Feuille de travaux pratiques - Python #5

Emeline LUIRARD

Ce TP est consacré aux tests statistiques : test du χ^2 et test de Kolmogorov-Smirnov.

1 Rappels sur les tests

Pour construire un test, il faut tout d'abord définir deux hypothèses :
- L'hypothèse nulle H_0, celle qu'on pense être vraie en général. Elle est la plus précise possible.
- L'hypothèse alternative H_1. On prend souvent le complémentaire de H_0.

Ensuite, on construit une statistique D, qui est une fonction de notre échantillon qui va vérifier :
- Sous H_0, D suit (asymptotiquement) une loi de fonction de répartition connue
- Sous H_1, D est (asymptotiquement) grand avec une grande probabilité.

Il va ensuite falloir définir une région de rejet pour construire la règle de décision :

$$D \in R \Rightarrow \text{on rejette } H_0$$
$$D \notin R \Rightarrow \text{on ne rejette pas } H_0$$

Pour trouver cette région, on définit le niveau du test $\alpha = \mathbb{P}_{H_0}(H_0 \text{ rejeté})$. Il doit être le plus petit possible. Alors $\alpha = \mathbb{P}_{H_0}(H_0 \text{ rejeté}) = \mathbb{P}_{H_0}(D \in R)$.

2 Les tests du χ^2

2.1 Test d'adéquation à une loi de probabilité sur un ensemble fini

Ce test a pour but de décider si un vecteur d'observations est une réalisation d'un échantillon de variables aléatoires de loi donnée. Cette dernière loi doit être à valeurs dans un ensemble **fini**. Soit $(x_1, \cdots, x_n)$ une réalisation d'un vecteur aléatoire $(X_1, \cdots, X_n)$ i.i.d. de loi inconnue $p = (p_1, \cdots, p_k)$, une probabilité sur $[\![1, k]\!]$. On note, pour $i \in [\![1, k]\!]$, $N_i(n) = Card\{j \in [\![1, n]\!], X_j = i\}$.

On suppose par ailleurs donnée une loi $p^0 = (p_1^0, \ldots, p_k^0)$. On souhaite tester l'hypothèse nulle $H_0 = \{p = p^0\}$ contre l'hypothèse alternative $H_1 = \{p \neq p^0\}$. On définit alors la statistique

$$D_n = D_n(p^0, X_1, \cdots, X_n) := n \times \sum_{i=1}^{k} \frac{(N_i(n)/n - p_i^0)^2}{p_i^0} = \sum_{i=1}^{k} \frac{(N_i(n) - np_i^0)^2}{np_i^0},$$

dont le comportement asymptotique est le suivant :

Théorème 2.1.

$$D_n = \sum_{i=1}^{k} \frac{(N_i(n) - np_i^0)^2}{np_i^0} \begin{cases} \xrightarrow{\mathcal{L}} \chi^2(k-1) & \textit{sous } H_0, \\ \xrightarrow[n\to+\infty]{p.s.} +\infty & \textit{sous } H_1. \end{cases}$$

La commande `D,p_valeur=scipy.stats.chisquare(f_obs, f_exp)` calcule la statistique D_n et calcule une p-valeur à partir du nombre d'occurrences observées `f_obs` et attendues `f_exp`.

Étant donné un niveau α (souvent $\alpha = 5\%$) et un réel η_α tel que $\mathbb{P}(\chi^2 > \eta_\alpha) = \alpha$, la zone de rejet $W_n = \{D_n > \eta_\alpha\}$ fournit alors un test de niveau asymptotique α pour $H_0 = \{p = p^0\}$ contre $H_1 = \{p \neq p^0\}$.

Exercice 1. *On suppose donnés une mesure de probabilité p^0 de support fini A, un vecteur de données $(x_i)_{1\leq i\leq n} \in A^n$ et un seuil $0 < \alpha < 1$. Ecrire un programme qui prend en entrées p, $(x_i)_{1\leq i\leq n}$ et α et qui en sortie donne le résultat du test du χ^2 d'adéquation de niveau α.*

N.B. *En pratique, on considère que l'approximation en loi par $\chi^2(k-1)$ est valide sous H_0 si $n \times \min_{1\leq j\leq k} p_j^0 \geq 5$. Si cette condition n'est pas satisfaite, on peut regrouper les classes à trop faibles effectifs afin d'atteindre ce seuil.*

Exercice 2. *En faisant appel deux cents fois consécutives à un générateur d'entiers pseudo aléatoires, avec un niveau de confiance de 99%, décider si le générateur fournit des données équiréparties.*

N.B. *Lorsque l'on a affaire à des lois sur $\mathbb{N}$, $\mathbb{R}$,..., on peut tout de même utiliser le test du χ^2 en découpant l'espace en un nombre fini de classes.*

2.2 Test d'adéquation à une famille de lois

On peut aussi se demander si la loi de l'échantillon appartient ou non à une famille de lois $(p(\theta))_{\theta\in\Theta}$, $\Theta \subset \mathbb{R}^d$. On note $\hat{\theta}_n$ l'estimateur du maximum de vraisemblance de θ. On va alors tester l'hypothèse nulle $H_0 = \{p \in \{p(\theta), \theta \in \Theta\}\}$ contre l'hypothèse alternative $H_1 = \{p \notin \{p(\theta), \theta \in \Theta\}\}$. On définit la statistique

$$D'_n = D'_n(p, X_1, \cdots, X_n) := \sum_{i=1}^{k} \frac{(N_i(n) - np_i(\hat{\theta}_n)^2}{np_i(\hat{\theta}_n)},$$

dont le comportement asymptotique est le suivant :

Théorème 2.2.

$$D'_n = \sum_{i=1}^{k} \frac{(N_i(n) - np_i(\hat{\theta}_n)^2}{np_i(\hat{\theta}_n)} \begin{cases} \xrightarrow{\mathcal{L}} \chi^2(k-d-1) & \text{sous } H_0, \\ \xrightarrow[n\to+\infty]{p.s.} +\infty & \text{sous } H_1. \end{cases}$$

Exercice 3. *On étudie le nombre de connexions 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 v.a. à valeurs dans $\mathbb{N}$ comptant le nombre de connexions par seconde. Peut-elle être considérée comme une loi de Poisson au niveau 5% ?

2.3 Test d'indépendance

On dispose d'un échantillon d'une loi $Z = (X, Y)$ et l'on souhaite déterminer si les variables X et Y sont indépendantes. Considérons donc n données $(z_1, \ldots, z_n) = ((x_1, y_1), \ldots, (x_n, y_n))$ dont

on suppose qu'elles sont les réalisations indépendantes et identiquement distribuées de variables aléatoires $(Z_1, \dots, Z_n) = ((X_1, Y_1), \dots, (X_n, Y_n))$ à valeurs dans des ensembles finis $[\![1, r]\!]$, $[\![1, s]\!]$. On note $p = (p_{ij}, 1 \leq i \leq r, 1 \leq j \leq s)$ la loi du couple $Z = (X, Y)$, c'est-à-dire :

$$p_{ij} = \mathbb{P}(Z = (i, j)) = \mathbb{P}(X = i, Y = j).$$

On introduit

$$N_{ij} = Card\{k, X_k = i, Y_k = j\}, \quad N_{i.} = N_{i1} + \cdots + N_{is}, \quad N_{.j} = N_{1j} + \cdots + N_{rj}.$$

Alors, avec l'hypothèse $H_0 = \{X$ et Y sont indépendants$\}$ et $H_1 = H_0^C$,

Théorème 2.3.

$$D_n = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(N_{ij} - \frac{Ni.N_{.j}}{n})^2}{\frac{Ni.N_{.j}}{n}} \begin{cases} \xrightarrow{\mathcal{L}} \chi^2((r-1)(s-1)) & \text{sous } H_0, \\ \xrightarrow[n \to +\infty]{p.s.} +\infty & \text{sous } H_1. \end{cases}$$

La commande `D, p_valeur, dlib, expected=scipy.stats.chi2_contingency(f_obs)` calcule la statistique D_n et la p-valeur, à partir d'un tableau à deux entrées contenant les nombres d'occurrences observées pour chaque coordonnée. Cela retourne également le degré de liberté et le nombre d'occurrences attendues.

À nouveau, étant donnés un niveau α et un réel η_α tel que $\mathbb{P}(\chi^2 \geq \eta_\alpha) = \alpha$, la zone de rejet $W_n = \{D_n > \eta_\alpha\}$ fournit un test de niveau asymptotique α de $H_0 = \{X$ et Y indépendantes$\}$ contre $H_1 = \{X$ et Y non indépendantes$\}$.

Exercice 4. *Supposons donnés un vecteur $(x_i, y_i)_{1 \leq i \leq n}$ et un seuil $0 < \alpha < 1$. Ecrire un programme qui prend en entrées le vecteur $(x_i, y_i)_{1 \leq i \leq n}$ et le seuil α et qui en sortie donne le résultat du test du χ^2 d'indépendance de niveau α.*

Exercice 5. *On désire étudier la répartition des naissances suivant le type du jour dans la semaine (jours ouvrables ou week-end) et suivant le mode d'accouchement (naturel ou par césarienne). Les données proviennent du "National Vital Statistics Report" et concernent les naissances aux USA en 1997.*

Naissances	*Naturelles*	*César.*	*Total*
J.O.	*2331536*	*663540*	*2995076*
W.E.	*715085*	*135493*	*850578*
Total	*3046621*	*799033*	*3845654*

Naissances	*Naturelles*	*César.*	*Total*
J.O.	60.6%	17.3%	77.9%
W.E.	18.6%	3.5%	22.1%
Total	79.2%	20.8%	100.0%

Tester au niveau 0.1% *l'hypothèse d'indépendance entre le type du jour de naissance (jour ouvrable ou week-end) et le mode d'accouchement (naturel ou césarienne).*

2.4 Test d'homogénéité

On dispose de ℓ échantillons différents $E_1, \cdots, E_\ell$ à valeurs dans $[\![1, k]\!]$. On se demande si ces échantillons ont la même loi. On note

$$O_{ij} = Card\{x \in E_j, x = i\} \quad O_{i.} = O_{i1} + \cdots + O_{i\ell}, \quad O_{.j} = O_{1j} + \cdots + O_{kj},$$

et $n = \sum_{i=1}^{k} \sum_{j=1}^{\ell} O_{ij}$. Alors, avec l'hypothèse $H_0 = \{$les échantillons sont issus de la même loi$\}$ et $H_1 = H_0^C$,

Théorème 2.4.

$$D_n = \sum_{i=1}^{k}\sum_{j=1}^{\ell} \frac{(O_{ij} - \frac{Oi.O_{.j}}{n})^2}{\frac{Oi.O_{.j}}{n}} \begin{cases} \xrightarrow{\mathcal{L}} \chi^2((k-1)(\ell-1)) & \text{sous } H_0, \\ \xrightarrow[n\to+\infty]{p.s.} +\infty & \text{sous } H_1. \end{cases}$$

La commande `D, p_valeur=scipy.stats.friedmanchisquare(echant1, echant2, echant3,...)` calcule la statistique D_n et p-valeur, à partir des échantillons.

3 Tests non paramétriques

Dans toute cette section μ désigne une mesure de probabilité sur $\mathbb{R}$ et F la fonction de répartition associée. On considère $(X_1, \ldots, X_n)$ un $n-$échantillon de loi μ et on note $(X_{(1)}, \ldots, X_{(n)})$ la statistique d'ordre associée.

3.1 Fonction de répartition empirique

La fonction de répartition empirique F_n associée à l'échantillon $(X_1, \ldots, X_n)$ est définie pour tout $x \in \mathbb{R}$ par

$$F_n(x) := \frac{Card\{1 \le k \le n, X_k \le x\}}{n},$$

ou encore

$$F_n(x) = \frac{k}{n} \text{ si } X_{(k)} \le x < X_{(k+1)}.$$

Le théorème de Glivenko–Cantelli assure que $||F_n - F||_\infty$ tend presque sûrement vers zéro lorsque n tend vers l'infini, d'autre part, le théorème de Kolmogorov-Smirnov assure que, si F est continue, $\sqrt{n}\,||F_n - F||_\infty$ converge en loi lorsque n tend vers l'infini vers une variable K dont la loi est appelée loi de Kolmogorov :

$$\lim_{n\to+\infty} \mathbb{P}\left(\sqrt{n}\,||F_n - F||_\infty \le x\right) = \mathbb{P}(K \le x) = 1 - 2\sum_{k=1}^{+\infty}(-1)^{k-1}e^{-2k^2x^2}.$$

Exercice 6. *L'objet de cet exercice est d'illustrer le théorème de Glivenko–Cantelli et le théorème de Kolmogorov–Smirnov. Choisir une fonction de répartition F parmi les fonctions de répartitions continues déjà implémentées dans Python.*

1. *Illustrer le fait que, pour tout $x \in \mathbb{R}$, la suite $(F_n(x))_n$ converge presque sûrement vers $F(x)$ lorsque n tend vers l'infini, en représentant sur un même graphique, la fonction F et plusieurs fonctions de répartition empiriques.*
2. *Montrer que*

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

3. *Illustrer les théorèmes de Glivenko–Cantelli et de Kolmogorov–Smirnov.*

3.2 Test d'adéquation de Kolmogorov–Smirnov

Grâce au théorème de Kolmogorov–Smirnov, on peut facilement mettre en oeuvre un test pour déterminer si un vecteur de données est ou non une réalisation d'un échantillon de loi prescrite. Si la loi en question a pour fonction de répartition F, on calcule la fonction de répartition empirique F_n associée aux données ainsi que la statistique

$$D_n = \sqrt{n}\,||F_n - F||_\infty\,.$$

À un seuil $0 < \alpha < 1$, on associe le nombre c_α tel que $\mathbb{P}(K \geq c_\alpha) = \alpha$. On a par exemple

α	0.10	0.05	0.025	0.01	0.005	0.001
c_α	1.22	1.36	1.48	1.63	1.73	1.95

La zone de rejet $W_n = \{D_n > c_\alpha\}$ fournit alors un test de niveau asymptotique α de l'hypothèse $H_0 = \{$les données sont des réalisation i.i.d. de loi de fonction de répartition $F\}$ et $H_1 = H_0^C$,

Théorème 3.1.

$$D_n = \sqrt{n}\,||F_n - F||_\infty \begin{cases} \overset{\mathcal{L}}{\Longrightarrow} \mu_{KS} & \textit{sous } H_0, \\ \underset{n\to+\infty}{\overset{p.s.}{\longrightarrow}} +\infty & \textit{sous } H_1. \end{cases}$$

La commande `D, p_valeur=scipy.stats.kstest(echant, loi)` calcule la statistique D_n et la p-valeur, à partir de l'échantillon. `loi` peut être le nom d'une loi dans `scipy.stats` ou alors le nom d'une fonction, que vous aurez créée, donnant la fonction de répartition F.

Exercice 7. *Au seuil de confiance 95%, décidez si les données `Z` téléchargeables ici sont des réalisations i.i.d. d'une variable exponentielle de paramètre 1.*

3.3 Comparaison d'échantillon, test de Smirnov

Soient $(X_1, \ldots, X_n)$ et $(Y_1, \ldots, Y_m)$ deux échantillons et F_n et G_m les fonctions de répartitions empiriques associées.
On cherche à tester l'hypothèse $H_0 = \{$Les deux échantillons proviennent d'une même loi continue$\}$ contre l'hypothèse alternative $H_1 = H_0^C$. Pour cela, on considère la statistique

$$D_{n,m} := \sqrt{\frac{mn}{m+n}} \times \sup_{x\in\mathbb{R}} |F_n(x) - G_m(x)|\,.$$

Théorème 3.2.

$$D_{n,m} := \sqrt{\frac{mn}{m+n}} \times \sup_{x\in\mathbb{R}} |F_n(x) - G_m(x)| \begin{cases} \overset{\mathcal{L}}{\Longrightarrow} \mu_{KS} & \textit{sous } H_0, \\ \underset{n\to+\infty}{\overset{p.s.}{\longrightarrow}} +\infty & \textit{sous } H_1. \end{cases}$$

La zone de rejet associée au test est du type $\{D_{m,n} \geq c_\alpha\}$ où $\mathbb{P}(K \geq c_\alpha) = \alpha$.
La commande `D, p_valeur=scipy.stats.ks_2samp(echant1, echant2)` calcule la statistique $D_{n,m}$ et la p-valeur, à partir des deux échantillons.

Références

[Bre07] Jean-Christophe Breton. Tests du χ^2, 2007. https://perso.univ-rennes1.fr/jean-christophe.breton/agreg/AGREG/COURS/khideux.pdf.

[CBCC16] Alexandre Casamayou-Boucau, Pascal Chauvin, and Guillaume Connan. *Programmation en Python pour les mathématiques - 2e éd.* Dunod, Paris, 2e édition edition, January 2016.

[Tas04] Philippe Tassi. *Méthodes statistiques*. Economica, Paris, 3e édition edition, October 2004.

[Vig18] Vincent Vigon. *python proba stat*. Independently published, October 2018.