Chapitre 4 Le Logiciel R et Statistique descriptive bivariée

Introduction

Nous présentons ici quelques représentations utiles dans un cadre bivarié. Nous allons fonder tous les exemples de ce chapitre sur le fichier de données *nutrition.csv*

Présentation du fichier:

Un échantillon de personnes âgées a été interrogé dans le cadre d'une enquête nutritionnelle. L'échantillon est constitué de 226 sujets.

Variables et codage

Description	Unité ou Codage	Variable
Sexe	2=Femme ; 1=Homme	sexe
Situation familiale	1=Vit seul 2=Vit en couple 3=Vit dans sa famille 4=Autre type de cohabitation	situation
Consommation journalière de thé	Nombre de tasses	the
Consommation journalière de café	Nombre de tasses	cafe
Taille	Cm	taille
Poids	Kg	poids
Âge le jour de l'entretien	Années	age
Consommation de viande	0=Jamais 1=Moins d'une fois par semaine 2=Une fois par semaine 3=2/3 fois par semaine 4=4/6 fois par semaine 5=Tous les jours	viande
Consommation de poisson	Idem	poisson
Consommation de fruits crus	Idem	fruit_crus
Consommation de fruits et légumes cuits	Idem	fruit_legume_cuits
Consommation de chocolat	Idem	chocol
Matière grasse préférentiellement utilisée pour la cuisson	1=Beurre 2=Margarine 3=Huile d'arachide 4=Huile de tournesol 5=Huile d'olive 6=Mélange d'huile (type Isio4) 7=Huile de colza 8=Graisse de canard ou d'oie	matgras

2.1 Tableaux croisant deux variables

2.1.1 Tableaux de contingence

Lorsque l'on dispose du tableau des données individuelles, on peut utiliser la fonction **table** () pour obtenir le tableau de contingence observe (encore appelé tri croise en effectifs) du couple *(X, Y)*.

```
> matable <- table(sexe,situation)
> matable
```

```
        situation
sexe     seul couple famille autre
  Homme    20     63       2     0
  Femme    78     56       7     0
```

Si l'on veut rajouter les marges a ce tableau, on peut utiliser la fonction addmargins () .

```
> table. complete <- addmargins(matable,FUN=sum,quiet=TRUE)
> table. complet
```

```
        situation
sexe     seul couple famille autre sum
  Homme    20     63       2     0  85
  Femme    78     56       7     0 141
  sum      98    119       9     0 226
```

2.1.2 Distribution conjointe

Le tableau de la distribution conjointe (encore appelé tri croise en fréquences relatives) du couple *(X, Y)* s'obtient a partir du tableau de contingence matable précèdent.

> **tableaufreq <- matable/sum(matable)**
> **tableaufreq**

```
        situation
sexe          seul      couple     famille       autre
  Homme 0.088495575 0.278761062 0.008849558 0.000000000
  Femme 0.345132743 0.247787611 0.030973451 0.000000000
```

> **Total <- sum**
> **addmargins(tableaufreq,FUN=Total,quiet=TRUE)**

```
       situation
sexe          seul      couple     famille       autre
  Homme 0.088495575 0.278761062 0.008849558 0.000000000
  Femme 0.345132743 0.247787611 0.030973451 0.000000000
  Total 0.433628319 0.526548673 0.039823009 0.000000000
       situation
sexe          Total
  Homme 0.376106195
  Femme 0.623893805
  Total 1.000000000
```

2.1.3 Distribution marginalesconjointe

L'obtention des marges d'une table de distribution tableaufreq (ou d'une table de contingence) s'obtient au moyen de la fonction margin.table().

> **margin.table(tableaufreq,l) # Marge de droite.**

```
sexe
    Homme     Femme
0.3761062 0.6238938
```

> **margin.table(tableaufreq,2) # Marge du bas.**

```
situation
      seul     couple    famille      autre
0.43362832 0.52654867 0.03982301 0.00000000
```

2.1.4 Distribution Conditionelle

Les tableaux des distributions conditionnelles s'obtiennent au moyen de la fonction prop. table ()

Distributions conditionnelles de situation sachant les valeurs de sexe

> **prop.table(matable,l)**

```
        situation
sexe          seul     couple    famille      autre
  Homme 0.23529412 0.74117647 0.02352941 0.00000000
  Femme 0.55319149 0.39716312 0.04964539 0.00000000
```

> **addmargins(prop.table(matable,1),margin=2,FUN=sum)**

```
        situation
sexe          seul     couple    famille      autre        sum
  Homme 0.23529412 0.74117647 0.02352941 0.00000000 1.00000000
  Femme 0.55319149 0.39716312 0.04964539 0.00000000 1.00000000
```

• Distributions conditionnelles de sexe sachant les valeurs de situation:

> **prop.table(matable,2)**

```
        situation
sexe          seul    couple   famille autre
  Homme 0.2040816 0.5294118 0.2222222
  Femme 0.7959184 0.4705882 0.7777778
```

> **addmargins(prop.table(matable,2),margin=1,FUN=sum)**

```
        situation
sexe          seul    couple   famille autre
  Homme 0.2040816 0.5294118 0.2222222
  Femme 0.7959184 0.4705882 0.7777778
  sum   1.0000000 1.0000000 1.0000000
```

2.2 Indicateurs Numériques

2.2.1 Mesures de liaison entre deux variables qualitatives

2.2.1.1 La statistique du Khi-deux de Pearson

```
> khi2 <- summary(table(sexe,matgras))$statistic
> khi2
[l] 15.15842
```

2.2.1.2 Coefficient de contingence de Pearson

```
> sexematgras <- table(sexe,matgras) # Tableau de contingence observé.
> q <- ncol(sexematgras)
```

2.2.2 Mesures de liaison entre deux variables ordinales

2.2.2.1 Le coefficient de Kendall

Ce coefficient est fonde sur la notion de concordance entre individus. Pour deux individus *i* et *j* et pour les variables ordinales X et *Y,* on dit que les paires *(Xi,Yi)* et *(Xj,Yj)* sont concordantes si sign(xj - *Xi)* =sign(yj - *Yi)* et discordantes si sign(xj - *Xi)* = -sign(Yj - *Yi).* Si *Xi = Xj* ou *Yi = Yj* (ou les deux), la paire correspondante n'est ni concordante ni discordante, et nous disons qu'il s'agit d'ex-*aequo.*

On calcule alors le coefficient de Kendall par la formule :

$$\tau_b = \frac{2(n_c - n_d)}{\sqrt{(N^2 - \sum_{i=1}^{p} n_{i\bullet}^2)(N^2 - \sum_{j=1}^{q} n_{\bullet j}^2)}}$$

$$\text{avec } 2(n_c - n_d) = \sum_{i=1}^{p} \sum_{j=1}^{q} \text{sign}(x_i - x_j)\text{sign}(y_i - y_j),\ n_{i\bullet} = \sum_{j=1}^{q} n_{ij} \text{ et } n_{\bullet j} = \sum_{i=1}^{p} n_{ij}.$$

Avec

Lorsqu'il n'y a pas d' *ex cequo,* cette formule se simplifie en

$$\tau = \frac{2(n_c - n_d)}{N(N-1)}.$$

Ces deux quantités se calculent en R au moyen de la fonction cor():

```
> cor(as.numeric(viande),as.numeric(poisson),method="kenda11")
[1J -0.1583088
```

2.2.3 Mesures de liaison entre deux variables quantitatives

2.2.3.1 Coefficient de corrélation de Pearson

L'indicateur de liaison approprie dans le cas de deux variables quantitatives est la corrélation. Il est défini comme le rapport entre la covariance des deux variables et le produit de leurs écarts types respectifs. Il se calcule au moyen de la fonction cor () .

```
> cor(taille,poids)
[1] 0. 6306576
```

2.2.3.2 Covariance

La covariance se calcule au moyen de la fonction cov () .

```
> cov(taille,poids)
[1J 68.32596
```

2.2.4 Mesures de liaison entre une variable quantitative une variable qualitative

2.2.4.1 Le rapport de corrélation

Le rapport de corrélation indique dans quelle mesure les variations d'une variable quantitative *Y* sont expliquées par les modalités d'une variable qualitative X à *p* modalités. En effet, on peut considérer que la variable X définit des groupes dans la population. Le rapport de

corrélation est alors défini comme le rapport entre la variance inter-groupes et la variance intra-groupe. Il se calcule au moyen de la formule suivante :

$$\eta^2_{Y|X} = \frac{\sum_{k=1}^{p} n_k(\overline{y}_k - \overline{y})^2}{\sum_{i=1}^{N}(y_i - \overline{y})^2}$$

dans laquelle n_k désigne le nombre d'observations y_i correspondant à la k-ième modalité de X.

Voici le programme R permettant de le calculer :

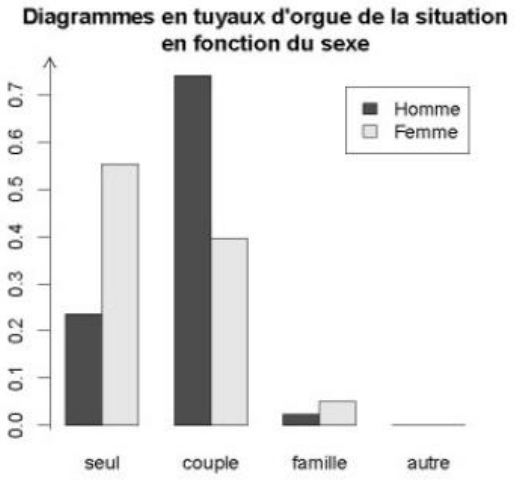
```
> eta2 <- function(x, gpe) {
                    moyennes <- tapply(x, gpe, mean)
                    effectifs <- tapply(x, gpe, length)
                     varinter <- (sum(effectifs * (moyennes - mean(x))A2))
                    vartot <- (var(x) * (length (x) - 1))
                     res <- varinter/vartot
                    return (res)
          }
> eta2(poids,sexe)
[1] 0.3325501
```

2.2 Représentations graphiques

2.2.1 Croisement de deux variables qualitatives

II est possible de superposer deux diagrammes en tuyaux d'orgue comme on peut le voir ici

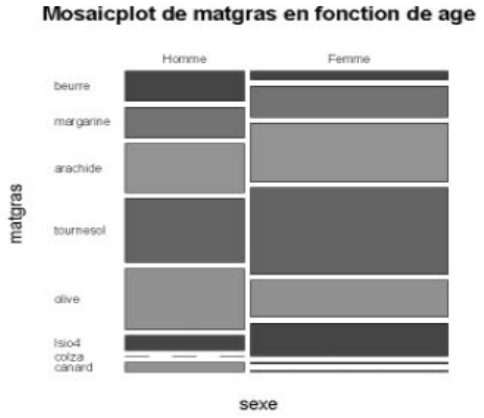
```
 > tss <- prop.table(table(sexe,situation),l)
> barplot(tss,bes=T,leg=T)
> title(paste("Diagrammes en tuyaux d'orgue de la situation", "en fonction du
+ sep="\n"))
> fleches(F,T)
```



.

Le diagramme mosaique peut aussi etre utile pour le croisement de deux variables qualitatives.

```
> par(las=l) # Ecriture horizontale des modalites.
> mosaicplot (sexe-matgras,color=brewer.pal (5, "Setl"),+ main="Mosaicplot de matgras en fonction de age")
```

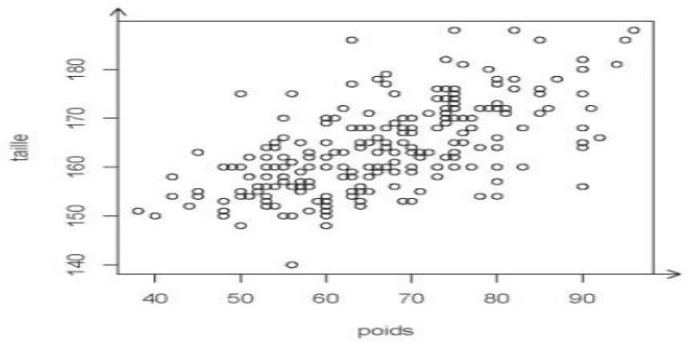


2.2.2 Croisement de deux variables quantitatives

La fonction autiliser dans ce contexte est la fonction plot O.

> **plot(taille-poids)**

> **fleches ()**



0

2.2.1 Croisement d'une variable qualitative et une variable quantitative

Croisement d'une variable qualitative et d'une variable **quan**titative

Dans ce contexte, il est intéressant de tracer des diagrammes en boite à moustaches *(boxplots)* de la variable quantitative pour chaque modalité de la variable qualitative. Si les variables ont été correctement structurées dans R, il

> **par(bty="n")**

> plot (cafe-sexe,col=brewer.pal(S, "Set2"),notch=T,varwidt**h=T,**

+ boxwex=O.3)

> **title(paste("Boxplot de la consommation de cafe",**

+ "en fonction du sexe",sep="\n"),family="Courier")

> **fleches(F,T)**