# Chapitre 2 : Données bidimensionnelles

## 1M09 Master MMAS

James Ledoux

Dépt de mathématiques, Univ. Poitiers

11 juillet 2009





## Introduction : liaison entre variables

- Étude simultanée de deux variables $X$ et $Y$
- **Objectif :**
  mettre en évidence une évolution simultanée : liaison
  - Causale mais pas toujours
  - (Z) liaison n'entraîne pas toujours une causalité
- Nuages de points, diagramme de profils, ...
- Covariance, corrélation linéaire, indice de concentration, régression linéaire, ...





**Sur un même échantillon de $n$ individus :**

$$X \longmapsto \text{série : } x_1, \ldots, x_n$$
$$Y \longmapsto \text{série : } y_1, \ldots, y_n$$

- Construction du nuage des points $\mathcal{N} := \{(x_i, y_i)\}_{i=1}^n$
  (ou « scatter-plot »)
  - Si on affecte un poids $p_i$ (souvent $1/n$) à chaque individu, le centre de gravité du nuage de points

$$G = \sum_{i=1}^{n} p_i (x_i, y_i) = (\overline{X}, \overline{Y})$$

  - Variables homogènes (même grandeur et même unité) alors repère orthonormé





- Variables hétérogènes
  - soit choisir une bonne échelle de représentation
  - soit centrage-réduction des données et repère orthonormé

**Opération de centrage d'une variable**
(même unité et $\overline{X - \overline{X}} = 0$)

$$\begin{array}{rcl} X & \longrightarrow & X - \overline{X}\mathbf{1}_n \\ \{x_i\}_{i=1}^n & \longrightarrow & \{x_i - \overline{X}\}_{i=1}^n \end{array}$$

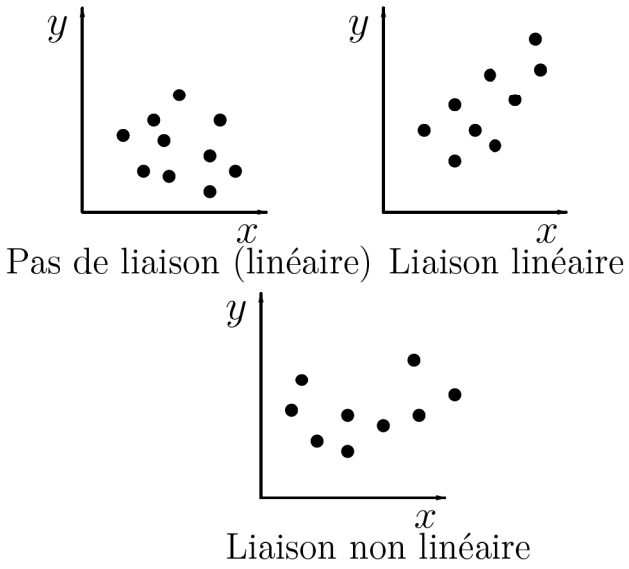
Centrage de $X$ et $Y$ : « déplacer l'origine de $\mathcal{N}$ en $G$ »
**Opération de réduction d'une variable**
(sans unité et $S_{X/s_X} = 1$)

$$\begin{array}{rcl} X & \longrightarrow & X/s_X \\ \{x_i\}_{i=1}^n & \longrightarrow & \{x_i/s_X\}_{i=1}^n \end{array}$$

**Centrage-réduction**
(sans unité, de moyenne emp. nulle et d'écart-type emp. unité)

$$\begin{array}{rcl} X & \longrightarrow & (X - \overline{X}\mathbf{1}_n)/s_X \\ \{x_i\}_{i=1}^n & \longrightarrow & \{(x_i - \overline{X})/s_X\}_{i=1}^n \end{array}$$

## Bilan visuel du nuage








### Exemple 9 (Consommation par CSP)

Pour cet exemple, il s'agit des évaluations, en franc, de dépenses (moyennes annuelles) de 8 catégorie sociaux-professionnelles (CSP) pour acquérir 8 denrées alimentaires.
On peut alors analyser chaque couple de variables.

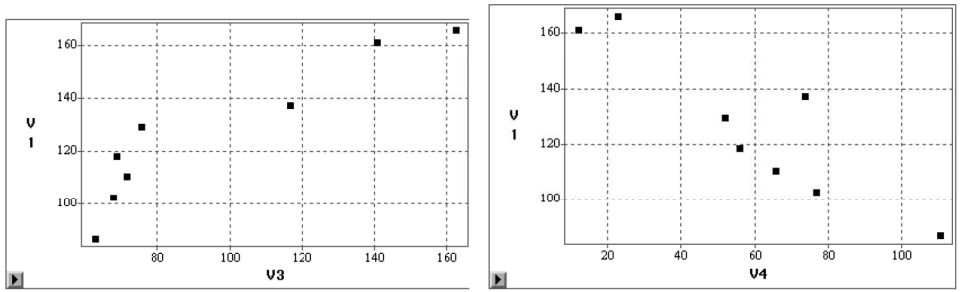



FIGURE 11: Scatter-plots résultant du croisement de deux variables





### Définition 13 (Covariance)

*La **covariance empirique** de $X$ et $Y$ est définie par*

$$s_{X,Y} := \sum_{i=1}^{n} p_i\,(x_i - \overline{X})(y_i - \overline{Y}) = \overline{(X - \overline{X})(Y - \overline{Y})} \tag{16}$$

- pour des v.a. $cov(X,Y) := \mathbb{E}\big[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\big]$
- La covariance se calcule aussi avec

$$s_{X,Y} = \overline{XY} - \overline{X}\,\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} x_i\,y_i - \overline{X}\,\overline{Y}$$

(cf $cov(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$)

- Indicateur très sensible aux valeurs aberrantes et qui s'exprime à partir des unités des variables





- **La covariance est une forme bilinéaire symétrique dont la variance est la forme quadratique associée** ($s_{X,X} = s_X^2$).
  En particulier, on en déduit les formules

$$\begin{array}{rrcl} & s_{X+Z,Y} & = & s_{X,Y} + s_{Z,Y} \\ \forall (a,b) \in \mathbb{R}^2 & s_{aX+b,Y} & = & a\,s_{X,Y} \\ & s^2_{X+Y} & = & s^2_Y + s^2_X + 2s_{X,Y} \\ & |s_{X,Y}| & \leq & s_Y \times s_X \quad \text{(Inégalité de Cauchy-Schwarz)} \end{array}$$

  On déduit de la seconde que la covariance est invariante par translation (par exemple par une opération de centrage)

- A mettre en rapport avec les formules sur les v.a.

$$\sigma^2(X+Y) = \sigma^2(X) + \sigma^2(Y) + 2cov(X,Y)$$
$$|cov(X,Y)| \leq \sigma(X)\sigma(Y) \quad \ldots$$

### Définition 14 (Coefficient de corrélation linéaire)

*Le **coefficient de corrélation linéaire empirique** est défini par :*

$$\rho_{X,Y} = \frac{s_{X,Y}}{s_X\, s_Y} = \frac{\sum_{i=1}^{n} p_i\,(x_i - \overline{X})(y_i - \overline{Y})}{s_X\, s_Y}$$

- Deux v.a.

$$\rho(X,Y) = \frac{C(X,Y)}{\sigma(X)\sigma(Y)} = \frac{E\left[(X - E[X])(Y - E[Y])\right]}{\sigma(X)\sigma(Y)}$$

- **Propriétés**
  - (a) $\rho_{X,Y}$ est symétrique
  - (b) $|\rho_{X,Y}| < 1$
  - (c) $|\rho_{X,Y}| = 1 \Longleftrightarrow$ **liaison linéaire exacte** : il existe $a, b, c \in \mathbb{R}$ tels que

$$\forall i, \quad ax_i + by_i + c = 0$$

  - (d) $\rho_{X,Y} = 0$ indique **une absence de liaison affine** entre les deux séries statistiques. Les variables $X$ et $Y$ sont dites non-corrélées.





- (e) $\rho_{X,Y} > 0$ : **Même tendance** dans les valeurs prises par $X$ et $Y$
  $\rho_{X,Y} < 0$ : **Opposition de tendance** dans les valeurs prises
  $\rho_{X,Y} \approx 1$ : pour beaucoup d'individus $i$ on a,

$$p_i \left(\frac{x_i - \overline{X}}{s_X}\right)\left(\frac{y_i - \overline{Y}}{s_Y}\right) >> 0$$

  Pour ces individus

$$\begin{cases} \text{soit : } x_i - \overline{X} >> 0 \ \textbf{ et } \ y_i - \overline{Y} >> 0 \\ \text{soit : } x_i - \overline{X} << 0 \ \textbf{ et } \ y_i - \overline{Y} << 0 \end{cases}$$

  $\rho_{X,Y} \approx -1$ : soit $\left( x_i - \overline{X} >> 0 \ \textbf{ et } \ y_i - \overline{Y} << 0 \right)$
  soit $\left( x_i - \overline{X} << 0 \ \textbf{ et } \ y_i - \overline{Y} >> 0 \right)$

- (f) On a $\rho_{aX+b,cY+d} = \text{signe}(ac)\rho_{X,Y}$.
  On en déduit que la **corrélation linéaire est invariante par un opération de centrage-réduction**

  et $\rho_{X,Y}$ est indépendante des unités de mesure de $X$ et $Y$





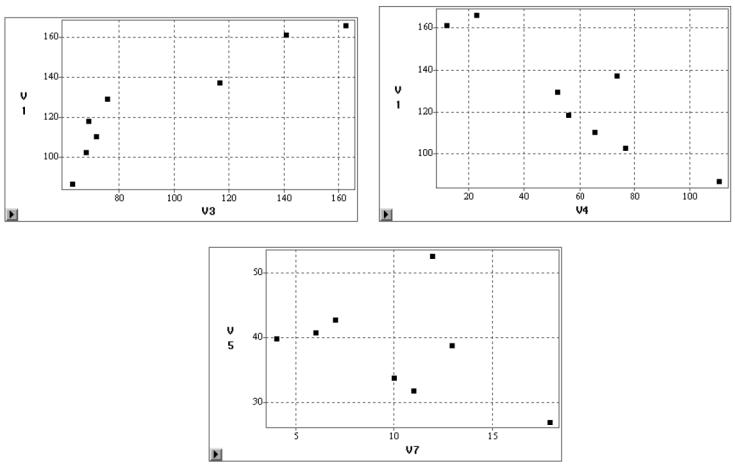



FIGURE 12: CSP : scatter-plots avec une corrélation respectivement fortement $> 0$, $< 0$ et enfin de faible amplitude





- **Données :** une variable continue explicative $X$
  une variable continue à expliquer $Y$
  Les deux variables sont observées sur $n$ individus
- **Objectif descriptif :**
  exploration d'une liaison, ici affine, entre $Y$ et une variable potentiellement explicative $X$
- **Modèle :**

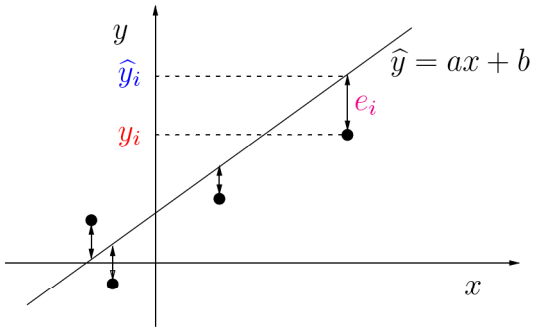
$$\boxed{Y = aX + b + E}$$

  où $a, b$ sont des constantes inconnues
  $E$ variable résiduelle (inconnue)

- **Travail**
  - ➡ Évaluer $a, b$ : $\hat{a}, \hat{b}$
    (ajustement du modèle aux données)
  - ➡ Déterminer la QLT de cet ajustement en fonction des objectifs de l'étude

## Critère des moindres carrés appliqué aux résidus observés

$i = 1, \ldots, n$ $\quad \boxed{e_i = y_i - \widehat{y}_i}$ où $\quad \widehat{y}_i = ax_i + b$




Chercher $\widehat{a}, \widehat{b}$ tels que

$$(\widehat{a}, \widehat{b}) = \arg \min_{(a,b)\in\mathbb{R}^2} \sum_i p_i\, e_i^2 = \arg \min_{(a,b)\in\mathbb{R}^2} \sum_i p_i \left\{y_i - (ax_i + b)\right\}^2$$





(17) $$\boxed{\widehat{a} = \frac{\sum_i p_i (x_i - \overline{X})(y_i - \overline{Y})}{\sum_i p_i (x_i - \overline{X})^2} = \frac{s_{X,Y}}{s_X^2}} \qquad \boxed{\widehat{b} = \overline{Y} - \widehat{a}\,\overline{X}}$$

### Définition 15 (Droite de régression linéaire)

*La **droite de régression** est donnée par*

$$\widehat{Y} = \widehat{a}X + \widehat{b}$$

*où $\widehat{a}$ et $\widehat{b}$, donnés par (17), sont les **coefficients de la régression**.*

- **Conséquences**
  - (a) **Les résidus sont centrées :** $\overline{E} = \sum_i p_i\, e_i = 0$ (si $b \neq 0$)
    $\Longrightarrow$ la droite de régression passe par le centre de gravité $G = (\overline{X}, \overline{Y})$ du nuage de points
  - (b) **Les résidus sont non-corrélés avec la variable explicative** ($\rho_{E,X} = 0$)





(c) $\widehat{a} = \dfrac{s_{X,Y}}{s_X^2} = \dfrac{s_Y}{s_X}\,\rho_{X,Y}$
$\Longrightarrow$ $\widehat{a}$ est proportionnel au coefficient de corrélation empirique $\rho_{X,Y}$

**Pente de la droite $\widehat{a}$ est du signe de $\rho_{X,Y}$**

- $\rho_{X,Y} > 0$ ➡ **Même tendance** dans les valeurs prises par $X$ et $Y$
- $\rho_{X,Y} < 0$ ➡ **Opposition de tendance** dans les valeurs prises par $X$ et $Y$

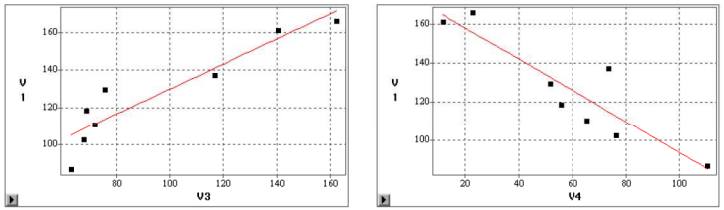



FIGURE 13: CSP : deux droites de régression





# Exploration des résidus $\{e_i\}_{i=1}^n$ :

On vérifie à partir des propriétés (a) et (b) :

### Théorème 1 (Formule de décomposition de la var. de $Y$)

*On a*

$$\begin{aligned} \overline{Y} &= \overline{\widehat{Y}} = \sum_{i=1}^n p_i\, \widehat{y}_i \\ (18)\quad \sum_i p_i (y_i - \overline{Y})^2 &= \sum_i p_i (y_i - \widehat{y}_i)^2 + \sum_i p_i (\widehat{y}_i - \overline{Y})^2 \\ s_Y^2 &= \underbrace{\sum_i p_i\, e_i^2}_{\text{var. résiduelle } s_E^2} + \underbrace{\sum_i p_i (\widehat{y}_i - \overline{Y})^2}_{\text{variance expliquée par } X \,:\, s_{Y:X}^2} \end{aligned}$$

(19) $$\boxed{s_Y^2 = s_E^2 + s_{Y:X}^2}$$

Lorsque $p_i = 1/n$ la relation (18) est souvent donnée sous la forme :

$$\text{SommeCarrésTotale} = \text{SommeCarrésRésiduelle} + \text{SommeCarrésExpliquée}$$

**Définition 16 (Coefficient de détermination)**

*À partir de (19), on a* $s_E^2 = (1 - R^2)s_Y^2$ *avec*

$$\boxed{R^2 = \frac{s_{Y:X}^2}{s_Y^2} = \frac{\sum_i p_i(\widehat{y}_i - \overline{Y})^2}{\sum_i p_i(y_i - \overline{Y})^2} \in [0,1]}$$

$R^2$ *est appelé le* ***coefficient de détermination de la régression*** *et représente la* ***proportion de la variance de*** $Y$ ***expliquée par la régression***.





**Proposition 1**

*On a*

$$R^2 = \rho_{Y,\widehat{Y}}{}^2.$$

*Dans le cas de la régression simple, on a également :*

$$R^2 = \rho_{X,Y}{}^2.$$

- $s_{Y:X}^2 = 0 \iff R^2 = 0 \iff \rho_{X,Y} = 0 \iff \widehat{a} = 0$ et la droite de régression se réduit à $\widehat{Y} = \overline{Y}$
  - ➡ « Plus $R^2$ est proche de 0, plus l'ajustement aux données est de mauvaise qualité du point de vue descriptif »
- À l'opposé,

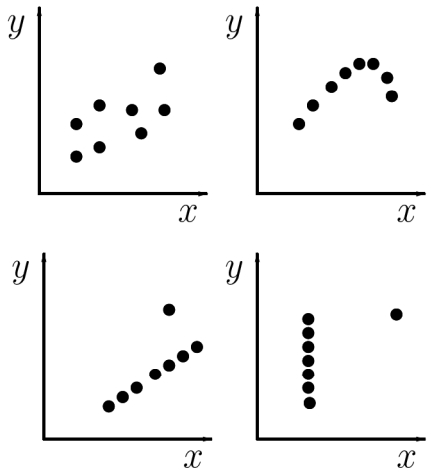
$$R^2 = 1 \iff s_{Y:X}^2 = s_Y^2 \iff E = 0 \iff Y = \widehat{a}\,X + \widehat{b} \iff |\rho_{X,Y}| = 1$$

  - ➡ « Plus $R^2$ est proche de 1, plus l'ajustement aux données est de qualité du point de vue descriptif »





**MAIS**




$\boldsymbol{\rho_{X,Y}}$, $\boldsymbol{\overline{X}}$, $\boldsymbol{\overline{Y}}$, $\boldsymbol{s_X^2}$, $\boldsymbol{s_Y^2}$ **sont identiques pour chaque nuage**





**Sur un même échantillon de $n$ individus :**

$X$ : qualitative à $r$ modalités

$Y$ : quantitative

- **Chaque modalité de $X$ définit une classe** ou un sous-échantillon $\mathcal{I}_l$ de taille $n_l$ de l'échantillon $\mathcal{I}$ des $n$ individus
- Calcul des résumés numériques associée à chaque classe :

$$l = 1, \ldots, r \quad \underbrace{\overline{Y}(l) := \frac{1}{n_l}\sum_{i \in \mathcal{I}_l} y_i}_{\text{moyenne intra-classe}} \quad \underbrace{s_Y^2(l) := \frac{1}{n_l}\sum_{i \in \mathcal{I}_l} \left(y_i - \overline{Y}(l)\right)^2}_{\text{variance intra-classe}}$$

# Introduction à l'analyse de variance à 1 facteur

## Exemple 10 (Effet du pentobarbital sur le rythme cardiaque)

19 chiens ont été prémédiqués au pentobarbital et on étudie l'effet de deux facteurs croisés sur leur rythme cardiaque.
L'effet est mesuré par le temps entre deux battements de coeur successifs (variable $Y$ en millisecondes).
Les deux facteurs à deux niveaux chacun pris en compte ne sont pas détaillés ici mais la variable qualitative représente les 4 modalités résultant de leur croisement.
Chaque chien a été observé dans les 4 conditions associées aux modalités. Cela donne un échantillon de 76 individus.

| | $\mathcal{I}_1$ | $\mathcal{I}_2$ | $\mathcal{I}_3$ | $\mathcal{I}_4$ | $\mathcal{I}$ |
|---|---|---|---|---|---|
| moyenne | 368.2 | 404.6 | 479.3 | 502.9 | 438.8 |
| écart-type | 51.7 | 86.9 | 80.6 | 68.0 | 91.1 |





## Boîtes parallèles

- Sur un même graphique, tracer les boîtes à moustaches associées à chacune des modalités de $X$
- Comparer les boîtes et mesurer l'influence de $X$ sur $Y$

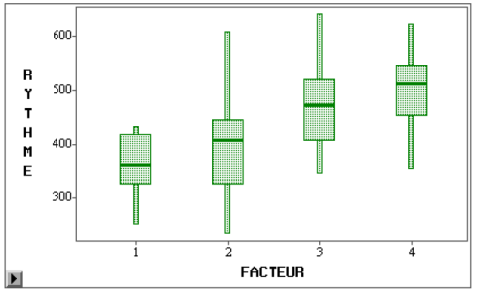



FIGURE 14: Box-plot des données de l'Exemple 10





## Théorème 2 (Formule de décomposition de la moy. et var. de $Y$)

*On a*

$$\overline{Y} = \frac{1}{n}\sum_{l=1}^{r} n_l \overline{Y}(l) = \sum_{l=1}^{r} f_l \overline{Y}(l)$$

$$s_Y^2 = \underbrace{\frac{1}{n}\sum_{l=1}^{r} n_l \left(\overline{Y}(l) - \overline{Y}\right)^2}_{\text{variance inter-classes}} + \underbrace{\frac{1}{n}\sum_{l=1}^{r} n_l\, s_Y^2(l)}_{\text{variance intra-classes}}$$

***variance inter-classes*** *ou* ***var. expliquée par*** $X$ : $\boldsymbol{s^2_{Y:X}}$

***variance intra-classes*** *ou* ***var. résiduelle*** $\boldsymbol{s_r^2}$

Pour des v.a. $X$ et $Y$

$$\begin{aligned}\mathbb{E}[Y] &= \mathbb{E}[\mathbb{E}[Y \mid X]] \\ \sigma^2(Y) &= \sigma^2(\mathbb{E}[Y \mid X]) + \mathbb{E}[\sigma^2(Y \mid X)]\end{aligned}$$





## Définition 17 (Rapport de corrélation)

*Le rapport de corrélation est un indice de liaison entre les variables $X$ et $Y$ défini par le rapport :*

$$r_{Y:X} := \sqrt{\frac{s^2_{Y:X}}{s^2_Y}} \tag{20}$$

- D'après Th 2, on a $0 \leq r_{Y:X} \leq 1$
- ${r_{Y:X}}^2$ représente la **proportion de variance expliquée par** $X$
- $r_{Y:X} = 1$ signifie que la variance sur $Y$ correspond exactement à la variance inter-classes ou encore est complètement expliquée par $X$. ($Y$ est constante dans chaque classe)
- À l'opposé $r_{Y:X} = 0$ signifie que la variance sur $Y$ n'est en aucun cas expliquée par $X$ $\quad(\overline{Y}(l) = \overline{Y}, l = 1, \ldots, r)$

| | $s^2_Y$ | $s^2_{Y:X}$ | $r_{Y:X}$ |
|---|---|---|---|
| **Exemple 10** | 8305.9 | 2973.94 | 0.6 |

## Données : 2 variables qualitatives

$X$ : $I = \{1, \dots, q\}$ $q$ modalités
$Y$ : $J = \{1, \dots, p\}$ $p$ modalités

| | $1$ | $\cdots$ | $j$ | $\cdots$ | $p$ | |
|---|---|---|---|---|---|---|
| $1$ | | | $\vdots$ | | | |
| $\vdots$ | | | $\vdots$ | | | |
| $i$ | $\cdots$ | $\cdots$ | $n_{ij}$ | $\cdots$ | $\cdots$ | $n_{i.}$ |
| $\vdots$ | | | $\vdots$ | | | |
| $q$ | | | $\vdots$ | | | |
| | | | $n_{.j}$ | | | $n$ |

$$n_{i.} = \sum_{j=1}^{p} n_{ij}$$
$$n_{.j} = \sum_{i=1}^{q} n_{ij}$$
$$n = \sum_{i} \sum_{j} n_{ij}$$

TABLE 1: **Table de contingence** ou tableau d'effectifs conjoints

**$X$ et $Y$ indépendantes ?**





| | | Couleur des cheveux | | | | |
|---|---|---|---|---|---|---|
| | | brune | châtain | roux | blond | Effectifs |
| **Couleur des yeux** | marron | 68 | 119 | 26 | 7 | $n_{1.} = 220$ |
| | noisette | 15 | 54 | 14 | 10 | $n_{2.} = 93$ |
| | vert | 5 | 29 | 14 | 16 | $n_{3.} = 64$ |
| | bleu | 20 | 84 | 17 | 94 | $n_{4.} = 215$ |
| | Effectifs | $n_{.1} = 108$ | $n_{.2} = 286$ | $n_{.3} = 71$ | $n_{.4} = 127$ | $n = 592$ |

TABLE 2: Table de contingence : ventilation d'une population de 592 femmes suivant leurs couleurs des yeux et des cheveux





## Tableau des fréquences conjointes

| | $1 \quad \cdots \quad p$ | Marge col. |
|---|---|---|
| $1$<br>$\vdots$<br>$q$ | $f_{ij} = \frac{n_{ij}}{n}$ | $f_{i.}$ |
| Marge lig. | $f_{.j}$ | $1$ |

avec $f_{.j} = \sum_{i \in I} f_{ij} = \frac{n_{.j}}{n}$ $\qquad f_{i.} = \sum_{j \in J} f_{ij} = \frac{n_{i.}}{n}$

**Interprétation :** tableau $(f_{ij})_{i,j}$ d'une loi de probabilité conjointe sur l'ensemble produit $I \times J$

Marge colonne $\equiv (f_{i.})_{i \in I}$ : loi marginale de la variable $X$
Marge ligne $= (f_{.j})_{j \in J}$ : loi marginale de la variable $Y$





- **Tableau des profils-lignes**

Lois conditionnelles $(Y|X = i)$

| | $1 \quad \cdots \quad j \quad \cdots \quad p$ | Total |
|---|---|---|
| $i$ | $\frac{f_{i1}}{f_{i.}} \quad \cdots \quad \frac{f_{ij}}{f_{i.}} \quad \cdots \quad \frac{f_{ip}}{f_{i.}}$ | $1$ |
| Profil moyen | $f_{.1} \quad \cdots \quad f_{.j} \quad \cdots \quad f_{.p}$ | $1$ |

- **Tableau des profils-colonnes**

Lois conditionnelles $(X|Y = j)$

| | $j$ | Profil moyen |
|---|---|---|
| $1$ | $\frac{f_{1j}}{f_{.j}}$ | $f_{1.}$ |
| $\vdots$ | $\vdots$ | $\vdots$ |
| $i$ | $\frac{f_{ij}}{f_{.j}}$ | $f_{i.}$ |
| $\vdots$ | $\vdots$ | $\vdots$ |
| $q$ | $\frac{f_{nj}}{f_{.j}}$ | $f_{q.}$ |
| Total | $1$ | $1$ |

| | | Couleur des cheveux | | | | |
|---|---|---|---|---|---|---|
| | | brune | châtain | roux | blond | Total |
| Couleur des yeux | marron | 0.31 | 0.54 | 0.12 | 0.03 | 1 |
| | noisette | 0.16 | 0.58 | 0.15 | 0.11 | 1 |
| | vert | 0.08 | 0.45 | 0.22 | 0.25 | 1 |
| | bleu | 0.09 | 0.39 | 0.08 | 0.44 | 1 |
| Profil moyen ou marge | | 0.18 | 0.48 | 0.12 | 0.22 | 1 |

| | | Couleur des cheveux | | | | Profil moyen |
|---|---|---|---|---|---|---|
| | | brune | châtain | roux | blond | ou marge |
| Couleur des yeux | marron | 0.63 | 0.42 | 0.37 | 0.06 | 0.37 |
| | noisette | 0.14 | 0.19 | 0.20 | 0.08 | 0.16 |
| | vert | 0.05 | 0.10 | 0.20 | 0.13 | 0.11 |
| | bleu | 0.19 | 0.29 | 0.24 | 0.74 | 0.36 |
| | Total | 1 | 1 | 1 | 1 | 1 |

TABLE 3: Tables des profils-ligne et profils-colonne





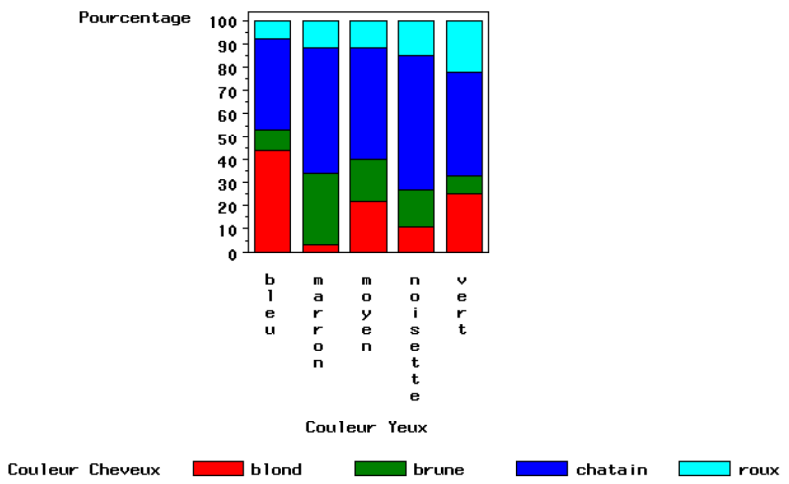



FIGURE 15: Profils lignes : famille de diagrammes en barre





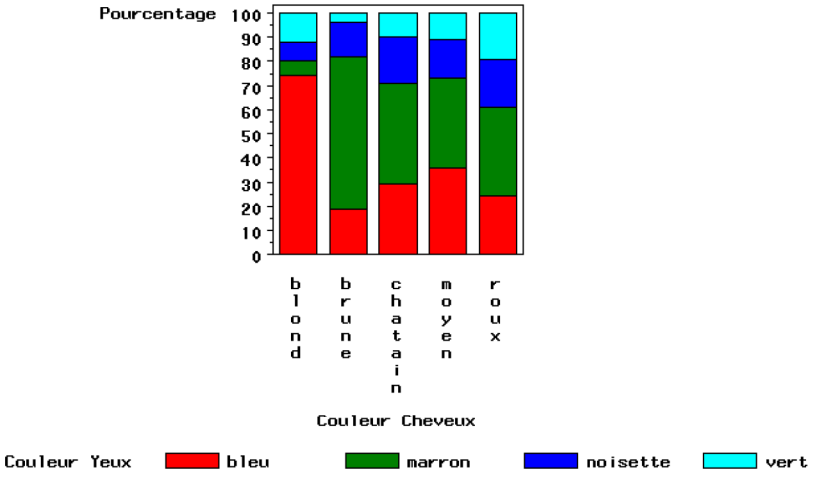



FIGURE 16: Profils colonnes : famille de diagrammes en barre





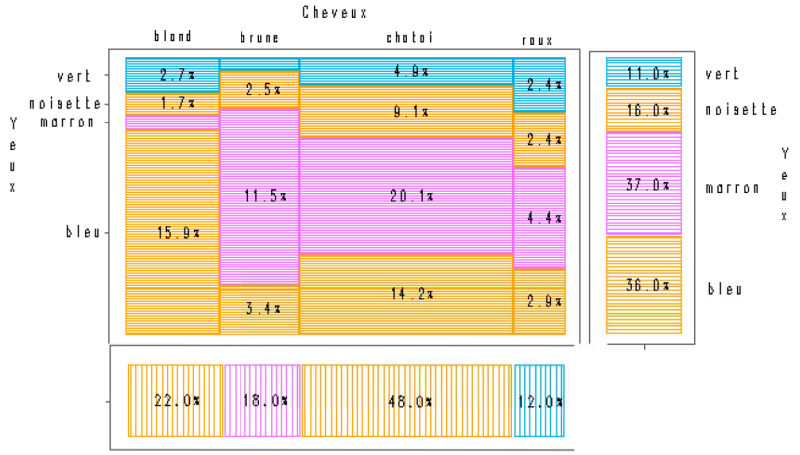



FIGURE 17: « Mosaïque »

**Indépendance de $X$ et $Y$ :**

(21) $$\boxed{\forall (i,j) \in I \times J \qquad f_{ij} = f_{i.} f_{.j}}$$

$\Longleftrightarrow$ **Tous les profils-lignes sont égaux**

$$\forall i \in I, \qquad \left(\frac{f_{ij}}{f_{i.}}\right)_{j\in J} = (f_{.j})_{j\in J} \qquad \text{i.e. Loi}(Y|X) = \text{Loi}(Y) = (f_{.j})_{j\in J}$$

$\Longleftrightarrow$ **Tous les profils-colonnes sont égaux**

$$\forall j \in J, \qquad \left(\frac{f_{ij}}{f_{.j}}\right)_{i\in I} = (f_{i.})_{i\in I} \qquad \text{i.e. Loi}(X|Y) = \text{Loi}(X) = (f_{i.})_{i\in I}$$

$\Longleftrightarrow$ $$\boxed{\forall (i,j) \in I \times J \qquad n_{ij} = \frac{n_{i.}\, n_{.j}}{n}}$$

**Remarque :** les lignes et les colonnes jouent un rôle totalement symétrique





- **Mesure de l'écart à l'indépendance** : indice du khi-deux

(22) $$\chi^2 = \sum_{i\in I}\sum_{j\in J} \frac{(n_{ij} - s_{ij})^2}{s_{ij}}$$

On peut montrer que $0 \leq \chi^2 \leq n \min(q-1, p-1)$

- Le **phi-deux** : $\Phi^2 := \chi^2/n$.
- Le **coefficient de Tschuprow** :

$$T := \sqrt{\frac{\Phi^2}{\sqrt{(q-1)(p-1)}}}.$$

- Le **coefficient de Cramer** :

$$V := \sqrt{\frac{\Phi^2}{d-1}} \quad d := \min(q,p).$$

On peut vérifier : $0 \leq T \leq V \leq 1$.

$n = 592$ et $q = p = 4$
$\chi^2 \simeq 138.29$, $\Phi^2 \simeq 0.23$, V de Cramer $\simeq 0.28$ : **conclusion ?**
Possible avec des résultats probabilistes sur certaines de ces statistiques





- **Comparer la table de contingence observée à une table de contingence construite sous l'hypothèse d'indépendance**

Table de référence $\equiv$ table des effectifs théoriques $(s_{ij})_{i,j}$ avec

$$s_{ij} = \frac{n_{i.} n_{.j}}{n}$$

**Remarque :** Les marges $(f_{i.})$ et $(f_{.j})$ sont fixées par le tableau initial

| | | Couleur des cheveux | | | | Profil moyen |
|---|---|---|---|---|---|---|
| | | brune | châtain | roux | blond | ou marge |
| Couleur des yeux | marron | 0.11 | 0.20 | 0.04 | 0.011 | $f_{1.} = 0.37$ |
| | noisette | 0.03 | 0.09 | 0.02 | 0.02 | $f_{2.} = 0.16$ |
| | vert | 0.01 | 0.05 | 0.02 | 0.03 | $f_{3.} = 0.11$ |
| | bleu | 0.03 | 0.14 | 0.03 | 0.16 | $f_{4.} = 0.36$ |
| Profil moyen ou marge | | $f_{.1} = 0.18$ | $f_{.2} = 0.48$ | $f_{.3} = 0.12$ | $f_{.4} = 0.22$ | 1 |

| | | Couleur des cheveux | | | | Profil moyen |
|---|---|---|---|---|---|---|
| | | brune | châtain | roux | blond | ou marge |
| Couleur des yeux | marron | 0.07 | 0.18 | 0.04 | 0.08 | $f_{1.} = 0.37$ |
| | noisette | 0.03 | 0.08 | 0.02 | 0.03 | $f_{2.} = 0.16$ |
| | vert | 0.02 | 0.05 | 0.01 | 0.02 | $f_{3.} = 0.11$ |
| | bleu | 0.07 | $f_{.2} f_{4.} = 0.18$ | 0.12 | 0.08 | $f_{4.} = 0.36$ |
| Profil moyen ou marge | | $f_{.1} = 0.18$ | $f_{.2} = 0.48$ | $f_{.3} = 0.12$ | $f_{.4} = 0.22$ | 1 |

TABLE 4: Tables des fréquences observées et théoriques