TP6 : TESTS STATISTIQUES

1. Tests paramétriques

Exercice 1 [Etude de la robustesse [1] d'un test]. Soit une suite d'observations $X_1, X_2, \ldots, X_n$ i.i.d. de loi μ. On note $\bar{X}_n$ la moyenne empirique et S^2 la variance empirique (version biaisée) :

$$\bar{X}_n = \frac{1}{n}\sum_{t=1}^{n} X_t, \qquad S^2 = \frac{1}{n}\sum_{t=1}^{n} \left(X_t - \bar{X}_n\right)^2 .$$

(1) On suppose que μ est une loi gaussienne, de moyenne m et de variance σ^2. Quelle est la loi de S^2 ?

(2) On pose $H_0 : \sigma^2 \leq 1$. Construire un test de niveau (exactement) α.

On se demande maintenant si le test construit précédemment s'étend à des lois non-gaussiennes, *id est*, si lorsque μ n'est plus gaussienne, le test précédent (avec la même statistique de test, avec le même choix pour la zone de rejet) est toujours, au moins asymptotiquement, de niveau α. On admet pour l'instant les résultats suivants :

– si $\chi^2_{n-1,\alpha}$ désigne le α-quantile de la loi du χ^2 à $n-1$ degrés de liberté, alors

$$\frac{\chi^2_{n-1,\alpha} - n}{\sqrt{2}\sqrt{n}} \longrightarrow u_\alpha \ ,$$

où u_α désigne le α-quantile de la loi gaussienne standard ;

– on a la convergence en loi suivante :

$$\sqrt{n}\left(\frac{S^2}{\sigma^2} - 1\right) \rightsquigarrow \mathcal{N}(0,\, \kappa - 1) \ ,$$

où κ désigne la kurtosis de μ, définie par

$$\kappa = \frac{\mu_4}{\mu_2^2},$$

avec, pour $k \in \mathbb{N}$, $\mu_k = \mathbb{E}\left[\left(X - \mathbb{E}[X]\right)^k\right]$ (notez qu'en particulier, $\mu_2 = \sigma^2$ est la variance de μ.)

(3) On note Φ la fonction de répartition [2] de la loi gaussienne. Prouvez qu'asymptotiquement, le test précédent est de niveau $1 - \Phi(u_\alpha\sqrt{2}/\sqrt{\kappa - 1})$, c'est-à-dire que

$$\limsup_n \sup_{\sigma^2 \leq 1} \mathbb{P}_{\sigma^2}(\text{rejet du test}) \leq 1 - \Phi\left(\frac{u_\alpha\sqrt{2}}{\sqrt{\kappa - 1}}\right).$$

On a donc prouvé la non-robustesse contre une modification de la valeur de la kurtosis.

1. La robustesse d'un test est définie comme la non-sensibilité de la procédure de test à la loi des observations. Le test asymptotique sur la moyenne fondé sur le TCL est ainsi robuste sur l'ensemble des lois admettant un moment d'ordre deux.

2. En particulier, on a donc $1 - \Phi(u_\alpha) = \alpha$.

(4) Mettons en évidence cette non-robustesse par voie de simulations. Pour $n = 20$, estimez par méthode de Monte–Carlo le niveau réel du test proposé à la question (2), pour $\alpha = 5\%$ et μ donné, d'une part par une loi de Laplace, et d'autre part par un mélange de gaussiennes $0.95\mathcal{N}(0,\,1) + 0.05\mathcal{N}(0,\,9)$ par exemple (il faut renormaliser ces deux lois pour être dans l'hypothèse nulle).

(5) Il faut encore prouver les deux résultats que nous avions admis. Ils découlent tous deux de la convergence en loi suivante, que je vous invite à vérifier et à prouver (on note m l'espérance de μ) :

$$\sqrt{n}\left(S^2 - \sigma^2\right) = \sqrt{n}\left(\frac{1}{n}\sum_{t=1}^{n}(X_t - m)^2 - \sigma^2\right) - \sqrt{n}\left(\bar{X}_n - m\right)^2 \rightsquigarrow \mathcal{N}(0,\,\mu_4 - \mu_2^2)\ .$$

2. Tests non paramétriques

Dans cette section, l'objectif est de savoir si une ou plusieurs lois de probabilité verifient certaines conditions, par exemple si un échantillon de loi inconnue a des chances de provenir d'une loi donnée, par exemple normale (test d'*ajustement*) ou si deux échantillons de données proviennent de variables indépendantes (test d'*indépendance*). Les tests portent sur des mesures de probabilité (caracterisées dans le cas continu par leur densité) et non plus sur des paramètres sur ces lois. On parle donc de tests *non-paramétriques*.

Exercice 2 [Tests d'ajustement et fonctions de répartition empiriques]. Nous allons tester si les générateurs de nombres aléatoires construits au TP2 satisfont les critères classiques de test sur les fonctions de répartition empiriques.

(1) Reprendre (ou réécrire rapidement) les générateurs du TP2 simulant respectivement une loi normale centrée réduite par la méthode de Box-Muller et une loi exponentielle de paramètre 1.

(2) Illustrez la convergence des fonctions de répartition empiriques F_n vers F pour chacune des lois précédentes. Quel théorème assure cette convergence et comment s'énonce-t-il ?

(3) Pour une taille d'échantillon n fixée, simulez N réalisations de la variable aléatoire $\sup_{x\in\mathbb{R}} |F_n(x) - F(x)|$ pour chacune des lois précédentes. On pourra remarquer que

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| = \max_{1\le i\le n}\left(\max\left(\left|\frac{i-1}{n} - F(X_{(i)})\right|, \left|\frac{i}{n} - F(X_{(i)})\right|\right)\right),$$

où $X_{(1)} \le \cdots \le X_{(n)}$. Comparez les lois des deux N-échantillons ainsi obtenus.

On constate que la quantité précédente $\sup_{x\in\mathbb{R}} |F_n(x) - F(x)|$ ne dépend pas de la loi F. Cependant, cette quantité dépend encore de n. On préfère donc s'intéresser à la statistique K_n suivante :

$$K_n = \sqrt{n}\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| = \sqrt{n}\|F_n - F\|_\infty.$$

C'est la statistique de Kolmogorov-Smirnov, qui converge en loi, si F est continue, vers la loi du même nom.

(4) Illustrez graphiquement la convergence en loi précédente pour les simulateurs de lois normale et exponentielle (vous pouvez utiliser la fonction `pks` de Stixbox).

(5) Proposez un test non-paramétrique d'ajustement de la loi de l'échantillon $X_1, \ldots, X_n$ à une loi F_0 donnée : quelles sont les hypothèses H_0 et H_1 du test, quand rejetez-vous H_0 ? Mettez en oeuvre le test, au niveau 0.05, dans les cas suivants :
   – échantillon de loi normale standard avec F_0 une loi normale standard,

– échantillon de loi exponentielle avec F_0 une loi exponentielle,
– échantillon de loi normale standard avec F_0 une loi Laplace.

Quel est le résultat du test, pour quelle taille d'échantillon ?

Les exercices suivants portent sur les tests non-paramétriques du χ^2 : test d'ajustement et test d'indépendance. Revoyez les résultats théoriques correspondants !

Exercice 3 [Générateurs pseudo-aléatoires]. Un ordinateur possède un générateur pseudo-aléatoire de nombres choisis au hasard dans l'ensemble des entiers de 0 à 9. On dispose d'un échantillon de taille $N = 1000$ de chiffres tirés par ce générateur. Les résultats sont répartis dans le tableau suivant.

Chiffres	0	1	2	3	4	5	6	7	8	9
Observations	120	87	115	103	91	109	92	112	94	77

(1) On veut tester l'hypothèse d'équiprobabilité pour chaque chiffre. Pour cela, mettez en oeuvre le test d'ajustement du χ^2. Quelle statistique de test considérer ? Quelle est sa loi limite sous H_0 ? Choisissez vous d'accepter l'hypothèse d'équiprobabilité pour l'échantillon précédent, et si oui pour quel niveau α ?

(2) Faites de même en remplaçant la table précédente par une table générée à partir de la fonction `rand` de Matlab.

Exercice 4 [Répartition de notes]. On dispose de la répartition de notes obtenues dans deux matières : Mathématiques X et Philosophie Y.

$X \setminus Y$	[0,4[	[4,8[	[8,12[	[12,16[	[16,20]
[0,4[	3	4	2	0	0
[4,8[	6	10	8	2	0
[8,12[	1	8	20	12	3
[12,16[	0	0	8	7	3
[16,20]	0	0	1	0	2

Testez l'hypothèse d'indépendance entre les notes obtenues en Mathématiques et en Philosophie. Pour cela, mettez en oeuvre un test du χ^2 d'indépendance, en répondant aux même questions qu'à l'exercice précédent.