Expérience de Mendel, génétique et tests du χ^2

Gendre Xavier - Verzelen Nicolas

exposé proposé par Yannick Baraud

23 Juin 2003

Présentation

Lors d'expériences scientifiques, les chercheurs sont souvent amenés à proposer des modèles pour décrire les phénomènes étudiés. Pour discuter de la validité de ces modèles les statistiques fournissent des outils permettant de décider en un certain sens si l'un d'entre eux est plus valable que d'autre. On s'intéresse ici à un test statistique, dit du χ^2, qui permet de justifier des lois discrètes. Nous le présenterons aux travers de sa construction et d'exemples liés à la génétique.

Table des matières

Chapitre 1

Introduction

1.1 Rappels et notations

1.1.1 Variables aléatoires gaussiennes

Si X est une variable aléatoire de carré intégrable, nous noterons $\mathbb{E}(X)$ son espérance et $V(X)$ sa variance. Si sa loi est $\mathcal{L}$, nous écrirons $X \sim \mathcal{L}$. Rappelons qu'une variable aléatoire gaussienne réelle est une variable aléatoire $X \sim \mathcal{N}(m, \sigma^2)$, avec $m \in \mathbb{R}$ et $\sigma \geqslant 0$, dont la fonction caractéristique vaut

$$\mathbb{E}[e^{i\xi X}] = \exp\left(i\xi m - \frac{\xi^2\sigma^2}{2}\right) \quad , \forall \xi \in \mathbb{R}$$

De simples calculs montrent que $\mathbb{E}(X) = m$ et $V(X) = \sigma^2$. De façon plus générale, on a la définition suivante :

Définition: Soit X une variable aléatoire à valeurs dans $E = \mathbb{R}^n$. On dit que X suit une loi gaussienne (ou que X est gaussienne) si pour tout $\alpha \in E$, la variable aléatoire réelle

$$\langle \alpha, X \rangle \triangleq \sum_{i=1}^{n} \alpha_i X_i$$

est gaussienne.

Si e est une base de E et si e^* est sa base duale, cette définition entraîne que $\langle e_i^*, X \rangle = X_i$ est gaussienne réelle pour tout i et donc de carré intégrable. Via la norme euclidienne $\|X\|^2 = X_1^2 + \cdots + X_n^2$, on voit que X est L^2. On notera m sa moyenne et A sa matrice de covariance (que l'on notera $Cov(X)$) qui est par définition la matrice symétrique $(Cov(X_i, X_j))_{1 \leqslant i,j \leqslant n}$ avec

$$Cov(Y_1, Y_2) = \mathbb{E}(Y_1 Y_2) - \mathbb{E}(Y_1)\mathbb{E}(Y_2)$$

Dans la suite, on notera $\mathcal{N}(m, A)$ cette loi. Rappelons la proposition suivante sur les gaussiennes :

Proposition 1.1 *Soit X une gaussienne sur E de moyenne m et de covariance A. Si F est la matrice dans la base canonique d'un endomorphisme de E, alors FX est gaussienne de moyenne Fm et de covariance FAF^* où F^* est la matrice duale de F.*

1.1.2 Rappels probabilistes

Nous rappellerons ici les deux théorèmes très importants en probabilités et en statistiques que sont la loi des grands nombres et le théorème central limite.

Théorème 1.2 (Loi forte des grands nombres) *Soit $(X_i)_{i\geqslant 0}$ une suite de variables aléatoires indépendantes et identiquement distribuées.*
Si $\mathbb{E}(|X|) < \infty$ alors

$$\tfrac{1}{n}(X_1 + \cdots + X_n) \xrightarrow[n\to\infty]{p.s} \mathbb{E}(X_1)$$

Une conclusion plus faible de la loi des grands nombres est évidemment la convergence en probabilité.

Théorème 1.3 (Théorème central limite) *Soit $(X_i)_{i\geqslant 0}$ une suite de variables aléatoires indépendantes et identiquement distribuées à valeurs dans un espace vectoriel E de dimension finie. On suppose que $a = Cov(X_1)$ existe et on note $m = \mathbb{E}(X_1)$. Alors*

$$\tfrac{1}{\sqrt{n}}(X_1 + \cdots + X_n - nm) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, a)$$

Un autre théorème utile pour prouver une convergence en loi est le suivant :

Théorème 1.4 (Théorème de Lévy) *Soit $(X_k)_{k\geqslant 0}$ une suite de variables aléatoires à valeurs dans E, alors*

$$\forall \xi \in \mathbb{R}^n,\ \lim_{k\to\infty} \mathbb{E}(e^{i\xi.X_k}) = \mathbb{E}(e^{i\xi.X}) \iff X_k \text{ converge en loi vers } X$$

1.1.3 Définitions des outils statistiques

Dans cette partie nous définirons les notions de statistique que nous allons utiliser pour construire le test du χ^2.

Définition: On appelle modèle statistique la donnée d'un triplet $(\Omega, \mathcal{A}, (P_\theta)_{\theta\in\Theta})$ où Ω est un ensemble, $\mathcal{A}$ est une tribu et $(P_\theta)_{\theta\in\Theta}$ est une famille de probabilités sur $(\Omega, \mathcal{A})$.

Pour notre exposé, on se limitera à $\Theta \subset \mathbb{R}^k$, on parle alors de statistique paramétrique. Si $P \in \{P_\theta ; \theta \in \Theta\}$, on appelle paramètre de P la valeur θ telle que $P = P_\theta$. En statistique, le but est d'obtenir des informations sur les paramètres qui définissent un phénomène aléatoire à partir d'observations de ce phénomène. Si ce phénomène prend ses valeurs dans l'espace mesurable $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ et que l'on a une famille de probabilités $(\mu_\theta)_{\theta\in\Theta}$ sur $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, il existe une façon canonique de construire un modèle statistique pour cette expérience. En effet, si on on observe n fois le phénomène aléatoire, considérons $\Omega = \mathcal{X}^n$ et les applications $X_i : \Omega \to \mathcal{X}$ de projection qui à un élément $\omega = (x_1, \ldots, x_n)$ de Ω associe $X_i(\omega) = x_i$. Chaque X_i correspond à une observation du phénomène, on les suppose indépendantes. Si on note $\mathcal{A} = \sigma(X_i ; 1 \leqslant i \leqslant n) = \mathcal{B}(\mathcal{X})^{\otimes n}$ et $P_\theta = \mu_\theta^{\otimes n}$, alors on a un modèle statistique que l'on appelle n-échantillon.

Dans la suite les quantités indicées par θ seront, par définition, définies sous la loi P_θ (par exemple $\mathbb{E}_\theta(X)$ est l'espérance de X calculée avec P_θ). Pour chercher des informations sur les paramètres d'une loi il semble naturel de travailler

avec des quantités qui approche "bien" le paramètre. C'est ce que formalise la définition ci-dessous.

Définition: Soit $g : \Theta \to \mathbb{R}^d$. On appelle estimateur de $g(\theta)$ au vu de l'observation $X = (X_1, \ldots, X_n)$ toute variable aléatoire $\sigma(X)$-mesurable $T : \Omega \to \mathbb{R}^d$. Si $\mathbb{E}_\theta(|T|) < \infty$, on appelle biais de l'estimateur la fonction

$$\begin{aligned} b_T : \Theta &\longrightarrow \mathbb{R}^d \\ \theta &\longmapsto \mathbb{E}_\theta(T) - g(\theta) \end{aligned}$$

On dit que T est sans biais si $b_T = 0$.

Il n'y a pas a priori de règle pour construire un estimateur, cependant on distingue deux méthodes particulières :

- **Méthode des moments :** si P_θ est une loi p-intégrable et que $g(\theta)$ est une fonction continue des p premiers moments de cette loi, alors

  $$g(\theta) = \psi(\mathbb{E}_\theta(X_1), \ldots, \mathbb{E}_\theta(X_1^p))$$

  On estime $g(\theta)$ par $\widehat{g(\theta)}_n = \psi(\overline{X_n^1}, \ldots, \overline{X_n^p})$ où

  $$\overline{X_n^k} = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

  Ces quantités s'appellent les moment empiriques, pour $k = 1$ on parle de moyenne empirique. ψ étant supposée continue, la loi des grands nombres donne

  $$\widehat{g(\theta)}_n \xrightarrow[n \to \infty]{p.s} g(\theta)$$

  On dit que la suite d'estimateurs $(\widehat{g(\theta)}_n)_{n \geqslant 0}$ est fortement consistante ou asymptotiquement sans biais.
- **Méthode du maximum de vraisemblance :** Soient $(\Omega, \mathcal{A}, (P_\theta)_{\theta \in \Theta})$ un modèle statistique et m une mesure σ-finie sur $(\Omega, \mathcal{A})$. On dit que le modèle est dominé par m si $\forall \theta \in \Theta,\ P_\theta \ll m$. Dans ce cas on pose $L_\theta = \frac{dP_\theta}{dm}$ la dérivée de Radon-Nicodym que l'on appelle vraisemblance. On dit que $\widehat{\theta}$ est un estimateur du maximum de vraisemblance de θ si

  $$L_{\widehat{\theta}(\omega)}(\omega) = \max_{\theta \in \Theta} L_\theta(\omega)$$

  Dans le cadre d'un n-échantillon, on verra plus tard que sous des hypothèses raisonnables, l'estimateur du maximum de vraisemblance $\widehat{\theta}_n$ est tel que

  $$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow[n \to \infty]{\mathcal{L}} \mathcal{N}(0, \sigma^2(\theta))$$

  avec $\sigma^2(\theta)$ optimal dans un certain sens.

1.1.4 Tests d'hypothèses

Le principe général est somme toute assez simple. On se place dans le modèle statistique $(\Omega, \mathcal{A}, (P_\theta)_{\theta \in \Theta})$. Soient Θ_0 et Θ_1 deux sous-ensembles disjoint de Θ, en observant ω, on veut tester $\theta \in \Theta_0$ contre $\theta \in \Theta_1$.

On appelle H_0 : "$\theta \in \Theta_0$" l'hypothèse principale ou l'hypothèse nulle et H_1 : "$\theta \in \Theta_1$" l'hypothèse alternative.

Définition: On appelle test toute variable aléatoire $d : \Omega \to \{0, 1\}$.

Un test s'utilise grâce à une règle de décision :

$$\text{On accepte } H_0 \iff d(\omega) = 0$$
$$\text{On rejette } H_0 \iff d(\omega) = 1$$

Pour discuter de la qualité des tests, introduisons les quelques définitions suivantes :

Définition: La fonction puissance d'un test d est $\beta : \Theta_1 \to [0, 1]$ définie par $\beta(\theta) = P_\theta(d = 1)$.
On dit que d est de niveau α si

$$\sup_{\theta \in \Theta_0} P_\theta(d = 1) \leqslant \alpha$$

A priori, un test ou un estimateur dépendent du nombre n d'observations. Ainsi on parlera de niveau asymptotique α si

$$\sup_{\theta \in \Theta_0} \left(\lim_{n \to \infty} P_\theta(d = 1) \right) \leqslant \alpha$$

La puissance représente la probabilité de rejeter H_0 sous H_1. Il semble donc normal de demander à la puissance d'un test d'être aussi prés de 1 que possible. Le niveau quant à lui représente une majoration de la probabilité de rejeter H_0 sous H_0, cette quantité doit donc être plutôt faible. Le niveau traduit l'idée de confiance que l'on pourra avoir dans les résultats du test. Il est important de remarquer qu'un test n'est pas symétrique, en effet un test de H_0 contre H_1 sera en général différent d'un test de H_1 contre H_0.

1.2 La loi du χ^2

1.2.1 La loi $\gamma_{p,a}$

Définition: La fonction gamma est définie sur $\{z \in \mathbb{C} | \Re(z) > 0\}$ par

$$\Gamma(z) = \int_0^{+\infty} x^{z-1} e^{-x} dx$$

Proposition 1.5 *On a* $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

Preuve: On a $\Gamma(\frac{1}{2}) = \int_0^{+\infty} x^{-\frac{1}{2}} e^{-x} dx$. En faisant le changement de variable $u = \sqrt{2x}$, on obtient $\Gamma(\frac{1}{2}) = \sqrt{2} \int_0^{+\infty} e^{-\frac{u^2}{2}} du = \sqrt{\pi}$

■

Proposition 1.6 (Formule des compléments) *Soient* $a, b > 0$.

$$\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \int_0^1 t^{a-1}(1-t)^{b-1}dt$$

Preuve: En posant $t = u^2$ on obtient

$$\Gamma(a) = \int_0^{+\infty} t^{a-1}e^{-t}dt = \int_0^{+\infty} u^{2a-2}e^{-u^2}2udu$$

Les intégrandes dans $\Gamma(a)$ et $\Gamma(b)$ étant positives, le théorème de Fubini-Tonelli nous donne

$$\Gamma(a)\Gamma(b) = 4\int_0^{+\infty}\int_0^{+\infty} e^{-(u^2+v^2)}u^{2a-1}v^{2b-1}dudv$$

Puis, par un changement de variables polaires, on a

$$\begin{aligned}
\Gamma(a)\Gamma(b) &= 4\int_0^{\frac{\pi}{2}}\int_0^{+\infty} e^{-\rho^2}\rho^{2a+2b-2}\sin^{2a-1}(\theta)\cos^{2b-1}(\theta)\rho d\rho d\theta \\
&= 4\int_0^{+\infty}\rho^{2a+2b-1}e^{-\rho^2}d\rho\int_0^{\frac{\pi}{2}}\sin^{2a-1}(\theta)\cos^{2b-1}(\theta)d\theta \\
&= 4\int_0^{+\infty} t^{a+b-1}e^{-t}\frac{dt}{2}\int_0^1 x^{a-1}(1-x)^{b-1}\sqrt{x}\frac{dx}{2\sqrt{x}} \qquad (1.1) \\
&= \Gamma(a+b)\int_0^1 t^{a-1}(1-t)^{b-1}dt
\end{aligned}$$

L'égalité **1.1** provenant des changements de variables $t = \rho^2$ et $\sin(\theta) = \sqrt{x}$.

■

Définition: Soient p et a des réels strictement positifs. On appelle loi gamma de paramètre de forme p et de paramètre d'échelle a, notée $\gamma_{p,a}$, la mesure de probabilité sur $\mathbb{R}_+^* =]0, +\infty[$ qui admet pour densité par rapport à la mesure de Lebesgue la fonction

$$\gamma_{p,a}(x) = \frac{1}{\Gamma(p)}x^{p-1}exp(-\frac{x}{a})a^{-p}\mathbf{1}_{]0,+\infty[}(x)$$

En choisissant les paramètres d'une loi $\gamma_{p,a}$ il est possible de retrouver un grand nombre de loi de probabilités classiques comme les lois exponentielles (prendre $p = 1$ et $a > 0$ pour avoir une loi $\mathcal{E}(a)$), ...

Proposition 1.7 *Soient X et Y deux variables aléatoires indépendantes telles que* $X \sim \gamma_{p,a}$ *et* $Y \sim \gamma_{q,a}$. *Alors* $X + Y \sim \gamma_{p+q,a}$.

Preuve: Notons $d_{X+Y}(z)$ la densité par rapport à la mesure de Lebesgue de la loi de $X+Y$. X et Y étant indépendantes, on obtient cette densité par convolution des densités des lois de X et Y :

$$\begin{aligned}
d_{X+Y}(z) &= \int_{-\infty}^{+\infty} \gamma_{p,a}(x)\gamma_{q,a}(z-x)dx \\
&= \frac{1}{\Gamma(p)\Gamma(q)} e^{-\frac{z}{a}} a^{-(p+q)} \int_{-\infty}^{+\infty} x^{p-1}(z-x)^{q-1}\mathbf{1}_{x>0}\mathbf{1}_{z>x}dx \\
&= \frac{1}{\Gamma(p)\Gamma(q)} e^{-\frac{z}{a}} a^{-(p+q)} z^{q-1} \int_0^z x^{p-1}(1-\frac{x}{z})^{q-1}dx\mathbf{1}_{]0,+\infty[}(z) \\
&= \frac{1}{\Gamma(p)\Gamma(q)} e^{-\frac{z}{a}} a^{-(p+q)} z^{p+q-1}\mathbf{1}_{]0,+\infty[}(z) \int_0^1 y^{p-1}(1-y)^{q-1}dy \\
&= \frac{1}{\Gamma(p+q)} z^{p+q-1}\exp(-\frac{z}{a}) a^{-(p+q)}\mathbf{1}_{]0,+\infty[}(z)
\end{aligned}$$

La dernière égalité venant de la formule des compléments et l'avant-dernière du changement de variable $y = \frac{x}{z}$. On reconnaît la densité d'une loi $\gamma_{p+q,a}$, d'où le résultat.

■

1.2.2 La loi du χ^2

Proposition 1.8 *Soit $(Z_i)_{1\leqslant i\leqslant n}$ une suite de variables aléatoires indépendantes et identiquement distribuées suivant une loi normale centrée réduite $\mathcal{N}(0,1)$. La loi de $Z_1^2 + \cdots + Z_n^2$ est une loi $\gamma_{\frac{n}{2},2}$, appelée loi du χ^2 à n degrés de liberté, notée $\chi^2(n)$.*

Preuve: Pour $n = 1$, on se donne $Z \sim \mathcal{N}(0,1)$ et f une fonction borélienne positive sur $\mathbb{R}$.

$$\begin{aligned}
\mathbb{E}[f(Z^2)] &= \int_{-\infty}^{+\infty} f(t^2)\frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}dt \\
&= 2\int_0^{+\infty} f(t^2)\frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}dt \\
&= \int_0^{+\infty} f(x)\frac{1}{\sqrt{2x\pi}} e^{-\frac{x}{2}}dx \\
&= \int_{-\infty}^{+\infty} f(x)\underbrace{\frac{1}{\Gamma(\frac{1}{2})} x^{\frac{1}{2}-1}e^{-\frac{x}{2}}2^{-\frac{1}{2}}\mathbf{1}_{]0,+\infty[}(x)}_{=\gamma_{\frac{1}{2},2}(x)}\,dx
\end{aligned}$$

Donc $Z^2 \sim \chi^2(1)$.
Ce résultat ainsi que la propriété 1.7 donne alors

$$Z_1^2 + \cdots + Z_n^2 \sim \chi^2(n)$$

■

Théorème 1.9 *Soient $E = \mathbb{R}^n$, H un sous-espace vectoriel de E de dimension $p \leqslant n$ et $Z = (Z_1, \ldots, Z_n) \sim \mathcal{N}(0, I_n)$. Notons Π la matrice dans la base canonique de la projection orthogonale de E sur H. Alors $\|\Pi(Z)\|^2 \sim \chi^2(p)$.*

Preuve:

Comme Π est un projecteur orthogonal, d'après la proposition 1.1 on obtient que $\Pi Z \sim \mathcal{N}(0, \Pi\Pi^*) = \mathcal{N}(0, \Pi)$
Soit u une base orthonormale de E telle que que $(u_1, \ldots, u_p)$ forme une base de H et soit U la matrice de changement de base de la base canonique à la base u.
Alors $U\Pi Z \sim \mathcal{N}(0, U\Pi U^*)$ où $U\Pi U^*$ vaut $\left[\begin{array}{c|c} I_p & 0 \\ \hline 0 & 0 \end{array}\right]$
On pose $U\Pi Z = (\widetilde{Z_1}, \ldots, \widetilde{Z_p}, 0, \ldots, 0)$ où $\widetilde{Z_1}, \ldots, \widetilde{Z_p} \sim \mathcal{N}(0, 1)$.

$$\|\Pi Z\|^2 = \|U\Pi Z\|^2 = \widetilde{Z_1}^2 + \cdots + \widetilde{Z_p}^2$$

Donc $\|\Pi Z\|_E^2 \sim \chi^2(p)$.

■

Chapitre 2

Test du χ^2 d'une hypothèse simple

2.1 Comportements asymptotiques

Soient E un ensemble fini de cardinal K et X une variable aléatoire à valeurs dans E. Notons $a_k,\ k \in \{1, \ldots, K\}$, les éléments de E et

$$p = (P(X = a_k),\ k \in \{1, \ldots, K\})$$

le vecteur des probabilités. Des considérations extérieures donnent pour ce vecteur une certaine valeur :

$$p^0 = (p_k^0,\ k \in \{1, \ldots, K\})$$

On dispose d'un n-échantillon $(X_1, \ldots, X_n)$ de variables aléatoires de même loi que X.
Posons $N_i^n = Card\{k | X_k = a_i\},\ \forall i \in \{1, \ldots, K\}$.

Définition: On appelle fréquences empiriques les quantités

$$\widehat{p_k^n} = \frac{N_k^n}{n}; k = 1..K$$

et on note $\widehat{p^n} = \{\widehat{p_k^n}; k = 1..K\}$ le vecteur des fréquences empiriques.

Définition: On définit la distance du χ^2 entre des lois sur E par

$$\chi^2(p, q) = \sum_{i=1}^{K} \frac{(p_k - q_k)^2}{q_k}$$

où p et q sont les vecteurs des probabilités relatifs à des lois sur E.

Remarquons que cette distance n'en est pas une à proprement dit puisque elle n'est pas symétrique mais elle traduit néanmoins une idée de "proximité" entre lois. Nous allons ainsi pouvoir comparer $\widehat{p^n}$ et p^0. Pour cela il nous faut

étudier la loi de $\chi^2(\widehat{p^n}, p^0)$ qui dépend a priori de K, de p^0 et de n. Le théorème suivant montre que, asymptotiquement en n, cette loi ne dépend en fait plus que de K.

Théorème 2.1 *Supposons $p_k \neq 0,\ \forall k \in \{1, \ldots, K\}$.*
Si on note

$$U_n = \sqrt{n}\left(\frac{\widehat{p_1^n} - p_1}{\sqrt{p_1}}, \ldots, \frac{\widehat{p_K^n} - p_K}{\sqrt{p_K}}\right)$$

Alors

$$U_n \xrightarrow[n \to \infty]{\mathcal{L}} \mathcal{N}(0, I_K - \sqrt{p}\sqrt{p}^T)$$

$$\text{où } \sqrt{p}\sqrt{p}^T = \begin{bmatrix} p_1 & \sqrt{p_1 p_2} & \cdots & \sqrt{p_1 p_K} \\ \sqrt{p_1 p_2} & p_2 & & \vdots \\ \vdots & & \ddots & \vdots \\ \sqrt{p_1 p_K} & \cdots & \cdots & p_K \end{bmatrix}$$

Preuve: Posons $Z^i = (Z_k^i)_{1 \leqslant k \leqslant K}$ avec $Z_k^i = \frac{\mathbf{1}_{\{X_i = k\}} - p_k}{\sqrt{p_k}}$. Étant fonction des X_i, les Z^i sont indépendantes et identiquement distribuées. On a :

$$\mathbb{E}[Z_k^i] = (p_k)^{-\frac{1}{2}}(\underbrace{P(X_i = k)}_{=p_k} - p_k) = 0$$

$$\mathbb{E}[{Z_k^i}^2] = \frac{1}{p_k}(p_k + {p_k}^2 - 2{p_k}^2) = 1 - p_k$$

Pour $k \neq k'$, $\mathbb{E}[Z_k^i Z_{k'}^i] = (p_k p_{k'})^{-\frac{1}{2}}(0 + p_k p_{k'} - p_k p_{k'} - p_k p_{k'}) = -\sqrt{p_k p_{k'}}$
Par le théorème central limite, on a

$$U_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n Z^i \xrightarrow[n \to \infty]{\mathcal{L}} \mathcal{N}(0, I_K - \sqrt{p}\sqrt{p}^T)$$

■

Lemme 2.2 *Soit $Z \sim \mathcal{N}(0, Q)$ où Q est la matrice d'un projecteur orthogonal. Alors $\|Z\|^2 \sim \chi^2(rg(Q))$.*

Preuve: C'est un corollaire du théorème 1.9. En effet, Q étant une projection orthogonale, on a $Q^* = Q^2 = Q$ et donc, si on prend $Y \sim \mathcal{N}(0, I_m)$ où m est la dimension de l'espace entier on peut écrire

$$Z = QY \sim \mathcal{N}(0, \underbrace{QQ^*}_{=Q^2=Q})$$

On conclut par le théorème 1.9 : $\|Z\|^2 \sim \|QY\|^2 \sim \chi^2(rg(Q))$

■

Utilisons maintenant ce lemme pour démontrer le résultat important suivant :

Théorème 2.3 *Sous l'hypothèse* $H_0 = \{p = p_0\}$, *si* $\forall k \in \{1, \ldots, K\},\ p_k^0 \neq 0$, *alors*

$$n\chi^2(\widehat{p^n}, p^0) \xrightarrow[n \to \infty]{\mathcal{L}} \chi^2(K-1)$$

L'hypothèse de non nullité des p_k^0 n'est pas restrictive car dans la perspective d'un test il n'est pas nécessaire de considérer des événements de probabilité nulle pour la "bonne" valeur de p.

Preuve: En reprenant les notations du théorème 2.1 on a $n\chi^2(\widehat{p^n}, p^0) = \|U_n\|^2$. Cette norme étant continue, le théorème 2.1 donne :

$$n\chi^2(\widehat{p^n}, p^0) \xrightarrow[n \to \infty]{\mathcal{L}} \|Z\|^2$$

où $Z \sim \mathcal{N}(0, I_K - \sqrt{p_0}\sqrt{p_0}^T)$. Montrons que la matrice $I_K - \sqrt{p_0}\sqrt{p_0}^T$ est celle d'un projecteur orthogonal de rang $K-1$, le lemme précédent permettra alors de conclure. Notons $(\sqrt{p_0}\sqrt{p_0}^T)^2 = (c_{ij})_{1 \leqslant i,j \leqslant K}$, on a

$$\forall i \in \{1, \ldots, K\},\ c_{ii} = \sum_{k=1}^{K} p_i p_k = p_i \underbrace{\sum_{k=1}^{K} p_k}_{=1} = p_i$$

$$\forall i \neq j \in \{1, \ldots, K\},\ c_{ij} = \sqrt{p_i p_j} \sum_{k=1}^{K} p_k = \sqrt{p_i p_j}$$

Donc $(\sqrt{p_0}\sqrt{p_0}^T)^2 = \sqrt{p_0}\sqrt{p_0}^T$ et $(I_K - \sqrt{p_0}\sqrt{p_0}^T)^2 = I_K - \sqrt{p_0}\sqrt{p_0}^T$. C'est bien la matrice d'un projecteur. De plus on a

$$(I_K - \sqrt{p_0}\sqrt{p_0}^T)\sqrt{p_0} = \sqrt{p_0} - \sqrt{p_0}\underbrace{\sqrt{p_0}^T\sqrt{p_0}}_{=1} = 0$$

Donc $Vect(\sqrt{p_0}) \subset \ker(I_K - \sqrt{p_0}\sqrt{p_0}^T)$. De plus, soit $x = (x_1, \ldots, x_K) \in \mathbb{R}^K$, on note

$$Y = (I_K - \sqrt{p_0}\sqrt{p_0}^T)x = x - \sqrt{p_0}\sum_{k=1}^{K} x_i\sqrt{p_0^k}$$

Alors $Y\sqrt{p_0} = x\sqrt{p_0} - x\sqrt{p_0} = 0$, d'où l'image de $I_K - \sqrt{p_0}\sqrt{p_0}^T$ est inclue dans $Vect(\sqrt{p_0}^{\perp})$ qui est de dimension $K-1$, le théorème du rang permet de conclure.

■

Théorème 2.4 *Sous l'hypothèse* $H_1 = \{p \neq p_0\}$ *et si* $\forall k \in \{1, \ldots, K\}, p_k^0 \neq 0$ *alors :*

$$n\chi^2(\widehat{p^n}, p^0) \xrightarrow[n \to \infty]{p.s} +\infty$$

Preuve: La loi des grands nombre donne $\widehat{p} \xrightarrow[n \to \infty]{p.s} p$. Donc si $p \neq p^0$, on a :

$$\chi^2(\widehat{p^n}, p^0) \xrightarrow[n \to \infty]{p.s} \chi^2(p, p^0) \neq 0$$

Donc $n\chi^2(\widehat{p^n}, p^0) \xrightarrow[n \to \infty]{p.s} +\infty$.

■

Illustration numérique du théorème 2.3

Nous nous proposons dans cette section d'illustrer numériquement la convergence donnée par le théorème 2.3 sur un n-échantillon de Bernoulli de paramètre 0.25. Nous utilisons le programme MAPLE pour réaliser cette illustration.
On s'intéressera surtout à la question : "A partir de quand n vaut-il $+\infty$?". Il n'y a à ce sujet aucun résultat théorique précis. Cependant des considérations heuristiques basées sur des simulations font qu'en général on considère que l'approximation asymptotique est satisfaisante dés que $\inf\{np_k;\ k = 1, \ldots, K\} > 5$. Ici $\inf\ p_k = \frac{1}{4}$, donc le critère nous proposerait de choisir $n > 20$. Nous détaillerons le code du programme pour le cas $n = 25$, puis nous considérerons les résultats pour $n = 15$ et $n = 20$.
Tout d'abord, on se donne la fonction $\gamma_{p,a}(x)$ ainsi que des procédures permettant le calcul de la distance du χ^2 et le tracé d'une courbe à partir d'une liste :

```
> restart;

  c:=1/(GAMMA(p))*1/a^p*x^(p-1)*exp(-x/a):
  c:=unapply(c,a,p,x):

  d:=proc(p,q)
  local i,K,s;
  K:=nops(p);
  s:=0;
  for i from 1 to K do
     s:=s+(p[i]-q[i])^2/q[i]
  od;
  return s;
  end:

  lisse:=proc(l,pas)
  local r,i;
  r:=[[0,b[1]]];
  for i from 1 to nops(l)-1 do
     r:=[op(r),[i*pas,(l[i]+l[i+1])/2]];
  od:
  return r:
  end:
```

Nous allons illustrer la convergence en loi via la convergence de la fonction de répartition de $n\chi^2(\widehat{p^n}, p^0)$ vers celle d'une loi $\chi^2(1)$, à un degré de liberté car ici

$K = 2$. Pour cela on se donne le vecteur p0 des probabilités d'une loi de Bernoulli de paramètre 0.25 ainsi que p n-échantillons tirés suivant une loi de Bernoulli de paramètre 0.25 grâce auxquels nous estimerons la fonction de répartition de $n\chi^2(\widehat{p^n}, p^0)$ (ici p=99 et n=25) :

```
> p0:=[0.75,0.25]:
  n:=25:
  p:=99:


  a:=[]:
  for k from 1 to p do
     L:=[stats[random,binomiald[1,0.25]](n)]:
     M:=[0,0]:
     for i from 0 to 1 do
        for j from 1 to n do
           if L[j]=i then M[1+i]:=M[1+i]+1:end if:
        od:
        M[i+1]:=M[i+1]/n:
     od:
     a:=[op(a),n*d(M,p0)]:
  od:

  b:=[]:
  for k from 1 to 48 do
     b:=[op(b),0]:
     for i from 1 to nops(a) do
        if a[i]<k/8 then b[k]:=b[k]+1:end if:
     od:
     b[k]:=b[k]/nops(a):
  od:
```

En traçant sur un même graphe la fonction de répartition d'un loi $\chi^2(1)$ et l'éstimé de celle de $n\chi^2(\widehat{p^n}, p^0)$, on obtient :

```
> plot([int(c(2,1/2,u),u=0..x),lisse(b,1/8)],x=0..6,color=[wheat,black]);
```

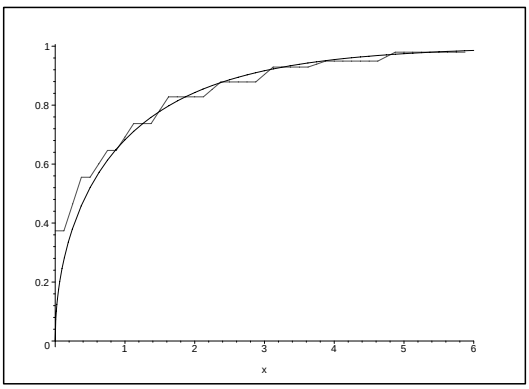

FIG. 2.1: *n=25*

La même procédure nous donne pour $n = 15$ et $n = 20$:

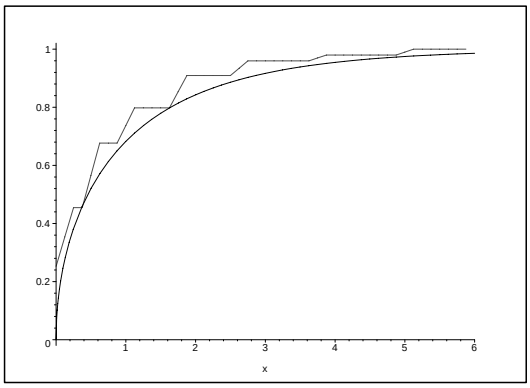

FIG. 2.2: *n=15*

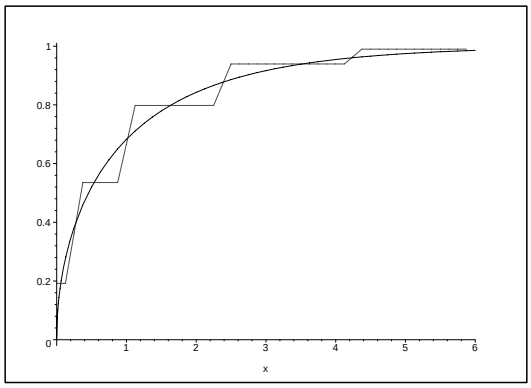

FIG. 2.3: *n=20*

Ces simulations semblent donc en accord avec le critère vu précédemment et $n = 20$ parait bien être un seuil.

2.2 Test du χ^2 d'une hypothèse simple

Les théorèmes de la partie précédente permettent de construire un test de niveau asymptotique $\alpha \in]0,1[$ de $H_0 = \{p = p_0\}$ contre $H_1 = \{p \neq p_0\}$. On se donne $\chi^2_\alpha(K-1)$ le fractile $1-\alpha$ de la loi $\chi^2(K-1)$, c'est-à-dire le réel vérifiant $P(X \leqslant \chi^2_\alpha(K-1)) = 1-\alpha$ où $X \sim \chi^2(K-1)$.
On construit ce test par la règle de décision suivante :

Si $n\chi^2(\widehat{p^n}, p^0) \leqslant \chi^2_\alpha(K-1)$, on accepte H_0.
Si $n\chi^2(\widehat{p^n}, p^0) > \chi^2_\alpha(K-1)$, on rejette H_0.

Cela définit bien un test de niveau asymptotique α car

$$P_{H_0}(RejeterH_0) = P_{H_0}(n\chi^2(\widehat{p^n}, p^0) > \chi^2_\alpha(K-1))$$

et le théorème 2.3 donne la convergence de cette probabilité vers

$$P_{H_0}(\chi^2(K-1) > \chi^2_\alpha(K-1)) = \alpha$$

car $]\chi^2_\alpha(K-1), +\infty[$ est un borélien dont la frontière est de mesure de Lebesgue nulle. De plus, la puissance de ce test tend vers 1 quand n tend vers l'infini, en effet le théorème 2.4 donne

$$P_{H_1}(RejeterH_0) = P_{H_1}(n\chi^2(\widehat{p^n}, p^0) > \chi^2_\alpha(K-1)) \xrightarrow[n \to \infty]{} 1$$

Illustration numérique de la convergence de la puissance

Reprenons les fonctions crées lors de la première simulation. En se replaçant dans le cas d'un n-échantillon de Bernoulli de paramètre $\frac{1}{4}$, nous allons maintenant regarder le comportement de la fonction $f : \theta \mapsto P_\theta(n\chi^2(\widehat{p^n}, p^0) > \chi^2_\alpha(1))$

quand $n \to \infty$. La théorie donne que $f(\theta)$ tend vers 1 si $\theta \neq \frac{1}{4}$ et vers le niveau du test α pour $\theta = \frac{1}{4}$. Nous faisons les simulations pour les valeurs de n=25, 100 et 500. Elles consistent à se donner 1000 n-échantillons comme précédemment puis d'estimer $P_\theta(n\chi^2(\widehat{p^n}, p^0) > \chi^2_\alpha(1))$ par la quantité

$$\frac{1}{1000}\sum_{j=1}^{1000} \mathbf{1}_{P_\theta(n\chi^2(\widehat{p^n},p^0)>\chi^2_\alpha(1))}$$

```
> N:=1000:
  graph:=[[],[],[]]:
  alpha:=0.05:
  F:=fractchi2(alpha,1):
  for t from 0 to 2 do
  n:=25-175*t/2+(325*t^2)/2:
     for z from 0.05 to 0.95 by 0.025 do
        p0:=[1-z,z]:
        lst:=[]:
        res:=0:
        for j from 1 to N do
           Obs:=[stats[random,binomiald[1,0.25]](n)]:
           pM:=[0,0]:
           for k from 1 to n do
              if Obs[k]=0 then pM[1]:=pM[1]+1:end if:
           od:
           pM[2]:=n-pM[1]:
           pM:=pM/n:
           lst:=[op(lst),n*d(pM,p0)]:
        od:
        for j from 1 to N do
            if lst[j]>F then res:=res+1:end if:
        od:
        res:=res/N:
        graph[t+1]:=[op(graph[t+1]),[z,evalf(res)]]:
     od:
  od:
```

Puis on trace les résultats sur un même graphe avec le niveau α :

```
> plot([graph[1],graph[2],graph[3],0.05],x=0..1,linestyle=[SOLID,DASH,DOT,DASHDOT],
  color=[red,red,red,black]);
```

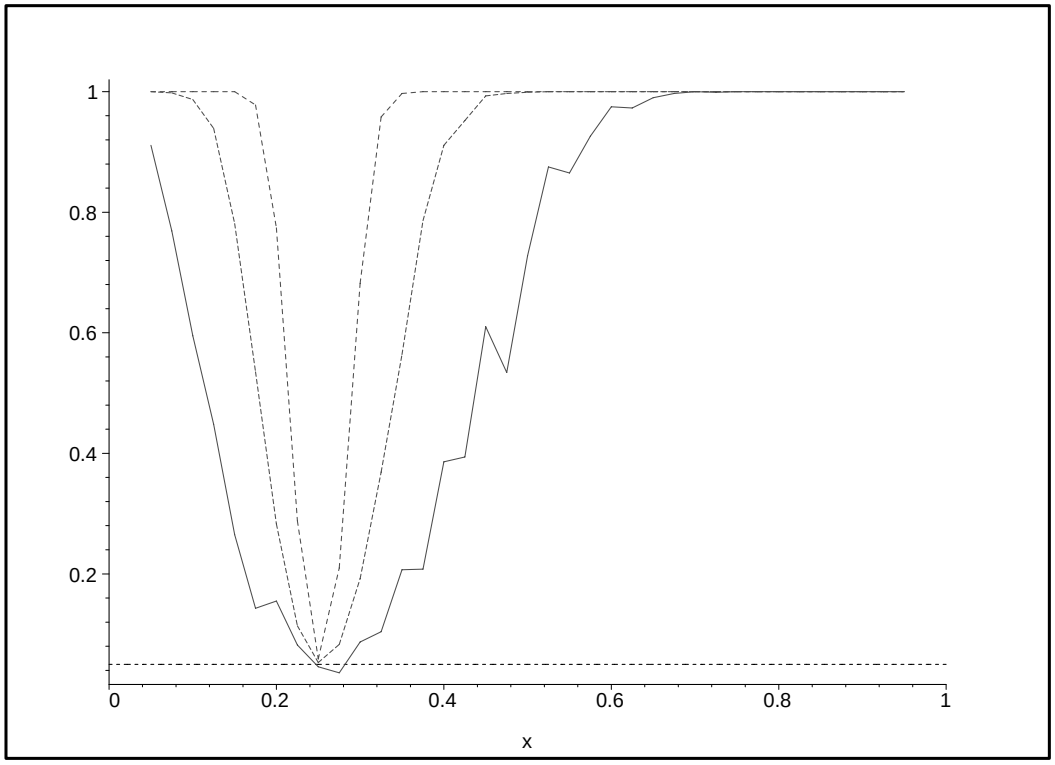

FIG. 2.4: *En continue pour n=25, en discontinu intermédiaire pour n=100 et en discontinu pour n=500*

2.3 L'expérience de Mendel

Gregor Mendel est reconnu comme le père fondateur de la génétique. Il a en effet découvert au milieu du XIX^e siècle les lois de transmission des caractères.

Pour ses expériences, il choisit de travailler sur des pois comestibles présentant sept caractères dont chacun peut se retrouver sous deux formes différentes, aisément identifiables : forme et couleur de la graine, couleur de l'enveloppe, forme et couleur de la gousse, position des fleurs et longueur de la tige. La première expérience qu'il décrira dans son article consiste à étudier les résultats d'hybridation obtenus pour l'une des paires de caractères seulement.

Pour cela, il croise deux variétés dont l'une présente des graines lisses et l'autre des graines ridées. Les résultats montrent que tous les hybrides produits (génération F1) ont des graines lisses. La saison suivante, Mendel sème les graines hybrides lisses et obtient, par auto-fécondation, une génération F2 qui présente à la fois des graines lisses et des graines ridées dans les proportions de 3 pour 1. Du fait que le caractère "ridé" soit réapparu à la seconde génération, sans intervention externe, Mendel déduit qu'il était resté présent dans l'hybride de manière latente. Comme le caractère "lisse" avait pris le pas sur le caractère "ridé", il appelle le premier "dominant", noté A, et le second "récessif", noté a. Il conclut que les hybrides reçoivent un facteur (allèle) de chacun des parents et rejette ainsi définitivement la notion d'hérédité par mélange : les caractères des

végétaux sont des quantités discrètes. Mendel déduit que la génération hybride F1 présente un génotype Aa (phénotype = caractère dominant) alors que le génotype de la génération F2 sont AA Aa ou aa (en proportions $\frac{1}{4}$, $\frac{1}{2}$ et $\frac{1}{4}$) ce qui correspond à une répartition en phénotype : $\frac{3}{4}$ dominant, $\frac{1}{4}$ récessif.

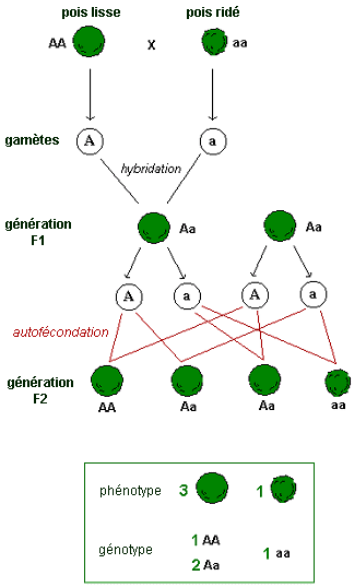

FIG. 2.5: *Hybridation*

Il s'intéresse ensuite à la transmission de plusieurs caractères et énonce la deuxième loi de Mendel :les caractères sont transmis de manière indépendante selon les mêmes rapports que ceux trouvés précédemment. Mendel veut confirmer ces lois par une expérience.

Pour cette expériences, dite de dihybridisme, Mendel passe tout d'abord 2 ans à sélectionner des races pures, c'est-à-dire des plantes qui, croisées entre elles, transmettent toujours le même caractère à leur descendance. Il croise deux variétés pures de pois différant par deux caractères. L'une était lisse et jaune, l'autre était ridée et verte. La première génération F1 donne des pois tous identiques, lisses et jaunes. La seconde génération F2 donne quant à elle des pois de couleur et de forme différentes. La première génération montre que le caractère jaune (J) est dominant sur vert (V) et lisse (l) sur ridé (r).

On peut donc résumer les résultats théoriques de l'expérience par le tableau suivant :

	Jl	**Jr**	**Vl**	**Vr**
Jl	JJ ll→**Jl**	JJ lr→**Jl**	JV ll→**Jl**	JV lr→**Jl**
Jr	JJ lr→**Jl**	JJ rr→**Jr**	JV lr→**Jl**	JV rr→**Jr**
Vl	JV ll→**Jl**	JV lr→**Jl**	VV ll→**Vl**	VV lr→**Vl**
Vr	JV lr→**Jl**	JV rr→**Jr**	VV lr→**Vl**	VV rr→**Vr**

Les résultats obtenus par Mendel sont les suivants :

Nombre de pois jaune et lisse	$N_{Jl} = 315$
Nombre de pois vert et lisse	$N_{Vl} = 108$
Nombre de pois jaune et ridé	$N_{Jr} = 101$
Nombre de pois vert et ridé	$N_{Vr} = 32$

Notons $p^0_{Jl} = \frac{9}{16}, p^0_{Jr} = \frac{3}{16}, p^0_{Vl} = \frac{3}{16}, p^0_{Vr} = \frac{1}{16}$ et $p^0 = (p^0_{Jl}\ ,\ p^0_{Jr}\ ,\ p^0_{Vl}\ ,\ p^0_{Vr})$. On veut tester $H_0 = \{p = p^0\}$ contre $H_1 = \{p \neq p^0\}$ en utilisant le test du χ^2. On se fixe un niveau $\alpha = 0.05$ et on lit sur les tables du χ^2 la valeur du fractile associé (ici K=4) : $\chi^2_\alpha(3) = 7.815$. Dans l'expérience de Mendel il y avait $n = 556$ pois et donc un vecteur des fréquences empiriques égal à :

$$\begin{aligned} \widehat{p^n} &= (\widehat{p^n_{Jl}}\ ,\ \widehat{p^n_{Jr}}\ ,\ \widehat{p^n_{Vl}}\ ,\ \widehat{p^n_{Vr}}) \\ &= (0.566\ ,\ 0.182\ ,\ 0.194\ ,\ 0.058) \end{aligned}$$

Le calcul de la distance du χ^2 donne :

$$\begin{aligned} \chi^2(\widehat{p^n}, p^0) &= \frac{(0.566 - 0.5625)^2}{0.5625} + \frac{(0.181 - 0.1875)^2}{0.1875} + \frac{(0.193 - 0.1875)^2}{0.1875} \\ &+ \frac{(0.057 - 0.0625)^2}{0.0625} \\ &= 8.924 \cdot 10^{-4} \end{aligned}$$

Donc $n\chi^2(\widehat{p^n}, p^0) = 0.496 < 7.815$, on accepte H_0.

Ici $\inf\{np_k;\ k = 1, \ldots, K\} = 34.75 > 5$ donc le niveau asymptotique est acceptable. On sait aujourd'hui que le modèle de Mendel est correct (comme le confirme le test), mais les résultats obtenus sont sujets à discussion. En effet, les expériences de Mendel portaient en fait sur 7 caractères différents. Or le pois ne possède que 7 paires de chromosomes (soit 14 chromosomes) et les gènes liés aux expériences de Mendel étaient précisément portés chacun sur une paire de chromosomes différente. Mendel a pu avoir de la chance mais il se peut aussi qu'il ait fait d'autres expériences dont les résultats était moins facilement interprétables. Ceci est directement lié à l'hypothèse sous-entendue de l'indépendance entre la transmission du gène de la couleur et celle du gène de la peau à laquelle Mendel ne pouvait pas pensait car, à l'époque, la notion de gène était inconnue. D'autres raisons d'ordre génétiques poussent aussi à penser que Mendel n'aurait publié que les résultats en accord avec sa théorie (ou simplement truqué ses résultats...) pour que celle-ci soit acceptée. De plus, d'un point de vue probabiliste, la valeur de $n\chi^2(\widehat{p^n}, p^0) = 0.496$ est surprenante car, si on considère que n est assez grand pour que la loi de $n\chi^2(\widehat{p^n}, p^0)$ soit assez proche de celle d'un $\chi^2(3)$, on a

$$\begin{aligned} P(n\chi^2(\widehat{p^n}, p^0) \leqslant 0.5) &= \int_0^{0.5} \frac{1}{2\sqrt{2}\Gamma(\frac{3}{2})} e^{-\frac{x}{2}} \sqrt{x} dx \\ &= \int_0^{0.5} \frac{1}{2\sqrt{2\pi}} e^{-\frac{x}{2}} \sqrt{x} dx \\ &\simeq 0.04055 \end{aligned}$$

Une probabilité si faible et donc une telle proximité avec p^0 sont suspectes, mais cependant Mendel reste connu pour avoir été le premier à proposer ce modèle de transmission des caractère.

Chapitre 3

Test d'indépendance

3.1 Motivations et notations

On observe deux caractères, l'un, X à valeurs dans $\{1, ..., I\}$ et l'autre, Y à valeurs dans $\{1, .., J\}$. Le couple (X,Y) est donc à valeurs dans $\{1, ..., I\} \times \{1, .., J\}$ et suit la loi $P(X = i, Y = j) = p_{ij}$. L'hypothèse nulle que l'on désire tester correspond à l'indépendance entre les variables X et Y.

On désire réaliser un test asymptotique de notre hypothèse. Ici, on ne peut plus évaluer la distance χ^2 entre la fréquence empirique et la valeur théorique. On va essayer de mesurer une distance χ^2 entre les fréquences empiriques et l'hypothèse nulle. On va ainsi étudier la distance entre les fréquences empiriques et le maximum de vraisemblance.
L'hypothèse nulle peut s'exprimer de la façon suivante :

$$p \in \Theta := \{(p_{ij})_{i=1..I, j=1..J}; p_{ij} = p_{i+}p_{+j}\}$$

où la notation "+"signifie la somme sur l'indice remplacé, par exemple

$$p_{i+} = \sum_{j=1}^{J} p_{ij}$$

On peut paramètrer l'ensemble Θ par a_i ; i=1,I; b_j ; j=1,J avec $\sum a_i = 1$, $\sum b_j = 1$ et $p_{ij} = a_i b_j$. On note encore N_{ij}^n l'effectif de la combinaison ij sur n observations indépendantes.
Etudions maintenant le maximum de vraisemblance d'un n-échantillon.

Lemme 3.1 *Sous l'hypothèse H_0 : $p \in \Theta$, l'estimateur du maximum de vraisemblance $\widetilde{p_{ij}}$ de p_{ij} vaut $\widetilde{p_{ij}} = \frac{(N_{i+}^n N_{+j}^n)}{n^2}$.*

Preuve: En prenant comme mesure de référence la mesure de comptage, on obtient la vraisemblance (à une constante près) :

$$V^n(a_1...a_I, b_1...b_J)(\omega) = \prod_{ij} a_i b_j{}^{N_{ij}^n} = \prod_i a_i{}^{N_{i+}^n} \prod_j b_j{}^{N_{+j}^n}$$

On peut donc rechercher $\widetilde{a_i}$ indépendamment de $\widetilde{b_j}$.
On remarque que si N_{i+}^n (resp. N_{+j}^n) est nul pour un certain i (resp. j), le

maximum est atteint pour $a_i = 0$ (resp. $b_j = 0$). On peut donc supposer que tous les $N^n_{i+}, N^n_{+j}, a_i, b_j$ sont non nuls.
Remplaçons a_I (resp. b_J) par $1 - \sum_{i=1}^{I-1} a_i$(resp. $1 - \sum_{j=1}^{J-1} b_j$) dans l'expression de la variance .En dérivant par rapport à a_i, on obtient

$$\frac{\partial V^n(p)}{\partial a_i} = V^n(p)(\frac{N^n_{i+}}{a_i} - \frac{N^n_{I+}}{a_I})$$

En dérivant par rapport à b_j,

$$\frac{\partial V^n(p)}{\partial b_j} = V^n(p)(\frac{N^n_{+j}}{b_j} - \frac{N^n_{+J}}{b_J})$$

par résolution de systême, on obtient :

$$\widetilde{a_i} = \frac{N^n_{i+}}{n} \quad ; \quad \widetilde{b_j} = \frac{N^n_{+j}}{n}$$

■

On veut maintenant étudier la loi limite de la statistique du χ^2 suivante $n\chi^2(\widehat{p^n}, \widetilde{p^n})$ qui mesure la distance entre la fréquence empirique et H_0.

Remarque : Le choix d'une telle statistique sera justifiée dans un cadre plus général dans le chapitre 4.

3.2 Changement de variable dans le théorème central limite

Commençons par démontrer deux lemmes généraux sur les convergences de variables aléatoires. Ils seront utiles aussi bien pour le changement de variables dans le théorème central limite que pour la généralisation du test du χ^2.

Lemme 3.2 (Slutsky) *Soient X_n et Y_n des variables aléatoires à valeurs dans $\mathbb{R}^k$ et $\mathbb{R}^l$*
Si $X_n \xrightarrow[n\to\infty]{\mathcal{L}} X$ et $Y_n \xrightarrow[n\to\infty]{\mathcal{L}} Y$ où Y est constante alors

$$(X_n, Y_n) \xrightarrow[n\to\infty]{\mathcal{L}} (X, Y)$$

Preuve: Soit $\xi \in \mathbb{R}^{k+l}$

$$\begin{aligned}\mathbb{E}(e^{i\xi(X_n,Y_n)}) &= \mathbb{E}(e^{i\xi(X_n,Y)}) + \mathbb{E}(e^{i\xi(X_n,Y_n)} - e^{i\xi(X_n,Y)}) \\ &= e^{i\xi_2 Y}\mathbb{E}(e^{i\xi_1 X_n}) + \mathbb{E}(e^{i\xi_1 X_n}(e^{i\xi_2(Y_n)} - e^{i\xi_2(Y)}))\end{aligned}$$

$|\mathbb{E}(e^{i\xi_1 X_n} e^{i\xi_2(Y_n - Y)})| \leq \mathbb{E}(|e^{i\xi_2(Y_n)} - e^{i\xi_2(Y)}|)$
Or $\phi : x \to |e^{i\xi_2(x)} - e^{i\xi_2(Y)}|$ est continue bornée donc le deuxième terme de la somme tend vers zéro

De plus $\mathbb{E}(e^{i\xi_1 X_n}) \xrightarrow[n\to\infty]{} \mathbb{E}(e^{i\xi_1 X})$ (théorème de lévy)

On obtient alors $\mathbb{E}(e^{i\xi(X_n,Y_n)}) \xrightarrow[n\to\infty]{} e^{i\xi_2 Y}\mathbb{E}(e^{i\xi_1 X}) = \mathbb{E}(e^{i\xi(X,Y)})$

Une nouvelle application du théorème de lévy nous donne, $(X_n, Y_n) \xrightarrow[n\to\infty]{\mathcal{L}}$ (X,Y)

■

Lemme 3.3 *Soient X_n et Y_n des variables aléatoires réelles qui vérifient les hypothèse du lemme précédent alors ;*

1. $(X_n + Y_n) \xrightarrow[n\to\infty]{\mathcal{L}} X+Y$
2. $X_nY_n \xrightarrow[n\to\infty]{\mathcal{L}} XY$

De plus si Y=0 alors $Y_nX_n \xrightarrow[n\to\infty]{(P)} 0$

Preuve: Les deux premiers résultats s'obtiennent en appliquant le lemme 3.2 et en remarquant qu'une fonction continue borné composée avec la fonction somme (resp. multiplication) est toujours continue bornée.

Démontrons le troisième résultat : en appliquant la première partie du lemme, on obtient que Y_nX_n converge en loi vers 0. Or, la fonction qui à x associe $|x| \wedge 1$ est continue bornée donc $\mathbb{E}(|X_nY_n| \wedge 1) \xrightarrow[n\to\infty]{} 0$ ce qui est équivalent à la convergence en probabilité de X_nY_n.

■

L'étude de la convergence en loi de $n\chi^2(\widehat{p^n}, \widetilde{p^n})$ va nécessiter des convergences en lois plus complexes que le théorème central limite. On va donc utiliser le théorème suivant :

Théorème 3.4 (Changement de variable dans le théorème central limite)

Soit T_n une suite de variables aléatoires à valeurs dans $\mathbb{R}^k$ telle que

$$\sqrt{n}(T_n - m) \xrightarrow[n\to\infty]{\mathcal{L}} N(0, \Gamma) \tag{3.1}$$

où m est un point de $\mathbb{R}^k$ et Γ une matrice (k,k). Soit g une fonction d'un voisinage U de m vers $\mathbb{R}^p$ de classe C^2, à dérivées secondes bornées. On note D_g la matrice jacobienne de g en m. Alors en définissant de manière quelconque g en dehors de U :

$$\sqrt{n}(g(T_n) - g(m)) \xrightarrow[n\to\infty]{\mathcal{L}} N(0, D_g\Gamma D_g^T)$$

Preuve: Supposons dans un premier temps que g est à valeurs dans $\mathbb{R}$ et soit $x \in U$ que l'on peut supposer être une boule ouverte. On écrit la formule de Taylor Lagrange :

$$g(x) - g(m) =< x - m, D_g > + \frac{1}{2}(x-m)^T D_g^2(m*)(x-m)$$

où m^* appartient au segment $[x, m]$ qui est inclus dans U. On en déduit

$$\begin{aligned}
\sqrt{n}(g(T_n) - g(m)) &= \sqrt{n} < T_n - m, D_g > \mathbf{1}_{T_n \in U} + \\
&+ \frac{1}{2}\sqrt{n}(T_n - m)^T D_g^2(m_n*)(T_n - m)\mathbf{1}_{T_n \in U} + \\
&+ \sqrt{n}(g(T_n) - g(m))\mathbf{1}_{T_n \notin U} \\
&= (A_n + B_n)\mathbf{1}_{T_n \in U} + C_n \mathbf{1}_{T_n \notin U} \qquad (3.2)
\end{aligned}$$

Comme D_g^2 est borné dans U, $|B_n| \leq (cste)\frac{1}{\sqrt{n}}\|\sqrt{n}(T_n - m)\|^2$ tend vers zéro en probabilité en utilisant l'équation (3.1) et le lemme 3.3
De même du lemme on tire

$$|T_n - m| \xrightarrow[n \to \infty]{(P)} 0 \qquad \text{et donc} \qquad \mathbf{1}_{T_n \notin U} \xrightarrow[n \to \infty]{(P)} 0$$

D'après le lemme 3.3, $\sqrt{n}(g(T_n - m))$ converge donc en loi vers la même limite que $< \sqrt{n}(T_n - m), D_g(m) >$. Or, on a

$$< \sqrt{n}(T_n - m), D_g > \xrightarrow[n \to \infty]{\mathcal{L}} N(0, D_g \Gamma D_g^T)$$

ce qui donne bien le résultat recherché.
Dans le cas général où g est à valeurs dans $\mathbb{R}^m$, en considérant $h =< v, g >$ pour tout $v \in \mathbb{R}^m$, on obtient :

$$\sqrt{n} < (g(T_n) - g(m)), v > \xrightarrow[n \to \infty]{\mathcal{L}} \mathrm{N}(0, v D_g \Gamma D_g^T v^T)$$

Ce qui en appliquant le théorème de Lévy en $\xi = 1$ donne

$$E(exp(i\sqrt{n} < (g(T_n) - g(m)), v >)) \xrightarrow[n \to \infty]{} exp(-\frac{v D_g \Gamma D_g^T v^T}{2})$$

ce qui donne d'après le théorème de Lévy :

$$\sqrt{n}(g(T_n) - g(m)) \xrightarrow[n \to \infty]{\mathcal{L}} \mathrm{N}(0, D_g \Gamma D_g^T)$$

■

Appliquons maintenant ce théorème pour étudier la convergence de $n\chi^2(\widehat{p^n}, \widetilde{p^n})$

3.3 Test d'indépendance

On utilise ici les notations définis au début de cette partie. En étudiant la distance $n\chi^2(\widehat{p^n}, \widetilde{p^n})$, on va essayer de réaliser un test du même type que dans la première partie.

Théorème 3.5 *Supposons qu'aucune des probabilités p_{ij} n'est nulle.*

1. *Sous H_0, $P \in \Theta$,*

$$n\chi^2(\widehat{p^n},\widetilde{p^n}) \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2((I-1)(J-1))$$

2. *Sous l'hypothèse générale* H_1 $P\notin\Theta$

$$n\chi^2(\widehat{p^n},\widetilde{p^n}) \xrightarrow[n\to\infty]{(P)} +\infty$$

Preuve:

1. Le principe de cette démonstration est le même que la démonstration du χ^2 simple :

   - on démontre la convergence en loi de $\sqrt{n}(\frac{\widehat{p_{ij}^n}-\widetilde{p_{ij}^n}}{\sqrt{p_{ij}}})_{ij}$ vers une normale centrée $\mathcal{N}(0,T)$
   - on démontre que T est un projecteur orthogonal de rang $(I-1)(J-1)$ et on conclut en prenant la norme $\|.\|_2$

   Remarque : Dans cette démonstration, les vecteurs seront de taille IJ, et les matrices carrées de tailles $IJ\times IJ$.

   soit $g:\mathbb{R}^{IJ}\to\mathbb{R}^{IJ}\quad (z_{ij})\mapsto(z_{i+}z_{+j})$
   On a bien $\widetilde{p}=g(\widehat{p})$
   Calculons sa diférentielle au point $(p_{ij}=a_ib_j)_{ij};\Sigma a_i=\Sigma b_j=1$
   Pour ce faire appliquons g à $((q_{ij})_{ij}=(p_{ij}+h\Delta_{ij})_{ij}$.
   On obtient $q_{i+}=p_{i+}+h\Delta_{i+}\,;\quad q_{+j}=p_{+j}+h\Delta+j$ et donc
   $q_{i+}q_{+j}=p_{i+}p_{+j}+hp_{i+}\Delta_{+j}+hp_{+j}\Delta_{i+}+o(h)$.
   Ce qui montre que la différentielle de g est l'application

   $$(\Delta_{ij})_{ij}\to(a_i\Delta_{+j}+b_j\Delta i+)_{ij}$$

   Notons Q la matrice de cet opérateur et $\bar{Q}=Id-Q$

   D'après le théorème central limite,
   $\sqrt{n}(\widehat{p_{ij}^n}-p_{ij})_{ij}\xrightarrow[n\to\infty]{\mathcal{L}}N(0,D(Id-\sqrt{p}\sqrt{p}^T)D)$ où $D=Diag(\sqrt{p_{ij}})$
   et $\sqrt{p}=(\sqrt{p_{ij}})_{ij}$

   En appliquant le théorème 3.4 à $h=id-g$, on obtient

   $$\sqrt{n}(\widehat{p_{ij}^n}-\widetilde{p_{ij}^n})_{ij}\xrightarrow[n\to\infty]{\mathcal{L}}N(0,\bar{Q}D(Id-\sqrt{p}\sqrt{p}^T)D\bar{Q}^T)$$
   $$\text{donc }\sqrt{n}(\tfrac{\widehat{p_{ij}^n}-\widetilde{p_{ij}^n}}{\sqrt{p_{ij}}})_{ij}\xrightarrow[n\to\infty]{\mathcal{L}}N(0,D^{-1}\bar{Q}D(Id-\sqrt{p}\sqrt{p}^T)D\bar{Q}^TD^{-1})$$

   Il reste encore à montrer que $D^{-1}\bar{Q}D(Id-\sqrt{p}\sqrt{p}^T)D\bar{Q}^TD^{-1}$ est bien un projecteur orthogonal de rang $(I-1)(J-1)$.

   Étudions d'abord $D^{-1}QD$ sur $(\sqrt{p}^\perp)$:
   – C'est un **projecteur** : $(D^{-1}QD)^2{=}D^{-1}QD$
   soit $u\in\sqrt{p}^\perp$, $Du=(\sqrt{p_{ij}}u_{ij})_{ij}=(z_{ij})_{ij}$ vérifie $\sum_{ij}z_{ij}=0$
   Soit $t=(a_iz_{+j}+b_jz_{i+})_{ij}=Qz$ alors $t_{i+}=\underbrace{a_iz_{++}}_{=0}+z_{i+}=z_{i+}$;

$t_{+j} = \underbrace{b_j z_{++}}_{=0} + z_{+j} = z_{+j}$ ce qui prouve que $Qt = t$ c'est à dire $D^{-1}Q^2D^{-1} = D^{-1}QD^{-1}$

- **image** En notant toujours $Du = z$, l'image est l'ensemble des $\frac{a_i z_{+j} + b_j z_{i+}}{\sqrt{a_i b_j}}$ où z_{i+} et z_{+j} peuvent prendre des valeurs quelconques sous la contrainte $\sum_i z_{i+} = \sum_j z_{+j} = 0$
  En effet, pour obtenir les deux vecteurs $(v_i)_i$ et $(w_j)_j$ quelconques vérifiant ces contraintes il suffit de poser $z_{ij} = (v_i/I) + (w_j/J)$. L'image est donc un espace de dimension $(I-1)+(J-1)$. Cela montre aussi que$(\sqrt{p}^{\perp})$ est stable par $D^{-1}QD$
- **noyau** Notons toujours $Du = z$. Remarquons que si $z_{i+} = z_{+j} = 0\,;\forall i \forall j\,;$ alors $Qz = 0$. Le noyau contient donc tous les vecteurs u tels que :

$$\forall i \sum_j u_{ij}\sqrt{p_{ij}} = 0 \quad ; \quad \forall j \sum_i u_{ij}\sqrt{p_{ij}} = 0 \tag{3.3}$$

  avec $p_{ij} = a_i b_j$. Ces vecteurs u sont tous des éléments de $\sqrt{p}^{\perp}$. On a un systême à $I+J$ équations, l'espace des solutions est donc de dimension supérieures à $(IJ-1)-(I+J) = (I-1)(J-1)$. Par application du théorème du rang, on a obtenu tout le noyau.
- **orthogonalité** Soient u appartenant au noyau et v appartenant à l'image,

$$v_{ij} = \frac{a_i c_j + b_j d_i}{\sqrt{a_i b_j}}; \quad avec \sum_i c_i = \sum_j d_j = 0$$

$$< u, v >= \sum_{ij} u_{ij} \frac{a_i c_j + b_j d_i}{a_i b_j}\sqrt{a_i b_j} = \sum_{ij} (\frac{c_j}{b_j} + \frac{d_i}{a_i}) u_{ij}\sqrt{a_i b_j} = 0,$$

  en utilisant la relation (3.3).

On a montré que $D^{-1}QD$ vu comme opérateur sur $(\sqrt{p}^{\perp})$ est un projecteur orthogonal de rang $(I-1)+(J-1)$. De plus $vect(\sqrt{p})$ est stable par $D^{-1}QD$. On obtient donc que $D^{-1}\bar{Q}D(Id - \sqrt{p}\sqrt{p}^T)D\bar{Q}^T D^{-1}$ est un projecteur orthogonal de rang $IJ - 1 - ((I-1)+(J-1)) = (I-1)(J-1)$.

En utilisant le lemme 2.2, on obtient bien

$$n\chi^2(\widehat{p^n}, \widetilde{p^n}) \quad \xrightarrow[n \to \infty]{\mathcal{L}} \quad \chi^2((I-1)(J-1))$$

2. Sous $P \notin \theta$, on a par la loi forte des grands nombres

$$\widehat{p^n}_{ij} \xrightarrow[n\to\infty]{p.s.} p_{ij} \text{ et}$$

$$\widetilde{p^n}_{ij} \xrightarrow[n\to\infty]{p.s.} \sum_k p_{ik} \sum_k p_{kj} = q_{ij}$$

   donc $\chi^2(\widehat{p^n}, \widetilde{p^n}) \xrightarrow[n \to \infty]{(P)} \chi^2(p,q)$,

   et on obtient $n\chi^2(\widehat{p^n}, \widetilde{p^n}) \xrightarrow[n \to \infty]{(P)} +\infty$

■

Remarque : Là encore, l'hypothèse aucune des p_{ij} est nulle n'est pas très contraignante.

Règle de test Le théorème 3.5 permet de construire un test asymptotique de l'hypothèse H_o contre H_1 par la règle de test suivante : Soit $\alpha \in [0,1]$ le niveau recherché et $\chi^2_\alpha((I-1)(J-1))$ le fractile 1-α de la loi $\chi^2((I-1)(J-1))$,

1. si $n\chi^2(\widehat{p^n}, \widetilde{p^n}) > \chi^2_\alpha((I-1)(J-1))$, on choisit H_1
2. sinon, on choisit H_0

Comme pour le test du χ^2 simple, on voit bien que le niveau asymptotique est α et la foncrion puissance tend vers 1 sur A^c.

Chapitre 4

Test de modèles réguliers

On cherchera dans cette partie à généraliser le test du χ^2 à des tests de modèles plus complexes. De manière intuitive, on va essayer de définir une sorte de distance entre nos fréquences empiriques et notre modèle. On va ainsi l'estimer par la distance χ^2 entre les fréquences empiriques et le maximum de vraisemblance.

Cependant ces résultats ne seront vrais qu'à certaines conditions sur le modèle. On introduira pour cela la notion de modèle régulier et on démontrera certaines propriétés de l'estimateur du maximum de vraisemblance.

Ce test permettra de redémontrer les résultats du chapitre 3 ou de construire un test statistique sur une modélisation du crossing-over.

4.1 Modèles réguliers

Soit Θ une ensemble paramètré et, pour tout $\theta \in \Theta$, F_θ une loi sur $(E, \mathcal{E})$. On observe une suite $(X_n)_{n\geq 1}$ d'observations indépendantes, de loi F_θ : on peut prendre $(\Omega,\ \mathcal{A},\ P_\theta) = (E,\ \mathcal{E},\ F_\theta)^N$, On suppose que pour tout θ, F_θ est dominée par une loi F sur $(E, \mathcal{E})$. Soit f_θ une vraisemblance, alors on a $F_\theta^{\otimes n} = f_\theta^{\otimes n} F^{\otimes n}$ en notant $f_\theta^{\otimes n}(x_1, ..., x_n) = f_\theta(x_1)...f_\theta(x_n)$. Dans la suite on notera $L_n(\theta) = f_\theta^{\otimes n}(X_1, ..., X_n)$

Remarque : Dans la suite, tous les modèles que nous considérerons seront dominés par la mesure de comptage car E sera fini.

4.1.1 Définition

Définition: Modèle régulier
Le modèle précédent où Θ est un ouvert de $\mathbb{R}^k$ est dit régulier, si l'on peut choisir une vraisemblance (L) satisfaisant les hypothéses suivantes :

- H1) Pour tout ω, $\theta \to L_\theta(\omega)$ est deux fois continûment différentiables sur Θ.

- H2) Pour tout $\theta \in \Theta$, $grad$ log $L(\theta)$ est une v.a. centrée de carré intégrable pour P_θ. Pour $1 \leq i, j \leq k$ et $\theta = (\theta_1, ..., \theta_k)$:

$$\mathbb{E}_\theta \left(\frac{\partial}{\partial \theta_i} \log L(\theta) \frac{\partial}{\partial \theta_j} \log L(\theta) \right) = -\mathbb{E}_\theta \left(\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log L(\theta) \right)$$

  Cette quantité est notée $I^{ij}(\theta)$. La matrice $I(\theta) = \{I^{ij}(\theta)\}_{1 \leq i,j \leq k}$ est appelée matrice d'information de Fisher en θ à l'instant n.
- H3) $I(\theta)$ est inversible pour tout θ

Remarque : L'hypothèse H2) est vérifiée si on peut échanger dérivations et intégrations de la log vraisemblance par rapport à P_θ

Remarque : Dans le cas que nous considérerons où la variable aléatoires prend un nombre fini de valeurs ($\|E\| < \infty$), le modèle est régulier si :
- La vraisemblance est de classe C^2 (la paramétrisation est bien régulière)
- Les vecteurs $(\partial_i p_j(\theta))_{1 \leq j \leq r}$ sont linéairement indépendants (on vérifie que le modèle est non redondant)

4.1.2 Echantillon d'un modèle régulier

Dans cette partie, on s'intéresse à la convergence du maximum de vraisemblance pour un n-échantillon, n tendant vers l'infini.

Considérons un modèle régulier en θ, $(E, \mathcal{E}, (F_\alpha)_{\alpha \in \Theta})$ dominé par F $(F_\alpha)_{\alpha \in \Theta}$ sa vraisemblance et $I(\theta)$ son information de Fisher en θ. On observe un échantillon canonique de ce modèle régulier ; on note donc $(\Omega,\ \mathcal{A},\ P_\theta) = (E,\ \mathcal{E},\ F_\theta)^N$ et X_p la p^{eme} coordonnée. Soit

$$l_n(\theta) = \log L_n(\theta) = \sum_{p=1}^{n} \log f(\theta, X_p)$$

Soit $Y_n^i = \partial_i l_n(\theta) = \sum_{p=1}^n \frac{\partial_i f(\theta, X_p)}{f(\theta, X_p)}$, $Y_n = (Y_n^i)_{1 \leq i \leq k} = grad\ l_n(\theta)$

Proposition 4.1 *Dans le cadre des hypothèses précédentes, on a :*

$$\frac{1}{\sqrt{n}} Y_n(\theta) \xrightarrow[n \to \infty]{\mathcal{L}} \mathcal{N}_k(0, I(\theta))$$

Preuve: Les vecteurs aléatoires $K_p = (\frac{\partial_i f(\theta, X_p)}{f(\theta, X_p)})_{1 \leq i \leq k}$ sont pour $p \geq 1$ centrés indépendants : Si on peut échanger dérivation et intégration, on obtient, en dérivant la fonction $\theta \to \mathbb{E}_\theta(1)$ identique à 1 : $\mathbb{E}_\theta(\frac{\partial_i f(\theta, X_p)}{f(\theta, X_p)})$=0.
De plus, la matrice de covariance de K_p est $I(\theta)$. Le théorème central limite vectoriel donne le résultat.

■

Proposition 4.2 *Soit un modèle $(E, \mathcal{E}, (F_\alpha)_{\alpha\in\Theta})$ régulier en θ, d'information de Fisher $I(\theta)$ et de vraisemblance f. On suppose qu'il existe un voisinage V de θ et une v.a. h sur $(E,\mathcal{E})$, F_θ-intégrable, et telle que, pour tout $x \in E$, $\alpha \in V, 1 \leq i,j \leq k$:*

$$|\partial_i \partial_j \log f(\alpha, x)| \leq h(x).$$

On observe alors un échantillon de ce modèle régulier. Si $\widehat{\theta_n}$ est un estimateur du maximum de vraisemblance qui converge en P_θ-probabilité vers θ, on a :

$$\sqrt{n}(\widehat{\theta_n} - \theta_n) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, I^{-1}(\theta))$$

$$I(\theta)\sqrt{n}(\widehat{\theta_n} - \theta_n) - \frac{1}{\sqrt{n}} grad\ l_n(\theta) \xrightarrow[n\to\infty]{(P)} 0$$

Pour la démonstration, on se référera à l'ouvrage de Dacunha-castelle.

Remarque : Dans le cadre de notre étude où la variable aléatoire ne prend qu'un nombre fini de valeurs ($E < \infty$), les hypothèses suivantes suffisent :
- le modèle est régulier en θ
- $\widehat{\theta_n}$ est consistant

Preuve: Comme le modèle est régulier en θ, $\partial_i\partial_j \log f(\alpha, x)$ est définie et continue sur un voisinage de θ, donc bornée sur un voisinage par $h(x)$. $h(x)$ est f_θ-intégrable car E est fini.

■

4.2 Test du χ^2 généralisé

Grâce à ces théorèmes de convergence du maximum de vraisemblance, on va pouvoir construire des tests plus élaborés sur le principe du test simple.

Théorème 4.3 *Soit $\Theta \subset \mathbb{R}^r$ où Θ est un ouvert et $\theta \to p(\theta) = (p_j(\theta))_{1\leq j\leq K}$ une fonction de Θ dans l'ensemble des probabilités sur $\{1, 2, ..., K\}$. On suppose que le modèle est régulier pour tout $\theta \in \Theta$, et que les $p_j(\theta)$ sont tous non nuls. alors la matrice d'information de Fisher est $I(\theta) = B(\theta)B(\theta)^T$, où $B(\theta)$ est la matrice $r \times K$ suivante :*

$$B(\theta) = \left(\frac{1}{\sqrt{p_j(\theta)}}\partial_i p_j(\theta)\right)_{1\leq i\leq r, 1\leq j\leq K}$$

Si de plus $\widehat{\theta_n}$, estimateur du maximum de vraisemblance est consistant pour tout θ, on obtient que sous H_0, $P \in \Theta$:

$$n\chi^2\left(\frac{N^n}{n}, p(\widehat{\theta_n})\right) \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2(K - 1 - r)$$

Preuve: Soit $\theta \in \Theta$
Notons $Z_n(\theta) = (\frac{N_i^n - np^i(\theta)}{\sqrt{np_i(\theta)}})_{1\leq i\leq K}$.

Le principe sera le même que précédemment : On va d'abord démontrer que $Z_n(\widehat{\theta_n})$ converge en loi vers une normale centrée de matrice de covariance A. Puis on montre que A est un projecteur orthogonal et on termine en appliquant le lemme 2.2 On calcule l'information de Fisher :

$$I^{ij}(\theta) = \sum_{u=1}^{K} p_u\theta \frac{\partial_i p_u(\theta)\partial_j p_u(\theta)}{[p_u(\theta)]^2};\ I(\theta) = B(\theta)B(\theta)^T$$

Ici on a $l_n(\theta) = \sum_{j=1}^{K} N_j^n \log p_j(\theta)$ donc

$$\frac{1}{\sqrt{n}} grad\ l_n(\theta) = \left(\frac{1}{\sqrt{n}} \sum_{u=1}^{K} N_u^n \frac{\partial_i p_u(\theta)}{p_u(\theta)} \right)_{1\leq i\leq k}$$

Remarquant que $B_\theta \sqrt{p_\theta}$ est nul, on a :

$$\frac{1}{\sqrt{n}} grad\ l_n(\theta) = B(\theta) Z_n(\theta)$$

De plus par le théorème central limite, $Z_n(\theta) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}_r(0, \Gamma(\theta))$
avec $\Gamma(\theta) = I_r - \sqrt{p(\theta)}\sqrt{p(\theta)}^T$.
Définissons $g : \mathbb{R}^K \to \mathbb{R}^{K+r}\ x \mapsto (x, I^{-1}(\theta)B(\theta)x)$. g est bien de classe C^2 à dérivées secondes bornées. On peut donc appliquer le théorème de changement de variable (3.4). $g(Z_n(\theta)) = (Z_n(\theta), I^{-1}(\theta)\frac{1}{\sqrt{n}} grad\ l_n(\theta))$ donc :

$$(Z_n(\theta), I^{-1}(\theta)\frac{1}{\sqrt{n}} gradl_n(\theta)) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}_{K+r}(0, C(\theta))$$

Comme on est bien dans les hypothèses du théorème 4.2 : (modèle régulier et estimateur consistant)

$$I(\theta)^{-1}\frac{1}{\sqrt{n}} gradl_n(\theta) - \sqrt{n}(\widehat{\theta_n} - \theta) \xrightarrow[n\to\infty]{P} 0$$

Par le lemme 3.2 de Slustky, on obtient :

$$(Z_n(\theta), \sqrt{n}(\widehat{\theta_n} - \theta_n)) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}_{K+r}(0, C(\theta)) \tag{4.1}$$

On calcule maintenant $C(\theta)$:

$$C(\theta) = \left[\begin{array}{c} I_K \\ \hline I^{-1}(\theta)B(\theta) \end{array}\right] \Gamma(\theta) \left[\begin{array}{c|c} I_K & B(\theta)^T I^{-1}(\theta) \end{array}\right] = \left[\begin{array}{c|c} \Gamma(\theta) & B(\theta)^T I^{-1}(\theta) \\ \hline I^{-1}(\theta)B(\theta) & I^{-1}(\theta) \end{array}\right]$$

On cherche maintenant à étudier :

$$\left(\frac{N_i^n - np_i(\widehat{\theta_n})}{\sqrt{np_i(\theta)}} \right)_{1\leq i\leq K} = Z_n(\theta) + \sqrt{n} \left(\sqrt{p_i(\theta)} - \frac{p_i(\widehat{\theta_n})}{\sqrt{p_i(\theta)}} \right)_{1\leq i\leq K}$$

Soit $h : (x_1, ...x_K, x_{K+1}...x_{K+r}) \mapsto \left(x_i - \frac{p_i((x_{K+1}...x_{K+r})+\theta)}{\sqrt{p_i(\theta)}} \right)_{1 \leq i \leq K}$
h est bien de la classe C^2 et à dérivées secondes bornées au voisinage $(\sqrt{p_i(\theta)}, 0)$ car le modèle est régulier. On a $h(\frac{N_i^n}{n\sqrt{p_i(\theta)}}, \widehat{\theta_n} - \theta_n) = \left(\frac{N_i^n - np_i(\widehat{\theta_n})}{n\sqrt{p_i(\theta)}} \right)_{1 \leq i \leq K}$ et $h(\sqrt{p_i(\theta)}, 0) = 0$
La matrice jacobienne $D(\theta)$ de h en $(\sqrt{p_i(\theta)}, 0)$ est la matrice $K \times (K + r)$: $D(\theta) = \left[\begin{array}{c|c} I_K & -B(\theta)^T \end{array} \right]$

$$DCD^T = \left[\begin{array}{c|c} I_K & -B^T \end{array} \right] \left[\begin{array}{c|c} \Gamma & B^T I^{-1} \\ \hline I^{-1}B(\theta) & I^{-1} \end{array} \right] \left[\begin{array}{c} I_K \\ \hline -B \end{array} \right]$$

$$= \left[\begin{array}{c|c} \Gamma - B^T I^{-1} B & B^T I^{-1} - B^T I^{-1} \end{array} \right] \left[\begin{array}{c} I_r \\ \hline -B \end{array} \right] = \Gamma - B^T I^{-1} B$$

Donc en réappliquant le théorème de changement de variable 3.4 à (4.1) ,on obtient que

$$\left(\frac{N_i^n - np_i(\widehat{\theta_n})}{\sqrt{np_i(\theta)}} \right)_{1 \leq i \leq K} \xrightarrow[n \to \infty]{\mathcal{L}} \mathcal{N}_K(0, DCD^T)$$

Comme $\frac{p(\widehat{\theta_n})}{p(\theta)}$ tend vers 1 en P_θ probabilité, on obtient :

$$Z_n(\widehat{\theta_n}) \xrightarrow[n \to \infty]{\mathcal{L}(P_\theta)} \mathcal{N}_K(0, I_K - \sqrt{p(\theta)}\sqrt{p(\theta)}^T - B(\theta)I^{-1}(\theta)B(\theta))$$

Soit $E(\theta) = B(\theta)I^{-1}(\theta)B(\theta)$. En remarquant que $B^T BE = B^T B$ et sachant que le rang de $B^T B$ vaut r et celui de E est inférieur ou égal à r, on obtient que E est de rang r. De plus, $E(\theta)E(\theta) = E(\theta)$. C'est un projecteur symétrique donc positif. Cette matrice est donc diagonalisable dans une base orthonormée et r de ses valeurs propres valent 1, les $(K - r)$ autres valent 0 ; $\sqrt{p(\theta)}$ est dans le noyau de $E(\theta)$. Soit $(V_1...V_r)$ une base orthonormée du sous-espace propre pour 1 et $V_{r+1} = \sqrt{p(\theta)}$ le vecteur orthogonal à $V_1...V_k$.

Soit $M(\theta) = I_K - \sqrt{p(\theta)}\sqrt{p(\theta)}^T - B(\theta)I^{-1}(\theta)B(\theta)$; $(v_1....v_{r+1})$ est une base orthonormée du noyau du projecteur orthogonal M. D'après le lemme 2.2, on a $\|Z_n(\widehat{\theta_n})\|$ converge en loi vers la loi $\chi^2(K - r - 1)$.

■

Remarque : Comme précédemment, on peut remplacer l'hypothèse régulier en θ par les hypothèses ici équivalentes : les $p_j(\theta)$ sont de classes C^2, non nulles et $(\partial_i p_j(\theta))_{1 \leq j \leq K}$ sont linéairement indépendants.

Proposition 4.4 *Sous les hypothèses précédentes, si l'échantillon suit une loi P tel que $P \notin \bar{A}$ où $A = \{p(\theta) | \theta \in \theta\}$, on a :*

$$n\chi^2 \left(\frac{N^n}{n}, p(\widehat{\theta_n}) \right) \xrightarrow[n \to \infty]{p.s.} +\infty$$

Preuve:
Comme $P \notin A$, il existe $k > 0$ tel que $\forall \theta \in \Theta, \chi^2(P, p(\theta)) > k$
Par la loi forte des grands nombres, $\frac{N^n}{n} \xrightarrow[n \to \infty]{p.s.}$ P
On obtient $p.s$, $lim\ inf \chi^2(P, p(\widehat{\theta_n})) \geq k$ donc,

$$n\chi^2(\frac{N^n}{n}, p(\widehat{\theta_n})) \xrightarrow[n \to \infty]{p.s.} +\infty$$

■

Remarque : Il reste encore à étudier le comportement à la frontière de A. On n'a pas de résultat général, cependant dans la pratique, ce sont des cas pathologiques qu'on discrimine assez bien (par exemple : cas où l'une des probabilités est nulle).
Ces théorèmes et propositions permettent ainsi de créer un test statistique de la même forme que dans le chapitre 2.

4.3 Cohérence avec le test d'indépendance

On va maintenant redémontrer les résultats de la partie précédente.
Rappelons les notations : le couple (X, Y) est à valeurs dans $\{1..., I\} \times \{1...J\}$ et suit la loi $(P(X = i, Y = j)) = p_{ij}$. L'hypothèse nulle correspond à l'indépendance entre les variables X et Y. Elle se traduit de la façon suivante :

$$p \in \theta := \{p_{ij}, i = 1..I, j = 1..J; p_{ij} = p_{i+}p_{+j}\}$$

On peut paramètrer θ par $a_i\ ;\ i = 1, I\ ;\ b_j\ ;\ j = 1, J$ avec $\sum a_i = 1,\ \sum b_j = 1$ et $p_{ij} = a_i b_j$

On voit bien que $\Theta \subset \mathbb{R}^{I+J-2}$. On considère $\Theta_1 = \overset{\circ}{\Theta}$ qui correspond à l'hypothèse où aucune des probabilités n'est nulle.

L'estimateur du maximum de vraisemblance est toujours $\widehat{\theta} = (\frac{N_{i+}}{n},\ \frac{N_{+j}}{n})$. Cet estimateur converge p.s. vers θ.
Il ne reste plus qu'à montrer que le modèle est régulier :
$L(\theta)(\omega) = \prod_{i=1}^{I-1}(a_i^{\mathbf{1}_{\omega=i,.}}) \prod_{j=1}^{J-1}(b_j^{\mathbf{1}_{\omega=.,j}})(1 - \sum_{i=1}^{I-1} a_i)^{\mathbf{1}_{\omega=I,.}}(1 - \sum_{j=1}^{J-1} b_j)^{\mathbf{1}_{\omega=.,J}})$
L est donc bien deux fois différentiable pour tout θ
Reste à vérifier que $I(\theta)$ est inversible :

$$\frac{\partial \log L(\theta)}{\partial a_i} = \frac{\mathbf{1}_{\omega=i,.}}{a_i} - \frac{\mathbf{1}_{\omega=I,.}}{a_I},\ \frac{\partial \log L(\theta)}{\partial b_j} = \frac{\mathbf{1}_{\omega=.,j}}{b_j} - \frac{\mathbf{1}_{\omega=.,J}}{b_J}$$

Ici $I(\theta) = \left[\begin{array}{c|c} A & 0 \\ \hline 0 & B \end{array}\right]$
où $A = \frac{1}{a_I}[1] + Diag(\frac{1}{a_i})_{i=1..I-1}$ et $B = \frac{1}{b_J}[1] + Diag(\frac{1}{b_j})_{j=1..J-1}$ On obtient bien que $I(\theta)$ est inversible pour tout $\theta \in \Theta_1$ (H3).

En appliquant le théorème précédent, on obtient bien que $n\chi^2(\frac{N_n}{n}, \widehat{\theta_n})$ converge en loi vers $\chi^2(IJ - 1 - (I + J + 2)) = \chi^2((I - 1)(J - 1))$.

4.4 Modélisation du crossing-over

On réitére l'expérience de Mendel avec deux caractères (V,v) et (C,c) codés par le même chromosome. On sélectionne deux races pures (V,C) et (v,c) qu'on croise. Puis, on étudie le croisement de deuxième génération. La loi de transmission devrait être celle d'un unique caractère :

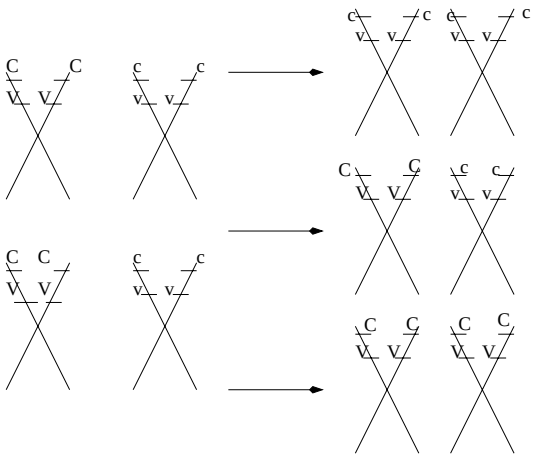

FIG. 4.1: *Hybridation*

Cependant, des expériences montrent l'obtention de phénotype (V,c) ou (v,C). De fait, un brassage intrachromosomique est assuré par les crossing-over. Il s'effectue lors de la première division de la méïose. Des échanges de segments de chromatides homologues peuvent alors intervenir. Les produits issus des chromatides ayant subi le crossing-over sont recombinés quant aux associatons de gènes : ils sont dits recombinés.Ceux n'ayant pas subis le crossing-over sont dits parentaux.

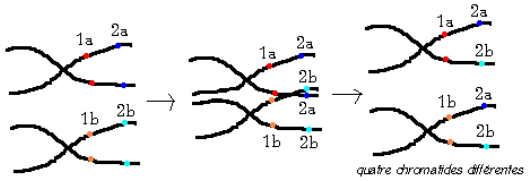

FIG. 4.2: *crossing-over*

On voit bien qu'on va obtenir deux chromatides parentaux et deux chromatides recombinés. Bien entendu, le crossing-over n'agira dans notre expérience que si il a lieu entre les deux gènes et que si les gènes en question sont hétérozygotes. De plus, s'il y a polusieurs crossing-overs, cela n'agira pour le génotype en question si il y a un nombre impair de crossing-over entre les deux gènes. Dans la suite, on appellera chiasma un tel événement.

L'hypothèse H_0 qu'on désire tester est que le crossing-over justifie à lui seul cet écart aux lois de Mendel, c'est à dire que les caractères suivent la loi suivante : Paramètrons H_o par η moitié de la probabilité d'avoir un chiasma pendant la

méïose, η est aussi la probabilité qu'un gamète ait un chromatide recombiné (pour c ou v).

Avec probabilité $(1-\eta)^2$ il n'y pas de chromatide recombiné :
on a (C,C)(V,V) :$\frac{1}{4}$, (C,c)(V,v) :$\frac{1}{2}$, (c,c)(v,v) :$\frac{1}{4}$
Avec probabilité $2\eta(1-\eta)$, il y a un chromatide sur un des deux parents :
(C,c)(V,V) :$\frac{1}{4}$, (C,C)(V,v) :$\frac{1}{4}$, (c,c)(v,V) :$\frac{1}{4}$, (c,C)(v,v) :$\frac{1}{4}$
Avec probabilité η^2, les chromatides sont recombinés chez les deux parents :
(c,c)(V,V) :$\frac{1}{4}$, (C,c)(V,v) :$\frac{1}{2}$, (C,C)(v,v) :$\frac{1}{4}$
Finalement, on obtient en terme de phénotype :
p_1=P(**C**,**V**)=$\frac{1}{4}\eta^2 - \frac{\eta}{2} + \frac{3}{4}$
p_2=P(**c**,**V**)=$\frac{1}{2}\eta - \frac{1}{4}\eta^2$
p_3=P(**C**,**v**)=$\frac{1}{2}\eta - \frac{1}{4}\eta^2$
p_4=P(**c**,**v**)=$\frac{1}{4}(1-\eta)^2$

On peut reparamètrer par $u = (1-\eta)^2$, on a alors

$$\Theta = \{(\frac{1}{4}u + \frac{1}{2}, -\frac{1}{4}u + \frac{1}{4}, -\frac{1}{4}u + \frac{1}{4}, \frac{1}{4}u), u \in]0,1[\}$$

L'estimateur du maximum de vraisemblance $\widehat{u}$ vérifie alors

$$\frac{N_1}{\widehat{u}+2} + \frac{N_2}{\widehat{u}-1} + \frac{N_3}{\widehat{u}-1} + \frac{N_4}{\widehat{u}} = 0$$

ce qui est équivalent à

$$\widehat{u}^2 n + \widehat{u}(-N_1 + 2(N_3+N_2) + N_4) - 2N_4 = 0$$

La première égalité montre qu'il existe un zéro entre 0 et 1 (théorème des valeurs intermédiaires) et la deuxième que l'autre zéro est négatif. $\widehat{u}$ est donc bien défini.

De plus, l'équation $\widehat{u}^2 + \widehat{u}(-p_1 + 2(p_3+p_2) + p_4) - 2p_4 = 0$ admet u comme solution. Donc comme $\frac{N_i^n}{n}$ converge p.s vers p_i et comme la fonction qui à (b,c) $(c<0)$ associe l'unique zéro strictement positif du polynôme (x^2+bx+c) est continue, $\widehat{u}$ converge p.s vers u.
Les p_i sont de classes C^2. calculons maintenant $I(u)$:

$$\log L_u(\omega) = \mathbf{1}_{\omega=1}\log(\frac{1}{4}u+\frac{1}{2}) + \mathbf{1}_{\omega=2}\log(-\frac{1}{4}u+\frac{1}{4}) + \mathbf{1}_{\omega=3}\log(-\frac{1}{4}u+\frac{1}{4}) + \mathbf{1}_{\omega=4}\log(\frac{1}{4}u)$$

$$\frac{\partial}{\partial u}\log L_u(\omega) = \frac{1}{u+2}\mathbf{1}_{\omega=1} + \frac{2}{u-1}\mathbf{1}_{\omega=2} + \frac{2}{u-1}\mathbf{1}_{\omega=3} + \frac{1}{u}\mathbf{1}_{\omega=4}$$

Calculons $I(u) = -\mathbb{E}(\frac{\partial^2}{\partial u^2}\log L_u) = \frac{1}{4}\left(\frac{1}{u+2} + \frac{2}{1-u} + \frac{1}{u}\right)$ qui est strictement positif pour $0 < u < 1$

On peut donc appliquer le théorème du χ^2 généralisé :

$$n\chi^2\left(p(\widehat{u}), \frac{N_i^n}{n}\right) \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2(2)$$

On peut donc définir un test de niveau asymptotique α et de puissance asymptotique 1 en prenant le fractile de la loi $\chi^2(2)$.

Fréquence des crossing-overs : Habituellement, un chromosome subit au moins un crossing-over. De façon générale, plus un chromosome est long et plus il subit de crossing-overs. Chaque type de chromosome dans une espèce donnée se caractérise par un nombre moyen de crossing-over. De même, on peut caractériser l'espace situé entre deux gènes d'un même chromosome (entre deux loci) par la fréquence des crossing-over qui s'y produisent. Plus cette distance est grande, plus la probabilité qu'un chiasma ait lieu entre ces points est grande.

Le nombre d'échantillon nécessaires pour ce test dépend fortement des gènes choisis. Si les deux gènes sont suffisament éloignés on pourra se contenter d'un vingtaine d'échantillon, par contre si ils sont très proches ce nombre peut exploser et tend vers l'infini.

Application : On teste la tansmission des caractères **H**(hairy) et **ST** (scarlet) chez la drosophile. On notera les gènes H et ST (qui sont dominants), les gènes récessifs associés seront notés +. Après 200 hybridations de deuxième génération, on trouve :
145 **(H,ST)**, 6 **(+,ST)**, 3 **(H,+)**, 46 **(+,+)**
on obtient $\widehat{u}$=0.9103 et $n\chi^2(\widehat{u}, \frac{N_i^n}{n}) = 1.01$.
Pour un test de niveau 0.05, le fractile est 5.05 donc l'hypothèse H_0 est acceptée. Cependant, d'autres hybridations montrent qu'il n'y a en fait des crossing-over que chez la femelle drosophile. Cela montre l'insuffisance de notre modèle qui aurait dû prendre deux paramètres η_1 et η_2.

Pour conclure, certes l'intervention de la statistique en biologie est importante pour valider des hypothèses, mais la modélisation et les expériences restent prépondérantes.

Bibliographie

- Toulouse, Paul S., *Thèmes de probabilités et statistique*
- Peter J. Bickel, Kjell A. Doksum, *Mathematical statistics*
- Didier Dacunha-Castelle, *Probabilités et statistiques. 2, Problèmes à temps mobile*