TP 2 : Analyse Factorielle des Correspondances (AFC)

Zaineb Smida

Vous disposez d'une base de données intitulée `assurance.txt`, qui provient d'une compagnie d'assurance et concerne 1776 jeunes conducteurs. Les variables d'intérêt sont les suivantes :

- `FORMULE` : formule du contrat d'assurance (A = `"minimum"`, B = `"moyenne"`, C = `"maximum"`)
- `VALEUR` : valeur marchande du véhicule assuré, codée en 4 classes : (1 = `"faible valeur"`, 2 = `"valeur moyenne"`, 3 = `"valeur élevée"`, 4 = `"valeur très élevée"`)

Packages nécessaires :

```
library("FactoMineR")
library("Factoshiny")
library("RColorBrewer")
```

0.1 Question

Importer le fichier `assurance.txt` sous **R**. Vous pouvez l'appeler `conducteurs`.

> Conseils
>
> Utiliser la fonction `read.table()`. Si la première ligne du fichier .txt contient le nom des variables, ajouter l'option `header = T` pour le préciser. Vous pouvez vérifier que l'importation a été correctement réalisée en utilisant `View(conducteurs)`.

> Solution

```
conducteurs <- read.table("assurance.txt", header = T)
```

0.2 Question

Transformer la variable qualitative `valeur` en `factor` et mettre des labels.

Conseils

Lorsque vous utilisez `table(conducteurs$valeur)`, vous constatez que cette variable est codée avec des entiers allant de 1 à 4. En réalité, 1 correspond à la modalité `faible valeur`, 2 à `valeur moyenne`, 3 à `valeur élevée` et 4 à `valeur très élevée`. La fonction `factor()` permet d'assigner des étiquettes à ces entiers pour la variable valeur. Pour ce faire, utilisez l'option `labels =` suivie d'un vecteur contenant les étiquettes ordonnées selon les entiers.

Solution

```
conducteurs$valeur = factor(conducteurs$valeur,
  labels=c("faible valeur","valeur moyenne","valeur élevée","valeur très élevée"))
```

0.3 Question

Construire la table de contigence des deux variables qualitatives.

Conseils

Vous pouvez utiliser la fonction `table()` sur deux variables qualitatives afin d'obtenir la table de contingence des effectifs.

Solution

```
tabcontin <- table(conducteurs$formule, conducteurs$valeur)
tabcontin
```

```
   faible valeur valeur moyenne valeur élevée valeur très élevée
A            61            388            61                 21
B            64            488           132                 23
C             0            346           118                 74
```

0.4 Question

Afficher les profils lignes et les profils colonnes.

Conseils

Pour obtenir les profils lignes (resp. colonnes), divisez les effectifs (la table créée précédemment) par la somme des lignes (resp. colonnes). Pour cela, utilisez la fonction `prop.table()` avec l'argument `margin = 1` (resp. `margin = 2`).

Solution

```r
tabPL <- prop.table(tabcontin, margin = 1) # profils-lignes
round(tabPL, digits = 2)
```

```
  faible valeur valeur moyenne valeur élevée valeur très élevée
A          0.11           0.73          0.11                0.04
B          0.09           0.69          0.19                0.03
C          0.00           0.64          0.22                0.14
```

```r
tabPC <- prop.table(tabcontin, margin = 2) # profils-colonnes
round(tabPC, digits = 2)
```

```
  faible valeur valeur moyenne valeur élevée valeur très élevée
A          0.49           0.32          0.20                0.18
B          0.51           0.40          0.42                0.19
C          0.00           0.28          0.38                0.63
```

0.5 Question

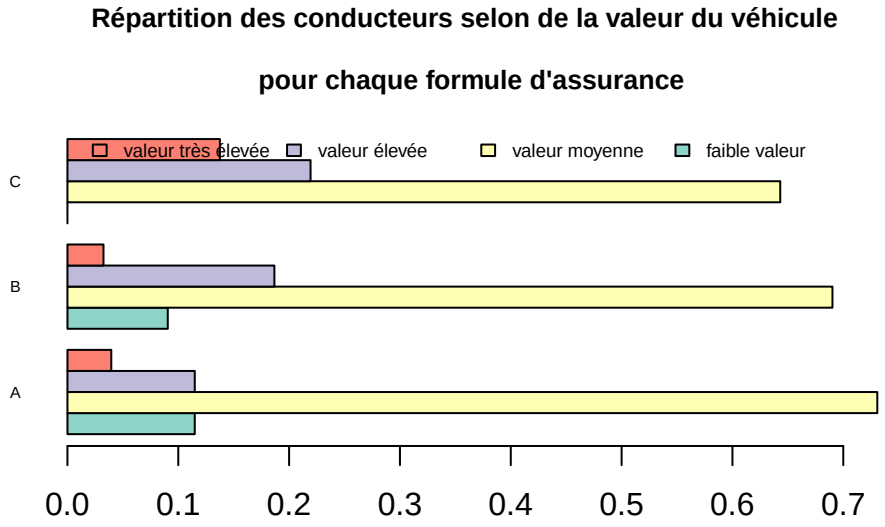
Statistique bivariée : apprécier le lien entre les variables à l'aide d'un outil graphique vu en cours.

Conseils

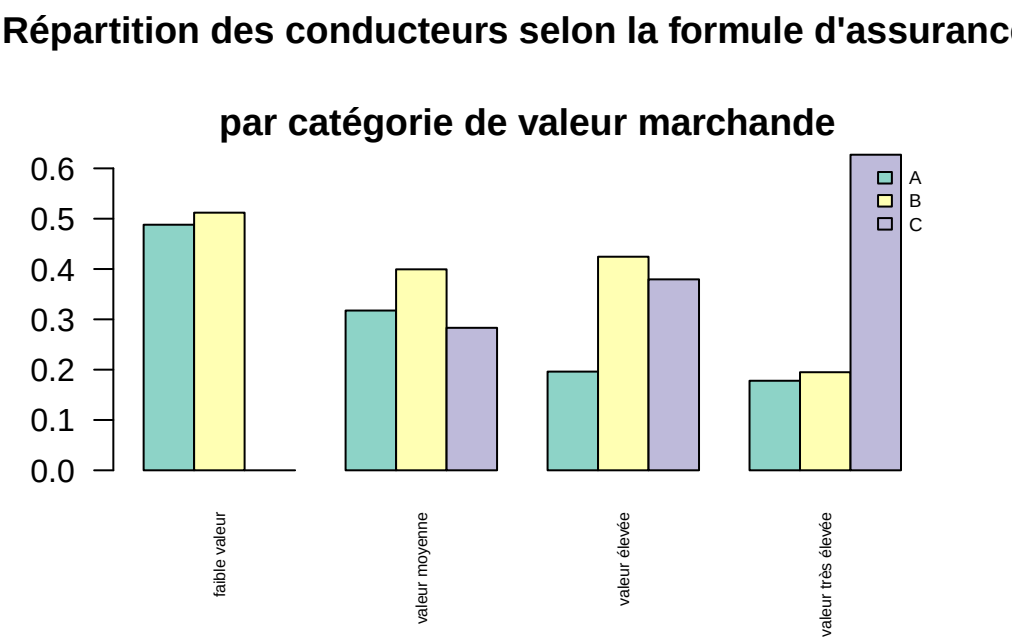
Pour représenter graphiquement les profils lignes ou colonnes, utilisez la fonction `barplot()`, appliquée sur les profils créés précédemment. Il existe plusieurs options pour personnaliser le graphique. Par exemple l'option `beside = T` permet de représenter les barres de manière juxtaposée.

Solution

```
library(RColorBrewer)
# profils lignes
barplot(t(tabPL), beside= T, horiz = T, las = 1, cex.names = 0.5,
         main = "Répartition des conducteurs selon de la valeur du véhicule
        \n pour chaque formule d'assurance", cex.main=0.8, legend.text = T,
         col= RColorBrewer::brewer.pal(4, name= "Set3"),
         args.legend = list(x = "top", ncol = 4, bty = "n", cex = 0.6))
```




```
# profils colonnes
 barplot(tabPC, beside= T, horiz = F, las = 2, cex.names = 0.5,
         main = "Répartition des conducteurs selon la formule d'assurance
         \n par catégorie de valeur marchande",
         legend.text = T, col= brewer.pal(3, name= "Set3"),
         args.legend = list(x = "topright", ncol = 1, bty = "n", cex = 0.6))
```



0.6 Question

Les deux variables sont-elles liées ? Utiliser le test statistique vu en cours permettant de répondre à cette question. Interpréter les sorties de R.

Conseils

Vous pouvez utiliser la fonction `chisq.test()`

Solution

```
resutest <- chisq.test(tabcontin)
resutest
```

```
    Pearson's Chi-squared test

data:  tabcontin
X-squared = 136.37, df = 6, p-value < 2.2e-16
```

```
# La statistique de test du Chi2 = 136.37
# (L-1)(C-1) = 6 degrés de liberté
#la p-valeur < 5%, nous rejetons l'hypothèse nulle.
#les deux variables ne sont donc pas indépendantes.
```

0.7 Question

Réaliser l'étude des écarts et des contributions au χ^2.

Conseils

Vous pouvez extraire les écarts de l'objet retourné par la fonction `chisq.test()` en utilisant la syntaxe `$residuals`.

Solution

```
ecarts <- resutest$residuals
ecarts
```

```
  faible valeur valeur moyenne valeur élevée valeur très élevée
A    3.86475602     1.18436762   -3.31694010        -2.40421750
B    2.01857766     0.06979619    0.73654762        -3.49794572
C   -6.15353484    -1.25664850    2.45094689         6.39840786
```

```
contrib <- ecarts^2
contrib
```

```
  faible valeur valeur moyenne valeur élevée valeur très élevée
A  14.936339059    1.402726650  11.002091609        5.780261769
B   4.074655765    0.004871508   0.542502389       12.235624283
C  37.865990991    1.579165463   6.007140648       40.939623150
```

```r
# Calcul du seuil de contribution
Seuil_contrib <- resutest$statistic / (nrow(tabcontin) * ncol(tabcontin))
# Le seuil de contribution est de 11.36425 dans ce cas

#On repère les 4 plus fortes contributions : (A, faible valeur),
#(C, faible valeur), (B, valeur très élevée), (C, valeur très élevée).

#analyse des écarts :
#pour  (A, faible valeur) : obs > att : sur-représentation des véhicules
#à faible valeur marchande associés à la formule d'assurance C.
#(C, faible valeur) : obs < att : sous-représentation
#(B, valeur très élevée) : obs < att : sous-représentation
#(C, valeur très élevée) : obs > att : sur-représentation
```

0.8 Question

Réaliser l'AFC sur les deux variables. Combien d'axes sont à retenir ? Justifier votre réponse.

Conseils

Pour réaliser une **AFC** sous `R`, chargez le package `FactoMineR` avec la commande `library(FactoMineR)` et utilisez la fonction `CA()`. Pour le moment, nous ne nous intéressons pas à afficher les résultats de l'AFC. Toutefois, vous pouvez utiliser l'argument `graph =` pour activer l'affichage graphique si nécessaire.

Solution

```r
resuafc=CA(tabcontin, graph = F)
# Etape 0 : liaison significative entre les 2 variables
           #(rejet de H0 dans le test du khi2)

# Etape 1 : Choix de la dimension
resuafc$eig
```

```
       eigenvalue percentage of variance cumulative percentage of variance
dim 1 0.071414257              93.00491                           93.00491
dim 2 0.005371213               6.99509                          100.00000
```

```
# on conserve seulement le 1er axe qui explique 93% de l'inertie totale
#selon le critère de la part d'inertie expliquée.
```

0.9 Question

Quels sont les profils ayant fortement contribué à l'apparition du (des) axe(s) retenu(s) ? Justifier votre réponse.

Conseils

On accède aux résultats de l'AFC en utilisant les syntaxes `$row$coord` et `$col$coord` pour les coordonnées, ainsi que `$row$contrib` et `$col$contrib` pour les contributions.

Solution

```
resuafc$row #tous les résultats sur les PL

$coord
       Dim 1       Dim 2
A -0.2320146  0.09243809
B -0.1303887 -0.08272060
C  0.4003431  0.01746997

$contrib
      Dim 1     Dim 2
A 22.537052 47.564299
B  9.477019 50.714422
C 67.985929  1.721278

$cos2
      Dim 1       Dim 2
A 0.8630104 0.136989558
B 0.7130213 0.286978687
C 0.9980994 0.001900611

$inertia
[1] 0.018649448 0.009491922 0.048644099

resuafc$col #tous les résultats sur les PC

$coord
```

```
                         Dim 1       Dim 2
faible valeur      -0.67349900  0.03761621
valeur moyenne     -0.04633951  0.01722818
valeur élevée       0.19102751 -0.14122655
valeur très élevée  0.68987033  0.15395419

$contrib
                        Dim 1     Dim 2
faible valeur      44.704995  1.854150
valeur moyenne      2.068932  3.802198
valeur élevée       8.947966 65.024614
valeur très élevée 44.278108 29.319039

$cos2
                        Dim 1       Dim 2
faible valeur      0.9968903 0.003109736
valeur moyenne     0.8785635 0.121436477
valeur élevée      0.6465947 0.353405280
valeur très élevée 0.9525604 0.047439557

$inertia
[1] 0.032025330 0.001681736 0.009882733 0.033195670
```

```
resuafc$row$coord[,1] #les nouvelles coodonnées des PL
```

```
         A          B          C
-0.2320146 -0.1303887  0.4003431
```

```
resuafc$col$coord[,1] #les nouvelles coodonnées des PL
```

```
     faible valeur     valeur moyenne      valeur élevée valeur très élevée
       -0.67349900        -0.04633951         0.19102751         0.68987033
```

```
# Etape 2 : les contributions des PL et des PC à l'inertie de l'axe
resuafc$row$contrib[,1]
```

```
        A         B         C
22.537052  9.477019 67.985929
```

```
#seuil =100/3=33 : la formule d'assurance C (maximum) a fortement contribué
#à l'inertie du premier axe

resuafc$col$contrib[,1]
```

```
     faible valeur     valeur moyenne      valeur élevée valeur très élevée
         44.704995           2.068932           8.947966          44.278108
```

```
#seuil=100/4 =25: les modalités faible et très élevée de la valeur marchande
#ont fortement contribué à l'inertie du premier axe
```

0.10 Question

Quels sont les profils bien représentés ? Justifier votre réponse.

Conseils

On accède aux cosinus 2 en faisant `$row$cos2` et `$col$cos2`pour étudier la qualité de la représentation.

Solution

```
# Etape 3 : Qualité de représentation des profils :
#On calcule le cos2 de chaque profil.
#Ici, sur l'axe 1 : cos2 > 0.5 est pris comme seuil.

resuafc$row$cos2[,1] # les 3 PL A, B, C
```

```
        A         B         C
0.8630104 0.7130213 0.9980994
```

```
resuafc$col$cos2[,1] # les 4 PC
```

```
     faible valeur     valeur moyenne      valeur élevée valeur très élevée
         0.9968903           0.8785635           0.6465947          0.9525604
```

```
# Sur l'axe 1 : tous les profils sont bien représentés.

# Les profils à la fois bien représentés et ayant une forte contribution
# sur l'axe 1 sont : la formule C, faible valeur et valeur très élevée.
```

0.11 Question

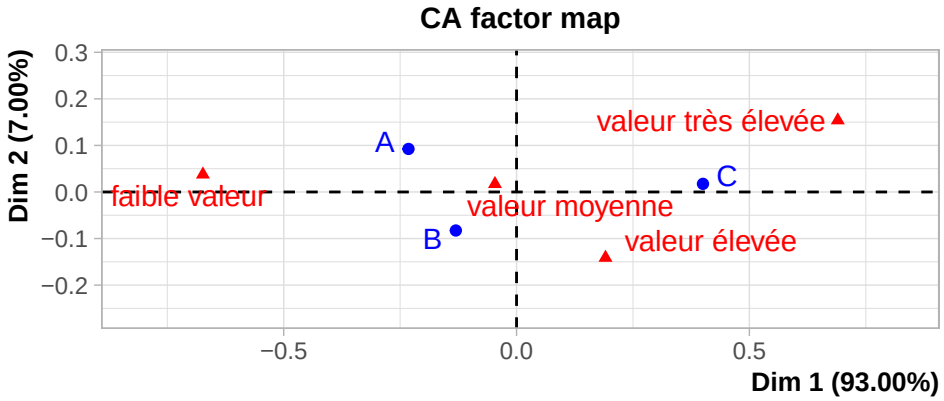
Réaliser le graphique simultané et commenter l'AFC.

Conseils

Pour réaliser le graphique, vous pouvez modifier l'argument `graph =` de la fonction `CA()`.

Solution

```
CA(tabcontin, graph = T)
```




```
**Results of the Correspondence Analysis (CA)**
The row variable has  3  categories; the column variable has 4 categories
The chi square of independence between the two variables is equal to 136.371 (p-value =  5.
```

```
*The results are available in the following objects:

   name              description
1  "$eig"            "eigenvalues"
2  "$col"            "results for the columns"
3  "$col$coord"      "coord. for the columns"
4  "$col$cos2"       "cos2 for the columns"
5  "$col$contrib"    "contributions of the columns"
6  "$row"            "results for the rows"
7  "$row$coord"      "coord. for the rows"
8  "$row$cos2"       "cos2 for the rows"
9  "$row$contrib"    "contributions of the rows"
10 "$call"           "summary called parameters"
11 "$call$marge.col" "weights of the columns"
12 "$call$marge.row" "weights of the rows"
```

```
# Étape 4 : Interprétation du graphique simultané des modalités (ou profils)
#On commente C, Faible et Très élevée : les jeunes conducteurs ayant
#un véhicule de valeur marchande très élevée sont associés à la formule
#d'assurance C (maximum) et s'opposent aux jeunes conducteurs ayant
#un véhicule de faible valeur.
#Remarque : On retrouve le commentaire des contributions.
```

0.12 Question

Refaire l'analyse en utilisant la fonction `CAshiny()` du package **Factoshiny**:

Solution

```
library(Factoshiny)
help(CAshiny)
CAshiny(resuafc)
```