Cursus Intégré
2004-2005

Rappels de statistique mathématique
Réponses question par question 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

Exercice corrigé 1

On dispose d'observations Y_i relatives au comportement de remboursement ou de non-remboursement d'emprunteurs :

$$Y_i = \begin{cases} 1 & \text{si l'emprunteur } i \text{ rembourse son crédit} \\ 0 & \text{si l'emprunteur } i \text{ est défaillant} \end{cases}$$

Afin de modéliser ce phénomène, on suppose l'existence d'une variable aléatoire Y_i^* normale, d'espérance m et de variance σ^2, qu'on appellera "capacité de remboursement de l'individu i", telle que :

$$Y_i = \begin{cases} 1 & \text{si } Y_i^* \geq 0 \\ 0 & \text{si } Y_i^* < 0 \end{cases}$$

☞ Q1 On note Φ la fonction de répartition de la loi $\mathcal{N}(0,1)$.
Exprimer la loi de Y_i en fonction de Φ.

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 Les paramètres m et σ^2 sont-ils identifiables ?

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).

Exercice corrigé 2

Un système S fonctionne en utilisant deux machines de types différents. Les durées de vie X_1 et X_2 des deux machines suivent des lois exponentielles de paramètres λ_1 et λ_2. Les variables aléatoires X_1 et X_2 sont supposées indépendantes.

☞ Q1 Montrer que
$$X \rightsquigarrow \mathcal{E}(\lambda) \Leftrightarrow \forall x \in \mathbb{R},\ \mathbb{P}(X > x) = \exp(-\lambda x)$$

– 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 Calculer la probabilité pour que le système ne tombe pas en panne avant la date t.
En déduire la loi de la durée de vie Z du système.
Calculer la probabilité pour que la panne du système soit due à une défaillance de la machine 1.

– 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 On dispose de n systèmes $S_1, ..., S_n$ identiques dont on observe les durées de vie $Z_1, ...Z_n$.

(a) Ecrire le modèle statistique correspondant et la vraisemblance des observations.
A-t-on suffisamment d'information pour estimer λ_1 et λ_2 ?

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) Si on observe à la fois les durées de vie des systèmes et la cause de la défaillance (machine 1 ou 2), écrire le modèle statistique correspondant et la vraisemblance des observations. A-t-on alors suffisamment d'information pour estimer λ_1 et λ_2 ?

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 Dans cette question, on considère un seul système S utilisant une machine de type 1 et une machine de type 2, mais on suppose que l'on dispose d'un stock de n_1 machines de type 1, de durées de vie $X_1^1, ..., X_1^{n_1}$, et d'un stock de n_2 machines de type 2, de durées de vie $X_2^1, ..., X_2^{n_2}$. Quand une machine tombe en panne, on la remplace par une machine du même type, tant que le stock de machines de ce type n'est pas épuisé. Quand cela arrive, on dit que le système S lui-même est en panne. On note toujours Z la durée de vie du système.
Le cas $n_1 = n_2 = 0$ correspond donc à la première question (pas de stock).

(a) Donner la loi de la somme de n variables indépendantes qui suivent une loi exponentielle de même paramètre λ.

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) Ecrire Z en fonction des X_j^i et en déduire $P(Z \geq t)$ en fonction de certaines lois gamma, dont on précisera les paramètres.

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

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

On note alors N le nombre de machines (des deux types) sorties du stocks quand le système tombe en panne, et Z_0 la durée écoulée avant la première panne d'une machine. On note Z_i la durée écoulée entre la i-ème panne et la $(i+1)$-ème panne. La durée de vie totale du système est donc :

$$Z = \sum_{i=0}^{N} Z_i$$

La $(N+1)$-ème panne est donc la panne fatale au système.

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

(c)

Montrer que les variables Z_i sont i.i.d. et donner leur loi.
On pourra utiliser (après l'avoir démontré) le résultat suivant :
Si X est une variable aléatoire de loi exponentielle de paramètre λ, alors

$$\mathbb{P}\left(X \geq s+t | X \geq s\right) = \mathbb{P}\left(X \geq t\right) = e^{-\lambda t}$$

(on dit que X est "sans mémoire").

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 :

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

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) Préciser l'ensemble des valeurs possibles pour la variable N et en donner la loi.

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 admet que N et les Z_i sont indépendantes. Calculer $\mathbb{E}(Z|N)$ en fonction de N, λ_1 et λ_2. Donner l'expression de $\mathbb{E}(Z)$ en fonction de $\mathbb{E}(N)$, λ_1 et λ_2.

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

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

Exercice corrigé 3

Ecrire la vraisemblance et déterminer une statistique exhaustive pour un échantillon de n observations i.i.d. de lois :

☞ Q1 loi de Poisson de paramètre λ :

$$P(X=k) = e^{-\lambda}\frac{\lambda^k}{k!}, \ \ k \in \mathbb{N}^*$$

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

Théorème de factorisation : T est exhaustive *ssi*

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

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

loi de Pareto de paramètres α et θ avec $\alpha > 1$, $\theta > 0$ de densité :

$$f(x) = \frac{\alpha - 1}{\theta} \left(\frac{\theta}{x}\right)^{\alpha} \mathbb{1}_{[\theta;+\infty[}(x)$$

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

loi de Weibull de paramètre α et θ avec $\alpha > 0, \theta > 0$ de densité :

$$f(x) = \alpha\theta x^{\alpha-1} e^{-\theta x^{\alpha}} \mathbb{1}_{[0;+\infty[}(x)$$

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 loi uniforme sur $[0, \theta]$ avec $\theta > 0$ inconnu.

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

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

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 $T(x) = \max_{i\in[\![1,n]\!]} x_i$

Exercice corrigé 4

On veut compter le nombre θ de poissons dans un lac fermé. Pour cela, on tire un poisson au hasard, on le marque et on le remet dans le lac. On tire un second poisson. S'il est déjà marqué, on en prend note et on le remet dans le lac. Sinon, on le marque à son tour et on le remet dans le lac. Et ainsi de suite.

On tire n poissons selon la procédure ci-dessus. Au n-ième tirage, l'observation consiste en une variable aléatoire Y_n qui vaut 1 si le poisson est déjà marqué, 0 sinon. Par définition, on a $Y_1 = 0$. Le but de l'exercice est de montrer que :

$$R_n = \sum_{i=1}^{n} Y_i$$

est une statistique exhaustive pour θ.

☞ Q1 Montrer que :

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

[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.

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 :

$$\begin{aligned} & \mathbb{P}(Y_n=y_n,\ldots,Y_1=y_1) \\ = \;& \mathbb{P}(Y_n=y_n|Y_{n-1}=y_{n-1,}\ldots,Y_1=y_1)\times\mathbb{P}(Y_{n-1}=y_{n-1,}\ldots,Y_1=y_1) \\ \vdots \;& \text{(par récurrence)} \\ = \;& \left\{\begin{array}{ll} & \mathbb{P}\left(Y_n=y_n|Y_{n-1}=y_{n-1},\ldots,Y_1=y_1\right) \\ \times & \mathbb{P}\left(Y_{n-1}=y_{n-1}|Y_{n-2}=y_{n-2},...,Y_1=y_1\right) \\ \vdots & \\ \times & \mathbb{P}\left(Y_1=y_1\right) \end{array}\right. \end{aligned}$$

☞ Q2

Montrer que la loi conditionnelle de Y_n sachant $R_{n-1}=r_{n-1}$ est une Bernoulli de paramètre :

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

et en déduire que la vraisemblance est proportionnelle à :

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

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

Montrer que l'expression `(1)` se réécrit :

$$\frac{1}{\theta^n} \frac{(\theta - 1)!}{(\theta - n - 1 + r_n)!}$$

On a successivement

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

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

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

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

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

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 En déduire que R_n est une statistique exhaustive pour θ.

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.