Statistique mathématique – recueil d'exercices

Luca Ganassali

Université Paris-Saclay

2025 – 2026

(Last update: January 13, 2026)

1. Outils probabilistes pour le statisticien

Cette première feuille d'exercices est particulièrement longue. Il n'est pas attendu de tout traiter.

Exercice 1.1 (Maximum de variables exponentielles i.i.d.)**.** Soient $(X_n)_{n\geq 1}$ sont i.i.d. et distribuées exponentiellement avec paramètre $\lambda > 0$, c'est-à-dire de densité $x \mapsto \lambda e^{-\lambda x}\mathbb{1}_{x>0}$ par rapport à la mesure de Lebesgue sur $\mathbb{R}$. Pour tout $n \geq 1$, on définit

$$M_n := \max(X_1, \ldots, X_n)\,.$$

Equivalent presque sûr de M_n (optionnel).

1. À l'aide du Lemme de Borel-Cantelli, montrer que $\liminf \frac{M_n}{\log n} \geq 1/\lambda$, p.s.
2. À l'aide du Lemme de Borel-Cantelli, montrer que $\limsup \frac{X_n}{\log n} = 1/\lambda$ p.s.
3. Montrer finalement que $\frac{M_n}{\log n} \xrightarrow[n\to\infty]{p.s.} 1/\lambda$. *On pourra faire une preuve "à la Cesaro".*

Développement limité de M_n puis de $1/M_n$.

4. Montrer que $M_n - \frac{\log n}{\lambda}$ converge en loi vers une loi notée G_λ que vous caractériserez. *On passera par la fonction de répartition.*
5. En appliquant la méthode Delta et la question précédente, montrer que

$$\log n \times \left(\frac{\log n}{M_n} - \lambda\right) \xrightarrow[n\to\infty]{(d)} H_\lambda,$$

et préciser la fonction de répartition de H_λ.

Application: intervalle de confiance pour λ.

6. Pour $\beta \in]0,1[$, calculer explicitement le quantile d'ordre β de la loi de H_λ, noté $h_{\lambda,\beta}$.
7. Montrer avec tout ce qui précède que pour tout $\alpha \in]0,1[$, en notant

$$I_1(X_1, \ldots, X_n; \alpha) := \left[\frac{\frac{\log n}{M_n}}{1 + \frac{\log(-\log(\alpha/2))}{\log n}}, \frac{\frac{\log n}{M_n}}{1 + \frac{\log(-\log(1-\alpha/2))}{\log n}}\right],$$

on a

$$\mathbb{P}\left(I_1(X_1, \ldots, X_n; \alpha) \ni \lambda\right) \xrightarrow[n\to\infty]{} 1 - \alpha\,.$$

$I_1(X_1, \ldots, X_n)$ est un intervalle aléatoire appelé intervalle de confiance asymptotique pour λ de probabilité de couverture $1 - \alpha$.

8. Quel est l'équivalent presque sûr de $\mathrm{diam}(I_1(X_1, \ldots, X_n; \alpha))$?
9. ($\star$ car moins guidé) Pourrait-on trouver un autre intervalle de confiance dont le diamètre décroit sensiblement plus vite en n que l'intervalle précédent ?

Exercice 1.2 (Empirical risk minimisation)**.** On observe $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d., de loi inconnue, avec X_i vecteurs aléatoires de $\mathbb{R}^d$ (*features*), et $Y_i \in \{0,1\}$ (*labels*). On appelle *classifieur* une fonction mesurable $h : \mathbb{R}^d \to \{0,1\}$. Soit $\mathcal{H}$ un ensemble fini de classifieurs. On considère le *classification loss* ℓ défini par

$$\ell(h(X), Y) = \mathbb{1}_{\{h(X)\neq Y\}}\,.$$

Pour $h \in \mathcal{H}$, on définit son *risque* $R(h)$ et son *risque empirique* $\widehat{R}_n(h)$, donnés par :

$$R(h) = \mathbb{E}[\ell(h(X), Y)], \qquad \widehat{R}_n(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(h(X_i), Y_i)\,.$$

Le but du statisticien est de trouver le meilleur classifieur $h \in \mathcal{H}$ au sens du risque R:

$$h^\star \in \arg\min_{h \in \mathcal{H}} R(h)\,.$$

Problème : R dépend de la loi des données qui est inconnue : on ne peut pas calculer $R(h)$. La seule chose à laquelle nous avons accès, ce sont les données. On définit l'estimateur ERM $\hat{h}_n$ (empirical risk minimizer) par :

$$\hat{h}_n \in \arg\min_{h \in \mathcal{H}} \widehat{R}_n(h)\,.$$

Notons qu'on a par définition $R(h^\star) \le R(\hat{h}_n)$. Le but de l'exercice est d'établir une inégalité dans l'autre sens (avec des termes en plus) afin de montrer que $\hat{h}_n$ n'est pas trop mauvais par rapport à $h^\star$ au sens du risque R.

1. A l'aide de l'inégalité de Hoeffding, montrer que pour tout $h \in \mathcal{H}$ et tout $\varepsilon > 0$,

$$\mathbb{P}\left(\left|\widehat{R}_n(h) - R(h)\right| \ge \varepsilon\right) \le 2e^{-2n\varepsilon^2}.$$

2. En utilisant une *union bound*, montrer que pour tout $\varepsilon > 0$,

$$\mathbb{P}\left(\sup_{h \in \mathcal{H}}\left|\widehat{R}_n(h) - R(h)\right| \ge \varepsilon\right) \le 2|\mathcal{H}|\, e^{-2n\varepsilon^2}.$$

   En déduire qu'avec probabilité au moins $1 - \delta$,

$$\sup_{h \in \mathcal{H}}\left|\widehat{R}_n(h) - R(h)\right| \le \sqrt{\frac{1}{2n}\log\left(\frac{2|\mathcal{H}|}{\delta}\right)}\,.$$

3. Montrer enfin qu'avec probabilité au moins $1 - \delta$,

$$R(\hat{h}_n) \le R(h^\star) + 2\sqrt{\frac{1}{2n}\log\left(\frac{2|\mathcal{H}|}{\delta}\right)}.$$

   Interpréter.

Exercice 1.3 (Loi et espérance conditionnelle)**.** Soit $X \sim \mathrm{Exp}(1)$, et Y une variable aléatoire dont la densité conditionnelle par rapport à la mesure de comptage sur $\mathbb{N}$ est donnée par :

$$p_{Y\,|\,X}(y\,|\,x) = \frac{x^y e^{-x}}{y!}\,.$$

1. Trouver la densité jointe de (X, Y) et identifier la mesure dominante.

2. Quelle est la densité marginale de Y ? En déduire $\mathbb{E}[Y]$.

3. Reconnaissez-vous la loi de Y conditionnellement à X ? Vérifier que la loi de l'espérance totale donne la même valeur pour $\mathbb{E}[Y]$.

4. Trouver la densité conditionnelle de X sachant $Y = y$.

Exercice 1.4 (Fonction caractéristique des variables gaussiennes)**.** Nous voulons montrer que si $X \sim \mathcal{N}(\mu, \sigma^2)$, alors pour tout $t \in \mathbb{R}$,

$$\Phi_X(t) := \mathbb{E}[e^{itX}] = \exp\left(i\mu t - \frac{\sigma^2}{2}t^2\right) .$$

0. Montrer le résultat lorsque X est dégénérée ($\sigma = 0$). Nous supposons désormais $\sigma > 0$.

1. Montrer que pour tout $t \in \mathbb{R}$, $\Phi_X(t) = e^{it\mu} f_\sigma(t)$, où

$$f_\sigma(t) := \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{i\sigma t y - y^2/2} dy .$$

2. Montrer que pour tout $t \in \mathbb{R}$, $f_\sigma(t) \in \mathbb{R}$, que f_σ est différentiable et trouver une équation différentielle qu'elle satisfait.

3. Conclure.

4. En déduire que si $X \sim \mathcal{N}(\mu, \sigma^2)$, et $a, b \in \mathbb{R}$, alors $aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.

5. En déduire que, si $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ et X_1 et X_2 sont indépendantes, alors

$$X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2) .$$

Exercice 1.5 (Méthode Delta pour une loi multinomiale)**.** *Cet exercice suppose une connaissance préalables des propriétés des vecteurs gaussiens (que nous reverrons plus tard dans le cours).* Soit

$$(A_n, B_n, C_n) \sim \text{Multinomial}(n; p_1, p_2, 1 - p_1 - p_2)$$

avec $p_1, p_2 \in (0, 1)$ et $p_1 + p_2 < 1$. Pour tout $n \geq 1$, on définit

$$U_n := \frac{A_n}{n}, \quad V_n := \frac{B_n}{n}, \quad Z_n := \log\left(\frac{U_n}{V_n}\right) .$$

1. Montrer que

$$\sqrt{n}\big((U_n, V_n) - (p_1, p_2)\big) \xrightarrow[n\to\infty]{(d)} \mathcal{N}\big(0, \Sigma\big),$$

   où

$$\Sigma = \begin{pmatrix} p_1(1 - p_1) & -p_1 p_2 \\ -p_1 p_2 & p_2(1 - p_2) \end{pmatrix} .$$

2. En appliquant la méthode Delta à la fonction

$$g(x, y) = \log(x) - \log(y),$$

   montrer que

$$\sqrt{n}\left(Z_n - \log\left(\frac{p_1}{p_2}\right)\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}\Big(0, \frac{1}{p_1} + \frac{1}{p_2}\Big).$$

   Interpréter la symétrie dans la variance limite en p_1, p_2.

Exercice 1.6 (Plus rapide que $1/\sqrt{n}$?)**.** Supposons que $X_1, \ldots, X_n$ sont i.i.d., centrées, de variance finie $\sigma^2 > 0$ inconnue. On note classiquement $\bar{X}_n$ la moyenne empirique des X_i et

$$\hat{\sigma}^2 := \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 .$$

1. Étudier la convergence presque sûre de $\bar{X}_n$ et de $\hat{\sigma}^2$.

2. Montrer que

$$\frac{\sqrt{n}\bar{X}_n}{\sqrt{\hat{\sigma}^2}} \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1)\,.$$

Dans toute la suite, on suppose que $\sigma^2 = 1$ et qu'il est connu.

3. Quelle est la limite en loi de $\sqrt{n}(\cos(\bar{X}_n) - 1)$?

4. Trouver une suite $(a_n)_{n\geq 1}$ telle que $a_n(\cos(\bar{X}_n) - 1)$ converge en loi vers une v.a. Z qui n'est pas presque sûrement constante. Caractériser la loi de Z.

Exercice 1.7 (Une inégalité de type Bennett)**.** On note h la fonction définie sur $\mathbb{R}_+$ par $h(x) = (1+x)\log(1+x) - x$, et ϕ définie sur $\mathbb{R}$ par $\phi(x) = \exp(x) - x - 1$. Soit $X \sim \mathrm{Poi}(\theta)$ avec $\theta > 0$.

1. Pour tout $\lambda \in \mathbb{R}$, calculer $\mathbb{E}[\exp(\lambda(X-\theta))]$ en fonction de ϕ.

2. Montrer que pour tout $x > 0$,

$$\mathbb{P}(X - \theta \geq x) \leq \exp(-\theta h(x/\theta))\,.$$

3. On donne $e^{-2h(4)} \sim 0.003$. Comment cette inégalité se compare-t-elle à Bienaymé–Tchebychev pour $\mathbb{P}(X \geq 10)$ lorsque $\theta = 2$?

Exercice 1.8 (Canaux gaussiens)**.** Supposons que $Y \sim \mathcal{N}(t, \tau^2)$ et que, conditionnellement à $Y = y$, $X_1, \ldots, X_n$ soient i.i.d. de loi $\mathcal{N}(y, \sigma^2)$. Montrer que

$$Y \mid (X_1, \ldots, X_n) \sim \mathcal{N}\left(\frac{t/\tau^2 + n\bar{X}_n/\sigma^2}{1/\tau^2 + n/\sigma^2}, \frac{1}{1/\tau^2 + n/\sigma^2}\right)\,.$$

Pour une variable gaussienne, on appelle *précision* l'inverse de sa variance. Quelle propriété de la précision a-t-on mise en évidence ici ?

Exercice 1.9 (Loi bêta)**.** Supposons que X et Y soient des v.a. positives et indépendantes.

1. Montrer que pour tout $0 < x < 1$,

$$\mathbb{P}\left(\frac{X}{X+Y} \leq x\right) = \mathbb{E}\left[F_X\left(\frac{xY}{1-x}\right)\right]\,.$$

2. Supposons que $\mathbb{E}[Y] < \infty$ et que X admette une densité p_X par rapport à μ. Montrer que $V = X/(X+Y)$ admet aussi une densité p_V par rapport à μ, donnée par

$$p_V(x) = \mathbb{E}\left[\frac{Y}{(1-x)^2} p_X\left(\frac{xY}{1-x}\right)\right]\,.$$

Pour $\alpha, \beta > 0$, la loi Gamma $\Gamma(\alpha, \beta)$ est la loi de densité

$$x \mapsto \frac{1}{\Gamma(\alpha)} \beta^\alpha x^{\alpha-1} e^{-\beta x}$$

par rapport à $\mathrm{Leb}_{\mathbb{R}_+}$, et la loi bêta $\mathrm{Beta}(\alpha, \beta)$ est la loi de densité

$$x \mapsto \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

par rapport à $\mathrm{Leb}_{[0,1]}$.

3. Montrer que si $X \sim \Gamma(\alpha, 1)$ et $Y \sim \Gamma(\alpha', 1)$ sont indépendantes, alors $\frac{X}{X+Y} \sim \mathrm{Beta}(\alpha, \alpha')$.

Exercice 1.10 (Absence de géométrie dans les graphes d'Erdős-Rényi)**.** Pour $n \geq 1$ on note $[n] = \{1, 2, \ldots, n\}$. Soit $G \sim G(n, p)$ un graphe aléatoire Erdős-Rényi d'ensemble de sommets $[n]$, où chaque arête $e = \{u, v\} \in \binom{[n]}{2}$ est présente dans G indépendamment avec probabilité $p \in [0, 1]$. On note $e(G)$ le nombre d'arêtes de G.

1. Existe-t-il une valeur de p telle que $G \sim G(n, p)$ soit uniforme sur l'ensemble des graphes d'ensemble de sommets $[n]$?

2. Pour $G \sim G(n, p)$, quelle est la loi de $e(G)$?

3. Pour $G \sim G(n, p)$, quelle est la loi conditionnelle de G sachant $e(G)$?

4. Expliquer le titre de l'exercice.

2. Modélisation statistique, suffisance et complétude

Exercice 2.1. On considère le modèle où $(X_1, \dots, X_n)$ sont i.i.d. avec $X_1 \sim \mathrm{Unif}([\theta, \theta+1])$, avec $\theta > 0$ inconnu.

1. Écrire le modèle sous la forme canonique du cours. Montrer qu'il est identifiable.
2. Montrer que $S(X) = (\min_i X_i, \max_i X_i)$ est suffisante.
3. Montrer que $A(X) = \max_i X_i - \min_i X_i$ est auxiliaire. *On pourra chercher la loi de $X_1 - \theta$ sous $\mathbb{P}_\theta$.*
4. En déduire que $S(X) = (\min_i X_i, \max_i X_i)$ n'est pas complète.

Exercice 2.2. On considère un modèle dans lequel l'observation X est de la forme $X = a(Y - Z)$ avec $Y \sim \mathrm{Exp}(\lambda)$ et $Z \sim \mathrm{Exp}(\mu)$ indépendantes, $\lambda, \mu > 0$ inconnus.

On suppose d'abord a connu et $a = 1$.

1. Quels sont ici les paramètres du modèle ?
2. En posant $b = 1/\lambda$ et $c = 1/\mu$, trouver deux équations satisfaites par b et c. En raisonnant sur ces équations, montrer que le modèle est identifiable.

Supposons à présent que a est inconnu dans $\mathbb{R}$.

3. En raisonnant sur la transformation $(Y, Z) \mapsto (Z, Y)$, montrer que le modèle n'est pas identifiable.

Supposons finalement que a est inconnu dans $\mathbb{R}_+$. L'argument précédent ne tient plus.

4. Montrer que si $Y \sim \mathrm{Exp}(\lambda)$ et $\alpha > 0$, $Y \sim \mathrm{Exp}(\lambda/\alpha)$.
5. En déduire que le modèle n'est pas identifiable.

Exercice 2.3. On considère le modèle où $(X_1, \dots, X_n)$ sont i.i.d. avec densité $x \mapsto C(\theta)\, x \mathbb{1}_{[0,\theta]}(x)$ pour un certain paramètre $\theta > 0$.

1. Trouver la valeur de $C(\theta)$ pour tout $\theta > 0$.
2. Trouver la densité de X_1/θ.
3. On trie les X_i par ordre croissant : $X_{(1)} \leq \dots \leq X_{(n)}$ sont appelées les statistiques d'ordre de l'échantillon. Montrer que $X_{(n)}$ et $X_{(1)}/X_{(n)}$ sont indépendantes. *On utilisera le théorème de Basu, après avoir montré que $X_{(1)}/X_{(n)}$ est auxiliaire à l'aide de la question précédente.*

Exercice 2.4 (Identifier une direction causale, $\star$)**.** On considère un modèle avec deux variables aléatoires X_1, X_2. On dit que X_1 *cause* X_2 si la loi du modèle est de la forme suivante :

$$X_1 = U_1 \quad X_2 = f(X_1, U_2),$$

où U_1 et U_2 sont des bruits indépendants et f est une fonction mesurable.

Une question fondamentale est la suivante : *en général, peut-on inférer quoi que ce soit sur la direction du lien causal entre X_1 et X_2 en observant quelques réalisations ?*

L'objectif de cet exercice est de répondre à la négative via un contre-exemple à ce fait. On considère le modèle

$$X_1 = U_1 \quad X_2 = X_1 + U_2,$$

où U_1 et U_2 sont des variables i.i.d. de loi $\mathrm{Exp}(1)$. Montrer qu'il existe un modèle causal inversé de la forme

$$X_2 = U_2' \quad X_1 = h(X_2, U_1'),$$

où U_1' et U_2' sont indépendants, et qui préserve la loi jointe de (X_1, X_2).

Exercice 2.5. On observe une réalisation de $G \sim G(n, p)$, c'est-à-dire un graphe aléatoire à n sommets (n est connu) dont chaque arête est présente indépendamment avec probabilité $p \in [0, 1]$ inconnu. Trouver une statistique suffisante et complète. Interpréter.

Exercice 2.6. On considère le modèle $(\mathbb{R}, \mathcal{B}(\mathbb{R}), (\mathrm{Unif}([0, \theta]))_{\theta \geq 0})$. Trouver une mesure dominante minimale pour ce modèle.

Exercice 2.7 ($\star$)**.** Reprendre l'exemple du modèle uniforme i.i.d. du cours. Montrer que $S(X) = \max_i(X_i)$ est suffisante sans utiliser le théorème de Neyman-Fisher, en donnant la distribution de X sachant $S(X)$. Interpréter la forme de cette distribution.