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

IS Math314 **Année 2007**

Corrigé du Devoir n° 2

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, le test du χ^2 consiste à rejetter l'hypothèse nulle si la statistique dite du χ^2 prend une valeur trop élevée. 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 $p \in \Theta$ 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

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

1) Commençons par établir une formule plus simple pour T_n. En développant le carré $(N_{n,i} - np_i)^2$, en sommant sur i chacun des trois termes obtenus, il vient après simplifications :

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

Rappelons que $p_1 + \cdots + p_d = 1$ et remarquons que $N_{n,1} + \cdots + N_{n,d} = n$: il y a n épreuves donnant chacune comme résultat l'un des x_i, le nombre total de résultats obtenus, répétitions comprises, est donc n. En reportant ceci dans (2), on obtient

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

2) Pour tout $\theta = (q_1, \ldots, q_d) \in \Theta$, les $\mathbf{1}_{\{X_j = x_i\}}$ sont des v.a. P_θ indépendantes et de même loi de Bernoulli de paramètre q_i. Leur somme $N_{n,i}$ suit donc sous P_θ la loi binomiale de paramètre n et q_i.

3) Quelle est la loi du *vecteur aléatoire* $N_n = (N_{n,1}, \ldots, N_{n,d})$ sous P_θ ? En utilisant la définition de l'addition vectorielle dans $\mathbb{R}^d$, on voit que :

$$N_n = \left(\sum_{j=1}^{n} \mathbf{1}_{\{X_j = x_1\}}, \ldots, \sum_{j=1}^{n} \mathbf{1}_{\{X_j = x_d\}} \right) = \sum_{j=1}^{n} \left(\mathbf{1}_{\{X_j = x_1\}}, \ldots, \mathbf{1}_{\{X_j = x_d\}} \right) = \sum_{j=1}^{n} V_j,$$

où les vecteurs aléatoires V_j de $\mathbb{R}^d$ sont P_θ-indépendants et de même loi sous P_θ, définie comme suit. Les valeurs vectorielles possibles pour V_j sont les vecteurs $e_1, \ldots, e_d$ de la base canonique de $\mathbb{R}^d$ et $P_\theta(V_j = e_i) = P_\theta(X_j = x_i) = q_i$. En effet l'égalité entre vecteurs $V_j(\omega) = e_i$ signifie que toutes les composantes de $V_j(\omega)$ d'indice $k \neq i$ sont nulles, autrement dit que $\mathbf{1}_{\{X_j(\omega) = x_k\}} = 0$ ou encore $X_j(\omega) \neq x_k$, *et* que la i^{e} composante de $V_j(\omega)$ vaut 1, c'est-à-dire $\mathbf{1}_{\{X_j(\omega) = x_i\}} = 1$ ou encore $X_j(\omega) = x_i$. Il y a donc équivalence entre $V_j(\omega) = e_i$ et $X_j(\omega) = x_i$. Ce raisonnement étant valable pour tout $\omega \in \Omega$ établit l'égalité d'évènements $\{V_j = e_i\} = \{X_j = x_i\}$, d'où l'égalité de leurs P_θ probabilités, quel que soit θ. La loi du vecteur aléatoire N_n sous P_θ est donc la loi multinomiale de paramètres n et $\theta = q = (q_1, \ldots, q_d)$.

4) Calculons $\mathbf{E}_\theta T_n$ pour $\theta = p$. On remarque que si $\theta = p$, $\mathbf{E}_\theta(N_{n,i} - np_i)^2$ est la variance sous P_θ de $N_{n,i}$ car cette v.a. suit alors la loi binomiale de paramètres n et p_i, d'espérance np_i. Comme cette variance vaut $np_i(1 - p_i)$, on en déduit en utilisant la deuxième égalité dans (1) et la linéarité de l'espérance que

$$\mathbf{E}_\theta T_n = \sum_{i=1}^{d} \frac{\mathbf{E}_\theta(N_{n,i} - np_i)^2}{np_i} = \sum_{i=1}^{d} \frac{\mathrm{Var}_\theta(N_{n,i})}{np_i} = \sum_{i=1}^{d} (1 - p_i) = d - \sum_{i=1}^{d} p_i = d - 1.$$

On voit ainsi que T_n a même espérance que $Y_1^2 + \cdots + Y_{d-1}^2$, où les Y_i sont i.i.d. $\mathfrak{N}(0, 1)$. En effet

$$\mathbf{E}(Y_1^2 + \cdots + Y_{d-1}^2) = (d - 1)\mathbf{E}Y_1^2 = (d - 1)\,\mathrm{Var}\, Y_1 = d - 1.$$

5) Montrons que pour tout $\theta \neq p$, T_n tend P_θ-presque sûrement vers $+\infty$. Si $\theta = (q_1, \ldots, q_d) \neq p = (p_1, \ldots, p_d)$, il y a *au moins un* indice $k \in \{1, \ldots, d\}$ tel que $q_k \neq p_k$. Choisissons donc un tel k et minorons T_n par le terme de rang k dans le second membre de (1) :

$$T_n \geq \frac{n}{p_k} \left(\frac{N_{n,k}}{n} - p_k \right)^2$$

Par la loi forte des grands nombres appliquée sur l'espace probabilisé $(\Omega, \mathcal{F}, P_\theta)$, $N_{n,k}/n$ converge P_θ-p.s. quand n tend vers l'infini vers $q_k = \mathbf{E}_\theta \mathbf{1}_{\{X_1 = x_k\}}$. Ceci s'écrit encore $P_\theta(\Omega') = 1$ en notant Ω' l'ensemble des $\omega \in \Omega$ tels que la suite de nombres réels $\big(N_{n,k}(\omega)/n\big)_{n\geq 1}$ converge vers q_k. On a donc

$$\forall \omega \in \Omega', \quad T_n(\omega) \geq \frac{n}{p_k}\Big(\frac{N_{n,k}(\omega)}{n} - p_k\Big)^2$$

et ce minorant apparaît comme le produit de n/p_k qui tend vers $+\infty$ avec n et d'un facteur qui tend vers $(q_k - p_k)^2$, constante *strictement* positive puisque $q_k \neq p_k$. Le minorant tend donc vers $+\infty$ et il en va de même pour $T_n(\omega)$. Ce raisonnement étant valide pour tout $\omega \in \Omega'$ et puisque $P_\theta(\Omega') = 1$, on conclut que T_n tend P_θ-p.s. vers $+\infty$.

6) Nous venons de voir le comportement asymptotique de T_n lorsque $\theta \neq p$. Le raisonnement ci-dessus n'est plus valable pour $\theta = p$, puisqu'en appliquant la LFGN, pour tout $k \in \{1, \ldots, d\}$, $(n/p_k)\big(N_{n,k}/n - p_k\big)^2$ présente p.s. quand n tend vers $+\infty$ une forme indéterminée du type « $\infty \times 0$ ».

Lorsque $\theta = p$, le vecteur espérance de $n^{-1}N_n$ est p. Le théorème limite central vectoriel appliqué à la loi multinomiale nous donne alors :

$$S_n^* := \sqrt{n}\Big(\frac{1}{n}N_n - p\Big) \xrightarrow[n\to+\infty]{\text{loi}} Z, \tag{3}$$

où $Z = (Z_1, \ldots, Z_d)$ est un vecteur aléatoire gaussien de loi $\mathfrak{N}(0, K)$, la matrice de covariance $K = [K_{i,j}]$ ayant pour coefficients :

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

Introduisons la fonction continue

$$g : \mathbb{R}^d \to \mathbb{R}, \quad x = (x_1, \ldots, x_d) \longmapsto \frac{x_1^2}{p_1} + \cdots + \frac{x_d^2}{p_d}.$$

Par conservation de la convergence en loi par image continue, on déduit de (3) que

$$g(S_n^*) \xrightarrow[n\to+\infty]{\text{loi}} g(Z) = \frac{Z_1^2}{p_1} + \cdots + \frac{Z_d^2}{p_d}.$$

On explicite $g(S_n^*)$ en notant que la i^{e} composante du vecteur S_n^* est

$$S_{n,i}^* = \sqrt{n}\Big(\frac{1}{n}N_{n,i} - p_i\Big),$$

d'où

$$g(S_n^*) = \sum_{i=1}^d \frac{(S_{n,i}^*)^2}{p_i} = \sum_{i=1}^d \frac{n}{p_i}\Big(\frac{1}{n}N_{n,i} - p_i\Big)^2 = T_n.$$

Nous avons donc établi sous $(\mathcal{H}_0)$ la convergence en loi de T_n vers $g(Z)$.

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 $\chi^2(d-1)$. Nous nous contenterons de vérifier cette affirmation dans le cas particulier $d=2$. Admettons provisoirement que $Z_1+Z_2=0$ p.s., ce qui s'écrit aussi $Z_2=-Z_1$ p.s. Alors

$$g(Z)=\frac{Z_1^2}{p_1}+\frac{Z_2^2}{p_2}=\frac{p_1+p_2}{p_1p_2}Z_1^2=\frac{Z_1^2}{p_1(1-p_1)},$$

parce que $p_1+p_2=1$. D'autre part Z_1 est gaussienne d'espérance nulle et de variance $K_{1,1}=p_1-p_1^2=p_1(1-p_1)$. La variable aléatoire $Y_1:=p_1^{-1/2}(1-p_1)^{-1/2}Z_1$ est donc gaussienne standard et $g(Z)=Y_1^2$. On vérifie ainsi que $g(Z)$ suit la loi $\chi^2(1)$, c'est bien la loi $\chi^2(d-1)$ pour $d=2$.

Il nous reste à justifier la nullité presque-sûre de Z_1+Z_2. En utilisant (3) appliquée avec $d=2$ et la forme linéaire continue $\ell:\mathbb{R}^2\to\mathbb{R}$, $(x_1,x_2)\mapsto x_1+x_2$, on obtient :

$$R_n:=\ell(Z)=\sqrt{n}\Big(\frac{1}{n}N_{n,1}-p_1\Big)+\sqrt{n}\Big(\frac{1}{n}N_{n,2}-p_2\Big)\xrightarrow[n\to+\infty]{\text{loi}}Z_1+Z_2.$$

En revenant à la définition de la convergence en loi, on a donc

$$\forall h:\mathbb{R}\to\mathbb{R}\text{ continue bornée},\quad \mathbf{E}h(R_n)\xrightarrow[n\to+\infty]{}\mathbf{E}h(Z_1+Z_2). \tag{4}$$

D'autre part,

$$R_n=\sqrt{n}\Big(\frac{N_{n,1}+N_{n,2}}{n}-(p_1+p_2)\Big)=\sqrt{n}\Big(\frac{n}{n}-1\Big)=0.$$

Ainsi R_n est simplement la v.a. nulle et donc $\mathbf{E}h(R_n)=\mathbf{E}h(0)=h(0)$. En reportant ceci dans (4), on voit que pour toute fonction h continue bornée sur $\mathbb{R}$, $\mathbf{E}h(Z_1+Z_2)=h(0)$, ce qui implique que Z_1+Z_2 a même loi (sous $(\mathcal{H}_0)$, c'est-à-dire sous P_p) que la v.a. nulle, d'où $P_p(Z_1+Z_2=0)=1$. En effet la famille des fonctions continues bornées est assez riche pour caractériser les lois de probabilité sur $\mathbb{R}$, cf. la caractérisation des lois par les h-moments dans le cours d'IPÉ.

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.

Pour cela, nous calculons T_n. Ici $n = 20\,000$, $p = (1/6, 1/6, 1/6, 1/6, 1/6, 1/6)$. Le tableau nous donne les $N_{n,i}$ observés pour $i = 1, \ldots, 6$. On trouve

$$T_n(\omega) = 94{,}189.$$

Ensuite on regarde dans la table des quantiles du χ^2, pour 5 degrés de liberté et on voit que la valeur t_α telle que $P_p(T_n > t_\alpha) = \alpha$ est pour $\alpha = 10^{-3}$, $t_\alpha = 20{,}515$, en négligeant l'erreur d'approximation due à la convergence de la loi de T_n vers $\chi^2(5)$. La valeur observée $T_n(\omega) = 94{,}189$ étant largement supérieure à cette valeur critique, nous rejettons au niveau 10^{-3} l'hypothèse d'équilibre du dé[1].

2) En reprenant le calcul de $T_n(\omega)$ à partir des données du tableau 2, on obtient

$$T_n(\omega) = 6{,}304.$$

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)

La lecture de la table des quantiles pour $\chi^2(5)$ nous donne $t_{0,1} = 9{,}236$. On peut donc accepter l'hypothèse d'équilibre du dé au niveau 10% (ainsi bien sûr qu'à tout niveau $\alpha < 0{,}1$). La même table nous donne $t_{0,5} = 4{,}351$, on pourrait donc aussi décider de rejetter $(\mathcal{H}_0)$ au niveau 0,5, mais ce choix du niveau et du rejet ne serait pas très pertinent car il y aurait alors une chance sur deux d'avoir pris la mauvaise décision.

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 semblent-elles en accord avec cette théorie ?

Ici on a une loi discrète sur une ensemble de cardinal $d = 4$. Il est commode de réarranger les données comme dans le tableau 4. On a $n = 315 + 101 + 108 + 32 = 556$.

	Lisse	Ridé
Jaune	315	101
Vert	108	32

TAB. 3 – Pois de Mendel, couleur et forme

Couplage	J-L	J-R	V-L	V-R
i	1	2	3	4
Effectif $N_{n,i}$	315	101	108	32
p_i	$\frac{9}{16}$	$\frac{3}{16}$	$\frac{3}{16}$	$\frac{1}{16}$

TAB. 4 – Pois de Mendel, effectifs observés et probabilités théoriques

On souhaite tester l'hypothèse nulle $\theta = p = (9/16, 3/16, 3/16, 1/16)$. On obtient après calcul $T_n(\omega) = 0{,}47$. La lecture de la table des quantiles de $\chi^2(3)$ nous indique que l'on peut accepter l'hypothèse d'accord des données avec la théorie de Mendel, et ce au niveau 0,9 (donc aussi à tout niveau $\alpha < 0{,}9$). En effet $t_{0,9} = 0{,}584$. D'autre part comme $t_{0,95} = 0{,}352$, on pourrait aussi décider de rejetter l'hypothèse au niveau 0,95, mais avec un risque de se tromper supérieur à 90%.

[1]La probabilité que nous ayions ainsi pris la mauvaise décision ne devrait pas excéder 10^{-3}.