Université de Perpignan – IUT de Carcassonne
Nathalie Villa-Vialaneix
Année 2010/2011

IUT STID, 1ère année & APPC

Statistique descriptive

Interrogation 2 : à rendre le Jeudi 18 novembre

Avant propos : Cette interrogation est à effectuer en binôme (les binômes sont les binômes de TP). Les réponses sont à compléter directement sur le fichier **`Interro2-etud.odt`**, téléchargeable sur mon site web (http://www.nathalievilla.org, en cliquant sur Enseignements / IUT STID Carcassonne / Statistique descriptive), et le fichier, renommé « Interro2-*nom1*-*nom2*.odt » (où *nom1* et *nom2* sont les noms des deux membres du binômes), est à envoyer à nathalie.villa@univ-perp.fr, **avant le jeudi 18 novembre à 20h**. Un accusé de réception de la bonne réception de votre fichier doit vous parvenir.

Noms : _____

Pour effectuer cette interrogation, vous utiliserez le fichier de données R **`TP1.RData`** et correspondant à l'étude tirée de l'article :

Desbois, D. (2008) Introduction to scoring methods: financial problems of farm holdings. *Case Studies in Business, Industry and Government Statistics*, **2**(1), 56-76.

Ce fichier est téléchargeable sur mon site web, http://www.nathalievilla.org, en cliquant sur Enseignements / IUT STID Carcassonne / Statistique descriptive, sous le nom `TP1.RData`. En particulier, l'objet de cette interrogation est l'étude de deux variables particulières de ce jeu de données, « OWNLAND » (qualitative) et « r1 » (quantitative). L'étude comprend une étude globale puis une comparaison selon les modalités de la variable d'intérêt principal du jeu de données qui est la variable « DIFF » indiquant un incident (ou non) de paiement dans le remboursement d'un crédit.

1 Étude de la variable « OWNLAND »....2
2 Étude de la variable « r1 »....3
3 Comparaison entre entreprises ayant eu un problème de remboursement et entreprises n'ayant pas eu de problème....5
3.1 Création de deux jeux de données....5
3.2 Comparaison de la variable « OWNLAND »....5
3.3 Comparaison de la variable « r1 »....6

1 Étude de la variable « OWNLAND »

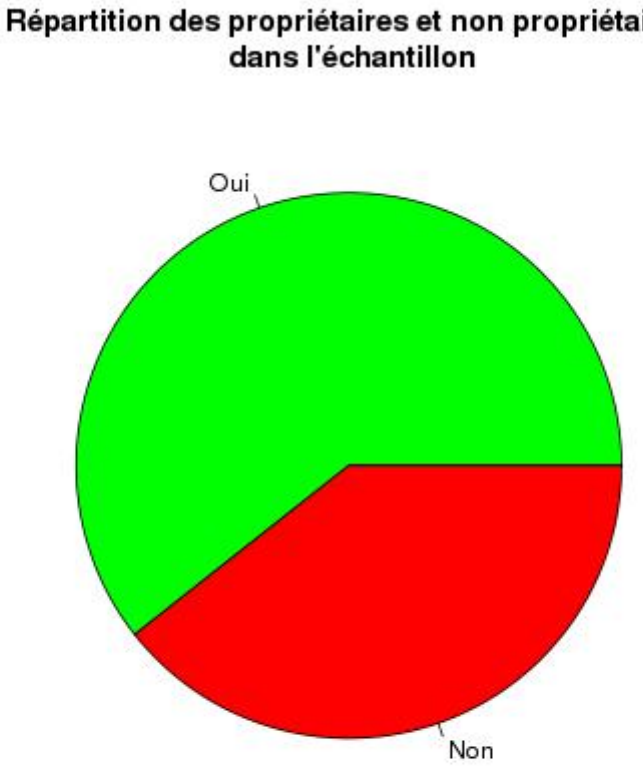
Dans cette partie, on se propose d'étudier la variable « OWNLAND » : donnez le tableau d'effectifs de cette variable ainsi qu'un graphique permettant de représenter sa distribution de manière adapté. Commentez les résultats obtenus.

On insèrera dans le cadre Code 1 le code généré pour effectuer ces deux éléments.

Tableau d'effectifs

Propriétaire	Oui	Non
Effectifs	764	496

Représentation graphique de la distribution de « OWNLAND »



Commentaires

La majorité des exploitants de l'échantillon sont des propriétaires (plus de 60%) mais la part des non propriétaires n'est pas négligeable puisqu'ils représentent plus d'un tiers des exploitants considérés.

```
# Tableau d'effectifs

.Table <- table(donnees$OWNLAND)

.Table

# Diagramme circulaire

pie(table(donnees$OWNLAND),labels=levels(donnees$OWNLAND),main="R
épartition des propriétaires et non propriétaires\n dans
l'échantillon",col=c("green","red"))
```

Code 1: Code généré pour effectuer le tableau d'effectifs et le graphique

2 Étude de la variable « r1 »

Dans cette seconde partie, on étudie la variable « r1 » qui est le ratio entre la dette totale et les avoirs de l'entreprise.

1. Déterminez la moyenne, la médiane, le minimum, le maximum et l'écart type de « r1 » : vous donnerez les valeurs de ces trois statistiques puis vous commenterez la différence entre moyenne et médiane. De plus, vous préciserez dans le cadre Code 2 le code généré pour déterminer ces trois valeurs.

Statistiques

La variable « r1 » a des valeurs variant de 0,119 (11,9%) à 3,494. La moyenne de la variable « r1 » est égale à 0,583 (58,3%), sa médiane à 0,532 (53,2%) et son écart type à 0,332 (33,2%).

Commentaires

La médiane de « r1 » est légèrement plus faible que sa moyenne ce qui est le signe d'une distribution sans doute légèrement étirée vers la droite : certaines entreprises doivent avoir des valeurs un peu atypiques vers les fortes valeurs pour la variable « r1 ».

```
numSummary(donnees[,"r1"],statistics=c("mean","sd","quantiles"),
quantiles=c(0,0.5,1))
```

Code 2 : Code généré pour déterminer les statistiques principales de la variable « r1 »

2. Que signifie concrètement, pour une entreprise, d'avoir une valeur de « r1 » supérieure à 1 ? Que peut-on attendre quant à l'influence de cette variable sur les incidents de paiement (variable DIFF) ?

Lorsqu'une entreprise agricole a une valeur de « r1 » supérieure à 1, cela signifie que le montant de ces dettes est supérieur au montant de ses biens. On peut imaginer que, dans ce cas, les entreprises sont plus souvent sujettes à avoir des incidents de paiement.

3. Effectuez un regroupement des valeurs de « r1 » en 10 classes pertinentes. Donnez le tableau d'effectifs de ces dix classes puis l'histogramme associé que vous commenterez (notamment en relation avec ce que vous avez dit à la question 1 ci-dessus).

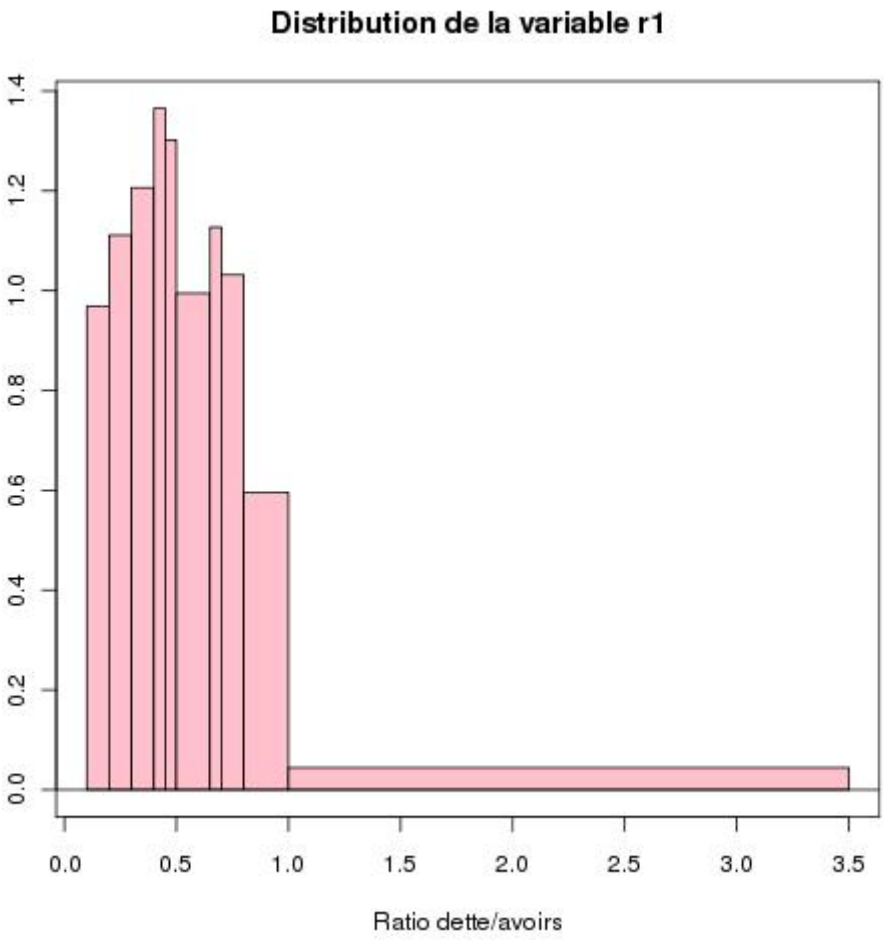
   Enfin, vous préciserez dans le cadre Code 3 le code généré pour effectuer le regroupement en classes, déterminer le tableau d'effectifs et construire l'histogramme.

Tableau d'effectifs du regroupement en classes

R1	[0,1;0,2[	[0,2;0,3[	[0,3;0,4[	[0,4;0,45[	[0,45;0,5[	[0,5;0,65[	[0,65;0,7[	[0,7;0,8[	[0,8;1[	[1;3,5[

Ef.	120	142	150	85	84	187	73	129	150	140

Histogramme



Commentaires

L'histogramme de la variable « r1 » montre un étalement marqué vers la droite : alors que la majorité des entreprises agricoles de l'échantillon ont une dette d'un montant inférieur à leurs avoirs, quelques entreprises ont des dettes représentant plus de 3 fois leurs avoirs.

```
# Découpage en 10 classes
donnees$classes.r1[donnees$r1<0.2]<-"[0,1;0,2["
donnees$classes.r1[donnees$r1<0.3&donnees$r1>=0.2]<-"[0,2;0,3["
donnees$classes.r1[donnees$r1<0.4&donnees$r1>=0.3]<-"[0,3;0,4["
donnees$classes.r1[donnees$r1<0.45&donnees$r1>=0.4]<-"[0,4;0,45["
donnees$classes.r1[donnees$r1<0.5&donnees$r1>=0.45]<-"[0,45;0,5["
donnees$classes.r1[donnees$r1<0.65&donnees$r1>=0.5]<-"[0,5;0,65["
donnees$classes.r1[donnees$r1<0.7&donnees$r1>=0.65]<-"[0,65;0,7["
donnees$classes.r1[donnees$r1<0.8&donnees$r1>=0.7]<-"[0,7;0,8["
donnees$classes.r1[donnees$r1<1&donnees$r1>=0.8]<-"[0,8;1["
donnees$classes.r1[donnees$r1>=1]<-"[1;3,5["
# Tableau d'effectifs
table(donnees$classes.r1)
# Histogramme
Hist(donnees$r1,scale="density",breaks=c(0.1,0.2,0.3,0.4,0.45,0.5
,0.65,0.7,0.8,1,3.5),col="pink",main="Distribution de la variable
r1",xlab="Ratio dette/avoirs",ylab="")
```

Code 3: Code généré pour effectuer le regroupement en classes, le tableau d'effectifs et l'histogramme

3 Comparaison entre entreprises ayant eu un problème de remboursement et entreprises n'ayant pas eu de problème

Le but de cette partie est de comparer les valeurs des variables « OWNLAND » et « r1 » entre les entreprises ayant eu un incident de paiement et celle n'en ayant pas eu.

3.1 Création de deux jeux de données

À partir du jeu de données initial, créez deux jeux de données : un nommé « donnees.sain » contenant les valeurs de toutes les variables pour les entreprises n'ayant pas eu d'incident de paiement (variable « DIFF » égale à « sain ») et un autre, nommé « donnees.incident » contenant les valeurs de toutes les variables pour les entreprises ayant eu un incident de paiement (variable « DIFF » égale à « incident »). Précisez dans le cadre Code 4 le code généré pour créer ces deux jeux de données.

```
donnees.sain <- subset(donnees, subset=DIFF=="sain")
donnees.incident <- subset(donnees, subset=DIFF=="incident")
```

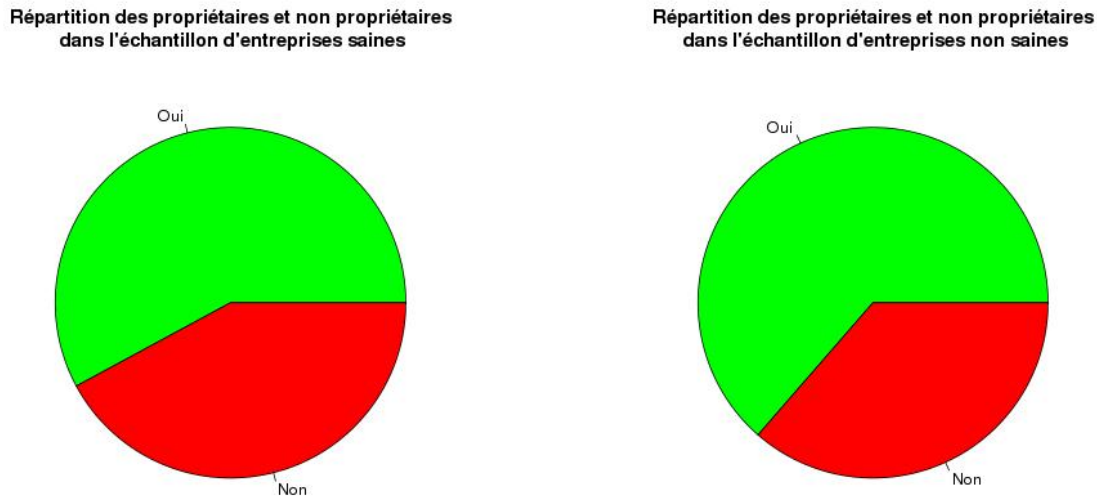
Code 4: Code généré pour créer deux jeux de données à partir des données initiales

3.2 Comparaison de la variable « OWNLAND »

Effectuez le diagramme circulaire de la distribution de la variable « OWNLAND » pour les

entreprises ayant eu un incident de paiement et pour les entreprises n'en ayant pas eu. Commentez les différences : les résultats étaient-ils prévisibles ? Si oui, pourquoi ?

Diagrammes circulaires



Commentaires

La proportion d'agriculteurs propriétaires parmi les agriculteurs n'ayant pas eu d'incident de paiement est légèrement plus faible que parmi les agriculteurs ayant eu un incident de paiement. Ce résultat peut paraître étonnant dans la mesure où l'on aurait pu imaginer qu'être propriétaire de son exploitation était une bonne indication de la santé financière de l'entreprise agricole : il semble donc que cela ne soit pas le cas.

3.3 Comparaison de la variable « r1 »

1. Déterminez la moyenne et l'écart type de la variable « r1 » pour les entreprises ayant eu un incident de paiement et pour celles n'en ayant pas eu. Commentez les différences ? Celles-ci étaient-elles prévisibles et si oui, pourquoi ?

Statistiques

Pour les agriculteurs n'ayant pas eu d'incident de paiement le ratio moyen entre dette et avoirs est de 37,6% (avec un écart type de 18,1%) alors que pour les agriculteurs ayant eu un incident de paiement, le ratio moyen entre dette et avoirs est égal à 80,6% (avec un écart type de 31,2%).

Commentaires

Comme on pouvait s'y attendre, le ratio moyen entre dette et avoirs est plus faible chez les agriculteurs n'ayant pas eu d'incident de paiement que chez ceux qui en ont eu un : les entreprises n'ayant pas d'incident de paiement ont effectivement une meilleure santé financière.

Par ailleurs, l'écart type de ce ratio est également plus faible chez les agriculteurs n'ayant pas eu d'incident de paiement ; c'est le signe que, parmi les agriculteurs ayant eu un incident de paiement, il y a une plus grande disparité de situation, avec sans doute des agriculteurs pour lesquels le ratio entre dette et avoirs est relativement élevé.

2. Déterminez la variable centrée réduite issue de « r1 » pour les entreprises ayant eu un incident de paiement et pour celles n'en ayant pas eu. Vous copierez le code permettant d'obtenir ces deux variables dans le cadre Code 5. Entre l'entreprise saine numéro 1 et

l'entreprise ayant eu un incident de paiement numéro 78, laquelle a la situation financière la plus favorable de son groupe ? Justifier.

```
.Z <- scale(donnees.sain[,c("r1")])
donnees.sain$Z.r1 <- .Z[,1]
remove(.Z)
.Z <- scale(donnees.incident[,c("r1")])
donnees.incident$Z.r1 <- .Z[,1]
remove(.Z)
```

Code 5: Code généré pour créer les deux variables centrées réduites

Comparaison entreprise 1 et entreprise 78

L'entreprise 1 (saine avec r1=0,449) a une valeur centrée réduite pour « r1 » égale à 0,403 et l'entreprise 78 (ayant eu un incident de paiement, avec r1=1,076) a une valeur centrée réduite pour « r1 » égale à 0,863. Ainsi, même si la situation financière de l'entreprise 1 est meilleure que celle de l'entreprise 78 en valeur absolue, sa situation, relativement aux entreprises saines, est moins bonne que la situation de l'entreprise 78 par rapport aux entreprises ayant eu un incident de paiement (car 0,403<0,863).