Projets MAP311

Modélisation bayésienne des genres

Xiang CHEN & Xin CHEN

26/06/2017

Modélisation bayésienne des genres

1. Préliminaires.

1) On a $f_{p,\theta}(x) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} 1_{]0,1[}(x)$

Par la definition:

$$\mathbb{E}_\theta[P] = \int_R x f_{p,\theta}(x)\, dx = \int_0^1 \frac{1}{B(a,b)} x^a (1-x)^{b-1} dx = \frac{B(a+1,b)}{B(a,b)}$$

On rappelle $\Gamma(n) = \int_0^{+\infty} t^{n-1} e^{-t} dt$

Par Intégration par partie

On a $\Gamma(n) = \left[\frac{1}{n} t^n e^{-t}\right]_0^{+\infty} + \int_0^{+\infty} \frac{1}{n} t^n e^{-t} dt = 0 + \int_0^{+\infty} \frac{1}{n} t^n e^{-t} dt = \frac{\Gamma(n+1)}{n}$

Comme $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$

Alors $\mathbb{E}_\theta[P] = \frac{B(a+1,b)}{B(a,b)} = \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} \times \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$

$$= \frac{\Gamma(a+1)}{\Gamma(a)} \times \frac{\Gamma(a+b)}{\Gamma(a+b+1)}$$

$$= \frac{a}{a+b}$$

On a $p_0 = \mathbb{E}_\theta[P] = \frac{a}{a+b}$

De même, $\mathbb{E}_\theta[P^v(1-P)^w] = \int_R x^v (1-x)^w f_{p,\theta}(x)\, dx$

$$= \int_0^1 \frac{1}{B(a,b)} x^{v+a-1}(1-x)^{w+b-1} dx$$

$$= \frac{B(a+v,b+w)}{B(a,b)}$$

2) On veut la loi conditionnelle de X_n sachant P

Si $P = p$

Alors $\mathbb{P}_\theta(X_n = k | P = p) = \binom{n}{k} p^k (1-p)^{n-k}$

Donc $\mathbb{P}_\theta(X_n = k | P) = \binom{n}{k} P^k (1-P)^{n-k}$

C'est-à-dire $X_n | P \sim B(n, P)$

3) On prend $g_1: \begin{array}{ccc} \mathbb{R}^+ & \to & \mathbb{R}^+ \\ x & \mapsto & x \end{array}$ mesurable positive

Alors on a $\mathbb{E}_\theta[\mathbb{E}_\theta[g_1(X_n)|P]] = \mathbb{E}_\theta[g_1(X_n)]$

Donc $\mathbb{E}_\theta[X_n] = \mathbb{E}_\theta\left[\mathbb{E}_\theta[X_n|P]\right]$

$= \mathbb{E}_\theta[nP]$ car $X_n|P \sim B(n,P)$

$= n\mathbb{E}_\theta[P]$

$= \frac{na}{a+b}$

On prend $g_2: \begin{array}{ccc} \mathbb{R}^+ & \to & \mathbb{R}^+ \\ x & \longmapsto & x \end{array}$ mesurable positive

De même $\mathbb{E}_\theta[\mathbb{E}_\theta[g_2(X_n)|P]] = \mathbb{E}_\theta[g_2(X_n)]$

Donc $\mathbb{E}_\theta\left[{X_n}^2\right] = \mathbb{E}_\theta\left[\mathbb{E}_\theta\left[{X_n}^2|P\right]\right]$

Or $\mathbb{E}_\theta\left[{X_n}^2|P\right] = \mathbb{V}ar_\theta[X_n|P] + (\mathbb{E}_\theta[X_n|P])^2 = nP(1-P) + n^2P^2$

Donc $\mathbb{E}_\theta\left[{X_n}^2\right] = \mathbb{E}_\theta\left[\mathbb{E}_\theta\left[{X_n}^2|P\right]\right]$

$= \mathbb{E}_\theta[nP(1-P) + n^2P^2]$

$= n^2\mathbb{E}_\theta[P^2] + n\mathbb{E}_\theta[P(1-P)]$

Alors $\mathbb{V}ar_\theta[X_n] = \mathbb{E}_\theta\left[{X_n}^2\right] - \mathbb{E}_\theta[X_n]^2$

$= n^2\mathbb{E}_\theta[P^2] + n\mathbb{E}_\theta[P(1-P)] - n^2\mathbb{E}_\theta[P]^2$

$= n\mathbb{E}_\theta[P(1-P)] + n^2\,\mathbb{V}ar_\theta[P]$

$\mathbb{E}_\theta[P^2] = \int_R x^2 f_{p,\theta}(x)\,dx = \int_0^1 \frac{1}{B(a,b)} x^{a+1}(1-x)^{b-1}dx = \frac{B(a+2,b)}{B(a,b)}$

$= \frac{\Gamma(a+2)\Gamma(b)}{\Gamma(a+b+2)} \times \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$

$= \frac{\Gamma(a+2)}{\Gamma(a+1)} \times \frac{\Gamma(a+1)}{\Gamma(a)} \times \frac{\Gamma(a+b)}{\Gamma(a+b+1)} \times \frac{\Gamma(a+b+1)}{\Gamma(a+b+2)}$

$= \frac{a(a+1)}{(a+\mathrm{b})(a+\mathrm{b}+1)}$

$= \frac{a+1}{a+\mathrm{b}+1}\mathbb{E}_\theta[P]$

$\mathbb{V}ar_\theta[P] = \mathbb{E}_\theta[P^2] - \mathbb{E}_\theta[P]^2$

Alors $\mathbb{V}ar_\theta[X_n] = n\mathbb{E}_\theta[P] - n\mathbb{E}_\theta[P^2] + n^2\,\mathbb{V}ar_\theta[P]$

$= n\mathbb{E}_\theta[P] - n\frac{a+1}{a+\mathrm{b}+1}\mathbb{E}_\theta[P] + n^2\frac{a+1}{a+\mathrm{b}+1}\mathbb{E}_\theta[P] - n^2\mathbb{E}_\theta[P]^2$

Comme $n^2\frac{a+1}{a+\mathrm{b}+1}\mathbb{E}_\theta[P] - n^2\mathbb{E}_\theta[P]^2$

$= n^2\mathbb{E}_\theta[P](\frac{a+1}{a+\mathrm{b}+1} - \frac{a}{a+b})$

$= n^2\frac{b}{(a+b)(a+b+1)}\mathbb{E}_\theta[P]$

On a $\mathbb{V}ar_\theta[X_n] = n\mathbb{E}_\theta[P] - n\frac{a+1}{a+\mathrm{b}+1}\mathbb{E}_\theta[P] + n^2\frac{b}{(a+b)(a+b+1)}\mathbb{E}_\theta[P]$

$$= \mathbb{E}_\theta[P](n - n\frac{a+1}{a+\mathrm{b}+1} + n^2\frac{b}{(a+b)(a+b+1)})$$

$$= \mathbb{E}_\theta[P](n\frac{b}{a+\mathrm{b}+1} + n^2\frac{b}{(a+b)(a+b+1)})$$

$$= \mathbb{E}_\theta[P](n\frac{b}{a+\mathrm{b}} \times \frac{a+b}{a+\mathrm{b}+1} + n^2\frac{b}{a+\mathrm{b}} \times \frac{1}{a+b+1})$$

$$= \mathbb{E}_\theta[P] \times n \times (1 - \frac{a}{a+\mathrm{b}})(\frac{a+b}{a+\mathrm{b}+1} + n\frac{1}{a+b+1})$$

$$= \mathbb{E}_\theta[P] \times (n - n\mathbb{E}_\theta[P]) \times \frac{a+b+n}{a+\mathrm{b}+1}$$

$$= \frac{1}{n}\mathbb{E}_\theta[X_n] \times (n - \mathbb{E}_\theta[X_n]) \times (1 + \frac{n-1}{a+\mathrm{b}+1})$$

2. Estimation des paramètres

1) Si $n = 1$

   D'après la question précédente

   On a $\mathbb{E}_\theta[X_1] = 1 \times \frac{a}{a+b} = p_0$

   Et $\mathbb{V}ar_\theta[X_n] = 1 \times \mathbb{E}_\theta[X_1] \times (1 - \mathbb{E}_\theta[X_1]) \times 1 = p_0(1 - p_0)$

   $\mathbb{P}_\theta(X_1 = 1) = p_0$

   $\mathbb{P}_\theta(X_1 = 1) = 1 - p_0$

   Donc la loi de X_1 ne dépend que p_0

   De plus on ne peut que estimer la valeur de $p_0 = \frac{a}{a+b}$

   Pour une valeur de p_0, on peut trouver plusieurs couples (a,b) qui vérifient $p_0 = \frac{a}{a+b}$

   D'où c'est impossible d'estimer la paramètre $\theta = (a, b)$

2) Si $n \geq 2$

   On veut construire $\widehat{\theta_k}$

   à partir de $M_k = \frac{1}{K}\sum_{k=1}^{K} X_n^{(k)}$ et $V_k = \frac{1}{K}\sum_{k=1}^{K}(X_n^{(k)})^2 - M_k^2$

   On note $\widehat{\theta_k} = (\widehat{a_k}, \widehat{b_k})$

   M_k est l'estimateur de moyenne

   V_k est l'estimateur de variance

   Comme $\begin{cases} \mathbb{E}_\theta[X_n] = \frac{na}{a+b} \\ \mathbb{V}ar_\theta[X_n] = \frac{1}{n}\mathbb{E}_\theta[X_n] \times (n - \mathbb{E}_\theta[X_n]) \times (1 + \frac{n-1}{a+\mathrm{b}+1}) \end{cases}$

On a $\begin{cases} M_k = \frac{n\widehat{a_k}}{\widehat{a_k}+\widehat{b_k}} \\ V_k = \frac{M_k}{n} \times (n - M_k) \times (1 + \frac{n-1}{\widehat{a_k}+\widehat{b_k}+1}) \end{cases}$

D'après méthode des moments

On peut en déduire $\widehat{a_k}\ et\ \widehat{b_k}$ en fonction de $M_k\ et\ V_k$

On obtient

$$\begin{cases} \widehat{a_k} = \frac{nM_k(n-M_k)-nV_k}{nV_k-M_k(n-M_k)} \times \frac{M_k}{n} \\ \widehat{b_k} = \frac{nM_k(n-M_k)-nV_k}{nV_k-M_k(n-M_k)} \times (1 - \frac{M_k}{n}) \end{cases}$$

3) D'après la loi de Grande nombre :

$$M_K = \frac{x_n^{(1)}+\cdots+x_n^{(K)}}{K} \xrightarrow[p.s\ k\to+\infty]{} E_\theta\left[x_n^{(1)}\right] = \frac{na}{a+b}$$

$$V_K = \frac{1}{K}\sum_{k=1}^{K}\left(x_n^{(k)}\right)^2 - M_K^2 \xrightarrow[p.s.\ k\to+\infty]{} E_\theta\left[\left(x_n^{(1)}\right)^2\right] - E_\theta^2\left[x_n^{(1)}\right] = Var_\theta\left(x_n^{(1)}\right)$$

$$= \frac{nab}{(a+b)^2}(1 + \frac{n-1}{a+b+1})$$

Donc $\widehat{a_k} \xrightarrow[p.s.\ k\to+\infty]{} \frac{n\times\frac{na}{a+b}\times\frac{nb}{a+b}-\frac{n^2ab}{(a+b)^2}\left(1+\frac{n-1}{a+b+1}\right)}{\frac{n^2ab}{(a+b)^2}\left(1+\frac{n-1}{a+b+1}\right)-\frac{na}{a+b}\times\frac{nb}{a+b}} \times \frac{na}{a+b} \times \frac{1}{n}$

$$= \frac{(n-1)(1-\frac{1}{a+b+1})}{1+\frac{n-1}{a+b+1}-1} \times \frac{a}{a+b} = a$$

De même $\widehat{b_k} \xrightarrow[p.s\ k\to+\infty]{} (a+b) \times \left(1 - \frac{a}{a+b}\right) = b$

Alors $\widehat{\theta_k} = (\widehat{a_k}, \widehat{a_k}) \xrightarrow[k\to+\infty]{} (a,b) = \theta$

4) On regroupe $x_n^{(k)}$ k=1,2,...,K

$\forall r \in [\![0,n]\!]$ on note l'ensemble $K_r = \{x_n^{(k)}: x_n^{(k)} = r\}$ alors $|K_r| = N_r$, $|K_r|$ représente cardinal, donc $\sum_{k=1}^{K} x_n^{(k)} = \sum_{r=0}^{n}\sum_{k=1}^{|K_r|}\left(x_n^{(k)} = r\right) = \sum_{r=0}^{n}\sum_{k=1}^{N_r} r = \sum_{r=0}^{n} rN_r$

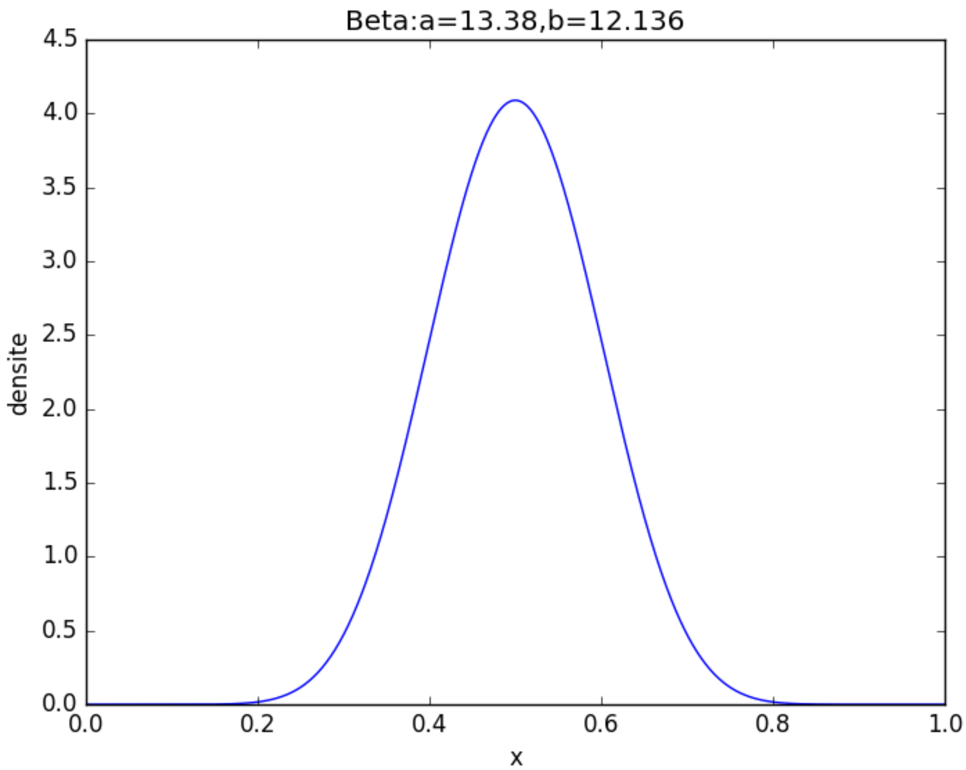
$$\sum_{k=1}^{K} {x_n^{(k)}}^2 = \sum_{r=0}^{n}\sum_{k=1}^{|K_r|}\left(x_n^{(k)} = r\right)^2 = \sum_{r=0}^{n}\sum_{k=1}^{N_r} r^2 = \sum_{r=0}^{n} r^2N_r$$

Valeur numérique : $M_K = 2.621875$, $V_K = 1.43514648437$,

$$\theta = (13{,}37995764, 12{,}13615216)$$

5)On utilise *scipy.stats.beta* pour tracer la densité et estime que $p_1 = P_\theta\left(P > 1/2\right) = 0.598253083758$

Figure 1 la densité



6) On choisit la nombre de famille K=50, 200, 500 et 1000, et simule N=1000 fois, tracer la densité, pour vérifier que θ est asymptotiquement normale.

Figure 2 K=50

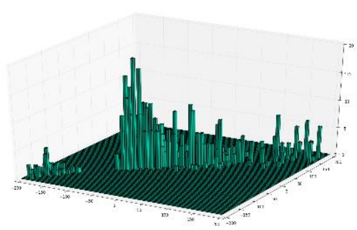


Figure 3 K=200

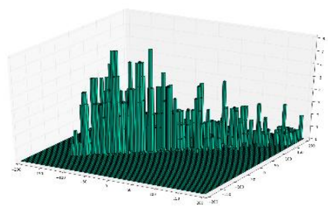


Figure 4 K=500

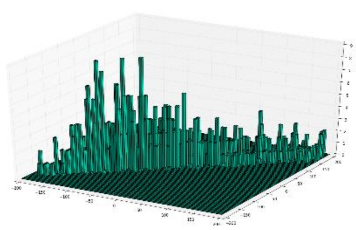
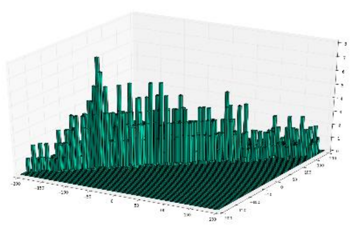


Figure 5 K=1000



Explication : pour obtenir la valeur moyenne M_k et et la variante V_k, on choisit que la nombre de famille k soit 50, 200, 500 et 1000, dans ces kfamilles, tout d'abord, on stimulant la probabilité d'avoir un garçon dans une famille, et puis, on simule le nombre de garçon dans cette famille, donc dans tous les k familles, on obtient $\left(x_5^{(1)},\ldots,x_5^{(k)}\right)$, en calculant M_k et V_k, or on a la relation entre (a, b) et (M_k, V_k), on obtient (a, b) d'où $\sqrt{K}\ \ (\widehat{\theta_K}-\theta)$. On le répète N=1000 fois, et puis on trace les résultats.

7) lorsque K=1000, N=1000, on obtient la matrice de covariance est :

$$\begin{bmatrix} 635.214404365997 & 600.6343725306089 \\ 600.6343725306089 & 568.5224295398386 \end{bmatrix}$$

3. Estimation de la probabilité d'avoir un garçon

1) On veut calculer $E_\theta[\mathrm{h(p)}|x_n=\mathrm{k}]=\frac{E_\theta[h(p)1_{\{x_n=k\}}]}{P_\theta(x_n=k)}$

En fait, $P_\theta(x_n=k)=E_\theta\left[1_{\{x_n=k\}}\right]=E_\theta\left[E_\theta\left[1_{\{x_n=k\}}\right]\middle|P\right]=E_\theta[P_\theta(x_n=k|P)]$

$$=E_\theta\left[\binom{n}{k}P^k(1-P)^{n-k}\right]$$

$$E_\theta\left[h(p)1_{\{x_n=k\}}\right]=E_\theta\left[E_\theta\left[h(P)1_{\{x_n=k\}}\middle|P\right]\right]=E_\theta[h(p)P_\theta(x_n=k|P)]$$

$$=E_\theta[h(P)\binom{n}{k}P^k(1-P)^{n-k}]$$

D'où $E_\theta[h(p)|x_n=k]=\frac{E_\theta[h(P)\binom{n}{k}P^k(1-P)^{n-k}]}{E_\theta\left[\binom{n}{k}P^k(1-P)^{n-k}\right]}=\frac{E_\theta[h(P)P^k(1-P)^{n-k}]}{E_\theta[P^k(1-P)^{n-k}]}$

2) D'après les résultats de question 1 de la partie 1, on a :

$$E_\theta[P^v(1-P)^w]=\frac{B(v+a,w+b)}{B(a.b)}$$

On prend $\mathrm{h}:\mathrm{x}\longmapsto x$ bornée sur [0, 1]

Donc $E_\theta[P|x_n=k]=\frac{E_\theta[P^{k+1}(1-P)^{n-k}]}{E_\theta[P^k(1-P)^{n-k}]}=\frac{B(k+a+1,n-k+b)}{B(k+a,n-k+b)}=\frac{\Gamma(\mathrm{k+a+1})\Gamma(\mathrm{n-k+b})}{\Gamma(\mathrm{k+a+1+n-k+b})}\times\frac{\Gamma(\mathrm{k+a+n-k+b})}{\Gamma(\mathrm{k+a})\Gamma(\mathrm{n-k+b})}=\frac{k+a}{a+b+n}$

De même, $E_\theta[P^2|x_n=k]=\frac{B(k+a+2,n-k+b)}{B(k+a,n-k+b)}=\frac{(k+a+1)(k+a)}{(a+b+n)(a+b+n+1)}$

Donc $\mathrm{Var[P}|X_n=\mathrm{k]}=E_\theta[P^2|x_n=k]-E_\theta^2[P|x_n=k]=\frac{k+a}{a+b+n}\times\frac{(a+k)(b+n-k)}{(a+b+n)(a+b+n+1)}$

Alors pour $\mathrm{P}|x_n=k, E_\theta[P|x_n=k]=\frac{a+k}{a+b+n}$,

$$\mathrm{Var[P}|x_n=k]=\frac{(a+k)^2(a+n-k)}{(a+b+n)^2(a+b+n+1)}$$

Pour P $E_\theta[P] = \frac{a}{a+b}, Var_\theta[P] = \frac{a^2 b}{(a+b)^2(a+b+1)}$

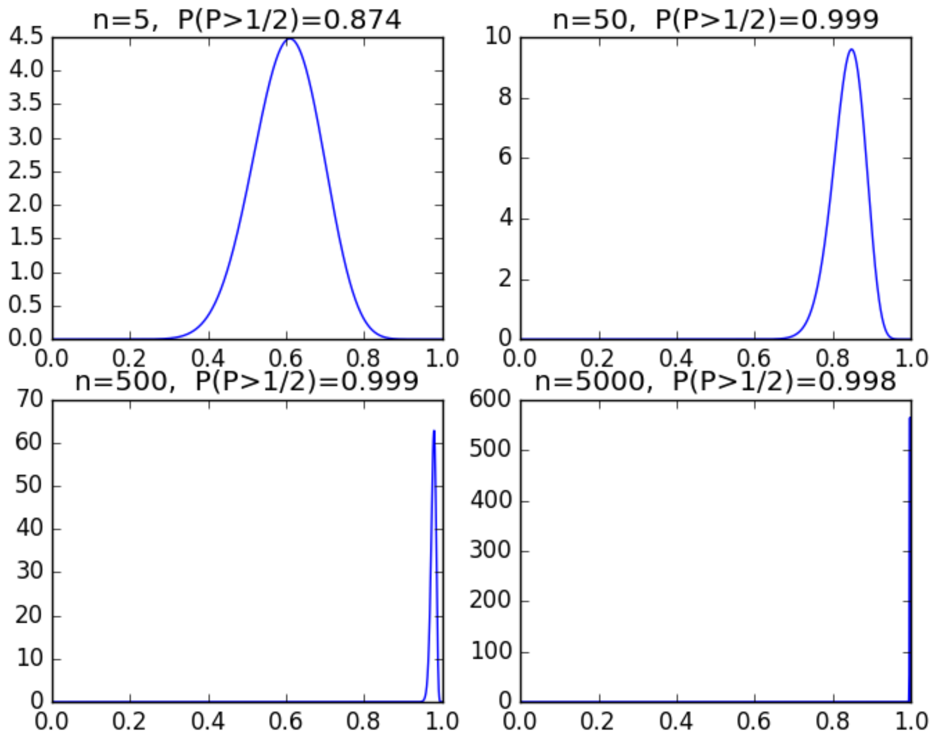
On fait analogue $\mathrm{a} + \mathrm{k} \leftrightarrow a,\ b + n - \mathrm{k} \leftrightarrow b$

Alors $\mathrm{P}|x_n = k$ suit une même loi que P avec la paramètre (a+k, b+n-k)

Donc $f_{P|x_n=k,\theta}(x) = \frac{1}{B(a+k,b+n-k)} x^{a+k-1}(1-x)^{b+n-k-1} \times 1_{]0,1[}(x)$

$$f_{P|x_n}(x) = \frac{1}{B(a+x_n,b+n-x_n)} x^{a+x_n-1}(1-x)^{b+n-x_n-1} 1_{]0,1[}(x)$$

3) Ci-dessous c'est la figure de densité pour n=5, 50, 500 et 5000



4) Merci à Python, on a calculé $P_\theta(P > 1/2|x_n = n)$ affiché ci-dessus.

Si les n précédents sont des garçons, alors le sexe du n+1-ème enfant est :

$$\begin{array}{c} garçon, \quad si\ n\ est\ grand \\ on\ ne\ peut\ pas\ prévoir, \quad si\ n\ n'\text{est pas grand} \end{array}$$

Liste des programmes :

- 2-5.py pour la question 2.5
- 2-6.py pour la question 2.6
- 2-7.py pour la question 2.7
- 3-3et3-4.py pour la question 3.3 et la question 3.4