

**ANALYSE DES DONNÉES**

Fiche de TD n°4

Une étude a été menée sur 592 femmes dans le but d'analyser une éventuelle liaison entre les variables "couleur des yeux" et "couleur des cheveux". Dans la suite, on désignera par  $X$  la variable "couleur des yeux" et par  $Y$  la variable "couleur des cheveux". Les résultats de l'étude sont contenus dans le tableau suivant :

X \ Y	Brun	Châtain	Roux	Blond
Marron	68	119	26	7
Noisette	15	54	14	10
Vert	5	29	14	16
Bleu	20	84	17	94

1. Le tableau suivant donne la distribution bivariée en pourcentage, ainsi que les distributions marginales elles-mêmes exprimées en pourcentage. Compléter alors ce tableau.

	Brun	Châtain	Roux	Blond	$f_i$
Marron	11.49	20.11	4.39	1.18	.
Noisette	.	9.12	2.36	1.69	15.70
Vert	0.84	4.90	2.36	2.70	10.80
Bleu	3.38	14.19	2.87	15.89	36.33
$f_{.j}$	.	48.32	11.98	21.46	.

2. Le tableau suivant donne les profils-lignes. Compléter ce tableau.

%	Brun	Châtain	Roux	Blond	
Marron	.	54.11	11.81	3.17	.
Noisette	16.11	.	15.03	10.76	.
Vert	7.78	45.37	21.85	25.00	.
Bleu	9.30	.	7.90	43.74	.
$G_\ell$	.	.	.	.	100

3. Indiquer dans quel espace sera représenté le nuage des profils-lignes. Ces profils vont en fait appartenir à un sous-espace. Préciser la dimension de ce dernier, puis indiquer la raison de cette propriété.

4. Le tableau suivant donne les profils-colonnes. Compléter ce tableau.

%	Brun	Châtain	Roux	Blond	$G_c$
Marron	62.99	41.62	36.64	5.50	.
Noisette	13.87	.	19.70	7.88	.
Vert	4.60	10.14	19.70	12.58	.
Bleu	.	29.37	23.96	.	.
	.	.	.	.	

5. Indiquer dans quel espace sera représenté le nuage des profils-colonnes. Ces profils vont en fait appartenir à un sous-espace. Préciser la dimension de ce dernier, puis indiquer la raison de cette propriété.
6. Pour chacun des nuages de profils, préciser la matrice des pondérations (i.e.  $D_n$  et  $D_p$ ), ainsi que les matrices associées aux métriques (i.e.  $M_p$  et  $M_n$ ).
7. Rappeler la formule matricielle permettant de calculer la matrice à diagonaliser, notée  $S$ , pour réaliser l'ajustement du nuage des profils-lignes. Pourquoi est-il inutile de chercher à écrire, puis à diagonaliser la matrice associée aux profils-colonnes ?
8. L'opération de diagonalisation de la matrice  $S$  par un logiciel de mathématiques a permis d'obtenir les valeurs propres suivantes :

$$\lambda_0 = 1 \quad \lambda_1 = 0.20877 \quad \lambda_2 = 0.02223 \quad \lambda_3 = 0.00260$$

Que pouvez-vous dire sur la valeur propre  $\lambda_0$  ? Sur quelles valeurs propres allez-vous concentrer votre analyse ?

9. Déterminer l'inertie associée à chacun des axes factoriels non-triviaux. En déduire le pourcentage d'inertie associé à chacun des axes, puis l'inertie cumulée.
10. Sachant que l'on veut conserver au moins 90% de l'inertie, indiquer le nombre d'axes factoriels que l'on doit sélectionner.
11. Les deux premiers vecteurs propres issus de la diagonalisation de la matrice  $S$  sont respectivement

$$u_1 = (-0.201436, -0.156799, -0.033982, 0.392135)$$

$$u_2 = (-0.262634, 0.105975, 0.256799, -0.100131)$$

Après avoir rappelé la formule permettant d'obtenir les coordonnées des profils-lignes dans la nouvelle base, compléter les matrices colonnes suivantes :

$$\Psi_1 = \begin{pmatrix} . \\ -0.2125 \\ 0.1617 \\ 0.5474 \end{pmatrix} \quad \text{et} \quad \Psi_2 = \begin{pmatrix} -0.0883 \\ 0.1673 \\ . \\ -0.0831 \end{pmatrix}$$

12. Rappeler la formule permettant d'obtenir la contribution d'un profil-ligne à la formation d'un axe. Compléter alors le tableau suivant :

Profils-lignes	$Cr_1(i)$	$Cr_2(i)$
1	.	13.04
2	3.40	19.79
3	1.35	.
4	52.14	11.26

13. Rappeler la formule permettant d'obtenir la qualité de représentation d'un profil-ligne sur un axe. Comment obtient-on alors facilement la qualité de représentation d'un profil-ligne sur le plan factoriel principal. Compléter alors le tableau suivant :

Profils-lignes	$Qual_1(i)$	$Qual_2(i)$	$Qual_{1 \times 2}(i)$
1	96.70	2.78	.
2	.	33.58	87.82
3	17.59	76.77	94.36
4	97.75	2.25	.

14. Indiquer la formule quasi-barycentrique permettant de calculer les coordonnées des profils-colonnes à partir des coordonnées des profils-lignes. Compléter alors les matrices colonnes suivantes :

$$\varphi_1 = \begin{pmatrix} . \\ -0.1483 \\ -0.1296 \\ 0.8349 \end{pmatrix} \quad \text{et} \quad \varphi_2 = \begin{pmatrix} -0.2148 \\ 0.0327 \\ . \\ -0.0696 \end{pmatrix}$$

15. Rappeler la formule permettant d'obtenir la contribution d'un profil-colonne à la formation d'un axe. Compléter alors le tableau suivant :

Profils-colonnes	$Cr_1(j)$	$Cr_2(j)$
1	22.25	.
2	5.09	2.32
3	0.96	55.13
4	.	4.67

16. Rappeler la formule permettant d'obtenir la qualité de représentation d'un profil-colonne sur un axe. Comment obtient-on alors facilement la qualité de représentation d'un profil-colonne sur le plan factoriel principal. Compléter alors le tableau suivant :

Profils-colonnes	$Qual_1(i)$	$Qual_2(i)$	$Qual_{1 \times 2}(i)$
1	.	15.12	.
2	86.59	4.21	90.80
3	13.36	.	94.62
4	99.14	0.69	.

17. Afin de proposer une interprétation contextuelle des axes, et en particulier le premier qui contient une part importante d'information, une technique consiste à sélectionner les profils dont la contribution est supérieure à leur pondération, puis à les ranger selon le signe de leur coordonnée sur l'axe considéré. Réaliser ce travail, puis synthétiser cette information à l'intérieur d'un tableau. En déduire une interprétation de chacun des facteurs (si possible!).
18. En faisant usage des relations quasi-barycentriques :
- représenter le nuage des profils-lignes  $\sqrt{\lambda_\alpha} \Psi_\alpha$  **en fonction** des profils-colonnes, puis déduire, si possible, les profils-colonnes qui attirent significativement les profils-lignes.
  - représenter le nuage des profils-colonnes  $\sqrt{\lambda_\alpha} \varphi_\alpha$  **en fonction** des profils-lignes, puis déduire, si possible, les profils-lignes qui attirent significativement les profils-colonnes.
19. Fournir une représentation simultanée des deux nuages. Vérifier que votre graphique est semblable à celui de la Figure 1.

20. Retrouve-t-on sur ce dernier les commentaires précédents? Justifier votre réponse.
21. Observe-t-on une indépendance entre les deux variables? Justifier votre réponse.
22. Calculer l'inertie totale des nuages. Que représente le produit de cette inertie totale par l'effectif total?
23. Réaliser un test d'indépendance du khi-deux au niveau  $\alpha = 5\%$ . Que décidez-vous? ( $c_{0,05} = 16.92$ ).
24. Analyser le nuage des profils-lignes.
25. Analyser le nuage des profils-colonnes.
26. Calculer les coefficients de distorsion  $1/\sqrt{\lambda_1}$  et  $1/\sqrt{\lambda_2}$ . Interpréter géométriquement leur rôle dans le cadre des relations quasi-barycentriques.
27. Analyser simultanément les deux nuages de profils selon le premier axe factoriel.
28. Vérifier vos commentaires à l'aide des écarts relatifs entre les profils et leur barycentre.

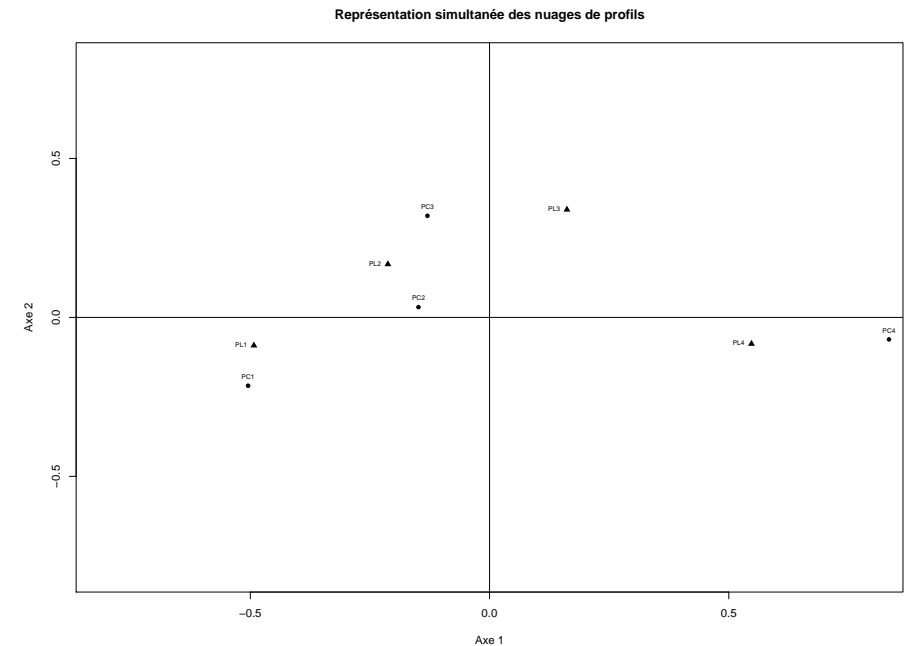


Fig. 1 - Représentation simultanée