MAT1978 – Probabilités et statistique
Automne 2023
Examen final, durée : 3 heures
Consignes :
- Deux feuilles d'aide mémoire (format lettre, manuscrites et recto/verso) sont autorisées.
- Les calculatrices (ou tout autre outil) sont interdites.
- Il y a 5 questions indépendantes, chacune comptant pour 20% de la note de l'examen.
- Les réponses non justifiées ne rapportent aucun point, sauf s'il est demandé explicitement de ne pas justifier.

On donne l'approximation (à deux décimales près) des constantes suivantes :

$$\sqrt{2} \approx 1.41\,, \qquad e \approx 2.72\,, \qquad \pi \approx 3.14\,, \qquad \log 2 \approx 0.69\,.$$

#1 On considère le jeu de hasard suivant. L'employé d'un casino place 2$ sur la table et lance une pièce de monnaie non-biaisée [1]. Si on observe *face* le jeu s'arrête et le joueur repart avec 2$ sinon l'employé de casino double la somme en jeu en plaçant 2$ additionnels sur la table. L'employé lance une nouvelle fois la pièce de monnaie. Si on observe *face*, le jeu s'arrête et le joueur repart avec 4$ sinon l'employé de casino double la somme en plaçant 4$ additionnels sur la table. Le jeu continue ainsi de suite tant que la pièce affiche *pile* et l'employé du casino continue de doubler le montant déjà présent sur la table après chaque *pile*. Le jeu se termine lorsqu'un *face* est finalement observé auquel cas l'employé du casino ne change pas le montant présent sur la table et le joueur repart avec le montant présent sur la table à ce moment.

(a) Soit T la variable aléatoire qui donne le nombre de fois que la pièce est lancée.

i. Donner l'univers de T. [3pts]

$\mathcal{S}_T = \mathbb{N}\backslash\{0\}$.

ii. Donner la loi de T. [3pts]

$T \sim \text{geo}(1/2)$.

iii. Quelle est la moyenne de T ? (ne pas redémontrer ce résultat) [2pts]

$\mathbb{E}(T) = 2$.

(b) Soit X la variable aléatoire donnant le gain du joueur (le montant avec lequel il repart lorsque le jeu s'arrête). Donner l'univers de X, noté $\mathcal{S}_X$. [2pts]

(c) Montrer que pour tout $x \in \mathcal{S}_X$, les événements $\{X = x\}$ et $\{T = \log(x)/\log(2)\}$ sont identiques. [2pts]

Si $x \in \mathcal{S}_X$ alors il existe $t \in \mathbb{N}\backslash\{0\}$ tel que $x = 2^t$ or

$$X = 2^t \Leftrightarrow T = t\,.$$

or $x = 2^t \Leftrightarrow t = \log(x)/\log(2)$ d'où l'égalité entre les deux événements

$$\{X = x\} = \{T = \log(x)/\log(2)\}\,.$$

1. C'est-à-dire que $\mathbb{P}(\text{pile}) = \mathbb{P}(\text{face}) = 1/2$.

(d) Déduire de la question précédente que

$$\mathbb{P}(X = x) = \mathbb{P}\left(T = \frac{\log(x)}{\log(2)}\right).$$

[2pts]

Si $A \subset B$, $\mathbb{P}(A) \leq \mathbb{P}(B)$ et si en plus $A \supset B$ alors $A = B$ et $\mathbb{P}(A) = \mathbb{P}(B)$. L'égalité des événements implique l'égalité des probabilités de ces événements.

(e) En utilisant la question précédente et la question (a), donner la fonction de masse de X. [3pts]

Comme $\mathbb{P}(T = k) = (1/2)^k$ on a que pour tout $x \in \mathcal{S}_X$:

$$\mathbb{P}(X = x) = \frac{1}{2}^{\log(x)/\log(2)} = e^{(\log(x)/\log(2))\log(1/2)} = e^{-\log(x)} = 1/x\,.$$

(f) Déduire le gain moyen d'un joueur de ce jeu et commenter. [3pts]

Par définition de l'espérance,

$$\mathbb{E}(X) = \sum_{i \in \mathcal{S}_X} i\mathbb{P}(X = i) = \sum_{i \in \mathbb{N}\setminus\{0\}} 2^i \frac{1}{2^i} = \sum_{i \in \mathbb{N}\setminus\{0\}} 1 = \infty\,.$$

En moyenne, on gagne une somme infinie !

#2 On considère une variable aléatoire continue X dont la fonction de densité est donnée par

$$x \in \mathbb{R}, \qquad f_X(x) = \begin{cases} 1/(x+2) & \text{si} \quad x \in (-1, b)\,, \\ 0 & \text{sinon}\,. \end{cases}$$

où b est un nombre réel positif. En utilisant la fonction indicatrice, cette fonction de densité s'écrit

$$x \in \mathbb{R}, \qquad f_X(x) = \frac{1}{x+2}\mathbb{1}[-1 < x < b]\,.$$

(a) Trouver b. [5pts]

On a $1 = \int f_X(x)\mathrm{d}x = \int_{-1}^{b} 1/(x+2)\mathrm{d}x = [\log(x+2)]_{x=-1}^{b} = \log(2+b)$ donc $b = e - 2$.

(b) Calculer $\mathbb{P}(X \in (0, 1/2))$ et $\mathbb{P}(X \in (0, 1))$. [5pts]

On a $\mathbb{P}(X \in (0, 1/2)) = [\log(x+2)]_{x=0}^{1/2} = \log(1/2+2) - \log(2) = \log(1 + 1/4)$. De plus $\mathbb{P}(X \in (0, 1)) = \mathbb{P}(X \in (0, b)) = \log(2+b) - \log(2) = 1 - \log(2)$.

(c) Soit Y une variable aléatoire telle que $Y = 3X$.

i. Montrer que Y est une variable aléatoire continue. Pour cela on reviendra à la définition d'une variable aléatoire continue. [5pts]

La fonction de répartition de Y est $F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(X \leq y/3) = F_X(y/3)$ qui est continue en y car $x \mapsto F_X(x)$ est continue. Donc Y v.a. continue.

ii. Montrer que la densité de Y vérifie

$$y \in \mathbb{R}, \qquad f_Y(y) = \frac{1}{y+6}\mathbb{1}[-3 < y < 3b]\,.$$

[5pts]

On a

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y}F_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y}F_X(y/3) = \frac{\mathrm{d}}{\mathrm{d}y}(y/3)F_X'(y/3) = 1/3 f_X(y/3) = \frac{1}{3}\frac{1}{(y/3)+2}\mathbb{1}[-1 < y/3 < b]\,.$$

#3 Soit $(X_1, \ldots, X_n) \sim_{iid} \mathcal{N}(\mu, 1)$ où $\mu \in \mathbb{R}$ est un paramètre inconnu.

(a) Quel argument permet de dire que $\bar{X}_n = (1/n)\sum_{i=1}^n X_i$ est une variable aléatoire de loi normale ? Démontrer ensuite que $\bar{X}_n \sim \mathcal{N}(\mu, 1/n)$. [5pts]

C'est la stabilité de la loi normale par addition et transformation linéaire. Ainsi, $\bar{X}_n \sim \mathcal{N}(a, b)$ avec $a = \mathbb{E}(\bar{X}_n) = (1/n)\sum_{i=1}^n \mathbb{E}(X_i) = \mu$ et $\mathbb{V}\mathrm{ar}(\bar{X}_n) = (1/n^2)\sum_{i=1}^n \mathbb{V}\mathrm{ar}(X_i) = n/n^2 = 1/n$ par indépendance.

(b) Donner la constante $K \in \mathbb{R}$ telle que le logarithme de la fonction de vraisemblance de μ pour le n-échantillon $(X_1, \ldots, X_n)$ vérifie

$$\mu \in \mathbb{R}, \qquad \log(V_{X_1,\ldots,X_n}(\mu)) = -\frac{1}{2}\sum_{i=1}^n (X_i - \mu)^2 + K\,. \tag{1}$$

[5pts]

Comme

$$V_{X_1,\ldots,X_n}(\mu) = \Pi_{i=1}^n \frac{1}{\sqrt{2\pi}}e^{-(X_i-\mu)^2/2} = \frac{1}{(2\pi)^{n/2}}e^{-(1/2)\sum_{i=1}^n (X_i-\mu)^2}\,.$$

donc $K = -(n/2)\log(2\pi)$.

(c) Donner l'estimateur du maximum de vraisemblance de μ basé sur $(X_1, \ldots, X_n)$, noté $\hat{\mu}$. [5pts]

En dérivant $f(\mu) = \log(V_{X_1,\ldots,X_n}(\mu))$ par rapport à μ on a

$$f'(\mu) = \sum_{i=1}^n (X_i - \mu)\,, \ f''(\mu) = -n < 0$$

donc avec $\hat{\mu} = \sum_{i=1}^n X_i/n$, $f'(\hat{\mu}) = 0$ et $\hat{\mu}$ est bien le maximum de f car f est concave.

(d) On précise maintenant que $\mu \in [0, 1]$, ce qui donne plus d'information que $\mu \in \mathbb{R}$. Donner l'estimateur du maximum de vraisemblance de μ basé sur $(X_1, \ldots, X_n)$ avec cette nouvelle information (on le notera $\tilde{\mu}$), c'est-à-dire que $\tilde{\mu}$ *doit* appartenir à $[0, 1]$. [5pts]

[Aide : On pourra exprimer $\log(V_{X_1,\ldots,X_n}(\mu))$ de l'Eq. (1) comme un polynôme du second degré et s'aider d'une représentation graphique de ce polynôme.]

On a que

$$f(\mu) = -\frac{n}{2}\mu^2 + \mu\sum_{i=1}^n X_i + K'\,.$$

On peut représenter graphiquement f et voir que si $\bar{X}_n \in [0, 1]$ alors $\hat{\mu} = \bar{X}_n$ si $\bar{X}_n > 1$ on voit que le maximum de la f qui est dans $[0, 1]$ est atteint en $\hat{\mu} = 1$ et si $\bar{X}_n < 1$, il est atteint en $\hat{\mu} = 0$.

#4 Dans le cours, nous avons vu que l'expérience de Buffon (qui consiste à laisser tomber une aiguille sur un plancher) [2] permet de donner une approximation du nombre π. Au cours de cette expérience aléatoire, un certain événement [3], noté E, *peut* se réaliser. Plus précisément, un calcul montrait que la probabilité de cet événement est de

$$\mathbb{P}(E) = \frac{1}{\pi}\,.$$

Par l'interprétation fréquentiste des probabilités, si on répéte un grand nombre de fois cette expérience aléatoire (disons n fois) et que l'on reporte le fait que E s'est réalisé à l'expérience i ($i \in \{1, 2, \ldots, n\}$) en définissant $X_i = 1$ et que l'on définit $X_i = 0$ si E ne s'est pas réalisé à l'expérience i, alors

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \approx \mathbb{P}(E) \tag{2}$$

et l'approximation devient meilleure et meilleure à mesure que n devient plus grand.

(a) Quelle est la loi de X_1 ? Préciser son espérance. [4pts]

Comme X_1 ne peut prendre que les valeurs 0 ou 1 et que $\mathbb{P}(X_1 = 1) = \mathbb{P}(E) = 1/\pi$ on a que $X_1 \sim$ bernoulli$(1/\pi)$ et $\mathbb{E}(X_1) = 1/\pi$.

(b) Quel théorème important permet de justifier l'approximation de l'Eq. (2) ? [4pts]

La loi des grands nombres

Un autre résultat central du cours permet de justifier que la variable aléatoire

$$Z_n = \frac{\sqrt{n}}{\sqrt{S^2}}\left(\bar{X}_n - \frac{1}{\pi}\right) \tag{3}$$

ressemble de plus en plus à une variable aléatoire de loi normale $\mathcal{N}(0, 1)$.

(c) Quel est ce théorème ? Rappeler ses hypothèses et comment est défini l'estimateur S^2. [4pts]

C'est le TCL. Il faut que les v.a. dont on fait la somme soit i.i.d. de variance finie. $S^2 = (1/n)\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$.

(d) Soit $F_n : \mathbb{R} \to (0, 1)$ la fonction de répartition de Z_n (de l'Eq. (3)) et $\Phi : \mathbb{R} \to (0, 1)$ la fonction de répartition de la loi $\mathcal{N}(0, 1)$. Sachant que pour tout $z \in \mathbb{R}$, $F_n(z) \to \Phi(z)$ quand $n \to \infty$, proposer un intervalle de confiance (approximatif et seulement valide quand n est grand) de niveau $1 - \varepsilon = 0.95$ pour π sachant que $\Phi^{-1}(0.975) \approx 1.96$. L'intervalle de confiance dépendra des statistiques $\bar{X}_n$, S^2 et n.

[8pts]

On a que

$$\mathbb{P}(Z_n \in (-\Phi^{-1}(1 - \varepsilon/2), \Phi^{-1}(1 - \varepsilon/2))) = F_n(\Phi^{-1}(1 - \varepsilon/2)) - F_n(-\Phi^{-1}(1 - \varepsilon/2)) \to 1 - \varepsilon$$

et en notant $r_\varepsilon = \Phi^{-1}(1 - \varepsilon/2)$

$$\{Z_n \in (-r_\varepsilon, r_\varepsilon)\} = \left\{\bar{X}_n - \frac{S}{\sqrt{n}} r_\varepsilon < \frac{1}{\pi} < \bar{X}_n + \frac{S}{\sqrt{n}} r_\varepsilon\right\}$$

2. les détails sur cette expérience sont inutiles pour cette question
3. là encore les détails sont inutiles

Supposons que $\{\bar{X}_n - (S/\sqrt{n})r_\varepsilon > 0\}$ (dont la probabilité tend vers 1 avec n) alors

$$\{Z_n \in (-r_\varepsilon, r_\varepsilon)\} = \left\{ \frac{1}{\bar{X}_n + \frac{S}{\sqrt{n}} r_\varepsilon} < \pi < \frac{1}{\bar{X}_n - \frac{S}{\sqrt{n}} r_\varepsilon} \right\} .$$

on peut donc donner comme intervalle de confiance approximatif

$$\left(\frac{1}{\bar{X}_n + \frac{S}{\sqrt{n}} 1.96}, \frac{1}{\bar{X}_n - \frac{S}{\sqrt{n}} 1.96} \right) .$$

#5 Un fabricant d'électronique fabrique une certaine composante, intervenant dans la construction d'un processeur, dont la longueur doit être de $\mu = 1$mm. Des défauts inhérents au processus de fabrication font que certaines réalisations de cette composante sont plus grandes que d'autres. Toutefois, on fait confiance au fabricant concernant le fait, qu'en moyenne, les réalisations de cette composante sont de longueur $\mu = 1$mm. Le fabricant prétend que son processus de fabrication est suffisamment précis et que l'écart-type des réalisations de cette composante est de $\sigma_0 = 0.05$mm. On ne tolérerait pas que l'écart-type soit trop grand, car cela risquerait de rendre certaines réalisations de la composante inaptes à être intégrées au processeur. On souhaite faire un test statistique (au niveau de significativité de $\alpha = 0.05$) pour savoir si l'hypothèse du fabricant concernant l'écart-type des réalisations de la composante est plausible.

(a) L'hypothèse nulle du test est clairement $\mathcal{H}_0 : \{\sigma = 0.05\}$. Quelle est l'hypothèse alternative qu'il faut choisir parmi les trois suivantes

$$\mathcal{H}_1 : \{\sigma < 0.05\}\,, \qquad \mathcal{H}_1 : \{\sigma \neq 0.05\}\,, \qquad \mathcal{H}_1 : \{\sigma > 0.05\}\,,$$

pour pouvoir effectuer le test dont il est question ici. [5pts]

On rejette $\mathcal{H}_0$ au profit de l'hypothèse qui met en défaut l'hypothèse du fabricant sur la précision de son processus. Ainsi, on choisit $\mathcal{H}_1 : \{\sigma > 0.05\}$.

(b) On suppose que les réalisations de la composante, notées $X_1, \ldots, X_n$, sont indépendantes et suivent une loi normale. Proposer deux statistiques de test pour le test dont il est ici question : une dont la loi sous l'hypothèse nulle est une loi du χ^2_{n-1} et l'autre dont la loi sous l'hypothèse nulle une loi du χ^2_n. [5pts]

On a comme statistique de test

$$S_1(X) = \frac{(n-1)\hat{S}^2}{\sigma_0^2} = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{0.05^2} \overset{\mathcal{H}_0}{\sim} \chi^2_{n-1}\,.$$

De plus

$$\frac{X_i - \mu}{\sigma_0} \overset{\mathcal{H}_0}{\sim} \mathcal{N}(0,1)$$

avec $\mu = 1$ connu donc

$$S_2(X) = \sum_{i=1}^n \left[\frac{X_i - 1}{0.05}\right]^2 \overset{\mathcal{H}_0}{\sim} \chi^2_n\,.$$

(c) Sachant que l'on dispose d'un échantillon de $n = 30$ réalisations de la composante pour lequel la mesure indique que

$$\sum_{i=1}^n (X_i - \bar{X}_n)^2 = (0.05 \times 4)^2\,,$$

avec $\bar{X}_n = \sum_{i=1}^n X_i/n$, l'hypothèse du fabricant sur σ est-elle plausible ? On utilisera le fait que $F^{-1}(0.05) = 13.121$ ou que $F^{-1}(0.95) = 42.557$, où F est la fonction de répartition (sous $\mathcal{H}_0$) de la statistique nécessaire pour cette question. Par ailleurs, $F^{-1} : (0,1) \to \mathbb{R}^+$ indique la réciproque de la fonction F. [5pts]

La région critique du test avec $S_1(X)$ est $(42.557, \infty)$ et

$$S_1(X) = \frac{(0.05 \times 4)^2}{0.05^2} = 16 < 42$$

donc $S_1(X)$ n'est pas dans la région critique, ainsi on ne peut rejeter $\mathcal{H}_0$ et l'hypothèse du fabricant est plausible.

(d) Sachant que pour tout $\alpha \in (0,1)$ et tout $n \in \mathbb{N}$, $F_n^{-1}(\alpha) > F_{n-1}^{-1}(\alpha)$, où F_n est la fonction de répartition d'une loi du χ^2_n, quelle statistique de test (parmi les deux proposées au (b)) est associée à la région critique la plus petite ? [5pts]

Avec $S_2(X)$, la région critique est $(F_n^{-1}(0.95), \infty) \subset (42.557, \infty)$, c'est donc $S_2(X)$.