Fiche 7 : Test du χ^2

Proposé par Pearson en 1900, le test du χ^2 est un test statistique permettant de tester l'adéquation d'une série de données à une famille de lois de probabilités ou de tester l'indépendance entre deux variables aléatoires.
Considérons une suite de variables aléatoire $X_1, X_2, \dots, X_n$ indépendantes et identiquement distribuées et à valeur dans un ensemble discret $\{B_1, \dots, B_r\}$ de probabilités $p_1, \dots, p_r$. Définissons les variables aléatoire de fréquence associées π_j :

$$\pi_j = \frac{1}{N}\text{card}\{i|X_i = j\}$$

Le théorème de Pearson affirme que : *la variable aléatoire* $T \sum_{j=1}^{r} \frac{N(\pi_j - p_j)^2}{p_j}$ *converge en loi (quand* N *tend vers l'infini) vers une loi du* χ^2 *à* $r-1$ *degrés de liberté.* Ce résultat repose sur le théorème central limite et est donc asymptotique mais est très rapidement une très bonne approximation.
Considérons maintenant un échantillon $x_1, \dots, x_n$ et notons $\hat{p}_j$ la fréquence empirique de la valeur j :

$$\hat{p}_j = \frac{1}{N}\text{card}\{i|x_i = j\}$$

Si H_0 est vraie (i.e., si notre échantillon a bien été tiré à partir d'une suite de variables aléatoires suivant la probabilité p), alors on peut tester s'il est raisonnable que la quantité $\hat{T} = \sum_{j=1}^{r} \frac{N(\hat{p}_j - p_j)^2}{p_j}$ ait été tiré selon χ^2_{r-1}. En se donnant un seuil de confiance α, on calcule donc c tel que $\alpha = \mathbb{P}(T > c) = \chi^2_{r-1}(c, \infty)$ et on compare $\hat{T}$ à c.
L'hypothèse alternative H_1 est que notre échantillon ait été tiré selon une loi p' distincte de p en au moins un point correspond pas. Dans ce cas, la quantité T va exacerber les différences de fréquences entre p et p' et T va tendre vers l'infini avec n. Il existe donc une zone de recouvrement entre ces deux hypothèses non négligeable quand n est petit mais qui décroît rapidement avec n.

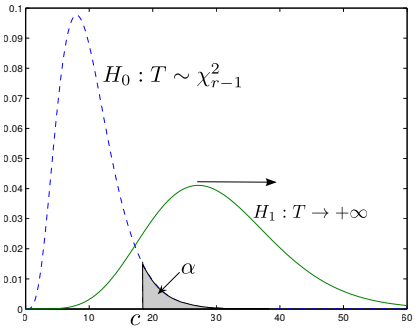

FIGURE 1 – Évolution de T selon H_0 ou H_1

La fonction R qui fait tous ce calculs pour vous s'appelle `chisq.test`.

Exercice 1. Illustration

1. On lance un dé 150 fois et on obtient la distribution suivante :

face	1	2	3	4	5	6
Nombre d'occurrences	22	21	22	27	22	36

   Notre dé est-il pipé ? S'il était uniforme, la probabilité de chacune des faces serait de $1/6$ et sur 150 lancers, on s'attendrait à avoir environ 25 occurrences de chaque face. Mais ici, le 6 est sorti 36 fois. Coïncidence ou dé pipé ?

2. Le pourcentage d'occurrence des 5 lettres les plus fréquentes de la langue anglaise est environ :

lettre	E	T	N	R	O
fréquence	29	21	17	17	16

   Ce type d'information peut s'avérer très utile en cryptographie pour casser des codes basés sur des décalages simples. Considérons maintenant un texte dont le nombre d'occurrences des lettres E,T,N,R et O est :

lettre	E	T	N	R	O
fréquence	100	110	80	55	14

   Effectuez un test du χ^2 pour tester l'hypothèse que la proportion des lettres dans ce texte corresponde à celle de la langue anglaise [1].

Exercice 2. Goodness-of-fit d'une loi continue Considérons une suite de variables aléatoire $X_1, X_2, \dots, X_n$ indépendantes et identiquement distribuées, à valeur dans $\mathbb{R}$ et de distribution p. On souhaite tester les hypothèses suivantes :

$$\begin{cases} H_0 : & p = p_0 \\ H_1 : & p \neq p_0 \end{cases},$$

où p_0 est une distribution donnée. Pour utiliser le test du χ^2, on échantillonne la loi p_0 comme on le ferait pour un histogramme. Le test du χ^2 reposant sur un passage à la limite, on s'assure en général que le nombre d'occurrences dans chaque intervalle est au moins 5. Dans le cas contraire, R émettra un warning.
Reprenons le DM sur la dynamique des populations [2] et testons :

1. l'adéquation d'un échantillon de 500 villes partant de l'état initial $(n_1(1), n_2(2) = (1, 1)$ avec une distribution uniforme sur $[0, 1]$;
2. l'adéquation d'un échantillon de 500 villes partant de l'état initial $(n_1(1), n_2(2) = (10, 10)$ avec une loi normale centrée en 0 et de variance correspondant à celle de cet échantillon.

1. C'est en réalité un peu plus compliqué que ça puisque le test du χ^2 repose sur l'hypothèse de l'indépendance de chaque lettre, ce qui n'est pas très raisonnable mais on fera avec...
2. http://rpubs.com/alegrand/10496