Tests du χ^2

L'une des fonctions des statistiques est de proposer, à partir d'observations d'un phénomène aléatoire (ou modélisé comme tel) une estimation de la loi de ce phénomène. C'est que nous avons fait en construisant des intervalles de confiance. Les statistiques servent aussi à prendre des décisions. Peut-on considérer qu'un médicament est plus efficace qu'un placebo ? Le nombre de consultations de Google par seconde suit-il une loi de Poisson ? Les gènes pilotant la couleur des yeux et celle des cheveux sont-ils sur les mêmes chromosomes ? Il y a deux points communs (au moins) à toutes ces questions : leurs réponses sont des oui-non et le phénomène sous-jacent est aléatoire. Les tests statistiques vont permettre d'apporter une réponse à des questions manichéennes en contrôlant l'aléa inhérent à la situation.

En statistique les deux éventualités sont appelées des hypothèses et sont notées H_0 (*hypothèse nulle*) et H_1 (*hypothèse alternative*). Souvent H_1 sera le contraire de H_0 ; dans tous les cas, le postulat est qu'une et une seule des deux hypothèses est vraie. Un test statistique est un algorithme qui conduit à accepter H_0 ou à rejeter H_0 à partir d'observations d'un phénomène aléatoire. Voici un exemple un peu naïf mais très instructif.

Exemple 1. On vient d'acheter un dé à six faces tout neuf sensé fournir des résultats distribués selon la loi uniforme sur $\{1, \ldots, 6\}$. On souhaite valider (ou invalider) cette affirmation du fabriquant à partir de l'observation de n lancers de dé. On est donc amené à poser les hypothèses :

$$H_0 \ : \ \text{le dé est équilibré} \quad \text{et} \quad H_1 \ : \ \text{le dé est pipé.}$$

Il y a deux éventualités pour la réalité et deux décisions possibles. Sur les quatre configurations, deux sont satisfaisantes (la prise de décision correspond à la réalité). On peut résumer la situation avec le tableau suivant :

DécisionRéalité	H_0 est vraie	H_1 est vraie
on accepte H_0	bonne décision	erreur de seconde espèce
on rejette H_0	erreur de première espèce	bonne décision

On distingue les deux erreurs car H_0 et H_1 ne jouent pas un rôle symétrique. L'hypothèse nulle représentera ce que l'on pense être la réalité. On souhaitera en premier lieu imposer la probabilité de commettre l'erreur de première espèce, c'est-à-dire de rejeter H_0 alors qu'elle est vraie. Cette probabilité est appelée niveau du test.

Remarque 2. On peut faire une analogie avec la justice qui pose comme principe la présomption d'innocence. On souhaite contrôler en priorité la probabilité d'envoyer un innocent en prison (erreur de première espèce) en négligeant pour l'instant celle de relâcher un coupable (erreur de seconde espèce). Dans l'exemple 1, on laisse le bénéfice du doute au constructeur de dés.

L'idée est de trouver une statistique (une fonction des observations) dont on connaît la loi si H_0 est vraie et qui ne se comporte pas de la même manière selon que H_0 ou H_1 est vraie. Les tests du χ^2 sont un exemple relativement simple de tests statistiques qui vont permettre de tester

1. l'adéquation à une loi de probabilité sur un ensemble fini : est-il raisonnable de penser que les résultats que j'observe sont des réalisations i.i.d. d'une loi $(p_1, \ldots, p_k)$ sur $\{1, \ldots, k\}$?
2. l'indépendance de deux caractères mesurés sur un même individu.
3. l'homogénéité de plusieurs échantillons : deux médicaments ont-ils le même effet (guérison, amélioration, état stationnaire) sur la population atteinte ?

1 Un peu de probabilités

Cette section retrace les grandes lignes du raisonnement qui permet de construire la statistique de test du χ^2. Elle peut être omise en première lecture ou par les non spécialistes à condition de retenir les notations, l'exercice 4, le théorème 8 et la proposition 9.

Soit $p = (p_1, \ldots, p_k)$ une loi de probabilité sur $\{1, \ldots, k\}$ et $X_1, \ldots, X_n$ un échantillon de loi p. On définit les variables aléatoires $(N_i(n))_{1\leq i\leq k}$ à valeurs dans $\{0, \ldots, n\}$ par $N_i(n) = \mathrm{Card}\{j = 1, \ldots, n,\ X_j = i\}$. On dit que le vecteur $N(n) = (N_1(n), \ldots, N_k(n))$ suit la loi multinomiale de paramètre (n, p).

Remarque 3. Attention n est un entier et p est un vecteur de probabilité. Pour Scilab, les paramètres de `grand` option `mul` sont n et un vecteur-colonne formé des $k-1$ premières coordonnées de p.

Exercice 4 (Loi multinomiale)**.** Soit $N(n)$ de loi multinomiale de paramètre (n, p).

1. Montrer que $\mathbb{P}(N_1(n) = n_1, \ldots, N_k(n) = n_k) = \begin{cases} \dfrac{n!}{n_1! \cdots n_k!} & \text{si } n_1 + \cdots + n_k = n \\ 0 & \text{sinon.} \end{cases}$
2. Quelle est la loi de $N_i(n)$ pour $i = 1, \ldots, k$?
3. Montrer que $\mathbb{E}(N_i(n)) = np_i$, $\mathbb{V}(N_i(n)) = np_i(1-p_i)$ et $\mathrm{cov}(N_i(n), N_j(n)) = -np_ip_j$ pour $i \neq j$.
4. Les variables aléatoires $N_1(n)$ et $N_2(n)$ sont-elles indépendantes ? Interpréter le signe de $\mathrm{cov}(N_1(n), N_2(n))$.
5. Montrer que $(N(n)/n)_n$ converge presque sûrement vers le vecteur p.

Proposition 5. *Soit $(X_n)_{n\in\mathbb{N}}$ des variables aléatoires i.i.d. de loi p. Avec les notations ci-dessus (en identifiant p à un vecteur colonne et en notant ${}^tp$ son transposé),*

$$\sqrt{n}\left(\frac{N(n)}{n} - p\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, K_p),$$

où $K_p = \Delta_p - p\,{}^tp$ en notant Δ_p est la matrice diagonale de diagonale p.

Idée de preuve. On applique le théorème limite central multidimensionnel aux variables aléatoires i.i.d. $(Y(n))_{n\in\mathbb{N}}$ à valeurs dans $\mathbb{R}^k$ définies par $Y_i(n) = \mathbf{1}_{\{X_n=i\}}$. Ces variables aléatoires suivent la loi multinomiale de paramètre $(1, p)$. La variable aléatoire $Y(1)$ admet p pour (vecteur-)espérance et K_p pour matrice de covariance. Remarquons de plus que $N(n) = Y(1) + \cdots + Y(n)$. □

Le résultat suivant, qui est à la base de l'idée des tests du χ^2, consiste à produire, à partir d'une variable aléatoire normale dans $\mathbb{R}^k$ de loi $\mathcal{N}(0, K_p)$ une variable aléatoire dont la loi ne dépend pas de K_p.

Proposition 6. *Si Z suit la loi $\mathcal{N}(0, K_p)$ alors la variable aléatoire $Z_1^2/p_1 + \cdots + Z_k^2/p_k$ suit une loi du χ^2 à $k-1$ degrés de liberté.*

Idée de preuve. En effet, si l'on note $U_i = Z_i/\sqrt{p_i}$ pour $i = 1, \ldots, k$ alors $U = (U_1, \ldots, U_k)$ suit la loi $\mathcal{N}(0, I - \sqrt{p}\,{}^t\!\sqrt{p})$ (où $\sqrt{p}$ désigne le vecteur(-colonne) des racines carrées des coefficients de p). Remarquons à présent que la matrice de covariance de U est la matrice de projection orthogonale sur l'hyperplan orthogonal au vecteur normé $\sqrt{p}$. Il existe donc O matrice orthogonale telle que $O(I - \sqrt{p}\,{}^t\!\sqrt{p})^tO$ soit égale à la matrice $\Delta_{(1,\ldots,1,0)}$ (avec la notation introduit dans la proposition 5). Le vecteur aléatoire OU suit alors la loi normale centrée de matrice de covariance

$$\mathbb{E}((OU)^t(OU)) = \mathbb{E}(OU^tU^tO) = O\mathbb{E}(U^tU)^tO = O\mathrm{cov}(U)^tO.$$

En d'autres termes, OU suit la loi $\mathcal{N}(0, \Delta_{(1,\ldots,1,0)})$, c'est-à-dire que les $k-1$ premières coordonnées de OU sont des variables aléatoires gaussiennes centrées réduites indépendantes tandis que la dernière est nulle. Ainsi la loi de la variable aléatoire

$$\frac{Z_1^2}{p_1} + \cdots + \frac{Z_k^2}{p_k} = \|U\|^2 = \|OU\|^2$$

est-elle la même que celle de la somme des carrées de $k-1$ variables aléatoires gaussiennes indépendantes centrées réduites, c'est-à-dire une loi du χ^2 à $k-1$ degrés de liberté. □

Lemme 7. *Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires définies sur $(\Omega, \mathcal{A}, \mathbb{P})$ à valeurs dans $\mathbb{R}^k$ qui converge en loi vers μ et f de $\mathbb{R}^k$ dans $\mathbb{R}^l$ continue. Alors $(f(X_n))_{n\in\mathbb{N}}$ converge en loi vers la mesure image de μ par f.*

Idée de preuve. Il suffit d'utiliser la caractérisation de la convergence en loi par la convergence des espérances des fonctions continues. □

Il ne reste plus qu'à rassembler les propositions 5 et 6 et le lemme 7 pour obtenir le résultat du jour.

Théorème 8. *Soit $(X_n)_{n\in\mathbb{N}^*}$ une suite de variables aléatoires i.i.d. de loi p alors*

$$D_n = n\sum_{i=1}^k \frac{(N_i(n)/n - p_i)^2}{p_i} = \sum_{i=1}^n \frac{(N_i(n) - np_i)^2}{np_i} \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2(k-1).$$

On a donc construit à partir de l'échantillon $X_1, \ldots, X_n$ une statistique D_n dont la loi ne dépend plus de p (au moins asymptotiquement). On est donc capable de choisir une région de $\mathbb{R}$ telle que D_n appartienne à cette région avec une probabilité (asymptotique) donnée.

Ajoutons à ce théorème de convergence, la proposition suivante. Elle n'est qu'une conséquence directe de la loi des grands nombres mais elle joue un rôle essentiel dans l'élaboration des tests du χ^2.

Proposition 9. *Soit q une loi de probabilité sur $\{1,\dots,k\}$ différente de p. Alors*

$$\sum_{i=1}^{k} \frac{(N_i(n)/n - q_i)^2}{q_i} \xrightarrow[n\to\infty]{p.s.} \sum_{i=1}^{k} \frac{(p_i - q_i)^2}{q_i}.$$

Conseils biblios 10. Pour retrouver tous ces résultats et bien d'autres, on pourra consulter [Tas85] ou [Sap90] (et [Mon82] pour les plus motivés).

2 Test d'adéquation à une loi donnée

On dispose d'observations que l'on considère comme des réalisations i.i.d. de loi p inconnue. On souhaite ici construire un test qui permette de répondre à la question suivante : la loi des observations est-elle p^0 ? En termes statistiques, on souhaite tester

$$H_0 \ : \ p = p^0 \quad \text{contre} \quad H_1 \ : \ p \neq p^0.$$

C'est par exemple le cas dans l'exemple 1 du dé à six faces avec p^0 la loi uniforme sur $\{1,\dots,6\}$.

2.1 Mise en place du test

On note α le niveau du test (en général $\alpha = 0.1, 0.05, 0.01$). En utilisant le théorème 8 et la proposition 9, on obtient le comportement asymptotique de D_n :

$$D_n = n\sum_{i=1}^{k} \frac{\left(\frac{N_i(n)}{n} - p_i^0\right)^2}{p_i^0} \begin{cases} \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2(k-1) & \text{si } H_0 \text{ est vraie,} \\ \\ \xrightarrow[n\to\infty]{p.s.} +\infty & \text{sinon.} \end{cases}$$

Le niveau étant fixé, on choisit une région de rejet égale à $\{D_n \geq x_{1-\alpha}\}$ où $x_{1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi $\chi^2(k-1)$. La règle de décision est la suivante. On calcule D_n grâce aux observations. Si $D_n \geq x_{1-\alpha}$ alors on rejette H_0, sinon on accepte H_0 (il vaut mieux dire que l'on ne rejette pas H_0).

Exemple 11 (Un peu de génétique)**.** Deux cobayes (génération 0) de lignées pures (c'est-à-dire qu'ils ont les mêmes caractéristiques que leurs ancêtres depuis des générations) dont les pelages sont gris et lisse pour le premier et blanc et rude pour le second ont donné une progéniture homogène au pelage gris et lisse. En croisant ces cobayes de la génération 1 entre eux, on a obtenu 64 descendants dont les pelages se répartissent de la manière indiquée dans le tableau suivant :

Pelage	gris et lisse	blanc et lisse	gris et rude	blanc et rude
Effectifs	33	13	15	3

Faisons les hypothèses de modélisation suivantes (on parle de modèle mendélien) :
— les cobayes sont des animaux diploïdes (ils possèdent deux versions d'un même chromosome) ;

— le gène responsable de la couleur du pelage est présent sous la forme de deux allèles, l'un dominant (A) associé au gris, l'autre récessif (a) associé au blanc ;
— le gène responsable de la texture du pelage est présent sous la forme de deux allèles, l'un dominant (B) associé au lisse, l'autre récessif (b) associé au rude ;
— les gènes responsables de la couleur et la texture du pelage sont sur des chromosomes différents ;
— chaque parent donne, au hasard, à son descendant une copie d'un des deux chromosomes de chaque paire, et ce indépendamment de l'autre parent.

1. Quel est le patrimoine génétique des cobayes de la génération 0 ? de la génération 1 ?
2. Montrer que la distribution théorique des cobayes de la génération 2 si le modèle mendélien tient est $(9/16, 3/16, 3/16, 1/16)$ où l'on a rangé les individus selon leur phénotype : 1 pour gris et lisse, 2 pour gris et rude, 3 pour blanc et lisse, 4 pour blanc et rude.
3. Ces résultats expérimentaux sont-ils conforment au modèle mendélien au niveau 0.05 ?

2.2 Quelques adaptations

Le test du χ^2 s'appuie sur deux résultats asymptotiques (une convergence en loi si H_0 est vraie et une convergence presque sure si H_1 est vraie). Or on ne dispose jamais que d'un nombre fini d'observations. Toute la question est de savoir si l'on a le droit de faire comme si la limite en loi était une égalité. En pratique, les livres recommandent la recette suivante : pour que le test soit valide, il faut que, pour tout $i = 1, \ldots, k$, np_i soit supérieur ou égal à 5. Si ce n'est pas le cas, il faut regrouper des classes à trop faibles effectifs pour atteindre le seuil exigé.

Le test du χ^2 peut aussi être utilisé pour tester l'adéquation à une loi sur $\mathbb{N}$, sur $\mathbb{R}$ ou même sur $\mathbb{R}^d$. Pour cela, il suffit de découper l'espace en un nombre fini de classes et faire fonctionner la moulinette χ^2. Pour une loi sur $\mathbb{N}$, on utilise le découpage suivant :

$$\mathbb{N} = \{0\} \cup \cdots \cup \{k\} \cup \{l \geq k+1\}.$$

2.3 Remarque très importante

Les hypothèses d'un test peuvent être vues comme des parties de l'ensemble des mesures de probabilité sur un certain espace. Dans notre cas, H_0 représente un singleton et H_1 son complémentaire. Les tests du χ^2 (comme tous les tests non paramétriques) ne donnent vraiment d'informations que si l'hypothèse H_0 est rejetée. En effet, si H_0 n'est pas rejetée, il se peut très bien que ce se soit parce que la loi p de l'échantillon est dans H_1 mais tout près de p^0. Ceci est encore renforcé lorsque l'on est obligé de regrouper des classes faute d'un échantillon trop petit ou de créer des classes pour des lois continues : des tas de lois fourniront les mêmes vecteurs de probabilité sur l'ensemble fini. **On se sert d'un test non paramétrique pour invalider un modèle**. Si H_0 est rejetée, alors il faut changer de modèle. Si H_0 n'est pas rejetée c'est que le modèle (bien que simpliste, approximatif... et vraisemblablement faux) est satisfaisant. Pensez au modèle des lois de la gravitation, parfaitement capable de décrire la trajectoire des astres mais en défaut sur la description des particules élémentaires.

Conseils biblios 12. Pour le principe de la méthode et de nombreux exemples, on pourra consulter [Tas85], [DRV01] et [CDD99]. L'exemple 11 est adapté de [DRV01, p. 123].

3 Test d'adéquation à une famille de lois

On souhaite ici mettre en place un test permettant de décider si la loi de l'échantillon appartient ou non à une famille de lois $(p(\theta))_{\theta\in\Theta}$ indexée par un paramètre θ à valeurs dans $\Theta \subset \mathbb{R}^d$. On suppose que l'on dispose de $\hat{\theta}_n$ l'estimateur du maximum de vraisemblance de θ. Il nous faut nous munir d'un théorème un peu plus puissant (et difficile) que le théorème 8. Sa démonstration est délicate car elle fait intervenir les propriétés des estimateurs du maximum de vraisemblance.

Théorème 13. *Soit $(X_n)_{n\in\mathbb{N}^*}$ une suite de variables aléatoires i.i.d. de loi $p(\theta)$ (avec $\theta \in \Theta$ inconnu) alors*

$$D'_n = \sum_{i=1}^{k} \frac{(N_i(n) - np_i(\hat{\theta}_n))^2}{np_i(\hat{\theta}_n)} \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2(k-d-1).$$

Remarque 14. Pour les familles classiques comme $(\mathcal{B}(n,\theta))_{\theta\in]0,1[}$, $(\mathcal{N}(\theta,1))_{\theta\in\mathbb{R}}$, $(\mathcal{P}(\theta))_{\theta>0}$, l'estimateur du maximum de vraisemblance est l'estimateur de la méthode des moments.

Remarque 15. Le fait que la loi limite soit toujours une loi du χ^2 pourrait paraître miraculeux. Cela tient au fait que l'estimateur du maximum de vraisemblance possède dans une très grande majorité des cas le comportement suivant : $\hat{\theta}_n$ converge p.s. vers θ (on dit qu'il est fortement consistant) et $\sqrt{n}(\hat{\theta}_n - \theta)$ converge en loi vers une mesure gaussienne.

Remarque 16. On peut toutefois faire un commentaire qualitatif sur le nombre de degrés de liberté de la loi limite. En effet, il est naturel que celui-ci soit plus petit que $k-1$ puisque l'on compare les fréquences empiriques, non plus à une loi fixée, mais à la loi la plus vraisemblable dans une famille paramétrée au vu des observations. Il paraîtrait logique que D'_n soit d'une certaine façon plus petit que D_n. C'est bien ce qui se passe puisque la fonction de répartition de la loi $\chi^2(k-1)$ est inférieure à celle de la loi $\chi^2(k-d-1)$ (penser à l'interprétation en terme de somme de carrés de variables aléatoires gaussiennes indépendantes). On dit que D'_n est stochastiquement inférieur à D_n.

Corollaire 17. *Pour tester $H_0 : p \in \{p(\theta), \theta \in \Theta\}$ contre $H_1 : p \notin \{p(\theta), \theta\in\Theta\}$ on utilise la statistique D'_n qui a les comportements suivants selon que H_0 ou H_1 est vraie :*

$$D'_n = \sum_{i=1}^{k} \frac{(N_i(n) - np_i(\hat{\theta}_n))^2}{np_i(\hat{\theta}_n)} \begin{cases} \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2(k-d-1) & \textit{si } H_0 \textit{ est vraie,} \\ \\ \xrightarrow[n\to\infty]{p.s.} +\infty & \textit{sinon.} \end{cases}$$

Exercice 18. Pour 10000 fratries de quatre enfants (exactement), on a relevé le nombre de garçons :

nombre de garçons	0	1	2	3	4
effectifs	572	2329	3758	2632	709

On modélise les naissances successives de la façon suivante.
— les naissances sont indépendantes ;

— à chaque naissance, la livraison est un garçon ou une fille avec probabilités respectives θ et $1-\theta$.

1. Dans ce modèle, quelle est la loi p du nombre de garçons dans une fratrie de quatre enfants ?
2. Tester l'hypothèse $H_0 : p = \mathcal{B}(4, 1/2)$ contre $H_1 : p \neq \mathcal{B}(4, 1/2)$ au niveau 0,05.
3. Tester l'hypothèse $H_0 : p \in \{\mathcal{B}(4, \theta), \theta \in]0, 1[\}$ contre $H_1 : p \notin \{\mathcal{B}(4, \theta), \theta \in]0, 1[\}$.
4. Conclusion ?

Exercice 19. On étudie le nombre de connexion à Google pendant la durée de temps unitaire d'une seconde. On fait 200 mesures.

nombre de connexion par seconde	0	1	2	3	4	5	6	7	8	9	10	11
effectif empirique	6	15	40	42	37	30	10	9	5	3	2	1

Soit X la variable aléatoire à valeurs dans $\mathbb{N}$ comptant le nombre de connexions par seconde. Peut-elle être considérée comme une loi de Poisson au niveau 5% ?

Conseils biblios 20. On trouvera un exercice semblable à l'exercice 18 dans [CDD99, p. 112]. L'exercice 19 est tiré de [Tas85, p. 313] avec un habillage différent.

4 Test d'indépendance

Soit $(Y_1, Z_1), \ldots, (Y_n, Z_n)$ des variables aléatoires i.i.d. avec $(Y_l)_{1\leq l\leq n}$ à valeurs dans $\{1, \ldots, r\}$ et $(Z_l)_{1\leq l\leq n}$ à valeurs dans $\{1, \ldots, s\}$. La loi de (Y_1, Z_1) est donnée par une matrice $P = (p_{ij})_{1\leq i\leq r, 1\leq j\leq s}$ à coefficients positifs dont la somme vaut 1 : $p_{ij} = \mathbb{P}(Y_1 = i,\ Z_1 = j)$. Notons, pour $i = 1, \ldots, r$ et $j = 1, \ldots, s$,

$$p_{i\cdot} = \mathbb{P}(Y_1 = i) = p_{i1} + \cdots + p_{is} \quad \text{et} \quad p_{\cdot j} = \mathbb{P}(Z_1 = j) = p_{1j} + \cdots + p_{rj}.$$

Les variables aléatoires Y_1 et Z_1 sont indépendantes si et seulement si, pour tous i et j, $p_{ij} = p_{i\cdot}p_{\cdot j}$. À partir de l'échantillon, définissons les variables aléatoires suivantes :

$$N_{ij} = \mathrm{Card}\{l = 1, \ldots, n\ ;\ (X_l, Y_l) = (i, j)\}, \quad N_{i\cdot} = N_{i1} + \cdots + N_{is} \quad \text{et} \quad N_{\cdot j} = N_{1j} + \cdots + N_{rj}$$

Proposition 21. *Avec les notations ci-dessus,*

$$D_n = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(N_{ij} - \frac{N_{i\cdot}N_{\cdot j}}{n}\right)^2}{\frac{N_{i\cdot}N_{\cdot j}}{n}} \begin{cases} \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2((r-1)(s-1)) & \textit{si } Y_1 \textit{ et } Z_1 \textit{ sont indépendantes,} \\ \xrightarrow[n\to\infty]{p.s.} +\infty & \textit{sinon.} \end{cases}$$

Remarque 22. Cette proposition peut être vue comme un corollaire du théorème 13. En effet on teste l'adéquation de la loi du couple à la famille paramétrique des lois produit sur $\{1, \ldots, r\} \times \{1, \ldots, s\}$ en estimant les paramètres par la méthode du maximum de vraisemblance. Le nombre de paramètres estimés est $(r-1) + (s-1)$ puisque la donnée des $r-1$ premiers coefficients de la loi de Y donne le dernier (et idem pour Z) et que la donnée des lois marginales d'une loi produit détermine la loi du couple. Sous l'hypothèse d'indépendance de Y et Z, la statistique D_n converge donc vers une loi du χ^2 à $rs - (r+s-2) - 1 = (r-1)(s-1)$ degrés de liberté.

Exercice 23 (Yeux et cheveux)**.** Depuis une terrasse de café ensoleillée, un statisticien en plein travail a noté les couleurs des yeux et des cheveux de 124 passants.

YeuxCheveux	blonds	bruns	roux	noirs
bleus	25	9	7	3
gris	13	17	7	10
marrons	7	13	5	8

Les deux critères sont-ils indépendants au niveau 0.05 ?

Encore d'autres exemples et la description de la méthode pour calculer la statistique de test D_n dans [Tas85], [DRV01] et [CDD99].

5 Test d'homogénéité

Les tests du χ^2 permettent aussi de tester l'homogénéité de plusieurs échantillons.

On étudie un caractère pouvant prendre k valeurs $A_1, \dots, A_k$ (ou k modalités ou à valeurs dans k classes). On dispose de l échantillons $E_1, \dots, E_l$ différents. Pour tout $i \in \{1, \dots, k\}$, pour tout $j \in \{1, \dots, l\}$, on connaît l'effectif observé O_{ij} de la valeur A_i dans l'échantillon E_j. On souhaite tester

H_0 : les échantillons sont issus de la même loi contre H_1 : les échantillons n'ont pas même loi.

La mise en place pratique du test est la même que pour le test d'indépendance. On définit

$$O_{i\cdot} = O_{i1} + \cdots + O_{il}, \quad O_{\cdot j} = O_{1j} + \cdots + O_{kj} \quad \text{et} \quad n = \sum_{i=1}^{k} \sum_{j=1}^{l} O_{ij} \left(= \sum_{i=1}^{k} O_{i\cdot} = \sum_{j=1}^{l} O_{\cdot j} \right)$$

Proposition 24. *Avec les notations ci-dessus,*

$$D_n = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{\left(O_{ij} - \frac{O_{i\cdot} O_{\cdot j}}{n}\right)^2}{\frac{O_{i\cdot} O_{\cdot j}}{n}} \begin{cases} \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2((k-1)(l-1)) & \text{si } H_0 \text{ est vraie,} \\ \xrightarrow[n\to\infty]{p.s.} +\infty & \text{sinon.} \end{cases}$$

Exercice 25 (Y-a un nouvel Omoooo ?)**.** On cherche à invalider le lieu commun qui affirme que toutes les lessives se valent. On utilise trois lessives appelées A, B et C et, à la sortie du lavage, on classe les vêtements en trois catégories : très sale (TS), légèrement sale (LS) et propre (P).

LessiveLinge	T.S.	L.S.	P.
A	30	65	205
B	23	56	121
C	75	125	300

Peut-on dire, au niveau 5%, que toutes les lessives sont identiques ?

Références

[CDD99] F. Couty, J. Debord et F. Daniel – *Probabilités et statistiques*, Dunod, 1999.

[DRV01] J.-J. Daudin, S. Robin et C. Vuillet – *Statistique inférentielle*, Presses Universitaires de Rennes, 2001.

[Mon82] A. Monfort – *Cours de statistique mathématique*, Économia, 1982.

[Sap90] G. Saporta – *Probabilités, analyse de données et statistique*, Éditions Technip, 1990.

[Tas85] P. Tassi – *Méthodes statistiques*, Économia, 1985.