STT1000 – Examen Intra 2

(Automne 2021)

Les calculatrices sont autorisées. Les documents sont en revanche interdits, sauf une page d'aide mémoire. L'examen dure 3 heures. Toute sortie avant la fin est autorisée, mais sera définitive.
La feuille propose 7 exercices et un barême approximatif est donné à titre indicatif. Les réponses doivent être reportées sur le cahier joint. Si vous utilisez 2 cahiers, merci de le mentionner, en indiquant 1/2 et 2/2 respectivement.
N'hésitez pas à faire des dessins pour vous aider, mais ne considérez pas un dessin comme une preuve. Si vous utilisez un résultat du cours dans votre preuve, nommez-le aussi précisément que possible.

Si $X \sim \mathcal{N}(0,1)$, on a les probabilités suivantes

$$\begin{cases} \mathbb{P}[X \leq -3] \approx 0.1350\% & \mathbb{P}[X \leq 1] \approx 84.1345\% \\ \mathbb{P}[X \leq -2] \approx 2.2750\% & \mathbb{P}[X \leq 2] \approx 97.7250\% \\ \mathbb{P}[X \leq -1] \approx 15.865\% & \mathbb{P}[X \leq 3] \approx 99.8650\% \\ \mathbb{P}[X \leq 0] = 50.0000\% & \mathbb{P}[X \leq 4] \approx 99.9968\% \end{cases}$$

Si $Q \sim \chi^2(50)$, on a les probabilités suivantes

$$\begin{cases} \mathbb{P}[Q \leq 35] \approx \ 5.31\% & \mathbb{P}[Q \leq 34.764] \approx \ 5.00\% \\ \mathbb{P}[Q \leq 40] \approx 15.67\% & \mathbb{P}[Q \leq 39.754] \approx 15.00\% \\ \mathbb{P}[Q \leq 50] \approx 52.66\% & \mathbb{P}[Q \leq 49.335] \approx 50.00\% \\ \mathbb{P}[Q \leq 60] \approx 84.27\% & \mathbb{P}[Q \leq 60.346] \approx 85.00\% \\ \mathbb{P}[Q \leq 70] \approx 96.76\% & \mathbb{P}[Q \leq 67.505] \approx 95.00\% \end{cases}$$

Si $X_i \sim \mathcal{N}(0,\sigma^2)$ sont i.i.d., $Q = kS^2/\sigma^2 \sim \chi^2(k)$ où $kS^2 = X_1^2 + \cdots + X_k^2$. De plus $\mathbb{E}(Q) = k$ et $\text{Var}[Q] = 2k$.
Pour rappel, la Δ-méthode permet d'affirmer que si $g : \mathbb{R} \to \mathbb{R}$ est une fonction dérivable, telle que $g'(\mu) \neq 0$, si

$$\sqrt{n}\Big(X_n - \mu\Big) \xrightarrow{\mathcal{L}} \mathcal{N}(0,\sigma^2), \text{ lorsque } n \to \infty,$$

alors

$$\sqrt{n}\Big(g(X_n) - g(\mu)\Big) \xrightarrow{\mathcal{L}} \mathcal{N}(0, g'(\mu)^2\sigma^2), \text{ lorsque } n \to \infty.$$

Si X est une variable aléatoire de fonction de répartition F, positive ($F(x) = 0$ pour $x < 0$), d'espérance finie

$$\mathbb{E}[X] = \int_0^\infty (1 - F(x))dx.$$

Si vous pensez que des hypothèses manquent pour répondre à la question, indiquez le dans le cahier. Si vous avez besoin d'introduire des objets mathématiques non définis dans l'énoncé, définissez les clairement.

Exercice 1 – ***(Test Gaussien)***

On considère un échantillon $\{x_1, x_2, \cdots, x_n\}$ tiré suivant une loi $\mathcal{N}(\theta, 1)$. Pour tester $H_0 : \theta \leq 5$ contre $H_1 : \theta > 5$, la region de rejet suivante est considérée,

$$\Big\{(x_1, x_2, \cdots, x_n) : \ \overline{x} > 5 + \frac{1}{\sqrt{n}}\Big\}$$

1. Quelle est la probabilité d'erreur de première espèce de ce test ?
2. Quelle est la probabilité d'erreur de seconde espèce de ce test ?

On avait presque fait cet exercice en classe...
1. La probabilité d'un erreur de première espère est

$$\alpha = \mathbb{P}\Big[\overline{X} > 5 + \frac{1}{\sqrt{n}}\Big|H_0\Big] \text{ où, sous } H_0,\ \overline{X} \sim \mathcal{N}\left(\theta, \frac{1}{n}\right)$$

avec $\theta = 5$ (ce qui correspond au pire des cas "$\theta \leq 5$"), soit

$$\alpha = \mathbb{P}\left[\overline{X} > 5 + \frac{1}{\sqrt{n}}\right] = \mathbb{P}\left[\frac{\overline{X} - \theta}{1/\sqrt{n}} > \frac{5 - \theta}{1/\sqrt{n}} + 1\right] \text{ où } Z = \frac{\overline{X} - \theta}{1/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

i.e.

$$\alpha = \mathbb{P}\left[Z > \sqrt{n}\big(5 - \theta\big) + 1\right] \text{ où } Z \sim \mathcal{N}(0, 1)$$

Si $\theta = 5$,

$$\alpha = \mathbb{P}\left[Z > 1\right] \text{ où } Z \sim \mathcal{N}(0, 1)$$

soit $\alpha = 15.9\%$.
2. Pour la fonction probabilité de seconde espèce,

$$\beta = \mathbb{P}\left[Z < \sqrt{n}\big(5 - \theta\big) + 1\right] \text{ où } Z \sim \mathcal{N}(0, 1),$$

comme c'est une erreur, on regarde, conditionnellement à H_1 la probabilité que $\overline{X} < 5 + \dfrac{1}{\sqrt{n}}$ (autrement dit, on ne rejette plus H_0). Soit

$$\beta = \mathbb{P}\left[Z < \sqrt{n}\big(5 - \theta\big) + 1\right] \text{ où } Z \sim \mathcal{N}(0, 1) = \Phi\Big(\sqrt{n}(5 - \theta) + 1\Big),$$

où Φ est la fonction de répartition de la loi normmale centrée et réduite, $\mathcal{N}(0, 1)$. Pour $\theta > 5$, $\beta \leq 1 - \alpha$. En fait, β décroit avec θ. Dans le "pire" de cas, si $\theta \approx 5$, $\beta \approx 1 - \alpha$. Au contraire, $\theta \to \infty$, $\beta \approx 0$.

Exercice 2 – *(Δ-méthode)*

On considère un échantillon de n observations $\{y_1, y_2, \cdots, y_n\}$ provenant de n variables aléatoires indépendantes de loi $f_\theta(y) = \theta x^{\theta-1}\mathbf{1}_{[0,1]}(y)$, où $\theta \in (0, \infty)$.

1. Calculer l'estimateur par la méthode des moments $\widehat{\theta}$ de θ, et donnez g telle que $\widehat{\theta} = g(\overline{y})$
2. En utilisant le théorème central limite, donnez la distribution approchée de $\sqrt{n}(\overline{Y} - \mu)$, où $\mu = \mathbb{E}[Y]$ sera un paramètre que l'on explicitera.
3. En utilisant la Δ-méthode, donne la distribution approchée de $\sqrt{n}(\widehat{\theta} - g(\mu))$
4. En déduire un intervalle de confiance à 95% pour θ.

---

1. L'espérance de Y est

$$\mathbb{E}[Y] = \int_0^1 y f_\theta(y)dy = \int_0^1 \theta y^{\theta-1} y dy = \int_0^1 \theta y^\theta dy = \theta \cdot \left[\frac{y^{\theta+1}}{\theta+1}\right]_0^1 = \frac{\theta}{\theta+1} = \mu$$

aussi, l'estimateur de la méthode des moments doit vérifier $\overline{y} = \dfrac{\widehat{\theta}}{\widehat{\theta}+1}$, de telle sorte que

$$\widehat{\theta} = \frac{\overline{y}}{1-\overline{y}} = g(\overline{y}) \text{ où } g(x) = \frac{x}{1-x}$$

2. Pour la variance de notre estimateur, nous avons besoin de la variance de Y, et de l'espérance de Y^2 (pour les calculs)

$$\mathbb{E}[Y^2] = \int_0^1 y^2 f_\theta(y)dy = \int_0^1 \theta y^{\theta+1} dy = \theta \cdot \left[\frac{x^{\theta+2}}{\theta+2}\right]_0^1 = \frac{\theta}{\theta+2}$$

donc

$$\text{Var}[Y] = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 = \frac{\theta}{(\theta+2)(\theta+1)^2} = \sigma^2,$$

de telle sorte que, grace au théorème central,

$$\sqrt{n}\left(\overline{Y} - \frac{\alpha}{1+\alpha}\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{\theta}{(\theta+2)(\theta+1)^2}\right).$$

3. Pour la Δ-method,

$$g(x) = \frac{x}{1-x} \text{ et } g'(x) = \frac{1}{(1-x)^2}$$

et la variance asymptotique de $\widehat{\theta}$ sera

$$\text{Var}(\widehat{\theta}) = g'(\mu)^2 \frac{\sigma^2}{n} = \cdots = \frac{\theta(\theta+1)^2}{n(\theta+2)} = \frac{\gamma^2}{n},$$

et on a alors

$$\sqrt{n}\left(\widehat{\theta} - \theta\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{\theta(\theta+1)^2}{(\theta+2)}\right).$$

4. On peut écrire
$$Z = \sqrt{n}\frac{\widehat{\theta}-\theta}{\sqrt{\gamma^2}} \approx \mathcal{N}(0,1),$$
de telle sorte que l'intervalle de confiance (symmétrique) à 95% s'écrit
$$\mathbb{P}\left(-1.96 \leq \sqrt{n}\frac{\widehat{\theta}-\theta}{\sqrt{\gamma^2}} \leq 1.96\right) = 95\%$$
soit
$$\mathbb{P}\left(\widehat{\theta} - \frac{1.96}{\sqrt{n}}\gamma \leq \theta \leq \widehat{\theta} + \frac{1.96}{\sqrt{n}}\gamma\right) = 95\%$$
autrement dit, l'intervalle de confiance à 95% sera, en replaçant γ par son estimateur
$$\left[\widehat{\theta} \pm \frac{1.96}{\sqrt{n}}\sqrt{\frac{\widehat{\theta}(\widehat{\theta}+1)^2}{(\widehat{\theta}+2)}}\right]$$

Exercice 3 – ***(Loi uniforme)***

On dispose de 5 observations, $\{0.95,\ 0.24,\ 0.83,\ 0.52,\ 0.69\}$, qu'on suppose tirées suivant une loi uniforme sur $[0,\theta]$.

1. Donner l'estimateur de la méthode des moments, $\widetilde{\theta}$, et sa valuer numérique.
2. Donner l'estimateur du maximum de vraisemblance $\widehat{\theta}$, et sa valuer numérique.
3. on veut tester $H_0 : \theta = 1$ contre $H_1 : \theta = 1.1$. Proposez un test

---

1. Si $X \sim \mathcal{U}([0,\theta])$,
$$\mathbb{E}[X] = \frac{\theta}{2} \text{ ou } \theta = 2\mathbb{E}[X]$$
et l'estimateur par la méthode des moments est alors $\widetilde{\theta} = 2\overline{x}$, soit, numériquement 1.292.
2. Pour l'estimateur de la méthode du maximum de vraisemblance, rappelons que $f_\theta(x) = \theta^{-1}\mathbf{1}(x \in [0,\theta])$ et
$$\mathcal{L}(\theta) = \prod_{i=1}^n f_\theta(x_i) = \frac{1}{\theta^n}\prod_{i=1}^n \mathbf{1}(x_i \in [0,\theta]) = \frac{1}{\theta^n}\mathbf{1}(x_i \in [0,\theta,\ \forall i]) = \frac{1}{\theta^n}\mathbf{1}(\theta > \max\{x_i\})$$
Comme $\theta \mapsto \mathcal{L}(\theta)$ n'est pas continue, on ne peut pas utiliser la condition du premier ordre. En revanche, on peut noter que
$$\begin{cases} \mathcal{L}(\theta) = 0 \text{ si } \theta < \max\{x_i\} \\ \mathcal{L}(\theta) = \theta^{-n} \text{ est décroissante } \text{ si } \theta \geq \max\{x_i\} \end{cases}$$
donc le maximum de $\mathcal{L}$ sera atteint en θ^{-n}. Aussi, l'estimateur du maximum de vraisenblance est alors $\widehat{\theta} = \max\{x_i\}$, soit, numériquement 0.95.

3. Pour tester $H_0 : \theta = \theta_0$ contre $H_1 : \theta = \theta_1$, intuitivement, on va rejeter H_0 (et accepter H_1) si $\widehat{\theta}$ est “trop grand”, autrement dit, on va rejeter H_0 si $\max\{x_1, \cdots, x_n\} > s$. Et s sera tel que

$$\mathbb{P}\Big[\max\{X_1, \cdots, X_n\} > s \Big| H_0\Big] = \alpha,$$

où le conditionnement par H_0 signifie que $U_i \sim \mathcal{U}([0, 1])$, aussi

$$\mathbb{P}\Big[\max\{X_1, \cdots, X_n\} > s\Big] = (s - 1)^n = \alpha \text{ donc } s = (1 - \alpha)^{1/n}.$$

Si $\alpha = 5\%$, et $n = 5$, $s = 99\%$.

$$\begin{cases} \text{si } \max\{x_1, \cdots, x_n\} \leq 0.99, \text{ on accepte } H_0 \\ \text{si } \max\{x_1, \cdots, x_n\} > .99, \text{ on rejette } H_0. \end{cases}$$

Exercice 4 – ***(Variance dans un modèle Gaussien)***

On dispose de n observations, $\{x_1, x_2, \cdots, x_n\}$, qu'on suppose tirées suivant une loi normale centrée, de variance θ, $\mathcal{N}(0, \theta)$.

1. Donner l'estimateur de la méthode des moments, $\widetilde{\theta}$, tel que $\widetilde{\theta}$ soit un estimateur sans biais de θ.
2. Donner l'estimateur du maximum de vraisemblance $\widehat{\theta}$
3. Montrer que l'information de Fisher vérifie $I_1(\theta) = 1/(2\theta)$.
4. Montrer que $\widetilde{\theta}$ est un estimateur efficace.
5. Donner la loi de $n\widetilde{\theta}/\theta$.
6. Proposer deux intervalles de confiance à 95% pour θ, en supposant $n = 50$: le premier de la forme $[0, a]$ et le second de la forme $[b, \infty)$.

---

1. Si $X \sim \mathcal{N}(0, \theta)$, notons que

$$\mathbb{E}[X] = 0 \text{ et } \text{Var}[X] = \mathbb{E}[X^2] = \theta.$$

On peut utiliser l'estimateur de la méthode des moments,

$$\widetilde{\theta} = \frac{1}{n}\sum_{i=1}^n x_i^2$$

qui sera un estimateur sans biais puisque

$$\mathbb{E}\big[\widetilde{\theta}\big] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n X_i^2\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[X_i^2\big] = \mathbb{E}\big[X^2\big] = \theta.$$

2. On peut écrire la vraisemblance

$$\mathcal{L}(\theta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\theta}} \exp\left(-\frac{x_i^2}{2\theta}\right) = \frac{1}{\sqrt{2\pi\theta}^n} \exp\left(-\frac{1}{2\theta}\sum_{i=1}^n x_i^2\right)$$

et la log-vraisemblance,

$$\log \mathcal{L}(\theta) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\theta - \frac{1}{2\theta} sum_{i=1}^n x_i^2.$$

On peut écrire la condition du premier ordre, qui nous dit qu'au maximum de la (log) vraisemblance, la dérivée de la (log) vraisemblance s'annule,

$$\left.\frac{\partial \log \mathcal{L}(\theta)}{\partial\theta}\right|_{\theta=\widehat{\theta}} = 0 \text{ avec } \frac{\partial \log \mathcal{L}(\theta)}{\partial\theta} = -\frac{n}{2} + \frac{1}{2\theta^2}\sum_{i=1}^n x_i^2$$

Aussi, $\widehat{\theta}$ vérifie

$$\frac{n}{2} = \frac{1}{2\widehat{\theta}^2}\sum_{i=1}^n x_i^2 \text{ soit } \widehat{\theta} = \frac{1}{n}\sum_{i=1}^n x_i^2 = \widetilde{\theta}.$$

3. Pour l'information de Fisher, on peut utiliser la définition de I_1 sur la base d'une seule observation (et donc en utilisant la densité au lieu de la vraisemblance),

$$I_1(\theta) = \mathbb{E}\left(-.\frac{\partial^2 \log f_\theta(X)}{\partial\theta^2}\right)$$

où,

$$\log f_\theta(x) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\theta - \frac{x^2}{2\theta}$$

$$\frac{\partial}{\partial\theta}\log f_\theta(x) = -\frac{1}{2\theta} + \frac{x^2}{2\theta^2}$$

et

$$\frac{\partial}{\partial\theta}\log f_\theta(x) = \frac{1}{2\theta^2} - \frac{x^2}{\theta^3}$$

De telle sorte que

$$I_1(\theta) = \mathbb{E}\left(-\frac{1}{2\theta^2} + \frac{X^2}{\theta^3}\right) \text{ où } X \sim \mathcal{N}(0,\theta)$$

or d'après le rappel en préambule du sujet d'examen, on nous disait que X^2/θ suit une loi du χ^2, à 1 degré de liberté (de moyenne 1), donc $\mathbb{E}[X^2] = \theta$, et

$$I_1(\theta) = -\frac{1}{2\theta^2} + \mathbb{E}\left(\frac{X^2}{\theta^3}\right) = -\frac{1}{2\theta^2} + \frac{\theta}{\theta^3} = \frac{1}{2\theta^2}.$$

4. La borne de Cramér-Rao est

$$\frac{1}{I_n(\theta)} = \frac{1}{nI(\theta)} = \frac{2\theta^2}{n},$$

compte tenu du calcul précédant. Or, pour montrer que $\widetilde{\theta}$ est efficace, étant donné qu'il est sans biais, il suffit de montrer que la variance de $\widetilde{\theta}$ coïncide avec la borne de Cramér-Rao. Or

$$\text{Var}[\widetilde{\theta}] = \text{Var}\left(\frac{1}{n}\sum_{i=1}^n X_i^2\right) = \frac{1}{n^2}\text{Var}\left(\sum_{i=1}^n X_i^2\right) = \frac{1}{n^2}\text{Var}\Big(\theta\cdot\underbrace{\frac{1}{\theta}\sum_{i=1}^n X_i^2}_{Q\sim\chi^2(n)}\Big) = \frac{n^2}{\theta^2}\cdot\text{Var}(Q)$$

où $\mathrm{Var}(Q) = 2n$, donc

$$\mathrm{Var}[\widetilde{\theta}] = \frac{\theta^2}{n^2} \cdot 2n = \frac{2\theta^2}{n} = \frac{1}{nI(\theta)}$$

qui coïncide avec la borne de Cramér-Rao, donc $\widetilde{\theta}$ est efficace.
5. Comme on l'a vu lors du calcul précédant,

$$\widetilde{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 = \frac{\theta}{n} \cdot \underbrace{\frac{1}{\theta}\sum_{i=1}^{n} X_i^2}_{Q\sim\chi^2(n)}$$

donc

$$Q = \frac{n\widetilde{\theta}}{\theta} \sim \chi^2(n).$$

6. Pour construire nos intervalles de confiance, on sait trouver α et β tels que

$$\mathbb{P}[Q \in [0, \alpha]] = \mathbb{P}[Q \in [\beta, \infty)] = 95\%$$

où a et b sont respectivement les quantiles à 95% et 5$ de la loi du chi-deux. Avec les valeurs données en introduction, comme on suppose $n = 50$,

$$\mathbb{P}(Q \in [0, 67.5]) = \mathbb{P}(Q \in [34.76, \infty)) = 95\%$$

soit

$$\mathbb{P}\left(\frac{n\widetilde{\theta}}{\theta} \in [0, 67.5]\right) = \mathbb{P}\left(\frac{n\widetilde{\theta}}{\theta} \in [34.76, \infty)\right) = 95\%$$

ou

$$\mathbb{P}\left(\frac{\theta}{n\widetilde{\theta}} \in [67.5^{-1}, \infty)\right) = \mathbb{P}\left(\frac{\theta}{n\widetilde{\theta}} \in [0, 34.76^{-1})\right) = 95\%$$

soit

$$\mathbb{P}\left(\theta \in [67.5^{-1}n\widetilde{\theta}, \infty)\right) = \mathbb{P}\left(\theta \in [0, 34.76^{-1}n\widetilde{\theta})\right) = 95\%$$

ce qui est exactement l'expression attendue pour un intervalle de confiance pour θ, aussi, comme $n = 50$

$$\begin{cases} \mathbb{P}\left(\theta \in [0, a]\right) = 95\%, \quad \text{avec } a = 50 \cdot 34.76^{-1} = 1.438\widetilde{\theta} \\ \mathbb{P}\left(\theta \in [b, \infty)\right) = 95\%, \quad \text{avec } b = 50 \cdot 67.5^{-1}\widetilde{\theta} = 0.741\widetilde{\theta} \end{cases}$$

Exercice 5 – *(Observations partielles de variables uniformes)*

0. Soit U une variable aléatoire, définie sur $[0, 2]$, avec la fonction de répartition suivante

$$F(u) = \begin{cases} 0 & \text{si } u \leq 1 \\ u/2 & \text{si } u \in (1, 2) \\ 1 & \text{si } u \geq 1 \end{cases}$$

Montrer que

$$\mathbb{E}[U] = \frac{5}{4}.$$

On dispose de n observations, $\{x_1, x_2, \cdots, x_n\}$, qu'on suppose tirées suivant une loi uniforme sur $[0, \theta]$, avec $\theta > 1$.

1. On suppose qu'on n'observe pas les x_i mais les y_i, où

$$y_i = \underbrace{\max\{1, x_i\}}_{=g(x_i)} = \begin{cases} 1 \text{ si } x_i \leq 1 \\ x_i \text{ si } x_i > 1 \end{cases}$$

2. Si X suit une loi uniforme sur $[0, \theta]$, et si $Y = g(X)$, donner la fonction de répartition de Y. En déduire la loi du maximum, $\max\{Y_1, \cdots, Y_n\}$.

3. En déduire un estimateur sans biais de θ construit à partir de $\{y_1, y_2, \cdots, y_n\}$

4. On suppose qu'on n'observe pas les x_i mais les z_i, où

$$z_i = \underbrace{\min\{1, x_i\}}_{=h(x_i)} = \begin{cases} x_i \text{ si } x_i \leq 1 \\ 1 \text{ si } x_i > 1 \end{cases}$$

5. Soit R le nombre d'observations Z plus petites (strictement) que 1. Donner la loi de R.

6. Proposez un estimateur sans biais de $\mathbb{P}[X > 1]$, construit à partir de R

---

0. On va utiliser l'expression donnée dans l'introduction. Cette écriture est en fait caractéristique des variables aléatoires positives, on peut la démontrer facilement en supposant U absolument continue, de densité $f = F'$,

$$\int_0^\infty (1 - F(u))du = \int_0^\infty \left(\int_u^\infty f(t)dt \right) du = \int_0^\infty \left(\int_0^t f(t)du \right) dt$$

en changeant les bornes d'intégration (car $u \in (t, \infty)$ quand $t \in (0, \infty)$ signifie $t \in (0, u)$ quand $u \in (0, \infty)$). Aussi

$$\int_0^\infty (1 - F(u))du = \int_0^\infty f(t) \left(\int_0^t du \right) dt = \int_0^\infty t f(t)dt = \mathbb{E}[U].$$

On peut l'utiliser ici

$$\mathbb{E}[U] = \int_0^\infty (1 - F(u))du = \int_0^1 \underbrace{(1 - F(u))}_{=1} du + \int_1^2 (1 - F(u))du + \int_2^\infty \underbrace{(1 - F(u))}_{=0} du = 1 + \int_1^2 1 - \frac{u}{2} du$$

soit

$$\mathbb{E}[U] = 1 + \int_1^2 1 - \frac{u}{2} du = 1 + \left[u - \frac{u^2}{4} \right]_1^2 = 1 + (2 - 1) - \left(\frac{2^2}{4} - \frac{1^2}{4} \right) = 1 + \frac{1}{4} = \frac{5}{4}.$$

1. La loi de Y (la variable aléatoire sous-jacente aux y_i) est un mélance discret / continu, avec une masse de probabilité en 1, et une densité sur $(1, \theta)$. On peut constuire sa fonction de répartition. Commençons par noter que

$$\mathbb{P}[Y < 1] = \mathbb{P}[\max\{X, 1\} < 1] = 0$$

mais (compte tenu de la discontinuité en 1),

$$\mathbb{P}[Y \leq 1] = \mathbb{P}[\max\{X, 1\} \leq 1] = \mathbb{P}[\max\{X, 1\} = 1] = \mathbb{P}[X \leq 1] = \frac{1}{\theta},$$

comme X est uniformément distribuée sur $[0, \theta]$. Ensuite, la fonction de répartition sera linéaire, pour atteindre 1 en θ, soit, pour $y \in (1, \theta)$,

$$\mathbb{P}[Y \leq y] = \frac{1}{\theta} + \frac{y-1}{\theta - 1} \cdots \left(1 - \frac{1}{\theta}\right) = \frac{1}{\theta} + \frac{y-1}{\theta} = \frac{y}{\theta}$$

(on aurait pu l'avoir simplement en traçant les deux fonctions de répartitions, celle de X et celle de Y, en notant qu'elles sont confondues au delà de 1). Bref, la fonction de répartition G vaut ici, sur $(0, \theta)$

$$G(y) = \frac{y}{\theta} \mathbf{1}(y \in (1, \theta)).$$

On peut maintenant calculer la fonction de répartition du maximum, $G^\star(y) = \mathbb{P}[Y^\star \leq y]$ où $Y^\star = \max\{Y_1, \cdots, Y_n\}$, puisque

$$G^\star(y) = \mathbb{P}[Y^\star \leq y] = \mathbb{P}[Y_i \leq y,\ \forall i] = \prod_{i=1}^{n} \mathbb{P}[Y_i \leq y]$$

et on va écrire

$$\mathbb{P}[Y_i \leq y] = \mathbb{P}[Y_i \leq y | Y^\star \leq 1] \cdot \mathbb{P}[Y^\star \leq 1 + \mathbb{P}[Y_i \leq y | Y^\star > 1] \cdot \mathbb{P}[Y^\star > 1]$$

On peut utiliser l'expression de l'espérance donnée en préambule,

$$\mathbb{E}[Y^\star] = \int_0^\infty (1 - G^\star(y)) dy = \frac{1}{\theta^n} + \frac{1}{\theta^n} \int_1^\theta n x^n dx$$

donc

$$\mathbb{E}[Y^\star] = \frac{1}{\theta^n} + \frac{1}{\theta^n} \int_1^\theta n x^n dx = \ldots = \frac{n}{n+1} \theta$$

3. Pour avoir un estimateur sans biais, on utiliser

$$\widehat{\theta} = \frac{n+1}{n} \max\{y_1, \cdots, y_n\}.$$

4. Soit $R = \mathbf{1}(Z_i < 1)$. Par construction, comme les X_i sont supposés indépendantes, les Z_i sont indépendantes, et donc R suit une loi binomiale, $R \sim \mathcal{B}(n, \mathbb{P}[Z_i < 1])$, avec

$$\mathbb{P}[Z_i < 1] = \mathbb{P}[X_i < 1] = \frac{1}{\theta}$$

Aussi,

$$R \sim \mathcal{B}\left(n, \frac{1}{\theta}\right)$$

Notons que, par construction

$$\mathbb{E}[R] = \frac{n}{\theta}.$$

5. Rappelons que

$$p = \mathbb{P}[X > 1] = \frac{\theta - 1}{\theta} = 1 - \frac{1}{\theta} = 1 - \frac{\mathbb{E}[R]}{n}$$

aussi, un estimateur naturel pour p est $1 - r/n$. Et cet estimateur estime sans biais p, puisque

$$\mathbb{E}\left[1 - \frac{R}{n}\right] = 1 - \frac{\mathbb{E}[R]}{n} = 1 - \frac{1}{\theta} = p$$

Aussi, $1 - r/n$ estime sans biais $\mathbb{P}[X > 1]$.

Exercice 6 – *(Échantillon de 1 et de 2)* [10 points]

On dispose de n observations, $\{x_1, x_2, \cdots, x_n\}$, qu'on suppose obtenues comme des réalisations de variables i.i.d. et prennant les valeurs 1 et 2, avec

$$\mathbb{P}[X_i = 1] = \frac{2-2\theta}{2-\theta} \text{ et } \mathbb{P}[X_i = 2] = \frac{\theta}{2-\theta}$$

où $\theta \in (0,1)$.

1. Calculer l'estimateur $\widehat{\theta}$ obtenu par la méthode des moments du paramètre θ
2. Montrez que
$$\sqrt{n}\big(\widehat{\theta} - \theta\big) \xrightarrow{\mathcal{L}} \mathcal{N}\left(a, \frac{(2-\theta)^2(2\theta-\theta^2-2)}{2}\right), \text{ quand } n \to \infty$$
où a sera une constante qu'on déterminera.

---

1. On nous demande d'utiliser la méthode des moments, or l'espérance vaut

$$\mathbb{E}[X_i] = \mathbb{P}[X_i = 1]\cdot 1 + \mathbb{P}[X_i = 2]\cdot 2 = \frac{2-2\theta}{2-\theta}\cdot 1 + \frac{\theta}{2-\theta}\cdot 2 = \frac{2}{2-\theta}$$

donc $\widehat{\theta}$ obtenu par la méthode des moments, vérifie

$$\overline{x} = \frac{2}{2-\widehat{\theta}} \text{ soit } \widehat{\theta} = 2(1-\overline{x}^{-1}).$$

2. On va utiliser ici la Δ-methode. Notons que

$$\text{Var}[X_i] = \mathbb{E}[X_i^2] - \mathbb{E}[X_i]^2 = \mathbb{P}[X_i = 1]\cdot 1^2 + \mathbb{P}[X_i = 2]\cdot 2^2 - \frac{2^2}{(2-\theta)^2} = \frac{2(1-\theta)}{2-\theta} + \frac{4\theta}{2-\theta} - \frac{4}{(2-\theta)^2} = \frac{4\theta - 2\theta^2 - 1}{(2-\theta)^2}$$

Or d'après le théorème central limite,

$$\sqrt{n}\frac{\overline{X} - \mathbb{E}[X_i]}{\sqrt{\text{Var}[X_i]}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \text{ ou } \sqrt{n}\left(\overline{X} - \frac{2}{2-\theta}\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{4\theta - 2\theta^2 - 4}{(2-\theta)^2}\right)$$

Pour répondre à la question, on va utiliser ici la Δ-methode, avec la fonction

$$g(x) = 2(1-x^{-1}) = 2 - \frac{2}{x} \text{ telle que } g'(x) = \frac{2}{x^2}$$

Par construction,

$$g(\overline{X}) = 2(1-\overline{X}^{-1}) = \widehat{\theta} \text{ et } g(\mathbb{E}[X_i]) = g\left(\frac{2}{2-\theta}\right) = 2\left(1 - \frac{2-\theta}{2}\right) = \theta$$

alors que

$$g'\left(\frac{2}{2-\theta}\right) = 2\left(\frac{2-\theta}{2}\right)^2 = \frac{(2-\theta)^2}{2}$$

de telle sorte que

$$g'\left(\frac{2}{2-\theta}\right)^2 \text{Var}[X_i] = \frac{(2-\theta)^4}{2^2}\cdot\frac{4\theta-2\theta^2-1}{(2-\theta)^2} = \frac{(2-\theta)^2(2\theta-\theta^2-2)}{2}$$

Or comme la Δ-methode permet d'écrire

$$\sqrt{n}\Big(g(\overline{X})-g(\mathbb{E}[X_i])\Big) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, g'\left(\frac{2}{2-\theta}\right)^2 \text{Var}[X_i]\right), \text{ lorsque } n\to\infty,$$

soit, par substitution

$$\sqrt{n}\Big(\widehat{\theta}-\theta\Big) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{(2-\theta)^2(2\theta-\theta^2-2)}{2}\right), \text{ quand } n\to\infty.$$

On notera au passage que $a=0$, ce que l'on savait déjà car l'estimateur de la méthode des moments est convergent.

Exercice 7 – ***(Accidents de la route)***

Dans une ville on donne la répartition du nombre de jours sans accident, avec un accident, deux accidents, etc. parmi 50 jours d'observation au cours d'une même année :

Nombre d'accidents	0	1	2	3	4
Nombre de jours	21	18	7	3	1

On suppose que le nombre d'accidents par jour, X suit une loi de Poisson. Donner un intervalle de confiance de niveau 95% pour le nombre moyen d'accidents par jour (on utilisera une approximation asymptotique). Expliquer rapidement quel test vous proposeriez pour tester $H_0 : X\sim\mathcal{P}(1)$.

Pour formaliser un peu, on avait un échantillon $\{x_1,\cdots,x_n\}$ avec $n=50$, et on nous donne une information partielle, à savoir

$$\sum_{i=1}^n \mathbf{1}(x_i=0)=21,\ \sum_{i=1}^n \mathbf{1}(x_i=1)=18,\ \sum_{i=1}^n \mathbf{1}(x_i=2)=7,\ \sum_{i=1}^n \mathbf{1}(x_i=3)=3,\ \sum_{i=1}^n \mathbf{1}(x_i=0)=1,$$

ce qui permettrait de construire un histogramme. Un peu de calcul à la main s'impose ici, le nombre moyen journalier d'accidents est

$$\overline{x} = \frac{21\times 0+18\times 1+7\times 2+3\times 3+1\times 4}{50} = \frac{45}{50} = 0.9$$

autrement dit, il y a un peu moins d'un accident par jour. Pour la loi de Poisson $\mathcal{P}(\lambda)$, l'estimateur de la méthode des moments et du maximum de vraisemblance coïncide, et $\widehat{\lambda}=\overline{x}$. Le théorème central limite (qui donnera l'approximation Gaussienne) permet d'écrire

$$\sqrt{n}\frac{\overline{X}-\lambda}{\sqrt{\lambda}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1),\ n\to\infty,$$

car si $X_i\sim\mathcal{P}(\lambda)$, $\mathbb{E}[X_i]=\text{Var}[X_i]=\lambda$, que l'on écrira

$$\sqrt{n}\frac{\overline{X}-\lambda}{\sqrt{\overline{X}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1),\ n\to\infty,$$

On peut alors écrire

$$\mathbb{P}\left[\overline{X}+q_{\alpha/2}\sqrt{\frac{\overline{X}}{n}}\leq\lambda\leq\overline{X}+q_{1-\alpha/2}\sqrt{\frac{\overline{X}}{n}}\right]\approx 1-\alpha$$

où classiquement q_u est le quantile de niveau u de la loi $\mathcal{N}(0,1)$, ou $q(u)=\Phi^{-1}(u)$. Dans le cas où $\alpha=5\%$,

$$\mathbb{P}\left[\overline{X}-1.96\sqrt{\frac{\overline{X}}{n}}\leq\lambda\leq\overline{X}+1.96\sqrt{\frac{\overline{X}}{n}}\right]\approx 95\%,$$

autrement dit, l'intervalle de confiance (au seuil $\alpha=5\%$) s'écrit

$$IC=\left[\overline{X}\pm 1.96\sqrt{\frac{\overline{X}}{n}}\right]=\left[0.9\pm 1.96\sqrt{\frac{0.9}{50}}\right]=[0.90\pm\pm 0.26]=[0.64;1.16]$$

C'est cette réponse qui était demandée.
Pour les puristes, pour un intervalle de niveau $1-\alpha$, on a

$$P\left(q_{\alpha/2}\leq\frac{\bar{X}-\lambda}{\sqrt{\frac{\lambda}{n}}}\leq q_{1-\alpha/2}\right)\approx 1-\alpha$$

que l'on peut aussi écrire

$$P\left(\frac{[\bar{X}-\lambda]^2}{\frac{\lambda}{n}}\leq q_{1-\alpha/2}^2\right)\approx 1-\alpha$$

ou encore

$$\mathbb{P}\left(\lambda^2-\lambda\left(2\bar{X}+\frac{q_{1-\alpha/2}^2}{2}n\right)+\bar{X}^2\leq 0\right)\approx 1-\alpha$$

on va alors résoudre cette équation de degré 2,

$$\Delta=\left(2\overline{X}+\frac{q_{1-\alpha/2}}{n}\right)^2-4\overline{X}^2=4\frac{\overline{X}q_{1-\alpha/2}^2}{n}+\frac{q_{1-\alpha/2}^4}{n^2}>0$$

donc le polynôme est négatif lorsque λ est entre les deux racines

$$\mathbb{P}\left(\bar{X}+\frac{q_{1-\alpha/2}^2}{2n}-\sqrt{\frac{\bar{X}q_{1-\alpha/2}^2}{n}+\frac{q_{1-\alpha/2}^4}{4n^2}}<\lambda<\bar{X}+\frac{q_{1-\alpha/2}^2}{2n}-\sqrt{\frac{\bar{X}q_{1-\alpha/2}^2}{n}+\frac{q_{1-\alpha/2}^4}{4n^2}}\right)\approx 1-\alpha$$

On retrouve l'expression précédante en négligeant le terme en $1/n^2$ dans la racine carrée, et le terme en $1/n$ avant la racine, qui sera de toutes façons plus petit que le terme en $1/\sqrt{n}$,

$$\mathbb{P}\left[\overline{X}-q_{1-\alpha/2}\sqrt{\frac{\overline{X}}{n}}\leq\lambda\leq\overline{X}+q_{1-\alpha/2}\sqrt{\frac{\overline{X}}{n}}\right]\approx 1-\alpha.$$