CHAPITRE 7 : TESTS DU χ^2

I Introduction

1) Motivations

Considérons l'exemple suivant : on croise des fleurs rouges avec des fleurs blanches. A la génération suivante elles sont toutes roses. On les croise de nouveau : on obtient 600 nouvelles fleurs, dont 141 sont rouges, 144 blanches et 315 roses.

Ce résultat s'explique par le fait que le gène codant la couleur possède deux allèles : rouge (R) et blanc (B), de telle sorte que

- une fleur RR est rouge,
- une fleur BB est blanche,
- une fleur RB ou BR est rose.

Si on fait l'hypothèse qu'un croisement consiste à prendre un gène de chacun des deux parents au hasard, on devrait obtenir

- une fleur rouge (RR) avec probabilité 1/4,
- une fleur blanche (BB) avec probabilité 1/4,
- une fleur rose (RB ou BR) avec probabilité 1/2.

Nous aimerions tester cette hypothèse à partir de cet échantillon.

2) La loi multinomiale

Définition 1. *Soient $n \geq 2$, $k \in \{2, \dots, n\}$ et $p_1, \dots, p_k$ des réels de $]0,1[$ tels que $p_1 + \cdots + p_k = 1$. Un vecteur aléatoire $(N_1, \dots, N_k)$ suit la loi multinomiale de paramètres n et $(p_1, \dots, p_k)$ si*

$$\forall (n_1, \dots, n_k) \in \mathbb{N}^k \qquad \mathbb{P}(N_1 = n_1, \dots, N_k = n_k) = \begin{cases} \dfrac{n!}{n_1! \dots n_k!} p_1^{n_1} \dots p_k^{n_k} & si \quad n_1 + \cdots + n_k = n \\ 0 & si \quad n_1 + \cdots + n_k \neq n. \end{cases}$$

Remarques : $\star$ C'est la généralisation de la loi binomiale. En effet, si $k = 2$ et $p_2 = 1 - p_1$, alors $N_1 \sim \mathcal{B}(n, p_1)$, $N_2 \sim \mathcal{B}(n, p_2)$ et $N_2 = n - N_1$.

$\star$ Si $n_1 + \cdots + n_k = n$ alors

$$\frac{n!}{n_1! \dots n_k!} =$$

Exemple : Soit $(X_1, \dots, X_n)$ un n-échantillon à valeurs dans $\{1, \dots, k\}$. Pour tout $j \in \{1, \dots, k\}$, nous posons

$$p_j = \mathbb{P}(X_1 = j) \qquad \text{et} \qquad N_j^n = \sum_{i=1}^{n} \mathbb{1}_{X_i = j} \quad (= \text{nombre de fois que } j \text{ est observé parmi les } X_i).$$

Alors $(N_1^n, \dots, N_k^n)$ suit la loi multinomiale de paramètres n et $(p_1, \dots, p_k)$. En effet :

Nous déduisons de cet exemple la proposition suivante :

Proposition 2. *Si $(N_1, \dots, N_k)$ suit la loi multinomiale de paramètres n et $(p_1, \dots, p_k)$ alors, pour tout $(i,j) \in \{1, \dots, k\}^2$ avec $i \neq j$,*

$$\mathbb{E}(N_i) = np_i, \qquad \mathbb{V}\mathrm{ar}(N_i) = np_i(1-p_i) \qquad et \qquad Cov(N_i, N_j) = -np_ip_j.$$

3) Un résultat de convergence

Proposition 3. *Soit $(X_1, \dots, X_n)$ un n-échantillon à valeurs dans $\{1, \dots, k\}$. Notons $p_j = \mathbb{P}(X_1 = j)$ et $N_j^n = \sum_{i=1}^n \mathbb{1}_{X_i = j}$ pour tout $j \in \{1, \dots, k\}$. Alors*

$$Z_n = \left(\frac{N_1^n - np_1}{\sqrt{np_1}}, \dots, \frac{N_k^n - np_k}{\sqrt{np_k}} \right) \xrightarrow[n \to +\infty]{\mathcal{L}} \mathcal{N}_k(0, I_k - \Gamma_p),$$

où Γ_p est la matrice carrée de taille k telle que $(\Gamma_p)_{i,j} = \sqrt{p_ip_j}$ pour tout $1 \leq i, j \leq k$.

Démonstration : Soit, pour $i \in \{1, \dots, n\}$,

$$Y_i = \left(\frac{\mathbf{1}_{X_i = 1} - p_1}{\sqrt{p_1}}, \dots, \frac{\mathbf{1}_{X_i = k} - p_k}{\sqrt{p_k}} \right)$$

Ce sont n vecteurs aléatoires i.i.d. et centrés. Le TCL vectoriel entraine que $Z_n = \sqrt{n}(\frac{Y_1 + \dots + Y_n}{n})$ converge en loi lorsque $n \to \infty$ vers une loi $\mathcal{N}_k(0, \Sigma_p)$ avec :

$$(\Sigma_p)_{ij} = \mathbb{E}[Y_1(i)Y_1(j)] = \frac{1}{\sqrt{p_ip_j}} \mathbb{E}(\mathbf{1}_{X_1 = i, X_1 = j} - p_i \mathbf{1}_{X_1 = j} - p_j \mathbf{1}_{X_1 = i} + p_ip_j)$$

$$= \frac{1}{\sqrt{p_ip_j}} (\mathbb{P}(X_1 = i, X_1 = j) - p_ip_j) = \delta_{ij} - \sqrt{p_ip_j}$$

II Test du χ^2 d'ajustement

Nous observons une variable aléatoire de loi multinomiale de paramètres n et $p = (p_1, \dots, p_k)$. Nous voulons tester H_0 : "$p = p^o$" contre H_1 : "$p \neq p^o$" pour un certain $p^o = (p_1^o, \dots, p_k^o)$.

Exemple : Dans l'exemple des fleurs, nous observons $n = 600$ et $p = (p_1, p_2, p_3)$ avec $\hat{p}_1 = 141/600$, $\hat{p}_2 = 144/600$ et $\hat{p}_3 = 315/600$. Nous voulons comparer p à $p^o = (1/4, 1/4, 1/2)$.

Théorème 4. *Soit $(X_1, \dots, X_n)$ un n-échantillon à valeurs dans $\{1, \dots, k\}$. Notons $p_j = \mathbb{P}(X_1 = j)$ et $N_j^n = \sum_{i=1}^n \mathbb{1}_{X_i = j}$ pour tout $j \in \{1, \dots, k\}$. Nous introduisons*

$$D_n^2 = \sum_{j=1}^k \frac{(N_j^n - np_j^o)^2}{np_j^o}$$

Alors : ⋆ *si $p = p^o$ alors $D_n^2 \xrightarrow[n \to +\infty]{\mathcal{L}} \chi^2(k-1)$,*

⋆ *si $p \neq p^o$ alors $D_n^2 \xrightarrow[n \to +\infty]{} +\infty$ presque sûrement.*

DÉMONSTRATION : **Etape 1** Supposons que $p = p^o$. Alors $D_n^2 = \|Z_n\|^2 \xrightarrow[n\to+\infty]{\mathcal{L}} \|Y\|^2$ avec $Y \sim \mathcal{N}_k(0, I_k - \Gamma_p)$. Notons $\sqrt{p} = (\sqrt{p_1}, \dots, \sqrt{p_k})$ et commencons par montrer que $I_k - \Gamma_p = \Pi_{Vect(\sqrt{p})^\perp}$, la projection orthogonale sur $Vect(\sqrt{p})^\perp$. Pour $i \in \{1, \dots, k\}$, posons $e_i = (0, \dots, 0, 1, 0, \dots, 0)$ le ième vecteur de la base canonique. Il existe $\lambda_i \in \mathbb{R}$ tel que $\Pi_{Vect(\sqrt{p})}(e_i) = \lambda_i \sqrt{p}$.
Nous avons $\langle (I_k - \Pi_{Vect(\sqrt{p})})e_i, \sqrt{p}\rangle = 0$, donc $\langle e_i, \sqrt{p}\rangle = \langle \lambda_i\sqrt{p}, \sqrt{p}\rangle$ c'est-à-dire
$\sqrt{p_i} = \lambda_i(p_1 + \dots + p_k) = \lambda_i$. Ainsi on a bien :

$$(\Pi_{Vect(\sqrt{p})^\perp}(e_i))_j = (e_i - \Pi_{Vect(\sqrt{p})}(e_i))_j = (e_i - \sqrt{p_i}\sqrt{p_j})_j = \delta_{ij} - \sqrt{p_i p_j}$$

Maintenant soit $W \sim \mathcal{N}_k(0, I_k)$, appliquons le théorème de Cochran à W et $E_1 = Vect(\sqrt{p})$, $E_2 = E_1^\perp$. On obtient $\Pi_{E_2}W \sim \mathcal{N}_k(0, I_k - \Gamma_p)$, $\|\Pi_{E_2}W\|^2 \sim \chi^2(\dim E_2)$ avec $\dim E_2 = k - 1$. Or Y et $\Pi_{E_2}W$ ont la même loi donc $\|Y\|^2 \sim \chi^2(k-1)$, ainsi $D_n^2 \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2(k-1)$.

Etape 2 Si $p \neq p^o$, alors il existe j tel que $p_j \neq p_j^o$. De plus, $D_n^2 \geq \frac{(N_j^n - np_j^o)^2}{np_j^o} = n\frac{(\frac{N_j^n}{n} - p_j^o)^2}{p_j^o}$. Or la loi forte des grands nombres nous assure que $\frac{N_j^n}{n} \xrightarrow[n\to+\infty]{p.s.} p_j$ donc $\frac{(\frac{N_j^n}{n} - p_j^o)^2}{p_j^o} \xrightarrow[n\to+\infty]{p.s.} \frac{(p_j - p_j^o)^2}{p_j^o} \neq 0$ et donc $D_n^2 \xrightarrow[n\to+\infty]{p.s.} +\infty$.

Corollaire 5. *Soit $\alpha \in]0,1[$. Le test du χ^2 d'ajustement de H_0 contre H_1 au niveau α est $\mathbb{1}_{D_n^2 > c_{1-\alpha}^{(k-1)}}$, où $c_{1-\alpha}^{(k-1)}$ est le quantile d'ordre $1-\alpha$ de $\chi^2(k-1)$.*

Notons que ce test est valable pour $n \geq 30$ et $np_j^o \geq 5$ pour tout $j \in \{1, \dots, k\}$.

Exemple : Dans l'exemple des fleurs,

$$D_n^2 = \frac{(141 - 600 \times 1/4)^2}{600 \times 1/4} + \frac{(144 - 600 \times 1/4)^2}{600 \times 1/4} + \frac{(315 - 600 \times 1/2)^2}{600 \times 1/2} \approx 1,53.$$

Si $\alpha = 5\%$ alors $c_{1-\alpha}^{(2)} \approx 5,99$. Nous avons $D_n^2 < c_{1-\alpha}^{(2)}$ donc nous acceptons l'hypothèse.

Remarque : La puissance asymptotique du test vaut 1 (ce qui est super ! !). En effet par la loi faible des grands nombres, nous avons

$$\forall p \neq p^o \qquad \frac{D_n^2}{n} \xrightarrow[n\to+\infty]{\mathbb{P}_p} \sum_{j=1}^k \frac{(p_j - p_j^o)^2}{p_j} > 0$$

donc, pour tous $p \neq p^o$ et $t \in \mathbb{R}$, $\mathbb{P}_p(D_n^2 \geq t) = \mathbb{P}_p(D_n^2/n \geq t/n) \xrightarrow[n\to+\infty]{} 1$.

III Variantes

1) Test du χ^2 d'ajustement à une famille paramétrée de loi

Théorème 6. *Soient $k > d+1$, $\Theta \subset \mathbb{R}^d$ et $p : \theta \in \Theta \longmapsto (p_1(\theta), \dots, p_k(\theta)) \in \mathbb{R}^k$ une application injective et de classe $\mathcal{C}^2$ sur Θ. Supposons que les coordonnées de p ne s'annulent jamais sur Θ et que, pour tout $\theta \in \Theta$, les vecteurs $\frac{\partial p}{\partial \theta_1}(\theta), \dots, \frac{\partial p}{\partial \theta_d}(\theta)$ sont libres.*

Soit $(X_1, \dots, X_n)$ un n-échantillon de loi $p(\theta)$ avec $\theta \in \Theta$. Notons $\widehat{\theta}_n$ l'EMV de θ et $N_j^n = \sum_{i=1}^n \mathbb{1}_{X_i = j}$ pour tout $j \in \{1, \dots, k\}$. Alors

$$D_n^2(\widehat{\theta}_n) = \sum_{j=1}^k \frac{\left(N_j^n - np_j(\widehat{\theta}_n)\right)^2}{np_j(\widehat{\theta}_n)} \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2(k-1-d)$$

Exemple : On souhaite tester si le nombre d'appels par jours à un standard suit une $\mathcal{P}(\theta)$. On estime d'abord θ par l'EMV $\widehat{\theta}_n$ puis on calcule $D_n^2(\widehat{\theta}_n)$ que l'on compare à $c_{1-\alpha}^{(k-1-d)}$ pour k et d bien choisis (cf exercices).

2) Test du χ^2 d'indépendance

Théorème 7. *Nous observons deux n-échantillons $(X_1,\dots,X_n)$ et $(Y_1,\dots,Y_n)$ respectivement à valeurs dans $\{1,\dots,k\}$ et $\{1,\dots,l\}$. Pour tous $i\in\{1,\dots,k\}$ et $j\in\{1,\dots,l\}$, posons*

$$N_{i,j}=\sum_{r=1}^{n}\mathbb{1}_{X_r=i}\mathbb{1}_{Y_r=j},\qquad N_{i\bullet}=\sum_{r=1}^{n}\mathbb{1}_{X_r=i}\qquad et\qquad N_{\bullet j}=\sum_{r=1}^{n}\mathbb{1}_{Y_r=j}.$$

Si $(X_1,\dots,X_n)$ est indépendant de $(Y_1,\dots,Y_n)$ alors

$$D_n^2=\sum_{i=1}^{k}\sum_{j=1}^{l}\frac{(N_{i,j}-N_{i\bullet}N_{\bullet j}/n)^2}{N_{i\bullet}N_{\bullet j}/n}\xrightarrow[n\to+\infty]{\mathcal{L}}\chi^2((k-1)(l-1)).$$

Une conséquence de ce théorème est qu'un test de H_0 : "les X_i sont indépendants des Y_j" contre H_1 : "les X_i ne sont pas indépendants des Y_j" est $\mathbb{1}_{D_n^2>c_{1-\alpha}^{(s)}}$, où $c_{1-\alpha}^{(s)}$ est le quantile d'ordre $1-\alpha$ de $\chi^2(s)$ avec $s=(k-1)(l-1)$.

3) Test du χ^2 d'homogénéité

Il s'agit d'un cas particulier du test du χ^2 d'indépendance. Nous observons un n-échantillon $(X_1,\dots,X_n)$ et un m-échantillon $(Y_1,\dots,Y_m)$. Nous supposons que ces deux échantillons sont indépendants et à valeurs dans $\{1,\dots,k\}$. Posons

$$\forall i\in\{1,\dots,k\}\qquad N_i=\sum_{j=1}^{n}\mathbb{1}_{X_j=i},\qquad M_i=\sum_{j=1}^{m}\mathbb{1}_{Y_j=i}\qquad \text{et}\qquad \widehat{p}_i=\frac{N_i+M_i}{n+m}$$

Si les deux échantillons ont la même loi alors

$$D_{n,m}^2=\sum_{i=1}^{k}\left(\frac{(N_i-n\widehat{p}_i)^2}{n\widehat{p}_i}+\frac{(M_i-m\widehat{p}_i)^2}{m\widehat{p}_i}\right)\simeq\chi^2(k-1).$$

Un test de H_0 : "$\mathbb{P}_{X_1}=\mathbb{P}_{Y_1}$" contre H_1 : "$\mathbb{P}_{X_1}\neq\mathbb{P}_{Y_1}$" est donné par $\mathbb{1}_{D_{n,m}^2>c_{1-\alpha}^{(k-1)}}$, où $c_{1-\alpha}^{(k-1)}$ est le quantile d'ordre $1-\alpha$ de $\chi^2(k-1)$.