Université Paris-Dauphine L3 CPES

DST1 – Statistique

Polycopié et calculatrice autorisés. Tout autre document interdit.
Durée 2h
Date : 27 octobre 2022

Dans tout le sujet pour tout $\alpha \in]0,1[$, q_α désigne le quantile d'ordre α de la loi normale $\mathcal{N}(0,1)$.
Approximations à 10^{-2} près : $q_{0.9} = 1.28$, $q_{0.95} = 1.64$, $q_{0.975} = 1.96$.

Exercice 1 (questions de cours (/5))

1. Quelles sont les différences entre la loi faible et la loi forte des grands nombres ?

   Solution: Ces deux théorèmes sont des résultats de convergence de la moyenne empirique d'une suite de variables aléatoires de même espérance et variance. La convergence de la loi faible est en probabilité et celle de la loi forte est presque sûre. Le résultat de la loi forte est vrai pour des hypothèses plus faibles (la loi faible nécessite l'existence d'une variance et celui de la loi forte seulement un moment d'ordre 1).

2. Soit $\widehat{\theta}$ un estimateur tel que

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, \sigma^2),$$

   où σ^2 est une quantité positive. Quelle est la variance asymptotique de $\widehat{\theta}$?

   Solution:
$$\mathrm{Var}_n(\widehat{\theta}) = \frac{\sigma^2}{n}.$$

3. Soit $(X_n)_{n\in\mathbb{N}^*}$ une suite de variables aléatoires telle que, pour tout $n \geq 3$,

$$\mathbb{P}(X_n = n) = \mathbb{P}(X_n = -n) = \frac{1}{n},$$

   et

$$\mathbb{P}(X_n = 0) = 1 - \frac{2}{n}.$$

   Monter que la suite $(X_n)_{n\in\mathbb{N}^*}$ converge en loi vers la variable aléatoire $X = 0$?

Solution: Soit $\varphi : \mathbb{R} \to \mathbb{R}$ une fonction continue et bornée. Nous avons, comme X_n est une variable discrète,

$$\mathbb{E}[\varphi(X_n)] = \sum_{x \in X(\Omega)} \varphi(x)\mathbb{P}(X_n = x) = \varphi(0)\left(1 - \frac{2}{n}\right) + \frac{\varphi(n)}{n} + \frac{\varphi(-n)}{n},$$

comme φ est bornée,

$$\lim_{n \to +\infty} \frac{\varphi(n)}{n} = \lim_{n \to +\infty} \frac{\varphi(-n)}{n} = 0.$$

Donc

$$\lim_{n \to +\infty} \mathbb{E}[\varphi(X_n)] = \varphi(0).$$

X_n convergence en loi vers 0, c'est-à-dire vers une variable aléatoire X telle que $\mathbb{P}(X = 0) = 1$.

4. Nous reprenons la suite $(X_n)_{n \in \mathbb{N}^*}$ de l'exercice précédent. La suite $(X_n)_{n \in \mathbb{N}}$ converge-t'elle en probabilité vers 0 ?

Solution: Nous avons, pour $\varepsilon > 0$,

$$\mathbb{P}(|X_n - 0| > \varepsilon) = \mathbb{P}(|X_n| > \varepsilon).$$

On fait tendre n vers $+\infty$ donc on peut prendre $n \geq \varepsilon$, dans ce cas

$$\mathbb{P}(|X_n| > \varepsilon) = \mathbb{P}(X_n = n) + \mathbb{P}(X_n = -n) = \frac{2}{n} \to 0.$$

Donc X_n converge en probabilité vers 0.

5. Nous reprenons la suite $(X_n)_{n \in \mathbb{N}^*}$ de la question 3. La suite $(X_n)_{n \in \mathbb{N}}$ converge-t'elle dans $\mathbb{L}^2$ vers 0 ?

Solution: Nous avons, comme X_n est une variable discrète

$$\begin{aligned} \mathbb{E}[(X_n - 0)^2] &= \mathbb{E}[X_n^2] = \sum_{x \in X(\Omega)} x^2 \mathbb{P}(X_n = x) \\ &= 0^2\mathbb{P}(X_n = 0) + n^2\mathbb{P}(X_n = n) + (-n)^2\mathbb{P}(X_n = n) \\ &= 2n. \end{aligned}$$

Cette quantité ne tend pas vers 0 donc $(X_n)_{n \in \mathbb{N}^*}$ ne converge pas dans $\mathbb{L}^2$ vers 0.

Exercice 2 (comparaison d'intervalles de confiance (/5)) Soit X une variable suivant la loi de Bernoulli de paramètre $p \in]0, 1[$ inconnu. Nous souhaitons estimer p à partir de l'observation d'un échantillon $X_1, \ldots, X_n$ de même loi que X.

1. Soit $\alpha \in]0,1[$. Donner (sans redémontrer le résultat du cours) un intervalle de confiance asymptotique de niveau α pour le paramètre p.

Solution:

$$\left[\widehat{p} \pm q_{1-\alpha/2}\frac{\sqrt{\widehat{p}(1-\widehat{p})}}{\sqrt{n}}\right],$$

où $\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

2. Soit $\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Utiliser l'inégalité de Bienaymé-Tchebychev pour montrer que, pour tout $t > 0$,

$$\mathbb{P}(|\widehat{p} - p| \geq t) \leq \frac{p(1-p)}{nt^2}.$$

Solution: D'après l'inégalité de Bienaymé-Tchebychev

$$\mathbb{P}(|\widehat{p} - \mathbb{E}[\widehat{p}]| \geq t) \leq \frac{\mathrm{Var}(\widehat{p})}{t^2}.$$

On obtient le résultat voulu en remarquant que

$$\mathbb{E}[\widehat{p}] = \mathbb{E}[X_1] = p$$

et

$$\mathrm{Var}(\widehat{p}) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_1) = \frac{p(1-p)}{n}.$$

3. En déduire que

$$\mathbb{P}(|\widehat{p} - p| \geq t) \leq \frac{1}{4nt^2}.$$

Solution: On peut faire un tableau de variations de la fonction $f(x) = x - x^2$. Nous avons $f'(x) = 1 - 2x$. f admet donc comme point critique $x^* = 1/2$, $f'(x)$ est positive si $x \leq x^*$, négative si $x \geq x^*$ donc f admet comme maximum $f(x^*) = 1/4$. Cela implique que $p(1-p) \leq 1/4$, ce qui nous donne le résultat voulu d'après 2.

4. En déduire un intervalle de confiance de niveau α pour p.

Solution: On cherche t tel que

$$\frac{1}{4nt^2} = \alpha,$$

i.e.

$$t = \frac{1}{2\sqrt{n\alpha}}.$$

D'après 3.,

$$\mathbb{P}\left(|\widehat{p} - p| \leq \frac{1}{2\sqrt{n\alpha}}\right) \leq \alpha,$$

ou, de manière équivalente,

$$\mathbb{P}\left(|\widehat{p} - p| < \frac{1}{2\sqrt{n\alpha}}\right) \geq 1 - \alpha$$

donc

$$\left[\widehat{p} \pm \frac{1}{2\sqrt{n\alpha}}\right]$$

est un intervalle de confiance de niveau α pour p.

5. Application numérique : lors d'une étude [1] réalisée aux États-Unis en 2020 auprès de 32 893 personnes, il a été relevé que 9.2% des personnes interrogées présentaient des troubles dépressifs. En utilisant les résultats des questions précédentes, donner les réalisations des intervalles de confiance asymptotique et non-asymptotique au niveau 5% de la proportion de personne présentant des troubles dépressifs dans la population totale.

Solution: On note p la proportion à estimer et $X_i = 1$ si la i-ème personne interrogée présente des troubles dépressifs et 0 sinon. D'après l'énoncé $\widehat{p} = 0.092$. L'intervalle de la question 1. donne

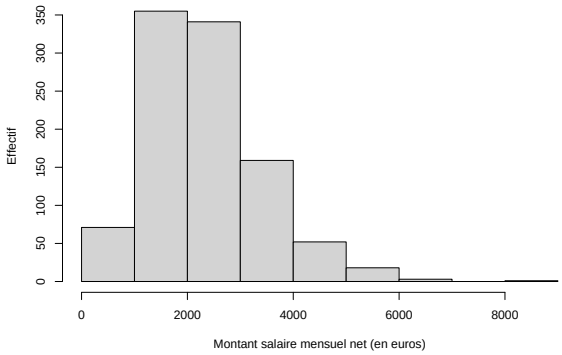
$$\left[0.092 \pm 1.96\frac{\sqrt{0.092 \times (1 - 0.092)}}{\sqrt{32893}}\right] = [0.089; 0.095],$$

en arrondissant les résultats à 10^{-3} près. Celui de la question 4. donne

$$\left[0.092 \pm \frac{1}{2\sqrt{32893 \times 0.05}}\right] = [0.080; 0.104].$$

Exercice 3 (Modélisation de la répartition des salaires (/10)) Nous souhaitons modéliser la répartition des salaires d'une entreprise comprenant un grand nombre de salariés. Pour cela nous avons accès seulement au données salariales d'un échantillon de 1000 employés de l'entreprise représentés par la Figure 1. Sur les données observées, le salaire moyen est de 2335€, le salaire médian de 2180 € et l'écart-type de la distribution est de 1058 €.

[1] Goodwin *et al.*(2022), Trends in U.S. Depression Prevalence From 2015 to 2020: The Widening Treatment Gap, *Am J Prev Med*, in press.



Salaire net mensuel (en €)	Effectif
$[0, 1000[$	71
$[1000, 2000[$	355
$[2000, 3000[$	341
$[3000, 4000[$	159
$[4000, 5000[$	52
$[5000, 6000[$	18
$[6000, 7000[$	3
$[7000, 8000[$	0
$[8000, 9000[$	1

Figure 1: Répartition des salaires de l'entreprise T. (données fictives)

1. Nous modélisons dans un premier temps les données observées comme une réalisation $X_1, \ldots, X_n$ d'un échantillon de variables aléatoires distribuées selon une loi normale de moyenne μ et de variance σ^2 inconnus.

   (a) Notons X une variable aléatoire de loi $\mathcal{N}(\mu, \sigma^2)$. Montrer que la médiane de X est égale à μ.

   Solution: Il suffit de montrer que $F_X(\mu) = 0.5$. En notant $Z = (X - \mu)/\sigma$,

   $$1 - F_X(\mu) = 1 - \mathbb{P}(X \leq \mu) = \mathbb{P}(X > \mu) = \mathbb{P}\left(\frac{X - \mu}{\sigma} > 0\right) = \mathbb{P}(Z > 0)$$

   or, par symétrie de la loi normale,

   $$\mathbb{P}(Z > 0) = \mathbb{P}(Z < 0) = \mathbb{P}\left(\frac{X - \mu}{\sigma} < 0\right) = \mathbb{P}\left(X < \mu\right) = F_X(\mu).$$

   Donc

   $$1 - F_X(\mu) = F_X(\mu)$$

   ce qui équivaut à dire que

   $$F_X(\mu) = 0.5.$$

   (b) Donner, en utilisant directement le résultat du cours, un intervalle de confiance de niveau α pour la moyenne μ et sa réalisation sur les données au niveau $\alpha = 5\%$.

Solution: D'après le cours, un IC de niveau α pour μ est

$$\left[\widehat{\mu} \pm q_{1-\alpha/2}\frac{\widehat{\sigma}}{\sqrt{n}}\right].$$

Sur les données observées, nous avons $\widehat{\mu}(\mathbf{x}) = 2335$ et $\widehat{\sigma}(\mathbf{x}) = 1058$, cela donne

$$\left[2335 \pm 1.96 \times \frac{1058}{\sqrt{1000}}\right] = [2269; 2401].$$

(c) En vous aidant de la Figure 1, et des réponses aux questions 1.(a) et 1.(b), pensez-vous qu'il est raisonnable de modéliser les salaires comme suivant une loi normale ?

Solution: A priori, les données observées ne suivent pas une loi normale : d'une part l'histogramme observé n'est pas symétrique, d'autre part la médiane des observations n'est pas dans l'intervalle de confiance que nous avons sur le salaire, ce qui signifie que la médiane n'est pas égale à μ, ce qui serait le cas si les observations suivaient une loi normale.

2. Nous supposons maintenant que les observations suivent une loi Γ de paramètres $k = 3$ et θ inconnu, c'est-à-dire que la densité de X est

$$f_X(x) = \frac{x^2 e^{-x/\theta}}{2\theta^3}\mathbf{1}_{\{x>0\}}.$$

(a) Vérifier sans calcul que

$$\int_0^{+\infty} x^2 e^{-x/\theta} dx = 2\theta^3.$$

Solution: Comme f_X est une fonction densité alors

$$\int_{\mathbb{R}} f_X(x)dx = 1,$$

d'où le résultat.

(b) Montrer que

$$\mathbb{E}[X] = 3\theta.$$

Solution:

$$\mathbb{E}[X] = \int_{\mathbb{R}} x f_X(x) dx = \int_0^{+\infty} \frac{x^3 e^{-x/\theta}}{2\theta^3} dx = \frac{1}{2\theta^3} \lim_{A\to+\infty} \int_0^A x^3 e^{-x/\theta} dx.$$

Par intégration par parties nous avons

$$\int_0^A x^3 e^{-x/\theta} dx = -A^3\theta e^{-A/\theta} + 3\theta \int_0^A x^2 e^{-x/\theta} \xrightarrow[A\to\infty]{} 0 + 3\theta \int_0^{+\infty} x^2 e^{-x/\theta} dx = 3\theta \times 2\theta^3,$$

en utilisant 2.(a). D'où le résultat.

(c) Montrer que l'estimateur

$$\widehat{\theta} = \frac{1}{3n} \sum_{i=1}^n X_i,$$

est un estimateur sans biais et convergent de θ.

Solution: Nous avons, en utilisant la linéarité de l'espérance et la question 2.(b),

$$\mathbb{E}[\widehat{\theta}] = \mathbb{E}\left[\frac{1}{3n} \sum_{i=1}^n X_i\right] = \frac{1}{3n} \sum_{i=1}^n \mathbb{E}[X_i] = \frac{1}{3n} \times n \times 3\theta = \theta,$$

donc $\widehat{\theta}$ est un estimateur sans biais de θ. En utilisant à nouveau la question 2.(b), par la loi des grands nombres, nous avons que

$$\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow[n\to\infty]{} \mathbb{E}[X] = 3\theta \quad p.s.$$

donc

$$\widehat{\theta} \xrightarrow[n\to\infty]{} \theta \quad p.s.$$

et $\widehat{\theta}$ est un estimateur fortement convergent donc convergent de θ.

(d) On admettra que $\mathrm{Var}(X) = 3\theta^2$. Notons

$$\widehat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i,$$

et $\mu = \mathbb{E}[X]$. Vérifier que

$$\sqrt{n}(\widehat{\mu} - \mu) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, 3\theta^2).$$

et en déduire que $\widehat{\theta}$ est asymptotiquement normal. Donner sa variance asymptotique.

Solution: La convergence en loi de $\widehat{\mu}$ découle directement du théorème central limite (l'énoncé nous dit bien que $\mathrm{Var}(X)$ est finie). Nous remarquons ensuite que $\widehat{\mu} = 3\widehat{\theta}$ et $\mu = 3\theta$ par 2.(b), ce qui donne

$$\sqrt{n}(3\widehat{\theta} - 3\theta) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, 3\theta^2)$$

ou, de manière équivalente

$$\frac{\sqrt{3}\sqrt{n}}{\theta}(\widehat{\theta} - \theta) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, 1).$$

$\widehat{\theta}$ est donc bien asymptotiquement normal, de variance asymptotique $\frac{\theta^2}{3n}$.

(e) Montrer que

$$\frac{\sqrt{3}\sqrt{n}}{\widehat{\theta}}(\widehat{\theta} - \theta) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, 1).$$

Solution: Comme $\widehat{\theta}$ est un estimateur convergent de θ, nous avons

$$\frac{\widehat{\theta}}{\theta} \xrightarrow[n\to+\infty]{\mathbb{P}} 1,$$

qui est une constante. En utilisant le théorème de Slutsky et la question 2.(d), nous avons

$$\frac{\frac{\sqrt{3}\sqrt{n}}{\theta}(\widehat{\theta} - \theta)}{\frac{\widehat{\theta}}{\theta}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, 1),$$

ce qui nous donne bien le résultat voulu.

(f) En déduire un intervalle de confiance asymptotique de niveau α pour le paramètre θ et donner sa réalisation sur les données au niveau $\alpha = 5\%$.

Solution: Le résultat de la question 2.(e) implique aussi que

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{\sqrt{3}\sqrt{n}}{\widehat{\theta}}(\widehat{\theta}-\theta)\right| \leq q_{1-\alpha/2}\right) = \mathbb{P}(|Z| \leq q_{1-\alpha/2}) = 1-\alpha,$$

où $Z \sim \mathcal{N}(0,1)$. En réécrivant l'équation précédente, nous avons

$$\lim_{n\to\infty} \mathbb{P}\left(\theta \in \left[\widehat{\theta} \pm q_{1-\alpha/2}\frac{\widehat{\theta}}{\sqrt{3n}}\right]\right) = 1-\alpha,$$

donc

$$\left[\widehat{\theta} \pm q_{1-\alpha/2}\frac{\widehat{\theta}}{\sqrt{3n}}\right]$$

est un intervalle de confiance asymptotique de niveau α pour θ.
Sur les données, nous avons, à 10^{-1} près,

$$\left[\widehat{\theta}(\mathbf{x}) \pm 1.96 \times \frac{\widehat{\theta}(\mathbf{x})}{\sqrt{3\times 1000}}\right] = \left[\frac{\widehat{\mu}(\mathbf{x})}{3} \pm 1.96 \times \frac{\widehat{\mu}(\mathbf{x})}{3\sqrt{3\times 1000}}\right] = [750; 806]$$

(g) Donner un intervalle de confiance asymptotique de niveau α pour μ dépendant de

$$\widehat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \widehat{\mu})^2}.$$

En déduire un autre intervalle de confiance asymptotique de niveau α pour θ et sa réalisation sur les données (en remarquant que $\widehat{\sigma}(\mathbf{x}) = 1058$). Comparer avec celui obtenu dans la question précédente.

Solution: D'après le cours, un intervalle de confiance asymptotique pour la moyenne μ est

$$\left[\widehat{\mu} \pm q_{1-\alpha/2}\frac{\widehat{\sigma}}{\sqrt{n}}\right].$$

Cela implique que

$$\lim_{n\to+\infty} \mathbb{P}\left(\mu \in \left[\widehat{\mu} \pm q_{1-\alpha/2}\frac{\widehat{\sigma}}{\sqrt{n}}\right]\right) = 1-\alpha.$$

Comme $\mu = 3\theta$ nous avons aussi

$$\lim_{n\to+\infty} \mathbb{P}\left(3\theta \in \left[\widehat{\mu} \pm q_{1-\alpha/2}\frac{\widehat{\sigma}}{\sqrt{n}}\right]\right) = 1-\alpha.$$

i.e.

$$\lim_{n\to+\infty} \mathbb{P}\left(\theta \in \left[\frac{\widehat{\mu}}{3} \pm q_{1-\alpha/2}\frac{\widehat{\sigma}}{3\sqrt{n}}\right]\right) = 1-\alpha.$$

Donc un intervalle de confiance asymptotique pour θ de niveau α est

$$\left[\frac{\widehat{\mu}}{3} \pm q_{1-\alpha/2}\frac{\widehat{\sigma}}{3\sqrt{n}}\right]$$

et sa réalisation sur les données au niveau $\alpha = 5\%$ est

$$\left[\frac{2335}{3} \pm 1.96 \times \frac{1058}{3\sqrt{1000}}\right] = [756; 800].$$

Le second intervalle est plus précis.