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

L3 **Année 2011**

Initiation à la Statistique
Corrigé du D.S. du 25 mars 2011

Ex 1. *Intervalle de confiance*

Le tableau ci-dessous donne un 100-échantillon observé d'une loi de Bernoulli de paramètre inconnu p.

1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	1	0	0	0	1	1	1	0	0	0	0	1	0	0	0	0	0
1	0	0	1	1	0	0	0	0	0	1	0	0	0	1	1	1	0	1	1
0	0	1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0

1) Si $S_n = X_1 + \cdots + X_n$ où les X_i sont indépendantes de même loi de Bernoulli et de paramètre p, l'intervalle de confiance pour p, au niveau 95%, bâti sur l'échantillon observé $X_1(\omega), \dots, X_n(\omega)$ par la méthode avec variance majorée est :

$$I(\omega) = \left[\frac{S_n(\omega)}{n} - \frac{1{,}96}{2\sqrt{n}}, \frac{S_n(\omega)}{n} + \frac{1{,}96}{2\sqrt{n}}\right]. \tag{1}$$

Ici $n = 100$ et les valeurs observées $X_1(\omega), \dots, X_{100}(\omega)$ sont données dans le tableau ci-dessus. Pour calculer $S_{100}(\omega)$, il suffit de compter le nombre de 1 dans ce tableau : $S_{100}(\omega) = 23$, d'où

$$I(\omega) = [0{,}132; 0{,}328].$$

Pour justifier la formule employée, on rappelle qu'en posant

$$S_n^* := \sqrt{\frac{n}{p(1-p)}}\left(\frac{S_n}{n} - p\right),$$

le théorème de de Moivre-Laplace nous donne pour tout $t > 0$,

$$P\left(-t \leqslant S_n^* \leqslant t\right) \xrightarrow[n\to+\infty]{} \Phi(t) - \Phi(-t) = 2\Phi(t) - 1,$$

où Φ désigne la fonction de répartition de la loi normale $\mathfrak{N}(0,1)$. Ceci peut se réécrire :

$$P\left(-t \leqslant S_n^* \leqslant t\right) = 2\Phi(t) - 1 + \varepsilon_n,$$

où l'erreur d'approximation gaussienne ε_n tend vers 0 quand n tend vers l'infini. On note ensuite que l'encadrement $-t \leqslant S_n^* \leqslant t$ se résout en un encadrement de p :

$$-t \leqslant S_n^* \leqslant t \iff \frac{S_n}{n} - \frac{t\sqrt{p(1-p)}}{\sqrt{n}} \leqslant p \leqslant \frac{S_n}{n} + \frac{t\sqrt{p(1-p)}}{\sqrt{n}}.$$

Cet encadrement ne fournit pas un intervalle de confiance, car les bornes dépendent du paramètre inconnu p. Pour y rémédier, on *élargit* cet encadrement en remarquant que $p(1-p)$ est maximal pour $p = 1/2$, d'où

$$P\left(\frac{S_n}{n} - \frac{t}{2\sqrt{n}} \leqslant p \leqslant \frac{S_n}{n} + \frac{t}{2\sqrt{n}}\right) \geqslant P\left(\frac{S_n}{n} - \frac{t\sqrt{p(1-p)}}{\sqrt{n}} \leqslant p \leqslant \frac{S_n}{n} + \frac{t\sqrt{p(1-p)}}{\sqrt{n}}\right).$$

On en déduit que

$$P\left(\frac{S_n}{n} - \frac{t}{2\sqrt{n}} \leqslant p \leqslant \frac{S_n}{n} + \frac{t}{2\sqrt{n}}\right) \geqslant 2\Phi(t) - 1 + \varepsilon_n.$$

En particulier on peut appliquer ceci avec $t = 1{,}96$ pour lequel $2\Phi(t) - 1 = 0{,}95$. En négligeant l'erreur d'approximation ε_{100}, on voit que la probabilité que l'intervalle aléatoire I défini par (1) contienne p est d'au moins 0,95.

Ex 2. *Dé*

On lance 3 600 fois un dé équilibré et on note nombre X de « cinq » obtenus. On cherche une valeur approchée de $P(578 \leqslant X \leqslant 622)$. Posons pour $i \in [\![1, 3600]\!]$,

$$A_i = \{\text{obtention d'un cinq au } i^{\text{e}} \text{ lancer}\}, \quad Y_i = \mathbf{1}_{A_i}.$$

Avec ces notations, il est clair que

$$X = \sum_{i=1}^{3\,600} Y_i.$$

Les A_i sont indépendants et de même probabilité $p = \frac{1}{6}$, les Y_i sont donc des variable aléatoires de Bernoulli indépendantes et de paramètre $p = \frac{1}{6}$. On en déduit que X suit la loi binomiale de paramètres $n = 3\,600$ et $p = \frac{1}{6}$. En particulier :

$$P(578 \leqslant X \leqslant 622) = \sum_{k=578}^{622} C_{3\,600}^k \left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{3\,600-k}.$$

Le calcul exact de cette probabilité n'est pas possible avec les moyens usuels.

On va donc en donner une valeur approchée en utilisant le théorème de de Moivre Laplace qui nous permet d'approximer la loi de

$$X^* := \frac{X - \mathbf{E}X}{\sqrt{\operatorname{Var} X}}$$

par la loi normale $\mathfrak{N}(0,1)$, puisque X est la somme d'un grand nombre de variables aléatoires de Bernoulli indépendantes et de même paramètre. En particulier pour tout $t > 0$,

$$P(-t \leqslant X^* \leqslant t) \simeq 2\Phi(t) - 1, \tag{2}$$

où Φ désigne la fonction de répartition de la loi $\mathfrak{N}(0,1)$. L'espérance et la variance de la variable aléatoire de loi binomiale X sont données par :

$$\mathbf{E}X = 3\,600 \times \frac{1}{6} = 600, \quad \operatorname{Var} X = 3\,600 \times \frac{1}{6} \times \frac{5}{6} = 500,$$

d'où

$$X^* = \frac{X - 600}{10\sqrt{5}}.$$

On remarque ensuite que

$$578 \leqslant X \leqslant 622 \;\Leftrightarrow\; \frac{578 - 600}{10\sqrt{5}} \leqslant X^* \leqslant \frac{622 - 600}{10\sqrt{5}} \;\Leftrightarrow\; -\frac{2{,}2}{\sqrt{5}} \leqslant X^* \leqslant \frac{2{,}2}{\sqrt{5}}.$$

En appliquant (2) avec $t = \frac{2{,}2}{\sqrt{5}}$, on en déduit que :

$$P(578 \leqslant X \leqslant 622) \simeq 2\Phi\left(\frac{2{,}2}{\sqrt{5}}\right) - 1 \simeq 2\Phi(0{,}984) - 1.$$

Ici l'auteur du sujet (et du corrigé) présente ses excuses aux étudiants pour avoir, par souci d'économie, inclus dans l'énoncé un fragment de la table de Φ, suffisant pour les exercices 1 et 3, mais pas pour achever le calcul numérique ci-dessus. La notation des copies tient évidemment compte de cette erreur.

Voici comment compléter à partir de l'extrait de table de Φ suivant :

x	0,90	0,91	0,92	0,93	0,94	0,95	0,96	0,97	0,98	0,99
$\Phi(x)$	0,815 9	0,818 6	0,821 2	0,823 8	0,826 4	0,828 9	0,831 5	0,834 0	0,836 5	0,838 9

La table donne $\Phi(0{,}98) = 0{,}836\,5$ et $\Phi(0{,}99) = 0{,}838\,9$. On calcule une valeur approchée de $\Phi(0{,}984)$ par interpolation linéaire.

$$\Phi(0{,}984) \simeq \Phi(0{,}98) + \frac{0{,}984 - 0{,}98}{0{,}99 - 0{,}98}\big(\Phi(0{,}99) - \Phi(0{,}98)\big) = 0{,}836\,5 + 0{,}4 \times 0{,}002\,4 = 0{,}837\,46.$$

Et finalement,

$$P(578 \leqslant X \leqslant 622) \simeq 0{,}674\,92 \simeq 0{,}675.$$

Ex 3. *Estimation du paramètre d'une loi géométrique*

La loi géométrique de paramètre $\theta \in]0,1[$ est celle du temps d'attente du *premier* succès dans une suite d'épreuves répétées indépendantes avec pour chaque épreuve probabilité de succès θ. Une variable aléatoire X sur l'espace probabilisé $(\Omega, \mathcal{F}, P_\theta)$ suit cette loi si

$$\forall k \in \mathbb{N}^*, \quad P_\theta(X = k) = (1-\theta)^{k-1}\theta. \tag{3}$$

On se propose dans cet exercice d'estimer θ.

1) Pour voir que (3) définit bien une loi de probabilité de variable aléatoire discrète, il sufit de vérifier que la série à termes positifs $\sum_{k=1}^{+\infty}(1-\theta)^{k-1}\theta$ est convergente de somme 1. Il s'agit d'une série géométrique de raison $q = 1-\theta$ et comme $0 < \theta < 1$, il en va de même pour q. La série est donc bien convergente. Sa somme se calcule en se ramenant à la série géométrique standard :

$$\sum_{k=1}^{+\infty}(1-\theta)^{k-1}\theta = \theta\sum_{j=0}^{+\infty}(1-\theta)^j = \frac{\theta}{1-(1-\theta)} = 1.$$

2) Calculons l'espérance

$$\mathbf{E}_\theta X = \sum_{k=1}^{+\infty} kP_\theta(X=k) = \sum_{k=1}^{+\infty} k(1-\theta)^{k-1}\theta = \theta\sum_{k=1}^{+\infty} k(1-\theta)^{k-1} = \theta\sum_{k=1}^{+\infty}\left[\frac{\mathrm{d}}{\mathrm{d}x}x^k\right]_{x=1-\theta}.$$

La série entière $f(x) = \sum_{k=0}^{+\infty} x^k$ a pour rayon de convergence 1 et est donc indéfiniment dérivable terme à terme sur $]-1,1[$. Ainsi, pour tout $x\in]-1,1[$, $f'(x) = \sum_{k=1}^{+\infty} kx^{k-1}$. Par ailleurs $f(x) = (1-x)^{-1}$, d'où $f'(x) = (1-x)^{-2}$. En appliquant ces égalités avec $x = 1-\theta$, il vient :

$$\mathbf{E}_\theta X = \theta\sum_{k=1}^{+\infty} k(1-\theta)^{k-1} = \frac{\theta}{(1-(1-\theta))^2} = \frac{1}{\theta}.$$

Pour calculer la variance de la loi de X sous P_θ, on utilise la formule de Koenig :

$$\operatorname{Var}_\theta X = \mathbf{E}_\theta X^2 - (\mathbf{E}_\theta X)^2,$$

qui nous ramène au calcul de $\mathbf{E}_\theta X^2$.

$$\mathbf{E}_\theta X^2 = \sum_{k=1}^{+\infty} k^2P_\theta(X=k) = \theta\sum_{k=1}^{+\infty} k^2(1-\theta)^{k-1} = \theta\sum_{k=1}^{+\infty}(k(k-1)+k)(1-\theta)^{k-1}.$$

On peut découper cette dernière série en somme de deux séries à termes positifs [1] :

$$\begin{aligned}
\theta\sum_{k=1}^{+\infty}(k(k-1)+k)(1-\theta)^{k-1} &= \theta(1-\theta)\sum_{k=2}^{+\infty} k(k-1)(1-\theta)^{k-2} + \theta\sum_{k=1}^{+\infty} k(1-\theta)^{k-1}\\
&= \theta(1-\theta)f''(1-\theta) + \frac{1}{\theta}\\
&= \theta(1-\theta)\frac{2}{(1-(1-\theta))^3} + \frac{1}{\theta}\\
&= \frac{2(1-\theta)}{\theta^2} + \frac{1}{\theta}.
\end{aligned}$$

1. Donc sans se préoccuper à ce stade de la convergence dans $\mathbb{R}$ qui sera obtenue en fin de calcul. Par ailleurs la série entière $\sum_{k=0}^{+\infty} Q(k)x^k$ où Q est un polynôme de degré d a aussi pour rayon de convergence 1, donc il n'y a pas à s'inquiéter.

En reportant ceci dans la formule de Koenig ci-dessus on obtient :

$$\operatorname{Var}_\theta X = \frac{2(1-\theta)}{\theta^2} + \frac{1}{\theta} - \frac{1}{\theta^2} = \frac{1-\theta}{\theta^2}.$$

3) On considère désormais le modèle statistique $(\Omega, \mathcal{F}, P_\theta)_{\theta\in]0,1[}$ et un n-échantillon $X_1, \dots, X_n$ associé. Autrement dit, pour toute valeur de θ, les variables aléatoires X_i sont P_θ-indépendantes et de même loi sous P_θ donnée par (3). On note $S_n = X_1 + \dots + X_n$ la somme de cet échantillon.

Montrons que n/S_n est un estimateur fortement consistant de θ, c'est-à-dire que :

$$\forall \theta \in]0,1[, \quad \frac{n}{S_n} \xrightarrow[n\to+\infty]{P_\theta-\text{p.s.}} \theta. \tag{4}$$

Réglons d'abord la question de l'éventuelle nullité du dénominateur S_n. Posons

$$\Omega_1 := \bigcap_{i\in\mathbb{N}^*} \{X_i \geqslant 1\}.$$

On vérifie que pour tout $\theta \in]0,1[$, $P_\theta(\Omega_1) = 1$. En effet en passant à l'évènement complémentaire, en notant que pour tout $i \in \mathbb{N}^*$, $P_\theta(X_i \geqslant 1) = 1$ et en utilisant la sous-σ-additivité de P_θ, on obtient :

$$P_\theta(\Omega_1^c) = P_\theta\left(\bigcap_{i\in\mathbb{N}^*}\{X_i < 1\}\right) \leqslant \sum_{i=1}^{+\infty} P_\theta(X_i < 1) = 0,$$

puisque tous les termes de cette série sont nuls [2]. Donc $P_\theta(\Omega_1^c) = 0$ et $P_\theta(\Omega_1) = 1$. On remarque alors que

$$\forall \omega \in \Omega_1, \forall n \geqslant 1, \quad S_n(\omega) \geqslant n > 0.$$

Sur l'espace probabilisé $(\Omega, \mathcal{F}, P_\theta)$, l'application de la loi forte des grands nombres à la suite des variables aléatoires X_i qui sont indépendantes et de même loi vérifiant $\mathbf{E}_\theta |X_1| = \mathbf{E}_\theta X_1 = \frac{1}{\theta} < +\infty$, nous donne la convergence :

$$\frac{S_n}{n} \xrightarrow[n\to+\infty]{P_\theta-\text{p.s.}} \frac{1}{\theta}.$$

Cela signifie qu'il existe un évènement $\Omega_{2,\theta}$ tel que $P_\theta(\Omega_{2,\theta}) = 1$ et

$$\forall \omega \in \Omega_{2,\theta}, \quad \frac{S_n(\omega)}{n} \xrightarrow[n\to+\infty]{} \frac{1}{\theta}.$$

Posons $\Omega'_\theta = \Omega_1 \cap \Omega_{2,\theta}$. Pour tout $\omega \in \Omega'_\theta$, $S_n(\omega)/n$ est non nul. Par continuité sur $\mathbb{R}^*$ de la fonction $t \mapsto 1/t$, la convergence ci-dessus entraîne alors que :

$$\forall \omega \in \Omega'_\theta, \quad \frac{n}{S_n(\omega)} \xrightarrow[n\to+\infty]{} \theta.$$

2. Le même argument montre plus gnéralement qu'une intersection au plus dénombrable d'évènements de probabilité 1 est encore un évènement de probabilité 1.

Or Ω'_θ est encore de P_θ-probabilité 1 comme intersection de deux évènements de P_θ-probabilité 1. Ce raisonnement étant valide pour tout $\theta \in]0,1[$, (4) est démontrée.

Montrons maintenant la convergence en loi suivante.

$$\forall \theta \in]0,1[, \quad S_n^* := \frac{\theta S_n - n}{\sqrt{n(1-\theta)}} \xrightarrow[n\to+\infty]{P_\theta-\text{loi}} Z, \quad Z \sim \mathfrak{N}(0,1). \tag{5}$$

En divisant numérateur et dénominateur par θ dans la formule définissant S_n^*, on remarque que :

$$S_n^* := \frac{S_n - \frac{n}{\theta}}{\sqrt{n\frac{1-\theta}{\theta^2}}} = \frac{S_n - n\mathbf{E}_\theta X_1}{\sqrt{n \operatorname{Var}_\theta X_1}} = \frac{S_n - \mathbf{E}_\theta X_1}{\sqrt{\operatorname{Var}_\theta S_n}}.$$

La convergence (5) apparaît alors comme une simple application du théorème limite central avec l'espace probabilisé $(\Omega, \mathcal{F}, P_\theta)$ et la suite $(X_i)_{i\geqslant 1}$ de variables aléatoires P_θ-indépendantes, de même loi sous P_θ, de carré intégrable et de variance $\operatorname{Var}_\theta X_i$ non nulle.

4) La convergence en loi (5) permet de construire un intervalle de confiance I au niveau 95% pour θ. Voici comment. Sur l'évènement Ω_1 de P_θ-probabilité 1 (quel que soit θ), S_n n'est jamais nul et l'encadrement $-t \leqslant S_n^* \leqslant t$ (pour $t > 0$) se résout en :

$$-t \leqslant S_n^* \leqslant t \Longleftrightarrow \frac{n}{S_n} - \frac{t\sqrt{n(1-\theta)}}{S_n} \leqslant \theta \leqslant \frac{n}{S_n} + \frac{t\sqrt{n(1-\theta)}}{S_n}.$$

Les bornes de cet encadrement dépendent du paramètre inconnu θ et ne peuvent donc pas être utilisées telles quelles pour constituer un intervalle de confiance. Mais comme $0 < \theta < 1$, on peut majorer $1-\theta$ par 1. On a donc sur Ω_1 l'implication :

$$|S_n^*| \leqslant t \Rightarrow \frac{n}{S_n} - \frac{t\sqrt{n}}{S_n} \leqslant \theta \leqslant \frac{n}{S_n} + \frac{t\sqrt{n}}{S_n},$$

d'où pour tout $\theta \in]0,1[$,

$$P_\theta\left(\frac{n}{S_n} - \frac{t\sqrt{n}}{S_n} \leqslant \theta \leqslant \frac{n}{S_n} + \frac{t\sqrt{n}}{S_n}\right) \geqslant P_\theta(|S_n^*| \leqslant t) \simeq 2\Phi(t) - 1,$$

pour n grand d'après (5). En choisissant $t = 1{,}96$, ceci nous permet de proposer

$$I(\omega) = \left[\frac{n}{S_n(\omega)} - \frac{1{,}96\sqrt{n}}{S_n(\omega)}\ ;\ \frac{n}{S_n(\omega)} + \frac{1{,}96\sqrt{n}}{S_n(\omega)}\right] \tag{6}$$

comme intervalle de confiance au niveau 95% pour θ.

Si on a la certitude que $\theta \geqslant \frac{3}{4}$, on peut améliorer cet intervalle en majorant $1-\theta$ par $1 - \frac{3}{4} = \frac{1}{4}$, ce qui nous donne l'intervalle deux fois plus court :

$$I'(\omega) = \left[\frac{n}{S_n(\omega)} - \frac{1{,}96\sqrt{n}}{2S_n(\omega)}\ ;\ \frac{n}{S_n(\omega)} + \frac{1{,}96\sqrt{n}}{2S_n(\omega)}\right].$$

5) On se propose de vérifier la convergence en loi suivante :

$$\forall\theta \in]0,1[, \quad T_n = \frac{\theta S_n - n}{\varepsilon_n + \sqrt{n\left(1-\frac{n}{S_n}\right)}} \xrightarrow[n\to+\infty]{P_\theta-\text{loi}} Z, \quad Z \sim \mathfrak{N}(0,1). \tag{7}$$

On obtient T_n en remplaçant le paramètre inconnu θ au dénominateur de S_n^* par l'estimateur fortement consistant n/S_n. L'ajout de ε_n au dénominateur sert seulement à éviter un dénominateur nul.

Pour traiter cette question, on admet la propriété suivante : si $(Z_n)_{n\geqslant 1}$ et $(Y_n)_{n\geqslant 1}$ sont deux suites de variables aléatoires définies sur le même espace probabilisé et telles que Z_n converge en loi vers Z et Y_n converge p.s. vers une constante c, alors Y_nZ_n converge en loi vers cZ. On appliquera cette propriété en prenant $Z_n = S_n^*$ et

$$Y_n := \frac{\sqrt{n(1-\theta)}}{\varepsilon_n + \sqrt{n\left(1-\frac{n}{S_n}\right)}}.$$

Le problème se réduit donc à justifier la convergence presque-sûre :

$$\forall\theta \in]0,1[, \quad Y_n \xrightarrow[n\to+\infty]{P_\theta-\text{p.s.}} 1. \tag{8}$$

Sur l'évènement Ω'_θ défini à la question 3, on a :

$$\forall\omega \in \Omega'_\theta, \quad \frac{n}{S_n(\omega)} \xrightarrow[n\to+\infty]{} \theta,$$

d'où en rappelant que la suite déterministe (ε_n) tend vers 0,

$$Y_n(\omega) = \frac{1}{\frac{\varepsilon_n}{\sqrt{n}} + \frac{1-\frac{n}{S_n(\omega)}}{1-\theta}} \xrightarrow[n\to+\infty]{} 1.$$

Comme $P_\theta(\Omega'_\theta) = 1$, (8) est démontré.

On peut déduire de la convergence en loi (7) un intervalle de confiance J pour θ au niveau 95%. En effet l'encadrement $-t \leqslant T_n \leqslant t$ pour $t > 0$ se résout en :

$$\frac{n}{S_n} - \frac{t\sqrt{n}}{S_n}\sqrt{1-\frac{n}{S_n}} - \frac{t\varepsilon_n}{S_n} \leqslant \theta \leqslant \frac{n}{S_n} + \frac{t\sqrt{n}}{S_n}\sqrt{1-\frac{n}{S_n}} + \frac{t\varepsilon_n}{S_n}.$$

Le seul intérêt de ε_n est d'empêcher que le dénominateur de $T_n(\omega)$ s'annule pour tout ω réalisant l'évènement de probabilité non nulle $\{S_n = n\}$. Pour l'application qui suit, on prend ε_n suffisamment petit pour être numériquement négligeable, par exemple $\varepsilon_n = 10^{-n}$. En prenant $t = 1{,}96$, on obtient alors pour θ l'intervalle de confiance $J(\omega)$ au niveau 95% :

$$J(\omega) = \left[\frac{n}{S_n(\omega)} - \frac{1{,}96\sqrt{n}}{S_n(\omega)}\sqrt{1-\frac{n}{S_n(\omega)}}\ ;\ \frac{n}{S_n(\omega)} + \frac{1{,}96\sqrt{n}}{S_n(\omega)}\sqrt{1-\frac{n}{S_n(\omega)}}\right]. \tag{9}$$

6) Le tableau ci-dessous donne un échantillon observé de taille 200.

1	2	1	1	1	1	1	1	2	1	1	1	1	1	3	2	1	2	1	1
1	1	1	1	1	1	1	1	4	1	1	1	2	1	1	1	1	2	2	1
2	1	1	1	1	1	1	1	2	1	1	1	2	1	1	1	1	1	2	2
2	2	1	1	3	1	4	1	1	1	1	1	1	2	1	2	1	1	1	1
2	1	1	1	1	6	1	1	1	1	2	2	1	1	1	1	1	1	2	1
1	1	1	1	4	1	2	1	1	1	1	1	1	1	1	2	1	2	1	1
1	1	1	3	1	1	1	1	2	1	1	2	2	2	1	1	1	1	2	1
1	1	1	3	2	1	1	1	2	1	3	1	1	1	1	1	1	1	1	2
3	2	1	1	1	2	1	1	2	1	1	1	2	3	2	1	1	1	1	4
1	2	4	2	1	1	1	1	1	1	1	1	1	1	1	1	2	5	1	2

a) La fonction de répartition empirique F_n associée à l'échantillon $X_1, \dots, X_n$ est définie par

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{X_i \leqslant t\}}.$$

C'est donc une fonction en escaliers dont les points de saut sont les valeurs observées des X_i, donc ici les entiers de 1 à 6. Le tableau suivant permet de dessiner sa représentation graphique pour les valeurs observées de l'échantillon.

Valeur	1	2	3	4	5	6
Effectif	146	40	7	5	1	1
Fréquence	0,73	0,20	0,035	0,025	0,005	0,005

b) On peut utiliser ce tableau pour calculer la valeur de l'estimateur de θ. En effet,

$$S_{200}(\omega) = \sum_{i=1}^{200} X_i(\omega) = 146 \times 1 + 40 \times 2 + 7 \times 3 + 5 \times 4 + 1 \times 5 + 1 \times 6 = 278.$$

Donc l'estimateur de θ calculé sur cet échantillon observé est

$$\frac{200}{S_{200}(\omega)} = \frac{200}{278} = 0{,}719.$$

c) En reportant ces valeurs numériques dans (6) et (9), on obtient pour les intervalles de confiance I et J calculés à partir de cet échantillon observé :

$$I(\omega) = [0{,}619\ ;\ 0{,}819], \qquad J(\omega) = [0{,}666\ ;\ 0{,}772].$$

On constate, comme on pouvait s'y attendre, que J est sensiblement meilleur que I.

7) Une autre idée pour estimer θ est de se rappeler que c'est le paramètre d'une variable aléatoire de Bernoulli Y_i valant 1 en cas de succès à la i^{e} épreuve et 0 en cas d'échec. Le tableau des $X_k(\omega)$ observés donné à la question 6) permet de reconstituer la suite des résultats $Y_i(\omega)$ des épreuves répétées. En effet si $X_k(\omega) = 4$ par exemple, cela signifie que lors de la k^{e} série d'épreuves, le premier succès est apparu à la 4^{e} épreuve. L'écriture $X_k(\omega) = 4$ est donc équivalente à $Y_{k,1}(\omega) = 0$, $Y_{k,2}(\omega) = 0$, $Y_{k,3}(\omega) = 0$, $Y_{k,4}(\omega) = 1$. La généralisation est immédiate, une valeur observée $X_k(\omega) = j$ correpond à $j-1$ zéros suivis d'un 1 pour la suite des $Y_i(\omega)$. On a ainsi en tout $N = 278$ observations

binaires et le nombre de 1 parmi elles est 200 (autant de 1 que de X_k dans l'échantillon). Le nombre de zéros est bien sûr 78. On en déduit immédiatement que

$$\overline{Y}(\omega) = \frac{200}{278} = 0{,}719.$$

L'intervalle de confiance à 95% par la méthode avec variance estimée pour θ est alors

$$K(\omega) = \left[\overline{Y}(\omega) - \frac{1{,}96\sqrt{\overline{Y}(\omega)(1-\overline{Y}(\omega))}}{\sqrt{S_n(\omega)}}\ ;\ \overline{Y}(\omega) + \frac{1{,}96\sqrt{\overline{Y}(\omega)(1-\overline{Y}(\omega))}}{\sqrt{S_n(\omega)}}\right],$$

en remarquant que la taille de l'échantillon binaire est $N = S_n(\omega)$. Comme $\overline{Y}(\omega) = n/S_n(\omega)$, on vérifie facilement que $K(\omega) = J(\omega)$.