LSTAT2100 - Exercices - Série 1

Solutions

Exercice 1

Nous nous intéressons à la couleur des yeux d'une certaine population. Pour $n = 300$ individus, pris au hasard, nous avons observé les chiffres suivants.

bleu	marron	noir	vert
48	122	95	35

(a) Ces couleurs sont-elles toutes réparties de manière uniforme (équiprobables) ? Répondez à cette question eu utilisant le test de LR. Calculez la $p-$valeur de ce dernier en utilisant (i) la théorie asymptotique, et (ii) des simulations.

Solution:

Soit p_k, $k = 1, \dots, 4$, les proportions des couleurs (bleu, vert, marron, noir). L'hypothèse à tester est la suivante

$$H_0: \, p_1 = p_2 = p_3 = p_4 = 1/4.$$

```r
# Les données
O <- c(48, 122, 95, 35)
E <- 300*rep(1 / 4, 4)
```

```r
# LR asymptotique
(2*sum(O * log(O/E))) |> print() |>  pchisq(df = 3, lower.tail = FALSE)
```

```
## [1] 67.4
## [1] 1.51e-14
```

```r
# LR via des simulations
{rmultinom(10000, size = 300, prob = rep(1 / 4, 4)) |>
apply(2, FUN = \(O) 2*sum(O * log(O/E))) >= 67.4} |> mean()
```

```
## [1] 0
```

Au seuil de 5%, nous rejetons donc 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) ? Utilisez le test de Pearson.

Solution:

Ici, le but est de tester $H_0: \, p_1 + p_4 = p_2 + p_3$.

```r
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
```

Au seuil de 5%, nous rejetons l'hypothèse H_0.

(c) Les proportions des yeux bleus et des yeux verts sont-elles les mêmes ? Utilisez un test LR. Détaillez votre approche et vos calculs.

Solution:

Ici, le but est de tester $H_0: \ p_1 = p_4$.
Pour réaliser le test, nous avons besoin des effectifs attendus sous H_0. 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 H_0. 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 remplaçant dans les expressions ci-dessus, on trouve,

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

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

O	48.0	122	95	35.0
E	41.5	122	95	41.5

```r
sum((E - O)^2 / E) |> pchisq(df = 1, lower.tail = FALSE)
```

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

$\Rightarrow$ Non-rejet de H_0, à 5%.

Remarque: Dans ce cas particulier, on peut tester la légalité des deux proportions en ne considérant que la sous-table composée des yeux verts et bleus

bleu	vert
48	35

```
Ob <- c(48, 35)
Eb <- c((Ob[1] + Ob[2]) / 2, (Ob[1] + Ob[2]) / 2)
sum((Eb - Ob)^2 / Eb) |> pchisq(df = 1, lower.tail = FALSE)
```

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

Cela fonctionne car les effectifs attendus des autres modalités (marron, noir) sont identiques aux effectifs observés (voir le tableau `t(cbind(O, E))` ci-dessous).

Exercice 2

Voici les fréquences ($Freq$) du nombre de passages de bus ($nBus$) par heure à un point d'arrêt. Ces données concernent une période de 30 heures réparties sur 5 jours de la 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: $\mu \neq 3$. Pour ce faire, utilisez les trois tests classiques vus dans le cours. Calculez un intervalle de confiance pour μ.

Solution:

```
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

L'EMV de μ n'est rien d'autre que la moyenne (voir cours)

```
n <- sum(Freq)
mu.hat <- {sum(Freq * nBus) / n} |> print()
```

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

En utilisant les formules vues au cours, nous calculons les trois statistiques et leurs $p-$valeurs de la manière suivante

```
mu.0 <- 3

# Wald
{(mu.hat - mu.0)^2*(n/mu.hat)} |> print() |> pchisq(df = 1, lower.tail = FALSE)
```

```
## [1] 0.59
```

```
## [1] 0.442
```

```
# Score
{(mu.hat - mu.0)^2*(n/mu.0)} |> print() |> pchisq(df = 1, lower.tail = FALSE)
```

```
## [1] 0.544
## [1] 0.461
```

```
# LR
{2 * n * (mu.hat * log(mu.hat / mu.0) - (mu.hat - mu.0))} |> print() |>
  pchisq(df = 1, lower.tail = FALSE)
```

```
## [1] 0.559
## [1] 0.455
```

Pour l'intervalle de confiance, nous pouvons utiliser le fait que $\hat{\mu} \sim_a N(\mu, \mu/n)$, pour établir l'intervalle (de Wald, asymptotique) suivant

```
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 ($H_0 : nBus \sim Pois(3)$) aux données ? *Proposez un test pertinent.*

Solution:

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

```
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

$$H_0 : p_k = p_k^0, \text{ pour } k = 0, \ldots, 5 \text{ et } p_{6+} = p_{6+}^0,$$

où $p_k = P(nBus = k), p_{6+} = P(nBus \geq 6), p_k^0 = P(Pois(3) = k)$, et $p_{+6}^0 = P(Pois(3) \geq 6)$.

```
# Effectifs observés/attendus
p0 <- dpois(x = 0:5, lambda = 3)
p0 <- c(p0, 1 - sum(p0))
E <- 30 * p0
O <- c(Freq, 0)
data.frame(O = O, E = E) |> t()
```

O	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

Nous pouvons alors calculer la statistique de Peasron et sa $p-$valeur comme suit

```
sum((O - E)^2 / E) |> print() |> pchisq(df = 6, lower.tail = FALSE)
```

```
## [1] 4.95
## [1] 0.551
```

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

```r
chisq.test(x = O, p = p0)
```

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

```
##
##  Chi-squared test for given probabilities
##
## data:  O
## X-squared = 4.95, df = 6, p-value = 0.55
```

Ce résultat doit être considéré avec prudence, car les effectifs attendus sont faibles (< 5). Au lieu de nous référer à la distribution asymptotique de χ^2_6, nous pouvons refaire le test en calculant la $p-$valeur par simulation

```r
chisq.test(x = O, p = p0, simulate.p.value = TRUE, B = 10000)
```

```
##
##  Chi-squared test for given probabilities with simulated p-value (based
##  on 10000 replicates)
##
## data:  O
## X-squared = 4.95, df = NA, p-value = 0.56
```

(c) Refaites le même test que celui de la question précédente mais, cette fois, sans supposer que $\mu = 3$; càd testez $H_0 : nBus \sim Pois(\mu)$, pour une μ fixe mais inconnue.

Solution:

À la différence du cas précédent il faudra ici estimer μ par la méthode de maximum de vraisemblance. Nous savons que l'EMV de μ est

```r
mu.hat
```

```
## [1] 2.7667
```

Par la suite on applique la même démarche que pour la question précédente, mais, attention, le degré de liberté change de 6 à 5 (à cause de l'estimation de μ).

```r
# Effectifs observés/attendus
p0 <- dpois(x = 0:5, lambda = mu.hat)
p0 <- c(p0, 1 - sum(p0))
E <- 30 * p0
data.frame(O = O, E = E) |> t()
```

O	1.0000	5.0000	6.0000	10.0000	4.0000	4.0000	0.0000
E	1.8861	5.2183	7.2187	6.6572	4.6046	2.5479	1.8672

```r
# Test de Pearson - asymptotique
sum((O - E)^2 / E) |> print() |> pchisq(df = 5, lower.tail = FALSE)
```

```
## [1] 5.084
```

```
## [1] 0.40572
```

```r
# Test de Pearson - simulation
{rmultinom(10000, size = 30, prob = p0) |>
apply(2, FUN = \(O) sum((O - E)^2 / E)) >= 5.084} |> mean()
```

```
## [1] 0.5346
```

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 son crédit ($Y = 1$) ou pas ($Y = 2$).

`data.csv` contient de nombreuses autres variables, mais nous n'en utiliserons que quelques-unes dans cet exercice. Nous découvrirons les variables d'intérêt à mesure que nous avançons, les autres sont à oublier.

Chargez les données avec la commande suivante, qui présume que votre répertoire de travail comprend `Data/data.csv`.

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

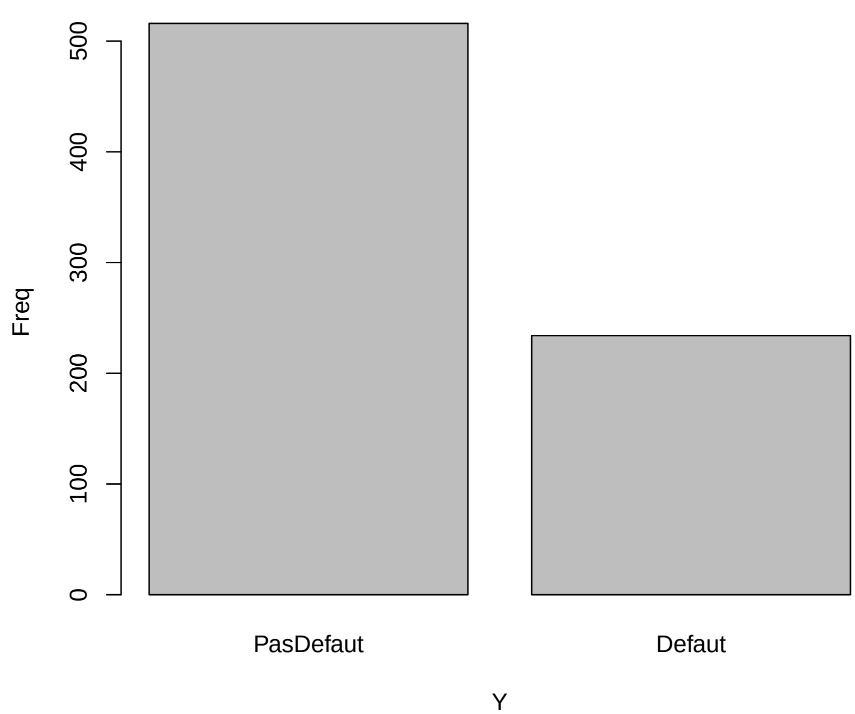
(a) Donnez le tableau des fréquences pour Y. Représentez ce tableau à l'aide d'un graphique approprié.

Solution:

```
tbl <- xtabs(~Y, data = mdata) |> print()      #ou table(mdata$Y)
```

```
## Y
## PasDefaut    Defaut
##       516       234
```

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



(b) La variable `check_ac` prend 4 modalités, correspondant aux flux moyens rentrants chaque mois sur le compte courant du client:

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

Construisez la table de contingence croisant les variables $X = check\_ac$ et Y et estimez les proportions $P(Y|X)$, pour les différentes valeurs de X et Y.

Solution:

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

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

```
##         Y
## check_ac PasDefaut Default
##      A11       103    105
##      A12       115     81
##      A13        33     10
##      A14       265     38
```

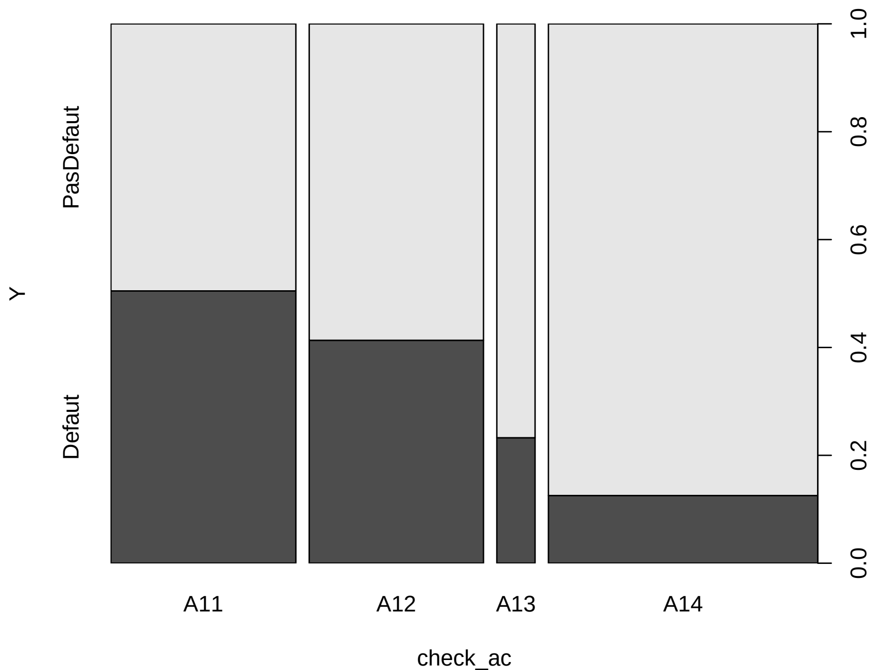
```r
proportions(tbl, "check_ac")
```

check_ac/Y	PasDefaut	Defaut
A11	0.49519	0.50481
A12	0.58673	0.41327
A13	0.76744	0.23256
A14	0.87459	0.12541

(c) Faites un graphique pour représenter la distribution (marginale) de X et la distribution (conditionnelle) de $Y|X$. Que suggère ce graphique quant à l'association entre X et Y ?

Solution:

```r
spineplot(tbl)
```



Ce graphique montre une forte association entre les deux variables étudiées.

(d) Tester l'indépendance entre X et Y.

Solution:

Nous avons plusieurs choix/fonctions pour effectuer le test:

```
summary(tbl)
# ou
chisq.test(tbl)
# ou
vcd::assocstats(tbl)$chisq_tests
```

	X^2	df	P(> X^2)
Likelihood Ratio	101.471	3	0
Pearson	95.793	3	0

$\Rightarrow$ il existe bien une association très significative (à 5%) entre les deux variables étudiées.

(e) La variable **tel** nous renseigne si le client a un numéro de téléphone (le jeu de données que nous étudions n'est pas très récent, mais cela n'a pas d'importance !). Cette variable prends deux valeurs:

- "A191": pour Non, le client n'a pas de téléphone
- "A192": pour Oui, le client a un téléphone

Calculez le rapport de cotes entre *tel* et Y. Quelle information peut-on en tirer sur l'association entre ces deux variables ? Compléter par un test statistique (d'indépendance) fondé sur le ratio calculé.

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) |> print()
```

```
##      Y
## tel   PasDefaut Defaut
##   Non       305    142
##   Oui       211     92
```

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

```
##  odds ratios for tel and Y
##
## [1] 0.93652
```

Le rapport de cotes est proche de 1, ce qui indique que la proportion de personnes qui remboursent leur crédit à temps (ou non) est similaire chez les détenteurs de téléphone et les non-détenteurs. En effet,

```
tbl |> proportions("tel")
```

tel/Y	PasDefaut	Defaut
Non	0.68233	0.31767
Oui	0.69637	0.30363

Nous pouvons construire un intervalle de confiance pour le rapport de cotes. Ensuite, il suffira de vérifier si cet intervalle contient la valeur 1 ou non pour confirmer ou infirmer l'indépendance.

```
confint(or)
```

	2.5 %	97.5 %
Non:Oui/PasDefaut:Defaut	0.68305	1.2841

L'intervalle du rapport des cotes contient 1, ce qui signifie que le fait de rembourser son crédit à temps est indépendant de la possession ou non d'un téléphone.

Exercice 4

Soit le tableau de de contingence suivant

X/Y	1	2
1	5	48
2	34	251

La signification des variables X et Y n'est pas importante pour la suite.

(a) Tester l'indépendance entre X et Y.

Solution:

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

X/Y	1	2
1	5	48
2	34	251

```
summary(tab)$p.value
```

```
## [1] 0.6015
```

(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

```r
mptab <- tab |> proportions() |> addmargins() |> print()
```

```
##      Y
## X              1        2      Sum
##   1   0.014793 0.142012 0.156805
##   2   0.100592 0.742604 0.843195
##   Sum 0.115385 0.884615 1.000000
```

```r
theta <- {(mptab[1, 3] + mptab[3, 1]) / 2} |> print()
```

```
## [1] 0.13609
```

(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).

```r
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] 2.7601
```

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

```
## [1] 0.25156
```