CORRECTION D.M. 6
VARIABLES ALEATOIRES A DENSITES

Remarque préliminaire : Ce sujet est largement inspiré du sujet Agro-Véto 2007 et, lorsque c'est possible, j'indiquerai les remarques faites par le jury sur cette épreuve.

Quelques remarques d'ordre général ont été faites sur les copies :

En probabilités :
Certains candidats n'ont pas compris ce qu'est une fonction de répartition : à la question I.1.b. ils écrivent que la fonction de répartition est nulle à l'extérieur de l'intervalle $[0,1]$ *ou se refusent à la calculer à l'extérieur de* $[0,1]$.
Quelques candidats pensent que deux variables aléatoires ayant la même loi sont égales.
La caractérisation des fonctions de répartition des variables à densité est trop souvent ignorée : certains candidats rappellent les propriétés générales des fonctions de répartition (croissante, $\mathcal{C}^1$ *par morceaux, limites en* $+\infty$ *et* $-\infty$*) mais ne répondent pas à la question posée.*
Quelques candidats confondent encore indépendance et incompatibilité.
On trouve dans quelques copies des densités négatives, des fonctions de répartition négatives.

En analyse :
La notion de fonction de classe $\mathcal{C}^1$ *sur* $\mathbb{R}$ *n'est pas toujours bien assimilée.*
La convergence des intégrales impropres est rarement traitée ; lorsqu'elle est traitée les théorèmes de convergence sont cités de façon incomplète ou fausse. Les intégrations par parties sont utilisées en oubliant de dire que les fonctions sont de classe $\mathcal{C}^1$*. Par ailleurs, l'écriture mathématique fait parfois défaut : des relations sont données sans précision de domaine, les éléments dx manquent aux intégrales...*
L'orthographe est déplorable dans quelques copies, négligée dans beaucoup.
A l'unanimité, les correcteurs déplorent la malhonnêteté dont font preuve certains candidats dans les calculs un peu délicats dont l'énoncé fournit la réponse (essentiellement la question II.2.d) : des erreurs disparaissent en fin de calcul ; des invocations magiques du type « par télescopage des termes » ou « grâce à la formule de binôme de Newton » remplacent des calculs non faits ou faux.

Première partie

On modélise le site de la cassure numéro i, $1 \leq i \leq n$ par une variable aléatoire $X_i \hookrightarrow \mathcal{U}_{]0,1[}$. Les variables aléatoires X_i sont supposées mutuellement indépendantes.

1. On s'intéresse tout d'abord aux sites de la première et de la dernière cassure ramenées à l'intervalle $]0,1[$, respectivement $Y_1 = min\{X_1, ..., X_n\}$ et $Y_n = max\{X_1, ..., X_n\}$.

   a) Une densité de $X_i \hookrightarrow \mathcal{U}_{]0,1[}$ est $f : x \longmapsto \mathbf{1}_{]0,1[}(x)$.
   Sa fonction de répartition est $F_X : x \longmapsto \begin{cases} 0 & \text{si } x \leq 0 \\ x & \text{si } 0 < x < 1 \\ 1 & \text{si } x \geq 1 \end{cases}$.
   $\mathbb{E}(X)$ et $\mathbb{E}(X^2)$ existent car les intégrales $\displaystyle\int_{-\infty}^{\infty} tf(t)dt$ et $\displaystyle\int_{-\infty}^{\infty} t^2 f(t)dt$ se réduisent à deux intégrales définies grâce à la relation de Chasles.
   En effet $\mathbb{E}(X) = \displaystyle\int_0^1 tdt = \frac{1}{2}$ et $\mathbb{E}(X^2) = \displaystyle\int_0^1 t^2 dt = \frac{1}{3}$.
   Soit, d'après la formule de Koënig-Huygens $\mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}^2(X) = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$.

b) *Écrivons une fonction Python* `simulCassure(p)` *permettant de retourner une liste* `Lc` *de* $n = 2*p$ *cassures prises au hasard dans* $]0,1[$ *selon le modèle des cassures aléatoires associé à des lois uniformes sur l'intervalle* $]0,1[$ *:* Il suffit pour ça de faire $2*p$ fois appel (boucle « Pour ») à la fonction `random()` de la bibliothèque `random` qui retourne un réel pris au hasard selon la loi uniforme dans l'intervalle $[0,1[$.

```
def simulcassure(p):
    C = [rdm.random() for k in range(2*p)]
    return C
```

c) *Montrons que pour tout* x *dans* $]0,1[$, $\mathbb{P}(Y_n \leq x) = x^n$:

$\forall x \in]0,1[$, $\mathbb{P}(Y_n \leq x) = \mathbb{P}[(X_1 \leq x) \cap ... \cap (X_n \leq x)] = \prod_{i=1}^{n} \mathbb{P}(X_i \leq x)$ car variables indépendantes.

Conclusion : $\boxed{\forall x \in]0,1[,\ \mathbb{P}(Y_n \leq x) = \prod_{i=1}^{n} F_{X_i}(x) = x^n}$

d) On commence par noter que $X_i(\Omega) =]0,1[$, $\forall 1 \leq i \leq n$ donc $\boxed{Y_n(\Omega) =]0,1[}$.
Il est donc immédiat que $F_{Y_n}(x) = 0$ si $x \leq 0$ et $F_{Y_n}(x) = 1$ si $x \geq 1$. En conséquence :

Conclusion : $\boxed{F_{Y_n}(x) = \begin{cases} 0 & \text{si } x \leq 0 \\ x^n & \text{si } 0 < x < 1 \\ 1 & \text{si } x \geq 1 \end{cases}}$

On vérifie que F_{Y_n} est **continue** sur $\mathbb{R}$, de **classe** $\mathcal{C}^1$ sur $\mathbb{R} \setminus \{1\}$ (puisqu'en 1 la dérivée à gauche vaut n et la dérivée à droite vaut 0) **croissante sur** $\mathbb{R}$, de **limite nulle en** $-\infty$ et de **limite** 1 **en** $+\infty$**.**

Conclusion : $\boxed{Y_n \text{ est une variable aléatoire à densité}}$

Lu dans le rapport de jury : « *En plus des erreurs déjà mentionnées sur la fonction de répartition, des candidats affirment que* F_n *est de classe* $\mathcal{C}^1$ *sur* $\mathbb{R}$*.* »

En dérivant F_{Y_n} sur $\mathbb{R} \setminus \{1\}$ et en posant $f_n(1) = 0$ on obtient une densité f_n de Y_n.
Conclusion : $\boxed{f_n \text{ définie sur } \mathbb{R} \text{ par } f_n(x) = nx^{n-1}\mathbb{I}_{]0,1[}(x) \text{ est une densité de } Y_n}$

e) *Montrons que pour tout* x *dans* $]0,1[$, $\mathbb{P}(Y_1 > x) = (1-x)^n$:
Commençons par rappeler que $X_i(\Omega) =]0,1[$, $\forall 1 \leq i \leq n$ donc $\boxed{Y_1(\Omega) =]0,1[}$.
On a dès lors $F_{Y_1}(x) = 0$ si $x \leq 0$ et $F_{Y_1}(x) = 1$ si $x \geq 1$.

$\forall x \in]0,1[$, $F_{Y_1}(x) = \mathbb{P}(Y_1 \leq x) = 1 - \mathbb{P}(Y_1 > x) = 1 - \mathbb{P}[(X_1 > x) \cap ... \cap (X_n > x)] = \prod_{i=1}^{n} \mathbb{P}(X_i > x)$

car variables indépendantes.
Or $\forall x \in]0,1[$, $\mathbb{P}(X_i > x) = 1 - \mathbb{P}(X_i \leq x) = 1 - F_{X_i}(x) = 1 - x$.

Conclusion : $\boxed{F_{Y_1}(x) = \begin{cases} 0 & \text{si } x \leq 0 \\ 1-(1-x)^n & \text{si } 0 < x < 1 \\ 1 & \text{si } x \geq 1 \end{cases}}$

$\forall x \in]0,1[$, $F'_{Y_1}(x) = n(1-x)^{n-1}$ et $\forall x \in \mathbb{R} \backslash]0,1[$, $F'_{Y_1}(x) = 0$. F est donc dérivable sur $\mathbb{R} \setminus \{0\}$ puisque la dérivée à gauche en 0 vaut 0 et sa dérivée à droite vaut n...

Il est immédiat de constater que F_{Y_1} est continue sur $\mathbb{R}$, de classe $\mathcal{C}^1$ sur $\mathbb{R}^*$, croissante sur $\mathbb{R}$, de limite nulle en $-\infty$ et de limite égale à 1 en $+\infty$. D'où $\boxed{Y_1 \text{ est une variable aléatoire à densité}}$.
Par dérivation de F_{Y_1} sur $\mathbb{R} \setminus \{0,1\}$, on obtient une densité f_1 de Y_1.

Conclusion : $\boxed{f_1 \text{ définie sur } \mathbb{R} \text{ par } f_1(x) = n(1-x)^{n-1}\mathbb{I}_{]0,1[}(x) \text{ est une densité de } Y_1}$

2. a) *Montrons que Y_n et $1-Y_1$ ont même loi* :
— $Y_n(\Omega) =]0,1[= 1 - Y_1(\Omega)$
— Montrons que Y_n et $1-Y_1$ ont même fonction de répartition :
$\forall x \in]0,1[$, $F_{1-Y_1}(x) = \mathbb{P}(1-Y_1 \leq x) = \mathbb{P}(Y_1 \geq 1-x) = \mathbb{P}(Y_1 > 1-x)$
Or $1-x \in]0,1[$ donc, d'après e) :$F_{1-Y_1}(x) = [1-(1-x)]^n = x^n = F_{Y_n}(x)$.

Comme ces deux fonctions valent respectivement 0 si $x < 0$ et 1 si $x > 1$, on vient de montrer que $\forall x \in \mathbb{R}$, $F_{1-Y_1}(x) = F_{Y_n}(x)$. Donc $1-Y_1$ et Y_n ont même fonction de répartition.

Une fonction de répartition caractérisant la loi d'une variable aléatoire à densité, on a :
Conclusion : $\boxed{Y_n \text{ et } 1-Y_1 \text{ ont même loi}}$.

Concernant l'espérance et la variance de Y_1 et Y_n, on commence par dire qu'elles existent car ce sont deux variables aléatoires de support fini (autre façon de dire que $Y_1(\Omega) = Y_n(\Omega) = [0,1]$ sont bornés et qu'en conséquence les intégrales permettant de calculer le moment d'ordre 1 et le moment centré d'ordre 2 sont des intégrales définies...)
Les lois étant les mêmes, on a :

$$\boxed{\begin{cases} \mathbb{E}(Y_n) &= 1 - \mathbb{E}(Y_1) \text{ par linéarité de l'espérance} \\ \mathbb{V}(Y_n) &= (-1)^2\mathbb{V}(Y_1) = V(Y_1) \end{cases}}$$

Lu dans le rapport de jury : « *Le début de la question n'est pas traité dans un nombre non négligeable de copies. Quelques candidats donnent des relations fausses, en particulier $\mathbb{V}(Y_n) = -\mathbb{V}(Y_1)$ ce qui donne une variance négative à la question 1.2.c) !!* ».

b) $\mathbb{E}(Y_n) = \displaystyle\int_0^1 nx^n dx = n\left[\frac{x^{n+1}}{n+1}\right]_0^1 = \frac{n}{n+1}$ et donc, d'après I.2.a), $\mathbb{E}(Y_1) = 1 - \dfrac{n}{n+1} = \dfrac{1}{n+1}$

c) D'après la formule de Koenig-Huygens, $\mathbb{V}(Y_n) = \mathbb{E}(Y_n^2) - \mathbb{E}^2(Y_n)$ avec
$\mathbb{E}(Y_n^2) = \displaystyle\int_0^1 nx^{n+1} dx = n\left[\frac{x^{n+2}}{n+2}\right]_0^1 = \frac{n}{n+2}$.
Donc $\mathbb{V}(Y_n) = \dfrac{n}{n+2} - \dfrac{n^2}{(n+1)^2} = \dfrac{n(n^2+2n+1) - n^2(n+2)}{(n+1)^2(n+2)}$.

Conclusion : $\boxed{\mathbb{V}(Y_n) = \dfrac{n}{(n+1)^2(n+2)} \text{ et donc } \mathbb{V}(Y_1) = \mathbb{V}(Y_n)}$

Lu dans le rapport de jury : « *Quelques candidats se lancent dans de longues justifications d'absolue convergence pour l'existence de l'espérance et de la variance ; d'autres font une intégration par parties pour les calculer.* »

d) Pour écrire une fonction `loiY1(m,p)`, on réalise grâce à une boucle « Pour » la création de `m` listes de $n = 2 * p$ cassures en appelant à chaque itération la fonction `simulCassure()`. A chaque

étape, on détermine le minimum de cette liste qu'on place dans une liste `LY1` répertoriant toutes les valeurs prises par $Y_1 = \min\{X_1, \cdots, X_n\}$.
Les calculs de la moyenne et de la variance se font avec les formules usuelles, à savoir :

$$\overline{y} = \frac{1}{m}\sum_{k=1}^{m} y_k \text{ et } s_y^2 = \frac{1}{n}\sum_{k=1}^{m} y_k^2 - \overline{y}^2$$

```
def loiY1b(m,p):
    LY1 = [min(simulCassure(p)) for k in range(m)]
    LY1_carre = [LY1[k]**2 for k in range(m)]
    moy = sum(LY1)/m
    var = sum(LY1_carre)/m - moy**2
    return LY1,moy,var
```

3. Pour tout $x \in [0,1]$ on pose U_i variable aléatoire discrète de Bernoulli de paramètre $p = \mathbb{P}(X_i \leq x) = x$, c'est-à-dire $U_i \hookrightarrow \mathcal{B}(1,x)$ et $U = \sum_{i=1}^{n} U_i$.
On rappelle que $U_i(\Omega) = \{0,1\}$ avec $\mathbb{P}(U_i = 1) = p = x$ et $\mathbb{P}(U_i = 0) = 1 - x$.

   a) *Loi de U :*
   — $U_i(\Omega) = \{0,1\}$ donc $U(\Omega) = [\![0,n]\!]$
   — Les variables U_i sont indépendantes car les variables aléatoires X_i le sont. En effet, $\forall i \neq j$, $(i,j) \in [\![1,n]\!]^2$:

$$\begin{aligned}\mathbb{P}[(U_i = 1) \cap (U_j = 1)] &= \mathbb{P}[(X_i \leq x) \cap (X_j \leq x)] = \mathbb{P}(X_i \leq x) \cdot \mathbb{P}(X_j \leq x) \\ &= \mathbb{P}(U_i = 1) \cdot \mathbb{P}(U_j = 1)\end{aligned}$$

   U est une somme de variables aléatoires de bernoulli indépendantes et de même paramètre x et désigne le nombre cassures ayant lieu avant le site $s = x$.

   Conclusion : $\boxed{U \hookrightarrow \mathcal{B}(n,x)}$

   Lu dans le rapport de jury : « *La justification de l'indépendance des variables U_i est trop souvent omise* »

   b) L'événement $(Y_k \leq x)$ est réalisé lorsque la k-ième cassure s'est présenté avant le site $s = x$ ce qui équivaut à dire que le nombre total de cassure qui on eu lieu avant $s = x$ est au moins égale à k ou encore que l'événement $(U \geq k)$ est réalisé.

   Conclusion : $\boxed{(Y_k \leq x) = (U \geq k)}$

   Dès lors : $\mathbb{P}(Y_k \leq x) = \mathbb{P}(U \geq x) = \sum_{j=k}^{n} \mathbb{P}(U = j) = \sum_{j=k}^{n} \binom{n}{j} x^j (1-x)^{n-j}$

   Conclusion : $\boxed{\forall x \in [0,1],\ \mathbb{P}(Y_k \leq x) = \sum_{j=k}^{n} \binom{n}{j} x^j (1-x)^{n-j}}$

4. On note F_k la fonction de répartition de Y_k et f_k sa densité continue sur $\mathbb{R}$ pout tout $k \in [\![2, n-1]\!]$.

   a) Soit $x \in [0,1]$ et $k \in [\![2, n-1]\!]$. On a $F_{n-k+1}(x) = \mathbb{P}(Y_{n-k+1} \leq x)$ donc d'après la question

précédente :

$$\begin{aligned}
F_{n-k+1}(1-x) &= \sum_{j=n-k+1}^{n} \binom{n}{j} (1-x)^j x^{n-j} = \sum_{i=0}^{k-1} \binom{n}{n-i} (1-x)^{n-i} x^i \text{ en posant } i = n-j \\
&= \sum_{i=0}^{k-1} \binom{n}{i} (1-x)^{n-i} x^i \text{ par propriété des coefficients binomiaux} \\
&= \sum_{i=0}^{n} \binom{n}{i} (1-x)^{n-i} x^i - \sum_{i=k}^{n} \binom{n}{i} (1-x)^{n-i} x^i \\
&= (1-x+x)^n - F_k(x) = 1 - F_k(x)
\end{aligned}$$

Conclusion : $\boxed{F_k(x) = 1 - F_{n-k+1}(1-x),\ \forall x \in [0,1]}$

Par dérivation, on obtient $\boxed{f_k(x) = f_{n-k+1}(1-x),\ \forall x \in [0,1]}$

Comme nous savons que Y_k est une **variable aléatoire à support fini** puisque par hypothèse $Y_k(\Omega) = [0,1]$, cela assure l'existence de $\mathbb{E}(Y_k)$ et $\mathbb{V}(Y_k)$.
$\mathbb{E}(Y_k) = \displaystyle\int_0^1 x f_k(x) dx = -\int_1^0 (1-y) f_k(1-y) dy$
En utilisant le changement de variable $y = 1 - x = \psi(x)$ qui est autorisé car la fonction ψ est de classe $\mathcal{C}^1$ sur $[0,1]$, on a :

$$\mathbb{E}(Y_k) = \int_0^1 (1-y) f_{n-k+1}(y) dy = 1 - \mathbb{E}(Y_{n-k+1})$$

Conclusion : $\boxed{\forall k \in [\![2, n-1]\!],\ \mathbb{E}(Y_{n-k+1}) = 1 - \mathbb{E}(Y_k)}$

Par ailleurs :

$$\mathbb{E}\left[(1-Y_k)^2\right] = \int_0^1 (1-x)^2 f_k(x) dx = -\int_1^0 y^2 f_k(1-y) dy$$

par le même changement admissible que précédemment. Soit :

$$\mathbb{E}\left[(1-Y_k)^2\right] = \int_0^1 y^2 f_{n-k+1}(y) dy = \mathbb{E}(Y_{n-k+1}^2)$$

Enfin, d'après la formule de Koënig-Huygens :

$$\begin{aligned}
\mathbb{V}(Y_{n-k+1}) &= \mathbb{E}(Y_{n-k+1}^2) - \mathbb{E}^2(Y_{n-k+1}) \\
&= 1 - 2\mathbb{E}(Y_k) + \mathbb{E}(Y_k^2) - (1 - \mathbb{E}(Y_k))^2 \\
&= 1 - 2\mathbb{E}(Y_k) + \mathbb{E}(Y_k^2) - (1 - 2\mathbb{E}(Y_k) + \mathbb{E}^2(Y_k)) = \mathbb{E}(Y_k^2) - \mathbb{E}^2(Y_k)
\end{aligned}$$

Conclusion : $\boxed{\mathbb{V}(Y_{n-k+1}) = \mathbb{V}(Y_k)}$

Lu dans le rapport de jury : « *La première partie de la question est rarement abordée. Les relations entre les espérances et les variances de Y_k et Y_{n-k+1} sont rarement justifiées. Dans certaines copies, les candidats posent $y = 1 - x$ sans plus de précision.* »

Remarque : En s'inspirant de la question I.2.a), on pouvait aussi dire :

$$\begin{aligned}
F_k(x) = 1 - F_{n-k+1}(1-x) &\Leftrightarrow F_k(x) = 1 - \mathbb{P}(Y_{n-k+1} \leq 1-x) \\
&\Leftrightarrow F_k(x) = 1 - \mathbb{P}(1 - Y_{n-k+1} \geq x) = \mathbb{P}(1 - Y_{n-k+1} < x) \\
&\Leftrightarrow F_k(x) = F_{(1-Y_{n-k+1})}(x)
\end{aligned}$$

D'où on déduit que Y_k et $1 - Y_{n-k+1}$ ont même loi.
On obtient alors directement :
$\boxed{\mathbb{E}(Y_k) = 1 - \mathbb{E}(Y_{n-k+1})}$ et $\boxed{\mathbb{V}(Y_k) = \mathbb{V}(Y_{n-k+1})}$

b) $\forall k \in [\![1, n]\!]$, on a par définition $\mathbb{E}(Y_k) = \displaystyle\int_0^1 x f_k(x) dx$.
Pour une intégration par parties, les fonctions u et v considérées étant de classe $\mathcal{C}^1$ sur l'intervalle $[0, 1]$, on pose :

$$\begin{cases} u = x & u' = 1 \\ v' = f_k(x) & v = F_k(x) - 1 \end{cases}$$

Ainsi :

$$\mathbb{E}(Y_k) = \big[x(F_k(x) - 1)\big]_0^1 - \int_0^1 (F_k(x) - 1) dx$$

Or, $F_k(1) = 1$ car $Y_k(\Omega) = [0, 1]$ et F_k continue sur $\mathbb{R}$.

Conclusion : $\boxed{\mathbb{E}(Y_k) = \displaystyle\int_0^1 (1 - F_k(x)) dx}$

Dès lors, $\forall k \in [\![1, n-1]\!]$:

$$\begin{aligned} \mathbb{E}(Y_{k+1} - Y_k) &= \int_0^1 \left[F_k(x) - F_{k+1}(x)\right] dx \text{ par linéarité de l'intégrale} \\ &= \int_0^1 \Big[\sum_{j=k}^{n} \binom{n}{j} x^j (1-x)^{n-j} - \sum_{j=k+1}^{n} \binom{n}{j} x^j (1-x)^{n-j}\Big] dx \text{ d'après I.3.b)} \\ &= \int_0^1 \binom{n}{k} x^k (1-x)^{n-k} dx = \binom{n}{k} \int_0^1 x^k (1-x)^{n-k} dx \end{aligned}$$

Conclusion : $\boxed{\mathbb{E}(Y_{k+1} - Y_k)^{=} \binom{n}{k} \displaystyle\int_0^1 x^k (1-x)^{n-k} dx, \forall k \in [\![1, n-1]\!]}$

Lu dans le rapport de jury : « *La première partie de la question est rarement traitée.* »

c) On introduit la suite $I_k(n) = \binom{n}{k} \displaystyle\int_0^1 x^k (1-x)^{n-k} dx$ pour $k \in [\![0, N]\!]$.
Montrons que $I_k(n) = I_{k+1}(n)$, $\forall k \in [\![0, n-1]\!]$:
$\forall k \in [\![0, n-1]\!]$, intégrons par partie l'intégrale $\displaystyle\int_0^1 x^k (1-x)^{n-k} dx$ qui définit $I_k(n)$. Pour cela, posons :

$$\begin{cases} u(x) = (1-x)^{n-k} & u'(x) = -(n-k)(1-x)^{n-k-1} \\ v'(x) = x^k & v(x) = \dfrac{x^{k+1}}{k+1} \end{cases}$$

où u et v sont deux fonctions de classe $\mathcal{C}^1$ sur $[0, 1]$. Alors :

$$\begin{aligned} I_k(n) &= \binom{n}{k}\left(\left[\frac{1}{k+1}(1-x)^{n-k}x^{k+1}\right]_0^1 + \frac{n-k}{k+1}\int_0^1 x^{k+1}(1-x)^{n-k-1}dx\right) \\ &= \binom{n}{k}\frac{n-k}{k+1}\int_0^1 x^{k+1}(1-x)^{n-k-1}dx \text{ car } 0^{n-k}=0,\ \forall k\in[\![0,n-1]\!] \\ &= \frac{n!}{k!(n-k)!}\frac{(n-k)}{k+1}\int_0^1 x^{k+1}(1-x)^{n-k-1}dx \\ &= \frac{n!}{(k+1)!(n-k-1)!}\int_0^1 x^{k+1}(1-x)^{n-k-1}dx = I_{k+1}(n) \end{aligned}$$

Conclusion : $\boxed{I_k(n) = I_{k+1}(n),\ \forall k\in[\![0,n-1]\!]}$

On en déduit que la suite $(I_k(n))_{k\in[\![0,n]\!]}$ est constante égale à $I_0 = \displaystyle\int_0^1 (1-x)^n dx$ avec :

$$\int_0^1 (1-x)^n dx\left[-\frac{(1-x)^{n+1}}{n+1}\right] = \frac{1}{n+1}$$

Conclusion : $\boxed{\forall k\in[\![0,n]\!],\ I_k(n) = \dfrac{1}{n+1}}$

d) $\forall k\in[\![1,n-1]\!],\ \mathbb{E}(Y_{k+1}-Y_k) = I_k(n) = \dfrac{1}{n+1}$.

D'où, par linéarité de l'espérance : $\forall k\in[\![1,n-1]\!],\ \mathbb{E}(Y_{k+1}) = \mathbb{E}(Y_k) + \dfrac{1}{n+1}$.

Autrement dit : $\big(\mathbb{E}(Y_k)\big)_{k\in[\![1,n]\!]}$ est une suite arithmétique de premier terme $\mathbb{E}(Y_1) = \dfrac{1}{n+1}$ et de raison $\dfrac{1}{n+1}$.

Conclusion : $\boxed{\mathbb{E}(Y_k) = \mathbb{E}(Y_1) + (k-1)\dfrac{1}{n+1} = \dfrac{1+k-1}{n+1} = \dfrac{k}{n+1}}$

Interprétation : Le modèle des cassures aléatoires conduite à penser que les cassures étant choisies au hasard selon une loi uniforme sur le génome, elles se trouveront en moyenne uniformément réparties, les n cassures découpant $n+1$ intervalles de longueur moyenne $\dfrac{N}{n+1}$.

Approche algorithmique.

Les travaux sur l'évolution du génome ont montré que les longueurs des segments obtenus par cassures présentent une grande diversité (de quelques Kb à plusieurs dizaines de Mb). Le modèle des cassures aléatoires dont on vient de tirer les conséquences (répartition uniforme des points de cassures et intervalles de même longueur moyenne) n'est-il pas en contradiction avec ces observations et en particulier rend-il compte de cette forte variance ?

1. *Montrons que, si n est suffisamment grand, il est possible d'approcher la loi de Y_1 par une loi exponentielle de paramètre n :*

   On note que si n est grand alors $\mathbb{E}(Y_1) = \dfrac{1}{n+1}$ est proche de 0. On peut donc considérer que pour de grandes valeurs de n, $Y_1(\Omega) = [0,a]$ avec a proche de 0. Dès lors :

   $$\forall x\in[0,a],\ f_1(x) = n(1-x)^{n-1} = ne^{(n-1)\ln(1-x)} \text{ avec } n-1 \underset{n\to\infty}{\sim} n \text{ et } \ln(1-x) \underset{x\to 0}{\sim} -x$$

soit :

$$(n-1)\ln(1-x) \underset{n\to\infty,x\to 0}{\sim} -nx \text{ et donc } f_1(x) \underset{n\to\infty,x\to 0}{\sim} ne^{-nx}$$

Remarque : En toute rigueur il faudrait démontrer, pour la composition de l'équivalence par l'exponentielle, que $(n-1)ln(1-x)+nx$ tend vers 0 quand n tend vers ∞ et x tend vers 0. C'est effectivement le cas puisqu'ici x est proche de $1/n$.

On rappelle que si $Z \hookrightarrow \exp(\lambda)$ alors $f_Z(x) = \lambda e^{-\lambda x}\mathbb{I}_{\mathbb{R}_+}(x)$.

Conclusion : Pour $n \in \mathbb{N}$ suffisamment grand, on peut considérer que $Y_1 \hookrightarrow \exp(n)$

2. *Pourquoi est-il légitime de penser que la loi des longueurs de segments $Y_{k+1}-Y_k$ peut également être approchée par une loi exponentielle de paramètre n ?*
On a vu à la question 4.c) que l'espérance de $Y_{k+1}-Y_k$ est égale à l'espérance de Y_1, autrement dit $\mathbb{E}(Y_{k+1}-Y_k) = \mathbb{E}(Y_1) = \dfrac{1}{n+1} \underset{n\to\infty}{\sim} \dfrac{1}{n}$ qui est l'espérance de $Z \hookrightarrow \exp(n)$.
La variable aléatoire égale à la longueur des segments a donc, quand n est grand, même espérance qu'une loi exponentielle de paramètre n.
Par ailleurs, Y_k étant pris au hasard sur l'intervalle $]0,1[$, on peut considérer que l'événement $(Y_k = a)$ est réalisé avec $0 < a < 1$ et

$$Y_{k+1} - Y_k = Y_{k+1} - a = min\{X_i - a, i \in [\![1,n]\!]/X_i - a \in]0,1[\}$$

Le comportement de $Y_{k+1}-Y_k$ est donc similaire à celui de Y_1 puisque, lorsque X suit une loi uniforme $X-a$ suit également une loi uniforme.

Ces éléments de réponse nous permettent de poser l'hypothèse (qu'il s'agira de valider) que $Y_{k+1}-Y_k$ peut être approchée par une loi exponentielle de paramètre n.

Conclusion : Hypothèse : $Y_{k+1} - Y_k \hookrightarrow \exp(n)$

3. *Montrons que si $Z \hookrightarrow \exp(\lambda)$ et a est un réel fixé ($a \in \mathbb{R}^*$), alors $T = aZ$ suit encore une loi exponentielle :*

— $T(\Omega) = aZ(\Omega) = \mathbb{R}_+$ puisque $a > 0$.
— On a immédiatement que : si $x < 0$, $F_T(x) = \mathbb{P}(T \leq x) = 0$.
— Si $x \geq 0$, alors $F_T(x) = \mathbb{P}(T \leq x) = \mathbb{P}(aZ \leq x) = \mathbb{P}(Z \leq x/a)$
Soit $F_T(x) = F_Z(x/a) = 1 - e^{-\lambda x/a}$

En dérivant on obtient que $f_T(x) = F_T'(x) = \dfrac{\lambda}{a} e^{-\lambda x/a}\mathbb{I}_{\mathbb{R}_+}(x)$.

Conclusion : Si $Z \hookrightarrow \exp(\lambda)$ alors $T = aZ \hookrightarrow \exp(\lambda/a)$.

On se place à nouveau sur le génome dont les cassures C_i sont prises indépendamment et uniformément sur $[\![1,N]\!]$ avec $N = 3.4e9$.
$(D_1, \cdots, D_n)$ désigne les n variables aléatoires ayant pour valeurs les valeurs des variables $(C_1, \cdots, C_n)$ ordonnées dans l'ordre croissant, ou encore les valeurs des variables $(NX_1, \cdots, NX_n)$ ordonnées dans l'ordre croissant.
Dès lors $D_{k+1} - D_k = NY_{k+1} - NY_k = N \cdot (Y_{k+1} - Y_k)$ avec $Y_{k+1} - Y_k \hookrightarrow \exp(n)$ par hypothèse.

Conclusion : Hypothèse : $D_{k+1} - D_k$ suit une loi exponentielle de paramètre $\lambda = n/N$.

✐ ***Conséquence :*** Si cette hypothèse est vérifiée, on doit avoir par propriété de la loi exponentielle $\mathbb{E}(D_{k+1} - D_k) = \dfrac{1}{\lambda} = \dfrac{N}{n}$ (ce qui, notons-le, est conforme aux résultats issus de la question 4. pour n grand) et $\mathbb{V}(D_{k+1} - D_k) = \dfrac{1}{\lambda^2} = \dfrac{N^2}{n^2}$.
Sachant que $N = 3.4e9$ et n est de l'ordre du millier, on mesure la dispersion des longueurs possibles des blocs de synthénie !

4. *Écrivons une fonction Python* `simulCassureGenome(N,p)` *permettant de retourner une liste* `Lc` *de* $n = 2 * p$ *cassures* $(C_1, \cdots, C_n)$ *uniformément distribué dans un génome de longueur fictive* N *:*
Il suffit pour ça de faire appel à la fonction `randint(a,b)` de la bibliothèque `random` qui retourne un entier pris au hasard selon la loi uniforme dans l'intervalle $[\![a, b]\!]$.
La fonction est dès lors très rapide :

```
def simulcassureGenome(N,p):
   Lc = [rdm.randint(1,N) for k in range(2*p)]
   return Lc
```

5. L'écriture d'une fonction de tri permettant de trier la liste `Lc` pour obtenir une liste `Ld` des mêmes valeurs organisées de façon croissante est une question de cours.
Nous choisissons pour notre part la méthode de tri rapide qui met en place une récursivité :

```
def trirapide(L):
    if L==[]:
        return []
    else:
        n=len(L)
        L1=[]
        L2=[]
        for k in range(1,n):
            if L[k]<=L[0]:
                L1.append(L[k]) #L1 reçoit les éléments les plus petits
            else:
                L2.append(L[k]) # L2 reçoit les éléments les plus grands
        L=trirapide(L1)+[L[0]]+trirapide(L2)
    return L
```

Il est dès lors facile d'en déduire une fonction `longueurSegments(N,p)` qui retourne la liste de toutes les longueurs rencontrées après p inversions puisqu'il suffit de retourner les distances entre chaque valeurs successives de la liste triée :

```
def longueurSegments(N,p):
   Lc = simulcassureGenome(N,p)
   Ld = trirapide(Lc)
   n1 = len(Ld)
   L = [C1[0]] # intialisation de la liste des longueurs
   for k in range(n1-1):
      L.append(C1[k+1]-C1[k])
   L.append(N-C1[n1-1]) # Pour tenir compte du dernier segment...
   return L
```

6. Après avoir répété 100 fois l'opération ci-dessus avec $N = 25000$ et $p = 320$, on obtient des segments de longueur comprise entre 0 et 150. Grâce à la fonction `histogram()` de la bibliothèque `numpy`, on obtient la liste `H` des fréquences relatives des longueurs de segments obtenus.

   On trouve ci-dessous l'histogramme de ces fréquences (`plt.bar(np.arange(150),H)`) :

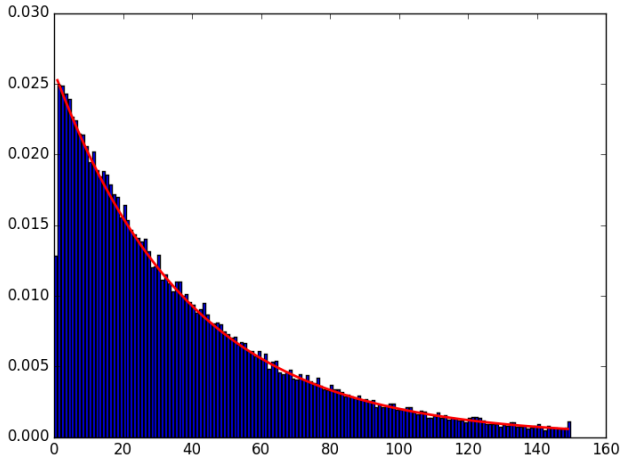

FIGURE 1 – `Histogramme des longueurs de segments séparés par` $n = 2p = 640$ `cassures.`

Nous souhaitons valider que la distribution de ces longueurs suit effectivement une loi exponentielle. Posons Z variable aléatoire à densité égale à la longueur des segments obtenus par cassures aléatoires. Si Z suit une loi exponentielle de paramètre λ, alors :

$$\forall n \in \mathbb{N}^*,\ H(n) \approx \mathbb{P}(n \leq Z < n+1) = F_Z(n+1) - F_Z(n)$$

Or, d'après le théorème des accroissements finis appliqué à F_Z qui est de classe $\mathcal{C}^1$ sur $\mathbb{R}_+$:

$$\exists c \in]n, n+1[/ F_Z(n+1) - F_Z(n) = (n+1-n)F_Z'(c) = f_Z(c) \approx f_Z(n)$$

Dès lors :

$$\forall n \in \mathbb{N}^*,\ H(n) \approx \lambda e^{-\lambda n} \Leftrightarrow \ln(H(n)) \approx \ln(\lambda) - \lambda n$$

En conséquence, le nuage de points $M_n(n, \ln(H(n))$ doit montrer un comportement linéaire et le coefficient de corrélation de la série statistique double formée des entiers naturels n de 1 à 150 et de $\ln(H(n))$, doit être proche de 1.
Par ailleurs si $y = ax + b$ est l'équation de la droite de régression de cette série double, alors $a = -\lambda$ et $b = \ln(\lambda)$. La détermination de a et de b doit donc permettre d'obtenir deux approximations concordantes de λ.

On obtient par le calcul : `np.corrcoef(np.arange(1,150),np.log(H[1:])=0.996`. Ce qui prouve le comportement exponentielle de la distribution des longueurs des segments obtenus par cassures aléatoires selon une loi uniforme.
Par ailleurs :
`a,b=np.polyfit(x1,np.log(H[1:]),1)` retourne $a = -0.0256$ et $b = \ln(\lambda) = -3.643$, soit $\lambda = 0.0256$ ou $\lambda = e^b = 0.0261$.

Conclusion : On peut estimer que $Z \hookrightarrow \mathcal{E}(\lambda)$ où $\lambda = 0.026$.

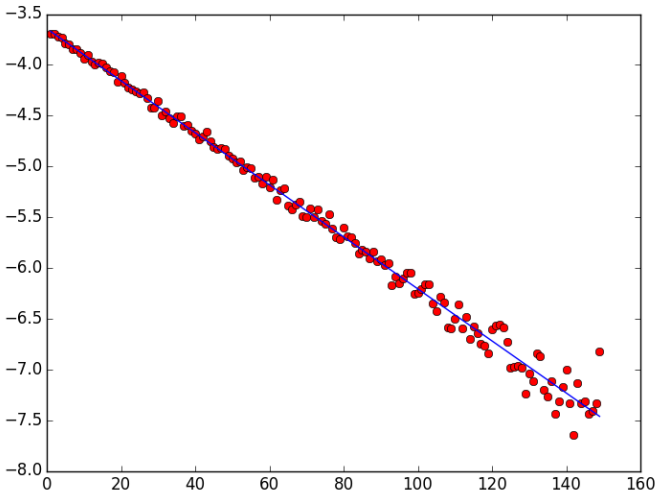

FIGURE 2 – `Tracé du nuage de points` $\{M_n(n, \ln(H(n)), n \in [\![1, 150]\!]\}$ `et de la droite de régression associée`

A la question de savoir si cela nous semble satisfaisant, il suffit de rappeler que $\mathbb{E}(Z) = \dfrac{1}{\lambda} = 38.46$ là où la première partie nous assure que $\mathbb{E}(Z) = \dfrac{N}{n+1} = 25000/(2 * 320 + 1) = 39.00$

✐ On retiendra que si n cassures ont lieu selon une loi uniforme sur l'intervalle $[\![1, N]\!]$, alors la longueur des $n + 1$ intervalles formés suit une loi exponentielle de paramètre n/N.

7. On observe ci-dessous l'histogramme des longueurs de séquences homologues supérieures à 1 million de nucléotides (blocs de synthénie) chez la souris et chez l'homme. Pouvez-vous désormais prendre partie dans la querelle qui oppose les tenants des cassures aléatoires et ceux des régions fragiles ?

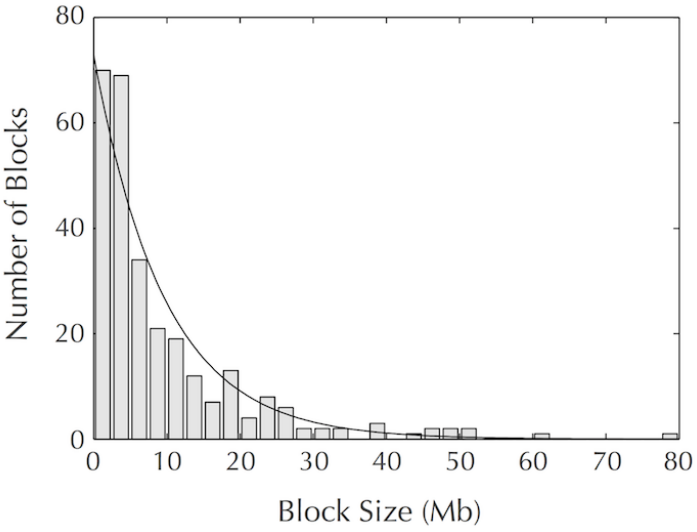

FIGURE 3 – `Histogramme des longueurs de blocs de synthénie homme-souris en 1996, 1416 marqueurs définissant 181 segments conservés (seuls les blocs dont la longueur est supérieure à 1 Mb sont pris en compte). L'histogramme est approché par une distribution exponentielle.`

L'adéquation obtenue en 1996 entre les fréquences des longueurs de 181 blocs de synthénie homme-souris (de longueur supérieur à 1 Mb) et une distribution exponentielle laisse penser que les deux génomes sont issus d'un ancêtre commun et formé chacun par inversion selon le modèle des cassures aléatoires.

Pour autant, la victoire des tenants de ce modèle sera de courte durée car à partir de 2003, les génomes de l'homme et de la souris sont entièrement séquencés et une première version de leur assemblage est disponible. Si à grande échelle, les résultats sont compatibles avec ceux obtenus avec des cartes génétiques et avec le modèle des cassures aléatoires, avec 281 segments conservés de plus de 1 Mb, par contre, à plus petite échelle, le nombre de petits segments conservés est plus important qu'attendu et ne peut être expliqué par le modèle de cassures aléatoires.
Il semble bien au bout du compte qu'il y ait effectivement des régions fragiles...