ESTIMATEURS — CORRECTION

SIMON COSTE

Exercice 1 • On appelle $X_1, \dots, X_{n_1}$ les variables aléatoires observées dans le premier échantillon, la variable X_i prenant la valeur 1 si le i-ème foyer possède le bien et 0 sinon, et $X'_1, \dots, X'_{n_2}$ les variables aléatoires construites de la même façon à partir des observations réalisées dans le deuxième échantillon.
Toutes les variables traitées ici prennent un nombre fini de valeurs, c'est pourquoi on ne se posera pas la question de l'*existence* de leur espérance et de leur variance.

(1) Les observations $X_1, \dots, X_{n_1}, X'_1, \dots, X'_{n_2}$ sont indépendantes et de même loi de Bernoulli de paramètre p. La somme

$$S_1 = \sum_{k=1}^{n_1} X_k$$

suit donc une loi binomiale de paramètres n_1 et p, tandis que la somme

$$S_2 = \sum_{k=1}^{n_2} X'_k$$

suit une loi binomiale de paramètres n_2 et p. Leur espérances sont

$$\mathbf{E}(S_1) = n_1 p \quad \text{et} \quad \mathbf{E}(S_2) = n_2 p$$

et leurs variances

$$\mathbb{V}(S_1) = n_1 p(1-p) \quad \text{et} \quad \mathbb{V}(S_2) = n_2 p(1-p)$$

(2) Par linéarité de l'espérance et quadraticité de la variance, on a

$$\mathbf{E}(F_1) = \frac{1}{n_1}\mathbf{E}(S_1) = p \quad \text{et} \quad \mathbf{E}(F_2) = \frac{1}{n_2}\mathbf{E}(S_2) = p$$

ainsi que

$$\mathrm{Var}(F_1) = \frac{1}{n_1^2}\mathrm{Var}(S_1) = \frac{p(1-p)}{n_1} \quad \text{et} \quad \mathrm{Var}(F_2) = \frac{1}{n_2^2}\mathrm{Var}(S_2) = \frac{p(1-p)}{n_2}$$

(3) F_1 et F_2 sont des estimateurs car ils sont définis uniquement en fonction des observations X_i et X'_i et peuvent être calculés à l'aide de ces observations sans connaître le paramètre p. Le fait qu'ils soient sans biais résulte de la question précédente puisque l'on a vu que leurs espérances étaient toutes deux égales à p.

Date: 2020-2021.

(4) Par linéarité de l'espérance,

$$\mathbf{E}(\mathrm{G}) = \frac{1}{2}\left(\mathbf{E}(\mathrm{F}_1) + \mathbf{E}(\mathrm{F}_2)\right) = \frac{1}{2}(p+p) = p$$

En tant que fonctionnelle d'estimateurs de p, G est un estimateur de p ; on vient de plus de voir qu'il est sans biais.

(5) F_1 et F_2 sont indépendantes, donc

$$\begin{aligned}\mathrm{Var}(\mathrm{G}) =& \frac{1}{4}\mathrm{Var}(\mathrm{F}_1 + \mathrm{F}_2)\\ =& \frac{1}{4}\left(\mathrm{Var}(\mathrm{F}_1) + \mathrm{Var}(\mathrm{F}_2)\right)\\ =& \frac{1}{4}\left(\frac{1}{n_1}p(1-p) + \frac{1}{n_2}p(1-p)\right)\\ =& \frac{1}{4}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)p(1-p)\end{aligned}$$

(6) À quel résultat doit-on s'attendre ? G est obtenu en faisant brutalement la moyenne des deux estimateurs F_1 et F_2, ce qui revient à leur donner la même importance. Pourtant, F_1 est obtenue à partir d'un échantillon plus grand que F_2 et devrait à ce titre être davantage pondérée pour obtenir un estimateur de meilleure qualité (c'est le sens de la question suivante !). Lorsque l'on calcule la moyenne de F_1 et F_2, on prend certes en compte plus d'information que lors du calcul de F_1, mais on donne à l'information issue du deuxième échantillon un poids démesuré. On s'attend donc à ce que G soit meilleur que F_1 (et donc que F_2) lorsque n_1 et n_2 ne sont pas trop éloignés : c'est la zone dans laquelle l'effet positif de l'ajout d'information prévaut. On s'attend par contre à perdre en précision du fait de la moyennisation avec F_2 lorsque F_1 est beaucoup plus précis que F_2, c'est-à-dire lorsque n_1 est bien plus grand que F_2. Le tout est de quantifier le seuil auquel ce phénomène arrive...
La qualité d'un estimateur se mesure (pour nous) à l'aide de son risque quadratique. G étant sans biais en tant qu'estimateur de p, on a

$$\mathrm{RQ}(\mathrm{G}, p) = \mathrm{Var}(\mathrm{G}) = \frac{1}{4}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)p(1-p)$$

Par ailleurs, F_1 et F_2 sont des estimateurs sans biais de p, donc leur risque quadratique est une fois encore égal à leur variance ; comme $n_1 > n_2$, le risque quadratique le plus bas est celui de F_1 d'après la question 2. G est donc un meilleur estimateur que F_1 et F_2 si et seulement s'il est meilleur que F_1, c'est-à-dire que

$$\mathrm{RQ}(\mathrm{G}, p) < \mathrm{RQ}(\mathrm{F}_1, p)$$

soit encore

$$\frac{1}{4}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)p(1-p) < \frac{1}{n_1}p(1-p)$$

soit, en divisant par p (on exclut le cas où $p = 0$ ou $p = 1$ dans lequel tous les estimateurs ont un risque quadratique nul) :

$$\frac{1}{4}\left(\frac{1}{n_1} + \frac{1}{n_2}\right) < \frac{1}{n_1}$$

ce qui équivaut à

$$\frac{1}{4}\left(1 + \frac{n_1}{n_2}\right) < 1$$

ou encore

$$1 + \frac{n_1}{n_2} < 4$$

c'est-à-dire

$$\frac{n_1}{n_2} < 3$$

G est donc meilleur que F_1 et F_2 si et seulement si $n_1 < 3n_2$.

(7) On s'attend cette fois à ce que les valeurs optimales des pondérations u et v reflètent les effectifs relatifs des deux échantillons. Vérifions-le !
Tout d'abord, si $u, v \in \mathbb{R}$ on a, par linéarité de l'espérance :

$$\mathbf{E}(u\mathrm{F}_1 + v\mathrm{F}_2) = u\mathbf{E}(\mathrm{F}_1) + v\mathbf{E}(\mathrm{F}_2) = (u+v)p$$

Pour que $u\mathrm{F}_1 + v\mathrm{F}_2$ (qui est bien un estimateur !) soit sans biais, il faut donc que $u + v = 1$. On a par ailleurs, en vertu de l'indépendance de F_1 et F_2 :

$$\mathrm{Var}(u\mathrm{F}_1 + v\mathrm{F}_2) = u^2\mathrm{Var}(\mathrm{F}_1) + v^2\mathrm{Var}(\mathrm{F}_2) = \left(\frac{u^2}{n_1} + \frac{v^2}{n_2}\right) p(1-p)$$

Dans le cas non dégénéré $p \in]0,1[$, minimiser cette variance revient à minimiser la quantité $\frac{u^2}{n_1} + \frac{v^2}{n_2}$. Mais on sait que l'on doit avoir $u + v = 1$, soit $v = 1 - u$, pour que $u\mathrm{F}_1 + v\mathrm{F}_2$ soit sans biais ; on cherche donc à minimiser la quantité

$$\mathrm{Q}(u) = \frac{u^2}{n_1} + \frac{(1-u)^2}{n_2}$$

Q est une fonction polynomiale du second degré en u dont le coefficient dominant est strictement positif. Elle admet donc un unique minimum sur $\mathbb{R}$, et ce minimum est atteint au point d'annulation de sa dérivée. Mais

$$\mathrm{Q}'(u) = 0 \;\Leftrightarrow\; \frac{2u}{n_1} - \frac{2(1-u)}{n_2} = 0 \;\Leftrightarrow\; n_2 u = (1-u)n_1 \;\Leftrightarrow\; u(n_1+n_2) = n_1$$

donc la variance minimale est réalisée pour

$$u = \frac{n_1}{n_1+n_2} \quad \text{et} \quad v = 1 - u = \frac{n_2}{n_1+n_2}$$

c'est-à-dire pour des pondérations reflétant les effectifs relatifs des deux échantillons... comme prévu.

Exercice 2 • (1) Pour tout individu $i \in \{1, dotsc, n\}$, on note X_i la variable aléatoire prenant la valeur 1 si l'individu soutient la politique de G.W. Bush et 0 sinon. Par hypothèse, les X_i sont des variables indépendantes de même loi de Bernoulli de paramètre p.

(2) 10000 est un grand nombre (rappel : en économétrie, un nombre est « grand » quand il dépasse 30). On utilise donc l'approximation gaussienne fournie par le théorème central limite, qui permet d'écrire que pour tout $(a,b) \in \mathbb{R}^2$ tel que $a < b$ on a :

$$\mathbf{P}\left(a \leqslant \sqrt{10000}\frac{\overline{\mathrm{X}}_{10000} - p}{\sqrt{p(1-p)}} \leqslant b\right) \approx \Phi(b) - \Phi(a)$$

où Φ est la fonction de répartition de la loi normale centrée réduite. Notons que compte tenu des résultats obtenus, on a $p \in]0,1[$, ce qui justifie que le quotient écrit ci-dessus existe bel et bien. En choisissant $b = \Phi^{-1}(0,975) \approx 1,96$ et $a = -b$ (relisez les exemples du cours si cela ne vous paraît pas évident !), on obtient :

$$\mathbf{P}\left(-1,96 \leqslant 100\frac{\overline{\mathrm{X}}_{10000} - p}{\sqrt{p(1-p)}} \leqslant 1,96\right) \approx 0,95$$

soit

$$\mathbf{P}\left(p - \frac{1,96\sqrt{p(1-p)}}{100} \leqslant \overline{\mathrm{X}}_{10000} \leqslant p + \frac{1,96\sqrt{p(1-p)}}{100}\right) \approx 0,95$$

La proportion $\overline{\mathrm{X}}_{10000}$ de sondés favorables à la politique de G.W. Bush se situe donc dans l'intervalle de fluctuation

$$\left[p - \frac{1,96\sqrt{p(1-p)}}{100}, p + \frac{1,96\sqrt{p(1-p)}}{100}\right]$$

avec une probabilité de 95%. Notons que le calcul d'un intervalle de confiance à partir de la loi exacte de $\overline{\mathrm{X}}_{10000}$ (c'est-à-dire du fait que $10000\overline{\mathrm{X}}_{10000}$ suit une loi binomiale de paramètres 10000 et p) aurait été pour le moins désagréable, puisqu'il aurait supposé d'évaluer la fonction de répartition de la loi $\mathcal{B}(10000, p)$!

(3) L'intervalle de confiance recherché se déduit de la relation

$$\mathbf{P}\left(-1,96 \leqslant 100\frac{\overline{\mathrm{X}}_{10000} - p}{\sqrt{p(1-p)}} \leqslant 1,96\right) \approx 0,95$$

puisque cette dernière implique que

$$\mathbf{P}\left(\left|100\frac{\overline{\mathrm{X}}_{10000} - p}{\sqrt{p(1-p)}}\right| \leqslant 1,96\right) \approx 0,95$$

soit

$$\mathbf{P}\left(\left|\overline{\mathrm{X}}_{10000} - p\right| \leqslant \frac{1,96\sqrt{p(1-p)}}{100}\right) \approx 0,95$$

d'où, comme $p(1-p) \leqslant \frac{1}{4}$:

$$\mathbf{P}\left(\left|\overline{\mathrm{X}}_{10000} - p\right| \leqslant \frac{1,96}{2 \cdot 100}\right) \geqslant 0,95$$

et donc

$$\mathbf{P}\left(p \in \left[\overline{\mathrm{X}}_{10000} - \frac{1,96}{200}, \overline{\mathrm{X}}_{10000} + \frac{1,96}{200}\right]\right) \geqslant 0,95$$

d'où l'intervalle de confiance recherché :

$$\left[\overline{\mathrm{X}}_{10000} - \frac{1,96}{200}, \overline{\mathrm{X}}_{10000} + \frac{1,96}{200}\right]$$

dont la réalisation empirique est

$$\left[0.352 - \frac{1,96}{200}, 0.352 + \frac{1,96}{200}\right] = [0.342, 0.362]$$

Notons que l'on aurait pu directement utiliser un résultat du cours et proposer l'intervalle de confiance (légèrement) plus performant

$$\left[\overline{\mathrm{X}}_{10000} - \frac{1,96\sqrt{\overline{\mathrm{X}}_{10000}(1-\overline{\mathrm{X}}_{10000})}}{100}, \overline{\mathrm{X}}_{10000} + \frac{1,96\sqrt{\overline{\mathrm{X}}_{10000}(1-\overline{\mathrm{X}}_{10000})}}{100}\right]$$

dont la réalisation empirique est

$$\left[0.352 - \frac{1,96\sqrt{0.352 \times (1-0.352)}}{100}, 0.352 + \frac{1,96\sqrt{0.352 \times (1-0.352)}}{100}\right] = [0.343, 0.361]$$

(4) Garantir l'indépendance et l'absence de biais des observations constitue la principale difficulté à laquelle doit faire face un institut de sondage. L'hypothèse de remise semble moyennement crédible (les individus acceptent-ils de répondre deux fois au même sondage ?), mais ce n'est pas la source d'autocorrélation la plus importante. On peut par contre parler :
   — De la circonscription géographique probable des observations dans le cas d'un sondage direct,
   — Du biais technologique, de la propension à répondre à un appel et de l'inégale présence au domicile dans le cas d'un sondage téléphonique,
   — Du phénomène d'auto-censure,
   — De la dépendance de la réponse à la formulation de la question,
   — Des non-réponses éventuelles,
   — Du scandale statistique que constitue le sondage s'il est réalisé en ligne,
   — De tant d'autres facteurs !

Exercice 3 • Il s'agit d'un grand classique autour de la formule de Bayes, que l'on rappelle tout de suite. Si A, B sont deux événements, cette formule dit que

$$\mathbf{P}(\mathrm{A}|\mathrm{B}) = \mathbf{P}(\mathrm{B}|\mathrm{A})\frac{\mathbf{P}(\mathrm{A})}{\mathbf{P}(\mathrm{B})}.$$

Dans notre cas, notons X la variable aléatoire qui vaut 1 si Jean-Victor est malade, et 0 sinon ; de même, notons T le résultat du test, 1 pour « positif » et 0 sinon.

Jean-Victor a manifestement été testé positif : T = 1. On cherche donc à calculer $\mathbf{P}(\mathrm{X}=1|\mathrm{T}=1)$. La formule ci-dessus nous indique que

$$\mathbf{P}(\mathrm{X}=1|\mathrm{T}=1)=\mathbf{P}(\mathrm{T}=1|\mathrm{X}=1)\frac{\mathbf{P}(\mathrm{X}=1)}{\mathbf{P}(\mathrm{T}=1)}.$$

La probabilité conditionnelle $\mathbf{P}(\mathrm{T}=1|\mathrm{X}=1)$ est la puissance de première espèce : c'est la probabilité que le test soit positif, à raison, ou encore $1-\mathbf{P}(\mathrm{T}=0|\mathrm{X}=1) = 1-$ faux négatifs, à savoir $100\%-9\%=91\%$.

Vous savez également que la probabilité d'être malade est de l'ordre de $\mathbf{P}(\mathrm{X}=1)=1\%$. Il reste donc à estimer la probabilité pour que le test soit positif $\mathbf{P}(\mathrm{T}=1)$, à laquelle on n'a pas directement accès. Il est pourtant facile de la retrouver à partir de la formule des probabilités totales ! Voici donc :

$$\mathbf{P}(\mathrm{T}=1)=\mathbf{P}(\mathrm{T}=1|\mathrm{X}=0)\mathbf{P}(\mathrm{X}=0)+\mathbf{P}(\mathrm{T}=1|\mathrm{X}=1)\mathbf{P}(\mathrm{X}=1).$$

La probabilité $\mathbf{P}(\mathrm{T}=1|\mathrm{X}=0)$ est précisément le taux de faux positifs, donc 10%, et $\mathbf{P}(\mathrm{X}=0)$ est la probabilité de ne pas être malade, donc 99%. D'autre part, $\mathbf{P}(\mathrm{T}=1|\mathrm{X}=1)$ est la probabilité pour que le test soit justement positif : 91%. Finalement, on obtient

$$\begin{aligned}\mathbf{P}(\mathrm{X}=1|\mathrm{T}=1) &= 91\% \times \frac{1\%}{10\% \times 99\% + 91\% \times 1\%}\\ &= 8.4\%.\end{aligned}$$

Cette probabilité est finalement assez faible, mais cela n'excuse rien : Jean-Victor doit mettre son masque.

Exercice 4 • Avant de commencer, rappelons à toutes fins utiles que si X suit une loi uniforme sur $[a,b]$ avec $a<b$, alors

$$\mathbf{E}(\mathrm{X}_1)=\frac{b-a}{2}$$

et

$$\mathrm{Var}(\mathrm{X}_1)=\frac{(b-a)^2}{12}$$

(1) Par linéarité de l'espérance, on a

$$\mathbf{E}\left(\overline{\mathrm{X}}_n\right)=\mathbf{E}(\mathrm{X}_1)=\frac{\theta}{2}$$

donc $\hat{\theta}_n = 2\overline{\mathrm{X}}_n$ est un estimateur sans biais de θ.

(2) $\hat{\theta}_n$ étant sans biais en tant qu'estimateur de θ, on a

$$\mathrm{RQ}(\hat{\theta}_n,\theta)=\mathrm{Var}(\hat{\theta}_n)=\frac{4}{n^2}\mathrm{Var}\left(\sum_{k=1}^{n}\mathrm{X}_k\right)=\frac{4}{n^2}\sum_{k=1}^{n}\mathrm{Var}(\mathrm{X}_k)=\frac{4}{n}\mathrm{Var}(\mathrm{X}_1)=\frac{4}{n}\frac{\theta^2}{12}=\frac{\theta^2}{3n}$$

par quadraticité de la variance et indépendance des X_i.

(3) Un réflexe à avoir en présence d'une variable définie comme le maximum (resp. le minimum) de variables indépendantes est de calculer sa fonction de répartition (resp. de queue). En effet, si $t\in\mathbb{R}$ on a :

$$\mathbf{P}(\tilde{\theta}_n \leqslant t)=\mathbf{P}\left(\forall i\in\{1,\dots,n\},\mathrm{X}_i\leqslant t\right)$$

d'où, par indépendance des X_i :

$$\mathbf{P}(\tilde{\theta}_n \leqslant t) = \prod_{i=1}^{n} \mathbf{P}(\mathrm{X}_i \leqslant t) = \mathbf{P}(\mathrm{X}_1 \leqslant t)^n$$

et donc

$$\mathbf{P}(\tilde{\theta}_n \leqslant t) = \begin{cases} 0 \text{ si } t < 0 \\ \left(\frac{t}{\theta}\right)^n \text{ si } t \in [0, \theta] \\ 1 \text{ si } t > \theta \end{cases}$$

La fonction de répartition de $\tilde{\theta}_n$ est de classe $\mathcal{C}^1$ par morceaux, donc $\tilde{\theta}_n$ admet une densité donnée par la dérivée de sa fonction de répartition aux points où cette dérivée existe, c'est-à-dire

$$f : x \mapsto \begin{cases} 0 \text{ si } t < 0 \\ \frac{nt^{n-1}}{\theta^n} \text{ si } t \in]0, \theta[\\ 0 \text{ si } t > \theta \end{cases}$$

(4) On calcule successivement l'espérance et la variance de $\tilde{\theta}_n$ grâce à l'expression de la densité que l'on vient de trouver :

$$\mathbf{E}(\tilde{\theta}_n) = \int_0^\theta t \cdot \frac{nt^{n-1}}{\theta^n} \mathrm{d}t = \int_0^\theta \frac{nt^n}{\theta^n} \mathrm{d}t = \frac{n}{\theta^n} \frac{\theta^{n+1}}{n+1} = \frac{n\theta}{n+1}$$

d'où l'expression du biais de $\tilde{\theta}_n$ en tant qu'estimateur de θ :

$$\mathbf{E}(\tilde{\theta}_n) - \theta = -\frac{\theta}{n+1}$$

Par ailleurs :

$$\mathbf{E}(\tilde{\theta}_n^2) = \int_0^\theta t^2 \cdot \frac{nt^{n-1}}{\theta^n} \mathrm{d}t = \frac{n\theta^2}{n+2}$$

par le théorème de transfert et un calcul similaire à celui réalisé pour l'espérance. On a donc :

$$\mathrm{Var}(\tilde{\theta}_n^2) = \mathbf{E}(\tilde{\theta}_n^2) - \mathbf{E}^2(\tilde{\theta}_n) = \frac{n\theta^2}{n+2} - \frac{n^2\theta^2}{(n+1)^2} = \frac{n(n+1)^2 - n^2(n+2)}{(n+2)(n+1)^2}\theta^2 = \frac{n}{(n+2)(n+1)^2}\theta^2$$

d'où enfin, par décomposition biais-variance du risque quadratique :

$$\mathrm{RQ}(\tilde{\theta}_n, \theta) = \left(-\frac{\theta}{n+1}\right)^2 + \frac{n}{(n+2)(n+1)^2}\theta^2 = \frac{n+2+n}{(n+2)(n+1)^2}\theta^2 = \frac{2}{(n+1)(n+2)}\theta^2$$

(5) Le risque quadratique de $\tilde{\theta}_n$ en tant qu'estimateur de θ, qui est de l'ordre de $\frac{1}{n^2}$, est asymptotiquement infiniment plus petit que celui de $\hat{\theta}_n$, qui est de l'ordre de $\frac{1}{n}$. On choisit donc $\tilde{\theta}_n$ pour construire un intervalle de confiance performant. On sait que l'on a nécessairement $\tilde{\theta}_n \leqslant \theta$ (reprenez la définition de $\tilde{\theta}_n$ si ce n'est pas clair), donc notre premier réflexe est de chercher à construire un intervalle de confiance pour θ de la forme $\left[\tilde{\theta}_n, \tilde{\theta}_n + a_n\right]$, où a_n est un terme à déterminer. On écrit donc que l'on souhaite avoir

$$\mathbf{P}\left(\theta \in \left[\tilde{\theta}_n, \tilde{\theta}_n + a_n\right]\right) = 95\%$$

soit

$$\mathbf{P}\left(\theta \leqslant \tilde{\theta}_n + a_n\right) = 95\%$$

puisque $\tilde{\theta}_n$ est nécessairement plus petit que θ, soit encore

$$\mathbf{P}\left(\theta - a_n \leqslant \tilde{\theta}_n\right) = 95\%$$

Mais on connaît la loi de $\tilde{\theta}_n$, et on peut donc calculer pour un $\varepsilon > 0$ quelconque

$$\mathbf{P}\left(\theta - \varepsilon \leqslant \tilde{\theta}_n\right) = \int_{\theta-\varepsilon}^{\theta} \frac{nt^{n-1}}{\theta^n}\mathrm{d}t = \left[\left(\frac{t}{\theta}\right)^n\right]_{\theta-\varepsilon}^{\theta} = 1 - \left(\frac{\theta-\varepsilon}{\theta}\right)^n$$

On veut donc $\left(\frac{\theta-a_n}{\theta}\right)^n = 0.05$ soit

$$\frac{\theta - a_n}{\theta} = 0.05^{\frac{1}{n}}$$

soit encore

$$1 - \frac{a_n}{\theta} = 0.05^{\frac{1}{n}}$$

d'où

$$a_n = \theta\left(1 - 0.05^{\frac{1}{n}}\right)$$

Malheur ! a_n s'exprime en fonction de θ et n'a donc pas sa place dans les bornes de notre intervalle de confiance. Qu'à cela ne tienne : on vient de montrer que

$$\mathbf{P}\left(\theta - \theta\left(1 - 0.05^{\frac{1}{n}}\right) \leqslant \tilde{\theta}_n\right) = 95\%$$

soit

$$\mathbf{P}\left(\theta \cdot 0.05^{\frac{1}{n}} \leqslant \tilde{\theta}_n\right) = 95\%$$

si bien que

$$\mathbf{P}\left(\theta \in \left[\tilde{\theta}_n, 0.05^{-\frac{1}{n}}\tilde{\theta}_n\right]\right) = 95\%$$

On a donc trouvé un intervalle de confiance unilatère *multiplicatif* pour θ. C'est un résultat tout à fait légitime, et on vérifie pour se donner bonne conscience que l'on a bien

$$0.05^{-\frac{1}{n}} = \frac{1}{0.05^{\frac{1}{n}}} = \sqrt[n]{\frac{1}{0.05}} = \sqrt[n]{20} > 1$$

On pourra bien sûr s'amuser à calculer un intervalle de confiance pour θ grâce au théorème central limite appliqué à $\hat{\theta}_n$ et remarquer combien cet intervalle est large comparé à celui que nous venons de déterminer.

Exercice 5 • La kurtosis est un paramètre intéressant qui mesure vaguement « l'aplatissement » d'une courbe.

(1) Appliquons l'inégalité de Cauchy-Schwarz aux variables aléatoires $A = |X - \mu|^2$ et $B = 1$. On obtient

$$\begin{aligned}\mathbf{E}[|X-\mu|^2] = \mathbf{E}[|AB|] &\leqslant \sqrt{\mathbf{E}[|A|^2]\mathbf{E}[|B|^2]} \\ &\leqslant \sqrt{\mathbf{E}[|X-\mu|^4]}.\end{aligned}$$

On en déduit immédiatement que $\kappa(X) \geqslant 1$. On a $\kappa(X) = 1$ si et seulement s'il y a égalité dans la première inégalité ci-dessus, qui est l'inégalité de Cauchy-Schwarz. Dans ce cas, les variables aléatoires $|X - \mu|^2$ et $B = 1$ sont proportionnelles, ce qui signifie que $|X - \mu|^2$ est constante, disons qu'elle est égale à c. Si $c = 0$, cela signifie que X est elle-même constante. Sinon, $c > 0$, et $X - \mu$ ne peut prendre que deux valeurs : $-c$ et $+c$. Par conséquent, X elle-même ne peut prendre que les deux valeurs $\mu \pm c$, et une variable qui ne peut prendre que deux valeurs est une variable aléatoire de Bernoulli.

(2) On rappelle maintenant un résultat fondamental en statistiques : si X_n converge en probabilité vers X et Y_n converge en probabilité vers Y, alors $X_n + Y_n$ converge en probabilité vers $X + Y$, $f(X_n)$ converge en probabilité vers $f(X)$ pour toute fonction continue f, et $1/X_n$ converge en probabilité vers $1/X$ si toutefois X et X_n sont presque sûrement strictement positives.

Pour estimer la kurtosis, on a envie d'utiliser la variable aléatoire suivante :

$$\frac{\frac{1}{n}\sum_{i=1}^n (X_i - \mu)^4}{\left(\frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2\right)^2}.$$

Ce n'est pas une mauvaise idée, mais on ne connaît pas μ : ce n'est donc pas une statistique. On va donc la remplacer par $\bar{X}_n = n^{-1}(X_1 + \cdots + X_n)$. On obtient ainsi l'estimateur

$$\hat{\kappa}_n = \frac{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^4}{\left(\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2\right)^2} = \frac{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^4}{(\hat{\sigma}_n^2)^2.}$$

On sait déjà par le cours que $\hat{\sigma}_n^2 \to \sigma^2$ en probabilité. On voudrait donc démontrer que le numérateur ci-dessus converge bien en probabilité vers $\mathbf{E}[|X-\mu|^4]$. Pour cela, on développe la puissance avec le binôme de Newton : pour chaque i, on a

$$(X_i - \bar{X})^4 = X_i^4 - 4X_i^3\bar{X} + 6X_i^2\bar{X}^2 - 4X_i\bar{X}^3\bar{X}^4$$

et donc le numérateur est égal à

$$\frac{\sum_{i=1}^n X_i^4}{n} - 4\bar{X}\frac{\sum_{i=1}^n X_i^3}{n} + 6\bar{X}^2\frac{\sum_{i=1}^n X_i^2}{n} - 4\bar{X}^3\frac{\sum_{i=1}^n X_i}{n} + \bar{X}^4.$$

En utilisant de façon répétée la loi des grands nombres [1] et les propriétés ci-dessus, on constate que cette somme converge en probabilité vers

$$\begin{aligned}\mathbf{E}[X^4] - 4\mu\mathbf{E}[X^3] + 6\mu^2\mathbf{E}[X^2] - 4\mu^3\mathbf{E}[X] + \mu^4 &= \mathbf{E}[X^4 - 4\mu X^3 + 6\mu^2X^2 - 4\mu^3X + \mu^4] \\ &= \mathbf{E}[(X-\mu)^4].\end{aligned}$$

1. On a supposé que les X_i possèdent des moments jusqu'à l'ordre 4 : on a donc bien $n^{-1}(X_1^\ell + \cdots + X_n^\ell) \to \mathbf{E}[X^\ell]$ pour $\ell \in \{1, 2, 3, 4\}$.

Par conséquent, on a bien $\hat{\kappa}_n \to \kappa(\mathrm{X})$ en probabilité.

Exercice 6 • (1) On considère que l'on observe 8 réalisations indépendantes d'une loi normale de paramètres μ (inconnu) et 5.01^2. Notons que la moyenne des 8 observations suit *exactement* une loi $\mathcal{N}\left(\mu, \frac{5.01^2}{8}\right)$ puisque la loi échantillonnée est gaussienne (ce résultat aurait été faux dans le cas général puisque 8 n'est pas exactement un grand nombre, même pour un économètre malhonnête).
On cherche *a priori* un intervalle de confiance de longueur minimale. On construit donc un intervalle bilatère selon la méthode usuelle :

$$\mathbf{P}\left(\left|\frac{\overline{\mathrm{X}}_8 - \mu}{\frac{5.01}{\sqrt{8}}}\right| \leqslant \Phi^{-1}(0.95)\right) = 0.9$$

puisque $\overline{\mathrm{X}}_8 - \mu \sim \mathcal{N}\left(0, \frac{5.01^2}{8}\right)$, d'où

$$\mathbf{P}\left(\left|\frac{\overline{\mathrm{X}}_8 - \mu}{\frac{5.01}{\sqrt{8}}}\right| \leqslant 1.64\right) \approx 0.9$$

l'approximation n'étant due qu'à l'erreur d'arrondi dans le quantile $\Phi^{-1}(0.95)$.
On en déduit :

$$\mathbf{P}\left(\left|\frac{\overline{\mathrm{X}}_8 - \mu}{\frac{5.01}{\sqrt{8}}}\right| \leqslant 1.64\right) \approx 0.9$$

d'où

$$\mathbf{P}\left(\mu \in \left[\overline{\mathrm{X}}_8 - \frac{5.01}{\sqrt{8}}1.64, \overline{\mathrm{X}}_8 + \frac{5.01}{\sqrt{8}}1.64\right]\right) \approx 0.9$$

ce qui montre que

$$\left[\overline{\mathrm{X}}_8 - \frac{5.01}{\sqrt{8}}1.64, \overline{\mathrm{X}}_8 + \frac{5.01}{\sqrt{8}}1.64\right]$$

est l'intervalle de confiance attendu. Sa réalisation numérique est

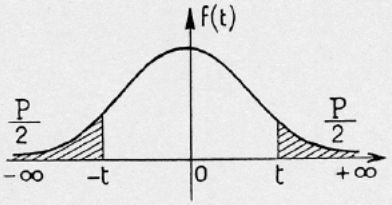
$$[11.93 - 2.90, 11.93 + 2.90] = [9.03, 14.83]$$

Il ne faut pas s'étonner du caractère très imprécis de cet intervalle de confiance compte tenu du nombre d'observations très faible et de l'écart-type très important de la distribution complète !
On pouvait ici encore utiliser directement la formule de l'intervalle de confiance donnée dans le cours, mais il est dans tous les cas nécessaire d'avoir en tête la construction ci-dessus.

(2) On utilise cette fois la formule du cours : si l'écart-type σ est inconnu, on le remplace par sa version empirique débiaisée

$$\hat{\sigma}_8 = \sqrt{\frac{1}{7}\sum_{k=1}^{8}\left(\mathrm{X}_i - \overline{\mathrm{X}}_8\right)^2}$$

dans la formule de l'intervalle de confiance, mais on remplace aussi le quantile de la loi normale $\Phi^{-1}(0.95)$ par le quantile correspondant de la loi de Student



ν \ P	0,90	0,80	0,70	0,60	0,50	0,40	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1	0,158	0,325	0,510	0,727	1,000	1,376	1,963	3,078	6,314	12,706	31,821	63,657	636,619
2	0,142	0,289	0,445	0,617	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925	31,598
3	0,137	0,277	0,424	0,584	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841	12,929
4	0,134	0,271	0,414	0,569	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604	8,610
5	0,132	0,267	0,408	0,559	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032	6,869
6	0,131	0,265	0,404	0,553	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707	5,959
7	0,130	0,263	0,402	0,549	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499	5,408
8	0,130	0,262	0,399	0,546	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355	5,041
9	0,129	0,261	0,398	0,543	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250	4,781
10	0,129	0,260	0,397	0,542	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169	4,587
11	0,129	0,260	0,396	0,540	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106	4,437
12	0,128	0,259	0,395	0,539	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055	4,318
13	0,128	0,259	0,394	0,538	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012	4,221
14	0,128	0,258	0,393	0,537	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977	4,140
15	0,128	0,258	0,393	0,536	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947	4,073
16	0,128	0,258	0,392	0,535	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921	4,015
17	0,128	0,257	0,392	0,534	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,898	3,965
18	0,127	0,257	0,392	0,534	0,688	0,862	1,067	1,330	1,734	2,101	2,552	2,878	3,922
19	0,127	0,257	0,391	0,533	0,688	0,861	1,066	1,328	1,729	2,093	2,539	2,861	3,883
20	0,127	0,257	0,391	0,533	0,687	0,860	1,064	1,325	1,725	2,086	2,528	2,845	3,850
21	0,127	0,257	0,391	0,532	0,686	0,859	1,063	1,323	1,721	2,080	2,518	2,831	3,819
22	0,127	0,256	0,390	0,532	0,686	0,858	1,061	1,321	1,717	2,074	2,508	2,819	3,792
23	0,127	0,256	0,390	0,532	0,685	0,858	1,060	1,319	1,714	2,069	2,500	2,807	3,767
24	0,127	0,256	0,390	0,531	0,685	0,857	1,059	1,318	1,711	2,064	2,492	2,797	3,745
25	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,316	1,708	2,060	2,485	2,787	3,725
26	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,315	1,706	2,056	2,479	2,779	3,707
27	0,127	0,256	0,389	0,531	0,684	0,855	1,057	1,314	1,703	2,052	2,473	2,771	3,690
28	0,127	0,256	0,389	0,530	0,683	0,855	1,056	1,313	1,701	2,048	2,467	2,763	3,674
29	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,311	1,699	2,045	2,462	2,756	3,659
30	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,310	1,697	2,042	2,457	2,750	3,646
40	0,126	0,255	0,388	0,529	0,681	0,851	1,050	1,303	1,684	2,021	2,423	2,704	3,551
80	0,126	0,254	0,387	0,527	0,679	0,848	1,046	1,296	1,671	2,000	2,390	2,660	3,460
120	0,126	0,254	0,386	0,526	0,677	0,845	1,041	1,289	1,658	1,980	2,358	2,617	3,373
∞	0,126	0,253	0,385	0,524	0,674	0,842	1,036	1,282	1,645	1,960	2,326	2,576	3,291

FIGURE 1. Table de la loi de Student pour différents degrés de liberté

de paramètre $8 - 1 = 7$, c'est-à-dire $\mathrm{F}_7(0.95) \approx 1.895$ (voir la figure 1). L'intervalle de confiance attendu est donc

$$\left[\overline{\mathrm{X}}_8 - \frac{\hat{\sigma}_8}{\sqrt{8}} 1.895, \overline{\mathrm{X}}_8 + \frac{\hat{\sigma}_8}{\sqrt{8}} 1.895\right]$$

dont la réalisation numérique est

$$[11.93 - 3.76, 11.93 + 3.76] = [8.17, 15.69]$$

puisque $\hat{\sigma}_8 \approx 5.61$. On obtient un intervalle de confiance plus large que dans la question précédente ; cela est conforme à l'intuition puisque l'on dispose *a priori* de moins d'information, mais il aurait tout à fait pu en être autrement si le petit échantillon dont nous disposons avait présenté une variance moindre que celle de l'ensemble des copies !

(3) On souhaite cette fois obtenir un intervalle de confiance unilatère de la forme $[\mathrm{S}, 20]$, où S est une statistique bien choisie. On utilise la formule du cours

et une table de quantiles de la loi normale [2] pour trouver, dans le cas où la variance est connue [3], l'intervalle suivant :

$$\left[\overline{\mathrm{X}}_8 - 1.28\frac{5.01}{\sqrt{8}}, 20\right]$$

soit

$$\left[11.93 - 1.28\frac{5.01}{\sqrt{8}}, 20\right] = [10.16, 20]$$

On constate avec soulagement que la borne inférieure de cet intervalle est plus grande que celle de l'intervalle bilatère correspondant, ce qui, cette fois, doit être systématiquement le cas.

2. En fait, nous n'avons besoin que de la table donnée dans la figure 1 pour déterminer le quantile que nous cherchons. Pourquoi ?

3. Le cas où la variance est inconnue vous est laissé en exercice !