball, Gym) : corrélation = 0.5173

Le nuage de points montre une tendance linéaire croissante modérée. Les villes avec des taux élevés en handball tendent à avoir des taux plus élevés en gymnastique. V3 (H.Ball=4467, Gym=34.2) et V5 se distinguent comme valeurs extrêmes.

Graphe 2 — (B.Ball, Natation) : corrélation = 0.5300

Corrélation modérée positive. Le nuage est plus dispersé. V20 (B.Ball=292.1) et V5 (B.Ball=287.0) sont des points atypiques (outliers) sur l'axe basket.

Graphe 3 — (Tennis, F.Ball) : corrélation = 0.5859

Le nuage montre une dispersion modérée avec une tendance croissante. V18 (Tennis=15.5, F.Ball=335.5) et V9 (F.Ball=449.1) sont des points remarquables.

Remarque générale : Aucun des trois nuages ne montre une dispersion aléatoire pure, ce qui confirme l'existence de corrélations entre ces paires de variables. Plus le nuage est « allongé », plus la corrélation est forte.