Feuille 1 : Rappels et estimateurs

Allan Merino

Master 1 Mathématiques, UE Statistiques, Janvier 2016

Quelques exercices de rappel à faire à la maison

Exercice 1

1. (Loi faible des grands nombres) Si $(X_n)_{n\geqslant 1}$ est une suite de variables aléatoires réelles indépendantes et identiquement distribuées telle que $\mathbb{E}(X_1) < +\infty$, alors, la suite de variables aléatoires $\left(\dfrac{S_n}{n}\right)_{n\geqslant 1}$ converge en loi vers la variable aléatoire constante égale à $\mathbb{E}(X_1)$.
2. (Théorème de la Limite Centrée) Si $(X_n)_{n\geqslant 1}$ est une suite de variables aléatoires réelles indépendantes et identiquement distribuées d'espérance μ ($\mu \in \mathbb{R}$) et de variance σ^2 ($\sigma \in \mathbb{R}_+^*$), alors, la suite de variables aléatoires réelles $\left(\dfrac{S_n - n\mu}{\sqrt{n}\sigma}\right)_{n\geqslant 1}$ converge en loi vers la loi normale de paramètres 0 et 1.
3. (Théorème de Moivre-Laplace) Si $(S_n)_{n\geqslant 1}$ est une suite de variables aléatoires réelles telle que pour tout $n \geqslant 1$, la variable aléatoire réelle S_n est de loi binomiale de paramètres n et p ($p \in]0,1[$), alors, la suite de variables aléatoires réelles $\left(\dfrac{S_n - np}{\sqrt{np(1-p)}}\right)_{n\geqslant 1}$ converge en loi vers la loi normale de paramètres 0 et 1.

Rappels : La variable aléatoire X est de loi Γ de paramètres a et λ (a, $\lambda \in \mathbb{R}_+^*$) si et seulement si sa loi de probabilité est absolument continue par rapport à la mesure de Lebesgue de dérivée de Radon-Nikodym

$$\frac{d\mathbb{P}_X(x)}{dx} = \frac{\lambda^a}{\Gamma(a)} x^{a-1} e^{-\lambda x} \mathbb{1}_{]0,+\infty[}(x).$$

On montre facilement que si X et Y sont deux variables aléatoires indépendantes de loi $\Gamma(a,\lambda)$ et $\Gamma(b,\lambda)$, alors, la variable aléatoire $X+Y$ suit une loi $\Gamma(a+b,\lambda)$.
Définition (loi du χ_2) : Une variable aléatoire X est de loi du chi-deux à n degrés de liberté ($n \geqslant 1$) si et seulement si X est de loi gamma de paramètres $\frac{n}{2}$ et $\frac{1}{2}$. On note souvent $X \sim \chi_{2,n}$.

Exercice 2

1. Soient X et Y deux variables aléatoires définies sur $(\Omega, \mathscr{F}, \mathbb{P})$. Déterminer la loi de probabilité de $X + Y$ dans les cas où
   — X et Y sont indépendantes de lois respectives $\epsilon(\alpha)$ et $\epsilon(\beta)$, $\alpha, \beta > 0$,
   — X et Y sont indépendantes de lois respectives $\mathscr{N}(\mu_1, \sigma_1^2)$ et $\mathscr{N}(\mu_2, \sigma_2^2)$, $\mu_1, \mu_2 \in \mathbb{R}, \sigma_1, \sigma_2 > 0$.
2. Plus généralement, montrer que si X et Y sont deux variables aléatoires indépendantes définies sur $(\Omega, \mathscr{F}, \mathbb{P})$, on a
$$f_{X+Y} = f_X \star f_Y.$$
3. Soit $X \sim \mathscr{N}(0, 1)$. Déterminer la loi de X^2.
4. Montrer que si $X_1, \ldots, X_n$ sont des variables aléatoires indépendantes identiquement distribuées de loi $\mathscr{N}(0, 1)$, alors $\sum_{k=1}^{n} \mathrm{X}_\mathrm{k}^2$ suit une loi du chi-deux à n degrés de liberté.

Exercice 3 : Critère des moindres carrés

Soient $x_j \in \mathbb{R}$ pour $j \in \{1, \ldots, n\}$ et $D = \{x_1, \ldots, x_n\}$. Posons $f_D : \mathbb{R} \to \mathbb{R}$ l'application définie par $f_D(x) = \frac{1}{n} \sum_{j=1}^{n} (x_j - x)^2$.
Montrer que l'unique minimum de f_D est $\bar{x}$, la moyenne des données D.

Définition Soit $p \in]0, 1[$.

1. Soit $D = \{x_1, \ldots, x_n\}$ un échantillon observé de taille n d'une variable statistique quantitative. On appelle p-quantile pour D un réel x_p tel que $Card(\{x_j; j \in \{1, \ldots, n\}$ et $x_j \leqslant x_p\} \geqslant np$ et $Card(\{x_j; j \in \{1, \ldots, n\}$ et $x_j \geqslant x_p\} \geqslant n(1 - p)$.
2. Soit f la densité d'une loi absolument continue sur $\mathbb{R}$. On dit que x_p est un p-quantile si $F(x_p) = \int_{-\infty}^{x_p} f(u)du = p$.

Exercice 4

1. Soit Y une variable aléatoire réelle de loi absolument continue, de densité f et de fonction de répartition F. Montrer que si F est strictement monotone sur $\{0 < F < 1\}$ et si $p \in]0, 1[$, alors il existe un unique p-quantile de Y.
2. Soit Z une variable aléatoire réelle qui suit une loi normale centrée réduite. Soient $z \in \mathbb{R}$ et $\{z_p, p \in]0, 1[\}$ l'ensemble de ses p-quantiles. Montrer que $\Phi(z) = 1 - \Phi(-z)$ (avec Φ la fonction de répartition de Z) ainsi que $z_p = -z_{1-p}$.
3. Soient $\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}_+^*$ et $X \sim \mathscr{N}(\mu, \sigma^2)$. Soit, de plus, x_p le p-quantile de X pour $p \in]0, 1[$. Montrer que $x_p = \mu + \sigma z_p$.
4. Expliquer la règle de "$68\% - 95\% - 99.7\%$" en calculant approximativement la probabilité pour que X prenne ses valeurs dans $[\mu - k\sigma, \mu + k\sigma]$ (appelée la $k\sigma$ - région), pour $k \in \{1, 2, 3\}$.

Quelques exercices pour débuter

Exercice 5 - Le modèle linéaire

Le but de cette étude consiste à rechercher une relation entre le pH x de divers lacs et le pH y de carpes vivant dans ces milieux. Le tableau ci-dessous présente la valeur x_i du pH du lac i et la valeur y_i du pH d'une carpe prélevée au hasard dans ce lac.

x_i	4	8	9	6	3.5	7	5
y_i	6.4	7.2	7.3	6.7	6.2	6.9	6.6

1. Représenter graphiquement le nuage de points correspondant à ces données. Quel type de relation semble lier les x_i et les y_i ?
2. La relation n'étant pas parfaite, on introduit un terme d'erreur ϵ_i. Nous supposons que les ϵ_i suivent une loi normale centrée et qu'ils sont indépendants entre eux. Exprimer, en fonction de x_i et ϵ_i, la loi de la variable aléatoire Y_i correspondant au pH d'une carpe du lac i.
3. Donner une structure statistique de cette expérience.

Rappels Une variable aléatoire absolument continue est de loi de Student à n degrés de liberté ($n \geqslant 1$) si et seulement si :

$$f_X(x) = \frac{\Gamma(\frac{(n+1)}{2})}{\Gamma(\frac{n}{2})} \frac{1}{\sqrt{n\pi}} \left(1 + \frac{x^2}{2}\right)^{-\frac{n+1}{2}} .$$

Exercice 6 - Loi de Student

Montrer que si X et Y sont deux variables aléatoires réelles indépendantes de lois respectives $\mathcal{N}(0,1)$ et $\chi_{2,n}$, alors,

$$T_n = \frac{\sqrt{n}X}{\sqrt{Y}}$$

est de loi de Student à n degrés de liberté.

Exercice 7

Soient $n \geqslant 2$ et $X_1, \ldots, X_n$ des variables aléatoires réelles indépendantes de moyenne μ et de variance $\sigma^2 > 0$. On pose :

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i, \quad \Sigma_\mathrm{n}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2,$$

$$\hat{\mathrm{S}}_\mathrm{n}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2, \quad \mathrm{S}_\mathrm{n}^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

On appelle $\bar{X}_n$ et S_n^2 l'espérance (reps. la variance) empirique.

1. Calculer l'espérance et la variance de $\bar{X}_n$ ainsi que l'espérance de S_n^2.
   A présent, on considère que pour tout j, X_j suit une loi normale $\mathcal{N}(\mu, \sigma^2)$.
2. Quelle est la loi de $\bar{X}_n$? Quelle est la loi de $\frac{n\Sigma_\mathrm{n}^2}{\sigma^2}$?
3. Montrer que $\frac{n\hat{\mathrm{S}}_\mathrm{n}^2}{\sigma^2} \sim \chi_{\mathrm{n}-1}^2$ et que $\bar{X}_n$ et $\hat{\mathrm{S}}_\mathrm{n}^2$ sont indépendantes.

4. En déduire la loi de $\sqrt{n}\dfrac{(\bar{X}_n - \mu)}{S_n}$.

Exercice 8

1. On effectue un comptage de population de bacilles dans cent cellules d'un individu. Les résultats sont présentés dans le tableau suivant :

Nombre x de bacilles par cellules	1	2	3	4	5	6	7	8	9	10	11
Nombre de cellules ayant x bacilles	5	15	21	20	18	10	2	3	3	1	2

   L'expérimentateur suppose que la variable aléatoire représentant le nombre de bacilles par cellule suit une loi de Poisson de paramètre λ inconnu. Donner la structure statistique correspondante.

2. On possède un dé (à 6 faces) truqué et on souhaite découvrir comment il est déséquilibré. On note $p_1, \ldots, p_6$ les probabilités respectives d'apparition des 6 faces. On effectue n lancers indépendants de ce dé et on observe, à chaque lancer, la face supérieure. Construire le modèle statistique adapté à ce modèle.

Estimation

Rappels Soit $\mathscr{S} = \{(\Omega, \mathscr{A}, \mathbb{P}_\theta)/\theta \in \Theta\}$ un modèle statistique. Soit $g : \theta \to \mathbb{R}^d$ une application. Soit $S : \mathscr{S} \to (\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$ une statistique d-dimensionnelle.

1. On dit que S est intégrable si
$$(\forall \theta \in \Theta) \qquad \mathbb{E}_\theta(\|S\|) = \int_\Omega \|S(\omega)\| d\mathbb{P}_\theta(\omega) < +\infty. \tag{1}$$
2. On dit qu'une statistique S est un estimateur sans biais de g si
$$(\forall \theta \in \Theta), \qquad \mathbb{E}_\theta(S) = g(\theta). \tag{2}$$
3. On dit qu'un estimateur sans biais de g est à dispersion minimale si pour tout $\theta \in \Theta$ et pour tout T estimateur sans biais de g, on a :
$$\mathbb{V}_\theta(S) \leqslant \mathbb{V}_\theta(T). \tag{3}$$

Par la suite, on utilisera la proposition suivante : Soit $S = (S_1, \ldots, S_d)$ un estimateur sans biais de g de carré intégrable. S est un estimateur sans biais à dispersion minimale de g si et seulement si pour tout $j = 1, \ldots, d$, pour tout $\theta \in \Theta$ et pour toute statistique X intégrable centrée (unidimensionnelle), on a $\mathbb{E}_\theta(S_j X) = 0$.

Soit $\mathscr{S} = (\Omega, \mathscr{A}, \mathbb{P}_\theta, \theta \in \Theta)$ un modèle statistique et $g : \Theta \to \mathbb{R}^d$ une application. Soit T_n une suite de statistique à valeurs dans $\mathbb{R}^d$. On dit que $(T_n)_{n\geqslant 1}$ est un estimateur de $g(\theta)$ faiblement convergeant si $T_n \to g(\theta)$ en probabilité par rapport à $\mathbb{P}_\theta$.

De nouveau, par la suite, on utilisera le lemme suivant : Soit $(T_n)_{n\in\mathbb{N}^*}$ une suite d'estimateurs sans biais de $g(\theta)$ telle que $(\forall \theta \in \Theta)$, $\lim\limits_{n\to+\infty} \mathbb{V}_\theta(T_n) = 0$. Alors, $(T_n)_{n\in\mathbb{N}^*}$ est faiblement convergeant.

Exercice 9

Soient $X_1, \ldots, X_n$ des variables aléatoires réelles indépendantes et identiquement distribuées.

1. Montrer que $\bar{X}_n$ est un estimateur sans biais de $g(\theta) = \mathbb{E}(X_1)$.
2. Montrer que $\bar{X}_n$ est un estimateur sans biais faiblement convergeant de $\mathbb{E}(X_1)$. Est-il fortement convergeant ?
3. Montrer que S_n^2 est un estimateur sans biais de $\mathbb{V}(X_1)$.
4. Montrer que S_n^2 est un estimateur sans biais faiblement convergeant de $g(\theta) = \mathbb{V}(X_1)$.

Exercice 10

Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes et identiquement distribuées d'espérance μ et de variance σ^2.

1. Donner la condition sur les constantes réelles $a_1, \ldots, a_n$ pour que $\sum\limits_{i=1}^{n} a_i X_i$ soit un estimateur sans biais de μ.
2. Parmi les estimateurs sans biais de cette forme, déterminer celui qui est de variance minimale et calculer sa variance.

Exercice 11

Soient $\Theta =]0, +\infty[$ et $\{\Omega, A, \mathbb{P}_\theta; \theta \in \Theta\} = \{(\mathbb{N}, \mathscr{P}(\mathbb{N}), \mathbb{P}_\theta); \theta \in \Theta\}$, avec $\mathbb{P}_\theta = \mathscr{P}(\theta)$, le modèle poissonien.

1. Donner un estimateur sans biais de θ.
2. Montrer qu'il n'existe pas d'estimateur sans biais de $\frac{1}{\theta}$.

Exercice 12

On considère le modèle constitué par les lois $\mathscr{B}(n, \theta)$ avec $\theta \in]0, 1[$ sur $\Omega = \{0, 1, \ldots, n\}$. On définit la variable X par $X(\omega) = \omega, (\forall \omega \in \Omega)$.

1. Soit f une fonction sur Ω telle que
$$(\forall \theta \in \Theta =]0, 1[, \mathbb{E}_\theta(f(X)) = 0.$$
Montrer que f est la fonction nulle.
2. Calculer $\mathbb{E}_\theta\left(\dfrac{X(X-n)}{n(n-1)}\right)$ et en déduire un estimateur sans biais de variance minimale de $-\theta(1-\theta)$.
3. On pose $n = 10$. Existe-t-il des estimateurs sans biais de variance minimale des fonctions suivantes du paramètre θ :
   (a) $\theta^2 - \theta^4$,
   (b) $\mathbb{P}_\theta(X \leqslant 4)$,
   (c) $exp(\theta)$.

Exercice 13

Un lac contient une proportion inconnue θ de poissons d'une certaine espèce E. On pêche des poissons dans ce lac jusqu'à obtenir n poissons de l'espèce E et on appelle X le nombre de poissons pêchés à ce moment-là. Autrement dit, le n^e poisson pêché de l'espèce E est le X^e poisson pêché. On admet que le nombre de poissons du lac est suffisamment grand pour que l'on puisse assimiler cette pêche à des tirages avec remise.

1. Calculer la loi de X.
2. Soient $\Theta = \{\theta \in]0,1[\}$ et $n > 1$. Montrer que $\mathbb{E}_\theta\left(\dfrac{n-1}{X-1}\right) = \theta$ et que $\mathbb{E}_\theta\left(\dfrac{n}{X}\right) \neq \theta$.
3. Montrer que pour $n > 1$, $\dfrac{n-1}{X-1}$ est un estimateur sans biais de variance minimale de θ.

Exercice 14

Soit $X_1, X_2, \ldots, X_n$ un échantillon de n valeurs mutuellement indépendantes, issu d'une population suivant une loi uniforme sur un intervalle $[a,b]$, où les bornes a et b sont inconnues. On pose $Z = min(X_1, \ldots, X_n)$, $T = max(X_1, \ldots, X_n)$ et $U = \dfrac{Z+T}{2}$.

1. Montrer que Z est un estimateur biaisé de a et asymptotiquement sans biais.
2. Montrer que T est un estimateur biaisé de a et asymptotiquement sans biais.
3. Montrer que U est un estimateur sans biais de a moyenne $\mu = \frac{a+b}{2}$.

Exercice 15

On considère un échantillon $X_1, \ldots, X_n$ de n variables aléatoires mutuellement indépendantes et suivant toutes une loi uniforme sur l'intervalle $[0,a]$, avec $a > 0$. Soit également X une variable aléatoire suivant la même loi uniforme sur $[0,a]$. Soit $\bar{X} = \sqrt[n]{X_1 \ldots X_n}$ la moyenne géométrique de l'échantillon.

1. Rappeler l'expression de $E(X)$.
2. Montrer que la fonction de densité de probabilité de la variable $Y = \sqrt[n]{X}$, de support $[0, \sqrt[n]{a}]$, est égale à $f_Y(x) = \frac{n}{a}x^{n-1}$.
3. Calculer $E(\bar{X})$. $\bar{X}$ est-il un estimateur sans biais de $E(X)$.
4. Calculer $E(\bar{X})$ en fonction de a pour $n = 1, 10, 100, 1000$. Que constatez-vous ?
5. Calculer la variance $V(\bar{X})$ de $\bar{X}$.
6. Déduire des questions précédentes, et à partir de $E(\bar{X})$, un estimateur sans biais de $E(X)$.