Feuille de TP n°8
Tests non paramétriques du chi-deux en modélisation

1 Test d'ajustement

Soit $(X_1, \ldots, X_n)$ un n-échantillon de loi inconnue μ. On veut tester H_0: « $\mu = \nu$ », contre H_1: « $\mu \neq \nu$ », où ν est une loi de probabilité fixée et connue. On considère une partition en k classes $(I_1, , \ldots, I_k)$ du support de ν. On note $\nu_1, \ldots, \nu_k$ les poids associés, que l'on supposera toujours strictement positifs. Lorsque ν est une loi discrète finie, les classes sont des singletons typiquement. Pour $i = 1, \ldots, k$, soit n_i l'effectif associé à la classe I_i. La statistique de test d'ajustement est

$$D_n = \sum_{i=1}^{k} \frac{(d_i - n_i)^2}{d_i}$$

avec $d_i = n\nu_i$. La LGN et le TCL ont pour conséquence le théorème suivant.

Théorème 1.1 (Khi-deux d'ajustement). *Lorsque n tend vers $+\infty$, la statistique D_n converge en loi vers $\chi^2(k-1)$ sous H_0, et presque sûrement vers $+\infty$ sous H_1.*

En pratique, pour n assez grand ($\inf_i d_i > 5$), on approche sous H_0 la loi de D_n par $\chi^2(k-1)$. On en déduit alors un test qui consiste à rejeter H_0 si $D_n > \chi^2_\alpha(k-1)$, où $\chi^2_\alpha(k-1)$ est le quantile $1-\alpha$ de la loi $\chi^2(k-1)$. Ce test est asymptotiquement de niveau α et de puissance 1 (i.e. convergent). En effet, on a $\mathbb{P}(D_n > \chi^2_\alpha(k-1)) \underset{n\to+\infty}{\longrightarrow} \alpha$ sous H_0, tandis que sous H_1, $\mathbb{P}(D_n > \chi^2_\alpha(k-1)) \underset{n\to+\infty}{\longrightarrow} 1$. Lorsque la loi ν n'est pas entièrement connue, on perd un degré de liberté pour chaque paramètre estimé, cf. [2, Chap. 15].

Exercice 1.2 (Générateurs pseudo-aléatoires). Un ordinateur possède un générateur pseudo-aléatoire de nombres choisis au hasard dans l'ensemble des dix premiers entiers. Les mille premiers résultats sont répartis dans le tableau suivant.

Chiffres	0	1	2	3	4	5	6	7	8	9
Observations	120	87	115	103	91	109	92	112	94	77

Peut-on accepter l'hypothèse d'équiprobabilité pour chacun des chiffres ? Faites de même en remplaçant la table précédente par une table générée à partir de la fonction `rand` de Matlab.

2 Test d'homogénéité

Le test d'homogénéité est une version élaborée du test d'ajustement. Considérons deux échantillons indépendants $(X_1, \ldots, X_n)$ et $(Y_1, \ldots, Y_p)$ de lois respectives μ_X et μ_Y. On suppose que chacun des supports de μ_X et μ_Y sont partitionnés en k classes $(I_1, \ldots, I_k)$ et $(J_1, \ldots, J_k)$ respectivement. Pour $i = 1, \ldots, k$, soient n_i et p_i les effectifs associés aux classes I_i et J_i. On veut tester H_0: « $\mu_X = \mu_Y$ », contre H_1: « $\mu_X \neq \mu_Y$ ». Sous H_0, il est naturel d'estimer la loi commune μ par $\widehat{\mu} = \{\widehat{\mu}_1, \ldots, \widehat{\mu}_k\}$ avec $\widehat{\mu}_i = (n_i + p_i)/(n+p)$. Le test d'homogénéïté repose sur la statistique $D_{n,p} := D_n^X + D_p^Y$ avec

$$D_n^X = \sum_{i=1}^{k} \frac{(d_i^X - n_i)^2}{d_i^X} \quad \text{et} \quad D_p^Y = \sum_{i=1}^{k} \frac{(d_i^Y - p_i)^2}{d_i^Y},$$

où $d_i^X = n\widehat{\mu}_i$ et $d_i^Y = p\widehat{\mu}_i$.

Théorème 2.1 (Khi-deux d'homogénéité). *Lorsque n et p tendent vers $+\infty$, la statistique $D_{n,p}$ converge en loi vers $\chi^2(k-1)$ sous* H_0*, et converge presque sûrement vers $+\infty$ sous* H_1*.*

En pratique, pour n et p assez grands, on peut sous H_0 approcher la loi de $D_{n,p}$ par $\chi^2(k-1)$. Ici encore, le test consistant à rejeter H_0 dès que $D_{n,p} > \chi^2_\alpha(k-1)$ est asymptotiquement de niveau α et de puissance 1.

Exercice 2.2. Écrire un programme permettant de générer deux échantillons de même loi, par exemple uniforme, normale, exponentielle ou binomiale et de tailles différentes $n = 1000$ et $p = 10000$. Tester ensuite l'homogénéïté de ces deux échantillons. Faire de même avec deux échantillons de lois distinctes mais de même support.

3 Test d'indépendance

Soient $(X_1, \dots, X_n)$ et $(Y_1, \dots, Y_n)$ deux n-échantillons dont la loi du couple $\mu_{(X,Y)}$ est inconnue. On note μ_X et μ_Y les lois marginales associées. On veut tester H_0: « $\mu_{(X,Y)} = \mu_X \otimes \mu_Y$ », contre H_1: « $\mu_{(X,Y)} \neq \mu_X \otimes \mu_Y$ ». On suppose que les supports de μ_X et μ_Y sont partitionnés en k classes $(I_1, \dots, I_k)$ et l classes $(J_1, \dots, J_l)$ respectivement. Pour $i = 1, \dots, k$ et $j = 1, \dots, l$, soit $n_{i,j}$ l'effectif associé aux classes I_i et J_j. On pose $n_{i*} = \sum_{j=1}^{l} n_{i,j}$ et $n_{*j} = \sum_{i=1}^{k} n_{i,j}$ les effectifs « marginaux ». Sous l'hypothèse H_0, il est naturel d'estimer $\mu_{(X,Y)}$ par $\widehat{\mu} = \left\{\widehat{\mu}_{i,j} := \frac{n_{i*}}{n}\frac{n_{*j}}{n}, 1 \leqslant i \leqslant k, 1 \leqslant j \leqslant l\right\}$. La statistique de test d'indépendance est

$$D_n = \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{(d_{i,j} - n_{i,j})^2}{d_{i,j}}$$

où $d_{i,j} = n\widehat{\mu}_{i,j} = n_{i*}n_{*j}/n$.

Théorème 3.1 (Khi-deux d'indépendance). *Lorsque n tend vers $+\infty$, la statistique de test D_n converge en loi vers un $\chi^2((k-1)(l-1))$ sous H_0 et converge presque sûrement vers $+\infty$ sous* H_1*.*

En pratique, pour n grand, on peut sous H_0 approcher la loi de D_n par $\chi^2((k-1)(l-1))$. Ici encore, le test consistant à rejeter H_0 dès que $D_n > \chi^2_\alpha((k-1)(l-1))$ est asymptotiquement de niveau α et de puissance 1.

Exercice 3.2 (Mathématiques philosophiques). Afin de savoir si les *Mathématiciens sont Philosophes*, on a relevé, sur 100 bacheliers, les notes obtenues en Mathématiques X et en Philosophie Y.

$X \setminus Y$	[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.

Exercice 3.3 (Visions de gauche et de droite). Les scores de vision aux deux yeux de 7477 femmes, âgées de 30 à 40 ans, ont été classés en quatre groupes notés de 1 à 4 par ordre décroissant.

$D \setminus G$	1	2	3	4
1	1520	266	124	66
2	234	1512	432	78
3	117	362	1772	205
4	36	82	179	492

Tester l'hypothèse d'indépendance puis de symétrie entre les deux yeux.

Références

[1] D. Dacunha-Castelle et M. Duflo – *Probabilités et statistiques. Tome 1*, Masson, Paris, 1982, Problèmes à temps ↓xe.
[2] G. Saporta – *Probabilités, analyse des données et statistique*, Technip, 1989.
[3] P. Toulouse – *Thèmes de probabilités et statistique*, Dunod, 1999, INTERDIT À L'ORAL.

