Feuille de travaux pratiques - Python #4

Ce TP est consacré aux tests statistiques : test du χ^2.
Lorsqu'on veut illustrer graphiquement une convergence en loi, via le tracé des densités ou des fonctions de répartition par exemple, c'est un plus de quantifier cette convergence numériquement. Cela peut être fait avec les test statistiques.

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.

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

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 $\mathbb{P}(\chi^2_{k-1} > \eta_\alpha) = \alpha$, 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 $(z_1, \dots, z_n) = ((x_1, y_1), \dots, (x_n, y_n))$ 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.

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

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.

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

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

Références

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

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