LSTAT2100 - Exercices - Série 1

Solutions

Exercice 1

On s'intéresse à la couleur des yeux d'individus dans un pays. Pour $n = 300$ individus pris au hasard, on a obtenu :

bleu	marron	noir	vert
48	122	95	35

(a) Peut-on dire que toutes les couleurs sont équiprobables ?

Solution:

```
O <- c(48, 122, 95, 35)
test <- chisq.test(O, p = rep(1 / 4, 4))
test
```

```
##
##  Chi-squared test for given probabilities
##
## data:  O
## X-squared = 66, df = 3, p-value = 3e-14
```

Rejet de H0 au seuil de 5%, donc nous rejetons l'hypothèse nulle que les proportions sont égales.

(b) Peut-on dire que les yeux foncés (marron et noir) sont deux fois plus probables que les yeux clairs (bleu et vert) ?

Solution:

```
chisq.test(c(O[1] + O[4], O[2] + O[3]), p = c(1 / 3, 2 / 3))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  c(O[1] + O[4], O[2] + O[3])
## X-squared = 4, df = 1, p-value = 0.04
```

(c) Peut-on dire que les yeux bleus et verts sont équiprobables ?

Solution:

Nous avons besoin des effectifs attendus sous l'hypothèse nulle:

$$H_0 : \ p_1 = p_4$$

Pour cela, calculons $\hat{p}_{01}, \hat{p}_{02}, \hat{p}_{03}$ et $\hat{p}_{04}$ les estimateurs de maximum de vraisemblance de p_1, p_2, p_3 et p_4 sous l'hypothèse nulle. La log-vraisemblance est

$$l_n = n_1 \log p_1 + n_2 \log p_2 + n_3 \log p_3 + n_4 \log p_4 + \text{const}, \ \text{avec } p_1 + p_2 + p_3 + p_4 = 1.$$

Sous H_0,

$$l_n = (n_1 + n_4) \log p_1 + n_2 \log p_2 + n_3 \log p_3 + \text{const}, \ \text{avec } p_3 = 1 - 2p_1 - p_2$$

Dès lors,

$$\frac{\partial l}{\partial p_1} = \frac{n_1 + n_4}{p_1} - 2\frac{n_3}{p_3} \text{ et } \frac{\partial l}{\partial p_2} = \frac{n_2}{p_2} - \frac{n_3}{p_3}.$$

En annulant ces deux dérivées partielles, on obtient

$$(n_1 + n_4)\hat{p}_{03} = 2n_3\hat{p}_{01} \text{ et et } n_2\hat{p}_{03} = n_3\hat{p}_{02}$$

En sommant les deux égalités, on obtient

$$(n_1 + n_2 + n_4)\hat{p}_{03} = n_3(2\hat{p}_{01} + \hat{p}_{02}) \Rightarrow (n_1 + n_2 + n_4)\hat{p}_{03} = n_3(1 - \hat{p}_{03}) \Rightarrow \hat{p}_{03} = \frac{n_3}{n}$$

Et en remplacant dans les expressions en haut, nous trouvons,

$$\hat{p}_{01} = \hat{p}_{04} = \frac{n_1 + n_4}{2n}; \ \hat{p}_{02} = \frac{n_2}{n}; \ \hat{p}_{03} = \frac{n_3}{n}$$

On peut en déduire:

```r
# 1. les effectifs attendus

E <- c((O[1] + O[4]) / 2, O[2], O[3], (O[1] + O[4]) / 2)
E
```

```
## [1]  41.5 122.0  95.0  41.5
```

```r
# 2. Stat. chi deux

X2 <- sum((E - O)^2 / E)
X2
```

```
## [1] 2.04
```

```r
# 3. pvaleur

pchisq(X2, df = 3 - 2, lower.tail = FALSE)
```

```
## [1] 0.154
```

Remarque: Si on compare les effectifs observés et attendus, on remarque qu'il n'y a que pour les yeux bleus et verts qu'on a des différences

```r
t(cbind(O, E))
```

O	48.0	122	95	35.0
E	41.5	122	95	41.5

On aurait pu donc prendre directement la sous-table composée des yeux verts et bleus et tester si les deux cas sont équiprobables:

```r
chisq.test(c(O[1], O[4]), p = c(1 / 2, 1 / 2))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  c(O[1], O[4])
## X-squared = 2, df = 1, p-value = 0.2
```

Exercice 2

Voici les fréquences (Freq) du nombre de passages de bus (nBus) par heure à un arrêt de bus. Ces données concernent une periode de 30 heures reparties sur 5 jours de semaine.

nBus	0	1	2	3	4	5
Freq	1	5	6	10	4	4

(a) Supposons que $nBus \sim Pois(\mu)$, $\mu > 0$. Tester l'hypothèse $H_0 : \mu = 3$ vs H_1: $\mu \neq 3$. Calculer un intervalle de confiance pour μ.

Solution:

```r
nBus <- 0:5
Freq <- c(1, 5, 6, 10, 4, 4)
data <- data.frame(nBus, Freq)
t(data)
```

nBus	0	1	2	3	4	5
Freq	1	5	6	10	4	4

```r
mu.0 <- 3
n <- sum(Freq)
# Le MLE de mu
mu.hat <- sum(Freq * nBus) / n
mu.hat
```

```
## [1] 2.77
```

```r
# Wald
z2 <- (sqrt(n) * (mu.hat - mu.0) / sqrt(mu.hat))^2
c(Stat.Wald = z2, pvaleur = pchisq(z2, df = 1, lower.tail = FALSE))
```

```
## Stat.Wald   pvaleur
##     0.590     0.442
```

```r
# Score
x2 <- (sqrt(n) * (mu.hat - mu.0) / sqrt(mu.0))^2
c(Stat.Score = x2, pvaleur = pchisq(z2, df = 1, lower.tail = FALSE))
```

```
## Stat.Score    pvaleur
##      0.544      0.442
```

```r
# LR
g2 <- 2 * n * (mu.hat * log(mu.hat / mu.0) - (mu.hat - mu.0))
c(Stat.LR = g2, pvaleur = pchisq(g2, df = 1, lower.tail = FALSE))
```

```
## Stat.LR pvaleur
##   0.559   0.455
```

```r
# Intervalle de confiance de Wald:
mu.hat + sqrt(mu.hat / n) * qnorm(0.975) * c(-1, 1)
```

```
## [1] 2.17 3.36
```

(b) En supposant que la fréquence moyenne de passage est de $\mu = 3$ par heure, testez l'ajustement d'une loi de Poisson aux données ?

Solution:

Tout d'abord, notez que l'information dont on dispose peut être écrite sous la forme suivante.

```r
cbind(t(data), c("+6", 0))
```

nBus	0	1	2	3	4	5	+6
Freq	1	5	6	10	4	4	0

Dès lors, l'hypothèse à tester est

$$H_0 : p_k = p_{0k}, \text{ pour } k = 0, \ldots, 5 \text{ et } p_6^+ = p_{06}^+,$$

où $p_k = P(nBus = k), p_k^+ = P(nBus \geq k), p_{0k} = P(Pois(3) = k)$, et $p_{0k}^+ = P(Pois(3) \geq k)$.

```r
# Effectifs attendus
p0 <- dpois(x = 0:5, lambda = 3)
p0 <- c(p0, 1 - sum(p0))
E <- 30 * p0
```

Freq	1.00	5.00	6.00	10.00	4.00	4.00	0.00
E	1.49	4.48	6.72	6.72	5.04	3.02	2.52

```r
x2 <- sum((c(Freq, 0) - E)^2 / E)
c(Stat.Pearson = x2, pvaleur = pchisq(x2, df = 6, lower.tail = FALSE))
```

```
## Stat.Pearson      pvaleur
##        4.947        0.551
```

ou, directement via la fonction `chisq.test` :

```r
chisq.test(x = c(Freq, 0), p = p0)
```

```
## Warning in chisq.test(x = c(Freq, 0), p = p0): Chi-squared approximation may be
## incorrect
```

```
##
##  Chi-squared test for given probabilities
##
## data:  c(Freq, 0)
## X-squared = 5, df = 6, p-value = 0.6
```

Remarque: Le résultat de ce test est à considérer avec prudence puisque les effectifs attendus sont faibles (voir cours). Refaites ce test en regroupant les modalités "nBus=0" et "nBus=1" d'une part et "nBus=5" et "nBus=6+" d'autre part, i.e., **refaites le test en considérant que les données sont**

nBus0	-1	2	3	4	+5
Freq0	6	6	10	4	4

Une meilleure façon d'éviter ce problème est d'effectuer un test exact par simulation (non vus dans le cours). Avec R cela donne:

```r
chisq.test(x = c(Freq, 0), p = p0, simulate.p.value = TRUE)
```

```
##
##  Chi-squared test for given probabilities with simulated p-value (based
##  on 2000 replicates)
##
## data:  c(Freq, 0)
## X-squared = 5, df = NA, p-value = 0.5
```

(c) Refaire le même test, sans supposer que $\mu = 3$.

Solution:

À la différence du cas précédent il faudra ici estimer μ par la méthode de maximum de vraisemblance. Le MLE de μ est

```r
mu.hat
```

```
## [1] 2.77
```

Par la suite on applique la même démarche que pour la question précédente, mais, attention, le degré de liberté change de 6 à 5.

```r
# Effectifs attendus
p0 <- dpois(x = 0:5, lambda = mu.hat)
p0 <- c(p0, 1 - sum(p0))
E <- 30 * p0
x2 <- sum((c(Freq, 0) - E)^2 / E)
c(Stat.Pearson = x2, pvaleur = pchisq(x2, df = 5, lower.tail = FALSE))
```

```
## Stat.Pearson      pvaleur
##        5.084        0.406
```

Exercice 3

Dans cet exercice, nous allons utiliser le jeu de données data.csv. Ce jeu de données comprend des observations liées à des clients qui ont contracté un crédit. Nous avons à notre disposition un certain nombre de variables, dont la variable Y qui nous dit si le client a pu rembourser dans les temps le crédit ($Y = 1$) ou pas ($Y = 2$).

Note: Bien qu'il y ait une multitude de variables disponibles, nous allons nous cantonner qu'à un sous ensemble de variables pour cet exercice.

Charger les données avec la commande suivante (cette commande suppose que votre répertoire de travail contient un répertoire `data` qui contient le fichier `data.csv`).

```r
mdata <- read.csv(file = "Data/data.csv")
mdata$Y <- factor(mdata$Y)
levels(mdata$Y) <- c("PasDefaut", "Defaut") # Pour plus de lisibilité.
```

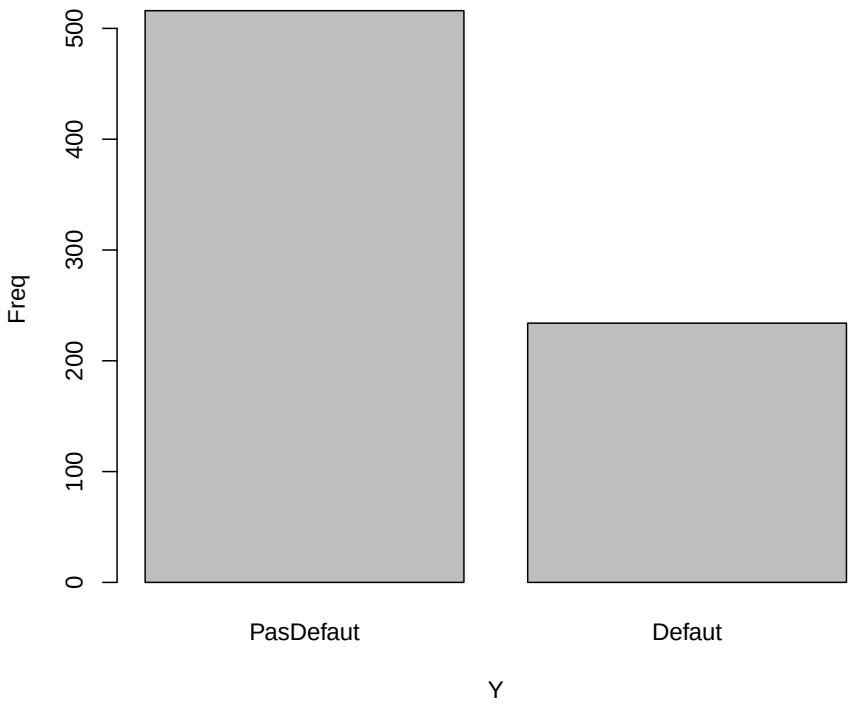
(a) À l'aide de la fonction `xtabs`, calculer le nombre d'observations dans chaque niveau de la variable Y. Visualisez cette information dans un graphique adéquat.

Solution:

```r
tbl <- xtabs(~Y, data = mdata)
tbl
```

PasDefaut	Defaut
516	234

```r
barplot(tbl, xlab = "Y", ylab = "Freq")
```



(b) Construisez la table de contingence croisant les variables $X = check\_ac$ et Y et calculez les fréquences $\hat{P}(Y = j|X = i)$.

Note: La variable *check_ac* prend 4 valeurs et correspond aux flux moyens rentrants chaque mois sur le compte courant des clients:

- A11: <0
- A12: $[0, 200)$
- A13: ≥ 200
- A14: Pas de compte courant.

Solution:

```
mdata$check_ac <- factor(mdata$check_ac)

tbl <- xtabs(~ check_ac + Y, data = mdata)
tbl
```

check_ac/Y	PasDefaut	Defaut
A11	103	105
A12	115	81
A13	33	10
A14	265	38

```r
p1tbl <- prop.table(tbl, 1) # 1: by row, 2: by col
p1tbl
```

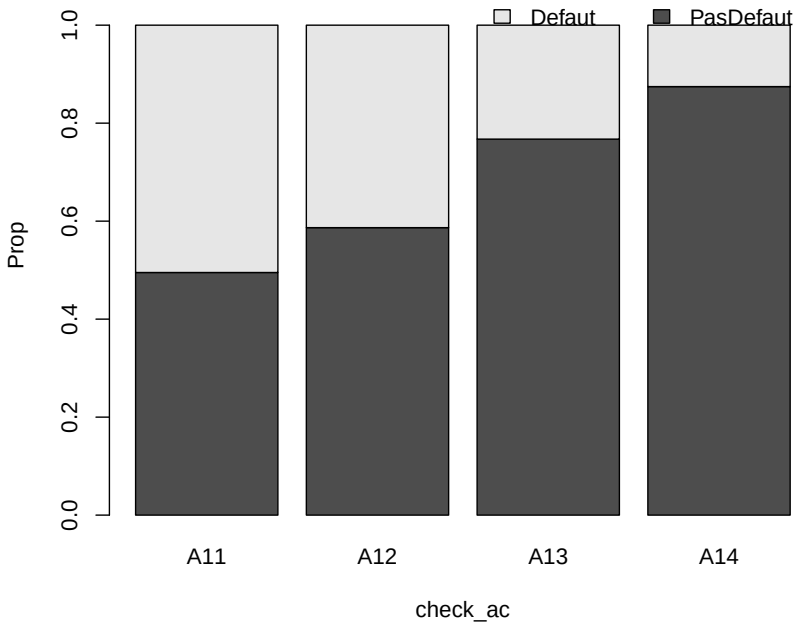
check_ac/Y	PasDefaut	Defaut
A11	0.495	0.505
A12	0.587	0.413
A13	0.767	0.233
A14	0.875	0.125

(c) Utiliser deux graphiques différents pour visualiser le lien entre ces deux variables.

Solution:

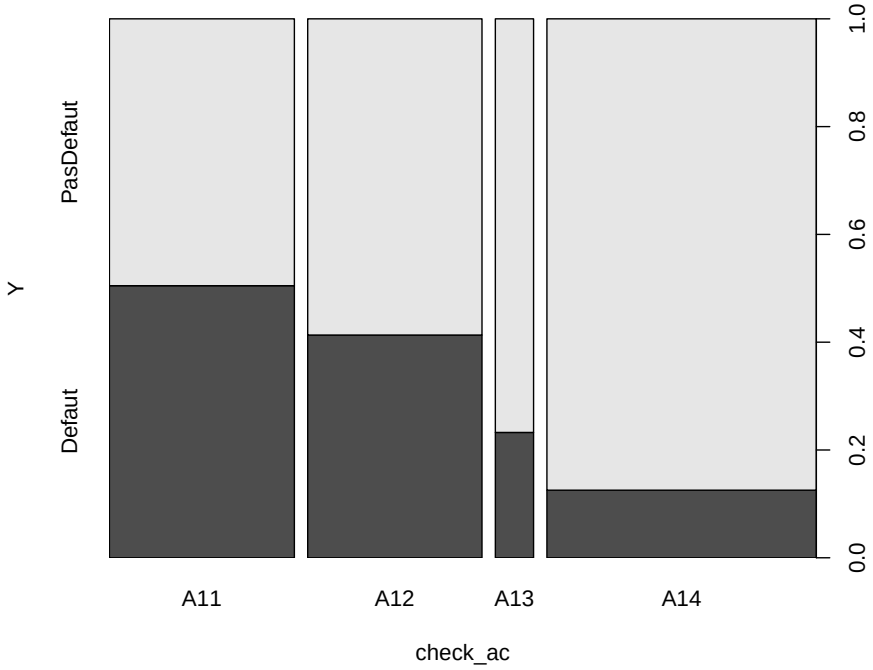
Barplot :

```r
barplot(t(p1tbl),
  ylim = c(0, 1.1), legend = TRUE, ylab = "Prop", xlab = "check_ac",
  args.legend = list(nco = 2, bty = "n")
)
```



spineplot/mosaicplot :

```r
spineplot(tbl)
```



(d) Tester l'indépendance entre ces deux variables.

Solution:

Vous avez plusieurs choix/fonctions pour effectuer le test:

```
summary(tbl)
```

```
## Call: xtabs(formula = ~check_ac + Y, data = mdata)
## Number of cases in table: 750
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 96, df = 3, p-value = 1e-20
```

```
# ou
chisq.test(tbl)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tbl
## X-squared = 96, df = 3, p-value <2e-16
```

```
# ou à l'aide du package vcd
assocstats(tbl)
```

```
##                      X^2 df P(> X^2)
## Likelihood Ratio 101.471  3        0
## Pearson           95.793  3        0
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.337
## Cramer's V        : 0.357
```

Rejet de H0 au seuil de 5%, donc nous rejetons l'hypothèse nulle d'indépendance entre les 2 variables; il existe bien une association significative entre ces variables.

(e) Etudier le lien entre les variables *tel* et Y, à l'aide de l'odds ratio. Interpréter.

Note: La variable **tel** nous renseigne si la personne a un numéro de téléphone (le jeu de données n'est pas tout récent...)

- A191 : Non
- A192 : Oui

Solution:

Nous pouvons utiliser la fonction `loddsratio` du package `vcd`

```
mdata$tel <- factor(mdata$tel)
levels(mdata$tel) <- c("Non", "Oui") # Pour plus de lisibilité.

tbl <- xtabs(~ tel + Y, data = mdata)
tbl
```

tel/Y	PasDefaut	Defaut
Non	305	142
Oui	211	92

```
or <- loddsratio(tbl, log = FALSE)
or
```

```
##  odds ratios for tel and Y
##
## [1] 0.937
```

L'odds ratio est inférieur à 1, donc l'événement "rembourser son crédit à temps" est moins fréquent dans le groupe des individus qui n'ont pas de téléphone que dans le groupe d'individus ayant un téléphone.

(f) Sur base de l'oddsratio, que pouvez-vous dire sur l'indépendance entre *tel* et Y.

Solution:

```r
# Nous pouvons construire un IC autour du OR et regarder s'il contient 1.
confint(or)
```

	2.5 %	97.5 %
Non:Oui/PasDefaut:Defaut	0.683	1.28

L'intervalle de l'odds ratio contient 1, donc rembourser son crédit à temps est indépendant du fait d'avoir un téléphone ou non.

Exercice 4

Des statistiques sur la réussite ou non des "Lancer francs" en Basketball ont été récoltées. Il y a eu 5 lancers où aucun des deux tirs n'a été un succès, 34 où uniquement le premier tir a été un succès, 48 où uniquement le deuxième tir a été un succès et 251 tirs avec deux succès.

Notons X la variable aléatoire binaire du **premier tir** ($1 =$ Échec, $2 =$ Succès), et Y la variable aléatoire binaire du **second tir** ($1 =$ Échec, $2 =$ Succès).

(a) Construisez un tableau de contingence à partir de ces statistiques.

Solution:

```r
O <- c(5, 34, 48, 251)
lancer <- data.frame(X = c(1, 2, 1, 2), Y = c(1, 1, 2, 2), Freq = O)
tab <- xtabs(Freq ~ X + Y, data = lancer)
tab
```

X/Y	1	2
1	5	48
2	34	251

(b) Tester l'indépendance entre les deux tirs (i.e. entre les variables X et Y.).

Solution:

```r
summary(tab)
```

```
## Call: xtabs(formula = Freq ~ X + Y, data = lancer)
## Number of cases in table: 338
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 0.27, df = 1, p-value = 0.6
```

(c) Soit $p_{ij} = P(X = i, Y = j)$, $i, j = 1, 2$. Considérons l'hypothèse suivante

$$H_0 : \ p_{11} = \theta^2, p_{12} = p_{21} = \theta(1-\theta), \text{ et } p_{22} = (1-\theta)^2$$

Montrer que sous H_0, X et Y sont indépendants et identiquement distribués.

Solution:

X et Y sont identiquement distribuées puisque

$$p_{1.} = p_{.1} = \theta \text{ et } p_{2.} = p_{.2} = 1 - \theta.$$

Ces variables sont indépendantes car

$$p_{ij} = p_{i.}p_{.j} \;\; i, j = 1, 2.$$

(d) En supposant un échantillonnage multinomial simple, donner l'estimateur du maximum de vraisemblance de θ et calculer le.

Solution:

Nous avons que le log-vraisemblance est donné par

$$l_n(\theta) = 2n_{11} \ln \theta + 2n_{22} \ln(1 - \theta) + (n_{12} + n_{21}) \ln \theta(1 - \theta) + const$$

Dès lors,

$$\begin{aligned} & \frac{dl_n}{d\theta}(\theta) = 0 \\ \iff & 2n_{11}(1 - \theta) - 2n_{22}\theta + (n_{12} + n_{21})(1 - 2\theta) = 0 \\ \iff & 2n\theta = 2n_{11} + n_{12} + n_{21} \\ \iff & \theta = \frac{n_{1.} + n_{.1}}{2n} = \frac{p_{1.} + p_{.1}}{2} \end{aligned}$$

Donc

$$\hat{\theta} = \frac{\hat{p}_{1.} + \hat{p}_{.1}}{2}.$$

On peut calculer cet estimateur "manuellement" ou en utilisant R à l'aide du code suivant

```
mptab <- addmargins(prop.table(tab))
mptab
```

X/Y	1	2	Sum
1	0.015	0.142	0.157
2	0.101	0.743	0.843
Sum	0.115	0.885	1.000

```
theta <- (mptab[1, 3] + mptab[3, 1]) / 2
theta
```

```
## [1] 0.136
```

(e) Proposer une statistique de test pour tester H_0. Effectuez le test et concluez.

Solution:

On peut utiliser le test du rapport de vraisemblance. Sa statistique est donnée par (voir cours)

$$G^2 = 2 \sum O \log \left(\frac{O}{E} \right),$$

avec $O = (N_{11}, N_{12}, N_{21}, N_{22})$ et $E = (n\hat{\theta}^2, n\hat{\theta}(1-\hat{\theta}), n\hat{\theta}(1-\hat{\theta}), n(1-\hat{\theta})^2)$. çàd

$$G^2 = 2 \left(N_{11} \ln \frac{N_{11}}{n\hat{\theta}^2} + N_{12} \ln \frac{N_{12}}{n\hat{\theta}(1-\hat{\theta})} + N_{21} \ln \frac{N_{21}}{n\hat{\theta}(1-\hat{\theta})} + N_{22} \ln \frac{N_{22}}{n(1-\hat{\theta})^2} \right).$$

Sous H_0, Cette variable suit asymptotiquement une distribution chi-deux de $3-1$ degrés de liberté. En effet, pour la vraisemblance non contrainte il y a trois paramètres à estimer (à savoir p_{11}, p_{12}, et p_{21}) alors qu'il n' y a qu'un seul paramètre à estimer sous H_0 (à savoir θ).

```
n <- sum(tab) # Nombre d'observations
E <- n * c(theta^2, theta * (1 - theta), theta * (1 - theta), (1 - theta)^2)
# la statistique de rapport de vraisemblance
g2 <- 2 * sum(O * log(O / E))
g2
```

```
## [1] 2.76
```

```
# pvaleur
pchisq(g2, df = 2, lower.tail = FALSE)
```

```
## [1] 0.252
```