Cursus Intégré
2004-2005

Rappels de statistique mathématique
Corrigé des travaux dirigés n° 1

Guillaume Lacôte
Bureau **E03**
✉ Guillaume.Lacote@ensae.fr
☞ http://ensae.no-ip.com/SE222/

Version du 20050421-15h38, révisée le 21 avril 2005

Corrigé de l'exercice 1

☞ Q1 Par définition $Y_i^* \rightsquigarrow \mathcal{N}(m, \sigma^2)$,
donc $\frac{Y_i^* - m}{\sigma} \rightsquigarrow \mathcal{N}(0,1)$
et par suite

$$\begin{aligned}\mathbb{P}\left(Y_i^* \geq 0\right) &= \mathbb{P}\left(\frac{Y_i^* - m}{\sigma} \geq -\frac{m}{\sigma}\right) \\ &= 1 - \Phi\left(-\frac{m}{\sigma}\right) \\ &= \Phi\left(\frac{m}{\sigma}\right)\end{aligned}$$

puisque la distribution de la loi normale est symétrique. Ainsi

$$Y_i \rightsquigarrow \mathcal{B}\left(1, \Phi\left(\frac{m}{\sigma}\right)\right)$$

☞ Q2 Le couple (m, σ^2) n'est clairement pas identifiable, car

$$\mathcal{B}\left(1, \Phi\left(\frac{m}{\sigma}\right)\right) = \mathcal{B}\left(1, \Phi\left(\frac{2m}{2\sigma}\right)\right)$$

En revanche, le rapport $\frac{m}{\sigma}$ l'est (en vertu du théorème de factorisation).

$$\begin{array}{ccc} & \star & \\ \star & & \star \end{array}$$

Corrigé de l'exercice 2

☞ Q1 – Si $X \rightsquigarrow \mathcal{E}(\lambda)$, alors $\forall x > 0,\ \mathbb{P}(X > x) = \int_x^{+\infty} \lambda \exp(-\lambda t) dt = e^{-\lambda x}$
– La réciproque s'obtient par dérivation de l'égalité précédente :
$\forall x \in \mathbb{R}, \quad f(x)$ existe et vaut $\lambda e^{-\lambda x} \mathbb{1}_{\mathbb{R}^+}(x)$,
donc toute variable aléatoire réelle vérifiant l'inégalité précédente admet une densité qui est celle de $\mathcal{E}(\lambda)$.

☞ Q2 – On a successivement

$$\begin{aligned}\mathbb{P}(Z > t) &= \mathbb{P}(X_1 > t \wedge X_2 > t) \quad \text{par définition de } Z \\ &= \mathbb{P}(X_1 > t)\mathbb{P}(X_2 > t) \quad \text{par indépendance} \\ &= e^{-(\lambda_1 + \lambda_2)t}\end{aligned}$$

Donc Z suit une loi $\mathcal{E}(\lambda_1 + \lambda_2)$.

– Par définition

$$\begin{aligned}\mathbb{P}(X_2 > X_1) &= \int\int \mathbb{1}_{(v>u)}.\lambda_1 e^{-\lambda_1 u}\lambda_2 e^{-\lambda_2 v} du dv \\ &= \int_0^{+\infty} e^{-\lambda_2 t}\lambda_1 e^{-\lambda_1 t} dt \\ &= \frac{\lambda_1}{\lambda_1+\lambda_2}\end{aligned}$$

☞ Q3 (a) Il s'agit de n observations i.i.d de loi $\mathcal{E}(\lambda_1+\lambda_2)$.
Modèle : $\{\mathbb{R}^{+n}, \mathcal{E}(\lambda_1+\lambda_2)^{\otimes n}, (\lambda_1,\lambda_2) \in \mathbb{R}^{+2}\}$
Vraisemblance : $L_n = (\lambda_1+\lambda_2)^n e^{-(\lambda_1+\lambda_2)\sum Z_i}$
Bien entendu seule la somme $(\lambda_1+\lambda_2)$ est identifiable.

(b) Cette fois, on dispose de la variable supplémentaire $S_i = \mathbb{1}_{(X_2>X_1)}$. Alors

$$\begin{aligned}\mathbb{P}(Z > t \,|\, S_i = 1) &= \mathbb{P}(X_2 > X_1 > t) \\ &= \int_{\mathbb{R}}\int_{\mathbb{R}} \mathbb{1}_{(u>v>t)}\lambda_1 e^{-\lambda_1 u}\lambda_2 e^{-\lambda_2 v} du dv \\ &= \int_t^{+\infty}\left(\int_v^{+\infty} \lambda_1 e^{-\lambda_1 u} du\right)\lambda_2 e^{-\lambda_2 v} dv \\ &= \int_t^{+\infty}\left(e^{-\lambda_1 v}\right)\lambda_2 e^{-\lambda_2 v} dv \\ &= \frac{\lambda_1}{\lambda_1+\lambda_2}\exp(-(\lambda_1+\lambda_2)t)\end{aligned}$$

De même, $\mathbb{P}(Z > t \,|\, S_i = 0) = \frac{\lambda_2}{\lambda_1+\lambda_2}\exp(-(\lambda_1+\lambda_2)t)$
Les densités correspondantes s'en déduisent (au signe près) par dérivation selon t :

$$\begin{aligned}L_n &= \prod_{s_i=1}\{\lambda_1 e^{-(\lambda_1+\lambda_2)z_i}\}\prod_{s_i=0}\{\lambda_2 e^{-(\lambda_1+\lambda_2)z_i}\} \\ &= \lambda_1^{n_1}\lambda_2^{n_2} e^{-(\lambda_1+\lambda_2)\sum_{i=1}^n z_i}\end{aligned}$$

où $n_1 = |\{i \,/\, s_i = 1\}|$ et $n_2 = |\{i \,/\, s_2 = 1\}| = n - n_1$.
Le couple (λ_1,λ_2) est donc identifiable, et on peut donc envisager d'estimer à la fois λ_1 et λ_2.

☞ Q4 (a) La loi $\Gamma(k,\lambda)$ admet pour densité $f(x) = \frac{\lambda^k}{(k-1)!}x^{k-1}e^{-\lambda x}\mathbb{1}_{x>0}$.
En particulier, toute loi $\mathcal{E}(\lambda)$ est une loi $\Gamma(1,\lambda)$.
On a en outre : *la somme de deux variables indépendantes de lois* $\Gamma(k,\lambda)$ *et* $\Gamma(l,\lambda)$ *suit une loi* $\Gamma(k+l,\lambda)$.

Cette propriété se prouve par récurrence sur k : [1]
soit $U \rightsquigarrow \Gamma(k,\lambda)$ et $V \rightsquigarrow \Gamma(1,\lambda)$; alors $W = U + V$ a pour densité :

$$\begin{aligned} f_W(w) &= \int_{u+v=w} \frac{\lambda^k}{(k-1)!} u^{k-1} e^{-\lambda u} \lambda e^{-\lambda v} du dv \\ &= \int_0^{+\infty} \frac{\lambda^k}{(k-1)!} u^{k-1} e^{-\lambda u} \lambda e^{-\lambda(w-u)} du \\ &= \frac{\lambda^{k+1}}{(k-1)!} e^{-\lambda w} \int_0^{+\infty} u^{k-1} du = \frac{\lambda^{k+1}}{k!} e^{-\lambda w} \end{aligned}$$

de sorte que $W \rightsquigarrow \Gamma(k+1,\lambda)$.
Ainsi la somme de $n \geq 1$ variables indépendantes qui suivent une loi exponentielle de même paramètre λ suit une loi $\Gamma(n,\lambda)$.

(b) On a $Z = \min\left(\sum_{i=0}^{n_1} X_1^i, \sum_{j=0}^{n_2} X_2^j\right)$ où $(\sum_{i=0}^{n_1} X_1^i) \rightsquigarrow \Gamma(n_1+1,\lambda_1)$ et $(\sum_{i=0}^{n_2} X_2^i) \rightsquigarrow \Gamma(n_2+1,\lambda_2)$.
Donc

$$\begin{aligned} \mathbb{P}(Z \geq t) &= \mathbb{P}\left(\left(\sum_{i=0}^{n_1} X_1^i \geq t\right) \text{ et } \left(\sum_{i=0}^{n_2} X_2^i \geq t\right)\right) \\ &= \mathbb{P}\left(\sum_{i=0}^{n_1} X_1^i \geq t\right) \times \mathbb{P}\left(\sum_{i=0}^{n_2} X_2^i \geq t\right) \quad \text{par indépendance} \\ &= \left(\int_t^{+\infty} \frac{\lambda_1}{n_1!} e^{-\lambda_1 x} (\lambda_1 x)^{n_1} dx\right) \times \left(\int_t^{+\infty} \frac{\lambda_2}{n_2!} e^{-\lambda_2 x} (\lambda_2 x)^{n_2} dx\right) \end{aligned}$$

(c) Si $X \rightsquigarrow \mathcal{E}(\lambda)$, alors :

$$\begin{aligned} \mathbb{P}(X \geq s+t | X \geq t) &= \frac{\mathbb{P}(X \geq s+t)}{\mathbb{P}(X \geq t)} \\ &= \frac{\exp(-\lambda(s+t))}{\exp(-\lambda t)} \\ &= e^{-\lambda s} \\ &= \mathbb{P}(X \geq s) \end{aligned}$$

Cette propriété caractérise la loi exponentielle.
En l'occurence, $Z_0 = \min(X_1^0, X_2^0) \rightsquigarrow \mathcal{E}(\lambda_1 + \lambda_2)$
Supposons que la première panne ait lieu au temps t (donc $Z_0 = t$) ; supposons (sans perte de généralité) qu'elle touche une machine de type 1. On a alors :

$$(Z_1 | Z_0 = t \wedge S_0 = 1) = (\min(X_1^1, X_2^0) | X_2^0 > t)$$

[1] Elle se montre aussi à l'aide des fonctions caractéristiques.

où X_1^1 désigne le temps de fonctionnement d'une nouvelle machine de type 1 (et suit donc une $\mathcal{E}(\lambda_1)$), et X_2^0 désigne le temps jusqu'à la première panne de l'"ancienne" machine de type 2.
Or $(X_2^0|X_2^0 > t)$ suit une loi $\mathcal{E}(\lambda_2)$ (la loi exponentielle est "sans mémoire"), et on a a nouveau

$$(Z_1|Z_0 = t \wedge S_0 = 1) \rightsquigarrow \mathcal{E}(\lambda_1 + \lambda_2)$$

Un raisonnement similaire dans le cas d'une panne de type 2 montre que la loi de Z_1 est indépendante de Z_0 et S_0 :

$$Z_i \rightsquigarrow \mathcal{E}(\lambda_1 + \lambda_2)(i.i.d)$$

(d) *Remarque* : Attention à la convention de l'exercice : il y a (n_1+1) machines de type 1 : n_1 sont en stock, et une est déjà présente dans le système. Il y a donc panne générale quand le stock est épuisé pour un type de machine, **et qu'en plus la dernière machine du même type tombe en panne**.
On a $\min(n_1, n_2) \leq N \leq n_1 + n_2 - 1$. En outre N ne peut prendre la valeur $k \in [\![\min(n_1, n_2, n_1 + n_2 - 1]\!]$ que dans deux cas :
– les n_1 machines de types 1 sont sorties du stock, $(k - n_1)$ machines de type 2 sont sorties du stock, et la dernière machine de type 1 tombe en panne (rappelons qu'il y a $(n_1 + 1)$ machines de type 1).
– cas symétrique en échangeant type 1 et type 2.
Notons $p_1 = \frac{\lambda_1}{\lambda_1+\lambda_2}$ et$p_2 = \frac{\lambda_2}{\lambda_1+\lambda_2}$. Une nouvelle panne a une probabilité p_1 de concerner une machine de type 1, p_2 de type 2.
On a alors

$$\forall k \geq n_1, n_2, \quad \mathbb{P}(N = k) = p_1 C_k^{n_1} p_1^{n_1} p_2^{k-n_1} + p_2 C_k^{n_2} p_2^{n_2} p_1^{k-n_2}$$

Ainsi, si par exemple $n_1 \leq k < n_2$, on a $\mathbb{P}(N = k) = p_1 C_k^{n_1} p_1^{n_1} p_2^{k-n_1}$.

(e) On a

$$\begin{aligned} \mathbb{E}(Z|N = n) &= \mathbb{E}\left(\sum_{i=0}^{n} Z_i\right) \\ &= (n+1)\mathbb{E}(Z_1) \\ &= \frac{n+1}{\lambda_1 + \lambda_2} \end{aligned}$$

Or $\mathbb{E}(Z) = \mathbb{E}(\mathbb{E}(Z|N))$, donc finalement

$$\mathbb{E}(Z) = \frac{\mathbb{E}(N) + 1}{\lambda_1 + \lambda_2}$$

$\star$
$\star \quad \star$

Corrigé de l'exercice 3

☞ Q1 *Rappel :* la statistique T est dite *exhaustive ssi* la loi conditionnelle $\mathbb{P}_\theta\left(X \in \cdot | T\left(X\right)\right)$ est indépendante de θ.

Théorème de factorisation : T est exhaustive *ssi*

$$\exists g_\theta, h : (\mathbb{R} \to \mathbb{R}) \ / \ \forall x \in \mathbb{R},\ p_\theta(x) = g_\theta\left(T\left(x\right)\right) h(x)$$

où g_θ dépend de θ mais pas h.

On retrouve les statistiques exhaustives en écrivant la vraisemblance du modèle et en appliquant le théorème de factorisation.

Dans le cas présent le paramètre du modèle est $\theta \in \Theta = \mathbb{R}$. On a pour $x \in \mathbb{R}^n$

$$L_n(x) = e^{-n\lambda} \frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}$$

D'où une statistique exhaustive : $S(x_1, ..., x_n) = \sum_{i=1}^n x_i$.

☞ Q2 Le paramètre du modèle étant $(\alpha, \theta) \in \Theta = \mathbb{R}^2$, on a

$$L_n(x) = \left(\frac{\alpha - 1}{\theta}\right)^n \frac{\theta^{n\alpha}}{\exp(\alpha \sum_{i=1}^n \log(x_i))} \mathbb{1}_{[\theta, +\infty]}(\min x_i)$$

D'où **une** statistique exhaustive :

$$S(x_1, ..., x_n) = \left(\sum_{i=1}^n \log(x_i), \min x_i\right)$$

☞ Q3 Cette fois le paramètre du modèle est $(\alpha, \theta) \in \Theta = \mathbb{R}^{+*2}$. On a

$$L_n(x) = \alpha^n \theta^n \left(\prod_{i=1}^n x_i\right)^{\alpha - 1} \exp\left(-\theta \sum_{i=1}^n x_i^\alpha\right) \ \mathbb{1}_{\mathbb{R}^+}(\min_i x_i)$$

Dans ce cas, on ne peut exhiber de statistique exhaustive autre que $T(X_1, \ldots, X_n) = (X_1, \ldots, X_n)$. En revanche, si on estimait uniquement θ (α constante connue), alors $T(x) = \sum_{i=1}^n x_i^\alpha$ serait une statistique exhaustive.

☞ Q4 La densité[2] de la loi $\mathcal{U}_{[0,\theta]}$ est

$$f_{\mathcal{U}_{[0,\theta]}}(x) = \frac{1}{\theta} \mathbb{1}_{x \geq 0} \mathbb{1}_{x \leq \theta}$$

[2]En toute rigueur le modèle $\left(\mathcal{U}_{[0,\theta]}\right)_{\theta \in \mathbb{R}^{+*}}$ n'est pas dominé (il n'y a pas une unique mesure μ qui domine toutes les probabilités $\mathcal{U}_{[0,\theta]}$).

Mais pour tout $M > 0$, le sous-modèle $\left(\mathcal{U}_{[0,\theta]}\right)_{\theta \in]0,M]}$ est dominé (par la mesure particulière qu'est la probabilité de $\mathcal{U}_{[0,M]}$). On exhibe donc ici une statistique $T(X)$ qui est exhaustive pour tout sous-modèle : elle est donc telle que

$$\forall M, \forall \theta \in]0, M], \text{ la loi de } X|T(X) \text{ ne dépend pas de } \theta$$

et donc

$$\forall \theta > 0, \text{ la loi de } X|T(X) \text{ ne dépend pas de } \theta$$

et donc $T(X)$ est bien exhaustive dans le modèle originel.

et on a donc

$$\begin{aligned} L_n(x_1,\ldots,x_n) &= \frac{1}{\theta^n}\mathbb{1}_{x_1,\ldots,x_n\geq 0}\mathbb{1}_{x_1,\ldots,x_n\leq\theta} \\ &= \frac{1}{\theta^n}\mathbb{1}_{\min x_i\geq 0}\mathbb{1}_{\max x_i\leq\theta} \end{aligned}$$

et une statistique exhaustive est donc $T(x)=\max_{i\in[\![1,n]\!]} x_i$

$\star$

$\star \quad \star$

Corrigé de l'exercice 4

☞ Q1 On applique $n-1$ fois la *formule du conditionnement* $\mathbb{P}(\mathcal{A}\cap\mathcal{B})=\mathbb{P}(\mathcal{A}|\mathcal{B})\mathbb{P}(\mathcal{B})$:
Ici :

$$\begin{aligned} & \mathbb{P}(Y_n=y_n,\ldots,Y_1=y_1) \\ = \;& \mathbb{P}(Y_n=y_n|Y_{n-1}=y_{n-1,}\ldots,Y_1=y_1)\times\mathbb{P}(Y_{n-1}=y_{n-1,}\ldots,Y_1=y_1) \end{aligned}$$

$\vdots$ (par récurrence)

$$= \begin{cases} & \mathbb{P}\left(Y_n=y_n|Y_{n-1}=y_{n-1},\ldots,Y_1=y_1\right) \\ \times & \mathbb{P}\left(Y_{n-1}=y_{n-1}|Y_{n-2}=y_{n-2},...,Y_1=y_1\right) \\ \vdots & \\ \times & \mathbb{P}\left(Y_1=y_1\right) \end{cases}$$

☞ Q2 Soit p le nombre de poissons distincts qui ont été tirés au cours des n premiers tirages. Supposons que $R_n=r_n$: il y a eu r_n retirages d'un poisson déjà tiré au moins une fois, donc $p+r_n=n$. Ainsi après $n-1$ tirages le nombre de poissons distincts tirés est $n-1-r_{n-1}$, et la proportion de poissons qui ont été tirés (et sont donc marqués) est

$$\boxed{\frac{n-1-r_{n-1}}{\theta}}$$

Par conséquent

$$\mathbb{P}\left(Y_n=y_n|R_{n-1}=r_{n-1}\right) = \left(\frac{n-1-r_{n-1}}{\theta}\right)^{y_n}\left(1-\frac{n-1-r_{n-1}}{\theta}\right)^{1-y_n}$$

Donc d'après le résultat précédent la vraisemblance s'écrit

$$\begin{aligned}
& \mathbb{P}(Y_1 = y_1, \ldots, Y_n = y_n; \theta) \\
= & \prod_{i=1}^{n} \mathbb{P}(Y_i = y_i | Y_{i-1} = y_{i-1}, \ldots, Y_1 = y_1) \\
= & \prod_{i=1}^{n} \mathbb{P}\left(Y_i = y_i | R_{i-1} = r_{i-1}, Y_{i-1} = (r_{i-1} - y_{i-1} - \cdots - y_1), Y_{i-2} = y_{i-2}, \ldots, Y_1 = y_1\right) \\
= & \prod_{i=1}^{n} \mathbb{P}(Y_i = y_i | R_{i-1} = r_{i-1}) \quad \text{car la loi de } Y_i | R_{i-1}, Y_{i-1}, \ldots, Y_1 \text{ ne dépend pas de } Y_1, \ldots, Y_{i-1} \\
= & \prod_{i=1}^{n} \mathbb{P}\left(Y_i = y_i | Y_{i-1} = y_{i-1}, \ldots, Y_1 = y_1\right) \\
= & \prod_{i=1}^{n} \left(\left(\frac{i - 1 - r_{i-1}}{\theta} \right)^{y_i} \left(\frac{\theta - i + 1 + r_{i-1}}{\theta} \right)^{1-y_i} \right) \\
= & \prod_{i=1}^{n} \left(\frac{1}{\theta} (i - 1 - r_{i-1})^{y_i} (\theta - i + 1 + r_{i-1})^{1-y_i} \right)
\end{aligned}$$

qui est proportionnel à

$$\boxed{\prod_{i=1}^{n} \tfrac{1}{\theta} (\theta - i + 1 + r_{i-1})^{1-y_i}}$$

par le facteur $\prod_{i=1}^{n} (i - 1 - r_{i-1})^{y_i}$ qui ne dépend pas de θ et n'a donc pas d'incidence sur la détermination d'une statistique exhaustive (d'après le théorème de factorisation).[3]

☞ Q3 On a successivement

$$\prod_{i=1}^{n} \frac{(\theta - i + 1 + r_{i-1})^{1-y_i}}{\theta} = \frac{1}{\theta^n} \prod_{i=1}^{n} (\theta - (i - 1 - r_{i-1}))^{1-y_i}$$

Or $\forall i \in [\![1, n]\!]$, $(y_i = 1)$:$(\theta - (i - 1 - r_{i-1}))^{1-y_i} = 1$, donc

$$\begin{aligned}
= & \frac{1}{\theta^n} \times \prod_{\left\{ \begin{array}{l} 1 \le i \le n \\ y_i = 1 \end{array} \right.} 1 \times \prod_{\left\{ \begin{array}{l} 1 \le i \le n \\ y_i = 0 \end{array} \right.} (\theta - (i - 1 - r_{i-1})) \\
= & \frac{1}{\theta^n} \prod_{\left\{ \begin{array}{l} 1 \le i \le n \\ j_i = i - 1 - r_{i-1} \\ y_i = 0 \end{array} \right.} (\theta - j_i)
\end{aligned}$$

[3] Ce facteur aurait bien entendu une influence pour le calcul de l'information de Fisher ou du maximum de vraisemblance.

Or $\phi : i \mapsto i - 1 - r_{i-1}$ associe au nombre de tirages le nombre de poissons distincts tirés. Donc restreinte à $\mathcal{I}_0 = \{i \leq n \,/\, y_i = 0\}$ (l'ensemble des i tels que le i-ième poisson tiré est un nouveau poisson pas encore marqué), c'est une bijetion de $\mathcal{I}_0$ sur $[\![\phi(1), \phi(i_0)]\!]$ où $i_0 = \max \mathcal{I}_0$. Donc

$$\prod_{\left\{\begin{array}{l} 1 \leq i \leq n \\ j_i = i - 1 - r_{i-1} \\ y_i = 0 \end{array}\right.} (\theta - j_i) = \prod_{1 \leq j \leq \phi(i_0)} (\theta - j)$$

Remarquons enfin que par définition de i_0, $\forall j > i_0, j \notin \mathcal{I}_0$ et donc $y_j = 1$ de sorte que

$$\phi(i_0) = i_0 - 1 - r_{i_0} = (i_0 + (j - i_0)) - 1 - (r_{i_0} + (j - i_0)) = j - 1 - r_{i_0 + (j - i_0)} = j - 1 - r_j$$

de sorte que $\phi(i_0) = n - 1 - r_n$ et en définitive la vraisemblance est proportionnelle à

$$\frac{1}{\theta^n} \prod_{1 \leq j \leq n-1-r_n} (\theta - j) \quad = \quad \boxed{\frac{1}{\theta^n} \frac{(\theta-1)!}{(\theta-n-1+r_n)!}}$$

☞ Q4 On a alors

$$l(y, \theta) = g_\theta(R_n(y)) \times h(y)$$

où

$$h(y) = \prod_{i=1}^{n} (i - 1 - R_{i-1}(y))^{y_i}$$

qui est indépendant de θ, et

$$g_\theta(R_n(y)) = \frac{1}{\theta^n} \frac{(\theta - 1)!}{(\theta - n - 1 + R_n(y))!}$$

qui ne dépend de y qu'au travers de $R_n(y)$.
Donc d'après le théorème de factorisation R_n est une statistique exhaustive. Ainsi pour estimer le nombre de poissons dans le lac, la connaissance de tous les tirages est superflue et le seul nombre total de poissons tirés plus d'une fois suffit.

⋆
⋆ ⋆