TD 1 : rappels de proba et modèles statistiques

Les questions marquées d'un astérisque **(*)** *sont facultatives. N'hésitez pas à vous manifester si vous avez le moindre doute sur une question, une définition, une preuve...*

Rappels. Les variables aléatoires $X_1, \dots, X_n$ sont indépendantes si pour chaque n-uplet d'ensembles mesurables $(A_1, \dots, A_n)$,

$$\mathbb{P}(X_1 \in A_1, \dots, X_n \in A_n) = \mathbb{P}(X_1 \in A_1) \dots \mathbb{P}(X_n \in A_n).$$

On a alors pour toutes fonctions mesurables $\varphi_1, \dots, \varphi_n$,

$$\mathbb{E}[\varphi_1(X_1) \dots \varphi_n(X_n)] = \mathbb{E}[\varphi_1(X_1)] \dots \mathbb{E}[\varphi_n(X_n)],$$

du moment que ces espérances ont un sens.

Exercice 1. (Indépendance) Pour chacun des cas suivants, préciser la loi marginale des variables aléatoires X et Y, et dire si X et Y sont indépendantes.

1. $Z = (X, Y)$ est une variable aléatoire uniforme sur $[0, 1]^2$.
2. $Z = (X, Y)$ est une variable aléatoire uniforme sur $\{(x, y) \in [0, 1]^2, x \leq y\}$.
3. $Z = (U, V)$ est une variable aléatoire uniforme sur $[0, 1]^2$, $X = \min(U, V)$ et $Y = \max(U, V)$.
4. On lance deux fois un dé à six faces équilibré. X est le numéro du premier lancer, Y le numéro du second.
5. On dispose d'une urne avec n boules donc $1 \leq k \leq n - 1$ boules rouges. On tire deux boules successivement sans remettre la première boule dans l'urne avant de tirer la seconde. On note X la couleur de la première boule, Y la couleur de la seconde.
6. $X = 0$ et Y est une variable aléatoire gaussienne centrée réduite.
7. **(*)** On dispose de deux urnes remplies de boules. L'une contient une proportion p de boules rouges, l'autre une proportion $q \neq p$ de boules rouge. Au début de l'expérience, on choisit une urne au hasard puis on tire deux boules (avec remise entre les deux). On note X la couleur de la première boule et Y la couleur de la seconde.

Exercice 2. (Variance et covariance)

1. Démontrer la formule de König-Huygens : pour toute variable aléatoire X de variance finie, $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.
2. Soient $X_1, \dots, X_n$ des variables aléatoires identiquement distribuées de variances finies, et soit $S = \sum_{i=1}^n X_i$. A-t-on (et si oui, le démontrer) :
   — $\mathbb{E}[S] = n\mathbb{E}[X_1]$?
   — $\mathrm{Var}(S) = n\mathrm{Var}(X_1)$?

   Est-il possible d'avoir $\mathrm{Var}(S) = n^2\mathrm{Var}(X_1)$?
3. Refaire la question 2 dans le cas où $X_1, \dots, X_n$ sont de plus indépendantes.

Exercice 3. (Contre-exemples)

1. Soient X et Y deux variables aléatoires de variances finies. La covariance de X et Y est alors définie par $\mathrm{Cov}(X,Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$. L'implication
$$X \text{ et } Y \text{ sont indépendantes} \Rightarrow \mathrm{Cov}(X,Y) = 0$$
est-elle vraie ?
2. Montrer que la réciproque de la question 1 est fausse. On pourra considérer le cas où X est une variable aléatoire d'espérance nulle sur $\mathbb{R}$, Z est une variable aléatoire uniforme sur $\{-1,1\}$ indépendante de X et $Y = ZX$. *(Affaire à suivre au cours 4, « Vecteurs Gaussiens » : si (X,Y) est un vecteur gaussien, la réciproque et vraie !)*
3. **(*)** Soient X et Y deux variables aléatoires indépendantes uniformes sur $\{-1,1\}$. Soit $Z = XY$. Montrer que X et Z (respectivement Y et Z) sont indépendantes. Les variables X, Y et Z sont-elles indépendantes ?

Exercice 4. (Inégalité de Markov)

1. Démontrer l'inégalité de Markov : soit X une variable aléatoire réelle presque sûrement positive, alors pour tout $a > 0$,
$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$
On pourra utiliser que $\mathbf{1}_{X \geq a} + \mathbf{1}_{X<a} = 1$, où on note $\mathbf{1}_A$ l'indicatrice de l'ensemble A.
2. En déduire l'inégalité de Bienaymé-Tchebychev : soit X une variable aléatoire réelle de variance finie, alors pour tout $a > 0$,
$$\mathbb{P}(|X - \mathbb{E}[X]| \geq a) \leq \frac{\mathrm{Var}(X)}{a^2}.$$
3. En déduire que si $X_1, X_2, \ldots$ est une suite de variables aléatoires i.i.d. (indépendantes et identiquement distribuées) de variances finies, alors la suite des moyennes empiriques $(\frac{1}{n}\sum_{i=1}^n X_i)_{n\geq 1}$ converge en probabilité vers $\mathbb{E}[X_1]$.
4. **(*)** Montrer que si $X_1, X_2, \ldots$ est une suite de variables aléatoires i.i.d. telles que $\mathbb{E}[(X_1 - \mathbb{E}[X_1])^3]$ est finie, alors la suite des moyennes empiriques $(\frac{1}{n}\sum_{i=1}^n X_i)_{n\geq 1}$ converge presque sûrement vers $\mathbb{E}[X_1]$.

Exercice 5. (Lois classiques) Pour chacune des lois suivantes, préciser son support, sa densité (pour les mesures continues) ou sa fonction de masse (pour les mesures discrètes) ainsi que ses deux premiers moments (c'est-à-dire $\mathbb{E}[X]$ et $\mathbb{E}[X^2]$) et sa variance. *Demandez si vous ne connaissez pas une de ces lois. Les phrases en italique sont des commentaires, il n'est pas nécessaire de les démontrer.*

1. $\mathcal{B}(p)$, loi de Bernoulli de paramètre p,
2. $\mathcal{U}([0,1])$, loi uniforme sur $[0,1]$,
3. $\mathcal{N}(\mu,\sigma^2)$, loi normale sur $\mathbb{R}$ de paramètres (μ,σ^2).
(* mais à connaître) Quelle loi suit la somme de deux variables Gaussiennes indépendantes de paramètres (μ_1,σ_1^2) et (μ_2,σ_2^2) ? *Piste : utiliser les fonctions caractéristiques.*
4. $\mathrm{Bin}(n,p)$, loi binomiale de paramètres (n,p).
Cela compte le nombre de succès lors de n expériences indépendantes, chacune ayant la même probabilité de succès p : si $X_1, \ldots, X_n$ sont des variables aléatoires i.i.d. suivant la loi de Bernoulli de paramètre p, alors $\sum_{i=1}^n X_i$ suit la loi Binomiale de paramètres (n,p).
(*) Soit X une variable aléatoire suivant la loi binomiale de paramètres (n,p) et Y une variable aléatoire suivant la loi binomiale de paramètres (X,q). Donner un sens à cette définition puis montrer que Y suit la loi binomiale de paramètres (n,pq).

5. $\mathcal{G}(p)$, loi géométrique de paramètre p.
   Une variable suivant la loi géométrique de paramètre p compte le nombre de tirages jusqu'au premier succès (inclus) dans une suite de tentatives i.i.d. de probabilité de succès p.

6. $\mathcal{P}(\lambda)$, loi de Poisson de paramètre λ,
   - **(*)** Montrer que si Z suit la loi de Poisson de paramètre λ et $(X_n)_{n\geq 1}$ est une suite de variables aléatoires telles que pour tout $n \geq 1$, X_n suit la loi binomiale de paramètres $(n, \lambda/n)$, alors pour tout $k \in \mathbb{N}$, $\mathbb{P}(X_n = k) \to \mathbb{P}(Z = k)$.
   - En utilisant la question précédente, justifier la phrase « la loi de Poisson compte le nombre d'occurrences d'un évènement rare ».
   - **(*)** Montrer que si X et Y sont des variables indépendantes suivant respectivement les lois de Poisson de paramètres p et q, alors $X + Y$ suit la loi de Poisson de paramètre $p + q$.

   Cette loi est très utilisée en pratique, par exemple pour le nombre de photons atteignant le miroir d'un télescope, le nombre de pièces défectueuses sortant d'une usine, le nombre de buts dans un match de foot, la répartition des impacts des bombes allemandes V1 sur Londres pendant la seconde guerre mondiale [Clarke, R. D. (1946), "An application of the Poisson distribution"]...

7. $\mathcal{E}(\lambda)$, loi exponentielle de paramètre λ.
   (*) Montrer que la loi exponentielle est la seule loi sans mémoire définissant une durée de vie d'espérance $1/\lambda$, c'est-à-dire que si X est une variable aléatoire positive d'espérance $1/\lambda$ telle que pour tout s et t, $\mathbb{P}(X > s + t | X > t) = \mathbb{P}(X > s)$, alors X suit la loi exponentielle de paramètre λ.

Exercice 6. (Modèles statistiques) Parmi les propositions suivantes, lesquelles définissent un modèle statistique ? Lesquels sont paramétriques ? **(*)** Lesquels sont identifiables ?

— $\{\mathcal{U}([0, x]), x \in \mathbb{R}_+^*\}$;
— $\{\mathcal{N}(0, \sigma^2), (\mu, \sigma) \in \mathbb{R}^2\}$;
— $\{\text{Bin}(n, p), n \in \mathbb{N}^*, p \in]0, 1]\}$;
— $\{\mathcal{E}(\lambda), \lambda > 0\}$ (lois exponentielles) ;
— L'ensemble des fonctions intégrables sur $\mathbb{R}$;
— L'ensemble des mesures de probabilité sur $\mathbb{R}$.

Proposer un modèle statistique associé aux paragraphes suivants :

« On s'intéresse à la moyenne de la contamination des anguilles adultes par divers polluants pour tester si elle dépasse le seuil réglementaire. On mesure pour cela la concentration de polluant dans les tissus des anguilles pêchées, qu'on suppose gaussienne. Une étude préliminaire montre que l'écart-type de cette mesure est inférieure à une valeur connue σ_0. »

« André arrive aux caisses de son super-marché préféré, mais il y a beaucoup de queue. Pour ne pas s'ennuyer, il décide d'estimer le temps qu'il mettra avant de payer. On considère que la durée de passage d'une personne en caisse suit une loi exponentielle de paramètre inconnu. »

Exercice 7. (Un peu de bricolage) Un statisticien amateur décide de faire un sondage pour estimer la proportion des gens dans le monde qui ont assisté à un concert des Beatles. Pour ce faire, il demande à ses contacts sur son réseau social préféré de lui indiquer s'ils ont assisté à un concert des Beatles et de poser la même question à leurs contacts. Il parvient ainsi à rassembler environ 400 réponses.

Quels commentaires pouvez-faire sur la fiabilité des résultats qu'il obtiendra ? *On demande ici un raisonnement qualitatif, pas une démonstration.*