

Croisement de deux variables qualitatives

A.B. Dufour & M. Royer

L'objectif de cette séance est d'étudier des couples de variables qualitatives. Pour aller plus loin, consulter les fiches `tdr32` et `tdr321`.

Table des matières

Principe Général	1
Exercice 1	3
Exercice 2	5

Principe Général

Pour étudier le relation entre deux variables qualitatives (ou discrètes), on construit un tableau appelé table de contingence, deux paramètres : le chi-deux de contingence et le coefficient de Cramer.

Table de contingence

Soient A et B , deux variables qualitatives ayant respectivement p et q modalités. Soit n , le nombre d'individus sur lesquels A et B ont été observées. 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	...	Bj	...	Bq	total
A1	n_{11}	...	n_{1j}	...	n_{1q}	$n_{1.}$
⋮	⋮	⋱	⋮	⋱	⋮	⋮
Ai	n_{i1}	...	n_{ij}	...	n_{iq}	$n_{i.}$
⋮	⋮	⋱	⋮	⋱	⋮	⋮
Ap	n_{p1}	...	n_{pj}	...	n_{pq}	$n_{p.}$
total	$n_{.1}$...	$n_{.j}$...	$n_{.q}$	$n_{..}$

Prenons par exemple, les deux variables suivantes : sexe (F, M) et examen (R réussi, E échoué) observées sur 20 étudiants.

	R	E
F	4	6
M	6	4

Les marges permettent de définir que la table de contingence contient 10 hommes, 10 femmes, 10 étudiants qui réussissent l'examen et 10 étudiants qui échouent.

Valeur du Chi-Deux de contingence

Pour ce faire, on construit la table de contingence théorique : répartition des 20 étudiants entre les différentes cases de la table s'il n'y a aucun lien entre les deux variables sexe et examen, tout en tenant compte des marges.

	R	E
F	5	5
M	5	5

La valeur du Chi-Deux de contingence compare les effectifs de la table de contingence observée (EO) avec les effectifs de la table de contingence théorique (ET).

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

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. Dans l'exemple proposé, nous avons :

$$\chi^2 = \frac{(4-5)^2}{5} + \frac{(6-5)^2}{5} + \frac{(6-5)^2}{5} + \frac{(4-5)^2}{5} = 0.8$$

Le coefficient de Cramer

Afin d'évaluer le degré de relation entre les deux variables qualitatives, divers indices ont été proposés. Nous avons retenu 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.

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

Dans l'exemple proposé, nous avons $V = 0.2$ ce qui laisse penser que la réussite ou l'échec à l'examen ne sont pas liés au sexe.

Sous


Afin de faciliter les calculs, vous pouvez écrire la fonction ci-dessous qui, partant de deux variables qualitatives (`factor`) retourne la valeur du coefficient de Cramer.

```
cramer <- function(x, y) {
  res <- chisq.test(x, y, correct = FALSE)
  chi2 <- as.numeric(res$statistic)
  n <- length(x)
  p <- length(levels(x))
  q <- length(levels(y))
  m <- min(p - 1, q - 1)
  V <- sqrt(chi2/(n * m))
  return(V)
}
sexe <- as.factor(rep(c("F", "M"), c(10, 10)))
examen <- as.factor(c("R", "R", "R", "R", "E", "E", "E", "E",
"E", "E", "R", "R", "R", "R", "R", "R", "R", "E", "E", "E", "E"))
cramer(sexe, examen)
[1] 0.2
```

Exercice 1

Lors d'une enquête menée par B. Vignal, P. Chazaud et A.B. Dufour (1994)[1], sur le tourisme et les pratiques de loisir en Ardèche, 2953 personnes ont été interrogées à l'aide d'un questionnaire bilingue français / anglais. 591 d'entre eux proviennent de la zone du moyen Vivarais. Peut-on mettre en évidence une différence d'âge entre campeurs et non campeurs ?

Les données peuvent se présenter sous deux formes.

- a) Les résultats des questionnaires sont rentrés sous la forme d'un tableau contenant en lignes les enquêtés et en colonnes leurs réponses aux différentes questions. Nous avons conservé dans un tableau les résultats aux deux questions : âge et type d'hébergement pendant les vacances. Il se trouve sur le site pédagogique dans le menu "fichiers de données" : `hebergement.rda` et comprend 591 lignes et 2 colonnes. Pour le télécharger dans , utiliser la commande :

```
load(url("http://pbil.univ-lyon1.fr/R/donnees/hebergement.rda"))
head(hebergement)

  age    logement
1 55-59    camping
2 20-24    camping
3 16-19 non_camping
4 25-29 non_camping
5 25-29    camping
6 40-49 non_camping

names(hebergement)
[1] "age"      "logement"
```

- b) Les données se présentent sous la forme d'une table de contingence.

	Campeurs	Non Campeurs
16-19 ans	7	4
20-24 ans	43	21
25-29 ans	37	37
30-39 ans	99	78
40-49 ans	79	62
50-54 ans	19	25
55-59 ans	15	15
60-65 ans	9	15
Plus de 65 ans	2	24

Dans une approche simplifiée, saisir les données dans 2 vecteurs appelés `campeurs` et `noncampeurs`, puis créer un tableau qui réunit le tout grâce à la commande :

```
matrix(c(campeur,noncampeur),nrow=9)
```

- 1) Choisir le premier mode de lecture des données et construire la table de contingence correspondant à la relation entre âge et mode d'hébergement. Cette table peut être notée `touristes`.
- 2) Calculer l'effectif de chaque ligne et de chaque colonne. Par exemple, combien d'individus ont entre 20 et 24 ans? Combien de touristes ont opté pour le camping?
- 3) Que fait la commande `margin.table(touristes,1)`?
`margin.table(touristes,2)`?
- 4) Donner une représentation graphique de la répartition des touristes selon les 2 variables à l'aide de la commande :

```
library(ade4)
table.cont(touristes, csize = 2)
```

- 5) Utiliser les commandes `prop.table(touristes,1)` et `prop.table(touristes,2)` pour commenter les données.
- 6) Si les variables *type de logement* et *âge des touristes* étaient indépendantes, quel tableau d'effectifs aurait-on obtenu? Afficher ce tableau grâce aux commandes suivantes :

```
res <- chisq.test(touristes)
res$expected
```

- 7) Calculer l'effectif de chaque ligne et de chaque colonne pour ce nouveau tableau d'effectifs théoriques.
- 8) Pour calculer le coefficient du χ^2 de contingence, taper la commande `res$statistic`.
- 9) En déduire le coefficient de Cramer et interpréter le résultat obtenu.

Complément. Choisir le second mode de lecture des données, donner un nom aux lignes et aux colonnes retrouver les résultats précédents.

Exercice 2

Nous allons étudier les réponses au questionnaire qui se trouvent dans le fichier L3APA06.txt.

- 1) En fonction du sport pratiqué, la latéralité totale ou partielle est plus ou moins importante. Étudions si les variables main d'écriture (mécriture) et main qui tient la fourchette (mfourchette) sont liées.
- 2) A l'aide de `table`, créer la table de contingence appelée latmain qui réunit les variables en question.
- 3) Les variables main d'écriture et main qui tient la fourchette sont-elles liées ? Pour ce faire, utiliser la fonction `cramer` de l'introduction.
- 4) Dans certains sports tels que le tir à l'arc, la latéralité est indispensable. Existe-t-il un lien entre les variables œil de visée (oeil) et main d'écriture (mécriture) ?
- 5) Considérer 2 variables qualitatives de votre choix parmi celles de votre questionnaire et étudier si ces 2 variables sont liées.

Références

- [1] P. Chazaud, A-B. Dufour, and B. Vignal. Vers une typologie des campeurs. l'exemple de l'ardèche. *Cahiers Espaces*, 36 :61–68, 1994.