CY TECH	ING1-GM RATTRAPAGE DE DATA EXPLORATION 2022-2023
Durée : 2h00	Examen papier 2 feuilles R/V manuscrites autorisées Calculatrice autorisée

Exercice 1 : 6 points

On étudie le jeu de données bodylight.csv constitué 5 variables quantitatives et 20 observations à l'aide d'une analyse en composantes principales dont les résultats se trouvent ci-dessous.

0.5 1) Quelle est la dimension de l'espace de représentation du nuage de points ? 5 dimensions

1 2) Quelle est l'inertie du nuage de point ? 3.65+0.783+.... =5

chaque valeur propre divisé par l'inertie totale qui est 5

0.5 3) Comment sont calculés les pourcentages d'inertie expliqué sur chaque axe (composante principale) ?

1 4) Comment sont calculés les nouveaux axes (composantes principales) ? Ci=X' Ui, X' matrice de donnes centree et reduite, et Ui le iéme vecteur propore.

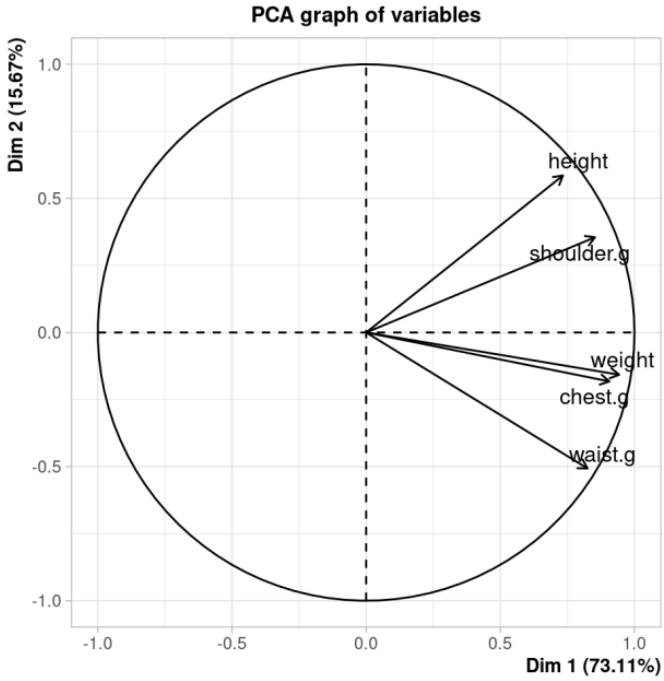
1 5) Combien d'axes faut-il retenir pour avoir une bonne représentation ? Les deux premiers fournissent 88.7% des informations donc suffisants

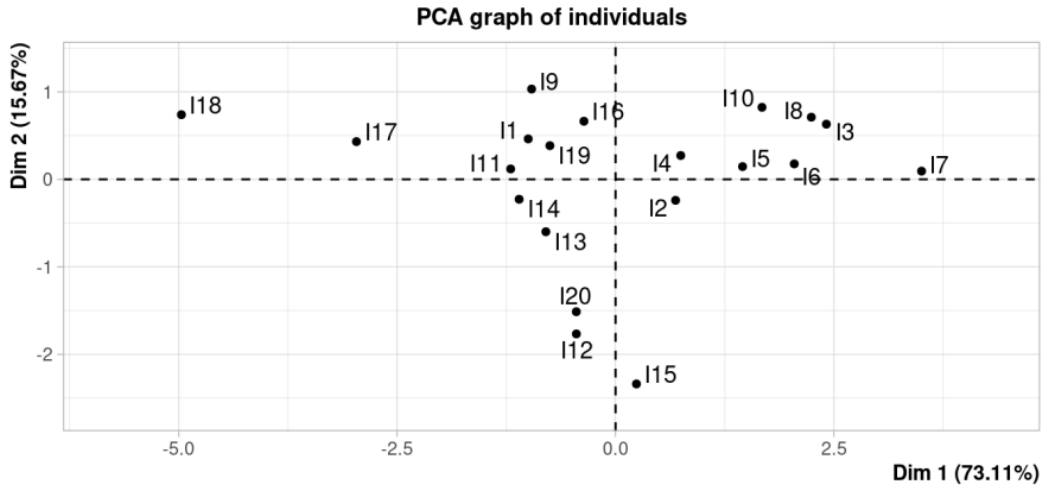
1 6) Que pouvez-vous dire du lien entre les variables « height » et « waist.g » ? correlé positivement

1 7) Comment pouvez-vous caractériser l'individu « I7 » ? il a l'absice le plus élevé donc elle a pris les valeurs les plus elevées pour les variables.

Justifiez chacune de vos réponses.

```
##        eigenvalue percentage of variance cumulative percentage of variance
## comp 1 3.65552473              73.110495                           73.11049
## comp 2 0.78341795              15.668359                           88.77885
## comp 3 0.33778275               6.755655                           95.53451
## comp 4 0.15560651               3.112130                           98.64664
## comp 5 0.06766806               1.353361                          100.00000
```





Exercice 2 : (14 points)

Le fichier JeuxVideo.csv sur Arel décrit 2066 jeux vidéo suivant plusieurs variables. Pour les questions suivantes, on retient les variables :

- Plateforme : MICROSOFT, NINTENDO, PC et SONY
- Genre : Action, Racing, Role-Playing, Shooter et Sports
- Taux des ventes en Europe : [0 ;1]

Partie 1 : Etude de l'impact de la plateforme sur le taux des ventes en Europe. 3 points

1) A partir du graphique suivant répondez aux questions suivantes :

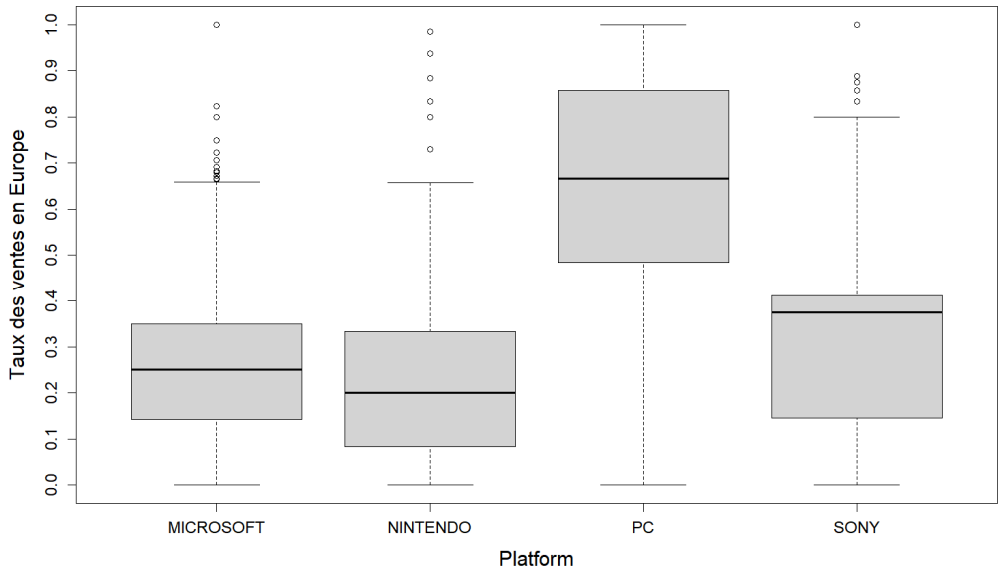
0.5 a) Quel est le taux médian des ventes en Europe pour la plateforme NINTENDO ? 0.2

0.5 b) Pour quelle(s) plateforme(s) 25% des jeux ont un taux de vente en Europe supérieur 40% ?

0.5 c) Le taux moyen des ventes en Europe pour la plateforme NINTENDO est-il à peu-près égal au taux moyen ? Justifiez.

0.5 d) Que pouvez-vous dire sur le lien entre les deux variables ? Liees

0.5 + 0.5 2) Quel indicateur numérique permet de mesurer le lien entre les variables ? Quelle est la formule qui permet de le calculer ? rapport de correlation



Partie 2 : Etude de l'impact de la plateforme sur le genre. (2.5 pts)

1) Lecture des tableaux

0.5 a) Quel est le pourcentage de jeux de sport sur plateforme SONY ? 13%

0.5 b) Quel est le pourcentage de plateformes SONY ? 41%

0.5 c) Quel est le pourcentage de plateformes SONY pour les jeux de sports ? 33%

0.5 d) Quel est le pourcentage de jeux de sport ? 30%

0.5 e) Quel pourcentage de jeux de sport SONY développe-t-il ?

2) Analyse des profils lignes et colonnes (3 points)

0.5 a) Dans le tableau des profils lignes à quoi correspond la dernière ligne « Profil moyen » ?

1 b) Quelles sont les plateformes qui s'éloignent le plus du profil moyen ? Donnez des exemples et commentez. PC

1 c) Quels sont les genres qui s'éloignent le plus du profil moyen ? Donnez des exemples et commentez. Role playing et shooter

0.5 d) Pensez-vous que les deux variables sont liées ? Oui

	Action	Racing	Role-Playing	Shooter	Sports	Sum
MICROSOFT	207	56	39	138	198	638
NINTENDO	133	46	18	48	131	376
PC	82	19	30	59	14	204
SONY	249	107	84	131	277	848
Sum	671	228	171	376	620	2066

Tableau de contingence

	Action	Racing	Role-Playing	Shooter	Sports	Sum
MICROSOFT	0.10	0.03	0.02	0.07	0.10	0.31
NINTENDO	0.06	0.02	0.01	0.02	0.06	0.18
PC	0.04	0.01	0.01	0.03	0.01	0.10
SONY	0.12	0.05	0.04	0.06	0.13	0.41
Sum	0.32	0.11	0.08	0.18	0.30	1.00

Tableau des fréquences

	Action	Racing	Role-Playing	Shooter	Sports
MICROSOFT	0.32	0.09	0.06	0.22	0.31
NINTENDO	0.35	0.12	0.05	0.13	0.35
PC	0.40	0.09	0.15	0.29	0.07
SONY	0.29	0.13	0.10	0.15	0.33
Profil moyen	0.32	0.11	0.08	0.18	0.30

Tableau des profils lignes

	Action	Racing	Role-Playing	Shooter	Sports	Profil moyen
MICROSOFT	0.31	0.25	0.23	0.37	0.32	0.31
NINTENDO	0.20	0.20	0.11	0.13	0.21	0.18
PC	0.12	0.08	0.18	0.16	0.02	0.10
SONY	0.37	0.47	0.49	0.35	0.45	0.41

Tableau des profils colonnes

3) Afin de déterminer numériquement si les variables sont liées, on calcule la distance du chi-deux. (3 points)

0.5 a) Rappelez à quoi correspond le tableau des effectifs théoriques. Effectif dans le cas d'independance

1 b) Expliquez le calcul qui permet d'obtenir l'effectif théorique des jeux d'action sur plateforme Microsoft. (671*638)/2066

	Action	Racing	Role-Playing	Shooter	Sports
MICROSOFT	207	70	53	116	191
NINTENDO	122	41	31	68	113
PC	66	23	17	37	61
SONY	275	94	70	154	254

Tableau des effectifs théoriques

1 c) Dans la sortie R ci-dessous, expliquez à quoi correspondent «X-squared » et « df » et comment ils sont calculés (on ne demande pas de faire les calculs). X-squared=(tij-nij)²/tij et df=(5-1)*(4-1)=12

```
        Pearson's Chi-squared test

data:  cont
X-squared = 103, df = 12, p-value <2e-16
```

0.5 d) Expliquez pourquoi on peut conclure que les deux variables sont liées. p-value < 0.05... donc variables liées

4) Afin de faire une analyse plus fine du lien entre les deux variables, on met en place une AFC. (2.5 pts)

1 a) Pourquoi n'y a-t-il que trois valeurs propres ? p=5, q=4, donc dimension de AFC est 4-1=3. donc 3 valeurs propores

0.5 b) A quoi correspond le pourcentage dans la colonne du milieu du tableau des valeurs propres ? % de khi deux explique par le deuxieme axe

1 c) Sur quel(s) axe(s) allez-vous analyser les résultats de l'AFC ? Justifiez. Les deux premiers axes, ils expliquent tous les deux 93.5% de la quantité du khi deux.

d) Interprétez le graphique de l'AFC.

Histogramme des valeurs propres

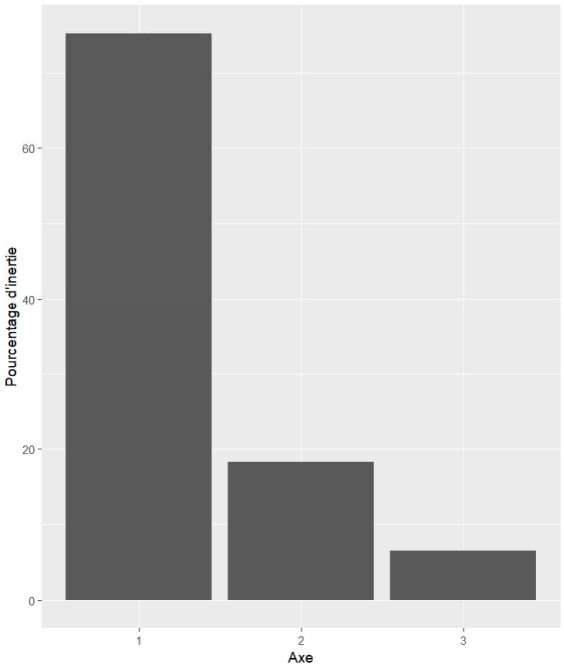


Tableau des valeurs propres

Axe	%	Cum. %
1	75.2	75.2
2	18.3	93.5
3	6.5	100.0