Feuille de TP n°6 – Tests du χ^2

1 Quelques rappels théoriques sur le modèle multinomial

Commençons par un petit rappel. On dit qu'une variable aléatoire X suit une loi du chi-deux à n degrés de liberté si sa loi admet

$$\frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} \exp(-x/2)\mathbf{1}_{\{x>0\}}$$

pour densité par rapport à la mesure de Lebesgue. C'est la loi de la somme de n carrés de variables gaussiennes centrées réduites indépendantes sur $\mathbb{R}$. C'est une loi gamma. En particulier la somme de deux v.a. suivant des lois du chi-deux de degré respectif m et n est encore une loi du chi-deux à $m+n$ degrés de liberté.

On considère une variable qualitative susceptible de prendre k valeurs (ou quantitative ventilée en k classes). On supposera que ces k valeurs sont l'ensemble $\{1,\dots,k\}$. On note $p := (\mathbb{P}(X=l),\ l=1,\dots,k)$ le vecteur des probabilités de chaque valeur possible. On suppose que l'on dispose d'un échantillon $(X_1,\dots,X_n)$ de loi $\mathcal{L}(X)$ et donc des variables aléatoires $N_1^n,\dots,N_k^n$ égales aux effectifs de chaque valeur possible pour X (N_1^n désigne le nombre de X_i qui ont pris la valeur 1). La loi du k-uplet $(N_1^n,\dots,N_k^n)$ est la loi multinomiale :

$$\mathbb{P}(N_1^n=n_1,\dots,N_k^n=n_k) = \frac{n!}{n_1!\dots n_k!} p_1^{n_1}\dots p_k^{n_k}, \quad \sum_{l=1}^k n_l = n.$$

Les variables aléatoires $(N_l^n)_l$ suivent des lois binomiales $(\mathcal{B}(n,p_l))_l$ dépendantes de covariance :

$$\operatorname{cov}(E_l^n, E_m^n) = -\frac{p_l p_m}{n}.$$

À partir des effectifs, on peut calculer les fréquences empiriques $\hat{p}_l^n := N_l^n/n$ pour $l=1,\dots,k$ et les comparer aux probabilités théoriques p. Pour cela, on introduit la *distance* du χ^2 entre lois sur $\{1,\dots,k\}$ qui est définie par :

$$D_{\chi^2}(p,q) := \sum_{l=1}^k \frac{(p_l-q_l)^2}{q_l}.$$

Attention : cette quantité n'est pas une vraie distance (en particulier elle n'est pas symétrique).

La loi de $D_{\chi^2}(\hat{p}^n,p)$ dépend en général de k, p et de n. Comme toutes ces quantités sont connues, rien n'empêche dans un cas particulier d'en approcher la loi par simulation. Il existe des tables pour le cas où p est la probabilité uniforme sur les k possibilités. Cependant, quand n est grand et après normalisation, la loi limite ne dépend plus que de K par le résultat suivant :

Theorème 1. *Soit p et q deux probabilités sur $\{1,\dots,k\}$ (avec $q_l > 0$ pour tout l), alors, quand la taille de l'échantillon n tend vers l'infini :*

$$nD_{\chi^2}(\hat{p}^n, q) = \sum_{l=1}^{k} \frac{(N_l^n - nq_l)^2}{nq_l} \begin{cases} \xrightarrow{\mathcal{L}} \chi^2(k-1) & si\ q = p, \\ \xrightarrow{p.s.} +\infty & si\ q \neq p. \end{cases}$$

►► Illustrer les deux conclusions du théorème précédent. Dans le cas où $p \neq q$, on pourra mettre en évidence la vitesse à laquelle $nD_{\chi^2}(\hat{p}^n, q)$ diverge en divisant par une fonction de n bien choisie (mais pas trop dure)...

2 Test d'adéquation à une loi discrète

Le théorème précédent permet de construire un test d'adéquation (de niveau asymptotique) à une loi de probabilité sur un ensemble fini, c'est-à-dire que l'on teste l'hypothèse H_0 : ■ $p = p^0$ ■ contre l'hypothèse H_1 : ■ $p \neq p^0$ ■.

Protocole du test. Soit $\alpha \in [0,1]$ le niveau de confiance souhaité et x_α le quantile $1-\alpha$ de la loi $\chi^2(k-1)$ (c'est-à-dire le réel x_α qui vérifie $\mathbb{P}(X \leq x_\alpha) = 1-\alpha$ où $\mathcal{L}(X) = \chi^2(k-1)$) :

- si $nD_{\chi^2}(\hat{p}^n, p^0) \leq x_\alpha$, on ne rejette pas H_0 au niveau α,
- si $nD_{\chi^2}(\hat{p}^n, p^0) > x_\alpha$, on rejette H_0 au niveau α,.

2.1 Préliminaires

►► Comment se débrouiller sans les tables ?

On souhaite donner une représentation de la puissance du test. Pour cela, on choisit p^0 égal à la loi uniforme sur $\{1,2,3,4\}$ et une taille d'échantillon $n = 1000$.

►► Tracer en fonction de ε variant de 0 à 1/4 la probabilité de rejet (estimée par simulation) de l'hypothèse H_0 lorsque la loi de l'échantillon est

$$p = (1/4, 1/4, 1/4 - \varepsilon, 1/4 + \varepsilon).$$

Dans la pratique, et surtout grâce à la puissance de calcul actuelle, on ne fait plus vraiment les tests de cette manière. On procède plutôt de la manière suivante : on calcule $nD_{\chi^2}(\hat{p}^n, p^0)$ puis on affiche la p-valeur associée qui est définie par

$$\mathbb{P}(X \geq nD_{\chi^2}(\hat{p}^n, p^0)),$$

où X suit une loi $\chi^2(k-1)$.

►► Comment interpréter ce nombre ?

2.2 Naissances

On a classé 10000 familles ayant exactement 4 enfants en fonction du nombre de garçons et l'on a obtenu les résultats suivants :

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

On se demande comment modéliser les naissances dans les familles nombreuses.

On propose dans un premier temps de supposer que les naissances sont indépendantes et la répartition garçon/fille équiprobable.

►► Quelle est alors la loi du nombre de garçons dans une fratrie de quatre enfants ? Mettre en place un test du χ^2 pour (in-)valider ce modèle. Quelle est la p-valeur ? Que doit-on en conclure ?

On relâche l'hypothèse de l'équirépartition.

►► Expliquer la façon de procéder. Quelle est la nouvelle p-valeur obtenue. Conclusion ?

►► Illustrer le théorème qui vous a permis de faire le test précédent (test d'adéquation à une famille de loi à un paramètre) et notamment la perte d'un degré de liberté pour la loi limite.

2.3 Étude du comportement asymptotique de la file M/M/1

On considère une file M/M/1 $(X_t)_{t\geq 0}$ récurrente positive ($\lambda < \mu$). On fait souvent l'hypothèse que le régime stationnaire est atteint très vite, ce qui permet de simplifier grandement les calculs. On souhaite valider cette approximation. Rappelons que la mesure invariante π est donnée par

$$\forall k \in \mathbb{N}, \quad \pi(k) = \left(\frac{\lambda}{\mu}\right)^k \left(1 - \frac{\lambda}{\mu}\right).$$

Voici une fonction Scilab qui permet de simuler une trajectoire de la file M/M/1 prenant comme paramètres λ, μ et l'instant final t. Elle donne aussi en sortie la matrice des temps de saut et des positions.

```
function a=MM1(l,m,t)
N=[0];
T=[0];
while (T($)<t)
    b=1/(l+m*(N($)>0));
    T=[T T($)+grand(1,1,"exp",b)];
    N=[N N($)+2*(rand(1,1)<l*b)-1];
end;
T($)=[];
N($)=[];
a=[T;N];
xbasc();plot2d2([T t],[N N($)+1],3);
endfunction;
```

Remarque 2. On souhaite utiliser le test d'adéquation du χ^2 pour illustrer le fait que la loi de X_t converge vers π quand t tend vers l'infini. Comme la loi de X_t est portée par $\mathbb{N}$ tout entier, il faut adapter un peu la méthode. L'idée est la suivante : on rassemble tous les entiers supérieurs ou égaux à un certain k_0 dans une même classe et on fait un test du χ^2 d'adéquation de la loi $X_t \wedge k_0$ à la loi

$$(\pi(0), \pi(1), \ldots, \pi(k_0 - 1), \pi([k_0, +\infty[)).$$

►► Écrire une fonction `test` qui prend en paramètres λ, μ, t et un entier p, génère p réalisations indépendantes de X_t, fait le test décrit ci-dessus pour $k_0 = 4$ au niveau $\alpha = 0.05$. Pour t grand, sur 1000 tentatives, combien de fois le test accepte H_0 ?

►► Adapter la fonction précédente en une fonction `multitest` pour tracer sur même graphique, en fonction de t, la probabilité pour qu'un échantillon de taille $p = 500$ passe le test[1].

2.4 Estimation et tests pour la matrice de transition d'une chaîne de Markov

On se donne une matrice de trnasition pour la chaîne de Markov sur deux points :

$$P = \begin{pmatrix} 1/3 & 2/3 \\ 3/4 & 1/4 \end{pmatrix}.$$

Étant donné une trajectoire $X_1, \ldots, X_n$ de la chaîne, on définit

$$N_i = \sum_{l=1}^{n} \mathbf{1}_{\{X_l=i\}} \quad \text{et} \quad N_{ij} = \sum_{l=1}^{n-1} \mathbf{1}_{\{X_l=i, X_{l+1}=j\}}.$$

►► Comment estimer les coefficients de la matrice à partir de l'observation d'une trajectoire de la chaîne ?
►► Illustrer par la simulation que

$$\sum_{i,j} \frac{(N_{ij}/N_i - P_{ij})^2}{N_{ij}/N_i} \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2(???).$$

Quel semble être le nombre de degrés de liberté ? Intrerprêter.
►► Comment pourrait-on utiliser ce résultat pour mettre en place un test répondant à la question suivante : la trajectoire que j'observe provient-elle de la matrice de transistion P_0 connue ?

3 Test d'indépendance de deux variables aléatoires discrètes

Soit (X, Y) un vecteur de deux variables aléatoires réelles discrètes à valeurs dans $\mathcal{X} \times \mathcal{Y}$ tel que $\mathrm{card}(\mathcal{X}) = a$ et $\mathrm{card}(\mathcal{Y}) = b$. On identifie donc $\mathbb{P}_{X,Y}$ à la matrice $P = (p_{ij}) \in \mathcal{M}_{ab}([0,1])$ définie par :

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

En notant les lois marginales

$$p_{i\bullet} = \sum_{j=1}^{b} p_{ij} = \mathbb{P}_X(\{i\}) = \mathbb{P}(X = i),$$

et

$$p_{\bullet j} = \sum_{i=1}^{a} p_{ij} = \mathbb{P}_Y(\{j\}) = \mathbb{P}(Y = j),$$

[1] On choisira quelques instants $t_1, \ldots, t_k$. Puis, pour chaque temps t_l, on fera passer le test ci-dessus à un certain nombre de p-échantillon et on comptera combien on réussit le test. Cette fonction risque d'être un peu gourmande en temps de calcul...

Si les deux variables aléatoires sont indépendantes alors, pour tout i et j, $p_{ij} = p_{i\bullet}p_{\bullet j}$. En pratique, on calcule une estimation de la distance du chi deux entre ces deux matrices comme il est expliqué dans les questions 3 et 4 suivantes. On sait que

$$d_n \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2((a-1)(b-1)),$$

s'il y a indépendance et que d_n tend p.s. vers $+\infty$ sinon. On utilise donc les quantiles de la loi du chi-deux pour déterminer la région d'acceptation de l'hypothèse d'indépendance

$$H_0 \ : \ \forall (i,j),\ p_{ij} = p_{i\bullet}p_{\bullet j} \quad \text{contre} \quad H_1 \ : \ \exists (i,j),\ p_{ij} \neq p_{i\bullet}p_{\bullet j}.$$

Le but des questions qui suivent est de montrer qu'il faut faire très attention à la taille de l'échantillon dont on dispose et que pour des valeurs trop faibles de n, assimilez d_n à une loi du chi-deux est très abusif.

Dans toute la suite, on pose $a = b = 5$ et on considère les deux variables aléatoires X et Y de loi respective sur $\{1,\dots,5\}$, (.1 .2 .3 .2 .2) et (.3 .4 .1 .1 .1). On note P' la matrice de terme général $p'_{ij} = p_{i\bullet}p_{\bullet j}$.

1. Générez un échantillon de $n = 500$ réalisations (X_k, Y_k) indépendantes de loi $\mathbb{P}_{X,Y} = P'$.
2. Calculez les effectifs $E = (e_{ij}) = (\text{card}\{k : (X_k, Y_k) = (i,j)\})$.
3. On définit le tableau de contingence $\hat{P} = E/n = (\hat{p}_{ij})$. Calculez la statistique du chi-deux mesurant la distance de la loi empirique au produit de ses propres lois marginales,

$$d_n = n\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{(\hat{p}_{ij} - \hat{p}_{i\bullet}\hat{p}_{\bullet j})^2}{\hat{p}_{i\bullet}\hat{p}_{\bullet j}} = \sum_{i=1}^{a}\sum_{j=1}^{b}\frac{(e_{ij} - e_{i\bullet}e_{\bullet j}/n)^2}{e_{i\bullet}e_{\bullet j}/n}.$$

4. Simuler $N = 1000$ réalisations indépendantes de d_n et illustrez le fait que la loi de d_n est proche d'une loi du chi-deux $\chi^2((a-1)(b-1))$.

Remarque. En pratique, on dit que le test est utilisable lorsque

$$n > \frac{10}{\min_{ij}(\min(p_{ij}, p'_{ij}))}.$$