TP de Modélisation Stochastique n^0 8 : Tests du Khi Deux

On rappelle les résultats suivants:

1. **Test d'ajustement:** On suppose que $(X_1, X_2, .., X_n)$ est un n-échantillon réel de loi inconnue f_X. On suppose que X prend ses valeurs sur k classes $(I_1, I_2, .., I_k)$ (k intervalles ou k nombres). On note $\hat{f}_i$ la fréquence empirique associée à la classe i: $\hat{f}_i = \frac{1}{n}\sum_{k=1}^{n} \mathbb{I}_{X_k \in I_i}$. On veut tester l'hypothèse

   H_0: X a pour loi f (connue)
   contre
   H_1: X n'a pas pour loi f.

   On note f_i (calculable a priori) la probabilité qu'une v.a. suivant la loi f appartienne à I_i. Alors:

   $$n \sum_{i=1}^{k} \frac{(f_i - \hat{f}_i)^2}{f_i} \underset{n \to +\infty}{\xrightarrow{\mathcal{L}}} \chi^2(k-1)$$

   (on utilise en pratique ce résultat lorsque pour tout i on a: $nf_i \geq 5$).

2. **Test d'homogénéité:** On suppose que $(X_1, X_2, .., X_n)$ est un n-échantillon réel de loi inconnue f^X et $(Y_1, Y_2, .., Y_m)$ un m-échantillon réel de loi inconnue f^Y, les deux échantillon étant indépendants l'un de l'autre. On suppose que X et Y prennent leurs valeurs sur k classes $(I_1, I_2, .., I_k)$. On note $\hat{f}_i^X$ et $\hat{f}_i^Y$ les fréquences empiriques associées à la classe i. On veut tester l'hypothèse

   H_0: X et Y ont la même loi
   contre
   H_1: X et Y n'ont pas la même loi.

   On note $\hat{f}_i = \frac{n\hat{f}_i^X + m\hat{f}_i^Y}{n+m}$, sorte de loi commune à X et Y. Alors:

   $$n \sum_{i=1}^{k} \frac{(\hat{f}_i - \hat{f}_i^X)^2}{\hat{f}_i} + p \sum_{i=1}^{k} \frac{(\hat{f}_i - \hat{f}_i^Y)^2}{\hat{f}_i} \underset{n \to +\infty}{\xrightarrow{\mathcal{L}}} \chi^2(k-1)$$

3. **Test d'indépendance:** On suppose que $((X_1, Y_1), (X_2, Y_2), .., (X_n, Y_n))$ est un n-échantillon réel d'un couple (X, Y) de loi inconnue. On suppose que X et Y prennent respectivement leurs valeurs sur k classes $(I_1, I_2, .., I_k)$ et l classes $(J_1, J_2, .., J_k)$. On note $\hat{f}_{ij}$ la fréquence empirique associée à (I_i, J_j), i.e. $\hat{f}_{ij} = \frac{1}{n}\sum_{k=1}^{n} \mathbb{I}_{(X_k \in I_i) \cap (Y_k \in J_j)}$. On veut tester l'hypothèse

   H_0: X et Y sont indépendantes
   contre
   H_1: X et Y ne sont pas indépendantes.

   On calcule f_{ij} fréquence obtenue si X et Y sont indépendantes, à savoir, $f_{ij} = \frac{1}{n}\sum_{k=1}^{n} \mathbb{I}_{(X_k \in I_i)} \cdot \frac{1}{n}\sum_{k=1}^{n} \mathbb{I}_{(Y_k \in J_j)}$.

   Alors:

   $$n \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(f_{ij} - \hat{f}_{ij})^2}{f_{ij}} \underset{n \to +\infty}{\xrightarrow{\mathcal{L}}} \chi^2(k-1)(l-1)$$

Exercices sur les tests du Khi Deux

1. Générer un échantillon de taille 100 de loi uniforme sur $[0, 1]$. Décomposer en k classes de même longueur. Faire un test du Khi Deux d'ajustement à la loi uniforme. Recommencer en augmentant la taille de l'échantillon, le nombre de classes, puis en changeant de loi (normale, binomiale, Poisson).

2. Générer un échantillon de taille 100 de loi exponentielle. Tester si l'échantillon est de loi normale, Poisson ou binomiale.

3. On peut aussi se servir du test d'ajustement pour tester s'il n'y a pas eu fraude (cas célèbre: Pascal, Mandel,...). On génère $X_k = |sin(k)|$, pour $k = 1, 2, .., n$. Pour différentes tailles de n, tester l'hypothèse H_0: $(X_1, X_2, .., X_n)$ est trop proche d'une loi uniforme pour ce soit possible.

4. Générer deux échantillons de même loi (uniforme, normale, exponentielle, binomiale, ...) et de tailles différentes (1000 et 10000 par exemple). Tester l'homogénéité de ces deux échantillons. Faire la même chose avec deux échantillons de loi distincte de même support.

5. On veut savoir si les mathématiciens sont philosophes. On a relevé sur 100 bacheliers les notes obtenues en mathématiques X et en philosophie Y.

X Y	[0,4[	[4,8[	[8,12[	[12,16[	[16,20]
[0,4[	3	4	2	0	0
[4,8[	6	10	8	2	0
[8,12[	1	8	20	12	3
[12,16[	0	0	8	7	3
[16,20]	0	0	1	0	2

Tester l'hypothèse d'indépendance entre les notes obtenues en mathématiques et en philosophie.

6. Générer un échantillon $(X_1, X_2, .., X_{1000})$ de taille 1000 de loi normale centrée réduite. Considérer $Y_i = X_i^2$. Tester l'indépendance de X et de Y. Faire la même chose avec d'autres lois (Bernoulli, exponentielle, uniforme...).

7. On s'intéresse au cas où l'asymptotique du Khi Deux ne s'applique pas. On reprend l'exercice suivant, exposé dans Tomassone et al., p.97: on veut savoir si la couleur du verre à une influence sur l'appréciation de la qualité d'un vin. D'où le tableau:

Couleur du verre (du moins foncé au plus foncé)	1	2	3	4	5
Nombre de personnes désignant le vin comme le meilleur	0	1	0	5	4

Simuler la loi de la distance entre la loi uniforme sur $\{1, 2, 3, 4, 5\}$ et la loi d'un échantillon équidistribué suivant une loi uniforme sur $\{1, 2, 3, 4, 5\}$. En déterminant les quantiles à partir de l'histogramme de cette distance, tester l'hypothèse d'influence (H_0) pour un niveau de confiance quelconque.

8. En reprenant les données de stixboxA111, faites des tests d'ajustement pour modéliser ces données par des lois connues.