

AFC

Yanice

20 novembre 2016

Contents

Intro - questions AFC	2
I. Visualisation des données	2
II. Construction du Tableau de contingence	3
II.1. Tableau de contingence (effectifs observées)	3
II.2. Tableau des fréquences relatives	4
II. Test du chi2 de contingence	4
II.1. Principe du test	4
II.2. Table de contingence théorique	5
II.3. Expression de l'inertie du nuage des individus	6
III. Représentations graphique de la table de contingence (effectifs observés)	6
IV.1. Score à priori (colonnes=cheveux)	10
V. Score optimum	10
V.1. AFC	10
Sortie AFC	11
VI. Représentations graphiques	12
V.2. Représentation simultannée des lignes et colonnes	13
VII. Aides à l'interprétation : inertia.dudi (pondérations non uniformes)	14
VII.1. Décomposition de l'inertie totale	14
VII.2. Contribution absolue des lignes (colonnes)	15
VII.3. Contribution relative des lignes (resp. colonnes)	15
Annexes : Profil lignes et profils colonnes : prop.table	16
Différences AFC ACP	17

Intro - questions AFC

- Disntace utilisé en AFC : CHi2
- Son apport ? Tableau de contingence + **échantillon représentatif (ce qu'on observe peut etre appliquée a la population contrairement a en ACP)**.

-** Poids profil ligne et colonnes : fréquences d'apparition des modalités dans la population**; Cela signifie que les moyennes sont des moyennes pondérées et pareil pour les variances.

- Centre de gravité des profils lignes = Quel est le profil ligne moyen? : fréquence des modalités données en colonné f.j
 - Déterminer l'intertie puis multiplier par le nombre d'individus puis afc
 - Quel ajustement lorsque les modalités ont les mêmes valeurs ou les mêmes distances ? On les regroupe, ça ne change rien
 - Que faire en plus du Chi2 pour manipuler les données ? S'assurer que le jeu de données représente bien la représentation.
 - Que faire si toutes les valeurs propres sont égales à l'unité ? On a au moins une dichotomie en quelque part. On sait quel modalité lignes associer à quelle modalité colonne.
 - Les valeurs propres proches de 1 traduisent une forte liaison entre les lignes et les colonnes.
 - Warning dans le test de chi2 : il ya de modalités dont les effectifs sont inférieurs à 5 (Trop, on ne peut pas interpréter)
 - Lorsque les variables présentent de nombreuses modalités, il est difficile d'extraire une info pertinente si on se contente d'observer le tableau de données. La technique de l'AFC est là pour pallier cette déficience.
- # Analyse Factorielle des Correspondances (AFC)

Considérons deux variables **qualitatives** V et W qui ont respectivement n et p modalités mesurées pour N individus. L'AFC consiste à s'intéresser à l'existence de relations entre les modalités des deux variables.

L'AFC permet de définir pour une table de contingence un score sur les colonnes tel que les scores moyens des lignes (obtenus en utilisant les fréquences des tableaux) soient les plus séparés possibles, au sens de la variance de ses scores moyens. Et inversement (on peut partir des colonnes plutôt que des fréquences des lignes).

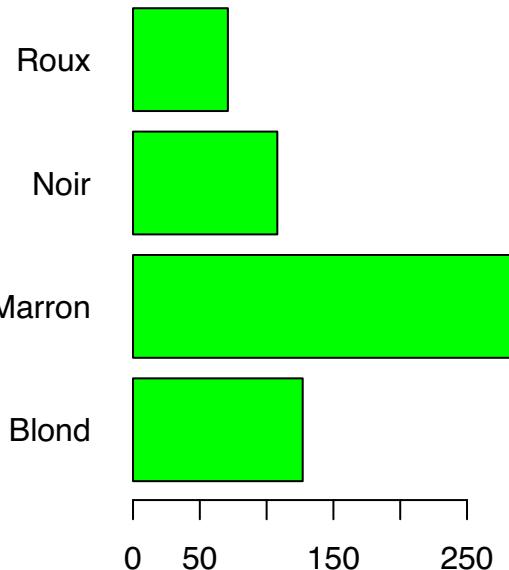
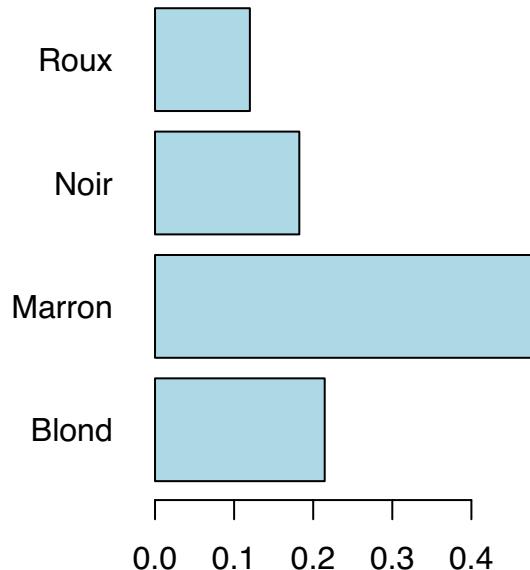
On va travailler sur la table de contingence. Analyser score à priori et score optimal. on va chercher une combinaison linéaire des modalités de la variable cheveux qui sépare le plus possible en moyenne de la couleur des yeux. score ou les coeff des combi linéaire de la couleur des yeux séparent le plus possible de la couleur des cheveux. Combi linéaire de modélisation pour optimiser la position de la couleur des yeux.

```
objet=
```

I. Visualisation des données

```
par(mfrow=c(1,2))

barplot(summary(objet[,1]),main="Fréquence absolue colonne1",col="green",horiz=TRUE, las=1)
barplot(summary(objet[,1])/length(objet[,1]),horiz=TRUE,las=1,main="fréquence relatives",col="lightblue")
```

Fréquence absolue colonne1**fréquence relatives**

II. Construction du Tableau de contingence

II.1. Tableau de contingence (effectifs observées)

Tableau croisant les modalités des 2 variables. Il contient les effectifs obtenu pour chaque couple de modalité. C'est un tableau croisé qui ventile la population entre les modalités des 2 variables qualitatives.

```
#Renvoie une table
(tc=table(yeux,cheveux))
```

```
##          cheveux
## yeux      Blond Marron Noir Roux
##   Bleu      94    84   20   17
##   Marron     7   119   68   26
##   Noisette   10    54   15   14
##   Vert       16    29    5   14
```

```
#Manipulation pour avoir un data.frame
(dftc=data.frame(unclass(tc)))
```

```
##          Blond Marron Noir Roux
## Bleu      94    84   20   17
## Marron     7   119   68   26
```

```
## Noisette    10     54     15     14
## Vert       16     29      5     14
```

Théoriquement :

- n_{ij} : effectif pour le couple de modalité (i,j)
- $n_{i\cdot}$: somme sur les colonnes
- $n_{\cdot j}$ somme des effectifs des lignes

Tout individu présente 1 unique modalité de chaque variable. Aucune modalité non nulle, sinon supprimé.

```
#Nombre total d'individu
n=sum(tc)
#Nombre de modalité pour la variable 1 (en ligne)
I=nrow(tc)
#Nombre de modalité pour la variable 2 (en colonne)
J=ncol(tc)
```

II.2. Tableau des fréquences relatives

```
freqR=tc/n
```

II. Test du chi2 de contingence

II.1. Principe du test

L'objectif est de déterminer si il existe un lien ou non entre les 2 variables.

H_0 : Indépendance entre les variables A et B contre H_1 : Présence d'un lien entre les deux variables.

- Table de contingence théorique (contenant les effectifs théoriques sous H_0) :

La fréquence théorique est définie par :

$$\mathbb{P}(A = i \cap B = j) = \mathbb{P}(A = i) \times (B = j) = \frac{n_{i\cdot}}{n} \times \frac{n_{\cdot j}}{n}$$

L'effectif théorique est donc

$$ET = n \times \mathbb{P}(A = i \cap B = j) = \frac{n_{i\cdot} \cdot n_{\cdot j}}{n}$$

- On compare les valeurs de la table de contingence théorique avec la table de contingence observée (tc ici).

$$Chi2 = \sum \sum \frac{(EO - ET)^2}{ET} = n \times \sum \sum \frac{(FO - ET)^2}{ET} \text{ avec } E = n \cdot F$$

V coefficient de cramer suit une Chi2(I-1, J-1) degrés de liberté.



```

testchi2=chisq.test(tc)
testchi2

##
## Pearson's Chi-squared test
##
## data: tc
## X-squared = 138.29, df = 9, p-value < 2.2e-16

#ou chisq.test(v1,v2)

#statistique de test
chi2=as.numeric(testchi2$statistic)
cramer=sqrt(chi2/(n*min(I-1,J-1)))

```

Il faudrait comparé la valeur théorique du chi2(I-1,J-1) avec la valeur trouvé par le test et si la valeur trouvé par le test Chi2 dépasse la valeur théorique, on rejette l'hypothèse d'indépendance.

Interprétation :

Chi2=0 ==> indépendance entre deux variables.

Chi2 petit ==> EO presques identiques à ET. Les variables sont peut liées entre elles. Chi2 grand ==> EO diff de ET. Les deux variables sont liées entre elles.

$p - \text{valeur} < \alpha$ on rejette H0: Il existe un lien entre les 2 variables qualitatives.

II.2.Table de contingence théorique

$$\frac{n i . n . j}{n}$$

```

# Table de contingence théorique
testchi2$expected

```

```

##           cheveux
## yeux      Blond   Marron   Noir    Roux
##   Bleu     46.12331 103.86824 39.22297 25.785473
##   Marron   47.19595 106.28378 40.13514 26.385135
##   Noisette 19.95101  44.92905 16.96622 11.153716
##   Vert     13.72973  30.91892 11.67568  7.675676

```

```

#Retrouver la table de contingence théorique
outer(margin.table(tc,1),margin.table(tc,2))/sum(tc)

```

```

##           cheveux
## yeux      Blond   Marron   Noir    Roux
##   Bleu     46.12331 103.86824 39.22297 25.785473
##   Marron   47.19595 106.28378 40.13514 26.385135
##   Noisette 19.95101  44.92905 16.96622 11.153716
##   Vert     13.72973  30.91892 11.67568  7.675676

```

```
#outer : multiplication de 2 tables
#margin.table : ni. et nj.
```

II.3. Expression de l'inertie du nuage des individus

Inertie du nuage des individus par rapport au centre de gravité.

$$\chi^2(x, k) = \frac{\left(\frac{1}{2}\right)^{k/2} x^{k-1} e^{-x/2}}{\Gamma(\frac{k}{2})}$$

$$Ig = \frac{\text{Chi2}(I-1, J-1)}{N}$$

III. Représentations graphique de la table de contingence (effectifs observés)

balloonplot

```
library(gplots)
```

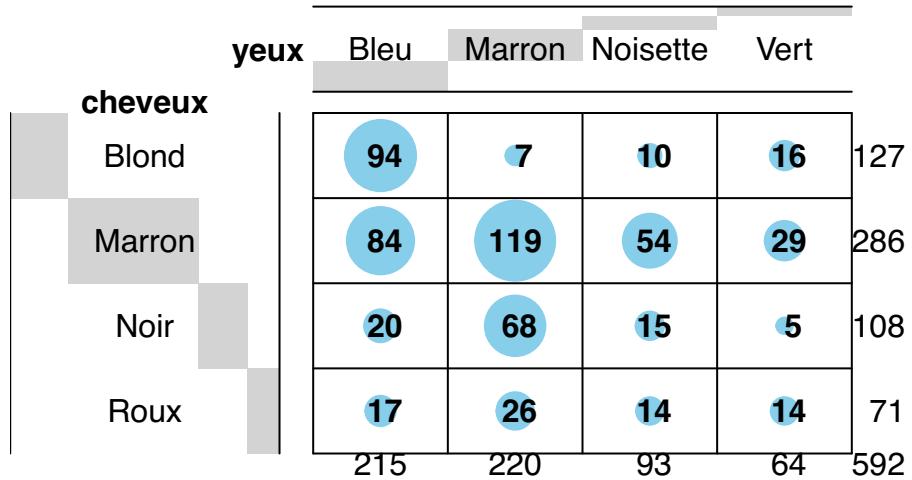
```
## Warning: package 'gplots' was built under R version 3.3.2
```

```
##
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':
##      lowess
```

```
balloonplot(tc)#si ce n'est pas sous le format table (as.table(tc))
```

Balloon Plot for x by y. Area is proportional to Freq.

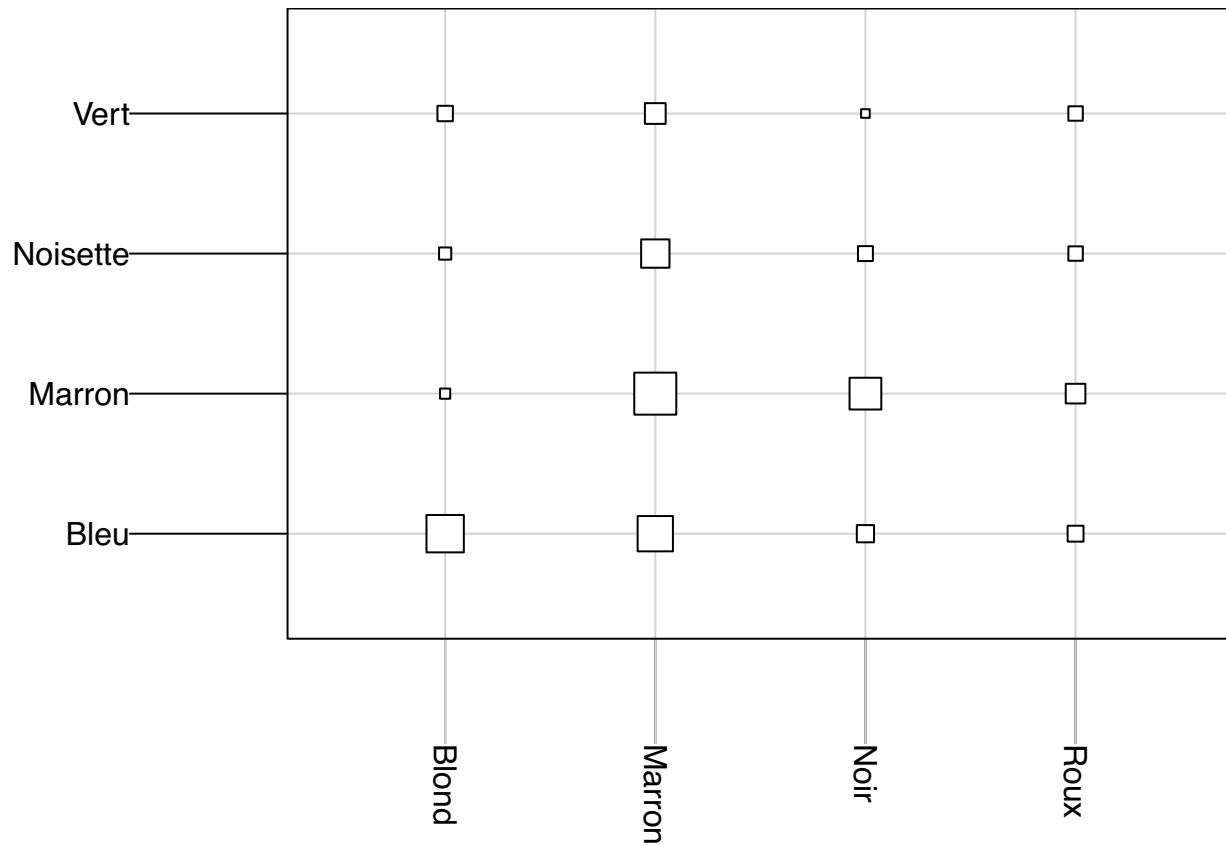


La surface des symboles est proportionnelle aux effectifs.

La surface grise représente les effectifs par variables.

table.cont

```
library(ade4)
table.cont(tc, row.labels = rownames(tc), col.labels=colnames(tc))
```

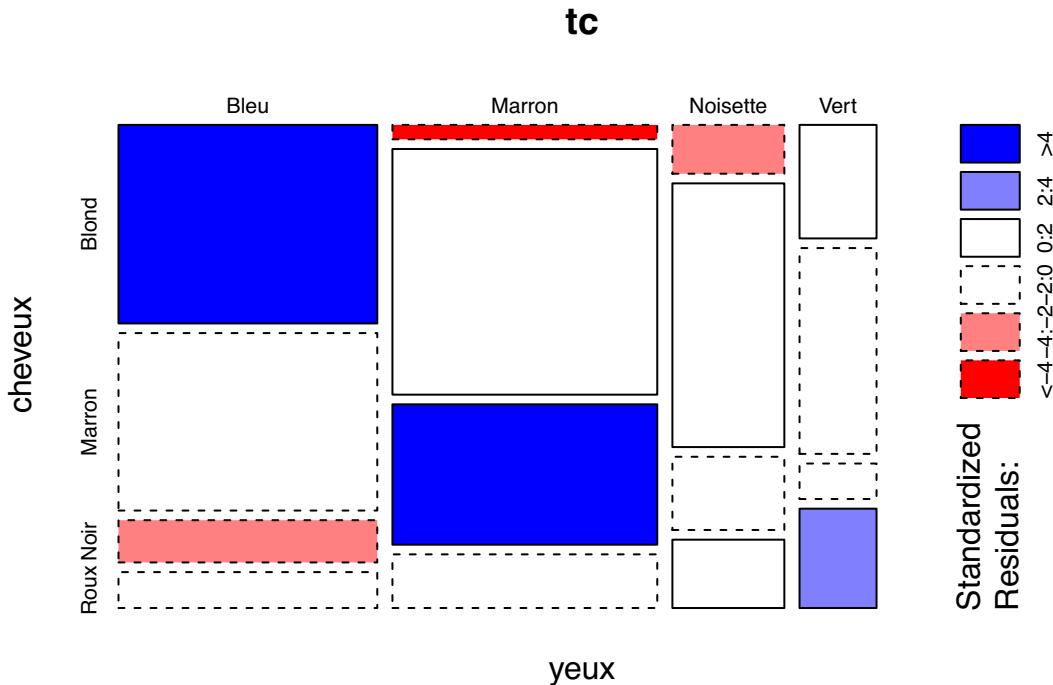


Principe : utiliser de symboles dont la surface est proportionnelle aux effectifs.

mosaicplot

Issu du test du chi2 de contingence. Permet de visualiser les écarts par rapport à l'hypothèse d'indépendance. (résidus) Permet de mettre en évidence les liens les eccarts les plus importants.

```
mosaicplot(tc, shade=TRUE)
```



- Se base sur $\frac{EO-ET}{\sqrt{ET}}$ qui sont les résidus standardisés. H0 : Il n'y a pas de liens entre les deux variables. Elles sont indépendantes.
- Mosaic coloré : écart entre observée et théorique important == chi 2 augmente et plus chi2 grand plus j'ai une relation entre mes deux variables.
- Bleu : Plus d'individus observé qu'attendu sous l'hypothèse d'absence de liens entre variables.(Excédent).
Surreprésentativité des effectifs observés par rapport aux théoriques.
- Rouge : Moins d'individus qu'attendu sous l'hypothèse d'absence de liens entre les variables.(Déficit).
Sous représentativité des effectifs observés par rapport aux théoriques.
- Table tte blanche: peu de chance que les variables soient liées.
- Bien quand on a de petit jeux de données
- La surface d'un élément est proportionnelle aux effectifs contenus dans la table de contingence.
- Lorsque les variables présentent de nombreuses modalités, il est difficile d'extraire une info pertinente si on se contente d'observer le tableau de données. La technique de l'AFC est là pour pallier cette déficience. #IV. Scoring

On veut essayer de trouver des similitudes entre lignes, colonnes, lignes et colonnes. Il faut donc transformer l'info qualitative contenue dans les données originales en une info quantitative, via la construction d'un score numérique.

IV.1. Score à priori (colonnes=cheveux)

On effectue un score a priori, si on semble connaitre bien nos données et voir qu'il ya clairement deux types de données opposé, ex: cheveux foncés et cheveux clairs. Sinon, on va au score optimum.

```
#Affectation d'un score a priori sur les colonnes
scoreApriori=c(1,-1,-1,1)
names(scoreApriori)=colnames(cheveux)
```

Score moyen pour les lignes

```
freqObservee=apply(dftc,1, function(x) x/sum(x));freqObservee
```

```
##           Bleu      Marron     Noisette      Vert
## Blond  0.43720930 0.03181818 0.1075269 0.250000
## Marron 0.39069767 0.54090909 0.5806452 0.453125
## Noir   0.09302326 0.30909091 0.1612903 0.078125
## Roux   0.07906977 0.11818182 0.1505376 0.218750
```

```
scoreMoyenLigne=apply(t(freqObservee),1,function(x) sum(x*scoreApriori))
scoreMoyenLigne
```

```
##           Bleu      Marron     Noisette      Vert
## 0.03255814 -0.70000000 -0.48387097 -0.06250000
```

Interprétation

-1 : cheveux foncés 1 : Cheveux clairs Donc MArron ayant un score à -0.70, les cheveux foncés dominent dans la population des gens ayant les yeux Marrons.

V. Score optimum

V.1. AFC

- Fréquences conjointes : $p_{ij} = \frac{n_{ij}}{n}$
- Fréquences marginales : $p_{i\cdot} = \frac{n_{i\cdot}}{n}$ et $p_{\cdot j} = \frac{n_{\cdot j}}{n}$
- P : tableau des p_{ij} et $P_0 = P - p_{i\cdot} * p_{\cdot j}$ le tableau centré
- DI et DJ les matrices diagonales des fréquences marginales

```
library(ade4)
afc=dudi.coa(dftc,scannf=FALSE,nf=ncol(dftc)-1)
```

On a $\min(\text{modalité}-1, \text{variable}-1)=3$ qui est le nombre de valeurs propres total de l'axe et le rang de la matrice analysée.

Le probleme de l'ACP des profils lignes et colonnes sont duaux.

$\lambda_1 \leq 1$

Choix des axes

La valeur propre associée à un axe mesure la significativité de la dépendance des profils projetés sur cet axe.



Sortie AFC

```



```

```
#centrés
round(sum(afc$co$Comp1*afc$cw),2)

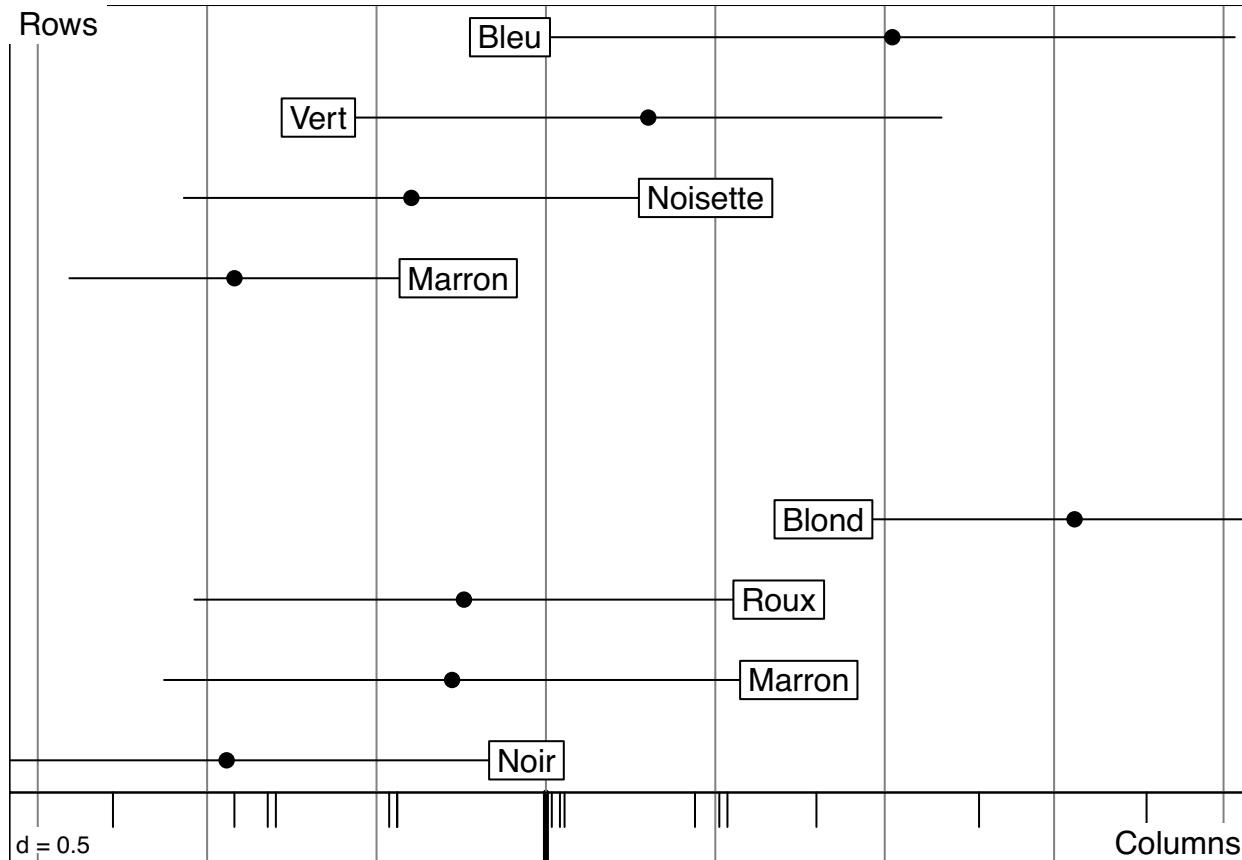
#Liens Chi2 et Inertie totale : It=Chi2/n
testchi2$statistic/n
sum(afc$eig)
```

li : Nouveaux individus centré (pondérés), de variance λ et de covariances nulles.

VI. Représentations graphiques

Lorsque je veux représenter qu'un seul axe, j'utilise la fonction score. score de mon analyse représente le 1er axe qui est l'axe qui a la plus grande inertie. score(ac,xax=1)

```
par(mfrow=c(1,1))
score(afc)
```



Elle contient l'info contenue sur les lignes(en haut) et les colonnes (en bas). En horizontal le premier facteur de l'analyse; le trait plus foncé représente le 0. Chaque point est la position de la modalité sur l'axe 1. C'est comme si je projette chaque point sur l'axe 1 et j'aurai sa position, on les a séparé pour pouvoir interpréter, en réalité il se trouve tous sur le même axe. La lisibilité sur un trait n'est pas top, on sépare les lignes des colonnes. les lignes verticales sont là pour faciliter la lecture (méridiennes). si plusieurs points appartiennent aux mêmes méridiens => variables liées entre elles. roux et marron sont liés à noisette. blond est le plus loin donc qui se distingue le plus = on lui affecte le plus fréquemment les yeux bleus.

Pas de plan quand on a qu'un seul axe...

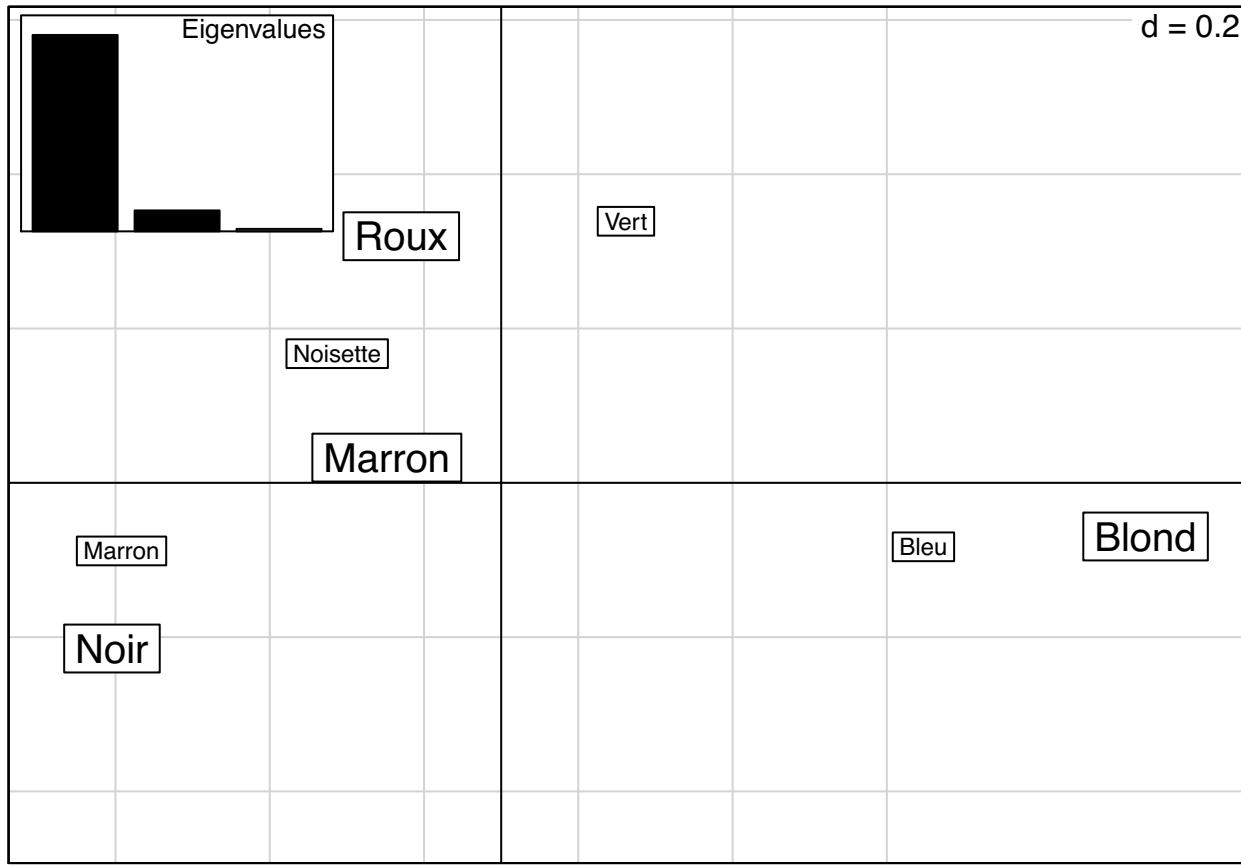
```
summary(afc)
```

```
## Class: coa dudi
## Call: dudi.coa(df = dftc, scannf = FALSE, nf = ncol(dftc) - 1)
##
## Total inertia: 0.2336
##
## Eigenvalues:
##      Ax1      Ax2      Ax3
## 0.208773 0.022227 0.002598
##
## Projected inertia (%):
##      Ax1      Ax2      Ax3
## 89.373   9.515   1.112
##
## Cumulative projected inertia (%):
##      Ax1    Ax1:2    Ax1:3
## 89.37   98.89   100.00
```

J'ai 89% de l'inertie totale contenue sur l'axe1 donc de l'info totale, l'axe2 est quasi négligeable. J'ai quand même de l'info mais très faible, je vais la regrouper mais il ne faut pas que je lui en donne trop d'importance. On est dans la population écossaise des années 70 :)

V.2. Représentation simultanée des lignes et colonnes

```
scatter(afc)
```



Représentation simultanées des lignes et colonnes. Couleurs des cheveux est liés aux yeux... La taille : faciliter la lecture pas plus de poids. MArron apparaît deux fois pour deux choses différente. verticale : Roux ressort un peu avec une tendance pour les yeux vert.

VII. Aides à l'interprétation : `inertia.dudi` (pondérations non uniformes)

Importantes dans les analyses à pondérations non uniformes (AFC par ex). Les statistiques d'inertie se trouvent dans `inertia.dudi`. Pour les Analyses à pondération uniforme, les résultats sont redondants avec les cartes factorielles.

```
INERTIA = inertia.dudi(afc, row.inertia=TRUE, col.inertia=TRUE)
```

VII.1. Décomposition de l'inertie totale

Inertie totale

$$I_T = \sum_{i=1}^r \lambda_k$$

r : rang de la matrice diagonalisée.

```
IT=sum(afc$eig)
```

Inertie relative du vecteur principal de rang k

$$\frac{\lambda_k}{I_T}$$

```
#Inertie relative : contribution des axes à l'inertie totale.  
IR=afc$eig/IT  
IR_pourcentage=IR*100
```

Sortie inertia.dudi\$TOT

```
INERTIA$TOT

#VAleurs propres de 1 à r
INERTIA$TOT$inertia

#Valeurs propres cumulés
INERTIA$TOT$cum

#Inertie relative cumulé du nuage sur les différentes dimensions
INERTIA$TOT$ratio
cumsum(IR_pourcentage)/100 #Pareil
```

VII.2. Contribution absolue des lignes (colonnes)

Contributions des individus à l'inertie de chaque axe.

```
#Méthode 1: avec inertia.dudi
INERTIA$row.abs#multiplié par 1000 pour faciliter la lecture
```

```
##          Axis1 Axis2 Axis3
## Bleu      5213   1124    31
## Marron    4312   1304   668
## Noisette   340   1980   6109
## Vert       135   5591   3192
```

```
#Méthode 2 : avec dudi.coa
afc$li[,1]*afc$li[,1]*afc$lw/afc$eig[1]
```

```
##        Bleu     Marron     Noisette      Vert
## 0.52128445 0.43115744 0.03400961 0.01354851
```

VII.3. Contribution relative des lignes (resp. colonnes)

CE sont des carrés de cosinus.

```
#Contribution relatives
```

```
#Méthode 1
```

```
INERTIA$row.rel
```

```
##          Axis1 Axis2 Axis3 con.tra
## Bleu      9775   -224    -1    4766
## Marron    -9670   -311     19    3985
## Noisette  -5424   3363  -1213     560
## Vert       1759   7726    516    689
```

```
#Méthode 2 : pour la 3eme ligne :
```

```
(afc$li[3, ] * afc$li[3, ])/(sum(afc$tab[3, ] * afc$tab[3, ] * afc$cw))
```

```
##          Axis1     Axis2     Axis3
## Noisette 0.5424487 0.3362865 0.1212648
```

- Indique si les individus sont bien représentés. Si ils sont >5000, ils sont bien représentés, donc on peut le commenter. Le signe moins indique si il est à gauche ou à droite de l'origine (axe).
- Les résultats sont multipliés par 1000
- La dernière colonne contient la contribution à la trace (voir page 14 td5 prof)

```
#Contribution relative cumulées
```

```
INERTIA$row.cum
```

```
##          Axis1 Axis2 Axis3 remain
## Bleu      9775  9999 10000     0
## Marron    9670  9981 10000     0
## Noisette  5424  8787 10000     0
## Vert       1759  9484 10000     0
```

Ce dernier tableau contient pour chaque ligne la somme des contributions relatives. (voir td prof)

Annexes : Profil lignes et profils colonnes : prop.table

Ce sont les fréquences conditionnelles associées aux profils lignes et colonnes profile ligne :

$$f_{j|i} = \frac{nij}{ni}$$

```
ProfLignes=prop.table(tc,1)
ProfColonnes=prop.table(tc,2)
#On vérifie également que la somme vaut 1
colSums(ProfColonnes)
```

```
## Blond Marron Noir Roux
##      1      1      1      1
```

Profil ligne : En sachant la ligne tu trouves la colonne

Différences AFC ACP

Pour une AFC, l'analyse des profils lignes centrés est équivalente à l'analyse des profils lignes **non centrées* si dans cette première nous retirons l'axe factoriel donné par le centre de gravité des lignes.

l'analyse des profils lignes centrés en retirant l'axe factoriel donné par le centre de gravité des lignes est équivalente à l'analyse des profils lignes **non centrées* contrairement à l'ACP.

AFC

L'analyse factorielle des correspondances (AFC ou CA pour *correspondence analysis* en anglais) est une extension de l'analyse en composantes principales pour analyser l'association entre deux variables qualitatives (ou catégorielles).

L'AFC permet de résumer et de visualiser l'information contenue dans le *tableau de contingence* formé par les deux variables catégorielles. Le tableau de contingence contient les fréquences formées par les deux variables.

L'AFC retourne les coordonnées des éléments des colonnes et des lignes du tableau de contingence. Ces coordonnées permettent de visualiser graphiquement l'association entre les éléments de lignes et de colonnes dans un graphique à deux dimensions.

Lors de l'analyse d'un tableau de contingence, une question typique est de savoir si certains éléments lignes sont associés à certains éléments colonnes. L'analyse factorielle par correspondance est une approche géométrique pour visualiser les lignes et les colonnes d'une table de contingence dans un graphique en nuage de points, de sorte que les positions des points lignes et celles des points colonnes correspondent à leurs associations dans le tableau.

Table de contingence

La table engendrée par le croisement de deux variables qualitatives s'appelle une table de contingence observée. Il est important de rappeler que :

- tout individu présente une modalité et une seule de chaque variable .
- chaque modalité doit avoir été observée au moins une fois, sinon elle est supprimée.

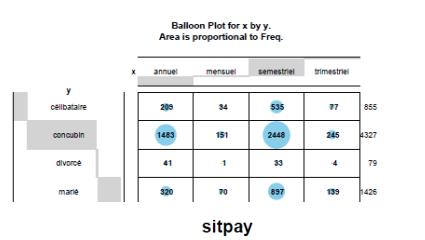
NB: $n_{i,j}$ = nb d'individu ayant comme modalité I pour la variable1 et la modalité J pour la variable2.

Variable2 \ Variable1	Modalité1	Modalité J
Modalité1	$n_{1,1}$		
....		
Modalité I			$n_{I,J}$

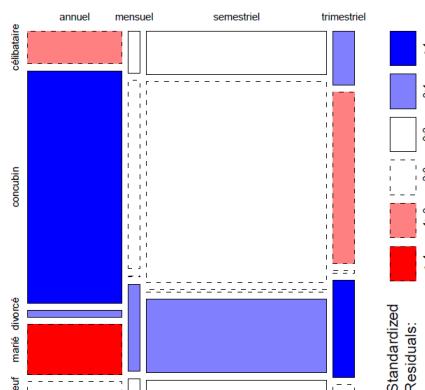
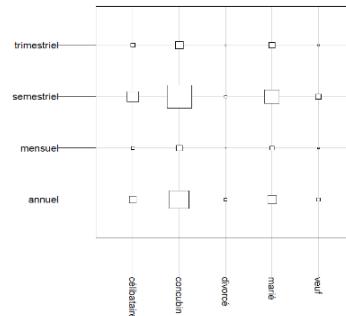
Représentation de la table de contingence X : Le principe est d'utiliser des symboles dont la surface est proportionnelle aux effectifs

`library(gplots)`

`balloonplot(as.table(X))`



`table.cont(X, row.labels = rownames(X), col.labels = colnames(X), csize = 2)`



La fonction mosaicplot permet de mettre en évidence les liens les plus importants.

`mosaicplot(X, shade = TRUE)`

$n <- sum(X)$: nb total d'individus n
 $I <- nrow(X)$: nb de modalités pour la variable ligne
 $J <- ncol(X)$: nb de modalités pour la variable colonne

Lien avec le test d'indépendance du Chi2

Le test du Chi-2 d'indépendance entre deux variables est caractérisé par les deux hypothèses :

- H_0 : les deux variables sont indépendantes
- H_1 : les deux variables sont liées

La statistique du test est la suivante : $\chi^2 = \sum_{i=1}^I \sum_{j=1}^J \frac{(n_{ij} - \frac{n_i n_j}{n})^2}{\frac{n_i n_j}{n}}$

Elle tend vers une loi du χ^2 à $(I-1)(J-1)$ degrés de liberté.

n : nb total d'individus

I : nb de lignes

J : nb de colonnes

Règle de décision :

- ❖ Si $p_{value} < \alpha$ (p-value est très faible), on H_0 . Les variables sont liées.
- ❖ Si $\chi^2_{obs} > \chi^2_{théorique (I-1)(J-1)ddi}$, on H_0 . Les variables sont liées. Il est alors intéressant d'explorer la structure de cette relation.

Statistique de test du Chi2 de la table de contingence X sur R :

```
reschi <- chisq.test(X)
```

```
reschi$statistic
```

Sous l'hypothèse nulle H_0 , $P(V2 = j | V1 = i) = P(V1 = i) * P(V2 = j)$.

Ainsi, sous H_0 , la fréquence théorique est égale à : $f_{i,j} = \frac{n_i}{n} * \frac{n_j}{n}$

On en déduit la table des effectifs théoriques (qui serait observée sous H_0), en conservant les effectifs marginaux observés.

$\frac{n_i * n_j}{n}$ avec :

```
reschi <- chisq.test(X)
```

OU

```
outer(margin.table(X, 1), margin.table(X, 2))/sum(X)
```

Conclusion : Que la liaison entre les 2 variables soit statistiquement significative ou non, on peut explorer la structure du tableau plus en détail. **Lorsque les variables présentent de nombreuses modalités**, il est **difficile d'extraire une information** pertinente si on se contente d'observer le tableau de données. La technique de l'Analyse Factorielle des Correspondances (**AFC**) est là pour pallier cette déficience.

Lorsqu'on connaît moins bien les données, l'AFC permet de regrouper les modalités qui ont des comportements similaires.

Résultats d'une AFC

Les résultats de l'AFC de la table de contingence permettent d'étudier le lien entre la variable1 et la variable2.

```
X <- as.data.frame(X)
```

```
AFC <- dudi.coa(X, scannf = F, nf = 3)
```

Les coordonnées des lignes dites axes principaux s'obtiennent par : `AFC$li`

Les coordonnées des colonnes dites composantes principales s'obtiennent par : `AFC$co`

Valeurs propres et Inertie

L'examen des valeurs propres permet de déterminer le nombre d'axes principaux à considérer. Les valeurs propres correspondent à la quantité d'informations retenue par chaque axe.

Valeurs propres

Le rang de la matrice analysée (=nb d'axes) est donné par $\min(I - 1 ; J - 1)$ soit : $\min(I - 1, J - 1)$ ou $\text{afc\$rank}$

Les valeurs propres issus de cette diagonalisation sont : $\text{afc\$eig}$

$$\underline{\text{Rappel du lien entre le Chi2 et l'inertie totale}} \quad I_T = \frac{\chi^2}{n}$$

```
reschi <- chisq.test(X)
                           ou
reschi$statistic
reschi$statistic/sum(X)
```

Décomposition de l'inertie totale

La somme des valeurs propres est égale à l'inertie totale du nuage de points : $I_T = \sum_{k=1}^r \lambda_k$ où r représente le rang de la matrice diagonalisée.

L'inertie relative du vecteur principal de rang k : $I_k = \frac{\lambda_k}{I_T}$

Sur R :

- ✓ inertie totale : $\text{sum(afc\$eig)}$
- ✓ inertie relative à chaque axe : $\text{afc\$eig} / \text{sum(afc\$eig)}$
- ✓ inertie relative cumulée : $100 * \text{cumsum(afc\$eig} / \text{sum(afc\$eig))}$

En résumé : $I_T = \sum_{k=1}^r \lambda_k = \frac{\chi^2}{n}$

Choix des axes

- ❖ Si **NRH0** (Non rejet hypothèse d'indépendance : les 2 variables sont indépendantes) : Méthodes de l'ACP

Critère de Kaiser : on ne garde que les axes qui ont des valeurs propres supérieures à la valeur propre moyenne.

Critère du coude ou d'infexion : sur l'histogramme des valeurs propres, on observe un décrochement (coude) suivi d'une décroissance régulière. On sélectionne les axes avant le décrochement.

- ❖ Si **RHO** (Les 2 variables sont liées) : Test d'indépendance des axes. On grade les **p-value < α** :

- Avec : α = seuil de significativité fixé et p-value = $\text{pchisq}()$
- $\text{pchisq(afc\$eig*sum(X), df=(I-1)*(J-1), lower.tail = FALSE)}$

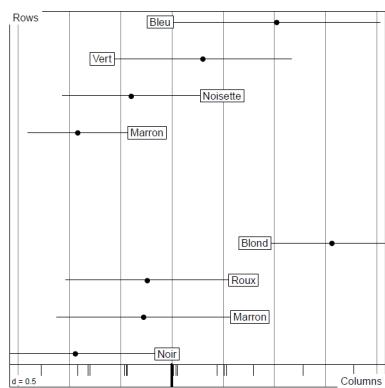
Représentation graphique



scatter(AFC)

La distance entre les points lignes **ou** entre les points colonnes donne une mesure de leur similitude (ou dissemblance). Les points lignes avec un profil similaire sont proches sur le graphique. Il en va de même pour les points colonnes.

Dans ce graphique, les lignes sont représentées par des petits rectangles et les colonnes par des grands rectangles.



score(AFC)

Ces graphiques montrent tout deux que :

- La lignes Vert et Noisettes sont associés le plus aux colonnes Roux et Marron.
- La ligne Bleu est associé le plus à la colonne Blond.
- La ligne Marron est associé le plus à la colonne Noir.

WARNING

- ❖ Le graphique ci-dessus (scatter) représente une *analyse symétrique* montrant les profils lignes et colonnes simultanément dans un espace commun. Dans ce cas, seule la distance entre les points lignes ou la distance entre les points colonnes peut être vraiment interprétée.
- ❖ La distance entre les points lignes et les points colonnes n'a pas de sens dans l'absolue ! Vous ne pouvez faire que des observations générales.

→ TABLE DE CONTINGENCE : `table()`
 $c \leftarrow \text{table}(v1, v2)$

↳ donne un tableau croisé qui renvoie les entrées entre les modèles des deux variables qualitatives

↳ on le transforme en data frame :

 $dfc \leftarrow \text{data.frame}(c)$

> Score

→ SCORE A PRIORI : on se base suivant sur une opposition naturelle des modèles

↳ Ex des ears :

 $\text{scorecheveux} \leftarrow c(1, -1, -1, 1)$ opposition force/clair

 $\text{names(scorecheveux)} \leftarrow \text{colnames(couleurs)}$
 scorecheveux

Blond	Narron	Noir	Rou
-------	--------	------	-----

1	-1	-1	1
---	----	----	---

Par chaque ligne de la table des contingence (couleur des yeux), on obtient la fréquence observée

 $dfcouleurs \leftarrow \text{data.frame}(couleurs(couleurs))$
 $dfcouleurs[["Bleu",]] / \text{sum}(dfcouleurs[["Bleu",]])$ fréquence des yeux Bleus par couleur de cheveux

Blond	Narron	Noir.	Rou
-------	--------	-------	-----

Bleu	0,44	0,39	0,09	0,08
------	------	------	------	------

Puis on calcule le score moyen par le modèle yeux Bleu

 $\text{yeux.Bleu} \leftarrow dfcouleurs[["Bleu",]] / \text{sum}(dfcouleurs[["Bleu",]])$ fréquences observées des yeux Bleus

Bleu	Blond	Narron	Noir	Rou
------	-------	--------	------	-----

Bleu	0,44	-0,39	-0,09	0,08
------	------	-------	-------	------

$\text{sum}(\text{yeux.Bleu} * \text{scorecheveux})$ ← score moyen par une modéliser (ci yeux Bleu)
 $0,0326$

On peut calculer le score moyen par toutes les couleurs des yeux

 $\text{frequyeux} \leftarrow \text{apply}(dfcouleurs, 1, \text{function}(x) x / \text{sum}(x))$

$t(frequyeux)$ a prend la transposée pour avoir les couleurs des yeux sur les lignes

 $\text{Scoreyeux} \leftarrow \text{apply}(t(frequyeux), 1, \text{function}(x) \text{sum}(x * \text{scorecheveux}))$

Bleu	Narron	Noisette	Vert	Score moyen de -0,7 : <0 donc cheveux foncés
------	--------	----------	------	--

Pour obtenir score moyen d'un modèle :

- 1) on attribue un score à priori
 - 2) par chaque ligne de la table de contingence, on calcule fréquence observée : $\frac{n}{\text{sum}(n)}$
 - 3) puis on obtient le score moyen du modèle en faisant :
- $$\text{sum}(\text{freq observées} * \text{score à priori})$$
- (sum, IV) about →

→ SCORE OPTIMUM

AFC = méthode qui donne un score sur les colonnes tq les scores moyens des lignes soient les plus séparés possibles (et vice versa)

library (ade4)

$ac \leftarrow \text{ade4::coa}(\text{dfcoloris}, \text{scannf} = F, \text{nf} = 3)$

$ac\$c1[, 1]$: scores optimaux par modèles des couleurs des cheveux
(plus généralement : les scores optimaux des modèles en colonne)

rownames($ac\$c1$) : donne les noms des modèles qui sont en colonne dans la table de contingence (ici Bleu Marron Noir Rose)

$ac\$11[, 1]$: scores optimaux par les modèles des couleurs des yeux
(plus généralement : les scores optimaux des modèles en ligne)

rownames($ac\$11[, 1]$) : donne le nom des modèles qui sont en ligne
score par l'interpolation dans la table de contingence (ici Bleu Marron Noir Vert)

- On peut trouver les scores moyens des modèles des couleurs de yeux (resp. des couleurs de cheveux) à partir des scores optimum des modèles des couleurs de cheveux en $ac\$c1[, 1]$ (resp. des modèles des yeux en $ac\$11[, 1]$).

- On peut directement trouver les scores moyens dans :

$ac\$li[, 1]$ par les modèles en ligne

$ac\$co[, 1]$ par les modèles en colonne

DONC

Score optimum	Score moyen
$ac\$c1[, 1]$: score optimum par les modèles de V_1	$ac\$co[, 1]$: score moyen par les modèles de V_1
$ac\$11[, 1]$: score optimum par les modèles de V_2	$ac\$li[, 1]$: score moyen par les modèles de V_2

Score moyen
$ac\$co[, 1]$: score moyen par les modèles de V_1
$ac\$li[, 1]$: score moyen par les modèles de V_2

Table de contingence

	Π_1^1	Π_2^1	Π_3^1
Π_1^2			
Π_2^2			
Π_3^2			

> TABLE DE CONTINGENCE

© Théo Jalabert



- Soit on a déjà le tableau / les vecteurs à partir duquel/ desquel on veut faire la table :

$$c \leftarrow \text{table}(v_1, v_2)$$

- Soit on a les données : ex :

→ mode de règlement : annuel, mensuel, semestriel, trimestriel

→ situation matinale : célibataire, concubin, devenue, marié ou veuf + les chiffres

et on construit nous même :

$$\text{table} \leftarrow \text{matrix}(c(\dots), \text{byrow}=T, \text{ncol}=\dots)$$

$$\text{colnames}(\text{table}) \leftarrow c(\dots)$$

$$\text{rownames}(\text{table}) \leftarrow c(\dots)$$

↳ informations de base :

$n \leftarrow \text{sum}(\text{table})$: nombre total d'individus

$I \leftarrow \text{nrow}(\text{table})$: nombre de modalités par la variable en ligne

$J = \text{ncol}(\text{table})$: nombre de modalités par la variable en colonne

↳ on peut construire le tableau des fréquences relatives : $f_{ij} = \frac{n_{ij}}{n}$

$$\text{freqtable} \leftarrow \text{table}/n$$

$$\text{round}(\text{freqtable}, \text{digits} = \alpha)$$

 ↳ nombre après la virgule

↳ Représentations : library(gplots)

• balloonplot(as.table(table))

• table.cont(table, row.labels = rownames(table), col.labels = colnames(table), csize = 2)

• mosaicplot(table, shade = TRUE)

→ PROFILS LIGNES / COLONNES : Variables V_1 / V_2

• LIGNES : $f_{j|i} = P(V_2=j | V_1=i)$

$$\text{i.e. } f_{j|i} = \frac{n_{ij}}{n_i}$$

$$\text{profLignes} \leftarrow \text{prop.table}(\text{table}, 1)$$

• COLUMNES : $f_{i|j} = P(V_1=i | V_2=j)$

$$\text{i.e. } f_{i|j} = \frac{n_{ij}}{n_j}$$

$$\text{profColonnes} \leftarrow \text{prop.table}(\text{table}, 2)$$