Université de Bordeaux - Master MIMSE - 2ème année

Notions de base pour l'analyse d'un tableau de contingence

Marie Chavent

http://www.math.u-bordeaux.fr/ machaven/

2014-2015

1 Notations et définitions

Un tableau de contingence est une matrice **K** de dimension $q \times m$ obtenue en croisant **deux variables qualitatives** X_1 et X_2 (ayant respectivement q et m modalités) observées sur un échantillon n individus.

$$
\begin{array}{c|ccc|c}
 & 1 \dots & s & \dots m & \\
\hline
1 & & & & \\
\vdots & & \vdots & & \\
i & \dots & n_{is} & \dots & n_{i.} \\
\vdots & & \vdots & & \\
q & & & & \\
\hline
 & & n_{.s} & & n_{..} = n
\end{array}
$$

Table 1 – Matrice de contingence **K**, effectifs marginaux.

On note :
— i une modalité de X_1
— s une modalité de X_2
— q le nombre de modalités de X_1
— m le nombre de modalités de X_2
— n_{is} le nombre d'individus de l'échantillon qui possèdent les modalités i et s
— $n_{i.} = \sum_{s=1}^{m} n_{is}$ est le nombre d'individus qui possèdent la modalité i
— $n_{.s} = \sum_{i=1}^{q} n_{is}$ est le nombre d'individus qui possèdent la modalité s
— $n_{..} = \sum_{i=1}^{q} \sum_{s=1}^{m} n_{is}$ est le nombre totel d'individus dans l'échantillon.

Exemple. On considère un tableau de contingence obtenu en ventilant 592 femmes suivant la couleur de leurs yeux et la couleur de leurs cheveux.

```
K <- read.table("couleurs.txt")
K

##           brun chatain roux blond
## marron      68     119   26     7
## noisette    15      54   14    10
## vert         5      29   14    16
## bleu        20      84   17    94

sum(K)

## [1] 592
```

On en déduit la matrice des fréquences $\mathbf{F}$ avec :

— $f_{is} = \frac{n_{is}}{n}$ le terme général de $\mathbf{F}$

— $f_{i.} = \sum_{s=1}^{m} f_{is} = \frac{n_{i.}}{n}$ les **masses des lignes**

— $f_{.s} = \sum_{i=1}^{q} f_{is} = \frac{n_{.s}}{n}$ les **masses des colonnes**

$$\begin{array}{c|ccc|c}
 & 1 \dots & s & \dots m & \\
\hline
1 & & & & \\
\vdots & & \vdots & & \\
i & \dots & f_{is} & \dots & f_{i.} \\
\vdots & & \vdots & & \\
q & & & & \\
\hline
 & & f_{.s} & & f_{..} = 1
\end{array}$$

TABLE 2 – Matrice des fréquences $\mathbf{F}$, fréquences marginales.

On notera par la suite :

— $\mathbf{r} = (f_{1.}, \dots, f_{i.}, \dots, f_{q.})^t$ le vecteur des poids des lignes (row)

— $\mathbf{c} = (f_{1.}, \dots, f_{.s}, \dots, f_{.m})^t$ le vecteur des poids des colonnes

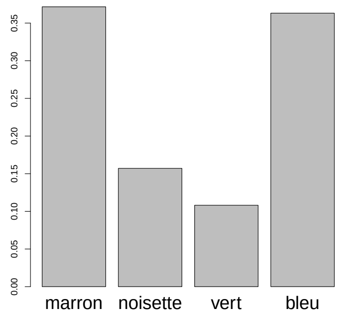
et

— $\mathbf{D}_r$=diag($\mathbf{r}$) la matrice diagonale de dimension $q \times q$ des poids des lignes

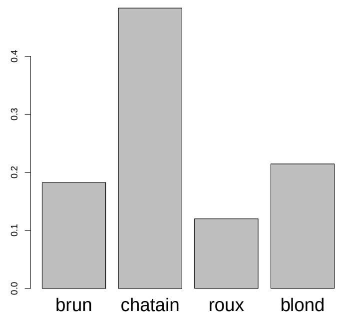
— $\mathbf{D}_c$=diag($\mathbf{c}$) la matrice diagonale de dimension $m \times m$ des poids des colonnes

Exemple.

```
#----------Calcul de la matrice des frequences F

F <- K/sum(K)
round(F*100,digit=2)

##            brun chatain roux blond
## marron    11.49   20.10 4.39  1.18
## noisette   2.53    9.12 2.36  1.69
## vert       0.84    4.90 2.36  2.70
## bleu       3.38   14.19 2.87 15.88

#----------Vecteurs r et c des poids des lignes et des colonnes (distributions marginales)

r <- apply(F,1,sum)
round(r,digit=2)

##   marron noisette     vert     bleu
##     0.37     0.16     0.11     0.36

c <- apply(F,2,sum)
round(c,digit=2)

##    brun chatain    roux   blond
##    0.18    0.48    0.12    0.21

barplot(r,cex.names=2)
```




```r
barplot(c,cex.names=2)
```



2 Matrice des profils lignes

La matrice des profils lignes **L** est obtenue en divisant chaque ligne i de **F** par son poids $f_{i.}$.

	$1 \ldots$	s	$\ldots m$
1			
$\vdots$			
i	$\ldots$	$f_{is}/f_{i.}$	$\ldots$
$\vdots$			
q			

TABLE 3 – Matrice des profils lignes **L**.

On note :

— $l_{is} = f_{is}/f_{i.} = n_{is}/n_{i.}$ le terme général de **L**,

— $\mathbf{l}_i = (l_{i1}, ..., l_{im})^t$ le vecteur de $\mathbb{R}^m$ décrivant la modalité i de X_1 (une ligne de **L**),

et on a :

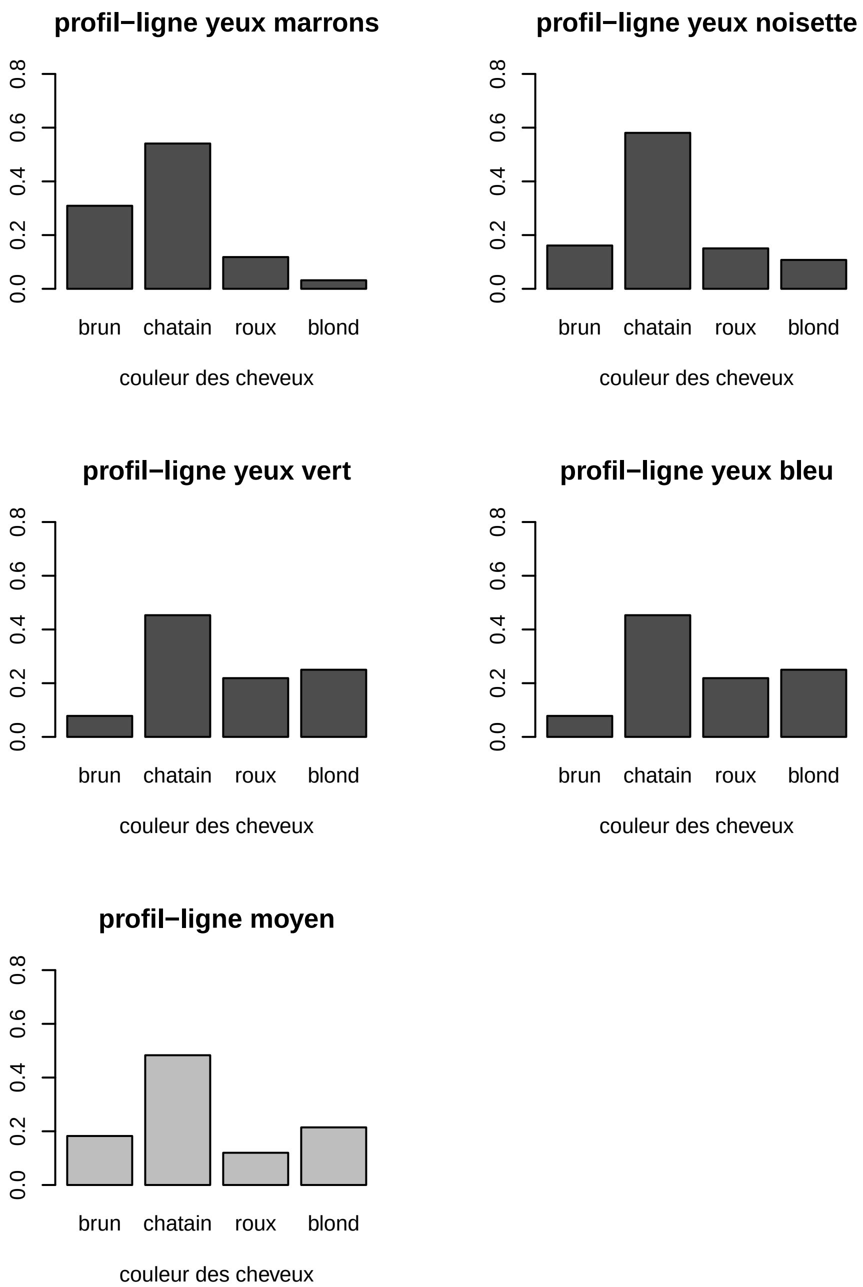
$$\mathbf{L} = \mathbf{D}_r^{-1}\mathbf{F}$$

Les q modalités de X_1 sont ainsi décrites par leurs profils lignes et ils forment un nuage de q vecteurs de $\mathbb{R}^m$ pondérés par les $f_{i.}$. On montre alors que le profil ligne moyen, centre de gravité de ce nuage, est **c** le vecteur des poids des lignes.

$\hookrightarrow$ **Preuve**.

Exemple.

```r
#---------Matrice des profils-lignes L (distributions conditionnelles en ligne)

L <- sweep(F,1,STAT=r,FUN="/")
round(L,digits=2)

##          brun chatain roux blond
## marron   0.31    0.54 0.12  0.03
## noisette 0.16    0.58 0.15  0.11
## vert     0.08    0.45 0.22  0.25
## bleu     0.09    0.39 0.08  0.44
```



On peut alors centrer **L** et le terme général de la matrice **L** centrée est :

$$f_{is}/f_{i.} - f_{.s} = \frac{f_{is} - f_{i.}f_{.s}}{f_{i.}}$$

On a alors la matrice $\mathbf{L}$ centrée qui vaut :

$$\mathbf{L} = \mathbf{D}_r^{-1}(\mathbf{F} - \mathbf{r}\mathbf{c}^t)$$

3 Matrice des profils colonne

La matrice des profils colonne $\mathbf{C}$ est obtenue en divisant chaque colonne s de $\mathbf{F}$ par son poids $f_{.s}$.

$$\begin{array}{c|ccc|}
 & 1\ldots & s & \ldots m \\
\hline
1 & & & \\
\vdots & & & \\
i & \ldots & f_{is}/f_{.s} & \ldots \\
\vdots & & & \\
q & & & \\
\hline
 & & & \\
\end{array}$$

Table 4 – Matrice des profils colonne $\mathbf{C}$.

On note :

— $c_{is} = f_{is}/f_{.s} = n_{is}/n_{.s}$ le terme général de $\mathbf{C}$,

— $\mathbf{c}_s = (c_{1s}, ..., c_{qs})^t$ le vecteur de $\mathbb{R}^q$ décrivant la modalité s de X_2 (une colonne de $\mathbf{C}$),

et on a :

$$\mathbf{C} = \mathbf{F}\mathbf{D}_c^{-1}$$

On considère maintenant que les m modalités de X_2 sont décrites par leurs profils colonne et qu'elles forment un nuage de m vecteurs de $\mathbb{R}^q$ pondérés par les $f_{.s}$. On montre alors que le profil colonne moyen, centre de gravité de ce nuage, est $\mathbf{r}$ le vecteur des poids des lignes. On peut alors centrer $\mathbf{C}$ et le terme général de la matrice $\mathbf{C}$ centrée est :

$$f_{is}/f_{.s} - f_{i.} = \frac{f_{is} - f_{i.}f_{.s}}{f_{.s}}$$

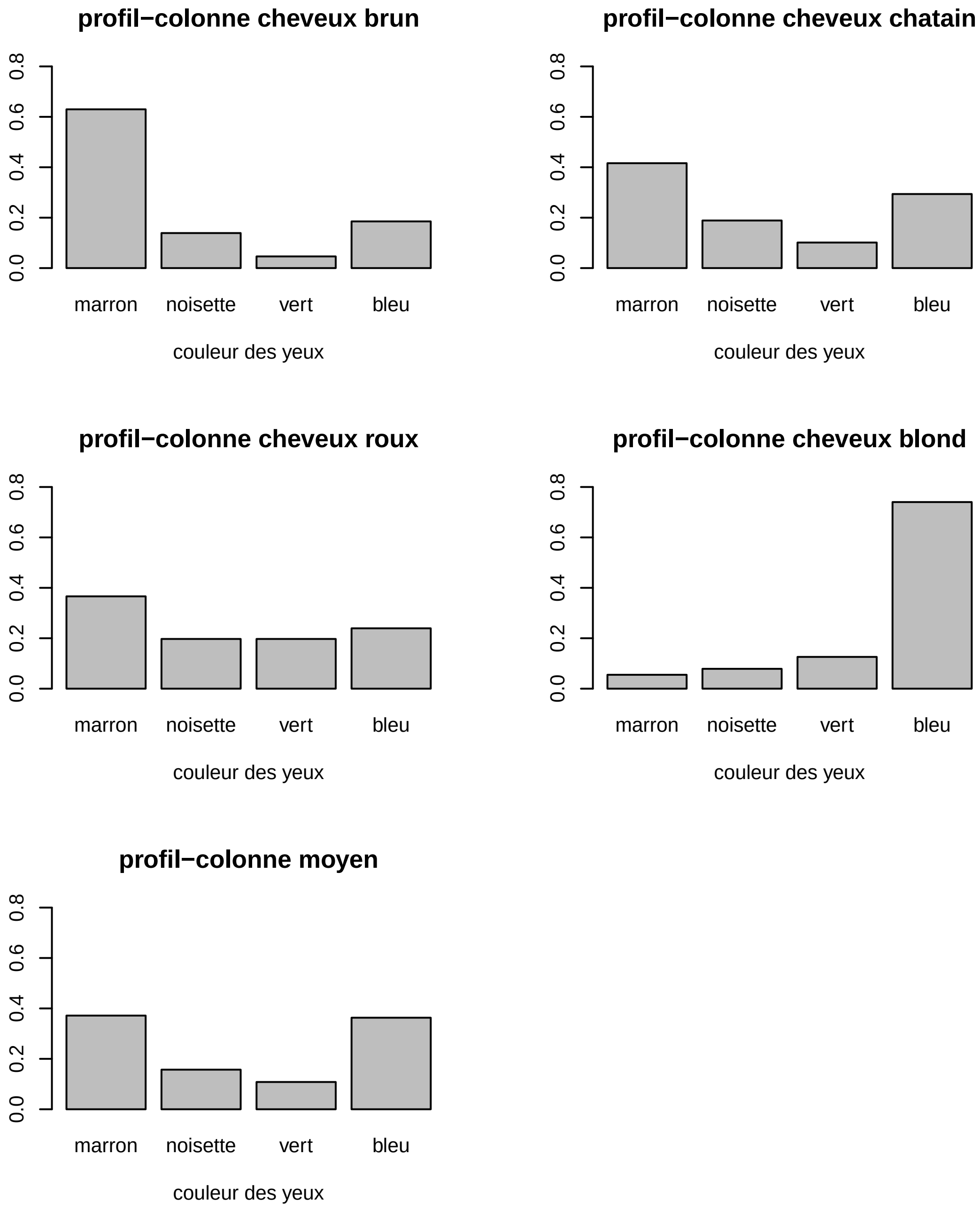
On a alors la matrice $\mathbf{C}$ centrée qui vaut :

$$\mathbf{C} = (\mathbf{F} - \mathbf{r}\mathbf{c}^t)\mathbf{D}_c^{-1}$$

Exemple.

```r
#---------Matrice des profils-lignes L (distributions conditionnelles en ligne)

L <- sweep(F,1,STAT=r,FUN="/")
round(L,digits=2)
```

```
##           brun chatain roux blond
## marron    0.31    0.54 0.12  0.03
## noisette  0.16    0.58 0.15  0.11
## vert      0.08    0.45 0.22  0.25
## bleu      0.09    0.39 0.08  0.44
```



4 Comparer les modalités d'une même variable

L'un des objectifs de l'analyse descriptive d'un tableau de contingence est d'analyser les "ressemblances" entre les modalités d'une même variable. Pour cela, on utilise une distance entre profils ligne pour comparer les modalités de X_1 et une distance entre profils colonnes pour comparer les modalités de X_2 :

— on utilise la métrique $\mathbf{D}_c^{-1}$ pour comparer deux profils lignes de $\Re^m$,

— on utilise la métrique $\mathbf{D}_r^{-1}$ pour comparer deux profils colonnes de $\Re^q$.

La distance entre deux profils lignes $\mathbf{l}_i$ et $\mathbf{l}_{i'}$ de $\mathbf{L}$ est donc :

$$\begin{aligned} d(\mathbf{l}_i, \mathbf{l}_{i'}) &= (\mathbf{l}_i - \mathbf{l}_{i'})^t \mathbf{D}_c^{-1} (\mathbf{l}_i - \mathbf{l}_{i'}) \\ &= \sum_{s=1}^{m} \frac{1}{f_{.s}} (\frac{f_{is}}{f_{i.}} - \frac{f_{i's}}{f_{i'.}})^2 \end{aligned}$$

Il s'agit de la distance Euclidienne entre deux lignes de $\mathbf{L}$, pondérée par l'inverse des poids des colonnes $\frac{1}{f_{.s}}$. La distance entre deux profils colonnes $\mathbf{c}_s$ et $\mathbf{c}_{s'}$ de $\mathbf{C}$ est de manière similaire :

$$\begin{aligned} d(\mathbf{c}_s, \mathbf{c}_{s'}) &= (\mathbf{c}_s - \mathbf{c}_{s'})^t \mathbf{D}_r^{-1} (\mathbf{c}_s - \mathbf{c}_{s'}) \\ &= \sum_{i=1}^{q} \frac{1}{f_{i.}} (\frac{f_{is}}{f_{.s}} - \frac{f_{is'}}{f_{.s'}})^2 \end{aligned}$$

Il s'agit de la distance Euclidienne entre deux colonnes de $\mathbf{C}$, pondérée par l'inverse des poids des lignes $\frac{1}{f_{i.}}$.

Ces distances sont appellées distances du χ^2 car l'inertie du nuage des profils ligne et l'inertie du nuage des profils colonne, calculée avec cette distance, est égale à $1/n$ près, au χ^2 entre les variables X_1 et X_2 (cf. section 6).

Exemple.

```
#---------distance du chi2 entre les profils lignes
sum((L[1,]-L[2,])^2/c) #carre de la distance entre marron et noisette

## [1] 0.1584588

sum((L[1,]-L[4,])^2/c) #carre de la distance entre marron et bleu

## [1] 1.081431

sum((L[1,]-c)^2/c) #carre de la distance entre marron et moyenne

## [1] 0.2504869
```

5 Liaison entre une modalité de X_1 et une modalité de X_2

En cas d'indépendance entre X_1 et X_2 on a :

$$f_{is} = f_{i.} f_{.s} \Leftrightarrow n_{is} = \frac{n_{i.} n_{.s}}{n} \tag{1}$$

avec $\begin{cases} f_{i.} f_{.s} = \text{fréquence théorique}, \\ \frac{n_{i.} n_{.s}}{n} = \text{effectif théorique}. \end{cases}$

On se sert de ce résultat pour définir une mesure "locale" de liaison entre une modalité i de X_1 et une modalité s de X_2. On dira que :

— si $f_{is} \approx f_{i.} f_{.s}$ alors il y a indépendance entre i et s,

— si $f_{is} > f_{i.}f_{.s}$ alors les modalités i et s s'attirent (car $\frac{f_{is}}{f_{i.}} > f_{.s}$ et $\frac{f_{is}}{f_{.s}} > f_{i.}$),

— si $f_{is} < f_{i.}f_{.s}$ alors les modalités i et s se repoussent (car $\frac{f_{is}}{f_{i.}} < f_{.s}$ et $\frac{f_{is}}{f_{.s}} < f_{i.}$).

Cela se mesure avec le taux de liaison :

$$t_{is} = \frac{f_{is} - f_{i.}f_{.s}}{f_{i.}f_{.s}}$$

Exemple.

```
#---------Matrice des taux de liaisons :
T <- (F-r%*%t(c))/(r%*%t(c))
round(T,digit=2)

##            brun chatain  roux blond
## marron     0.69    0.12 -0.01 -0.85
## noisette -0.12    0.20  0.26 -0.50
## vert      -0.57   -0.06  0.82  0.17
## bleu      -0.49   -0.19 -0.34  1.04
```

Sur cet exemple :

— La fréquence des bruns aux yeux marron est 69,4 au dessus de ce qu'elle serait s'il y avait indépendance entre les deux variables couleurs des yeux et couleur des cheveux. De plus ces deux modalités s'attirent (le vérifier sur les tableaux de profils ligne et colonne).

— La fréquence des blonds aux yeux marron est 85,2 au dessous de ce qu'elle serait s'il y avait indépendance entre les deux variables. En plus ces deux modalités se repoussent (le vérifier sur les tableaux de profils ligne et colonne).

6 Inertie et liaison entre X_1 et X_2

On se sert aussi de (1) pour définir une mesure "globale" de liaison entre X_1 et X_2 :

$$\chi^2 = \sum_{i=1}^{q}\sum_{s=1}^{m} \frac{(n_{is} - \frac{n_{i.}n_{.s}}{n})^2}{\frac{n_{i.}n_{.s}}{n}} = n \underbrace{\sum_{i=1}^{q}\sum_{s=1}^{m} \frac{(f_{is} - f_{i.}f_{.s})^2}{f_{i.}f_{.s}}}_{\Phi^2}$$

Remarques :

— En cas d'indépendance, on a $\chi^2 = \Phi^2 = 0$.

— Si on pondère chaque taux de liaison t_{is} par $f_{i.}f_{.s}$ on a $\bar{t} = \sum_i \sum_s f_{i.}f_{.s}t_{is} = 0$ et donc $\mathrm{Var}(t) = \sum_{i,s} f_{i.}f_{.s}t_{is}^2 = \frac{\chi^2}{n} = \Phi^2$.

Théorème : On a :

$$\frac{\chi^2}{n} = I(\mathbf{L}) = I(\mathbf{C})$$

où $I(\mathbf{L})$ est l'inertie du nuage des profils lignes et où $I(\mathbf{C})$ est l'inertie du nuage des profils colonnes calculées avec la distance du χ^2.

$\hookrightarrow$ **Preuve**. Il faut utiliser la définition de l'inertie d'une nuage de points pondérés.

Exemple.

```
#----------Chi2 et inertie
chisq.test(K)$statistic

## X-squared
##  138.2898

distsq <- function(x) #fonction pour calculer le carre de la distance du chi2 au profil ligne moyen.
{
sum((x-c)^2/c)
}

distsq(L[1,]) #carre de la distance du chi2 en profil yeux marron et profil moyen

## [1] 0.2504869

sum((L[1,]-c)^2/c) #carre de la distance entre marron et moyenne

## [1] 0.2504869

apply(L,1,distsq) #vecteur des carres des ecarts

##     marron   noisette       vert       bleu
## 0.25048691 0.08332116 0.14878530 0.30656555

sum(apply(L,1,distsq)*r)#somme des carres des ecarts ponderes

## [1] 0.2335977

sum(apply(L,1,distsq))*sum(K) #chi2

## [1] 467.1821
```