EXAMEN 17 janvier 2019
Statistique et apprentissage
M2 Probabilités et Modèles Aléatoires

Durée 3h.

Exercice 1. Rappel. Les intégrales $\int_0^{+\infty} x^{t-1}e^{-x}\mathrm{d}x$ sont notées $\Gamma(t)$ pour $t>0$. On sait que $\Gamma(t+1)=t\,\Gamma(t)$ et que si $t\in\mathbb{N}$, $\Gamma(t+1)=t!$

On observe n variables aléatoires réelles $X_1,\dots,X_n$ i.i.d. de densité commune $\frac{1}{2\lambda}e^{-|x|/\lambda}$, où λ est un paramètre réel strictement positif. On veut estimer λ en utilisant diverses méthodes.

(1) Préciser, pour $q\in\mathbb{N}^\star$, l'espérance et la variance de $|X_i|^q$.

(2) On pose

$$S_n=\frac{1}{nq!}\sum_{i=1}^{n}|X_i|^q.$$

Donner la vitesse et la loi limite de S_n. On pourra utiliser la notation

$$\sigma_q^2=\frac{(2q)!}{(q!)^2}-1.$$

(3) On fixe $q\in\mathbb{N}^\star$.

3.a Donner un estimateur $\hat{\lambda}_n$ consistant de λ.

3.b Trouver la vitesse et la loi limite de $\hat{\lambda}_n$.

3.c Proposer un test asymptotique de niveau $\alpha\in]0,1[$ de l'hypothèse $H_0:\lambda=1$ contre $H_1:\lambda>1$.

(4) On note $X_{(n)}=\max(X_1,\dots,X_n)$.

4.a Calculer la fonction de répartition de X_i et de $X_{(n)}$.

4.b Montrer que lorsque n tend vers l'infini, $X_{(n)}-\lambda\ln n$ a une loi limite dont on donnera la fonction de répartition.

4.c En déduire un second estimateur $\bar{\lambda}_n$ consistant de λ. Que pensez-vous de sa qualité ?

(5) Donner, à l'aide de l'inégalité de Bienaymé-Tchébychev et en utilisant à nouveau S_n, un intervalle de confiance de niveau $1-\alpha$ pour λ **lorsque** n **n'est pas grand**.

Solution de l'exercice 1.

(1) La moyenne est

$$\mathbb{E}\left[|X_i|^q\right]=\frac{1}{2\lambda}\int_0^{\infty}|x|^q\,e^{-|x|/\lambda}\,\mathrm{d}x=\lambda^q\,\Gamma(q+1)=\lambda^q\,q!.$$

On en déduit que la variance de $|X_i|^q$ est

$$\lambda^{2q}\left[\Gamma(2q+1)-(\Gamma(q+1))^2\right]=\lambda^{2q}((2q)!-(q!)^2).$$

(2) Par le CLT, la variable

$$\frac{\sum_{i=1}^n |X_i|^q - n\lambda^q\, q!}{\sqrt{n\,\lambda^{2q}((2q)! - (q!)^2)}} = \sqrt{n}\,\frac{S_n - \lambda^q}{\lambda^q\,\sigma_q}$$

converge en loi vers une $\mathcal{N}(0,1)$ à vitesse $\sqrt{n}$.

(3) 3.a Le calcul de la moyenne suggère de considérer

$$\hat{\lambda}_n := (S_n)^{\frac{1}{q}}$$

qui converge p.s. vers λ quand $n \to +\infty$ pour tout $\lambda > 0$.

3.b Nous avons vu que $\hat{\lambda}_n := (S_n)^{\frac{1}{q}} = J(S_n)$, avec $J'(s) = \frac{1}{q} s^{\frac{1}{q}-1}$. Par la δ-méthode $\hat{\lambda}_n$ a vitesse $\sqrt{n}$ et loi limite $J'(\lambda)\mathcal{N}(0,1)$

3.c On pose comme statistique de test la variable $\hat{\lambda}_n$ définie ci-dessus, et la région de rejet $R :=]k, +\infty[$, avec k à déterminer de façon à ce que pour $\lambda = 1$

$$\mathbb{P}(\hat{\lambda}_n > k) \leq \alpha.$$

Or pour n assez grand, on sait que la loi de $(\hat{\lambda}_n)^q$ est proche de $N(\lambda^q, \frac{1}{n}\lambda^{2q}\sigma_q^2)$ et donc pour $\lambda = 1$ et $Z \sim \mathcal{N}(0,1)$

$$\mathbb{P}\left(\hat{\lambda}_n > k\right) = \mathbb{P}\left((\hat{\lambda}_n)^q > k^q\right) \approx \mathbb{P}\left(1 + \frac{\sigma_q}{\sqrt{n}} Z > k^q\right)$$

et donc on peut choisir

$$k = \left(1 - q_\alpha \frac{\sigma_q}{\sqrt{n}}\right)^{1/q}$$

où $\mathbb{P}(Z < q_\alpha) = \alpha$.

(4) 4.a On a pour $t \leq 0$

$$\mathbb{P}(X_i \leq t) = \frac{1}{2\lambda} \int_{-\infty}^t e^{\frac{x}{\lambda}}\, \mathrm{d}x = \frac{1}{2} e^{\frac{t}{\lambda}}$$

et pour $t > 0$

$$\mathbb{P}(X_i \leq t) = \frac{1}{2} + \frac{1}{2\lambda} \int_0^t e^{-\frac{x}{\lambda}}\, \mathrm{d}x = 1 - \frac{1}{2} e^{-\frac{t}{\lambda}}.$$

Donc pour $t \leq 0$

$$\mathbb{P}\left(X_{(n)} \leq t\right) = (\mathbb{P}(X_1 \leq t))^n = \left(\frac{1}{2} e^{t/\lambda}\right)^n$$

et pour $t > 0$

$$\mathbb{P}\left(X_{(n)} \leq t\right) = (\mathbb{P}(X_1 \leq t))^n = \left(1 - \frac{1}{2} e^{-t/\lambda}\right)^n$$

4.b Donc pour tout $t \in \mathbb{R}$ et n suffisamment grand

$$\mathbb{P}\left(X_{(n)} \leq t + \lambda \ln n\right) = \left(1 - \frac{1}{2n} e^{-t/\lambda}\right)^n \to \exp\left(-\frac{1}{2} e^{-t/\lambda}\right)$$

c'est à dire une loi du Gumbel.

4.c Nous obtenons que $\bar{\lambda}_n := \frac{X_{(n)}}{\ln n}$ converge en probabilité vers λ pour toute valeur de $\lambda > 0$. On vient de voir que

$$\ln n \left(\bar{\lambda}_n - \lambda\right)$$

converge en loi. L'estimateur $\bar{\lambda}_n$ a donc vitesse $\ln n$, alors que $\hat{\lambda}_n$ a vitesse $\sqrt{n}$ par la δ-méthode.

(5) On a pour $t > 0$

$$\mathbb{P}(|S_n - \lambda^q| > t) \leq \frac{\mathrm{Var}(S_n)}{t^2} = \frac{\lambda^{2q}\,\sigma_q^2}{nt^2}.$$

Pour $t = \frac{\lambda^q\,\sigma_q}{\sqrt{n\alpha}}$ alors nous obtenons que

$$\mathbb{P}(|S_n - \lambda^q| \leq t) \geq 1 - \frac{\lambda^{2q}\,\sigma_q^2}{nt^2} = (1-\alpha).$$

Donc

$$\mathbb{P}\left(\lambda^q \in \left[S_n - \frac{\lambda^q\,\sigma_q}{\sqrt{n\alpha}}, S_n + \frac{\lambda^q\,\sigma_q}{\sqrt{n\alpha}}\right]\right) \geq 1 - \alpha.$$

Or, nous avons

$$\lambda^q \geq S_n - \frac{\lambda^q\,\sigma_q}{\sqrt{n\alpha}} \Longrightarrow \lambda^q \left(1 + \frac{\sigma_q}{\sqrt{n\alpha}}\right) \geq S_n$$

$$\lambda^q \leq S_n + \frac{\lambda^q\,\sigma_q}{\sqrt{n\alpha}} \Longrightarrow \lambda^q \left(1 - \frac{\sigma_q}{\sqrt{n\alpha}}\right) \leq S_n$$

mais il est fort probable que $\left(1 - \frac{\sigma_q}{\sqrt{n\alpha}}\right) < 0$. Nous obtenons au moins que

$$\mathbb{P}\left(\lambda^q \geq \left(1 - \frac{\sigma_q}{\sqrt{n\alpha}}\right)^{-1} S_n\right) \geq 1 - \alpha$$

et donc

$$\mathbb{P}\left(\lambda \geq \left(1 - \frac{\sigma_q}{\sqrt{n\alpha}}\right)^{-\frac{1}{q}} S_n^{\frac{1}{q}}\right) \geq 1 - \alpha.$$

Exercice 2. Une chaîne de télévision a programmé une série avec K émissions. Pour suivre l'audience, on dispose d'un panel de téléspectateurs indépendants entre eux, de taille n supposée élevée, de l'ordre de quelques milliers d'individus (en pratique, $n = 6000$). Ce panel permet, pour l'émission numéro k ($k = 1, \ldots, K$) de la série, de connaître le nombre N_k d'individus ayant regardé cette émission k. On suppose que chaque individu a probabilité $p \in]0, 1[$ de regarder l'émission numéro k de la série, pour chaque $k = 1, \ldots, K$.

1. Pour k fixé entre 1 et K, on considère dans les trois questions qui suivent le modèle statistique à une observation N_k.

1.a Quelle est la loi de N_k ? Que valent $\mathbb{E}N_k$ et $\mathrm{Var}(N_k)$?

1.b Calculer $\hat{p}_k$, l'estimateur du maximum de vraisemblance de p.

1.c Donner un intervalle de confiance asymptotique de niveau $1 - \alpha = 95\%$ pour p.

2. On considère maintenant le modèle à K observations $N_1, \ldots, N_K$, supposées indépendantes.

   2.a Calculer $\hat{p}$, l'estimateur de p par la méthode du maximum de vraisemblance.

   2.b Exprimer $\hat{p}$ en fonction de $\hat{p}_1, \ldots, \hat{p}_K$.

   2.c Donner la loi asymptotique de $\hat{p}$ lorsque n tend vers l'infini, K restant fixé. En déduire un intervalle de confiance approché, de niveau $1 - \alpha = 95\%$, pour p.

Solution de l'exercice 2.

(1) N_k est une binomiale de paramètres (n, p). La fonction de vraisemblance ici est

$$L_n(n_k, p) = \binom{n}{n_k} p^{n_k}(1-p)^{n-n_k},$$

et on obtient que l'estimateur du maximum de vraisemblance de p est

$$\hat{p}_k = \frac{N_k}{n}.$$

Donc on sait que pour n assez grand un intervalle de confiance de niveau $(1-\alpha)$ pour p est

$$\left[\hat{p}_k - q_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_k(1-\hat{p}_k)}{n}}, \hat{p}_k + q_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_k(1-\hat{p}_k)}{n}}\right]$$

(2) La fonction de log-vraisemblance de ce modèle statistique est

$$\log L(n_1, \ldots, n_K; p) = \sum_{i=1}^{L}(n_i \ln p + (n - n_i)\ln(1-p)) + C(n, n_1, \ldots, n_K)$$

et nous obtenons donc

$$\hat{p} = \frac{N_1 + \cdots + N_K}{nK} = \frac{1}{K}\sum_{i=1}^{K}\hat{p}_k.$$

La loi asymtotique de $\hat{p}_k$ est $\mathcal{N}(p, \frac{p(1-p)}{n})$ et la loi asymptotique de $\hat{p}$ est donc $\hat{p} \approx \mathcal{N}(p, \frac{p(1-p)}{Kn})$. Donc on sait que pour n assez grand un intervalle de confiance de niveau $(1-\alpha)$ pour p est

$$\left[\hat{p} - q_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{Kn}}, \hat{p} + q_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{Kn}}\right].$$

Exercice 3. On considère un couple de variables aléatoires (X, Y) à valeurs dans $\mathbb{R}^3 \times \{0, 1\}$. Les trois composantes de X sont notées T, B et E, respectivement. La variable T modélise le nombre d'heures hebdomadaires moyen qu'un étudiant passe devant la télévision et la variable B le nombre moyen de sorties festives nocturnes qu'il effectue chaque semaine. La composante E représente quant à elle une quantité positive inaccessible, qui regroupe quantitativement tous les facteurs négatifs pouvant conduire notre étudiant à l'échec (manque de travail, absence de motivation, etc.). Pour des raisons évidentes, la variable aléatoire E ne peut donc être mesurée.

Finalement, la variable aléatoire Y représente simplement le succès universitaire de l'étudiant : 1 s'il réussit son examen et 0 s'il échoue. On suppose que

$$Y = \begin{cases} 1 & \text{si } T+B+E<7 \\ 0 & \text{autrement.} \end{cases}$$

On fait également l'hypothèse que T, B et E sont indépendantes et suivent chacune une loi exponentielle de paramètre 1. La règle de Bayes associée au couple $((T,B),Y)$ est notée $g^\star(T,B)$.

(1) Que vaut $L^\star$, l'erreur de Bayes associée au couple $((T,B,E),Y)$?

(2) Donner l'expression de $\mathbb{P}(Y=1\,|\,T,B)$.

(3) En déduire l'expression de $g^\star(T,B)$.

(4) Prouver que, pour $x\geq 0$,

$$\mathbb{P}(T+B>x)=\int_x^{+\infty} ue^{-u}\mathrm{d}u=(1+x)e^{-x}.$$

Calculer explicitement $\mathbb{P}(g^\star(T,B)\neq Y)$.

(5) Reprendre les questions précédentes avec cette fois-ci $\mathbb{P}(Y=1\,|\,T)$, $g^\star(T)$ et $\mathbb{P}(g^\star(T)\neq Y)$.

(6) Calcul l'erreur que commet un étudiant (présomptueux) qui décide que $Y=1$, indépendamment des valeurs prises par T et B.

Solution de l'exercice 3.

(1) Ici $L^\star=0$, car la fonction $g(t,b,e)=\mathbf{1}_{(t+b+e<7)}$ satisfait $\mathbb{P}(g(T,B,E)\neq Y)=0$.

(2) Nous avons

$$\mathbb{P}(Y=1\,|\,T,B)=\mathbb{P}(E<7-B-T\,|\,T,B)=\max\left(0,1-e^{-(7-T-B)}\right).$$

(3) La règle de Bayes $g^\star(T,B)$ est donnée par

$$g^\star(T,B)=\begin{cases} 1 & \text{si } \mathbb{P}(Y=1\,|\,T,B)>1/2 \\ 0 & \text{sinon} \end{cases} = \mathbf{1}_{(e^{-(7-T-B)}<1/2)}=\mathbf{1}_{(T+B<7-\ln 2)}$$

(4) On a, en notant $Y:=T+B$,

$$\begin{aligned}
&\mathbb{P}(g^\star(T,B)\neq Y)=\mathbb{P}\left(\mathbf{1}_{(Y<7-\ln 2)}\neq \mathbf{1}_{(Y+E<7)}\right)= \\
&=\mathbb{P}(Y<7-\ln 2,\ Y+E\geq 7)+\mathbb{P}(Y\geq 7-\ln 2,\ Y+E<7) \\
&=\mathbb{E}\left[\mathbf{1}_{(Y<7-\ln 2)}\,e^{-(7-Y)}\right]+\mathbb{E}\left[\mathbf{1}_{(7-\ln 2\leq Y<7)}\left(1-e^{-(7-Y)}\right)\right] \\
&=\int_0^{7-\ln 2} xe^{-x}e^{-7+x}\,dx+\int_{7-\ln 2}^{7} xe^{-x}\left(1-e^{-7+x}\right)\,dx \\
&=e^{-7}\left[\frac{(7-\ln 2)^2}{2}+2(8-\ln 2)-8-\frac{1}{2}\left(7^2-(7-\ln 2)^2\right)\right]
\end{aligned}$$

(5) Nous avons

$$\mathbb{P}(Y=1\,|\,T)=\mathbb{P}(E+B<7-T\,|\,T)=\left(1-(8-T)e^{-(7-T)}\right)\mathbf{1}_{(T<7)}.$$

La règle de Bayes $g^\star(T)$ est donnée par

$$g^\star(T)=\begin{cases} 1 & \text{si } \mathbb{P}(Y=1\,|\,T)>1/2 \\ 0 & \text{sinon} \end{cases} = \mathbf{1}_{(T<7,\,(8-T)e^{-(7-T)}<1/2)}=\mathbf{1}_{(T<t_0)}$$

où $t_0 \in]0,7]$ est la seule solution de $(8-t)e^{-(7-t)} = 1/2$. Donc $\mathbb{P}(g^\star(T) \neq Y)$ est donné, en notant $Z := B + E$, par

$$\begin{aligned}
&\mathbb{P}\left(\mathbf{1}_{(T<t_0)} \neq \mathbf{1}_{(T+B+E<7)}\right) = \mathbb{P}(T \geq t_0,\, T+B+E<7) + \mathbb{P}(T<t_0,\, T+B+E \geq 7)\\
&= \mathbb{P}(t_0 \leq T < 7,\, Z < 7-T) + \mathbb{P}(T<t_0,\, Z \geq 7-T)\\
&= \int_{t_0}^{7} e^{-x}\left(1-(8-x)e^{-7+x}\right)dx + \int_0^{t_0} e^{-x}e^{-7+x}\,dx\\
&= e^{-t_0} - e^{-7} - 8(7-t_0)e^{-7} + \frac{7^2-t_0^2}{2}e^{-7} + t_0 e^{-7}
\end{aligned}$$

(6) On a

$$L = \mathbb{P}\left(1 \neq \mathbf{1}_{(T+B+E<7)}\right) = \mathbb{P}(T+B+E \geq 7) = \frac{1}{2}\int_7^\infty u^2 e^{-u}\,\mathrm{d}u.$$