Processus de Poisson homogènes
Application à des données génomiques

Mélisande ALBERT - Nicolas OGOREK

P. REYNAUD-BOURET

Table des matières

Introduction

Les études de phénomènes aléatoires au cours du temps sont aujourd'hui légions, que ce soit dans le domaine de la physique nucléaire, de la biologie cellulaire ou bien encore dans des situations concrètes de la vie courante, comme pour l'étude du congestionnement d'une centrale téléphonique (dépendant du processus des appels téléphoniques qui se produisent à des instants aléatoires).

Les processus de Poisson (du nom du mathématicien français Siméon Denis Poisson, XIX$^{\text{ème}}$ siècle) sont des processus ponctuels, les plus simples à étudier. Nous nous contenterons d'approfondir ceux dits homogènes, c'est à dire de paramètre constant : l'apparition d'évènements est équilibrée au cours d'une période d'étude (concrètement, on peut imaginer que le nombre de coups de fils au cours d'une journée n'augmente pas brusquement à l'heure du déjeuner), s'opposant ainsi aux processus de Poisson dits inhomogènes.

Les évènements particuliers modélisés seront par la suite appelés des tops : ils peuvent être temporels, par exemple s'ils représentent le moment d'entrée d'une personne dans un établissement donné (comme une banque), l'apparition d'un tremblement de terre ; ou bien spatiaux, comme la position des gènes sur la chaîne d'ADN...

Ainsi, après avoir défini ce qu'est réellement un processus de Poisson homogène, nous illustrerons ce concept en essayant de modéliser la position des gènes et de séquences promotrices dans une bactérie (Escherichia Coli) à partir de son ADN, et ainsi voir si la présence des ces gènes peut être ou non assimilé à un processus de Poisson. Pour cela nous utiliserons les tests de Kolmogorov - Smirnov (voir [5]) et le logiciel de statistique R (voir [6]) développé par le Cran (the Comprehensive R Archive Network) (voir [7]).

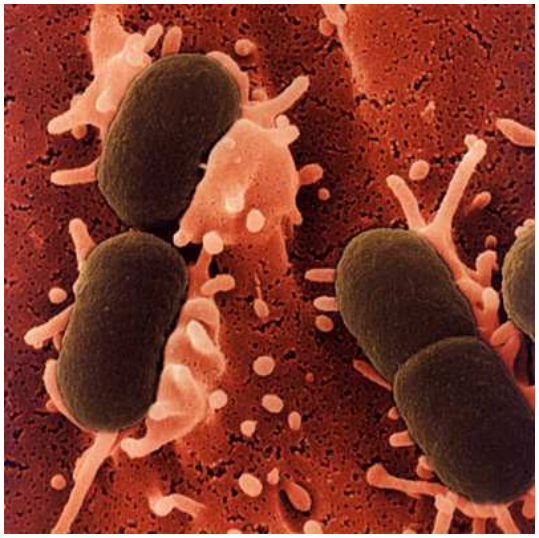


Fig. 1 – Escherichia Coli

1 Processus de Poisson homogènes

1.1 Définitions préliminaires

Commençons par énoncer quelques définitions utiles, trouvées dans les livres de références [1], [2] et [3] :

Définition.

Un processus stochastique est une fonction aléatoire $t \mapsto X_t$.

Définition.

Désignons par $N(t)$ le nombre de tops se produisant dans l'intervalle de temps $[0, t]$, et supposons que $N(0) = 0$. Le processus $\{N(t)\ ;\ t \geq 0\}$, est appelé processus de comptage et vérifie :

- $\star$ $\forall t \geq 0,\ N(t) \in \mathbb{N}$;
- $\star$ $t \mapsto N(t)$ est croissante ;
- $\star$ $\forall\ 0 < a < b,\ N(b) - N(a)$ représente le nombre de tops se produisant dans l'intervalle de temps $]a, b]$.

Définition.

Un processus de comptage est dit à accroissements stationnaires si la loi de probabilité du nombre de tops se produisant dans un intervalle de temps donné ne dépend que de la longueur de celui-ci.

Définition.

Un processus de comptage est dit à accroissements indépendants si les nombres de tops se produisant dans des intervalles de temps disjoints sont indépendants.

1.2 Processus de Poisson homogènes - Première approche

Définition 1.

Un processus de comptage $\{N(t)\ ;\ t \geq 0\}$ est appelé processus de Poisson d'intensité $\lambda > 0$ si :

a) $N(0) = 0$;

b) le processus est à accroissements indépendants ;

c) le nombre de tops se produisant dans un intervalle de temps de longueur $t \geq 0$ suit une loi de Poisson de paramètre λt, *ie*

$$\forall s \geq 0, \forall t \geq 0, \forall n \in \mathbb{N}, \quad \mathbb{P}(N(s+t) - N(s) = n) = e^{-\lambda t} \frac{(\lambda t)^n}{n!}$$

Définition 2.

Un processus de comptage $\{N(t)\ ;\ t \geq 0\}$ est appelé processus de Poisson d'intensité $\lambda > 0$ si :

i) $N(0) = 0$;

ii) le processus est à accroissements indépendants, et stationnaires ;

iii) $\mathbb{P}(N(h) = 1) = \lambda h + o(h)$ pour $h \to 0$;

iv) $\mathbb{P}(N(h) \geq 2) = o(h)$ pour $h \to 0$.

Remarque : La définition 2 est plus générale que la définition 1 car elle demande principalement que le processus soit indépendant et stationnaire. Comme nous allons le voir, la loi de Poisson découle des hypothèses iii) et iv).

Théorème 1.

Les définitions 1 et 2 sont équivalentes.

Démonstration :

(1 ⇒ 2) (cf. référence[3])

Soit $\{N(t)\ ;\ t \geq 0\}$ défini par la définition 1 et montrons qu'il vérifie les propriétés de la définition 2.

i) C'est a).

ii) On sait que le processus est à accroissements indépendants par b), et le processus est à accroissements stationnaires car on voit bien que seule la longueur de l'intervalle t intervient dans c).

iii) On fait un développement limité pour $h \to 0$:

$$\begin{array}{rcll} \mathbb{P}(N(h)=1) & = & \lambda h e^{-\lambda h} & \text{d'après c)} \\ & = & \lambda h(1+o(1)) & \text{(développement limité de } e^{-\lambda h} \text{ pour } h \to 0) \\ & = & \lambda h + o(h) & \end{array}$$

iv) On a, pour h au voisinage de 0 :

$$\begin{array}{rcll} \mathbb{P}(N(h)\geq 2) & = & \sum_{k\geq 2} \mathbb{P}(N(h)=k) & \\ & = & \sum_{k\geq 2} e^{-\lambda h}\frac{(\lambda h)^k}{k!} & \text{(d'après c))} \\ & = & e^{-\lambda h}(\sum_{k\geq 0}\frac{(\lambda h)^k}{k!} - 1 - \lambda h) & \text{(on somme sur } \mathbb{N} \text{ puis on retire les deux premiers termes)} \\ & = & e^{-\lambda h}(e^{\lambda h} - 1 - \lambda h) & \\ & = & 1 - e^{-\lambda h}(1+\lambda h) & \\ & = & 1 - (1-\lambda h + o(h))(1+\lambda h) & \text{(développement limité de } e^{-\lambda h} \text{ pour } h \to 0) \\ & = & 1 - 1 - \lambda h + \lambda h + o(h) & \\ & = & o(h) & \end{array}$$

(2 ⇒ 1) (cf. référence[1])

Réciproquement, considérons $\{N(t)\ ;\ t \geq 0\}$ défini par la définition 2.
Montrons qu'il vérifie les propriétés de la définition 1.

a) C'est i).

b) C'est d'après ii).

c) Pour montrer qu'une variable aléatoire $N(t)$ vérifiant la définition 2 suit une loi de Poisson, nous utiliserons le fait que la transformée de Laplace caractérise la loi.

Tout d'abord, calculons la transformée de Laplace d'une loi de Poisson.
Soit X une variable aléatoire réelle suivant une loi Poisson de paramètre $\lambda t > 0$.
On a alors $\forall u \geq 0$:

$$\begin{array}{rcl} \mathbb{E}[\ e^{-uX}\ ] & = & \sum_{n\in\mathbb{N}} e^{-un}\mathbb{P}(X=n) \\ & = & \sum_{n\in\mathbb{N}} e^{-un} e^{-\lambda t}\frac{(\lambda t)^n}{n!} \\ & = & e^{-\lambda t}\sum_{n\in\mathbb{N}}\frac{(\lambda t e^{-u})^n}{n!} \\ & = & e^{-\lambda t} e^{\lambda t e^{-u}} \\ & = & e^{\lambda t(e^{-u}-1)} \end{array}$$

Soit $N(t)$ vérifiant la définition 2. Calculons sa transformée de Laplace :
fixons $u \geq 0$ et définissons $g(t) = \mathbb{E}[\ e^{-uN(t)}\ ]$.

- $\forall h > 0$ on calcule :

$$\begin{array}{rcll} g(t+h) & = & \mathbb{E}[\ e^{-uN(t+h)}\ ] & \\ & = & \mathbb{E}[\ e^{-uN(t)} e^{-u(N(t+h)-N(t))}\ ] & \\ & = & \mathbb{E}[\ e^{-uN(t)}\ ]\mathbb{E}[\ e^{-u(N(t+h)-N(t))}\ ] & \text{(accroissements indépendants)} \\ & = & g(t).\mathbb{E}[\ e^{-u(N(h)-N(0))}\ ] & \text{(accroissements stationnaires)} \\ & = & g(t).\mathbb{E}[\ e^{-uN(h)}\ ] & \text{(car } N(0)=0) \end{array}$$

En outre, par iii) on a $\mathbb{P}(N(h)=1) = \lambda h + o(h)$ pour $h \to 0$,
et, par iv) on a $\mathbb{P}(N(h)\geq 2) = o(h)$ pour $h \to 0$ d'où :

$$\begin{array}{rcl} \mathbb{P}(N(h)=0) & = & 1 - \mathbb{P}(N(h)\geq 1) \\ & = & 1 - [\mathbb{P}(N(h)=1) + \mathbb{P}(N(h)\geq 2)] \\ & = & 1 - \lambda h + o(h) \end{array}$$

Ainsi on obtient :

$$\begin{array}{rcl}\mathbb{E}[\ e^{-uN(h)}\ ] & = & \sum_{n\geq 0} e^{-un}\mathbb{P}(N(h)=n) \\ & = & \mathbb{P}(N(h)=0)+e^{-u}\mathbb{P}(N(h)=1)+\sum_{n\geq 2} e^{-un}\mathbb{P}(N(h)=n)\end{array}$$

Or, $\forall n \geq 2 \quad \mathbb{P}(N(h)\geq 2) = \sum_{k\geq 2} \underbrace{\mathbb{P}(N(h)=k)}_{\geq 0} \geq \mathbb{P}(N(h)=n)$,

D'où on a :

$$\begin{array}{rcl}\mathbb{E}[\ e^{-uN(h)}\ ] & \leq & 1-\lambda h+o(h)+e^{-u}(\lambda h+o(h))+(\sum_{n\geq 0} e^{-un})\mathbb{P}(N(h)\geq 2) \\ & = & 1-\lambda h+o(h)+e^{-u}(\lambda h+o(h))+(\sum_{n\geq 0} e^{-un})o(h) \\ & = & 1-\lambda h(1-e^{-u})+o(h)\end{array} \tag{1}$$

Ainsi on a $g(t+h) = g(t)[1-\lambda h(1-e^{-u})+o(h)]$

- En procédant de même que dans le cas $h>0$, $]0,t+h]$ et $]t+h,t]$ étant disjoints, on obtient :

$\forall h<0,\ g(t) = g(t+h).\mathbb{E}[\ e^{-uN(-h)}\ ]$ <u>ie</u> $g(t+h) = \frac{g(t)}{\mathbb{E}[\ e^{-uN(-h)}\ ]}$
Or d'après 1, on a :

$$\frac{1}{\mathbb{E}[\ e^{-uN(-h)}\ ]} = \frac{1}{1+(\lambda h(1-e^{-u})+o(h))} = 1-\lambda h(1-e^{-u})+o(h)$$

Ainsi, on obtient $g(t+h) = g(t)[1-\lambda h(1-e^{-u})+o(h)]$ (comme pour le cas $h>0$)

Finalement, $\forall h \in \mathbb{R}^*$, h petit,

$$\frac{g(t+h)-g(t)}{h} = \lambda g(t)(e^{-u}-1)+\frac{1}{h}o(h)$$

donc la limite existe quand $h \to 0$. Ainsi g est dérivable et

$$g'(t) = \lambda g(t)(e^{-u}-1)$$

En outre, $\forall\ t \geq 0,\ g(t) > 0$ d'où $\frac{g'(t)}{g(t)} = \lambda(e^{-u}-1)$
<u>ie</u> en intégrant : $\ln(|g(t)|) = \lambda t(e^{-u}-1)+C$ avec $C \in \mathbb{R}$
or $g(0)=1$ d'où $C=0$.
Ainsi on a :

$$g(t) = e^{\lambda t(e^{-u}-1)}$$

On reconnaît la transformée de Laplace d'une loi de Poisson.
On en déduit que N(t) suit une loi de Poisson, d'où c).

□

1.3 Une construction des processus de Poisson homogènes

Dans cette section, nous allons montrer que de tels processus existent.

<u>Définition</u> 3.

Soit $(T_n)_{n\geq 0}$ une suite croissante de variables aléatoires réelles positives telles que $(T_1, T_2-T_1, \cdots, T_n-T_{n-1}, \cdots)$ soit une suite de variables aléatoires indépendantes et de même loi, suivant une loi exponentielle de paramètre $\lambda > 0$. On lui associe le processus de comptage $\{N(t)\ ;\ t \geq 0\}$, avec $N(t) = \sum_{n\geq 1} \mathbb{1}_{\{T_n \leq t\}}$.

<u>Remarque</u> : On peut vérifier que c'est bien un processus de comptage.

Remarque : On a alors : $N(t) = \begin{cases} 0 & si \quad t < T_1 \\ n & si \quad T_n \leq t < T_{n+1} \end{cases}$

Théorème 2.

Un processus de comptage défini par la définition 3 est un processus de Poisson au sens de la définition 1.

Commençons d'abord par montrer quelques lemmes utiles.

Lemme 1.

$\forall t \geq 0, \ \forall n \geq 1,$

$$\int_{\mathbb{R}^n} \mathbb{1}_{\{0<t_1<\cdots<t_n\leq t\}} \ dt_1 \cdots dt_n = \frac{t^n}{n!}$$

Preuve : Soit $n \geq 1$ et soit $t \geq 0$. Alors :

$$\begin{array}{lll}
\int_{\mathbb{R}^n} \mathbb{1}_{\{0<t_1<\cdots<t_n\leq t\}} \ dt_1 \cdots dt_n & = & \int_{\mathbb{R}^{n-1}} \mathbb{1}_{\{0<t_2<\cdots<t_n\leq t\}} \ [\int_0^{t_2} dt_1] \ dt_2 \cdots dt_n \qquad \text{(Fubini - Tonelli)} \\
& = & \int_{\mathbb{R}^{n-1}} \mathbb{1}_{\{0<t_2<\cdots<t_n\leq t\}} \ t_2 \ dt_2 \cdots dt_n \\
& \vdots & \text{(par récurrence, pour } 2 \leq k \leq n-1) \\
& = & \int_{\mathbb{R}^{n-k+1}} \mathbb{1}_{\{0<t_k<t_{k+1}\cdots<t_n\leq t\}} \frac{1}{(k-1)!} \ (t_k)^{k-1} \ dt_k \cdots dt_n \\
& = & \int_{\mathbb{R}^{n-k}} \mathbb{1}_{\{0<t_{k+1}<\cdots<t_n\leq t\}} \ [\int_0^{t_{k+1}} \frac{(t_k)^{k-1}}{(k-1)!} dt_k] \ dt_{k+1} \cdots dt_n \qquad \text{(Fubini - Tonelli)} \\
& = & \int_{\mathbb{R}^{n-k}} \mathbb{1}_{\{0<t_{k+1}\cdots<t_n\leq t\}} \frac{1}{k!} \ (t_{k+1})^k \ dt_{k+1} \cdots dt_n \\
& \vdots & \text{(par récurrence, pour } k = n-1) \\
& = & \int_{\mathbb{R}} \mathbb{1}_{\{0<t_n\leq t\}} \frac{1}{(n-1)!} \ (t_n)^{n-1} \ dt_n \\
& = & \frac{t^n}{n!}
\end{array}$$

□

Lemme 2.

Soit $\{N(t) \ ; \ t \geq 0\}$ un processus de comptage défini par la définition 3. Alors $N(t)$ suit une loi de Poisson de paramètre λt. En outre, sachant que $N(t) = n$, $\{T_i\}_{1\leq i\leq n}$ est un n-échantillon de loi uniforme sur le segment $[0, t]$, de densité

$$(t_1, \cdots, t_n) \mapsto \frac{n!}{t^n} \mathbb{1}_{\{0<t_1<\cdots<t_n\leq t\}}$$

Preuve : (inspirée de la référence [4])

- Commençons par calculer la loi de $N(t)$.

Tout d'abord, le n-uplet $(T_1, T_2, \cdots, T_n)$ a pour densité (par rapport à la mesure de Lebesgue)

$$(t_1, \cdots, t_n) \mapsto \mathbb{1}_{\{0<t_1<\cdots<t_n\}} \lambda^n e^{-\lambda t_n}$$

En effet, soit f une fonction mesurable, positive. Posons $f(T_1, \cdots, T_n) = g(T_1, T_2 - T_1, \cdots, T_n - T_{n-1})$.

$$\begin{array}{lll}
\mathbb{E}[\ f(T_1, \cdots, T_n) \ ] & = & \mathbb{E}[\ g(t_1, t_2 - t_1, \cdots, t_n - t_{n-1}) \ ] \\
& = & \int_{\mathbb{R}^n} g(t_1, t_2 - t_1, \cdots, t_n - t_{n-1}) \mathbb{1}_{\{0<t_1<\cdots<t_n\}} \lambda e^{-\lambda t_1} \lambda e^{-\lambda(t_2-t_1)} \cdots \lambda e^{-\lambda(t_n-t_{n-1})} \ dt_1 \cdots dt_n \\
& = & \int_{\mathbb{R}^n} f(t_1, \cdots, t_n) \mathbb{1}_{\{0<t_1<\cdots<t_n\}} \lambda^n e^{-\lambda t_n} \ dt_1 \cdots dt_n
\end{array}$$

Ainsi on obtient, $\forall t \in \mathbb{R}_+, \ \forall n \in N$,

$$\begin{array}{lll}
\mathbb{P}(N(t) = n) & = & \mathbb{P}(T_n \leq t < T_{n+1}) \\
& = & \int_{\mathbb{R}^{n+1}} \mathbb{1}_{\{t_n\leq t<t_{n+1}\}} \ d(t_1, \cdots, t_{n+1}) \\
& = & \int_{\mathbb{R}^{n+1}} \mathbb{1}_{\{0<t_1<\cdots<t_n\leq t<t_{n+1}\}} \lambda^{n+1} e^{-\lambda t_{n+1}} \ dt_1 \cdots dt_{n+1} \\
& = & \int_{\mathbb{R}^n} \mathbb{1}_{\{0<t_1<\cdots<t_n\leq t\}} \lambda^{n+1} \ [\int_t^\infty e^{-\lambda t_{n+1}} dt_{n+1}] \ dt_1 \cdots dt_n \qquad \text{(par Fubini - Tonelli)} \\
& = & e^{-\lambda t} \lambda^n \int_{\mathbb{R}^n} \mathbb{1}_{\{0<t_1<\cdots<t_n\leq t\}} \ dt_1 \cdots dt_n \\
& = & e^{-\lambda t} \frac{(\lambda t)^n}{n!} \qquad \text{(par le lemme 1)}
\end{array}$$

On obtient donc la première partie du lemme :

$$\mathbb{P}(N(t) = n) = e^{-\lambda t} \frac{(\lambda t)^n}{n!}$$

- Calculons la loi de $(T_1, \cdots, T_n)$ sachant $N(t) = n$:

$\forall\ \Gamma \subset \mathbb{R}^n$ mesurable, on a, comme $\forall t > 0,\ \forall n \in \mathbb{N},\ \mathbb{P}(N(t) = n) \neq 0$,

$$\begin{aligned}
\mathbb{P}((T_1, \cdots, T_n) \in \Gamma \mid N(t) = n) &= \frac{\mathbb{P}(N(t)=n, (T_1, \cdots, T_n) \in \Gamma)}{\mathbb{P}(N(t)=n)} \\
&= \frac{n!}{e^{-\lambda t}(\lambda t)^n} \int_{\Gamma \times \mathbb{R}} \mathbb{1}_{\{0<t_1<\cdots<t_n\leq t<t_{n+1}\}} \lambda^{n+1} e^{-\lambda t_{n+1}}\ dt_1 \cdots dt_{n+1} \\
&= \frac{n!}{e^{-\lambda t} t^n} \int_{\Gamma} \mathbb{1}_{\{0<t_1<\cdots<t_n\leq t\}} \left[\int_t^{\infty} \lambda e^{-\lambda t_{n+1}} dt_{n+1}\right] dt_1 \cdots dt_n \quad \text{(Fubini - Tonelli)} \\
&= \frac{n!}{t^n} \int_{\Gamma} \mathbb{1}_{\{0<t_1<\cdots<t_n\leq t\}}\ dt_1 \cdots dt_n
\end{aligned}$$

Ainsi $\{(T_1, \cdots, T_n) | N(t) = n\}$ a pour densité $(t_1, \cdots, t_n) \mapsto \frac{n!}{t^n} \mathbb{1}_{\{0<t_1<\cdots<t_n\leq t\}}$ par rapport à la mesure de Lebesgue sur $\mathbb{R}^n$.

- Soient $S_1, \cdots, S_n$ des variables aléatoires indépendantes, de loi uniforme sur le segment $[0, t]$, (*ie* de densité $x \mapsto \frac{1}{t}\mathbb{1}_{\{0\leq x\leq t\}}(x)\ dx$).

Si on note $\{S'_1, \cdots, S'_n\}$ les variables aléatoires obtenues en réordonnant $\{S_1, \cdots, S_n\}$ dans l'ordre croissant, alors le n-uplet $(S'_1, \cdots, S'_n)$ a pour densité : $(s'_1, \cdots, s'_n) \mapsto \frac{n!}{t^n}\ ds'_1 \cdots ds'_n$. On reconnaît la densité de $\{(T_1, \cdots, T_n) | N(t) = n\}$, d'où le lemme 2. □

Lemme 3.

Soit $\{N(t)\ ;\ t \geq 0\}$ un processus de comptage défini par la définition 3.
$\forall\ 0 = r_0 < r_1 < \cdots < r_k = t$ subdivision de $[0, t]$ et $\forall\ (n_j)_{1\leq j\leq k}$ / $\sum_{j=1}^k n_j = n$, on a :

$$\mathbb{P}(N(r_j) - N(r_{j-1}) = n_j\ ;\ j \in \{1, 2, \cdots, k\}) = \prod_{j=1}^{k} \frac{e^{-\lambda(r_j - r_{j-1})}[\lambda(r_j - r_{j-1})]^{n_j}}{n_j!}$$

Preuve : (inspirée de la référence [4])

Soit $0 = r_0 < r_1 < \cdots < r_k = t$ une subdivision quelconque de $[0, t]$. Avec les mêmes notations que dans la preuve du lemme 2, si on note $(N'_b - N'_a)$ le nombre de variables aléatoires S_k dans l'intervalle $]a, b]$ (ce qui est également le nombre de S'_k dans $]a, b]$ car on ne fait que reprendre les mêmes variables aléatoires mais dans un ordre différent), on a :

$\mathbb{P}(N'_{r_j} - N'_{r_{j-1}} = n_j\ ;\ j \in [\![1, k]\!]) = n! \prod_{i=1}^k \frac{1}{n_j!} \left(\frac{r_j - r_{j-1}}{t}\right)^{n_j}$.

En effet, puisque $S_1, \cdots, S_n$ iid $\sim \mathcal{U}([0, t])$, alors $(N'_1, \cdots, N'_k)$ forme un vecteur de loi multinomiale de paramètres :

- ⋆ n (nombre de tirages indépendants) avec $\sum_{i=1}^k n_i = n$;
- ⋆ $p_j = \frac{r_j - r_{j-1}}{t}$ $(1 \leq j \leq k)$ (probabilité de la $j^{\text{ème}}$ issue n_j).

Ainsi, $\{(T_1, \cdots, T_n) | N(t) = n\}$ ayant même loi que $\{S'_1, \cdots, S'_n\}$, on obtient donc que :

$$\mathbb{P}(\underbrace{N(r_j) - N(r_{j-1}) = n_j}_{n_j \text{ tops dans }]r_{j-1}, r_j]}\ ;\ j \in [\![1, k]\!] \mid N(t) = n) = n! \prod_{i=1}^k \frac{1}{n_j!} \left(\frac{r_j - r_{j-1}}{t}\right)^{n_j}$$

(avec $n = \sum_{j=1}^k n_j$, sinon cette probabilité est nulle).

Pour se débarrasser du conditionnement, on multiplie par $\mathbb{P}(N(t) = n)$. Rappelons que $\mathbb{P}(N(t) = n) = e^{-\lambda t}\frac{(\lambda t)^n}{n!}$ (cf. lemme 2).

Ainsi, on a :

$$\begin{aligned}
\mathbb{P}(N(r_j) - N(r_{j-1}) = n_j\ ;\ j \in [\![1, k]\!]) &= \mathbb{P}(N(t) = \textstyle\sum_{j=1}^k n_j).\mathbb{P}(N(r_j) - N(r_{j-1}) = n_j\ ;\ j \in [\![1, k]\!] \mid N(t) = n) \\
&= e^{-\lambda t} \frac{(\lambda t)^{(\sum_{j=1}^k n_j)}}{\not{n}!} \cdot \not{n}! \prod_{j=1}^k \frac{1}{n_j!} \left(\frac{r_j - r_{j-1}}{t}\right)^{n_j} \\
&= e^{-\lambda[\sum_{j=1}^k (r_j - r_{j-1})]} \prod_{j=1}^k \frac{(\lambda t)^{n_j}}{n_j!} \frac{(r_j - r_{j-1})^{n_j}}{\not{t}^{n_j}} \\
&= \prod_{j=1}^k \frac{e^{-\lambda(r_j - r_{j-1})}[\lambda(r_j - r_{j-1})]^{n_j}}{n_j!}
\end{aligned}$$

et ce $\forall\ 0 = r_0 < r_1 < \cdots < r_k = t$ subdivision de $[0, t]$, d'où le lemme 3. □

Démontrons à présent le théorème 2.

Démonstration : Soit $\{N(t)\ ;\ t \geq 0\}$ un processus de comptage défini par la définition 3.

a) $N(0) = 0$ car $T_1 > 0$ presque sûrement.

Montrons que $\{N(t)\ ;\ t \geq 0\}$ est à accroissements indépendants, et que le nombre de tops dans l'intervalle $[s, s+t]$, $N(s+t) - N(s)$ suit une loi de Poisson de paramètre λt, *ie*

$$\mathbb{P}(N(s+t) - N(s) = n) = e^{-\lambda t}\frac{(\lambda t)^n}{n!}$$

D'après le lemme 3, on a : $\forall\ 0 = r_0 < r_1 < \cdots < r_k = t$ subdivision de $[0,t]$ et $\forall\ (n_j)_{1\leq j\leq k}\ /\ \sum_{j=1}^k n_j = n$,

$$\mathbb{P}(N(r_j) - N(r_{j-1}) = n_j\ ;\ j \in \{1,2,\cdots,k\}) = \prod_{j=1}^{k} \frac{e^{-\lambda(r_j - r_{j-1})}[\lambda(r_j - r_{j-1})]^{n_j}}{n_j!}$$

Ainsi, $\forall\ s,t > 0$, (pour $r_0 = 0, r_1 = s, r_2 = s+t$), on a $\mathbb{P}(N(s) = k, N(s+t) - N(s) = n) = e^{-\lambda s}\frac{(\lambda s)^k}{k!} \cdot e^{-\lambda t}\frac{(\lambda t)^n}{n!}$
et en sommant sur les k, on obtient :

$$\begin{array}{rcl} \mathbb{P}(N(s+t) - N(s) = n) & = & \sum_{k\in\mathbb{N}} \mathbb{P}(N(s) = k, N(s+t) - N(s) = n) \\ & = & \sum_{k\in\mathbb{N}} [e^{-\lambda s}\frac{(\lambda s)^k}{k!}] \cdot e^{-\lambda t}\frac{(\lambda t)^n}{n!} \\ & = & e^{-\lambda t}\frac{(\lambda t)^n}{n!} \end{array}$$

Finalement, $\forall s,t > 0,\ N(s+t) - N(s)$ suit une loi de Poisson, d'où c).
Et de plus, on obtient que $\mathbb{P}(\bigcap_{1\leq j\leq k}\{N(r_j) - N(r_{j-1}) = n_j\}) = \prod_{j=1}^k \mathbb{P}(N(r_j) - N(r_{j-1}) = n_j)$.
On en déduit donc que le processus est bien à accroissements indépendants, d'où b).
Cela conclut donc la démonstration du théorème 2.

□

Théorème 3.
Soit $\{N(t)\ ;\ t \geq 0\}$ défini par la définition 1. Notons $(X_n)_{n\geq 1}$ les sauts (X_n représente l'instant du $n^{ième}$ top). Alors la suite (X_n) ainsi définie vérifie la définition 3.

Remarque : La démonstration de ce théorème est à peine évoquée dans [4]. Ce qui suit constitue notre travail théorique principal.

Démonstration :
Reprenons les mêmes notations que dans le théorème 3, et considérons $(T_n)_{n\geq 1}$ vérifiant la définition 3.
On peut alors définir un processus $\{N'(t)\ ;\ t \geq 0\}$ au sens de la définition 3.
Montrons que $\forall n \geq 1, (X_k)_{1\leq k\leq n}$ et $(T_k)_{1\leq k\leq n}$ ont même loi.

- 1ère étape : Montrons que $\forall n \geq 1, \forall\ \Gamma \subset \mathbb{R}^n$ mesurable, on a :

$$\mathbb{P}((X_1,\cdots,X_n) \in \Gamma \mid N(t) = n) = \mathbb{P}((T_1,\cdots,T_n) \in \Gamma \mid N'(t) = n)$$

Par le lemme 2, on sait que $N'(t)$ suit une loi de Poisson, et que $\forall\ \Gamma \subset \mathbb{R}^n$ mesurable,

$$\mathbb{P}((T_1,\cdots,T_n) \in \Gamma \mid N'(t) = n) = \frac{n!}{t^n}\int_\Gamma \mathbb{1}_{\{0<t_1<\cdots<t_n\leq t\}}\ dt_1\cdots dt_n$$

▷ 1ère sous-étape : **Cas particulier des pavés rangés.**

Soit $0 = r_0 < r_1 < \cdots < r_k = t$ une subdivision de l'intervalle $[0,t]$, soit $(n_j)_{1\leq j\leq k}\ /\ \sum_{j=1}^k n_j = n$.
Considérons $\Gamma = \bigotimes_{j=1}^k]r_{j-1}, r_j]^{n_j}$.
Montrons que $\mathbb{P}((X_1,\cdots,X_n) \in \Gamma \mid N(t) = n) = \frac{n!}{t^n}\int_\Gamma \mathbb{1}_{\{0<t_1<\cdots<t_n\leq t\}}\ dt_1\cdots dt_n$.

Posons $P = \mathbb{P}((X_1, \cdots, X_n) \in \Gamma \mid N(t) = n)$.

On a : $\{N(r_j) - N(r_{j-1}) = n_j\} = \{X_{(\sum_{i=1}^{j-1} n_i)+m} \in]r_{j-1}, r_j] \ , \ m \in [\![1, n_j]\!]\}$
$= \{(X_{(\sum_{i=1}^{j-1} n_i)+m})_{1 \leq m \leq n_j} \in]r_{j-1}, r_j]^{n_j}\}$

D'où on obtient :

$P = \mathbb{P}(\{(X_{(\sum_{i=1}^{j-1} n_i)+m})_{1 \leq m \leq n_j} \in]r_{j-1}, r_j]^{n_j} \ , \ j \in [\![1, k]\!]\} \mid N(t) = n)$
$= \mathbb{P}(\{N(r_j) - N(r_{j-1}) = n_j \ , \ j \in [\![1, k]\!]\} \mid N(t) = n)$

Ainsi :

$$P = \frac{\mathbb{P}(\{N(r_j) - N(r_{j-1}) = n_j \ , \ j \in [\![1, k]\!]\} \ , \ N(t) = n)}{\mathbb{P}(N(t) = n)} = \frac{\mathbb{P}(\{N(r_j) - N(r_{j-1}) = n_j \ , \ j \in [\![1, k]\!]\})}{\mathbb{P}(N(t) = n)}$$

Or, d'après la définition 1, on sait que $\forall t \geq 0, \ N(t)$ suit une loi de Poisson de paramètre λt, et que le processus est à accroissements indépendants. Les intervalles $]r_{j-1}, r_j]$ étant disjoints, on obtient :

$$\mathbb{P}(\{N(r_j) - N(r_{j-1}) = n_j \ , \ j \in [\![1, k]\!]\}) = \prod_{j=1}^{k} \frac{e^{-\lambda(r_j - r_{j-1})} \ [\lambda(r_j - r_{j-1})]^{n_j}}{n_j!} = e^{-\lambda t} \frac{(\lambda t)^n}{n!} \ n! \prod_{j=1}^{k} \frac{1}{n_j!} \left(\frac{r_j - r_{j-1}}{t}\right)^{n_j}$$

D'où $P = \frac{n!}{t^n} \ \prod_{j=1}^{k} \frac{(r_j - r_{j-1})^{n_j}}{n_j!}$.

En outre, d'après le lemme 1, $\forall j \in [\![1, k]\!], \frac{(r_j - r_{j-1})^{n_j}}{n_j!} = \int_{\mathbb{R}^{n_j}} \mathbb{1}_{\{0 < t_{j_1} < \cdots < t_{j_{n_j}} \leq r_j - r_{j-1}\}} \ dt_{j_1} \cdots dt_{j_{n_j}}$ d'où

$P = \frac{n!}{t^n} \ \prod_{j=1}^{k} [\int_{\mathbb{R}^{n_j}} \mathbb{1}_{\{0 < t_{j_1} < \cdots < t_{j_{n_j}} \leq r_j - r_{j-1}\}} \ dt_{j_1} \cdots dt_{j_{n_j}}]$
$= \frac{n!}{t^n} \ \int_{\mathbb{R}^n} \prod_{j=1}^{k} \mathbb{1}_{\{0 < t_{j_1} < \cdots < t_{j_{n_j}} \leq r_j - r_{j-1}\}} \ dt_{1_1} \cdots dt_{1_{n_1}} dt_{2_1} \cdots dt_{2_{n_2}} \cdots dt_{k_1} \cdots dt_{k_1} \cdots dtk_{n_k}$

$\forall j \in [\![1, k]\!], \forall i \in [\![1, n_j]\!]$, posons $x_{j_i} = t_{j_i} + r_{j-1}$.

La mesure de Lebesgue étant invariante par translation, on obtient :

$P = \frac{n!}{t^n} \ \int_{\mathbb{R}^n} \prod_{j=1}^{k} \mathbb{1}_{\{r_{j-1} < x_{j_1} < \cdots < x_{j_{n_j}} \leq r_j\}} \ dx_{1_1} \cdots dx_{k_{n_k}}$
$= \frac{n!}{t^n} \ \int_{\mathbb{R}^n} \mathbb{1}_{\{0 < x_{1_1} < \cdots x_{1_{n_1}} \leq r_1 < x_{2_1} < \cdots < x_{2_{n_2}} \leq r_2 < \cdots < r_{k-1} < x_{k_1} < \cdots < x_{k_{n_k}} \leq r_k\}} \ dx_{1_1} \cdots dx_{k_{n_k}}$
$= \frac{n!}{t^n} \ \int_{\bigotimes_{j=1}^{k}]r_{j-1}, r_j]^{n_j}} \mathbb{1}_{\{0 < x_{1_1} < \cdots x_{1_{n_1}} < x_{2_1} < \cdots < x_{k_{n_k}} \leq t\}} \ dx_{1_1} \cdots dx_{k_{n_k}}$

<u>*ie*</u>

$$\mathbb{P}((X_1, \cdots, X_n) \in \Gamma \mid N(t) = n) = \frac{n!}{t^n} \int_{\Gamma} \mathbb{1}_{\{0 < x_1 < \cdots < x_n \leq t\}} \ dx_1 \cdots dx_n \tag{2}$$

et ce pour tout Γ de la forme $\bigotimes_{j=1}^{k}]r_{j-1}, r_j]^{n_j}$, avec $(r_j)_{1 \leq j \leq k}$ subdivision de $[0, t]$.

▷ <u>2ème sous-étape</u> : **Cas des pavés quelconques.**

Montrons que pour tout pavé $\Gamma = \bigotimes_{i=1}^{n}]\alpha_i, \beta_i]$, on a :

$$\mathbb{P}((X_1, \cdots, X_n) \in \Gamma \mid N(t) = n) = \frac{n!}{t^n} \int_{\Gamma} \mathbb{1}_{\{0 < t_1 < \cdots < t_n \leq t\}} \ dt_1 \cdots dt_n$$

Considérons $(\alpha_i, \beta_i)_{1 \leq i \leq n}$ des réels tels que $\forall i \in [\![1, n]\!], \alpha_i < \beta_i$.
On a bien sûr, $\Gamma = \bigotimes_{i=1}^{n}]\alpha_i, \beta_i]$ mesurable dans $\mathbb{R}^n$ muni de la mesure de Lebesgue.
Calculons $\int_{\Gamma} \mathbb{1}_{\{0 < t_1 < \cdots < t_n \leq t\}} \ dt_1 \cdots dt_n$ et $\mathbb{P}((X_1, \cdots, X_n) \in \Gamma \mid N(t) = n)$.
Considérons à présent $(r_j)_{1 \leq j \leq k}$ le réordonnement dans l'ordre croissant et sans répétition des α_j, β_j.
Alors, $\forall i \in [\![1, n]\!], \exists \ m_i \geq 0, \ \exists \ j_i \in [\![1, k]\!] \ / \ ]\alpha_i, \beta_i] =]r_{j_i}, r_{j_i + m_i}] = \bigcup_{l_i=0}^{m_i - 1}]r_{j_i + l_i}, r_{j_i + l_i + 1}]$.
Ainsi, on a :

$\Gamma = \bigotimes_{i=1}^{n}]\alpha_i, \beta_i]$
$= \bigotimes_{i=1}^{n} \bigcup_{l_i=0}^{m_i - 1}]r_{j_i + l_i}, r_{j_i + l_i + 1}]$
$= \bigcup_{l_1=0}^{m_1 - 1} \cdots \bigcup_{l_n=0}^{m_n - 1} \{\bigotimes_{i=1}^{n}]r_{j_i + l_i}, r_{j_i + l_i + 1}]\}$ (on développe)

Les $\bigotimes_{i=1}^{n}]r_{j_i + l_i}, r_{j_i + l_i + 1}]$ sont tous disjoints car le produit d'une partition est une partition du produit.

Posons maintenant $\Delta_0 = \{(p_1 \cdots, p_n) \ / \ \forall i \in [\![1, n]\!], j_i \leq p_i < j_i + m_i\}$. Ainsi, pour $p_i = j_i + l_i$, on obtient $\Gamma = \bigcup_{p_1=j_1}^{j_1+m_1-1} \cdots \bigcup_{p_n=j_n}^{j_n+m_n-1} \{\bigotimes_{i=1}^n]r_{p_i}, r_{p_i+1}]\}$.

$$\Gamma = \bigcup_{(p_1,\cdots,p_n)\in\Delta_0} \bigotimes_{i=1}^n]r_{p_i}, r_{p_i+1}]$$

Ainsi : $\left\{ \begin{array}{rcl} \int_\Gamma \mathbb{1}_{\{0<t_1<\cdots<t_n\leq t\}}\ dt_1 \cdots dt_n & = & \sum_{(p_1,\cdots,p_n)\in\Delta_0} \int_{\bigotimes_{i=1}^n]r_{p_i}, r_{p_i+1}]} \mathbb{1}_{\{0<t_1<\cdots<t_n\leq t\}}\ dt_1 \cdots dt_n \\ \mathbb{P}((X_1, \cdots, X_n) \in \Gamma \mid N(t) = n) & = & \sum_{(p_1,\cdots,p_n)\in\Delta_0} \mathbb{P}((X_1, \cdots, X_n) \in \bigotimes_{i=1}^n]r_{p_i}, r_{p_i+1}] \mid N(t) = n) \end{array} \right.$

En outre, pour tous $1 \leq i_0 < j_0 \leq n$ tels que $p_{i_0} > p_{j_0}$, (*ie* $p_{i_0} \geq p_{j_0+1}$ et $r_{p_{i_0}} \geq r_{p_{j_0+1}}$),
quelque soit $(p_1, \cdots, p_{i_0}, \cdots, p_{j_0}, \cdots, p_n) \in \Delta_0$ fixé, on a :

- $\forall (t_1, \cdots, t_n) \in \bigotimes_{i=1}^n]r_{p_i}, r_{p_i+1}],\ t_{j_0} \leq r_{p_{j_0+1}} \leq r_{p_{i_0}} < t_{i_0}$ d'où $\mathbb{1}_{\{0<t_1<\cdots<t_n\leq t\}}(t_1, \cdots, t_n) = 0$.
  Et donc $\int_{\bigotimes_{i=1}^n]r_{p_i}, r_{p_i+1}]} \mathbb{1}_{\{0<t_1<\cdots<t_n\leq t\}} dt_1 \cdots dt_n = 0$.
- De même, par croissance des X_i, on a :
  $\mathbb{P}((X_1, \cdots, X_n) \in \bigotimes_{i=1}^n]r_{p_i}, r_{p_i+1}] \mid N(t) = n) = 0$.

Ce qui nous amène donc à poser $\Delta = \{(p_1, \cdots, p_n) \in \Delta_0 \ / \ p_1 \leq \cdots \leq p_n\}$, et $D = \bigcup_{(p_1,\cdots,p_n)\in\Delta} \bigotimes_{i=1}^n]r_{p_i}, r_{p_i+1}]$.
On vient de montrer que :
$\left\{ \begin{array}{rcl} \int_\Gamma \mathbb{1}_{\{0<t_1<\cdots<t_n\leq t\}} dt_1 \cdots dt_n & = & \int_D \mathbb{1}_{\{0<t_1<\cdots<t_n\leq t\}} dt_1 \cdots dt_n \\ \mathbb{P}((X_1, \cdots, X_n) \in \Gamma \mid N(t) = n) & = & \mathbb{P}((X_1, \cdots, X_n) \in D \mid N(t) = n) \end{array} \right.$

Or $\forall\ (p_1, \cdots, p_n) \in \Delta$, $\bigotimes_{i=1}^n]r_{p_i}, r_{p_i+1}]$ est de la forme $\bigotimes_{j=1}^k]r_{j-1}, r_j]^{n_j}$ avec $0 = r_0 < r_1 < \cdots < r_k = t$, les n_j pouvant être nuls, et $\sum_{j=1}^k n_j = n$ par construction donc, d'après l'équation 2 de la première sous-étape, on a :

$$\mathbb{P}((X_1, \cdots, X_n) \in \bigotimes_{i=1}^n]r_{p_i}, r_{p_i+1}] \mid N(t) = n) = \frac{n!}{t^n} \int_{\bigotimes_{i=1}^n]r_{p_i}, r_{p_i+1}]} \mathbb{1}_{\{0<t_1<\cdots<t_n\leq t\}}\ dt_1 \cdots dt_n$$

Finalement, on obtient :

$$\begin{array}{rcl} \mathbb{P}((X_1, \cdots, X_n) \in \Gamma \mid N(t) = n) & = & \mathbb{P}((X_1, \cdots, X_n) \in D \mid N(t) = n) \\ & = & \sum_{(p_1,\cdots,p_n)\in\Delta} \mathbb{P}((X_1, \cdots, X_n) \in \bigotimes_{i=1}^n]r_{p_i}, r_{p_i+1}] \mid N(t) = n) \\ & = & \sum_{(p_1,\cdots,p_n)\in\Delta} \frac{n!}{t^n} \int_{\bigotimes_{i=1}^n]r_{p_i}, r_{p_i+1}]} \mathbb{1}_{\{0<t_1<\cdots<t_n\leq t\}}\ dt_1 \cdots dt_n \\ & = & \frac{n!}{t^n} \int_D \mathbb{1}_{\{0<t_1<\cdots<t_n\leq t\}} dt_1 \cdots dt_n \\ & = & \frac{n!}{t^n} \int_\Gamma \mathbb{1}_{\{0<t_1<\cdots<t_n\leq t\}} dt_1 \cdots dt_n \\ & = & \mathbb{P}((T_1, \cdots, T_n) \in \Gamma \mid N'(t) = n) \end{array}$$

et ce quelque soit $\Gamma \subset \mathbb{R}^n$ pavé mesurable.

▷ 3ème sous-étape : **Cas des boréliens.**

L'ensemble des pavés forme une famille σ-finie qui engendre les boréliens de $\mathbb{R}^n$.
D'où $\forall\ \Gamma$ borélien, on a $\mathbb{P}((X_1, \cdots, X_n) \in \Gamma \mid N(t) = n) = \mathbb{P}((T_1, \cdots, T_n) \in \Gamma \mid N'(t) = n)$.

- 2ème étape : **Déconditionnement**

Ainsi, $\forall n \geq 1,\ \forall\ \Gamma \subset \mathbb{R}^n$, on a :

$$\begin{array}{rcl} \mathbb{P}((X_1, \cdots, X_n) \in \Gamma) & = & \mathbb{P}((X_1, \cdots, X_n) \in \Gamma,\ N(t) = n) \\ & = & \mathbb{P}((X_1, \cdots, X_n) \in \Gamma \mid N(t) = n) \times \mathbb{P}(N(t) = n) \\ & = & \mathbb{P}((T_1, \cdots, T_n) \in \Gamma \mid N'(t) = n) \times \mathbb{P}(N'(t) = n) \\ & = & \mathbb{P}((T_1, \cdots, T_n) \in \Gamma,\ N'(t) = n) \\ & = & \mathbb{P}((T_1, \cdots, T_n) \in \Gamma) \end{array}$$

On en déduit que $\forall n \in \mathbb{N}, (X_1 \cdots, X_n)$ et $(T_1, \cdots, T_n)$ ont même loi, *ie* $\{N(t); t \geq 0\}$ vérifie les propriétés de la définition 3.
Ceci termine donc la démonstration du théorème 3. □

2 Application à un exemple concret

2.1 Position du problème

Nous venons donc de voir ce qu'était précisément un processus de Poisson homogène, avec différentes définitions équivalentes. Nous allons maintenant essayer de les appliquer dans le domaine concret, en nous servant plus particulièrement de la définition 3. Le domaine d'étude se situe dans le cadre de la biologie, avec des applications à des données génomiques.

Escherichia Coli (ou "colibacille") est une bactérie intestinale des mammifères, pouvant entraîner des infections urinaires, des gastro-entérites, des méningites... L'atout de cette bactérie est qu'il s'agit de l'un des organismes vivants les plus étudiés de nos jours. Son patrimoine génétique a été entièrement séquencé en 1997. Son génome comprend 4,6 millions de paires de bases codant environ 4200 protéines. Les génomes de Escherichia Coli diffèrent selon leur type et cela dans des proportions très importantes (40% seulement de gènes en communs pour trois catégories différentes de Escherichia Coli).

Les données en notre possession, fournies par l'INRA (Institut National de la Recherche Agronomique), sont de deux sortes. Nous savons d'un côté à quelles positions sur la chaîne d'ADN se situe un des types de sites promoteurs du gène (présence du mot "TATAAT", où A et T représentent les bases azotées Adénine et Thymine), et nous connaissons également où commencent les gènes.

Notre but est de savoir si oui ou non, ces deux phénomènes peuvent être associés à des processus de Poisson homogènes. Pour cela nous allons tester, à partir des tests de Kolmogorov - Smirnov (que nous allons définir par la suite), si les répartitions des sites promoteurs (respectivement des débuts de gènes) suivent des lois uniformes, ou si encore les écarts entre chaque début d'expression (top) peuvent être assimilés à une distribution exponentielle.

2.2 Test de Kolmogorov - Smirnov

Ici, nous voulons tester si la répartition des séquences promotrices et des gènes peut être modélisée par un processus de Poisson homogène. Nous allons donc voir si ces points peuvent être vus comme des variables aléatoires uniformes indépendantes et de même loi, rangées dans l'ordre croissant (cf. lemme 2), ou si les écarts entre les gènes ou les TATAAT répondent à la définition 3, à savoir qu'ils sont indépendants, et suivent tous une loi exponentielle de paramètre λ.

Pour cela, nous appliquerons le test de Kolmogorov - Smirnov (cf. [5]) qui a pour but de déterminer si un échantillon suit bien une loi donnée connue par sa fonction de répartition continue (qui rappelons-le, caractérise la loi d'une variable aléatoire), ou bien si deux échantillons suivent la même loi.

Ce test repose sur la quantité définie pour F et G deux fontions de répartition : $D(F,G) = \sup_{t\in\mathbb{R}} |G(t)-F(t)|$.

Commençons par énoncer une propriété remarquable du cours de master 1 de statistiques (voir [5] pour preuve) :

Propriété 1.

Soit $(X_1, X_2, \cdots, X_n)$ un n-échantillon de variables aléatoires indépendantes et de même loi, de fonction de répartition F, et considérons la fonction de répartition empirique $F_n(t) = \frac{1}{n}\sum_{k=1}^{n} \mathbb{1}_{\{X_i\leq t\}}$ (estimateur sans biais et fortement consistant de $F(t)$).

Si F est continue, alors $D(F,F_n) = \sup_{t\in\mathbb{R}} |F_n(t) - F(t)|$ ne dépend pas de F.

Ainsi, on peut donc construire une table de Kolmogorov - Smirnov représentant les quantiles d'ordre α, notés par la suite $\xi_{n,\alpha}$, en fonction de n et de α. $\forall$ F fonction de répartition continue d'un n-échantillon,

$$\mathbb{P}(\sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \leq \xi_{n,\alpha}) = \mathbb{P}(\sup_{t\in\mathbb{R}} |F_n^U(t) - F^U(t)| \leq \xi_{n,\alpha}) = \alpha$$

où F^U et F_n^U sont respectivement les fonctions de répartition et de répartition empirique associées à la loi uniforme sur $[0,1]$.

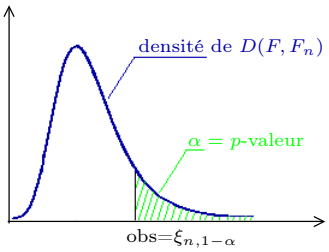
Principe du Test.

On teste l'hypothèse nulle : H_0 : "$F = F_0$" contre l'alternative : $H_1 =$ "$F \neq F_0$" au niveau α.
On accepte H_0 si on a :

$$\forall t \in \mathbb{R}, |F_n(t) - F_0(t)| \leq \xi_{n,1-\alpha}$$

où $\xi_{n,1-\alpha}$ est le quantile d'ordre $1-\alpha$ extrait de la table de Kolmogorov - Smirnov.

En effet, sous H_0, on a :
$\mathbb{P}(\forall t \in \mathbb{R}, |F_n(t) - F_0(t)| \geq \xi_{n,1-\alpha}) = \mathbb{P}(\sup_{t\in\mathbb{R}} |F_n(t) - F_0(t)| \geq \xi_{n,1-\alpha}) = \mathbb{P}(D(F_n, F) \geq \xi_{n,1-\alpha}) = 1 - (1-\alpha) = \alpha$ donc c'est bien un test de niveau α.



La p-valeur (voir figure ci-contre) est la valeur de α (le niveau du test) pour laquelle l'observation $\sup_{t\in\mathbb{R}} |F_n(t) - F_0(t)|$ est à la limite entre les zones d'acceptation et de rejet de l'hypothèse nulle. Elle représente la vraisemblance d'observer ceci sous H_0, ainsi pour une petite p-valeur (< 10 %) l'hypothèse nulle n'est pas vraisemblable, alors que pour une grande p-valeur (> 20 %), elle semble généralement convaincante en l'absence d'autres observations.

2.3 R et résultats

Pour effectuer les tests, nous avons utilisé le logiciel de statistiques R, en nous appuyant sur les références [6] et [7].

2.3.1 Test des TATAAT

D'après le lemme 2, on a vu que sachant le nombre total de tops d'un processus de Poisson à l'instant t, ceux-ci sont uniformément répartis sur le segment $[0, t]$. On va donc tester ici si les $\mathtt{T_i}$ suivent une loi uniforme sur le segment $[0, 9288442]$, où 9288442 représente la longueur de la chaîne d'ADN étudiée. En effet, on note une forte ressemblance entre le graphe de la fonction de répartition empirique des positions des TATAAT, et celui de la fonction de répartition d'un n-échantillon de variables aléatoires uniformes sur $[0, 9288442]$ (ici $n = 1036$) :

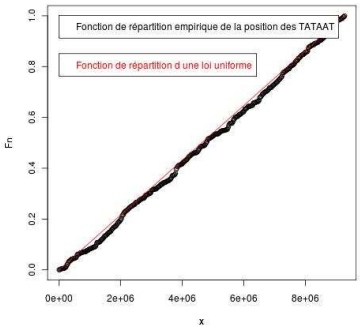


FIG. 2 – Position des TATAAT

Premièrement on récupère les données dans un tableau que l'on appelle `tataat`, puis on transforme la colonne qui nous intéresse en un vecteur `T` :

```
> tataat=read.table("Bureau/poisson/donnees/ecoliK12-tataat-d10000.pos")
> T=tataat[[1]]
```

Ensuite, on teste si les $\mathtt{T_i}$ suivent bien une loi uniforme sur $[0, 9288442]$ par le test de Kolmogorov - Smirnov :

```
> ks.test(T,punif,0,9288442)

        One-sample Kolmogorov-Smirnov test
```

```
data : T
D = 0.0358, p-value = 0.1410
alternative hypothesis : two-sided
```

La p-valeur étant comprise entre 10% et 20%, on est dans la zone de flou. Ce test ne nous permet pas de conclure.

Nous allons donc regarder les écarts entre les TATAAT, et tester s'ils sont exponentiels (en se ramenant à la définition 3). En effet, lorsque l'on trace les graphes de la fonction de répartition empirique des écarts entre les TATAAT, et de la fonction de répartition d'une loi exponentielle (de paramètre $\frac{1035}{9288442}$, cf. plus bas), on remarque une forte ressemblance, d'où notre désir de modéliser la position des TATAAT par un processus de Poisson.

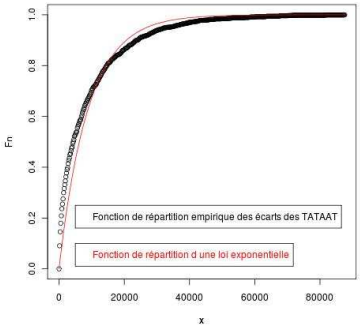


FIG. 3 – Ecarts TATAAT

Calculons à présent les différences des valeurs consécutives dans le nouveau vecteur `t` représentant donc les écarts :

```
> t=diff(T)
```

On veut tester si les $\mathtt{t_i} = \mathtt{T_i} - \mathtt{T_{i-1}}$ suivent bien une loi exponentielle (cf. définition 3). Ne connaissant pas le paramètre, mais sachant que la moyenne d'une loi exponentielle de paramètre λ vaut $\frac{1}{\lambda}$, on va approximer la moyenne par :

$$\frac{1}{\frac{1}{n}\sum_{i=1}^{n} \mathtt{t_i}} = \frac{n}{\sum_{i=1}^{n} \mathtt{T_{i+1}} - \mathtt{T_i}} = \frac{n}{T_{n+1}} \approx \frac{n}{9288442} =: \mathtt{lt}$$

Sachant que la longueur de l'ADN étudié est de 9288442, on peut supposer que $\mathtt{T_{n+1}} \approx 9288442$.

```
> n=length(t)
[1] 1035
> lt=1035/9288442
```

Finalement, on teste si les $\mathtt{t_i}$ suivent bien une loi exponentielle de paramètre `lt` par le test de Kolmogorov - Smirnov :

```
> ks.test(t,pexp,lt)

     One-sample Kolmogorov-Smirnov test

data : t
D = 0.1598, p-value < 2.2e-16
alternative hypothesis : two-sided

Warning message :
In ks.test(t, pexp, lt) :
impossible de calculer les p-values correctes avec des ex-aequos
```

Ainsi on obtient une p-valeur tout à fait négligeable ($< 2,2 \times 10^{-16}$), et on peut rejeter l'hypothèse nulle ; les écarts entre les TATAAT ne suivent pas une loi exponentielle. On peut expliquer ce résultat par le fait que les TATAAT sont auto-recouvrants (self overlapping), c'est-à-dire qu'on peut trouver deux fois le mot TATAAT à la suite. On en déduit donc que la répartition des TATAAT ne peut être modélisée par un processus de Poisson homogène.

Remarque : On a un message d'erreur qui nous dit que les p-valeurs ne sont pas exactes car on a des doublons (par exemple, `t` contient 3 fois 31), ce qui ne devrait pas être le cas puisque la fonction de répartition d'une loi exponentielle de paramètre `lt` ($F_l(t) = 1 - e^{-lt}\mathbb{1}_{\{t>0\}}$) est continue. Toutefois, le test de Kolomogorov - Smirnov est bien basé sur la différence entre les deux courbes de la figure 3. Le fait d'avoir des données entières et donc qui ne peuvent être qu'approximativement exponentielles n'explique pas une p-valeur aussi petite. L'approximation elle-même par une loi exponentielle est donc a priori fausse.

```
> which(t==31)
[1] 375 399 651
```

Remarque : Le test de Kolmogorov - Smirnov nécessite d'avoir vraiment F_0, or ici nous avons estimé le paramètre. Si la calibration est théoriquement nécéssaire, les quantiles étant légèrement faussés, la petitesse de la p-valeur nous conforte dans le fait que la loi exponentielle est fausse. En effet, en pratique, créons un n-échantillon `x` de loi exponentielle de paramètre `lt`, et estimons le paramètre par l'inverse de sa moyenne `m`. On teste alors si les $\mathtt{x}_\mathtt{i}$ suivent bien une loi exponentielle de paramètre `m` :

```
> x=rexp(1035,lt)
> m=1/mean(x)
> ks.test(x,pexp,m)

     One-sample Kolmogorov-Smirnov test

data : x
D = 0.019, p-value = 0.85
alternative hypothesis : two-sided
```

On obtient ainsi une très grande p-valeur. De manière générale, cette procédure a juste tendance à accepter H_0 plus souvent, ce qui justifie ici notre estimation du paramètre.

2.3.2 Test des gènes

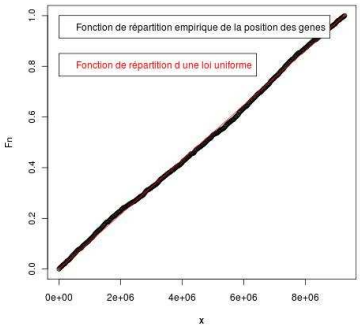


FIG. 4 – Position des gènes

De même que précédemment, on remarque une nette ressemblance entre les graphes de la fonction de répartition empirique de la position des gènes et de la fonction de répartition d'une loi uniforme sur le segment $[0, 9288442]$. On effectue donc exactement les mêmes tests que pour les TATAAT :

```
> genes=read.table("Bureau/poisson/donnees/ecoliK12-genes-d10000.pos")
> G=genes[[1]]
```

```
> ks.test(G,punif,0,9288442)

     One-sample Kolmogorov-Smirnov test

data : G
D = 0.0213, p-value = 0.04052
alternative hypothesis : two-sided
```

Contrairement au test sur les TATAAT, on obtient une p-valeur qui n'est pas dans la zone de flou ($< 10\%$) ce qui nous permet de rejeter l'hypothèse nulle. Ceci est dû au fait que nous avons beaucoup plus de données pour les gènes.

Nous pouvons aussi regarder les écarts entre les gènes.

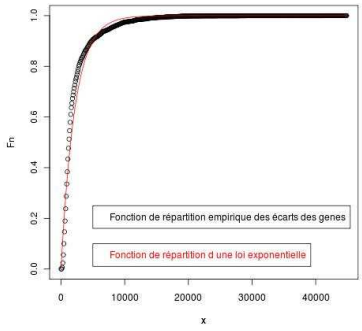


FIG. 5 – Ecarts gènes

De même que précédemment, on remarque une nette ressemblance (peut-être même encore plus forte) entre les graphes respectifs de la fonction de répartition empirique des écarts entre les gènes et de la fonction de répartition d'une loi exponentielle de paramètre $\frac{4289}{9288442}$. On effectue donc le même test que pour les TATAAT :

```
> genes=read.table("Bureau/poisson/donnees/ecoliK12-genes-d10000.pos")
> G=genes[[1]]
> g=diff(G)
> length(g)
[1] 4289
> lg=4289/9288442
> ks.test(g,pexp,lg)

     One-sample Kolmogorov-Smirnov test

data : g
D = 0.1123, p-value < 2.2e-16
alternative hypothesis : two-sided

Warning message :
In ks.test(g, pexp, lg) :
impossible de calculer les p-values correctes avec des ex-aequos
```

De même que lors du test des TATAAT, on obtient une p-valeur tout à fait négligeable. On rejette donc l'hypothèse selon laquelle les gènes sont répartis selon un processus de Poisson homogène, confirmant ainsi le résultat du test sur l'uniformité des positions.

En conclusion, on a pu voir grâce au test de Kolmogorov - Smirnov que ni les gènes, ni les séquences promotrices TATAAT de la bactérie Escherichia Coli ne sont répartis selon un processus de Poisson homogène. Il a été montré par d'autres tests qu'ils peuvent être modélisés par un processus plus complexe liant l'apparition des TATAAT et des gènes (cf. [8]).

Références

[1] Sheldon M.ROSS, *Introduction to probability models*, Academic Press, 8$^{\text{th}}$ edition, 2003.

[2] Sheldon M.ROSS, *Initiation aux probabilités*, Presse Polythechniques et universitaires romandes, 7$^{\text{ème}}$ édition, 2007.

[3] Dominique FOATA, Aimé FUCHS, *Processus stochastiques*, DUNOD, 2002.

[4] Jacques NEVEU, *Cours de probabilités*, Ecole Polythechnique, 1974.

[5] Vincent RIVOIRARD, *Préparation à l'agrégation - Notes de cours*, 2010.

[6] Magalie FROMONT, *Introduction à R*, ENSAI, 2008.

[7] Site Web du Cran, `http ://cran.r-project.org/`.

[8] Gaëlle GUSTO and Sophie SCHBATH, *FADO : a statistical method to detect favored or avoided distances between occurrences of motifs using the Hawkes' model*, INRA, 2005.