

A - Statistiques Univariées (Total = $3\frac{1}{2}$)

1. $Q_1 = 3,93$; 25% des avocats ont un prix inf ou égal à 3,93 \$ le kilo.

$Q_2 = 4,58$; 50% " " (1) "

$Q_3 = 5,2$; 75% " " " "

2° - Etendue = $Max - Min = 7,76 - 2,6 = 5,16$

$$IQ = Q_3 - Q_1 = 5,2 - 3,93 = 1,27 \quad \left(\frac{1}{2}\right)$$

3° -

$$\begin{aligned} m &= Q_1 - 1,5 (Q_3 - Q_1) \\ &= 3,93 - 1,5 (5,2 - 3,93) \\ &= 2,025 \end{aligned} \quad (1)$$

$$\begin{aligned} M &= Q_3 + 1,5 (Q_3 - Q_1) \\ &= 5,2 + 1,5 (5,2 - 3,93) \\ &= 7,105 \end{aligned} \quad (1)$$

4° Oui, il ya des valeurs atypiques (au moins une valeur); car $M = 7,105$ alors que le maximum est $7,76$. $\left(\frac{1}{2}\right)$

5° Moyenne = $4,723$ \$

Mediane = $4,58$ \$ $\left(\frac{1}{2}\right)$

La moyenne donne une idée sur le prix dans toute la population, alors que la médiane sur 50% de la population

B - Analyse Bivariables. Total = $5 + 2\left(\frac{3}{4}\right) + 4\frac{3}{4} = 12\frac{1}{2}$

① 1° $r_{xy} = \frac{\text{cov}(x,y)}{S_x \cdot S_y} = \frac{324,6}{\sqrt{1,151} \times \sqrt{423428}} = 0,465$ $\left(\frac{1}{2}\right)$

les deux variables sont faiblement liées de même sens.

2° $y = ax + b$ $\left\{ \begin{array}{l} \hat{a} = \frac{C_{xy}}{S_x^2} = \frac{324,6}{1,151} = 282,02 \\ \hat{b} = \bar{y} - \hat{a}\bar{x} = 522,7787 - (282,02 \times 4,723) = -809,2 \end{array} \right.$ $\left(\frac{1}{2}\right)$

$\Rightarrow \boxed{y = 282,02 X - 809,2}$

3° - $X = 3,92$; $Y = 27,357$ tonnes

a) $\hat{Y} = ??$

$$\hat{Y} = 282,02 (3,92) - 809,2 = 296,32 \quad \left(\frac{1}{2}\right)$$

b) erreur = $Y - \hat{Y} = 27,357 - 296,32 = -268,963$ $\left(\frac{1}{2}\right)$

4° $R^2 = (r_{xy})^2 = (0,465)^2 = 0,216$ $\textcircled{1}$

21,6 % de la variabilité de vente est expliquée par la prise.

5° Les valeurs aberrantes (atypiques) ont un résidu centré qui n'est pas dans l'intervalle $[-2,2]$. $\left(\frac{1}{2}\right)$

On remarque que $\text{Max} = 2,572$ qui n'est pas dans $[-2,2]$, donc il y a au moins une valeur atypique.

6° On remarque que le coefficient de détermination est loin de 1; et le coeff de corrélation est de 0,465 $\left(\frac{1}{2}\right)$
ce qui implique que le lien entre ces deux variables n'est pas assez fort pour qu'on estime Y en fonction de X . On remarque aussi qu'il y a des valeurs atypiques ce qui peut influencer sur la construction de ce modèle. $\textcircled{2}$

②

1° En comparant les extrémités de 4 boeuflets, on peut remarquer qu'ils ont la même longueur (on constate), donc les distributions sont presque les mêmes dans les 3 premières régions mais un peu différent de 4^e $\Rightarrow$ on remarque qu'il y a pas de lien entre le prix et les régions

2° $V_{\text{inter}} = \frac{2,383}{130} = 0,0183$

$$V_{\text{intra}} = \frac{126}{130} = 0,969$$

$$V_{\text{total}} = V_{\text{inter}} + V_{\text{intra}}$$

$$= 0,0183 + 0,969 = 0,987$$

3° $\pi = \frac{V_{\text{inter}}}{V_{\text{total}}} = \frac{0,0183}{0,987} = 0,0185$

1,85% de la variabilité des prix est expliquée par les régions.

③

	Center	North	South	West	Total
Conventi.	8	20	22	34	84
Organic	4	4	7	31	46
Total	12	24	29	65	130

1° - tableau (Total). $\frac{1}{2}$

2° - $f_{12} = \frac{n_{12}}{n} = \frac{20}{130} = 0,154$. ①

15,4% des avocats conventionnel sont vendu en region "North".

$$f_{.1} = \frac{n_{.1}}{n} = \frac{12}{130} = 0,092$$

9,2% des avocats sont vendu en Region "Center".

$$f_{2/1} \rightarrow f_{i=2/j=1} = \frac{n_{21}}{n_{.1}} = \frac{4}{12} = 0,33$$

ou

$$\rightarrow f_{j=2/i=1} = \frac{n_{12}}{n_{1.}} = \frac{20}{84} = 0,238$$

③

3° - profil Moyenne ligne:

$$\left(\frac{12}{130}, \frac{24}{130}, \frac{29}{130}, \frac{65}{130} \right)$$

$\frac{1}{2}$

$$= (0,092; 0,185; 0,223; 0,5).$$

4° - Tableau profil ligne:

a)

	Center	North	South	West	total
Convo	0,095	0,24	0,262	0,404	1
Organic	0,087	0,087	0,152	0,674	1

$\frac{1}{2}$

b) La répartition des avocats dans la sous-population "Organic" s'éloigne beaucoup de la répartition dans toute la populat° (profil moyen ligne); alors que cette répartition est proche de profil moyen ligne pour la sous-population "Organic".

c) La répartition dans la populat° et les sous populations lignes "Conventionnel" et "Organic" n'est pas la même, alors, on peut remarquer que les variables sont liées entre elles!

5° - a) Test d'indépendance de Khi-deux $\left(\frac{1}{4}\right)$

b) Tableau des eff. théoriques: $\left(\frac{1}{2}\right)$

	Center	North	South	West	Total
Conven	7,75	15,50	18,74	42	84
Organic	4,25	8,49	10,26	23	46
Total	12	24	29	65	130

c) $\chi^2 = 9,6106$

d.d.l. = $(4-1)(2-1) = 3$

Seuil = 7,82.

$\chi^2 > \text{Seuil} \Rightarrow$ les variables sont liées.

d) Toutes les eff théoriques sont ≥ 5 ; donc le test d'indépendance de Khi-deux est efficace.

C - Analyse Multivariée (ACP) (Total = $4\frac{1}{2}$)

1° - Il ya 130 points. Les points sont représentés en dimension 7.

2° - La quantité d'information est l'inertie qui mesure la dispersion totale du nuage des points; c'est la somme

des distances au carré entre les observations et le centre de gravité du nuage.

- On peut la calculer à partir de la matrice de var-cov ou la matrice de Corrélat°.

$$I = \sum_{i=1}^p S_{X_i}^2 \quad (X_i \text{ les variables étudiées})$$

$$= p \quad (\text{si les variables sont normalisées}).$$

3° $\lambda_3 = ??$

$$I = p = 7$$

①

on sait que: $\frac{\lambda_3}{I} \times 100 = 10,354 \text{ (\% de variance)}$

$$\Rightarrow \frac{\lambda_3}{7} \times 100 = 10,354$$

$$\Rightarrow \boxed{\lambda_3 = 0,725}$$

$\lambda_4 = ??$

$$\lambda_4 = I - (\lambda_1 + \lambda_2 + \lambda_3 + \lambda_5 + \lambda_6 + \lambda_7)$$

$$\Rightarrow \boxed{\lambda_4 = 0,2588}$$

4° les deux premières composantes principales contiennent 83,70% de l'inertie totale.

②

5° - poids moyen:

$$\left\{ \begin{array}{l} \text{variable} = \frac{1}{7} = 0,143 \\ \text{Individu} = \frac{1}{130} = 7,69 \times 10^{-3} = 0,00769 \end{array} \right.$$

$\frac{1}{2}$

6° - On peut remarquer que toutes les variables sont bien représentées en Dim2 ; sauf la variable "Price".

$$\cos^2_1(\text{Price}) + \cos^2_2(\text{Price}) = 0,2871 + 0,098 = 0,3851$$

très loin de 1 ; Contrairement à toutes les autres variables qui ont $\cos^2_1 + \cos^2_2$ proche de 1.

$\frac{1}{2}$

(où aussi car cette variable est représentée par un point proche de l'origine dans le cercle de corrélation).

7° - Cercle de corrélation :

$\frac{1}{2}$

On peut remarquer que les variables (Total.Bags; Hars et volume) sont fortement corréliées positivement entre elles ; et sont faiblement corréliées avec la variable (Year). ~~et fortement~~

5

Year corréliée fortement de sens inverse avec les variables (Fuerle et Bacon) qui sont très corréliées positivement entre elles!

8-

On diagonalise la Matrice de Var - Cov:



On calcule les valeurs propres

$$\lambda_1 > \lambda_2 > \dots > \lambda_p$$

$$\left(\frac{1}{2}\right)$$



On calcule les vecteurs propres associés.

$$u_1, u_2, \dots, u_p$$



$$C_1 = \tilde{X} \cdot u_1 \quad C_2 = \tilde{X} \cdot u_2 \dots$$

Donc; C_1 représente le produit entre la matrice des données centrées et réduite et le premier vecteur propre; et C_2 " " " " et le deuxième vecteur propre.