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 f_X. On suppose que X prend ses valeurs dans k classes $(I_1, , \ldots, I_k)$. La démarche est similaire dans le cas discret en remplaçant les classes par les valeurs prises par X. Pour $i = 1, \ldots, k$, soit n_i l'effectif associé à la classe I_i. Pour une loi de probabilité $f = \{f_1, \ldots, f_k\}$ donnée, on veut tester H_0: « X a pour loi f », contre H_1: « X n'a pas pour loi f ». Le test d'ajustement est basé sur la statistique

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

avec $d_i = nf_i$.

Theorème 1.1 (Khi-deux d'ajustement). *Sous* H_0, D_n *converge en loi vers un khi-deux* $\chi^2(k-1)$.

En pratique, pour n assez grand, on peut sous H_0 approcher la loi de D_n par un $\chi^2(k-1)$. De plus, on peut montrer que sous H_1, D_n tend vers l'infini. On a donc une région d'acceptation $\mathcal{A} = [0, a]$ avec, pour un niveau de confiance $1-\alpha$ donné, $\mathbb{P}(D_n \leqslant a) = 1 - \alpha$.

Remarque 1.2. Il peut arriver que la loi f ne soit pas entièrement connue. Pour chaque paramètre estimé de f, on perd un degré de liberté.

Exercice 1.3 (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. Soit $(X_1, \ldots, X_n)$ un n-échantillon de loi inconnue f_X et soit $(Y_1, , \ldots, Y_p)$ un p-échantillon de loi inconnue f_Y. On suppose que les deux échantillons sont indépendants et que X et Y prennent respectivement leurs valeurs dans k classes $(I_1, \ldots, I_k)$ et $(J_1, \ldots, J_k)$. 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: « X et Y ont la même loi », contre H_1: « X et Y n'ont pas la même loi ». Sous l'hypothèse H_0, soit f la loi commune à X et Y. Il est naturel d'estimer f par $\widehat{f} = \left\{\widehat{f}_1, \ldots, \widehat{f}_k\right\}$ avec $\widehat{f}_i = (n_i + p_i)/(n+p)$. Le test d'homogénéité repose sur la statistique $D_n^p = D_n^X + D_p^Y$ avec

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

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

Theorème 2.1 (Khi-deux d'homogénéité). *Sous* H_0, D_n^p *converge en loi vers un khi-deux* $\chi^2(k-1)$.

En pratique, pour n et p assez grand, on peut sous H_0 approcher la loi de D_n^p par un $\chi^2(k-1)$. De plus, on peut montrer que sous H_1, D_n^p tend vers l'infini. On a donc une région d'acceptation $\mathcal{A} = [0, a]$ avec, pour un niveau de confiance $1-\alpha$ donné, $\mathbb{P}(D_n^p \leqslant a) = 1-\alpha$.

Exercice 2.2. Créer un code Matlab 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'homogénéité de ces deux échantillons. Faire de même avec deux échantillons de lois distinctes mais de même support.

3 Test d'indépendance

Soient $(X_1, \ldots, X_n)$ et $(Y_1, \ldots, Y_n)$ deux n-échantillons dont la loi du couple $f_{(X,Y)}$ est inconnue. On suppose que X et Y prennent respectivement leurs valeurs dans k classes $(I_1, , \ldots, I_k)$ et l classes $(J_1, \ldots, J_l)$. Pour $i = 1, \ldots, k$ et $j = 1, \ldots, l$, soit n_{ij} l'effectif associé aux classes I_i et J_j. On pose $n_{i*} = \sum_{j=1}^{l} n_{ij}$ et $n_{*j} = \sum_{i=1}^{k} n_{ij}$. On veut tester H_0: « X et Y sont indépendantes », contre H_1: « X et Y ne sont pas indépendantes ». Sous l'hypothèse H_0, il est naturel d'estimer $f_{(X,Y)}$ par $\widehat{f} = \left\{\widehat{f}_{ij} = f_{i*}f_{*j} \text{ avec } 1 \leqslant i \leqslant k, 1 \leqslant j \leqslant l\right\}$ où $\widehat{f}_{i*} = n_{i*}/n$, $\widehat{f}_{*j} = n_{*j}/n$. Le test d'indépendance est basé sur la statistique

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

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

Theorème 3.1 (Khi-deux d'indépendance). *Sous* H_0, D_n *converge en loi vers un khi-deux* $\chi^2(k-1)(l-1)$.

En pratique, pour n assez grand, on peut sous H_0 approcher la loi de D_n par un $\chi^2(k-1)(l-1)$. De plus, on peut montrer que sous H_1, D_n tend vers l'infini. On a donc une région d'acceptation $\mathcal{A} = [0, a]$ avec, pour un niveau de confiance $1-\alpha$ donné, $\mathbb{P}(D_n \leqslant a) = 1-\alpha$.

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.

