UNIVERSITÉ DE NICE SOPHIA-ANTIPOLIS | Année Universitaire 2015/2016
L2 MASS | **Statistique**
TD de Statistique

Contrôle continu 2
Correction du sujet B

Correction 1. (1) On commence par déterminer la valeur de $\mathbb{P}(X_1 = 1)$. Comme $\mathbb{P}(X_1 = -1) + \mathbb{P}(X_1 = 0) + \mathbb{P}(X_1 = 1) = 1$, on a donc $\mathbb{P}(X_1 = 1) = 1 - \gamma$. Ensuite,

$$\mathbb{E}[X_1] = -1 \times (\gamma(1-\gamma)) + 1 \times (1-\gamma) = (1-\gamma)^2$$

et

$$\mathbb{E}\left[X_1^2\right] = (-1)^2 \times (\gamma(1-\gamma)) + 1^2 \times (1-\gamma) = 1 - \gamma^2.$$

(2) Les deux fonctions g_1 et g_2 vont de $]0,1[$ dans $]0,1[$. Soient x et y dans $]0,1[$ tels que $g_1(x) = y$. On raisonne par équivalences :

$$g_1(x) = y \Leftrightarrow (1-x)^2 = y \Leftrightarrow 1 - x = \sqrt{y} \text{ (car } 1 - x \geq 0 \text{ et } y \geq 0) \Leftrightarrow x = 1 - \sqrt{y}.$$

Ainsi g_1 est inversible sur $]0,1[$ et son inverse et $g_1^{-1} :]0,1[\to]0,1[$ qui à y associe $1 - \sqrt{y}$. Soient x et y dans $]0,1[$ tels que $g_2(x) = y$. On raisonne par équivalences :

$$g_2(x) = y \Leftrightarrow 1 - x^2 = y \Leftrightarrow x^2 = 1 - y \Leftrightarrow x = \sqrt{1-y} \text{ (car } x \geq 0 \text{ et } 1 - y \geq 0).$$

Ainsi g_2 est inversible sur $]0,1[$ et son inverse et $g_2^{-1} :]0,1[\to]0,1[$ qui à y associe $\sqrt{1-y}$.

On a $\mathbb{E}[X_1] = g_1(\gamma)$, donc l'estimateur donné par la méthode des moments (en utilisant le moment d'ordre 1) de γ est $B_1 = g_1^{-1}(\frac{1}{n}\sum_{i=1}^n X_i) = 1 - \sqrt{\frac{1}{n}\sum_{i=1}^n X_i}$. **Remarque :** la variable aléatoire $\frac{1}{n}\sum_{i=1}^n X_i$ peut prendre des valeurs négatives et dans ce cas là, l'estimateur que l'on a donné n'est pas bien défini. Cependant, on sait que $\sum_{i=1}^n X_i$ est proche de $\mathbb{E}[X_1] = (1-\gamma)^2 > 0$ quand n est grand. On peut donc espérer que ce cas là ne se produise pas souvent quand l'échantillon est assez grand.

On a $\mathbb{E}\left[X_1^2\right] = g_2(\gamma)$, donc l'estimateur donné par la méthode des moments (en utilisant le moment d'ordre 2) de γ est $B_2 = g_2^{-1}(\frac{1}{n}\sum_{i=1}^n X_i^2) = \sqrt{1 - \frac{1}{n}\sum_{i=1}^n X_i^2}$. **Remarque :** L'estimateur B_2 est bien défini dans tous les cas. En effet, $\frac{1}{n}\sum_{i=1}^n X_i^2 \leq 1$ en tant que variable aléatoire.

(3) La vraisemblance de l'échantillon est

$$L(x_1, \ldots, x_n, \gamma) = \prod_{i=1}^n \mathbb{P}(X_i = x_i) = [\gamma(1-\gamma)]^{n_{-1}} \gamma^{2n_0} (1-\gamma)^{n_1} = \gamma^{n_{-1}+2n_0}(1-\gamma)^{n_{-1}+n_1}.$$

(4) On peut considérer $\log(L(x_1, \ldots, x_n, \gamma)) = (n_{-1} + 2n_0)\log(\gamma) + (n_{-1} + n_1)\log(1-\gamma)$. Sa dérivée par rapport à γ est $\frac{n_{-1}+2n_0}{\gamma} - \frac{n_{-1}+n_1}{1-\gamma}$. Elle est positive si et seulement si $\gamma \leq \frac{n_{-1}+2n_0}{2n_{-1}+2n_0+n_1}$. Donc l'estimateur par maximum de vraisemblance est $B_{MV} = \frac{N_{-1}+2N_0}{2N_{-1}+2N_0+N_1}$ où N_{-1}, N_0 et N_1 sont les variables aléatoires correspondantes aux réalisations n_{-1}, n_0 et n_1. C'est à dire, $N_{-1} = \sum_{i=1}^n \mathbf{1}_{X_i=-1}$, $N_0 = \sum_{i=1}^n \mathbf{1}_{X_i=0}$ et $N_1 = \sum_{i=1}^n \mathbf{1}_{X_i=1}$.

Correction 2. (1) On effectue un sondage au sein d'une population de grande taille. Ainsi, la probabilité qu'une personne de notre échantillon ait été touché par la grippe est q (la proportion dans la population totale). Comme $X_1 = 1$ si la première personne interrogée a été touchée par la grippe (ce qui arrive avec probabilité q), la loi de X_1 est la loi de Bernoulli de paramètre q. **Remarque :** Ne surtout pas écrire $q = 0,6$. Ici, q est inconnu et on cherche à l'estimer.

(2) L'estimateur classique de la proportion théorique q est la proportion empirique qui s'écrit $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$ avec les notations de l'énoncé et donne l'estimation $\hat{q} = 0,6$. **Remarque :** On différencie donc la variable aléatoire $\bar{X}_n$ de sa réalisation $\hat{q}$ mesurée par l'enquêteur. De plus, il ne faut surtout pas écrire $q = \bar{X}_n$ ou $q = 0,6$.

(3) Les variables $(X_1, \ldots, X_n)$ forment un échantillon de loi $\mathcal{B}(q)$. Par le Théorème Central Limite ($n \geq 30$), on peut approximer la variable aléatoire $\bar{X}_n - q$ par la variable $Y \sim \mathcal{N}(0, \sqrt{\frac{q(1-q)}{n}})$.

(4) On commence par centrer et réduire nos variables. On sait que $\bar{X}_n - q$ peut être approximer par Y donc $\sqrt{n}\frac{\bar{X}_n - q}{\sqrt{q(1-q)}}$ peut être approximer par $\sqrt{n}\frac{Y}{\sqrt{q(1-q)}} \sim \mathcal{N}(0,1)$. Comme $F(z_\alpha) = 1 - \alpha/2$, on sait que $\mathbb{P}\left(-z_\alpha \leq \sqrt{n}\frac{Y}{\sqrt{q(1-q)}} \leq z_\alpha\right) = 1 - \alpha$ par symétrie de la loi normale. Or, l'approximation nous dit que

$$\mathbb{P}\left(-z_\alpha \leq \sqrt{n}\frac{\bar{X}_n - q}{\sqrt{q(1-q)}} \leq z_\alpha\right) \approx \mathbb{P}\left(-z_\alpha \leq \sqrt{n}\frac{Y}{\sqrt{q(1-q)}} \leq z_\alpha\right)$$

donc

$$\mathbb{P}\left(\bar{X}_n - z_\alpha\sqrt{\frac{q(1-q)}{n}} \leq q \leq \bar{X}_n + z_\alpha\sqrt{\frac{q(1-q)}{n}}\right) \approx 1 - \alpha.$$

(5) La réponse à la question précédente ne permet pas de donner directement un intervalle de fluctuation pour q. En effet, les bornes de l'encadrement autour de q dépendent de q. On connait plusieurs techniques pour contourner ce problème : méthode de l'ellipse, méthode par excès, ou bien le remplacement de q par son estimation $\hat{q}$ justifié par le Lemme de Slutsky.

Correction 3. On considère l'échantillon probabiliste $(X_1, \cdots, X_n)$, $n = 4$ de loi mère $\mathcal{N}(\theta, \sigma)$. Cette échantillon est la modélisation aléatoire des tailles des 4 nouveau-nés données dans l'enoncé.

(1) L'estimateur classique de la moyenne théorique θ est $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$ et donne l'estimation $\hat{\theta} = 50$cm. **Remarque :** On différencie donc la variable aléatoire $\bar{X}_n$ de sa réalisation $\hat{\theta}$ mesurée par l'infirmier. De plus, il ne faut surtout pas écrire $\theta = \bar{X}_n$ ou $\theta = 50$cm.

(2) Soit α dans $]0,1[$. **Remarque :** Comme les variables aléatoires X_i sont i.i.d. de loi $\mathcal{N}(\theta, \sigma)$, on sait que :

$$Z_n := \sqrt{n}\frac{\bar{X}_n - \theta}{\sigma} \sim \mathcal{N}(0,1).$$

Cependant, cette variable ne nous est d'aucune utilité compte tenu du fait que l'on ne connait pas σ.

Posons

$$T_n := \sqrt{n}\frac{\bar{X}_n - \theta}{\sqrt{V_n}},$$

où $V_n = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$. D'après le cours, la variable T_n suit une loi de Student de paramètre $n-1$. On note ϕ la fonction de répartition de la loi de Student de paramètre

$n-1$. Soit t_α tel que $\phi(t_\alpha) = 1-\alpha/2$ (**Remarque :** en pratique, on lit t_α sur la table, mais il faut dire dès cette question quel est le lien entre α et t_α). On a donc, par symétrie de la loi de Student,

$$\mathbb{P}\left(-t_\alpha \leq T_n \leq t_\alpha\right) = 1-\alpha,$$

et donc

$$\mathbb{P}\left(\bar{X}_n - \frac{t_\alpha\sqrt{V_n}}{\sqrt{n}} \leq \theta \leq \bar{X}_n + \frac{t_\alpha\sqrt{V_n}}{\sqrt{n}}\right) = 1-\alpha.$$

Ainsi, $\left[\bar{X}_n - \frac{t_\alpha\sqrt{V_n}}{\sqrt{n}}, \bar{X}_n + \frac{t_\alpha\sqrt{V_n}}{\sqrt{n}}\right]$ est un intervalle de fluctuation pour θ au niveau de confiance $1-\alpha$.

(3) Un niveau de confiance de 95% correspond à $\alpha = 0.05$. Sur la table de Student de paramètre égal à 3, on lit $t_{0.05} = 3.182$. De plus, on calcule la réalisation de la moyenne empirique donnée par l'échantillon, $\bar{x}_n = \hat{\theta} = 50$ ainsi que la réalisation de la variance empirique donnée par l'échantillon, $v_n = 0.313$. Ainsi, $[49,11; 50,89]$ est un intervalle de confiance pour θ au niveau 95%.

(4) L'intervalle de fluctuation que l'on a trouvé à la question 2 est centré sur la moyenne empirique. Notons ℓ sa longueur. On a $\ell = 2\frac{t_\alpha\sqrt{V_n}}{\sqrt{n}}$, et on en déduit $t_\alpha = \frac{l\sqrt{n}}{2\sqrt{V_n}}$. Avec $n=4$, $v_n = 0.313$ et $\ell = 1.193$, on trouve $t_\alpha = 2,131$. On lit sur la table de la loi de Student de paramètre égal à 3, $\phi(2,131) \approx 0,95$. Or, on a la relation $\phi(t_\alpha) = 1-\alpha/2$, donc on en déduit $\alpha \approx 0,1$. Ainsi, le niveau de confiance d'un intervalle de fluctuation de longueur $1,193$cm centré en la moyenne empirique est approximativement de 90%.

Correction 4. (1) Comme les variables aléatoires X_i sont i.i.d. de loi $\mathcal{N}(m,s)$, on sait que les variables $\frac{nV_n^*}{s^2}$ et $\frac{(n-1)V_n}{s^2}$ suivent respectivement une loi du khi-deux à n degrés de libertés et une loi du khi-deux à $n-1$ degrés de libertés. On peut le justifier simplement pour $\frac{nV_n^*}{s^2}$. En effet, $\frac{nV_n^*}{s^2} = \sum_{i=1}^n (\frac{X_i - m}{s})^2$ est la somme des carrés de n variables gaussiennes $\mathcal{N}(0,1)$ indépendantes. Donc, par définition de la loi du khi-deux, $\frac{nV_n^*}{s^2} \sim \chi^2(n)$.

(2) Soit α dans $]0,1[$. Comme la moyenne m est connue, on va utiliser la variable $\frac{nV_n^*}{s^2}$. D'après la question 1, $\frac{nV_n^*}{s^2} \sim \chi^2(n)$. On note Φ la fonction de répartition de la loi $\chi^2(n)$. Soient u_α et t_α tels que $\Phi(t_\alpha) = \alpha/2$ et $\Phi(t_\alpha) = 1-\alpha/2$ (**Remarque :** en pratique, on lit u_α et t_α sur la table, mais il faut dire dès cette question quels sont les liens entre α et (u_α, t_α)). Ainsi, on a

$$\mathbb{P}\left(u_\alpha \leq \frac{nV_n^*}{s^2} \leq t_\alpha\right) = 1-\alpha,$$

et donc

$$\mathbb{P}\left(\frac{nV_n^*}{t_\alpha} \leq s^2 \leq \frac{nV_n^*}{u_\alpha}\right) = 1-\alpha.$$

Ainsi, $\left[\frac{nV_n^*}{t_\alpha}, \frac{nV_n^*}{u_\alpha}\right]$ est un intervalle de fluctuation pour s^2 au niveau de confiance $1-\alpha$.

(3) Un niveau de confiance de 90% correspond à $\alpha = 0.1$. Sur la table de la loi du khi-deux, on lit $u_{0.1} = 1,145$ et $t_{0.1} = 11,07$. De plus, on a $m = 0$ et on calcule la réalisation de V_n^* donnée par l'échantillon, $v_n^* = 0,664$. Ainsi, $[0,30; 2,90]$ est un intervalle de confiance au niveau 90%.