Université des Sciences et Technologies de Lille
U.F.R. de Mathématiques Pures et Appliquées

IS Math314 **Année 2007**

Devoir n° 2, à rendre en T.D. la semaine du ? ? mai 2007

Ex 1. *La statistique du* χ^2

Pour tester l'hypothèse qu'une variable aléatoire X suit une loi discrète sur un ensemble fini $A := \{x_1, \dots, x_d\}$ de cardinal d, il est usuel de faire le test du χ^2 qui consiste à rejetter l'hypothèse nulle si la statistique dite du χ^2 prend une valeur trop élevée. Par exemple on peut tester ainsi que les faces d'un dé ont même probabilité d'apparition (dans ce cas X est la loi des points obtenus lors d'un lancer et on teste que cette loi est la loi uniforme sur $A = \{1, \dots, 6\}$).

Le modèle statistique sous-jacent est le suivant. Chaque loi discrète sur A est caractérisée par le vecteur $\theta = (q_1, \dots, q_d)$ des probabilités des x_i en sorte que $P_\theta(X = x_i) = q_i$ pour $i = 1, \dots, d$. L'ensemble des paramètres est ici

$$\Theta := \{\theta = (q_1, \dots, q_d) \in \mathbb{R}_+^{*d};\ q_1 + \cdots + q_d = 1\}.$$

On fixe un certain $p \in \Theta$, par exemple dans le cas du dé $p = (1/6, \dots, 1/6)$ et on considère l'hypothèse nulle

$$(\mathcal{H}_0) \ : \qquad \theta = p.$$

On dispose d'un échantillon $X_1, \dots, X_n$, *i.e.* de variables aléatoires X_j, $j = 1, \dots, n$ qui pour tout $\theta \in \Theta$, sont P_θ-mutuellement indépendantes et de même loi sous P_θ. Pour $i = 1, \dots, d$, on définit les v.a.

$$N_{n,i} := \sum_{j=1}^{n} \mathbf{1}_{\{X_j = x_i\}}.$$

Ainsi $N_{n,i}$ est le nombre d'obtentions de x_i dans l'échantillon. À partir des $N_{n,i}$, on peut calculer la statistique

$$T_n := \sum_{i=1}^{d} \frac{n}{p_i}\Big(\frac{N_{n,i}}{n} - p_i\Big)^2 = \sum_{i=1}^{d} \frac{(N_{n,i} - np_i)^2}{np_i}.$$

1) Vérifiez que

$$T_n = \sum_{i=1}^{d} \frac{N_{n,i}^2}{np_i} - n.$$

2) Quelle est la loi de $N_{n,i}$ sous P_θ ?

3) Quelle est la loi du *vecteur aléatoire* $N_n = (N_{n,1}, \ldots, N_{n,d})$ sous P_θ ?

4) Calculez $\mathbf{E}_\theta T_n$ pour $\theta = p$. Comparez avec l'espérance $\mathbf{E}(Y_1^2 + \cdots + Y_{d-1}^2)$, où les Y_i sont i.i.d. $\mathfrak{N}(0,1)$.

5) Montrez que pour tout $\theta \neq p$, T_n tend P_θ-presque sûrement vers $+\infty$.

6) En utilisant le théorème limite central dans $\mathbb{R}^d$, montrez que *sous* $(\mathcal{H}_0)$, T_n converge en loi vers la variable aléatoire $g(Z)$ donnée par

$$g(Z) = \frac{Z_1^2}{p_1} + \cdots + \frac{Z_d^2}{p_d}, \quad \text{où} \quad Z = (Z_1, \ldots, Z_d) \sim \mathfrak{N}(0, K),$$

la matrice de covariance $K = [K_{i,j}]$ ayant pour coefficients

$$K_{i,j} = p_i \delta_{i,j} - p_i p_j, \quad 1 \leq i, j \leq d.$$

7) Nous admettrons dans la suite que *sous* $(\mathcal{H}_0)$, la loi de $g(Z)$ est la même que celle de $Y_1^2 + \cdots + Y_{d-1}^2$ où les Y_i sont i.i.d. $\mathfrak{N}(0,1)$, autrement dit, que $g(Z)$ suit la loi du χ^2 (« loi du khi2 ») à $d-1$ degrés de liberté, notée $\chi^2(d-1)$. Vérifiez cette affirmation dans le cas particulier $d = 2$. Vous pourrez utiliser après l'avoir justifiée, l'égalité $Z_1 + Z_2 = 0$ p.s.

Ex 2. *Le test du χ^2*

Cet exercice propose une application élémentaire des résultats du précédent au test de l'hypothèse $(\mathcal{H}_0)$. On conserve toutes les notations introduites ci-dessus. Nous venons de voir que si $(\mathcal{H}_0)$ n'est pas vérifiée (donc si $\theta \neq p$), T_n tend p.s. vers $+\infty$, tandis que si $(\mathcal{H}_0)$ est vérifiée $(\theta = p)$, T_n a pour loi limite $\chi^2(d-1)$. En particulier si on se donne α proche de 0, on peut grâce à la table des quantiles de cette loi trouver un réel t_α tel que $P(T_n > t_\alpha) \simeq \alpha$. La conséquence pratique de ceci est que pour n grand, on aura tendance à observer de « grandes » valeurs de $T_n(\omega)$ si $(\mathcal{H}_0)$ est fausse et des valeurs « petites » si $(\mathcal{H}_0)$ est vraie. Dans ce contexte, on considèrera comme « grandes » les valeurs supérieures à t_α et comme « petites » les autres. Ceci nous conduit à la règle de décision suivante pour le test du khi2 de $(\mathcal{H}_0)$ au niveau α :

- si $T_n(\omega) \leq t_\alpha$, on accepte $(\mathcal{H}_0)$,
- si $T_n(\omega) > t_\alpha$, on la rejette.

La probabilité de rejetter $(\mathcal{H}_0)$ lorsqu'elle est vraie (erreur de première espèce) est donc α.

1) On a lancé 20 000 fois un dé et obtenu les résultats du tableau 1. On se demande

1	2	3	4	5	6
3 407	3 631	3 176	2 916	3 448	3 422

TAB. 1 – 20 000 lancers d'un dé (1)

si le dé est équilibré. Proposez une réponse en effectuant un test du khi2 au niveau $\alpha = 10^{-3}$.

1	2	3	4	5	6
3 259	3 403	3 292	3 276	3 373	3 397

TAB. 2 – 20 000 lancers d'un dé (2)

	Lisse	Ridé
Jaune	315	101
Vert	108	32

TAB. 3 – Pois de Mendel, couleur et forme

2) Reprendre la même question avec les données du tableau 2. On vous laisse libre du choix du niveau.

3) Lors d'une expérience, Mendel a observé la forme et la couleur d'un échantillon de 556 pois. Ses observations sont résumées dans le tableau 3. La seconde loi d'hérédité de Mendel prévoit que dans les conditions de cette expérience, les couplages de caractères ont les probabilités 9/16 pour jaune-lisse, 3/16 pour jaune-ridé, 3/16 pour vert-lisse et 1/16 pour vert-ridé. Les observations vous semblent-elles en accord avec cette théorie ?