

EXPÉRIENCES DE MENDEL

Certaines expériences de Mendel pour étudier l'hérédité sont présentées ainsi que la théorie qu'il proposa pour les expliquer, pour laquelle nous mettons en œuvre des tests statistiques. Dans une première partie, des expériences portant sur l'hérédité d'un seul caractère font intervenir la loi binomiale. Dans une seconde partie, une expérience sur l'hérédité d'un couple de caractères amène au test du chi-deux pour la loi multinomiale.

Gregor Mendel découvrit les principaux mécanismes de l'hérédité en reproduisant des pois de jardin dans le cadre d'expériences soigneusement planifiées. À la suite de ces recherches, Mendel proposa une théorie particulière de l'hérédité. Dans la théorie de Mendel, les caractères sont déterminés par des unités discrètes qui se transmettent intactes au fil des générations. L'importance des idées de Mendel ne fut reconnue qu'aux environs de 1900, bien après sa mort. Le travail de Mendel est le prototype de l'analyse génétique. Il établit les règles d'une approche expérimentale et logique qui est toujours d'usage aujourd'hui.

1. L'APPROCHE EXPÉRIMENTALE DE MENDEL

Mendel étudia le pois de jardin pour deux raisons principales :

- il existe un vaste éventail de variétés de pois, de formes et de couleurs distinctes, facilement identifiables;
- les pois peuvent s'autoféconder ou être croisés, au choix de l'expérimentateur.

Mendel choisit d'étudier 7 caractères (ou traits) différents chez le pois : la forme de la graine dont le pois provient, la couleur de la graine, la couleur de la fleur, la forme de la gousse, la couleur de la gousse, la position des fleurs sur la plante, la longueur des tiges.

Pour chacun de ces caractères, il se procura deux variétés de pois différentes pour le caractère en question et les cultiva pendant deux ans pour s'assurer d'obtenir des *lignées pures*. Une lignée pure est une population dont les individus donnent des descendants identiques à eux-mêmes en ce qui concerne la forme du caractère considéré, appelée *phénotype*, que ce soit par autofécondation ou par croisement de deux individus de la lignée.

Par exemple, deux lignées obtenues par Mendel étaient pures pour la couleur de la fleur : une à fleurs pourpres, l'autre à fleurs blanches. Une plante de la

lignée à fleurs pourpres donnait des graines, après autofécondation ou croisement avec une plante de la même lignée, qui se développaient en des plantes à fleurs pourpres., elles-mêmes donnant des descendants à fleurs pourpres, et ainsi de suite. De même la lignée à fleurs blanches produisait uniquement des pois à fleurs blanches.

Au total Mendel obtint soit 7 paires de lignées correspondant à 7 paires de phénotypes :

Caractère	Phénotypes	
1) Forme de la graine	ronde	anguleuse
2) Couleur de la graine	jaune	verte
3) Couleur des fleurs	pourpre	blanche
4) Forme des gousses	arrondie	ridée
5) Couleur des gousses	verte	jaune
6) Position des fleurs	axiale	terminale
7) Longueur des tiges	longue	courte

Pour chaque caractère, Mendel a ensuite effectué une pollinisation croisée entre les deux variétés de pois de lignée pure; par exemple, en ce qui concerne le caractère couleur des fleurs, il provoqua la pollinisation entre des pois à fleurs pourpres et des pois à fleurs blanches. Ce croisement entre deux variétés est appelé *hybridation*. On nomme génération P (parentale) les parents de lignée pure, et on appelle génération F1 (première génération filiale) les hybrides qui en sont issus, qualifiés aussi de *monohybrides* du fait que leurs parents se différencient par un seul trait. En permettant l'autofécondation de ces hybrides F1, on obtient une génération F2 (deuxième génération filiale). C'est l'étude des plantes de la génération F2 qui a permis à Mendel de formuler le principe fondamental de l'hérédité aujourd'hui connu sous le nom de loi de ségrégation.

Afin d'étudier comment un couple de caractères présents chez une même plante se transmet à ses descendants, Mendel suivit une démarche similaire. Il obtint deux lignées pures relativement à deux caractères simultanément et qui présentaient deux phénotypes distincts pour chaque caractère, à savoir une lignée de pois dont la graine est arrondie et jaune, et une lignée de pois dont la graine est anguleuse et verte. Les hybrides obtenus par croisement de ces deux lignées sont nommées *dihybrides*. Par l'observation de la génération F2 des dihybrides, Mendel a pu formuler un deuxième principe : la loi d'assortiment indépendant des caractères.

2. LES EXPÉRIENCES DE CROISEMENT DES MONOHYBRIDES

2.1. Description des résultats. Tous les descendants issus du croisement d'une plante à fleurs pourpres avec une plante à fleurs blanches eurent des

fleurs pourpres. Aucun hybride ne présenta de couleur intermédiaire. Mendel qualifia le phénotype “fleurs pourpres” de *dominant* et le phénotype “fleurs blanches” de *récessif*.

Le même phénomène se produisit lors du croisement des deux lignées pures de chacune des 7 paires : les hybrides présentèrent tous le même phénotype, identique à celui de l’un des deux parents, le phénotype dominant par définition.

Ensuite, Mendel fit se reproduire les hybrides par autofécondation. Dans la génération F₂, deux phénotypes étaient présents : les phénotypes des deux lignées pures parentes. Le phénotype récessif était réapparu. Mendel compta alors le nombre de plantes correspondant à chaque phénotype :

Génération F₂ des monohybrides

Expérience	Effectifs			
	Phénotype dominant		Phénotype récessif	
1) Forme de la graine	ronde	5474	anguleuse	1850
2) Couleur de la graine	jaune	6022	verte	2001
3) Couleur des fleurs	pourpre	705	blanche	224
4) Forme des gousses	arrondie	882	ridée	299
5) Couleur des gousses	verte	428	jaune	152
6) Position des fleurs	axiale	651	terminale	207
7) Longueur des tiges	longue	787	courte	277

Le fait remarquable est que le rapport entre le nombre de plantes de phénotype dominant et le nombre de plantes de phénotype récessif est proche de 3 dans toutes les expériences.

De plus, des études supplémentaires montrèrent que les plantes de phénotype récessif étaient de lignée pures et que parmi les plantes de phénotype dominant il y avait en réalité deux groupes : un premier tiers correspondait à des plantes de lignée pure et les deux tiers restant à des plantes similaires aux hybrides F₁ c’est-à-dire dont les descendants portent les caractères dominants et récessifs dans le rapport 3 à 1.

2.2. La théorie de Mendel. Pour expliquer ces résultats, le botaniste proposa de modéliser la transmission d’un caractère donné comme suit :

- L’existence des gènes : il existe des déterminants de l’hérédité de nature particulière, les gènes. Un gène a deux formes possibles, ou allèles, chacune correspondant à un phénotype, soit **A** l’allèle du phénotype dominant et **a** l’allèle du phénotype récessif.

- Une plante adulte possède une paire (non ordonnée) de gènes, le *génotype* de la plante, qui peut donc être : **AA**, **Aa**, **aa**.

- Le phénotype d’une plante est déterminé par son génotype : le phénotype récessif est observé seulement si le génotype est **aa**, les autres génotypes

donnent le phénotype A .

- Le génotype d'une plante dépend de celui de ses parents de la manière suivante : chacune des deux gamètes (le pollen et l'ovule) intervenues lors de sa création, est porteuse d'un des deux allèles du parent dont elle provient, chaque allèle étant équiprobable, et ce indépendamment de l'autre gamète; le génotype du descendant est la réunion des deux allèles portés par les gamètes.

Le dernier point est nommée aujourd'hui loi de ségrégation (des gènes au niveau des gamètes) et a pour conséquence qu'une plante hérite d'un des deux allèles (avec équiprobabilité) de chacun de ses parents (qui peuvent être la même plante), les deux transmissions étant indépendantes. Le principe de ségrégation comporte aussi l'idée (implicite) que tous les phénomènes de ségrégation considérés sont indépendants les uns des autres, autrement dit toutes les transmissions de génotype ont lieu indépendamment les unes des autres, en particulier celles concernant les descendants d'une même plante.

Dans le modèle de Mendel, une plante de génotype AA (ou aa) aura, par autofécondation, des descendants de même génotype. Une plante de génotype Aa aura un descendant, par autofécondation, présentant un des trois génotypes possibles avec les probabilités suivantes :

Génotype	AA	aa	Aa
Probabilité	$1/4$	$1/4$	$1/2$

Les génotypes AA et aa sont donc ceux des lignées pures, et par suite les hybrides F_1 ont comme génotype Aa . Un individu de la génération F_2 présente donc un phénotype dominant A avec une probabilité de $3/4$ et un phénotype récessif a avec une probabilité $1/4$, ce qui justifie le rapport de 3 observé dans les expériences 1-7.

2.3. Analyse statistique de l'expérience 1. On présente un test de la théorie de Mendel basé uniquement sur les données de l'expérience 1. Bien sûr la démarche serait la même pour un test basé sur une des autres expériences.

2.3.1. *Le modèle statistique.* Sous l'hypothèse de Mendel, le nombre de plantes présentant un phénotype dominant (graine ronde) est une réalisation d'une variable aléatoire de loi binomiale $\mathcal{B}(n, p_0)$ avec $p_0 = 3/4$ (et $n = 5474 + 1850$). Sous l'hypothèse alternative, si on ne fait de suppositions supplémentaires, on peut juste affirmer que le nombre de plantes rondes sous forme de graine, est une réalisation d'une variable aléatoire à valeurs dans $\{1, \dots, n\}$. Formellement le modèle statistique associé est

$$\mathcal{E}_n = (\{1, \dots, n\}, \mathcal{P}(\{1, \dots, n\}), (P_{n,\theta})_{\theta \in \Theta})$$

où Θ est l'ensemble des probabilités sur $\{1, \dots, n\}$ muni de la tribu discrète et $\mathcal{P}_\Theta^n = \theta$, dans lequel on veut effectuer le test de

$$H_0 = \{\mathcal{B}(n, p_0)\} \quad \text{contre} \quad H_1 = \Theta \setminus H_0.$$

2.3.2. *Test pour une loi binomiale.* On s'intéresse à la construction d'un test de niveau α donné et pour ce faire on adopte une approche asymptotique : n est considéré comme grand . On note X l'identité de $\{1, \dots, n\}$,

$$Z_n = \sqrt{\frac{n}{p_0(1-p_0)}} \left(\frac{X}{n} - p_0 \right)$$

et Φ la fonction de répartition de la loi normale centrée réduite $\mathcal{N}(0, 1)$.

Proposition 1. *Pour tout $\alpha \in]0, 1]$, la suite des tests de région critique*

$$W_{n,\alpha} = \{|Z_n| > \Phi^{-1}(1 - \alpha/2)\}$$

est asymptotiquement (quand $n \rightarrow \infty$) de niveau α .

On peut appliquer ce résultat de la manière suivante : $n = 5474 + 1850$ est grand donc $1_{W_{n,\alpha}}$ est de niveau à peu près α . Au lieu de choisir une valeur de α , on donne simplement la valeur observée de $2(1 - \Phi(|Z_n|))$, appelée P-valeur, puisque l'on a

$$W_{n,\alpha} = \{2(1 - \Phi(|Z_n|)) < \alpha\}.$$

Ici on observe une P-valeur de 0,61. L'hypothèse H_0 n'est donc pas rejetée pour tout $\alpha < 0,61$. On en conclut que le résultat de l'expérience 1 n'est pas en contradiction avec la théorie de Mendel.

2.3.3. *Spécification de l'alternative.* Vu la forme de l'hypothèse alternative H_1 des modèles statistiques \mathcal{E}_n , la suite de tests de la proposition ?? n'est pas convergente et on a même que les fonctions puissances s'annulent en certains points de l'alternative, ce qui n'est pas satisfaisant.

Afin de palier à ce manque, on se propose d'introduire une hypothèse supplémentaire, vraie à priori, sur les phénotypes observés dans la génération F2:

On supposera par la suite que les phénotypes des individus de la génération F2 sont des réalisations de variables indépendantes et de même loi.

Cette hypothèse n'est complètement gratuite :

- L'hypothèse d'équidistribution est naturelle puisque les individus de la génération F2 ont été créés dans des conditions identiques : tout a été fait pour ces individus soient à priori indiscernables.

- L'hypothèse d'indépendance est plus forte : elle revient à accepter le fait que les transmissions des propriétés des parents à un descendant s'opèrent indépendamment les unes des autres, une hypothèse présente dans le modèle de Mendel. Elle est aussi justifiée par un souci de simplicité puisque il paraît difficile de concevoir et de préciser la forme d'une quelconque dépendance entre ces phénomènes.

2.4. Étude de la puissance dans le sous-modèle. En admettant le principe énoncé dans le paragraphe ??, on peut maintenant affirmer que le nombre de plantes présentant un phénotype dominant (graine ronde) est une réalisation d'une variable aléatoire de loi binomiale $\mathcal{B}(n, p)$ avec $p \in [0, 1]$ inconnu. Le modèle statistique correspondant s'écrit

$$\mathcal{E}'_n = (\{1, \dots, n\}, \mathcal{P}(\{1, \dots, n\}), (\mathcal{B}(n, p))_{p \in [0, 1]})$$

qui est un sous-modèle de \mathcal{E}_n après le changement de paramètres $p \mapsto \mathcal{B}(n, p)$. Pour l'hypothèse nulle $\{p_0\}$ contre l'alternative $[0, 1] \setminus \{p_0\}$, les tests $1_{W_{n, \alpha}}$ sont de niveau asymptotique α et la proposition suivante précise les propriétés de la suite des fonctions puissance

$$\eta_{n, \alpha}(p) = \mathcal{B}(n, p)(W_{n, \alpha}).$$

Proposition 2. *Pour tout $\alpha \in]0, 1]$,*

- (1) *la suite de tests $(1_{W_{n, \alpha}})$ est convergente i.e*

$$\forall p \in [0, 1] \setminus \{p_0\} \quad \eta_{n, \alpha}(p) \xrightarrow{n \rightarrow \infty} 1.$$

- (2)

$$\forall h \in \mathbb{R} \quad \eta_{n, \alpha}(p_0 + h/\sqrt{n}) \xrightarrow{n \rightarrow \infty} \varphi_\alpha\left(\frac{h}{\sqrt{p_0(1-p_0)}}\right)$$

où $\varphi_\alpha(x) = \mathbb{P}\left(|Z + x| > \Phi^{-1}(1 - \alpha/2)\right)$ et Z désigne une variable aléatoire gaussienne $\mathcal{N}(0, 1)$.

Une traduction concrète de cette proposition pour l'expérience 1 est que pour $\alpha = 5\%$ (par exemple) la puissance est plus grande que 0,975 (à une petite erreur près) dès que $|p - p_0| > 0,02$.

2.4.1. Preuve de la proposition ??. Le point (1) est une conséquence directe de la loi faible des grands nombres. La démonstration du point (2) exposée ici repose sur le théorème de Lindeberg qui est une extension du théorème central-limite et que nous admettrons.

Théorème 1 (de Lindeberg). *Soit $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)_{n \in \mathbb{N}}$ une famille d'espaces probabilisés; on suppose donnée, pour chaque $n \in \mathbb{N}$, une suite $(\xi_i^n)_{1 \leq i \leq n}$ une suite de vecteurs aléatoires d -dimensionnels définis sur $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$, indépendants,*

de carré intégrable et centrés. Notons \mathbf{K}_i^n la matrice de covariance de ξ_i^n . Sous les conditions

- (1) la suite $\sum_{i=1}^n \mathbf{K}_i^n$ converge vers une matrice \mathbf{K}
- (2)

$$\forall \varepsilon > 0 \quad \sum_{i=1}^n \mathbb{E}(\|\xi_i^n\|^2 \mathbf{1}_{\{\|\xi_i^n\| > \varepsilon\}}) \xrightarrow{n \rightarrow \infty} 0$$

les vecteurs $\sum_{i=1}^n \xi_i^n$ convergent en loi vers une loi gaussienne $\mathcal{N}(0, \mathbf{K})$.

À l'aide du théorème de Lindeberg on démontre le résultat d'approximation de la loi binomiale par la loi normale suivant :

Lemme 1. Si (p_n) est une suite de réels de $]0, 1[$ telle que $np_n(1-p_n) \rightarrow \infty$, les variables

$$\sqrt{\frac{n}{p_n(1-p_n)}} \left(\frac{X}{n} - p_n \right)$$

convergent en loi sous $\mathcal{B}(n, p_n)$ vers une loi gaussienne $\mathcal{N}(0, 1)$.

Preuve. Pour chaque n , considérons n variables Y_i^n , $1 \leq i \leq n$, i.i.d. de loi de Bernoulli de paramètre p_n . Puisque $\sum_{i=1}^n Y_i^n$ suit une loi $\mathcal{B}(n, p_n)$, il s'agit de montrer que les variables $\sum_{i=1}^n \xi_i^n$ convergent en loi vers $\mathcal{N}(0, 1)$ où

$$\xi_i^n = \frac{1}{\sqrt{np_n(1-p_n)}} (Y_i^n - p_n).$$

On est clairement dans le cadre d'application du théorème de Lindeberg, il suffit donc de prouver que

$$\sum_{i=1}^n \mathbb{E}((\xi_i^n)^2) \rightarrow 1,$$

$$\forall \varepsilon > 0 \quad \sum_{i=1}^n \mathbb{E}((\xi_i^n)^2 \mathbf{1}_{\{|\xi_i^n| > \varepsilon\}}) \rightarrow 0.$$

La première convergence est évidente car $\mathbb{E}((\xi_i^n)^2) = 1/n$. La seconde découle de la majoration

$$\mathbb{E}((\xi_i^n)^2 \mathbf{1}_{\{|\xi_i^n| > \varepsilon\}}) \leq \frac{1}{\varepsilon^2} \mathbb{E}((\xi_i^n)^4) \leq \frac{1}{\varepsilon^2 n^2 p_n(1-p_n)}.$$

□

On est maintenant en mesure de terminer la démonstration du point (2) de la proposition ???. Soit $h \in \mathbb{R}$. On a juste à montrer que les variables Z_n

convergent en loi sous $\mathcal{B}(n, p_0 + h/\sqrt{n})$ vers $\frac{h}{\sqrt{p_0(1-p_0)}} + Z$ où Z suit une loi $\mathcal{N}(0, 1)$. D'après le lemme ??, les variables

$$Z_n^h = \sqrt{\frac{n}{(p_0 + h/\sqrt{n})(1 - p_0 - h/\sqrt{n})}} \left(\frac{X}{n} - p_0 - h/\sqrt{n} \right)$$

convergent en loi vers Z . De plus

$$Z_n = \sqrt{\frac{(p_0 + h/\sqrt{n})(1 - p_0 - h/\sqrt{n})}{p_0(1 - p_0)}} Z_n^h + \frac{h}{\sqrt{p_0(1 - p_0)}}.$$

Le lemme Slutsky permet de conclure.

3. CROISEMENT DES DIHYBRIDES

3.1. Description des résultats. Les dihybrides de la génération F1 issus du croisement des deux lignées pures (une lignée à graine ronde et jaune, et une lignée à graine anguleuse et verte) étaient tous ronds et jaunes au stade de graine, ce qui est cohérent avec les expériences 1 et 2. Dans la génération F2, les deux phénotypes des deux caractères sont présents et Mendel dénombra les effectifs suivants

Génération F₂ des dihybrides

Graines	Rondes	Anguleuses	Total
Jaunes	315 (56,7%)	101 (18,2%)	416 (74,8%)
Vertes	108 (19,4%)	32 (5,8%)	140 (25,2%)
Total	423 (76,1%)	133 (23,9%)	556

On remarque immédiatement que les proportions marginales observées sont proches de 3/4 et 1/4.

3.2. L'hypothèse de Mendel. L'hypothèse que fit Mendel pour décrire l'hérédité d'un couple de caractères est simplement que l'hérédité de chaque caractère est gouvernée par les principes décrits dans la section ??, *les deux transmissions se déroulant indépendamment l'une de l'autre.*

3.3. Analyse statistique. Nous faisons dans la suite une hypothèse analogue à celle de ?? mais qui concerne non seulement chacun des deux caractères mais le couple de caractères chez les individus de F2.

3.3.1. Le modèle statistique. Les effectifs du tableau de ?? sont la réalisation d'un vecteur aléatoire $\mathbf{N} = (N_1, \dots, N_d)$, avec $d = 4$, de loi multinomiale de paramètres $n = 556$ et $\theta = (\theta_1, \dots, \theta_d) \in \Theta$, notée $P_{n,\theta}$, avec

$$\Theta = \left\{ \theta = (\theta_1, \dots, \theta_d) \in [0, 1]^d \mid \sum_{i=1}^d \theta_i = 1 \right\}.$$

Selon l'hypothèse de Mendel on a $\theta = \theta_0 = (9/16, 3/16, 3/16, 1/16)$. On choisit $H_0 = \{\theta_0\}$ comme hypothèse nulle.

3.3.2. *Test du chi-deux d'adéquation.* Le test du chi-deux d'adéquation est basé sur le comportement asymptotique des statistiques

$$D_n = n \sum_{i=1}^d \frac{1}{\theta_{0,i}} \left(\frac{N_i}{n} - \theta_{0,i} \right)^2$$

décrit par le théorème de Pearson. On désigne par $F_{\chi_d^2}$ la fonction de répartition de la loi du chi-deux à d degrés de liberté.

Théorème 2 (de Pearson). *Les variables D_n converge en loi sous $P_{\theta_0, n}$ vers une loi du chi-deux à $d - 1$ degrés de liberté.*

Par conséquent, pour tout $\alpha \in]0, 1]$, la suite des tests de région critique

$$\left\{ D_n > F_{\chi_{d-1}^2}^{-1}(1 - \alpha) \right\}$$

est asymptotiquement de niveau α .

On trouve ici comme réalisation de D_n la valeur $d_n \approx 0,47$ ce qui donne la P-valeur $1 - F_{\chi_{d-1}^2}(d_n) \approx 0,93$. Le modèle de Mendel est donc acceptable au vu de l'expérience sur les dihybrides.

3.3.3. *Puissance du test du chi-deux d'adéquation.* Tout d'abord on donne un résultat simple.

Proposition 3. *Pour tout $\alpha \in]0, 1]$, la suite de tests $\left(\mathbf{1}_{\left\{ D_n > F_{\chi_{d-1}^2}^{-1}(1-\alpha) \right\}} \right)$ est convergente.*

Pour décrire une propriété plus fine de la suite des fonctions puissance on a besoin de la définition suivante :

Proposition-Définition 1. *Soit Z un vecteur gaussien standard à valeurs dans \mathbb{R}^k et $\xi \in \mathbb{R}^k$. Si $\|\cdot\|$ désigne la norme euclidienne de \mathbb{R}^k , la loi de $\|Z + \xi\|^2$ ne dépend que de k et $\lambda = \|\xi\|^2$ et est appelée loi du chi-deux décentrée de paramètres k et λ .*

Le résultat suivant est une extension du théorème Pearson. On note

$$\eta_{n,\alpha}(\theta) = P_{n,\theta}(D_n > F_{\chi_{d-1}^2}^{-1}(1 - \alpha))$$

les fonctions puissance.

Proposition 4. *Soit $\mathbf{h} = (h_1, \dots, h_d) \in \mathbb{R}^d$ tel que $\sum_{i=1}^d h_i = 0$. Les variables D_n convergent en loi sous $P_{n,\theta_0+h/\sqrt{n}}$ vers une variable Y qui suit une loi du*

chi-deux décentrée de paramètres $d - 1$ et $\sum_{i=1}^d \frac{1}{\theta_{0,i}} h_i^2$.
Par conséquent

$$\eta_{n,\alpha}(\theta_0 + \mathbf{h}/\sqrt{n}) \rightarrow \mathbb{P}(Y > F_{\chi_{d-1}^2}^{-1}(1 - \alpha)).$$

3.3.4. *Estimation du niveau réel par la méthode de Monte-Carlo.* Il est légitime de vérifier si l'approche asymptotique est justifiée. Pour ce faire on peut par exemple estimer le niveau réel du test par la méthode de Monte-Carlo.

On propose un algorithme pour simuler un vecteur de loi multinomiale qui suppose que l'on sache déjà simuler une variable de loi binomiale.

Soit à simuler un vecteur $\mathbf{N} = (N_1, \dots, N_d)$ de loi multinomiale de paramètres n et (p_1, \dots, p_d) où $0 < p_i < 1$ pour tout i . Posons

$$q_i = \frac{p_i}{p_i + p_{i+1} + \dots + p_d}, \quad 1 \leq i < d.$$

L'algorithme suivant permet de simuler une réalisation de \mathbf{N} :

```

Initialiser le vecteur N de longueur d au vecteur nul
s = 0
i = 1
Tant que i < d et s < n faire
    Simuler une variable de loi  $\mathcal{B}(n - s, q_i)$  et
    affecter le résultat à N(i)
    s = s + N(i)
    i = i + 1
Fin de boucle
N(d) = n - s
Renvoyer le vecteur N

```

Par exemple, pour $\alpha = 5\%$, un intervalle de confiance pour le niveau réel est $0,0505 \pm 0,005$ au niveau de confiance $99,9\%$. On peut donc affirmer dans ce cas que le niveau réel du test du chi-deux est raisonnablement proche du niveau asymptotique.

Fin du texte

4. PROPOSITIONS DE DÉVELOPPEMENTS

- (1) Le candidat pourra donner une minoration non asymptotique de la puissance des tests de la proposition ???. Afin d'illustrer la proposition ??, il pourra faire une représentation graphique.
- (2) Le candidat pourra rappeler la définition d'une loi multinomiale et justifier pourquoi elle intervient dans le modèle statistique de la section 3.3.1.
- (3) Tout ou partie des démonstrations du théorème ?? et des propositions de la section ??
- (4) Le candidat pourra implémenter l'algorithme de la section ?? et obtenir des intervalles de confiance pour la puissance du test du chi-deux. Il pourra proposer un autre algorithme pour la simulation d'un vecteur de loi multinomiale.

4.1. Le modèle statistique - L'estimateur du maximum de vraisemblance.

$$\mathcal{E}_n = (\Omega_n, \mathcal{F}_n, (P_\theta^n)_{\theta \in \Theta})$$

$$\Omega_n = \left\{ (n_1, \dots, n_d) \in \mathbb{N}^d \mid \sum_{i=1}^d n_i = n \right\}$$

$$\Theta = \left\{ \theta = (\theta_1, \dots, \theta_d) \in [0, 1]^d \mid \sum_{i=1}^d \theta_i = 1 \right\}$$

Soit P_θ^n la loi multinomiale de paramètres \mathbf{n} et θ .

$$P_\theta^n(\{(n_1, \dots, n_d)\}) = \frac{n!}{n_1! \dots n_d!} \prod_{i=1}^d \theta_i^{n_i}$$

$$\lambda^n(\{(n_1, \dots, n_d)\}) = \frac{n!}{n_1! \dots n_d!}$$

Proposition 5. *La dérivée de Radon-Nikodym de P_θ^n par rapport à λ^n est donnée par*

$$L_n(\theta) = \prod_{i=1}^d \theta_i^{N_i}$$

L'estimateur du maximum de vraisemblance existe et est unique. Il est donné par

$$\hat{\theta}_n = \left(\frac{N_1}{n}, \dots, \frac{N_d}{n} \right)$$

4.2. Le test du χ^2 d'adéquation.

4.3. **Le test du χ^2 pour une hypothèse de dimension k .** H_0 est un sous-variété de \mathbb{R}^d de dimension $k < d - 1$.

$$H_0 \subset \text{Int}(\Theta)$$

Pour tout $\theta \in H_0$, il existe un voisinage V de θ dans \mathbb{R}^d , un ouvert U de \mathbb{R}^k et une application $\varphi : U \rightarrow V$ continûment différentiable telle que

- φ est un homéomorphisme entre U et $V \cap H_0$,
- la différentielle de φ en tout point de U est de rang k .

Puisque L_n est continue sur Θ (en tout point $\omega \in \Omega$), L admet un maximum sur l'adhérence de \bar{H}_0 de H_0 . Il s'ensuit qu'il existe au moins estimateur $\hat{\tau}_n$ à valeurs dans \bar{H}_0 tel que $L_n(\hat{\tau}_n) \geq L_n(\xi)$ pour tout $\xi \in \bar{H}_0$.

Théorème 3. *Dans le sous-modèle $(P_\theta)_{\theta \in H_0}$ il existe un estimateur du maximum de vraisemblance $\hat{\tau}^n$. Pour tout $\theta \in H_0$, les variables aléatoires*

$$D_n = n \sum_{i=1}^d \frac{1}{\hat{\tau}_{n,i}} (\hat{\theta}_{n,i} - \hat{\tau}_{n,i})^2$$

convergent en loi sous P_θ^n vers la loi du χ^2 à $d - 1 - k$ degrés de liberté.

Corollaire 1. *Pour tout $\alpha > 0$, la suite des tests de régions critiques*

$$W_n = \{\hat{\tau}_n \in \text{Int}(\Theta)\} \cap \{D_n > c(d - k - 1, 1 - \alpha)\}$$

est de niveau asymptotique α .

Lemme 2.

$$\forall \xi \in \text{Int}(\Theta) \quad \Lambda_n(\xi) - \Lambda_n(\hat{\theta}_n) = - \sum_{i=1}^d N_i \int_0^1 ds \int_0^s dr \frac{(\xi_i - \hat{\theta}_{n,i})^2}{(r\xi_i + (1-r)\hat{\theta}_{n,i})^2}.$$

En déduire que pour tous $\theta \in \text{Int}(\Theta)$, $K \geq 0$:

$$\sup_{\|\xi - \theta\| \leq K/\sqrt{n}} \left| \Lambda_n(\xi) - \Lambda_n(\hat{\theta}_n) + \frac{1}{2} \sum_{i=1}^d \frac{1}{\theta_i} (\xi_i - \hat{\theta}_i^n)^2 \right| \xrightarrow[n \rightarrow \infty]{P_\theta^n} 0.$$

$$\forall \xi \in \Theta \quad \Lambda_n(\xi) - \Lambda_n(\hat{\theta}_n) \leq -\frac{1}{2} \sum_{i=1}^d N_i (\xi_i - \hat{\theta}_i^n)^2$$

En déduire que pour tout $\theta \in \Theta_0$ les vecteurs aléatoires $\sqrt{n}(\hat{\theta}^{0,n} - \hat{\theta}_n)$ sont tendues sous P_θ^n i.e :

$$\sup_n P_\theta^n (\|\sqrt{n}(\hat{\theta}^{0,n} - \hat{\theta}_n)\| \geq K) \xrightarrow[K \rightarrow \infty]{} 0.$$

Pour $x \in \mathbb{R}^d$, notons

- $\pi(x)$ la projection orthogonale de x sur l'espace affine $\theta + \text{Im}(D_0\varphi)$, relativement au produit scalaire canonique de \mathbb{R}^d ,
- $\tilde{\pi}(x)$ la projection orthogonale de x sur l'espace affine $\theta + \text{Im}(D_0\varphi)$, relativement au produit scalaire

$$(x, y) \mapsto \sum_{i=1}^d \frac{x_i y_i}{\theta_i}.$$

Lemme 3.

$$\sqrt{n}(\hat{\tau}_n - \pi(\hat{\tau}_n)) \xrightarrow[n \rightarrow \infty]{P_\theta^n} 0.$$

Lemme 4.

$$\sqrt{n}(\hat{\tau}_n - \tilde{\pi}(\hat{\theta}_n)) \xrightarrow[n \rightarrow \infty]{P_\theta^n} 0.$$

4.4. Le test du χ^2 d'indépendance.