LSTAT2100 - Exercices - Série 1

Énoncés

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.

(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.

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

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.

(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.*

(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.

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.

(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.

(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 ?

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

(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.

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.

(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.

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

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