Statistiques décisionnelles : TD 3

1 Complément sur le test du χ^2

En cours, nous avons vu que le test du χ^2 permettait d'obtenir un test asymptotique pour l'adéquation à une loi de support fini (un nombre fini d de variables). Nous verrons en cours qu'on peut adapter cette démarche pour tester l'adéquation à une loi discrète à support dans $\mathbb{N}$. Pour obtenir un nombre fini de classes, on regroupe en une seule classe toutes celles dont les effectifs sont trop petits. Par exemple, pour une loi géométrique de paramètre $1/2$ avec 50 variables observées, on pourra créer le classe $X \geq 5$ et y regrouper toutes les observations plus grandes que 5. Les classes considérées doivent toutes satisfaire la condition $np_i(1-p_i) \geq 5$.

2 Utilisation de R

Voici quelques commandes utiles pour faire ces exercices à l'aide de R.

— $X = c(0.1, 0.2, 0.7)$ crée le vecteur $(0.1, 0.2, 0.7)$ et lui attribue le nom X.

— `chiq.test` est une fonction qui fait beaucoup de choses différentes (nous verrons au moins 3 utilisations dans ce semestre). Regardez la documentation. Vous voulez appliquer le test du χ^2 avec un seul échantillon et une probabilité cible. Pour un vecteur d'occurences X et un vecteur de probabilité $p1$, la commande `chisq.test(X,p=p1)` applique le test du χ^2 avec pour probabilité cible p_1.

3 Exercices

Exercice 1. *Une enquête sur* 160 *familles de quatre enfants tirées au hasard dans une population de familles permet d'établir la répartition suivante :*

Nombre de filles	*0*	*1*	*2*	*3*
Nombre de familles	*20*	*48*	*62*	*30*

1. *Sous l'hypothèse de l'équirépartition du sexe à la naissance et de l'indépendance du sexe lors de naissances différentes, quelle est la loi du nombre de filles d'une famille de* 3 *enfants ?*
2. *Donner un test de niveau* 10% *pour tester si cette répartition est compatible avec les données observées.*

Un nouvelle enquête sur 160 *familles est réalisée et on obtient les données suivantes*

Nombre de filles	*0*	*1*	*2*	*3*	*4*
Nombre de familles	*16*	*48*	*62*	*30*	*4*

Pouvez-vous appliquer le test du χ^2 *avec ces* 5 *valeurs ?*

Exercice 2. *On a récolté le nombre d'accidents mortels par heure sur un long week-end aux Etats-Unis.*

Nombre d'accidents mortels par h.	*0 ou 1*	*2*	*3*	*4*	*5*	*6*	*7*	*8 ou plus*
Nombre d'heures	*5*	*8*	*10*	*11*	*11*	*9*	*8*	*10*

Peut-on considérer que le nombre d'accidents mortels par heure survenant un long week-end suit une loi de Poisson de paramètre 5 *?*

Si $X \sim Poiss(5)$*, on pourra calculer avec R les valeurs de* $\mathbb{P}(X = k)$ *pour* k *entre* 0 *et* 7 *et calculer* $\mathbb{P}(X \geq 8)$ *de manière approximative, ou bien utiliser 1-ppois(7,5).*

Exercice 3. *Une questionnaire a été distribué aux* 56 *étudiants et étudiantes de L3MIASHS de l'université de Lille pour connaître leur niveau de satisfaction sur le cours de R.Butez, noté de* 1 *à* 5[1]*, et on a représenté les résultats dans le tableau ci-dessous.*

Niveau de satisfaction	*1*	*2*	*3*	*4*	*5*
Nombre de réponses	*5*	*6*	*10*	*15*	*20*

1. *Ce tableau est-il compatible avec une distribution uniforme des réponses ?*
2. *On soupçonne l'enseignant d'avoir changé certaines notes basses en notes hautes pour se faire bien voir de sa hiérarchie. Si on estime qu'il a transféré la moitié des* 1 *et* 2 *vers* 4 *et* 5*, qu'elle serait la loi observée si la distribution de départ était uniforme ?*
3. *Les réponses observées sont-elles compatibles avec une telle manipulation ?*

Exercice 4. *On observe le vecteur suivant*

0.67 0.70 1.6 82.06 2.31 2.80 3.10 4.79 8.09 8.38

Ces observations sont-elles compatibles avec une loi du $\chi^2(4)$ *(pour un niveau* 10%*) ? Quel test appliquez-vous ? On s'aidera de la fonction* `pchisq` *en R, où pchisq(5.5,4) calcule la probabilité qu'une variable du* $\chi^2(4)$ *soit inférieure ou égale à* 5.5*.*

Exercice 5 (χ^2 pour les lois continues)**.** *Il est possible d'adapter le test du* χ^2 *pour tester l'adéquation à une loi continue, et c'est le but de cet exercice dans le cas particulier de lois uniformes sur* $[0, 1]$*.*

On dispose de 100 *réalisations indépendantes* X_i *de loi inconnue, et on souhaite tester si elles proviennent d'une loi uniforme sur* $[0, 1]$*. Malheureusement, il vous est interdit d'utiliser le test de Kolmogorov-Smirnov. Vous disposez uniquement du tableau récapitulatif ci-dessous.*

Intervalles	$[0, 0.2[$	$[0.2, 0.4[$	$[0.4, 0.6[$	$[0.6, 0.8[$	$[0.8, 1]$
Nombre de variables	*19*	*20*	*16*	*23*	*22*

1. *Si on numérote les intervalles du tableau de* 1 *à* 5 *et qu'on construit les variables*

$$Y_i = \begin{cases} 1 \text{ si } X_i \in [0, 0.2[\\ 2 \text{ si } X_i \in [0.2, 0.4[\\ \dots \\ 5 \text{ si } X_i \in [0.8, 1]. \end{cases}$$

Quelle est la loi des Y_i *sous l'hypothèse* H_0 *: "les* X_i *sont de loi uniforme sur* $[0, 1]$ *?*

1. Où 1= satisfait, 2= très satisfait , ... , 5= très très très très satisfait.

2. *En déduire un test asymptotique de niveau α pour l'adéquation des X_i à une loi uniforme sur $[0,1]$ ".*
3. *Existe-t-il d'autres lois de probabilités sur les X_i qui mènent à la même loi sur les Y_i ? Si oui, construisez en une.*
4. *Qu'en déduisez-vous sur la puissance de ce test ?*

Exercice 6. *On dispose de* 10 *résultats de simulation de la loi uniforme sur* $[0,1]$ *:*

$$0,134 \quad 0,628 \quad 0,789 \quad 0,905 \quad 0,250 \quad 0,563 \quad 0,790 \quad 0,470 \quad 0,724 \quad 0,569.$$

1. *Étudiez si cet échantillon conduit à rejeter l'hypothèse nulle selon laquelle le tirage a bien eu lieu selon la loi uniforme sur* $[0,1]$*. Quel test choisissez-vous ? et pourquoi ?*

4 Exercice d'approfondissement

Exercice 7 (Test de Lilliefors)**.** *On dispose de n observations de variables aléatoires indépendantes $X_1, \ldots, X_n$ de loi inconnue, et on souhaite déterminer s'il existe m et σ^2 tels que la loi des X_i soit une loi $\mathcal{N}(m, \sigma^2)$. Le but de cet exercice est, pour un niveau α donné, de réaliser un test pour*
H_0 :La loi des X_i est une gaussienne $\mathcal{N}(m, \sigma^2)$ pour un certain couple (m, σ).
H_1 : La loi des X_i n'est pas gaussienne.

1. *Rappeler les estimateur du maximum de vraissemblance pour m et σ^2. On les notera $\bar{X}_n$ et V_n.*
2. *Si $Y_1, \ldots, Y_n$ sont des variables i.i.d. de loi $\mathcal{N}(0,1)$, quelle est la loi des variables $\sigma Y_i + b$?*
3. *Si on note $F_{(m,\sigma^2)}$ la fonction de répartition d'une variable de loi $\mathcal{N}(m, \sigma^2)$, rappeler la relation entre $F_{(m,\sigma^2)}$ et $F_{(0,1)} = \phi$.*
4. *On considère la statistique de test*

$$h(F_n, F_{(\bar{X}_n, V_n)}) = \sup_{t \in \mathbb{R}} \Big| \frac{1}{n} \sum_{i=1}^{n} 1_{X_i \leq t} - F_{(\bar{X}_n, V_n)} \Big|.$$

Montrer que $h(F_n, F_{(\bar{X}_n, V_n)})$ a la même loi que

$$\sup_{t \in \mathbb{R}} \Big| \frac{1}{n} \sum_{i=1}^{n} 1_{Y_i \leq t} - F_{(\bar{Y}_n, W_n)} \Big|$$

où $\bar{Y}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i$ et $W_n = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y}_n)^2$. On pourra relier la loi de $\bar{X}_n$ et $\bar{Y}_n$, puis celle de V_n et W_n.
5. *En déduire que la loi de $h(F_n, F_{(\bar{X}_n, V_n)})$ ne dépend ni de m, ni de σ^2 mais uniquement de n.*
6. *Construire un test non asymptotique de niveau α (et même de taille α) pour la normalité des X_i. Le test ainsi construit est le test de Lilliefors (*1968*).*

Exercice 8 (Exercice de R)**.** *Simulez* 10 *puis* 100 *variables de la loi de votre choix. Décidez d'un nombre fini de classes et choisissez une loi cible. Testez si votre échantillon suit la loi cible en appliquant le test du* χ^2*. Recommencez en changeant les classes et les lois.*

Cet exercice a pour but de vous faire simuler des variables avec R, de créer des tableaux remplissant des conditions (la commande `sum((X<0.3)*(X>0.2))` *donne le nombre d'éléments de X entre* 0.2 *et* 0.3*), et ainsi vous familiarisez avec l'utilisation de R. Le second but est de vous faire comprendre la plus grande faiblesse du test du* χ^2 *: le choix des classes. Pour une distribution de probabilité continue, il n'existe pas de choix de classes privilégié, et le résultat du test peut dépendre du découpage. J'aimerais que vous arriviez à trouver une situation où vous acceptez et rejetez la même hypothèse avec les mêmes données, mais pour deux découpages différents.*