Tests du Khi-deux

Retour au plan du cours

1 Test d'ajustement du χ^2

Soit $X_1, \dots X_n$ i.i.d. à valeurs dans un ensemble fini $\{a_1, \dots, a_r\}$ avec $r > 1$. On suppose que

$$\forall j = 1, \dots, r,\ p_j = \mathbb{P}(X_i = a_j) > 0.$$

Le problème que l'on se pose est le suivant : étant donné $\pi = (\pi_1, \dots, \pi_r)'$ vérifiant $\forall j,\ \pi_j > 0$ et $\sum_{j=1}^r \pi_j = 1$, comment tester $p = \pi$ à partir du n-echantillon, $X_1, \dots, X_n$?

Pour tout $j = 1, \dots, r$ on pose

$$N_{jn} = \sum_{i=1}^n \mathbb{1}\{X_i = a_j\},$$

le nombre de X_i prenant la valeur a_j.

PROPOSITION 1. — *La v.a. $N_n = (N_{1n}, \dots, N_{rn})'$ suit une loi multinomiale $\mathcal{M}(n, p_1, \dots, p_r)$ sur $\mathbb{N}^r$, c'est à dire que pour tout $(n_1, \dots, n_r) \in \mathbb{N}^r$ on a*

$$\mathbb{P}[N_{1n} = n_1, \dots, N_{rn} = n_r] = \begin{cases} \frac{n!}{n_1! \dots n_r!} p_1^{n_1} \dots p_r^{n_r} & \text{si } \sum_{j=1}^r n_j = n \\ 0 & \text{sinon.} \end{cases}$$

PROPOSITION 2. — *Soit $\sqrt{p} = \left(\sqrt{p_1}, \dots, \sqrt{p_r}\right)'$. On a*

$$Y_n = \left(\frac{N_{1n} - np_1}{\sqrt{np_1}}, \dots \frac{N_{rn} - np_r}{\sqrt{np_r}} \right) \xrightarrow{\text{Loi}} \mathcal{N}_r(0, \Gamma),$$

où $\Gamma = I_r - \sqrt{p}\sqrt{p}'$.

THÉORÈME 3. — *Sous l'hypothèse que $X_1, \dots, X_n$ sont i.i.d. de loi $p = (p_1, \dots, p_r)$,*

$$Z_n = \sum_{j=1}^r \frac{(N_{jn} - np_j)^2}{np_j} \xrightarrow{\text{Loi}} \chi^2_{(r-1)}.$$

Le test d'ajustement du χ^2 consiste à rejeter l'hypothèse $p = \pi$ au niveau α si

$$T_n = \sum_{j=1}^r \frac{(N_{jn} - n\pi_j)^2}{n\pi_j} > x_{r-1,1-\alpha}$$

où $x_{r-1,1-\alpha}$ est le $1-\alpha$ quantile d'un χ^2 à $r-1$ degrés de liberté. D'après le résultat précedent on obtient un test de niveau *asymptotique* α. En montre que l'approximation de la loi de T_r (sous l'hypothèse) par un χ^2 à $r-1$ degrés de liberté est bon dès lors que $n\pi_j \geq 5$ pour tout j.

Que peut-on dire de la puissance du test ? En notant par $\lceil(.,.)$ la distance euclidienne sur $\mathbb{R}^r$, on a que

$$\frac{T_n}{n} \geq \lceil^2(N_n/n, \pi) \xrightarrow{\text{p.s.}} \lceil^2(p, \pi)$$

par la loi des grands nombres et donc $T_n \xrightarrow{\text{p.s.}} +\infty$. La puissance du test tend donc vers 1 quand n tend vers $+\infty$.

Exemple. — **Expérience de Mendel**

On croise 2 populations (pures) de pois : l'une jaune et ronde l'autre verte et ridée. Selon sa prédiction au bout de 2 croisements la proportion de pois

JR jaunes et ronds est $9/16$

Jr jaunes et ridés est $3/16$

vR verts et ronds est $3/16$

vr verts et ridés est $1/16$

Pour chacun des phénotypes il obtient les résultats suivants : $N_{JR} = 315$, $N_{Jr} = 101$, $N_{vR} = 108$, $N_{vr} = 32$. Ici $r = 4$ et l'on obtient que $T_n = 0.47$ et $x_{3,0.95} = 7.82$. On accepte donc très largement l'hypothèse de Mendel.

2 Test du χ^2 d'adéquation à une famille de lois

Soit Θ un ouvert de $\mathbb{R}^k$ avec $1 \leq k < r$ et

$$\mathcal{P}(\Theta) = \{\pi(\theta) mboxtelque\ \theta \in \Theta\},$$

une famille de lois discrètes sur $\{a_1, \ldots, a_r\}$. On aimerait tester l'hypothèse $p \in \mathcal{P}(\Theta)$ contre $p \notin \mathcal{P}(\Theta)$.

Or, on a le résultat (admis) suivant :

THÉORÈME 4. — *Supposons que*
- *Pour tout $j = 1, \ldots, r$, $\theta \mapsto \pi_j(\theta)$ est $\mathcal{C}^2$ sur Θ et vérifie pour tout $\theta \in \Theta$, $\pi_j(\theta) \neq 0$.*
- *Pour tout $\theta \in \Theta$, les vecteurs $v_i = (\partial_i \pi_j(\theta), \ldots, \partial_i \pi_r(\theta))'$ pour $i = 1, \ldots, k$ forment une famille libre de $\mathbb{R}^r$ (bonne paramétrisation).*
- *Pour tout θ, si $X_1, \ldots, X_n$ sont i.i.d. de loi $\pi(\theta)$ alors l'estimateur du maximum de vraisemblance $\hat{\theta}_n$ est consistant vers θ.*

Sous ces conditions, si $X_1, \ldots, X_n$ sont i.i.d. de loi $\pi(\theta)$ alors

$$\sum_{j=1}^{r} \frac{\left(N_{jn} - n\pi_j(\hat{\theta}_n)\right)^2}{n\pi_j(\hat{\theta}_n)} \xrightarrow{\text{Loi}} \chi^2(r-k-1).$$

On construit le test du χ^2 d'adéquation à $\mathcal{P}(\Theta)$ de la manière suivante : on rejette l'hypothèse si

$$T_n = \sum_{j=1}^{r} \frac{\left(N_{jn} - n\pi_j(\hat{\theta}_n)\right)^2}{n\pi_j(\hat{\theta}_n)} > x_{r-k-1,1-\alpha}.$$

Sous l'alternative :

$$\frac{T_n}{n} \geq \lceil^2 (N_n/n, \mathcal{P}(\Theta)) \xrightarrow{\text{p.s.}} \lceil^2 (p, \mathcal{P}(\Theta)),$$

et donc la puissance tend vers 1 dès que $\lceil^2 (p, \mathcal{P}(\Theta)) > 0$.

2.1 Test du χ^2 d'indépendance

On suppose que le support de μ est un ensemble fini doublement indicé $\{a_{i,j},\ i = 1, \ldots I, j = 1, \ldots J\}$ pour lequel les indices i, j représentent 2 facteurs (par exemple : couleur des cheveux/ couleur des yeux). On veut tester l'indépendance des 2 facteurs. On prend

$$\Theta = \left\{\theta = (u_1, \ldots, u_{I-1}, v_1, \ldots, v_{J-1}) \text{ tel que } u_i > 0,\ v_j > 0,\ \sum_{i=1}^{I-1} u_i < 1, \sum_{j=1}^{J-1} v_j < 1\right\},$$

$$\mathcal{P}(\Theta) = \{\pi_{i,j}(\theta) = u_i v_j \text{ tel que } u_I = 1 - \sum_{i=1}^{I-1} u_i,\ v_J = 1 - \sum_{j=1}^{J-1} v_j,\ \theta \in \Theta\}.$$

L'estimateur du maximum de vraisemblance de θ est donné par

$$\hat{u}_i = N_{i.} = \frac{1}{n}\sum_{j=1}^{J} N_{i,j} \quad \text{et} \quad \hat{v}_j = N_{.j} = \frac{1}{n}\sum_{i=1}^{I} N_{i,j}.$$

Les hypothèses du théorème sont vérifées et l'on rejette donc l'hypothèse d'indépendance si

$$T_n = n\sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(N_{i,j} - N_{i,.}N_{.,j}/n)^2}{N_{i,.}N_{.,j}} > x_{(I-1)(J-1),1-\alpha}.$$

Montrons que l'estimateur du maximum de vraisemblance est bien celui-là : on a à maximiser en u_i et v_j :

$$\begin{aligned}
\sum_{i,j} N_{i,j}\log(u_i v_j) &= \sum_{i=1}^{I} N_{i,.}\log(u_i) + \sum_{j=1}^{J} N_{.,j}\log(v_j) \\
&= \sum_{i=1}^{I-1} N_{i,.}\log(u_i) + N_{I,.}\log(1 - \sum_{i=1}^{I-1} u_i) \\
&\quad + \sum_{j=1}^{J-1} N_{.,j}\log(v_j) + N_{.,J}\log(1 - \sum_{j=1}^{J-1} v_j)
\end{aligned}$$

En dérivant en u_i on obtient que

$$\frac{N_{i,.}}{u_i} = \frac{N_{I,.}}{u_I}$$

et donc que les membres de droites sont constants $= c$ indépendants de i. Comme $\sum_i N_{i,.} = n$ et $\sum_i u_i = 1$ on obtient que $c = n$ et donc $u_i = N_{i,.}/n$. Il en est de même pour les v_j. On obtient ainsi un maximum sur Θ quand les $N_{i,.}$ et $N_{.,j}$ sont non nuls. ■

2.2 Test d'homogénéité

Soit $X_1, \ldots, X_n$ et $X'_1, \ldots, X'_n$ 2 échantillons indépendants de 2 lois discrètes, μ (correspondants à p), μ' (correspondants à p'). sur $\{a_1, \ldots, a_r\}$. On veut tester $\mathrm{H}_0 : \ p = p'$ (p étant inconnu). On utilise la statistique de test

$$T_n = \sum_{i=1}^{r} \frac{(N_i - N'_i)^2}{(N_i + N'_i)}.$$

Sous l'hypothèse H_0, $T_n \xrightarrow{\text{Loi}} \chi^2(r-1)$. Pour tester au niveau α, on rejette donc quand $T_n > x^2_{r-1,1-\alpha}$.

Montrons la convergence en loi. Par le lemme de Slutsky il suffit de montrer que

$$\tilde{T}_n = \frac{n}{2} \sum_{i=1}^{r} \frac{(N_i/n - N'_i/n)^2}{p_i} \xrightarrow{\text{Loi}} \chi^2(r-1),$$

car pour tout i, sous H_0, $(N_i + N'_i)/(2n) \xrightarrow{\text{p.s.}} p_i$.

$$\begin{aligned} \tilde{T}_n &= \frac{1}{2} \| \left(\frac{N_{1n} - np_1}{\sqrt{np_1}}, \ldots \frac{N_{rn} - np_r}{\sqrt{np_r}} \right) - \left(\frac{N'_{1n} - np_1}{\sqrt{np_1}}, \ldots \frac{N'_{rn} - np_r}{\sqrt{np_r}} \right) \|^2 \\ &= \frac{1}{2} \| Y_n - Y'_n \|^2 \end{aligned}$$

où Y_n et Y'_n sont i.i.d. et convergent en loi vers la loi $\mathcal{N}_r(0, \Gamma)$. On en déduit que

$$Y_n - Y'_n \xrightarrow{\text{Loi}} \mathcal{N}_r(0, 2\Gamma).$$

$$\frac{1}{\sqrt{2}}(Y_n - Y'_n) \xrightarrow{\text{Loi}} \mathcal{N}_r(0, \Gamma).$$

$$\frac{1}{2} \| Y_n - Y'_n \|^2 \xrightarrow{\text{Loi}} \chi^2_{(r-1)}$$

par le théorème de l'application continue.

Remarque. — Ce test d'homogénéité correspond à un test d'indépendance entre l'appartenance à l'un ou l'autre des 2 groupes et le caractère étudié.