Fiche TD avec le logiciel  : tdr1102

Seconds pas vers l'analyse de données ...

A.B. Dufour & D. Clot

Cette fiche comprend des exercices portant sur les paramètres descriptifs principaux et les représentations graphiques associés aux croisement de deux variables, qu'elles soient quantitatives ou qualitatives. L'objectif est de se placer dans un contexte d'Analyse de Données.

Table des matières

1 Introduction

L'objectif de cette fiche est de rappeler les outils principaux de l'**analyse descriptive bivariée** sous R. Les fichiers de données analysées sont les mêmes que ceux de la fiche `tdr1101.pdf`.

1. Les crimes violents aux U.S.A.

```
CSD <- read.csv("http://pbil.univ-lyon1.fr/R/donnees/CrimeStateDate.csv", header=TRUE)
```

2. La sécurité routière dans les départements français.

```
SR0910 <- read.table("http://pbil.univ-lyon1.fr/R/donnees/SecRoutiere0910.txt", header=T)
```

3. Les clients d'une agence bancaire.

```
library(ade4)
data(banque)
```

On distingue deux types de variables donc trois croisements possibles.

Croisement	Paramètre	Graphique
Quantitatif × Quantitatif	covariance coefficient de corrélation coefficient de détermination	nuage de points
Qualitatif × Qualitatif	Chi-Deux Coefficient de Cramer	mosaïque représentation en "ballons"
Quantitatif × Qualitatif	Rapport de Corrélation	représentation inter et intragroupes boîtes à moustaches

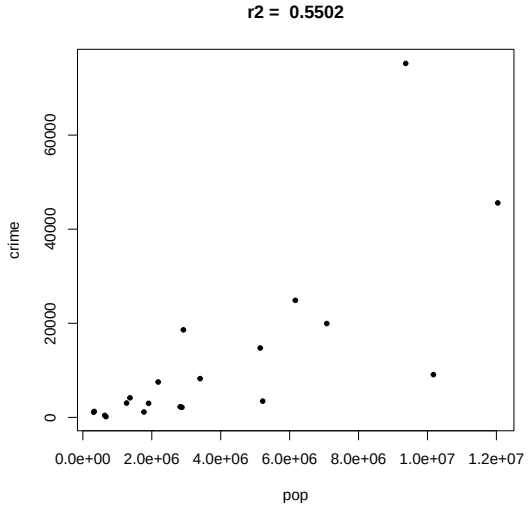
Les définitions des paramètres et des graphiques sont présentées à titre de rappel dans le TD et non dans le cadre d'un cours associé aux croisements des variables.

2 Croisement entre deux variables quantitatives

2.1 Principe par l'expérimentation

On étudie la relation entre le nombre d'habitants et le nombre de crimes violents dans les 51 états d'Amérique du Nord.
Dans cette phase du travail, on propose une fonction `quantquant()` qui échantillonne 20 lignes au hasard dans le data frame `CSD`. On représente le nuage de points et on affiche un coefficient de lien entre les deux variables dit coefficient de détermination (carré du coefficient de corrélation).

```
quantquant <- function() {
  nligne <- sample(1:2341,20)
  pop <- CSD$Population[nligne]
  crime <- CSD$Crime_Violent[nligne]
  cor12 <- cor(pop,crime)^2
  plot(pop,crime,pch=20,main=paste("r2 = ",round(cor12,4)))
}
quantquant()
```



Exercice. Réitérer la fonction `quantquant()` plusieurs fois, observer la forme du nuage de points et le coefficient de détermination associé. Discuter la relation entre les deux.

2.2 Définitions et mises en garde

On considère deux variables quantitatives X et Y mesurées sur n individus. Le lien entre ces deux variables est la **covariance** :

$$cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Si la covariance est élevée (en valeur absolue), les deux variables sont liées. Mais elle s'exprime dans les unités des deux variables et cela la rend difficile à interpréter. On lui préfère donc le **coefficient de corrélation** :

$$r = \frac{cov(X,Y)}{s_X s_Y} \quad \text{où} \quad s_X \quad \text{et} \quad s_Y \quad \text{sont les écarts types de} \quad X \quad \text{et} \quad Y.$$

Le coefficient de corrélation est compris entre -1 et 1. Le signe indique le sens de la relation. On lui préfère parfois son carré dit **coefficient de détermination**.

La représentation graphique associée à la relation entre deux variables quantitatives est appelée **nuage de points**. Un bon nuage de points présente une forme ellipsoïdale. Lier coefficient et graphique est fondamental comme le montre F. Anscombe (1973) dans l'exercice ci-dessous.

Exercice

```
data(anscombe)
names(anscombe)
```
```
[1] "x1" "x2" "x3" "x4" "y1" "y2" "y3" "y4"
```

Calculer les coefficients de détermination et représenter, dans une même fenêtre graphique, les nuages de points des couples de variables suivants : (x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4). Conclure.

2.3 Les accidents de la route dans les départements français

1. Etudier la relation entre la population dans les départements français et les blessés en 2009 : coefficient de détermination et nuage de points.
2. Que peut-on dire de la dispersion de la variable 'blessés en 2009' en fonction de l'augmentation du nombre d'habitants dans les départements ?
3. Transformer les deux variables précédentes en leur logarithme népérien. Construire le nouveau nuage de points et calculer le coefficient de détermination.
4. Conclure.

3 Croisement de deux variables qualitatives

3.1 Principe par l'expérimentation

On étudie la relation entre la somme déposée sur un livret d'épargne et le sexe des clients d'une banque. La variable 'somme' possède trois modalités : nulle, faible, forte ; la variable 'sexe' est binaire 'femme', 'homme'. Voici un exemple de table observée :

```
tabcont <- matrix(c(437,95,26,185,49,18),byrow=TRUE,ncol=3)
rownames(tabcont) <- c("homme","femme")
colnames(tabcont) <- c("nulle","faible","forte")
tabcont
```

```
      nulle faible forte
homme   437     95    26
femme   185     49    18
```

Les marges de la table observée sont les effectifs associés aux modalités de chaque variable.

```
apply(tabcont,1,sum)
```

```
homme femme
  558   252
```

```
apply(tabcont,2,sum)
```

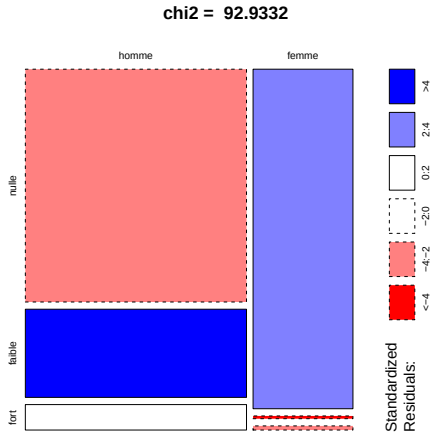
```
nulle faible  forte
  622    144     44
```

Dans cette phase du travail, on propose une fonction `qualqual()` qui construit différentes tables avec les mêmes marges que la table de contingence originale. On réalise une représentation en mosaïque de la table et on affiche un coefficient de lien entre les deux variables dit Chi-Deux de Contingence.

```
qualqual <- function() {
  caseEF <- sample(0:188,1)
  caseBC <- 188-caseEF
  caseA <- 558-caseBC
  caseD <- 622-caseA
#
        neomin <- min(44,caseEF,188-caseEF)
if (neomin == (188-caseEF))
{ caseC<-sample(0:neomin,1); caseF <- 44-caseC}else{ caseF <- sample(0:neomin,1); caseC <- 44-caseF}
        caseB <- caseBC-caseC
        caseE <- 144-caseB
#
result <- c(caseA,caseB,caseC,caseD,caseE,caseF)
tabcont <- matrix(result,ncol=3,byrow=TRUE)
colnames(tabcont) <- c("nulle","faible","fort")
```

```
rownames(tabcont) <- c("homme","femme")
print(tabcont)
#
res <- chisq.test(tabcont, correct = FALSE)
chi2 <- as.numeric(res$statistic)
mosaicplot(tabcont,shade=T, main=(paste("chi2 = ",round(chi2,4))))
}
qualqual()
```

```
      nulle faible fort
homme   375    142   41
femme   247      2    3
```



Exercice. Réitérer la fonction `qualqual()` plusieurs fois, observer les couleurs associées à la mosaïque et le coefficient de Cramer associé. Discuter la relation entre les deux.

3.2 Définitions

Pour étudier la relation entre deux variables qualitatives (ou discrètes), le paramètre de lien est le chi-deux de contingence auquel on peut associer un paramètre descriptif borné entre 0 et 1 : le coefficient de Cramer. Deux représentations graphiques permettent de visualiser la table de contingence : une représentation en cercles ou carrés basée sur les effectifs observés et une représentation en mosaïque basée sur les écarts entre les effectifs théoriques et les effectifs observés que l'on va rappeler ci-dessous.

Table de contingence

Soient A et B, deux variables qualitatives ayant respectivement p et q modalités observées sur n individus. La table de contingence observée est un tableau croisé où les colonnes correspondent aux q modalités de la variable B et les lignes aux p modalités de la variable A. On note n_{ij} le nombre d'individus possédant à la fois la modalité i de la variable A et la modalité j de la variable B.

	B1	$\cdots$	Bj	$\cdots$	Bq	total
A1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1q}	$n_{1.}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
Ai	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{iq}	$n_{i.}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
Ap	n_{p1}	$\cdots$	n_{pj}	$\cdots$	n_{pq}	$n_{p.}$
total	$n_{.1}$	$\cdots$	$n_{.j}$	$\cdots$	$n_{.q}$	n

Valeur du Chi-Deux de contingence

Afin de définir le lien entre les deux variables, on construit une table de contingence dite table théorique. Cette table est basée sur l'hypothèse d'indépendance entre les deux variables c'est-à-dire par l'équiprobabilité au sein des différentes cases de la table conditionnée par ses marges soit :

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

La valeur du Chi-Deux de contingence compare les effectifs de la table de contingence observée ($EO = n_{ij}$) avec les effectifs de la table de contingence théorique ($ET = \frac{n_{i.}n_{.j}}{n}$).

$$\chi^2 = \sum \frac{(EO - ET)^2}{ET}$$

Si $\chi^2 = 0$, il y a indépendance entre les deux variables.
Si χ^2 est petit, les effectifs observés sont presque identiques aux effectifs théoriques. Les deux variables sont peu liées entre elles.
Si χ^2 est grand, les effectifs observés sont différents des effectifs théoriques. Les deux variables sont liées entre elles.

Le coefficient de Cramer

Comme il est difficile de définir si la valeur du Chi-Deux est grande ou non, des paramètres descriptifs ont été définis. On retient l'indice de Cramer qui varie entre 0 et 1. Si le coefficient est proche de 0, les variables ne sont pas liées. Si le coefficient est proche de 1, les variables sont liées.

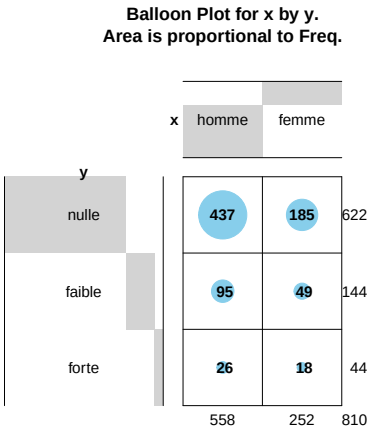
$$V = \sqrt{\frac{\chi^2}{n \times \min(p-1, q-1))}}$$

Mais ce paramètre, bien que borné et proche du coefficient de détermination dans son interprétation, est peu utilisé.

La représentation en "ballons"

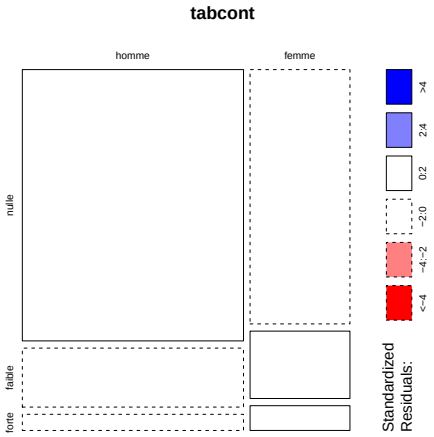
Cette représentation simple se base sur les effectifs observés n_{ij}.

```
library(gplots)
balloonplot(as.table(tabcont))
```



La représentation en mosaïque
Cette représentation plus complexe se base sur les écarts entre effectifs observés et théoriques $(EO - ET)/\sqrt{ET}$. Elle donne une lisibilité au Chi-Deux.

```
mosaicplot(tabcont,shade=T)
```



3.3 Les clients de la banque

On étudie les relations entre l'âge des clients et les différentes épargnes proposées par la banque : l'assurance vie (`assurvi`), l'épargne logement (`eparlog`) et le livret d"épargne (`eparliv`).

```
data(banque)
names(banque)
```

```
 [1] "csp"      "duree"    "oppo"     "age"      "sexe"     "interdit" "cableue"
 [8] "assurvi"  "soldevu"  "eparlog"  "eparliv"  "credhab"  "credcon"  "versesp"
[15] "retresp"  "remiche"  "preltre"  "prelfin"  "viredeb"  "virecre"  "porttit"
```

Réaliser les différents croisements possibles et conclure en se mettant dans la peau du directeur de la banque.

Remarque. Utiliser l'instruction `chisq.test()$expected` afin de visualiser les effectifs théoriques de la table de contingence. Pour une bonne analyse, il est souhaitable que ces derniers soient supérieurs à 5.

4 Croisement d'une variable quantitative et d'une variable qualitative

4.1 Principe par l'expérimentation

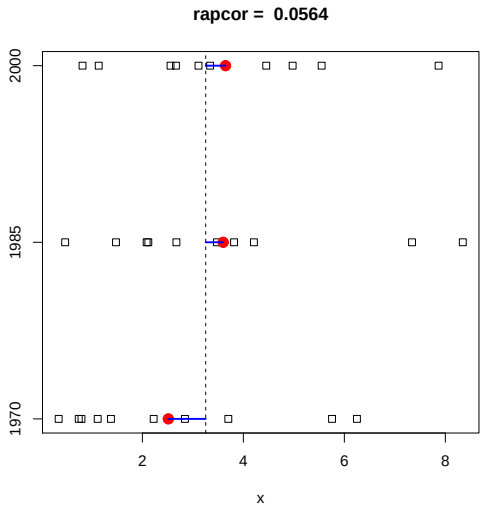
On étudie la relation entre le nombre de crimes violents aux U.S.A. et le temps. L'année est considérée comme une variable qualitative.

Afin de visualiser la relation entre le nombre de crimes et l'année, on propose la représentation suivante.
- Les années sont représentées sur l'axe vertical, le nombre de crimes sur l'axe horizontal.
- Un carré blanc représente un individu.
- Les points rouges représentent les moyennes pour chaque année.
- La ligne en pointillé représente la moyenne de l'ensemble des individus.
- Les traits bleus représentent les écarts entre les moyennes des groupes et la moyenne de l'ensemble.

Dans cette phase du travail, on propose une fonction `quantqual()` qui échantillonne au hasard 10 états américains pour les années 1970, 1985 et 2000, calcule le rapport : crimes violents sur population en pour mille. On réalise la représentation graphique explicitée ci-dessus et on affiche un coefficient de lien entre les deux variables dit rapport de corrélation.

```
variation <- function(x) var(x)*(length(x)-1)
varinter <- function(x,gpe) {
    moyennes <- tapply(x,gpe,mean)
    effectifs <- tapply(x,gpe,length)
    res <- (sum(effectifs*(moyennes-mean(x))^2))
    }
#
graphnf <- function(x,gpe) {
    rapcor <- varinter(x,gpe)/variation(x)
    stripchart(x~gpe, main=paste("rapcor = ",round(rapcor,4)))
    points(tapply(x,gpe,mean),1:length(levels(gpe)),col="red",pch=19,cex=1.5)
    abline(v=mean(x),lty=2)
    moyennes <- tapply(x,gpe,mean)
    traitnf <- function(n) segments(moyennes[n],n,mean(x),n,col="blue",lwd=2)
    sapply(1:length(levels(gpe)),traitnf)
    }
#
rapport <- CSD$Crime_Violent/CSD$Population
rapport <- rapport*1000
rap1970 <- rapport[CSD$Date=="1970"]
```

```
rap1985 <- rapport[CSD$Date=="1985"]
rap2000 <- rapport[CSD$Date=="2000"]
annees <- factor(c(rep("1970",10),rep("1985",10),rep("2000",10)))
quantqual <- function(){
nlignes <- sample(1:51,10)
crimes <- c(rap1970[nlignes],rap1985[nlignes],rap2000[nlignes])
graphnf(crimes,annees)
}
quantqual()
```



Exercice. Réitérer la fonction `quantqual()` plusieurs fois, observer les relations entre les moyennes et le rapport de corrélation. Discuter la relation entre les deux.

4.2 Définitions

Pour étudier la relation entre une variable qualitative et une variable quantitative, on décompose la variation totale en variation intergroupe et en variation intragroupe. Pour mesurer l'intensité de la relation (toujours d'un point de vue descriptif), on peut calculer un paramètre appelé rapport de corrélation.

La notion de variation

La variance d'une variable quantitative peut être perçue selon le point de vue descriptif ou le point de vue inférentiel. Ce terme général peut désigner :

- ⋆ la variance descriptive mesurée sur une groupe de n individus

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

- ⋆ la variance estimée de la population à partir d'un échantillon de n individus

$$\widehat{\sigma^2} = \frac{n}{n-1}s^2$$

que l'on peut encore écrire

$$\widehat{\sigma^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

C'est pourquoi, on préfèrera travailler sur la variation totale c'est-à-dire la somme des carrés des écarts à la moyenne :

$$variation = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

Sous R , nous avons écrit
soit :

```
variation <- function(x) sum((x-mean(x))^2)
```

soit :

```
variation <- function(x) var(x)*(length(x)-1)
```

La notion de variation intergroupe

On va calculer le carré des écarts entre la moyenne du groupe et la moyenne globale. Cette quantité est appelée *variation inter-groupe* (C'est la longueur du trait bleu.)

$$varinter = \sum_{k=1}^{p} n_k(\overline{x_k} - \bar{x})^2$$

où $\overline{x_k}$ désigne le nombre moyen de crimes pour l'année k et n_k, le nombre d'états ayant été échantillonnés l'année k.
Sous R ,nous avons écrit :

```
varinter <- function(x,gpe) {
    moyennes <- tapply(x,gpe,mean)
    effectifs <- tapply(x,gpe,length)
    res <- (sum(effectifs*(moyennes-mean(x))^2))
    return(res)
    }
```

Le rapport de corrélation

Pour étudier la relation entre une variable qualitative et une variable quantitative, on calcule le rapport de corrélation noté η^2 :

$$\eta^2 = \frac{\sum_{k=1}^{p} n_k(\overline{x_k} - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Si le rapport est proche de 0, les deux variables ne sont pas liées.
Si le rapport est proche de 1, les variables sont liées.

Le rapport de corrélation a donc le même sens interprétatif que le coefficient de détermination et le coefficient de Cramer.
Sous R ,nous écrirons :

```
eta2 <- function(x,gpe) {res <- varinter(x,gpe)/variation(x) ; return(res)}
```

Remarque.

Variation Totale = Variation inter-groupes + Variation intragroupe

La variation intragroupe peut donc s'écrire soit comme la différence entre la variation totale et la variation inter-groupe, soit comme la somme pondérée des variances calculées à l'intérieur de chaque groupe.

$$varintra = \sum_{k=1}^{p} n_k s_k^2$$

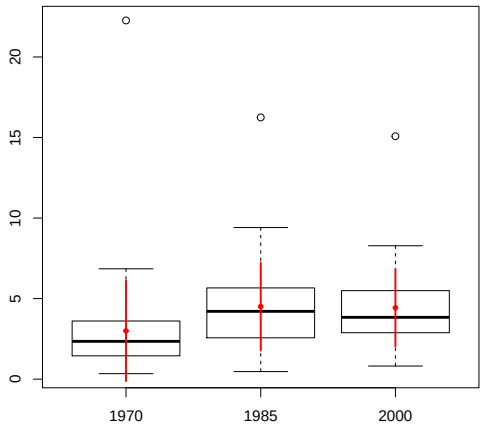
où s_k^2 est la variance descriptive au sein du groupe k.

Il n'existe pas vraiment de représentation graphique associée à la visualisation de la relation entre une variable qualitative et une variable quantitative. C'est pourquoi nous avons choisi la représentation ci-dessus.

Une représentation graphique liant une variable quantitative et une variable qualitative classiquement utilisée est la boîte à moustaches. Elle visualise les quartiles alors que le rapport de corrélation est basée sur des moyennes et des variances. Si la variable étudiée est de nature symétrique, cela ne pose aucun problème. Si la variable étudiée est asymétrique, ce qui est observée graphiquement est différent de ce qui est calculée.

```
rap1970 <- rapport[CSD$Date=="1970"]
rap1985 <- rapport[CSD$Date=="1985"]
rap2000 <- rapport[CSD$Date=="2000"]
raptot <- c(rap1970,rap1985,rap2000)
date <- factor(c(rep("1970",51),rep("1985",51),rep("2000",51)))
boxplot(raptot~date)
points(1,mean(rap1970),col="red",pch=20)
segments(x0=1,y0=mean(rap1970)+sd(rap1970),y1=mean(rap1970)-sd(rap1970),col="red",lwd=2)
points(2,mean(rap1985),col="red",pch=20)
segments(x0=2,y0=mean(rap1985)+sd(rap1985),y1=mean(rap1985)-sd(rap1985),col="red",lwd=2)
points(3,mean(rap2000),col="red",pch=20)
segments(x0=3,y0=mean(rap2000)+sd(rap2000),y1=mean(rap2000)-sd(rap2000),col="red",lwd=2)
eta2(raptot, date)
```

```
[1] 0.0600497
```



4.3 Accidents de la route et régions françaises

1. On conserve les régions françaises ayant au moins 6 départements : le Centre, l'Ile de France, la région Midi-Pyrénées, la région Provence Alpes Côte d'Azur et la région Rhône Alpes. Préparer le nouveau jeu de données en prenant soin de bien redéfinir les modalités de la variables région.

```
(int <- levels(SR0910$region)[summary(SR0910$region)>=6])
```

```
[1] "Centre"                    "Ile_De_France"
[3] "Midi_Pyrenees"             "Provence_Alpes_Cote_dAzur"
[5] "Rhone_Alpes"
```

2. Etudier le lien entre ces régions et chacune des variables suivantes : les accidents corporels en 2009, les blessés en 2009, les tués en 2009 et le ratio.

3. Que conseiller aux présidents de région ?