

Chapitre 3

Tests du χ^2

Les tests du χ^2 font partie des tests les plus utilisés en statistique. La raison en est qu'ils peuvent s'appliquer dans beaucoup de situations (paramétriques, ou **non paramétriques**). La plupart de ces tests peuvent être utilisés pour vérifier qu'un modèle choisi est en adéquation avec la réalité, et peuvent donc aider le statisticien dans le choix de son modèle.

3.1 Test du χ^2 d'ajustement.

La problématique des tests d'ajustement est la suivante. Étant donné un n -échantillon d'une loi μ **inconnue** sur un espace E (connu), on désire vérifier si $\mu = \mu_0$, où μ_0 est une **probabilité** sur E **donnée**. Par exemple, on veut savoir si le nombre quotidien d'accidents de voitures dans une ville suit une loi de Poisson de paramètre 10. Dans cet exemple, $E = \mathbb{N}$, et μ_0 est la loi $\mathcal{P}(10)$. Dans cette situation, le paramètre du modèle est l'ensemble des probabilités sur E .

De façon plus générale, on veut savoir si μ appartient à un certain ensemble de lois, ou pas. Par exemple, on veut savoir si le nombre quotidien d'accidents de voitures dans une ville suit une loi de Poisson.

Le test du χ^2 d'ajustement permet d'aborder cette question. Sachez toutefois qu'il existe d'autres tests d'ajustement, comme le test de Kolmogorov, dont nous ne parlerons pas ici.

3.1.1 Test du χ^2 d'ajustement à une loi sur un espace fini.

C'est la version la plus simple (et paramétrique) du test du χ^2 . L'observation est une réalisation d'un n -échantillon (X_1, \dots, X_n) de la loi $p = \sum_{j=1}^d p_j \delta_{\xi_j}$, où les ξ_j sont connus, et les p_j sont les paramètres du modèle. Autrement dit, on observe n variables indépendantes qui peuvent prendre un nombre fini de valeurs connues (notées ξ_j), ξ_j ayant une probabilité p_j inconnue.

Soit $\pi = \sum_{j=1}^d \pi_j \delta_{\xi_j}$ (π_j connus). On veut tester (H_0) : “ $p = \pi$ ” contre (H_1) : “ $p \neq \pi$ ”.

Par exemple, dans le cas de l'exercice 1, on veut savoir si la couleur de certaines fleurs (les mufliers) est gérée par un couple d'allèles. Si tel est le cas, on devrait observer 3 couleurs pour ces mufliers : rouge en proportion 1/4, ivoire en proportion 1/4 et pâle en proportion 1/2. Dans ce cas, le nombre d'issues possibles est $d = 3$, les issues possibles sont $\{R, I, P\}$ pour Rouge, Ivoire et Pâle, les proportions de chacune des couleurs sont notes p_R, p_I, p_P et on veut tester (H_0) : “ $(p_R, p_I, p_P) = (1/4, 1/4, 1/2)$ ” contre (H_1) : “ $(p_R, p_I, p_P) \neq (1/4, 1/4, 1/2)$ ”.

Si n est grand, les fréquences empiriques N_j/n où $N_j = \sum_{i=1}^n \mathbb{1}_{X_i=\xi_j}$ ($j \in \{1, \dots, d\}$) est le nombre de fois où on rencontre ξ_j dans l'échantillon, sont proches de $P(X_1 = \xi_j) = p_j$ par la loi des grands nombres. Karl Pearson a donc introduit la statistique

$$T_n = \sum_{j=1}^d \frac{1}{n\pi_j} (N_j - n\pi_j)^2$$

qui devrait être proche de 0 dans l'hypothèse (H_0) .

L'idée est alors de construire un test dont la région de rejet serait du type $\{T_n \geq t\}$. A cette fin, on a besoin de connaître la loi de T_n sous (H_0) .

Théorème 3.1.1 (*admis*) On suppose que $\forall j \in \{1, \dots, d\}$, $\pi_j > 0$. Si les variables $(X_i)_{i=1, \dots, n}$ sont i.i.d. de loi π ,

$$T_n \xrightarrow[n \rightarrow \infty]{(loi)} Z, \quad \text{où } Z \sim \chi_{d-1}^2.$$

Région de rejet du test du χ^2 d'ajustement à π . Soit α fixé, et t_α défini par $P[X > t_\alpha] = \alpha$, où $X \sim \chi_{d-1}^2$. $\mathcal{R} = \left\{ \sum_{j=1}^d \frac{(N_j - n\pi_j)^2}{n\pi_j} > t_\alpha \right\}$ est une zone de rejet de (H_0) : “ $p = \pi$ ” contre (H_1) : “ $p \neq \pi$ ” de niveau α , si **n est assez grand**. En pratique, on impose $n \geq 30$, $n\pi_j \geq 5$, pour tout $j \in \{1, \dots, d\}$. de façon traditionnelle, les variables N_j sont appelés les **effectifs observés**, les nombres $n\pi_j$ sont appelés les **effectifs théoriques**.

Région de confiance pour π . Le théorème 3.1.1 peut s'interpréter en disant que la variable T_n est un pivot asymptotique pour l'estimation de la loi des

X_i . On en déduit donc des régions de confiance pour cette loi. Soit t_α défini comme précédemment. On a alors, si n est assez grand,

$$\forall \pi, P_\pi \left[\sum_{i=1}^d \frac{(N_i - n\pi(i))^2}{n\pi(i)} \leq t_\alpha \right] \simeq 1 - \alpha.$$

3.1.2 Test du χ^2 d'ajustement à une loi à densité.

Le test du χ^2 d'ajustement peut également être utilisé pour tester l'ajustement à une loi à densité Q sur \mathbb{R}^k , i.e. pour tester

$$\begin{aligned} (H_0) : & \quad (X_1, \dots, X_n) \text{ est un } n - \text{échantillon de loi } Q, \\ \text{contre } (H_1) : & \quad (X_1, \dots, X_n) \text{ n'est pas un } n - \text{échantillon de loi } Q. \end{aligned}$$

Q est ici une loi **connue** (par exemple, la loi $\mathcal{N}(0, 1)$). Pour cela, on commence par choisir un entier d et une application $\phi : \mathbb{R}^k \rightarrow \{1, \dots, d\}$, et on applique ensuite le test du χ^2 à l'échantillon $(Y_1, \dots, Y_n) = (\phi(X_1), \dots, \phi(X_n))$ qui est maintenant un échantillon de variables prenant d valeurs.

Se donner une telle application ϕ , revient à se donner une partition V_1, \dots, V_d de \mathbb{R}^k et à poser $\phi(x) = j, \forall x \in V_j$. Sous (H_0) , on a alors

$$\begin{aligned} \pi_j &= P(Y_1 = j) = P(X_1 \in V_j) = Q(V_j), \\ N_j &= \sum_{i=1}^n \mathbb{1}_{Y_i=j} = \sum_{i=1}^n \mathbb{1}_{X_i \in V_j} = \text{nombre de } X_i \text{ qui tombent dans } V_j. \end{aligned}$$

Au niveau α , la région de rejet du test du χ^2 d'ajustement à la loi Q est alors

$$\mathcal{R} = \left\{ \sum_{i=1}^d \frac{(N_j - nQ(V_j))^2}{nQ(V_j)} \geq t_\alpha \right\}, \text{ où } t_\alpha \text{ est tel que } P[Z \geq t_\alpha] = \alpha, Z \sim \chi_{d-1}^2.$$

Notez que ce test dépend de la partition choisie. Il y a donc autant de tests du χ^2 dans cette situation, que de choix de partitions. La seule contrainte sur le choix de la partition est que $\forall j, nQ(V_j) \geq 5$. Les conclusions du test peuvent être très différentes suivant le choix de la partition. Par exemple, nous présentons dans la figure 3.1, les Pvaleurs des tests du χ^2 d'ajustement à la loi $\mathcal{N}(0, 1)$ pour différentes partitions de \mathbb{R} (au nombre de 1000), pour **une même réalisation** $(X_1(\omega), \dots, X_{1000}(\omega))$ (générée avec le générateur de nombres aléatoires $\mathcal{N}(0, 1)$ de matlab).

3.1.3 Test du χ^2 d'ajustement à une famille de loi.

Le test du χ^2 peut être généralisé au cas où on se permet d'estimer un certain nombre de paramètres de la loi. Commençons par présenter la situation dans le cas où les observations ne peuvent prendre qu'un nombre fini de valeurs connues. Dans ce cas, on suppose que l'observation (X_1, \dots, X_n) est un

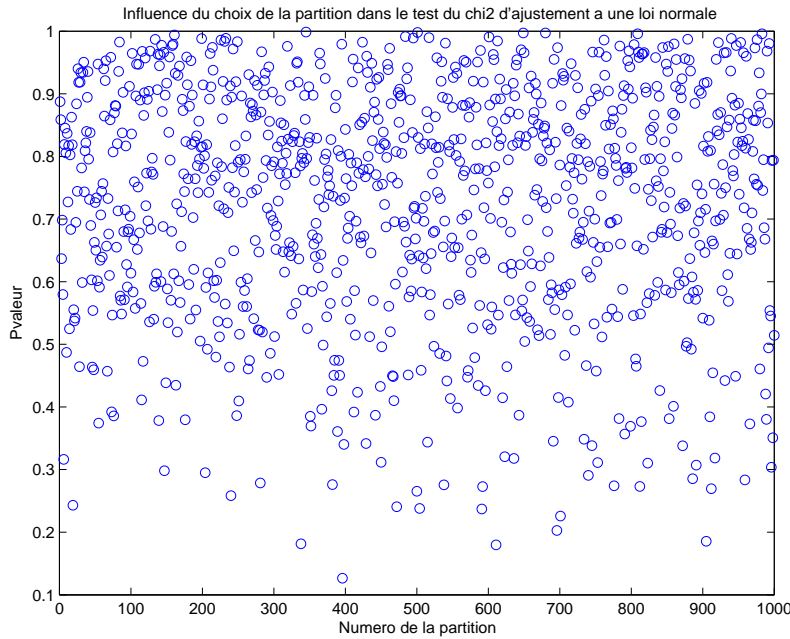


FIGURE 3.1 – Évolution de la Pvaleur en fonction de la partition.

n -échantillon de loi $\mu(\theta) = \sum_{j=1}^d p_j(\theta) \xi_j$, les ξ_j étant connus et $\theta \in \Theta \subset \mathbb{R}^k$ étant un paramètre inconnu. On veut alors tester

(H_0) : (X_1, \dots, X_n) est un n – échantillon d’une loi de la famille $\{\mu(\theta), \theta \in \Theta\}$
 contre (H_1) : La loi des X_i n’est pas dans la famille $\{\mu(\theta), \theta \in \Theta\}$

Pour cela, on introduit

$$T_n(\theta) = \sum_{j=1}^d \frac{1}{np_j(\theta)} (N_j - np_j(\theta))^2$$

$T_n(\theta)$ n’est pas une statistique puisqu’elle dépend du paramètre θ . On estime alors θ par l’estimateur du maximum de vraisemblance $\hat{\theta}_n$ et on utilise $T_n(\hat{\theta}_n)$ comme statistique de test. Cela revient à tester l’adéquation à la loi la plus vraisemblable dans la famille $\{\mu(\theta), \theta \in \Theta\}$. On choisit comme région de rejet du test

$$R(X) = \left\{ T_n(\hat{\theta}_n) \geq t \right\}$$

Pour déterminer t , on a besoin de connaître la loi de $T_n(\hat{\theta}_n)$ sous (H_0) . On admettra le résultat suivant :

Théorème 3.1.2 *Supposons que*

- Θ est un ouvert de \mathbb{R}^k ;
- $\forall j \in \{1, \dots, d\}$, $p_j : \Theta \rightarrow [0; 1]$ est de classe \mathcal{C}^2 ;

- $\forall \theta \in \Theta$, $\left(\frac{\partial p_j}{\partial \theta_l}\right)_{\substack{j=1 \dots d \\ l=1 \dots k}}$ est de rang k (ce qui impose $k \leq d$);
- l'estimateur du maximum de vraisemblance $\hat{\theta}_n$ de $\theta \in \Theta$ existe pour tout n , et vérifie pour tout $j \in \{1, \dots, d\}$, $\forall n$, $p_j(\hat{\theta}_n) > 0$.

Alors, sous (H_0) , $T_n(\hat{\theta}_n) \xrightarrow{(loi)} Z$, où $Z \sim \chi_{d-k-1}^2$.

La formule presque incantatoire à retenir, est que le degré du χ^2 est égal au

nombre de classes - 1 - nombre de paramètres estimés ($d - 1 - k$).

Ainsi, la région de rejet du test du χ^2 de niveau α de

(H_0) : (X_1, \dots, X_n) est un échantillon d'une loi de la famille $\{\mu(\theta), \theta \in \Theta\}$
contre (H_1) : **La loi des X_i n'est pas dans la famille $\{\mu(\theta), \theta \in \Theta\}$**

est donnée par

$$\mathcal{R} = \left\{ T_n(\hat{\theta}_n) \geq t_\alpha \right\}, \text{ où } t_\alpha \text{ est solution de } P[Z \geq t_\alpha] = \alpha, Z \sim \chi_{d-k-1}^2.$$

Là encore, on peut généraliser à des lois qui ne sont plus nécessairement à support fini, en projetant les données sur une partition.

Ce test peut s'utiliser pour tester l'ajustement des données à une famille de lois à densité. Par exemple, le chimiste peut se demander si la série de dosages $(X_1(\omega), \dots, X_n(\omega))$ qu'il a obtenue, est bien la réalisation d'un échantillon gaussien. Il va donc tester

$$(H_0) : \exists (m, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+ \text{ tel que } X_i \sim \mathcal{N}(m, \sigma^2) \\ \text{contre } (H_1) : \forall (m, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+, X_i \text{ ne suit pas } \mathcal{N}(m, \sigma^2)$$

On a déjà vu que dans ce contexte, l'estimateur du maximum de vraisemblance de (m, σ^2) est $(\bar{X}_n, \tilde{\sigma}_n^2)$, où $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ et $\tilde{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$. Pour effectuer son test, le chimiste va commencer par choisir une partition V_1, \dots, V_d de \mathbb{R} . Sous (H_0) ,

$$P_{(H_0)} [X_1 \in V_j] = \int_{V_j} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-m)^2}{2\sigma^2}\right) dy \triangleq p_j(m, \sigma^2).$$

Si on désigne par N_i le nombre de $(X_k; k = 1, \dots, n)$ qui tombent dans V_i ($N_i \triangleq \sum_{k=1}^n \mathbb{1}_{X_k=i}$), la statistique de test que le chimiste va utiliser est donc :

$$T_n = \sum_{j=1}^d \frac{(N_j - np_j(\bar{X}_n, \tilde{\sigma}_n^2))^2}{np_j(\bar{X}_n, \tilde{\sigma}_n^2)}.$$

Sous (H_0) , T_n tend en loi vers une variable du χ_{d-2-1}^2 quand la taille n de l'échantillon est suffisamment grande. La région de rejet d'un test de niveau α est donc $\mathcal{R} = \{T_n \geq t_\alpha\}$, où t_α est solution de $P[Z \geq t_\alpha] = \alpha$, avec $Z \sim \chi_{d-3}^2$.

3.2 Test du χ^2 d'indépendance.

Le théorème 3.1.2 peut être utilisé pour tester l'indépendance entre deux échantillons de données. On est ici dans une situation où l'observation est constituée de deux n -échantillons (X_1, \dots, X_n) et (Y_1, \dots, Y_n) , et il s'agit de savoir si ces deux échantillons sont indépendants ou pas. Il revient au même de dire que l'observation est constituée d'un n -échantillon d'un couple de variables (X, Y) à valeurs dans $E \times F$. On veut tester (H_0) : “ X et Y sont indépendantes ” contre (H_1) : “ X et Y ne sont pas indépendantes ”. Pour cela, on choisit des partitions de E et F , $E = \cup_{i=1}^k E_i$, $F = \cup_{j=1}^m F_j$, et on désigne par N_{ij} le nombre de points (X_l, Y_l) qui tombent dans $E_i \times F_j$:

$$N_{ij} = \sum_{l=1}^n \mathbb{1}_{X_l \in E_i} \mathbb{1}_{Y_l \in F_j}.$$

Si on connaissait les lois marginales du couple (X, Y) , i.e la loi de X et celle de Y , on connaîtrait $p_i = P[X \in E_i]$ et $q_j = P[Y \in F_j]$. Sous l'hypothèse (H_0) d'indépendance, on aurait alors $\frac{N_{ij}}{n} \sim P(X_1 \in E_i; Y_1 \in F_j) = p_i q_j$, et on pourrait utiliser comme statistique de test

$$T = \sum_{i=1}^k \sum_{j=1}^m \frac{(N_{ij} - np_i q_j)^2}{np_i q_j}.$$

En effet, sous (H_0) , T tend en loi vers un χ_{km-1}^2 . Ici, on ne connaît pas les lois de X et de Y . On les estime donc par leurs estimateurs du maximum de vraisemblance sous (H_0) , i.e on estime p_i par $\hat{p}_i \triangleq \frac{N_{i\bullet}}{n} = \frac{1}{n} \sum_{j=1}^m N_{ij}$, et q_j par

$\hat{q}_j \triangleq \frac{N_{\bullet j}}{n} = \frac{1}{n} \sum_{i=1}^k N_{ij}$. On utilise alors la statistique

$$S = \sum_{i=1}^k \sum_{j=1}^m \frac{(N_{ij} - n\hat{p}_i \hat{q}_j)^2}{n\hat{p}_i \hat{q}_j}.$$

On a ainsi estimé $(k-1) + (m-1)$ paramètres (les “-1” viennent des deux relations $\sum_i p_i = \sum_j q_j = 1$). On en déduit que S tend en loi vers un χ^2 de degré $km - 1 - (k-1) - (m-1) = (k-1)(m-1)$. Au niveau α , la région de rejet du test est donc $\mathcal{R} = \{S \geq t_\alpha\}$, où t_α est déterminé par l'équation $P[Z \geq t_\alpha] = \alpha$, avec $Z \sim \chi_{(k-1)(m-1)}^2$.

3.3 Exercices.

Exercice 1. Partant de races pures, Bauer a croisé des mufliers ivoires avec des mufliers rouges. A la deuxième génération, après autofécondation des plantes de la première, il a obtenu : 22 mufliers rouges, 52 pâles et 23 ivoires. Si la couleur des fleurs est gérée par un couple d'allèles, la répartition théorique est : Rouge $\rightarrow 1/4$, Pâle $\rightarrow 1/2$, Ivoire $\rightarrow 1/4$.

Que conclure ?

Exercice 2. On se propose de tester si la variable X suit une loi normale $\mathcal{N}(2, 1)$, connaissant l'observation suivante d'un 21-échantillon :

0,3 0,7 0,9 1,2 1,4 1,4 1,5 1,5 1,6 1,9 2,0
2,1 2,1 2,3 2,5 2,6 2,7 3,0 3,8 3,9 4,0.

Exercice 3. On veut vérifier si le nombre quotidien d'accidents de voiture dans une ville suit une loi de Poisson de paramètre 1.

Nombre d'accidents par jour :	0	1	2	3	plus de 4
Nombre de jours observé :	35	40	17	6	2

Exercice 4. On dispose de quatre catégories A, B, C, D de thermomètres. On sait que la distribution des thermomètres construits par un fabricant est

A	B	C	D
0.87	0.09	0.03	0.01

Un nouveau lot de 1336 thermomètres va être utilisé. Sa répartition en catégories est

A	B	C	D
1188	91	47	10

Ce lot diffère-t-il des anciens ?

Exercice 5. Sur deux populations I et II , on étudie la répartition des 4 groupes sanguins. On observe

	O	A	B	AB
I	121	120	79	33
II	118	95	121	30

Cette répartition est-elle identique pour les deux groupes ?

Exercice 6. *Test du χ^2 d'indépendance.* On a classé 217 enfants d'après leurs performances dans des tests de langage (L) et d'équilibre physique (E). Tester l'hypothèse de l'indépendance des performances de langage et d'équilibre.

	L1	L2	L3
E1	45	26	12
E2	32	50	21
E3	4	10	17

Exercice 7. *Test du χ^2 de symétrie.* On étudie un n -échantillon d'un couple de caractères X et Y à valeurs dans $\{1, \dots, k\}$; on observe N_{ij} fois $(i, j) \in \{1, \dots, k\}^2$. On teste la symétrie de la loi de (X, Y) . Montrer qu'on peut utiliser la statistique

$$\sum_{1 \leq i < j \leq k} \frac{(N_{ij} - N_{ji})^2}{N_{ij} + N_{ji}}$$

et donner sa loi asymptotique ainsi qu'un test convenable de niveau voisin de α de

$$\begin{aligned} (H_0) : & \quad \text{“Le couple } (X, Y) \text{ est symétrique”} \\ \text{contre } (H_1) : & \quad \text{“Le couple } (X, Y) \text{ n'est pas symétrique.”} \end{aligned}$$

Application. Le degré de vision de chacun des deux yeux de 7477 femmes âgées de 30 à 40 ans a été classé en quatre groupes désignés par 1, 2, 3 et 4 par ordre croissant du pire au meilleur. Les abréviations G et D font référence respectivement aux yeux gauches et droits.

	G1	G2	G3	G4
D1	1520	266	124	66
D2	234	1512	432	78
D3	117	362	1772	205
D4	36	82	179	492

Faire le test de niveau $\alpha = 0,05$. Pour quels α accepte-t'on la symétrie ? Tester l'indépendance.

Exercice 8. Un groupe d'étudiants a observé le fonctionnement d'une roulette de casino pendant deux jours, soit au total 4000 essais. Le résultat de leurs observations est consigné dans le tableau suivant où sont indiqués les nombres de fois où sont sortis chacun des 38 numéros de la roulette :

Numéro	1	2	3	4	5	6	7	8	9	10	11
Rouge	121	89	123	112	130	130	113	118	95	88	113
Noir	107	92	89	119	117	131	94	106	108	91	97

Numéro	12	13	14	15	16	17	18	0	00
Rouge	125	107	90	97	93	98	105	109	111
Noir	100	102	81	92	108	97	102		

Cette roulette est-elle biaisée ?

