Étude des liaisons entre deux variables

Université Assane SECK de Ziguinchor

UFR des Sciences et techniques

Département Informatique

Licence 2 : Ingénierie Informatique

Sommaire

Sommaire

Dans le processus de décision ou comme support à l'aide à la décision ou dans le cadre d'une expérimentation technologique, il arrive fréquemment que les conclusions et les recommandations soient basées sur l'existence d'une liaison d'ordre fonctionnelle ou statistique entre deux ou plusieurs variables.

Exemple

Le responsable de la mise en marché d'une entreprise, après avoir établi une relation entre les dépenses en publicité et le volume des ventes, peut effectuer une prévision du volume des ventes selon un niveau de dépenses en publicité.

Dans la suite de ce chapitre, on traitera le cas où l'on dispose de deux caractères X et Y observés sur les mêmes individus. Plusieurs cas de figures sont envisageables :

Sommaire

Sur chaque individu de l'échantillon sont mesurées maintenant deux variables quantitatives X et Y. Après l'analyse de chacune de ces variables, on procède à l'étude de la relation entre ces deux variables.

- Existe-t-il une liaison entre les deux variables ?
- Cette relation est-elle linéaire ou non linéaire ?

Nuage de points

La description de la liaison entre les deux variables se fait en premier lieu par un examen du nuage de points $(x_i,\ y_i)$. L'observation du nuage permet de justifier la linéarité ou non de la relation (des tests existent mais nécessitent des conditions d'application rarement atteintes).

Coefficient de corrélation linéaire de Peason

On procède dans le cas où la liaison est linéaire au calcul d'un paramètre de la liaison linéaire entre les deux variables, le coefficient de corrélation de Pearson : $r = \dfrac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y}$

- Plus r est proche de $+1$, plus les variables sont positivement corrélées, lorsqu'une variable augmente l'autre a tendance à augmenter en parallèle.
- Plus r est proche de -1, plus les variables sont négativement corrélées, lorsqu'une variable augmente l'autre a tendance à diminuer en parallèle.
- Quand r est proche de 0, les variables ne sont pas significativement corrélées, les variations d'une variable n'apporte pas d'information sur les variations de l'autre variable.

Test de corrélation linéaire

- Supposons que l'on soit dans une situation où l'emploi du coefficient de corrélation linéaire est justifié, et que l'on observe une valeur "élevée" de $|r|$.
- Quand peut-on dire que cette valeur est significativement non-nulle ?
- Il existe un test statistique permettant de tester l'hypothèse nulle $H_0 \ : \ r = 0$ dans le cas d'une liaison linéaire. Dans le cas où H_0 est rejetée, on réalise en général une **régression linéaire** (chapitre II) de Y en X ou de X en Y.
- Dans R, le coefficient de corrélation est obtenu par la fonction cor. Le test correspondant s'obtient à partir de la fonction cor.test.

Etude de l'exemple arbre

Une étude de la liaison entre l'âge (années) et le diamètre (cm) dans une population d'arbres

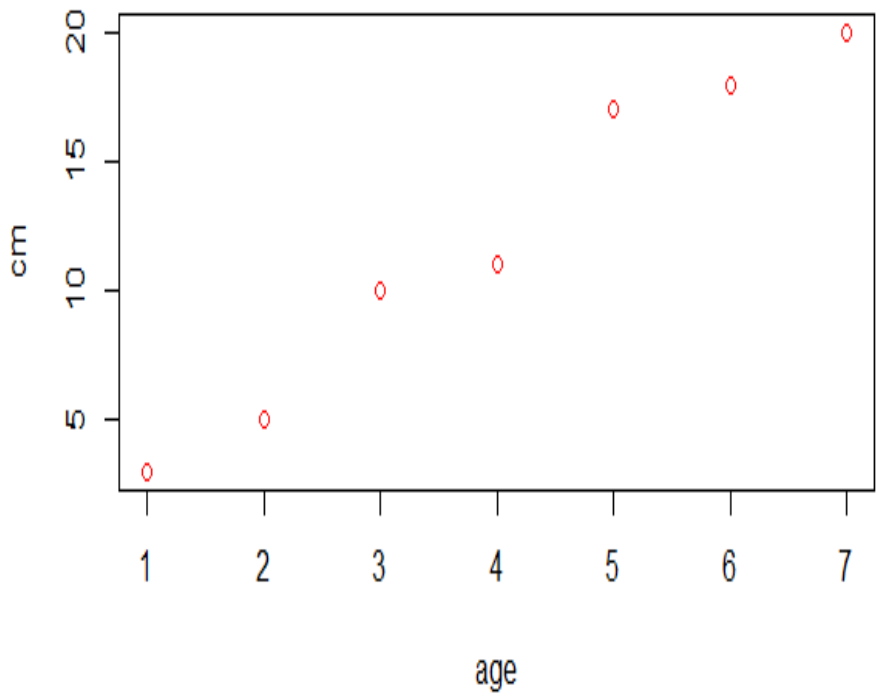
Age	1	2	3	4	5	6	7
Diamètres	3	5	10	11	17	18	20

Mise en oeuvre sous R

```
age = 1 :7
cm = c(3,5,10,11,17,18,20)
plot(age,cm,col="red")
cor(age,cm)
cor.test(age,cm)
```

Etude de l'exemple arbre

plot(age,cm,col="red")



Etude de l'exemple arbre

cor.test(age,cm)

Pearson's product-moment correlation
data : age and cm
$t = 12.55$, $df = 5$, $p.value = 5.702e-05$
alternative hypothesis : true correlation is not equal to 0
95 percent confidence interval :
0.8948996 0.9978013
sample estimates :
cor
0.9844952

Etude de l'exemple arbre

Interprétation des résultats

- $p.value = 5.702.10^{-05} < 5\% \Rightarrow$ on rejette l'hypothèse H_0. Ainsi les variables Age et Diamètres sont linéairement dépendents.
- Le coefficient de corrélation est estimé à 0.9844952 qui est positif et très proche de 1.
- Les variables Age et Diamètre sont fortement corrélées et évoluent dans le même sens.

Sommaire

Analyse de la variance

Principe de l'analyse

C'est une technique qui a été développée au début du siècle par l'agronome Fischer. L'objectif est toujours de mesurer l'effet d'un facteur qualitatif sur un caractère d'intérêt souvent quantitatif. Nous citons à titre d'exemples les cas où :

- On épand des niveaux d'engrais {faible, moyen, fort} (caractère X) sur des parcelles de blé et on relève le rendement à l'hectare (caractère Y).
- On nourrit des rats avec trois sources de protéines {boeuf, porc, soja} (caractère X) et on relève la prise de poids au bout d'un mois (caractère Y).

Représentation graphique

- Pour avoir une idée des distributions conditionnelles, on peut faire le graphe des boîtes à moustaches de la variable *y* sur chaque sous-population. On peut alors dessiner sur un même graphe les boîtes à moustaches de la variable *y* pour chaque modalité de la variable *x*.

- Dans l'hypothèse où x et y sont indépendantes, toutes ces boîtes à moustaches se ressemblent. Les différences visibles entre ces boîtes permettent de se faire une idée de l'influence de la variable y sur la variable x.

Rapport de corrélation

Une autre quantité qui permet de juger de la liaison entre x et y est le rapport dit de corrélation empirique entre la variance intercatégorie et la variance empirique totale :

$$e^2 = \frac{\frac{1}{n}\sum_{j=1}^{r} n_j(\bar{y}_j - \bar{y})^2}{\sigma_n^2(y)}$$

1. $\bar{y}_1, \ldots, \bar{y}_r$ les moyennes empiriques de la variable y sur chaque sous-échantillon.
2. $\sigma_n^2(y)$ est la variance empirique totale qui se décompose :

$$\sigma_n^2(y) = \frac{1}{n}\sum_{j=1}^{r} n_j(\bar{y}_j - \bar{y})^2 + \frac{1}{n}\sum_{j=1}^{r} n_j\sigma_j^2(y)$$

3. $\sigma_1^2(y), \ldots, \sigma_r^2(y)$ les variances empiriques de la variable y sur chaque sous-échantillon.

Exemple

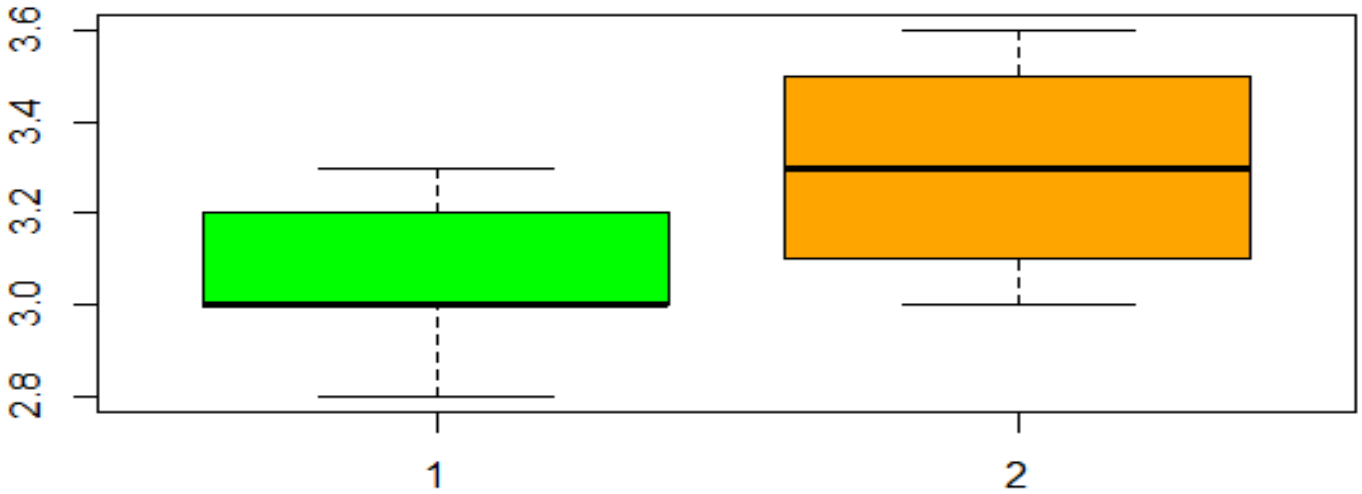
On étudie le poids des bébés en fonction du sexe et on obtient pour 5 filles $2.8;\ 3;\ 3;\ 3.2;\ 3.3$ et pour 4 garçons $3;\ 3.2;\ 3.4;\ 3.6$. On souhaite étudier l'influence du sexe sur le poids des bébés.

Mise en oeuvre sous R

```
> femme =c(2.8,3,3,3.2,3.3)
> homme=c(3,3.2,3.4,3.6)
> boxplot(femme,homme,col=c("green","orange"))
```

Exemple

Interprétation des résultats



Au vu de ce graphe, il semble bien qu'il y ait une influence du sexe sur le poids des bébés.

Sommaire

Représentation graphique

On peut représenter graphiquement les profils-lignes (ou les profils-colonnes) sous forme de diagrammes en barres parallèles. Les différences visibles entre ces barres permettent de se faire une idée de la liaison des variables x et y.

Test du χ_2 d'indépendance

Pour juger de la liaison entre X et Y , on peut aussi faire un test du χ_2 d'indépendance , basé sur la statistique

$$t_n = \sum_{i=1}^{r} \sum_{j=1}^{l} \frac{\left(n_{ij} - \frac{n_{i\bullet} n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet} n_{\bullet j}}{n}}$$

Exemple

Des pièces défectueuses ont été classées selon le type de défectuosité A, B, C ou D et la division 1, 2 ou 3 où la pièce a été produite. Les résultats suivants ont été obtenus

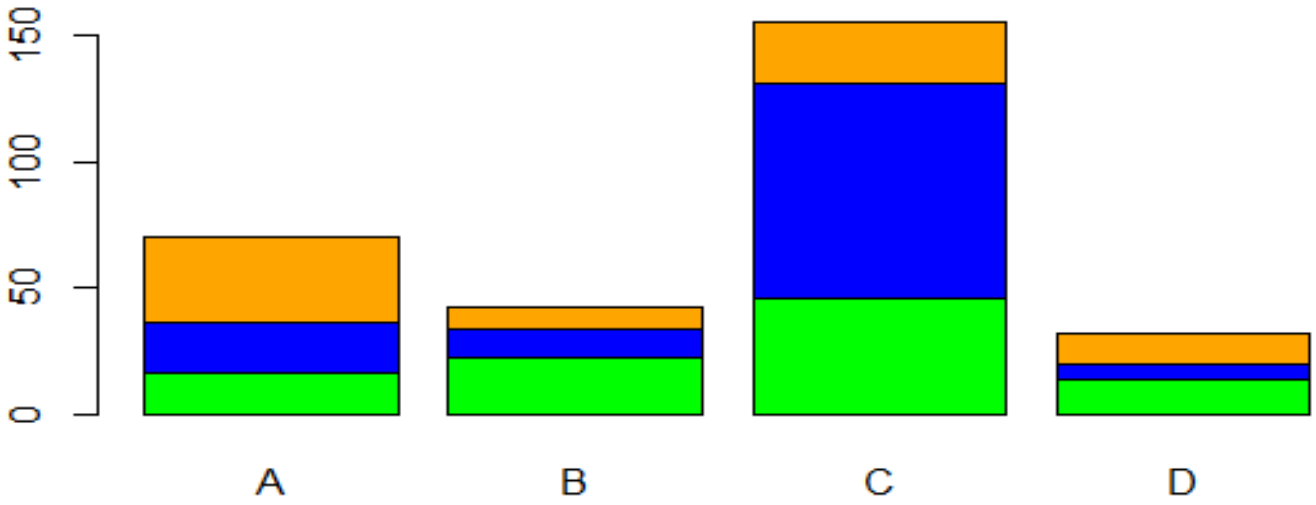
Division/Type de défectuosité	A	B	C	D
Division 1	16	22	46	14
Division 2	20	12	85	6
Division 3	34	8	24	12

Tester au seuil de 5% l'hypothèse d'indépendance du type de défectuosité et de la division.

Mise en oeuvre sous R

```
> tabcont = as.table(rbind(c(16,22,46,14),
c(20,12,85,6),c(34,8,24,12)))
> dimnames(tabcont) <- list(Division = c("D1", "D2","D3"), type =
c("A","B", "C","D"))
> barplot(tabcont,col=c("green","blue","orange"))
> chisq.test(tabcont)
```

Interprétation des résultats



Au vu de ce graphique, les deux variables "Division" et "type de défectuosité" semblent liées.

Interprétation des résultats

Pearson's Chi-squared test
data : tabcont
X-squared = 46.569, df = 6, p-value = 2.281e-08

$p.value = 2.281.10^{-08} < 5\% \Rightarrow$ on rejette l'hypothèse nulle H_0 d'indépendance. Ainsi, les variables "Division" et "type de défectuosité" sont liées.

Sommaire

Récapitulatif

Variable X explicative

Variable Y expliquée	Quantitatif	Qualitatif
Quantitatif	- Régression simple - Corrélation simple	Analyse de la variance à un facteur
Qualitatif	Régression logistique	Test du khi-deux d'indépendance