UNIVERSITÉ DE THIES

UFR DES SCIENCES ECONOMIQUES ET SOCIALES

UFR DES SCIENCES ET TECHNOLOGIQUES

Master 1 Science Des Données et Applications

Par

JOHANA BINTA VITALE FAYE

ALMAMY YOUSSOUF LY

COUMBA SY

Projet 1 de Statistiques

Professeur : Dr. L. Nging

Année universitaire 2019-2020

Note :

Ceci est un rapport du projet 1 de Statistique.

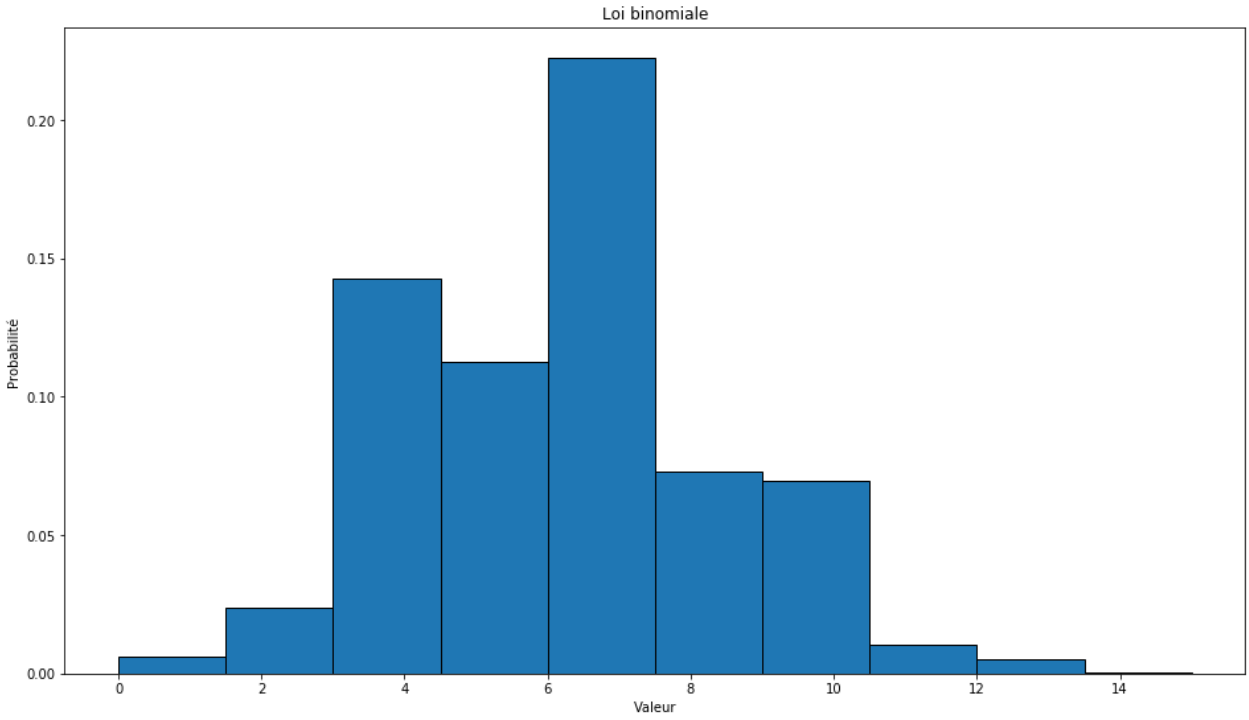
Pour plus détails veuillez-vous référer à nos notebooks.

Question 1 : *Simulation de lois*

1. Simuler un échantillon de taille 1000 suivant une loi binomiale (30, 0.2).

   Notre simulation donne le résultat suivant : [5, 4, 7, ... ,8, 8, 5]

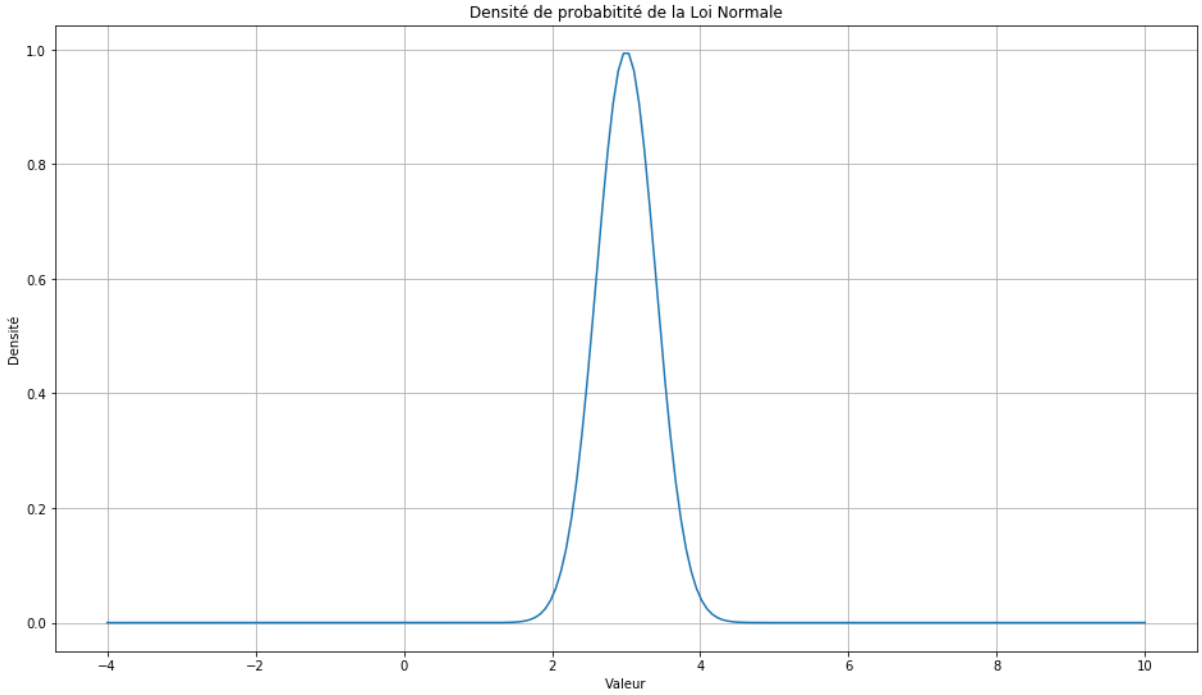
   - Tracer de l'histogramme :



2. Simuler un échantillon de taille 10000 suivant une loi normale N (3, .4)

   Notre simulation donne le résultat suivant : [3.33710088 ,2.85619027, 3.2388986, ... ,2.46561294 ,3.27344374, 2.26258547]

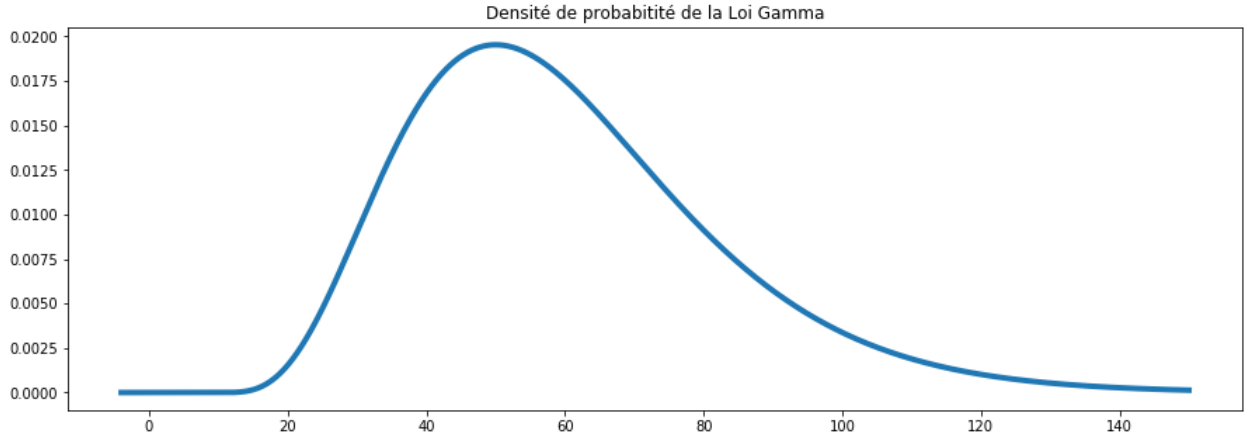
- Tracer de la fonction de densité



3. Simuler un échantillon de taille 10000 suivant une loi gamma (10, .5)

   Notre simulation donne le résultat suivant : [6.11340493, 5.41062849, 4.61776275, ... ,3.6 9655337, 3.03225534, 4.53054297]

   - Tracer la fonction de densité

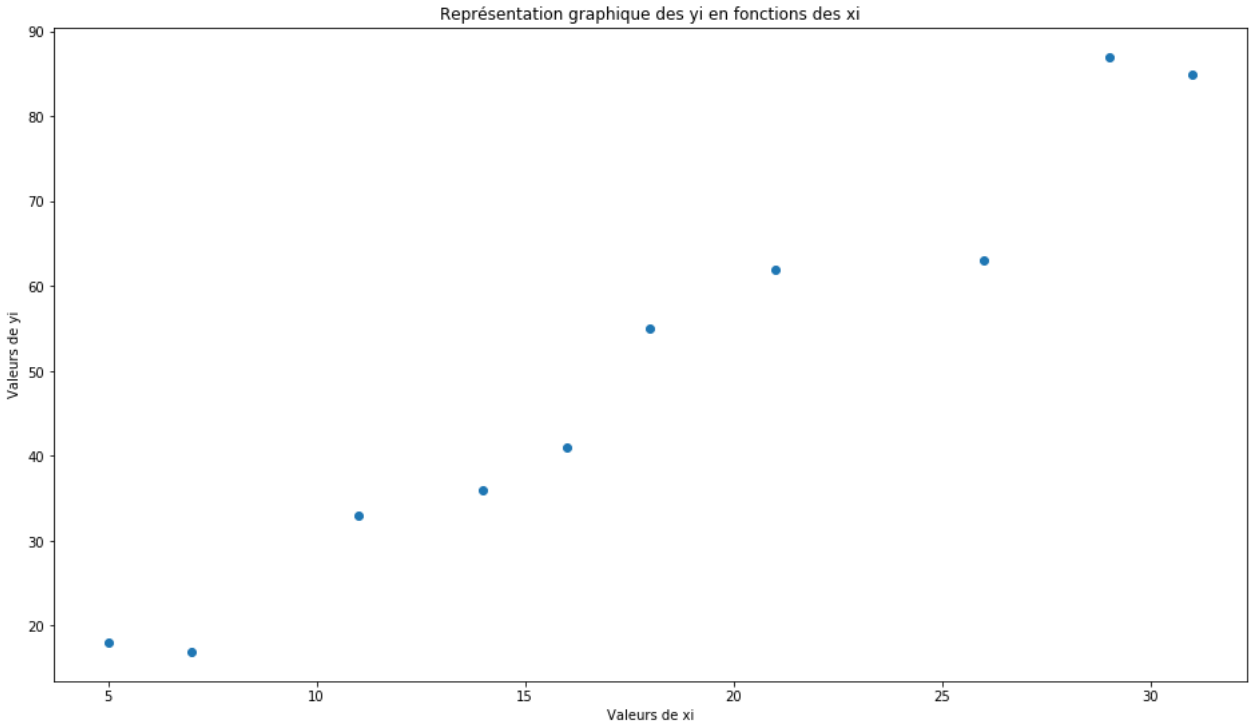


Question 2 : ***Régression linéaire simple***

1. Enregistrement dans un format adapté

	xi	yi
0	18	55
1	7	17
2	14	36
3	31	85
4	21	62
5	5	18
6	11	33
7	16	41
8	26	63
9	29	87

2. Représentation de yi = f(xi) et constat



- Constat : D'après la représentation graphique nous pouvons soupçonner une liaison linéaire entre ces deux variables car on a l'impression que les point sont situés de part et d'autre suivant l'allure d'une droite.

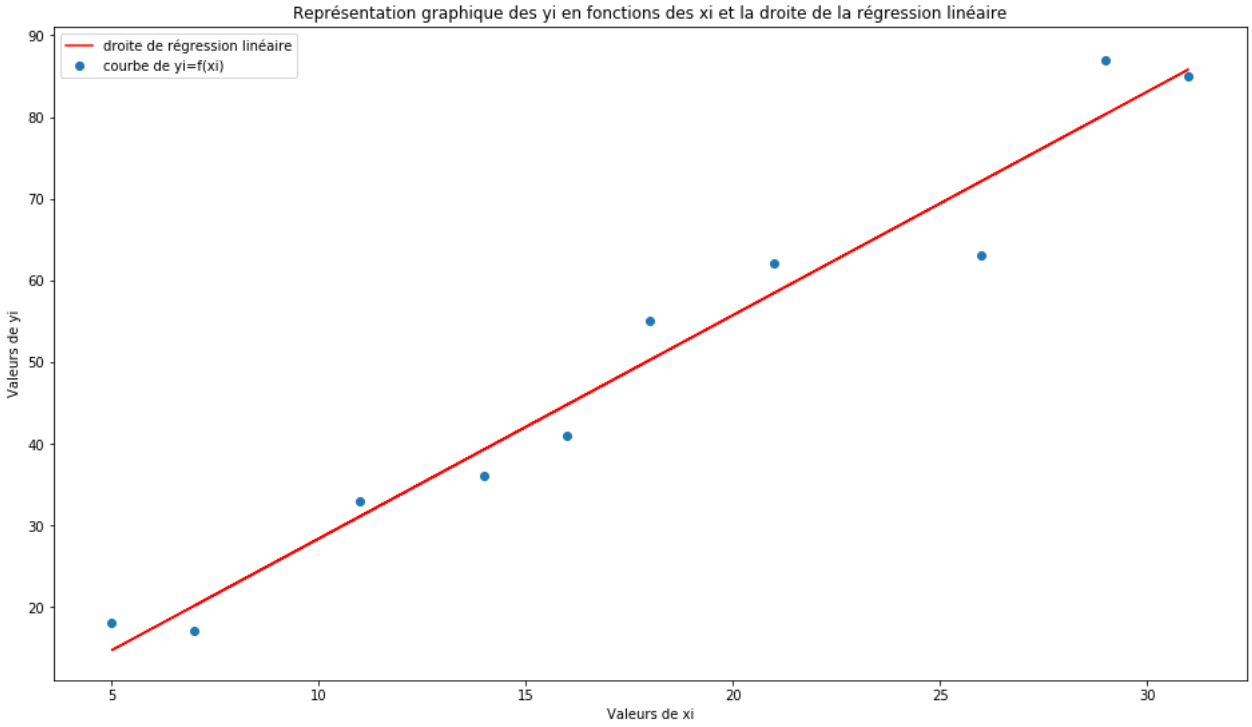
3. Déterminons les coefficients de la droite des MCO

   Les coefficients a et b sont respectivement : (2.735, 1.021)

4. Donner les ordonnées des *yi* calculés par la droite des moindres carrés correspondant aux différentes valeurs des *xi*

   Les différentes de yi sont : [50.251, 20.166, 39.311, 85.806, 58.456, 14.696, 31.106, 44.781, 72.131, 80.336]

5. Tracer de la droite sur le même graphique



6. Estimation plausible de *Y* à *xi* = 21
   La valeur de Y à xi = 21 est : 58.456

7. Ecart entre la valeur observée de Y à *xi* = 21 et la valeur estimée avec la droite des Moindres carrés et appellation de l'écart

   Cet écart est appelé l'erreur entre la valeur observée de Y à xi = 21 et celle estimée à la même abscisse avec la droite des moindres carrées ordinaires. Sa valeur est : 3.544

8. Vérifions que la droite des moindres carrés obtenue en 2 passe par le point moyen

   La droite des moindres carrés passe par le point de moyen de coordonnés : (17.8 ,49.7)

   - Généralisation :
     Vue que la formule du coeficient directeur de la droite de régression est donné par b = y_moy - a*x_moy et en remplaçant b dans l'équation de la droite de régression y = ax + b on tombe sur a = y - y_moy / x - x_moy on peut donc généraliser que p our n'importe droite de régression passe par le point moyen.
     NB : x_moy : moyenne des valeurs de x
     y_moy: moyenne des valeurs de y

Question 3 : ***Données réelles***

1. Enregistrement et vérification

   smp = pd.read_table("smp.csv", sep=";")

   Nous avons 799 observations et 26 variables.

2. Changement des types des variables

```
age              float64
prof            category
duree           category
discip          category
n.enfant         float64
n.fratrie          int64
ecole           category
separation      category
juge.enfant     category
place           category
abus            category
grav.cons       category
dep.cons        category
ago.cons        category
ptsd.cons       category
alc.cons        category
subst.cons      category
scz.cons        category
char            category
rs              category
ed              category
dr              category
suicide.s        float64
suicide.hr      category
suicide.past    category
dur.interv       float64
dtype: object
```

3. Calculer la moyenne, la variance, et l'écart type

   Pour chacune des variables ci-dessous la moyenne, la variance et l'écart-type sont :

   age (38.9, 176.16, 13.27)

   n.enfant (1.76, 3.36, 1.83)

   n.fratrie (4.29, 11.83, 3.44)

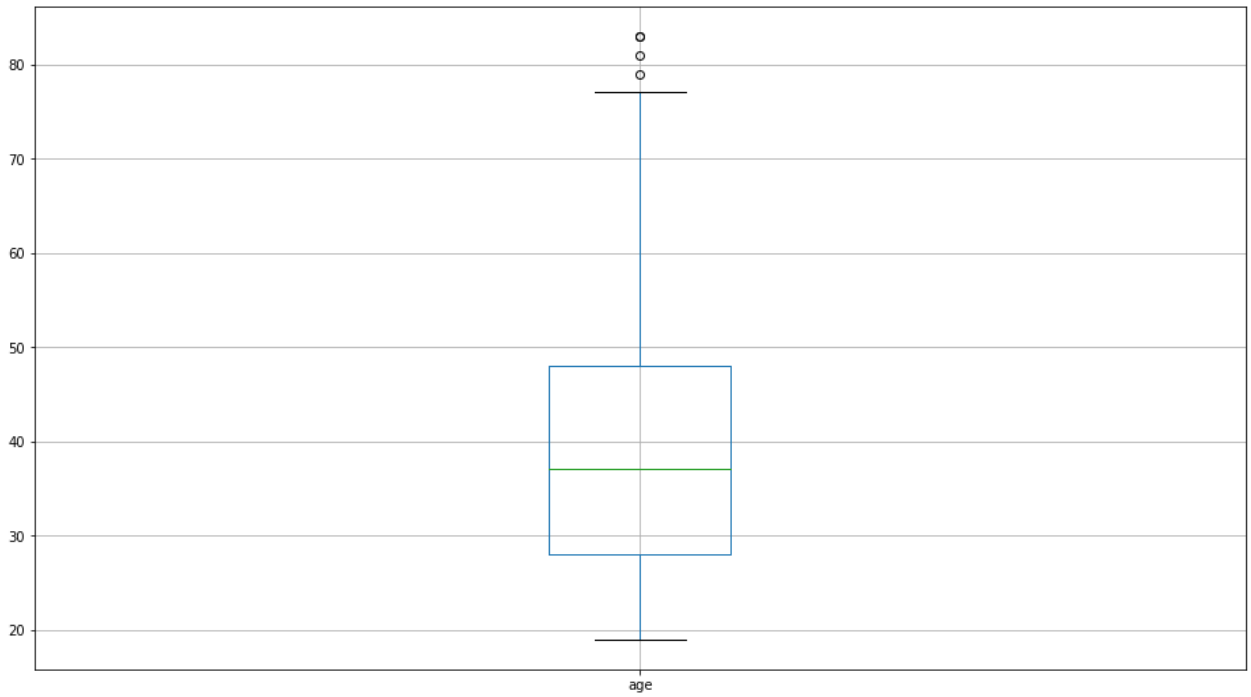
   suicide.s (0.79, 2.06, 1.43)

   dur.interv (61.89, 386.38, 19.66)

- Les 3 premières quantiles de la variable âge sont :

  count 797.000000
  mean 38.899624
  std 13.280978
  min 19.000000
  25% 28.000000 = Q1
  50% 37.000000 = Q2
  75% 48.000000 = Q3
  max 83.000000
  Name: age, dtype: float64

4. Tracer le boxplot pour la variable âge



On a un boxplot qui représente la distribution des âges de cette population carcérale.
La médiane (trait en vert) est de 37 ans environ ce qui signifie que 50% des détenus sont âgés de moins de 37 ans et 50 % sont âgés de plus de 37 ans. L'âge maximale est environ de 83 ans (le point le plus haut). L'âge minimale est de 19 ans (trait en bas).

5. Données des agriculteurs ayant plus de 2 enfants

:

	age	prof	duree	discip	n.enfant	n.fratrie	ecole	separation	juge.enfant	place	...	subst.cons	scz.cons	char	rs	ed	dr	suicide.s	suicide
14	64.0	agriculteur	NaN	0.0	3.0	2	1.0	0.0	0.0	0.0	...	0	0	1.0	1.0	1.0	3.0	0.0	
311	42.0	agriculteur	4.0	0.0	3.0	6	1.0	0.0	0.0	0.0	...	0	0	2.0	1.0	3.0	2.0	3.0	
390	36.0	agriculteur	4.0	1.0	3.0	4	3.0	1.0	1.0	1.0	...	1	0	1.0	NaN	3.0	1.0	0.0	
441	79.0	agriculteur	5.0	0.0	5.0	6	2.0	0.0	0.0	0.0	...	0	0	1.0	2.0	1.0	1.0	0.0	

4 rows × 26 columns

Nous avons donc 4 agriculteurs qui ont plus de deux enfants.

6. Calcul des fréquences des modalités de la var profession

La fréquence (%) des modalité de la variable profession sont:

ouvrier 28.63

sans emploi 27.99

employe 17.02

artisan 11.35

prof.intermediaire 7.31

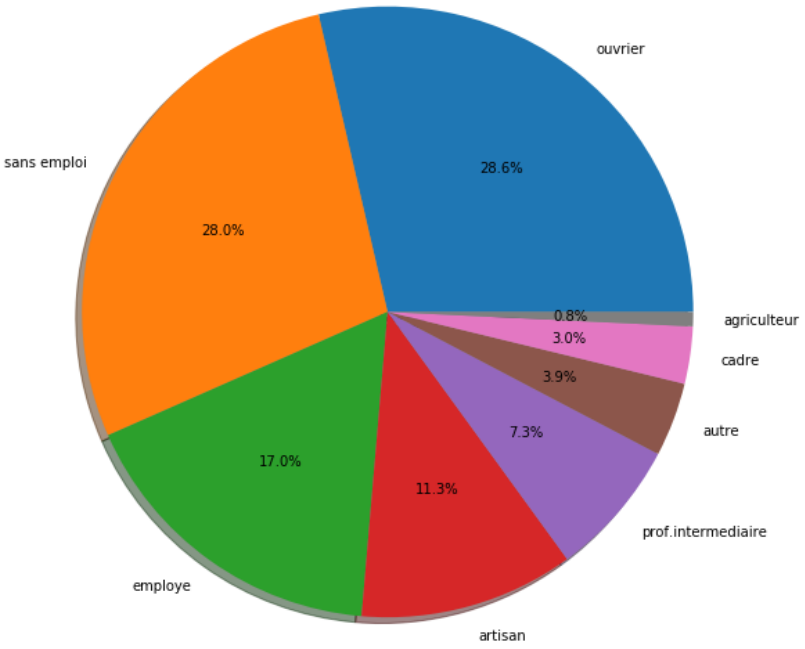
autre 3.91

cadre 3.03

agriculteur 0.76

- La catégorie modale : on a une variable qualitative nominale

7. Tracer le diagramme circulaire de la variable profession



8. Moyenne des âges par profession

	age	n.enfant	n.fratrie	suicide.s	dur.interv
	mean	mean	mean	mean	mean
prof					
agriculteur	48.833333	2.666667	3.833333	1.600000	78.750000
artisan	45.111111	2.386364	4.077778	0.517647	63.825581
autre	34.935484	1.483871	3.548387	0.500000	64.230769
cadre	50.083333	2.166667	2.791667	0.708333	56.956522
employe	38.711111	1.534351	3.940741	0.766129	62.053435
ouvrier	37.396476	1.746606	5.022026	0.895928	61.731481
prof.intermediaire	43.258621	2.107143	3.810345	0.465517	63.075472
sans emploi	35.896396	1.469484	4.283784	0.960976	61.088235

9. Donner la table des effectifs pour la var prof incluant les NaN

```
ouvrier              227
sans emploi          222
employe              135
artisan               90
prof.intermediaire    58
autre                 31
cadre                 24
NaN                    6
agriculteur            6
Name: prof, dtype: int64
```

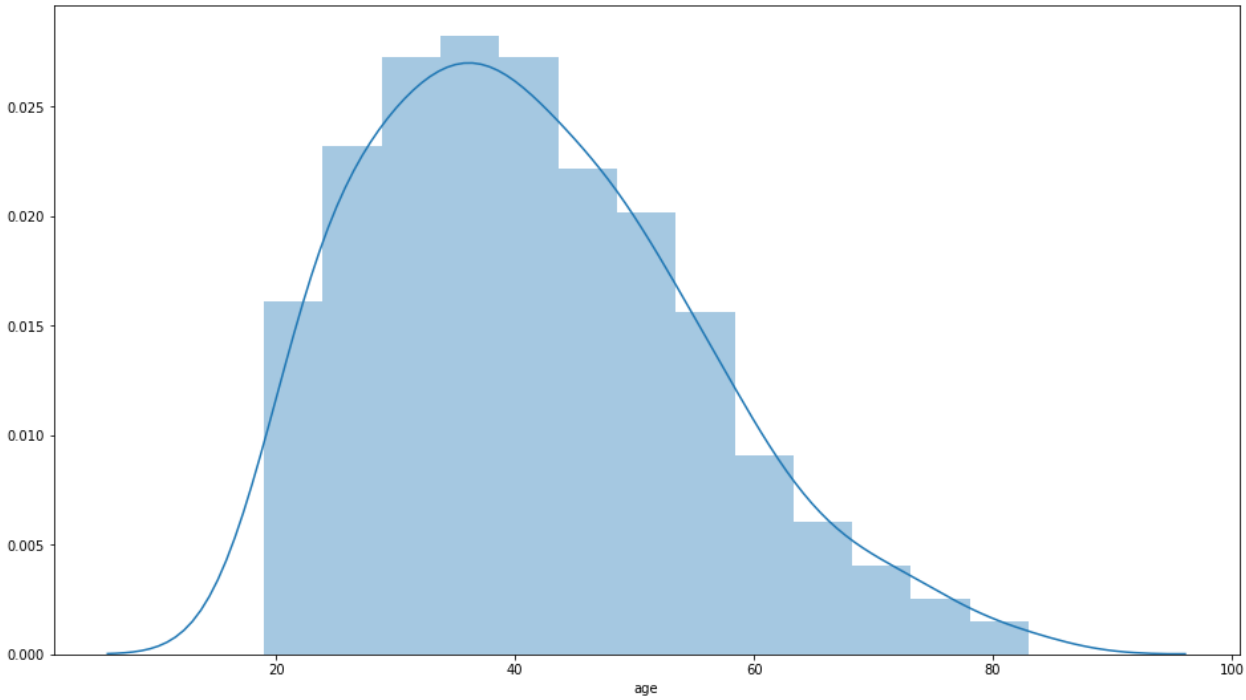
10. Donner le nombre de NaN pour chaque variable

```
age              2
prof             6
duree          223
discip           6
n.enfant        26
n.fratrie        0
ecole            5
separation      11
juge.enfant      5
place            7
abus             7
grav.cons        4
dep.cons         0
ago.cons         0
ptsd.cons        0
alc.cons         0
subst.cons       0
scz.cons         0
char            96
rs             103
ed             107
dr             111
suicide.s       41
suicide.hr      39
suicide.past    14
dur.interv      50
dtype: int64
```

11. Suppression des NaN

```
age             0
prof            0
duree           0
discip          0
n.enfant        0
n.fratrie       0
ecole           0
separation      0
juge.enfant     0
place           0
abus            0
grav.cons       0
dep.cons        0
ago.cons        0
ptsd.cons       0
alc.cons        0
subst.cons      0
scz.cons        0
char            0
rs              0
ed              0
dr              0
suicide.s       0
suicide.hr      0
suicide.past    0
dur.interv      0
dtype: int64
```

12. Tracer de l'histogramme et la densité de la variable âge sur la même figure



13. Discrétisation de la var aaléatoire âge

(48.0, 83.0] 117
(37.0, 48.0] 108
(28.0, 37.0] 100
(18.999, 28.0] 78
Name: age_classe, dtype: int64

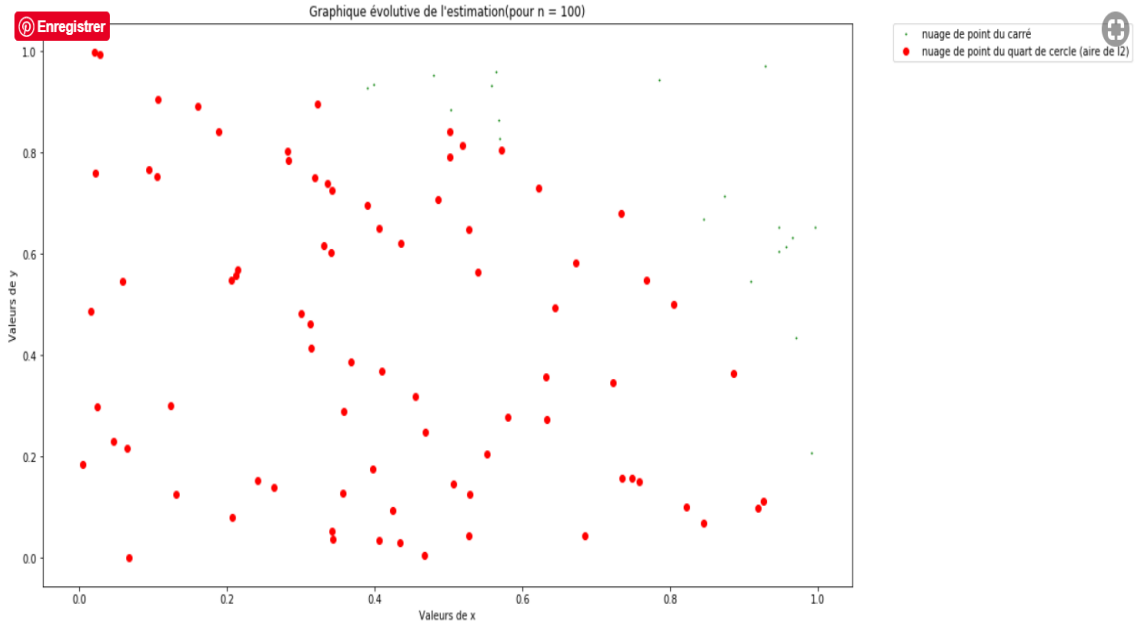
14. Donnez les fréquences des modalités de la nouvelle variable age_classe

Les fréquences (%) des modalité de la variable age_classe sont:
(48.0, 83.0] 29.03
(37.0, 48.0] 26.80
(28.0, 37.0] 24.81
(18.999, 28.0] 19.35
Name: age_classe, dtype: float64

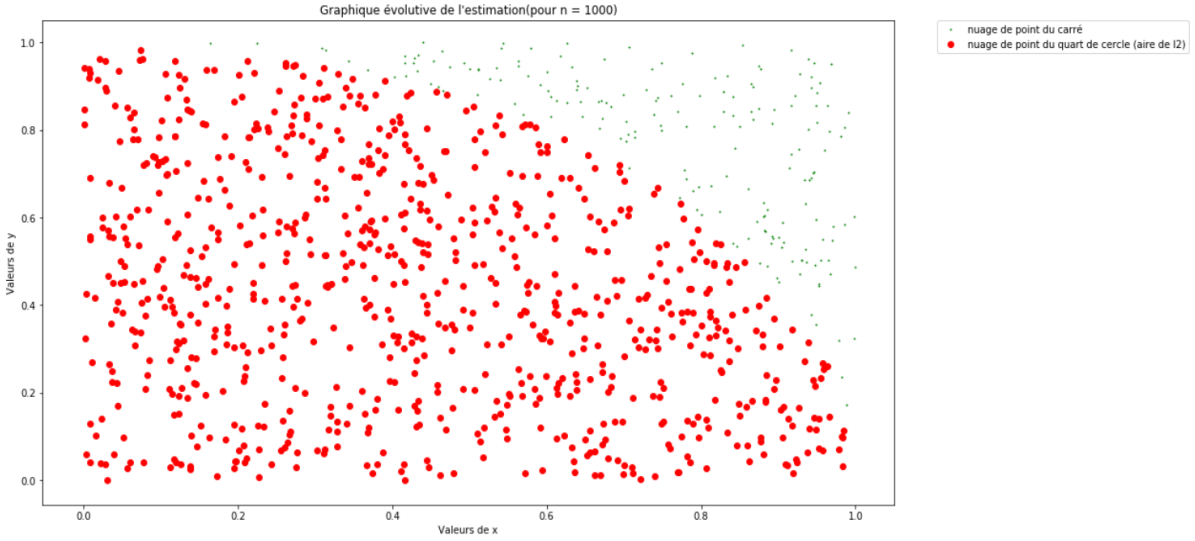
Question 4 : ***Méthode de Monté Carlos***

- Estimer I2 par une méthode de Monte Carlos avec n = 10000
  Après exécution de la fonction estimer(n), on obtient : 0.7837
- Observer par graphique l'évolution de cette estimation lorsque n varie et vérifier la cohérence avec la valeur théorique de I2

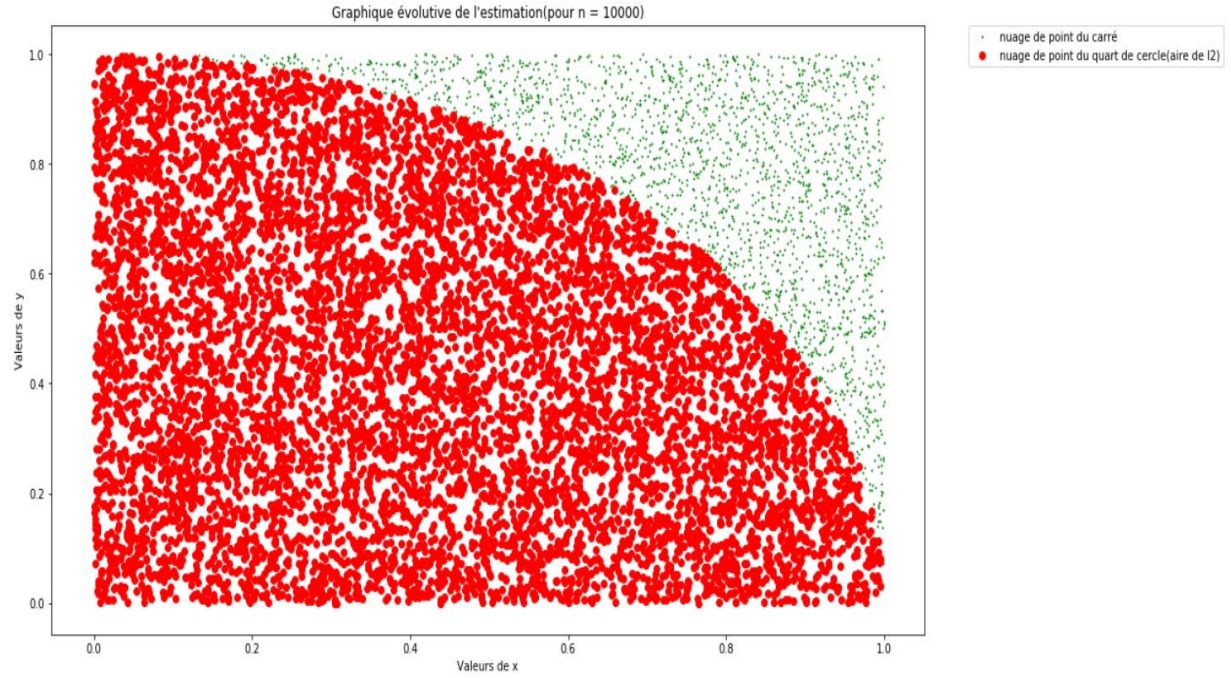
  Graphique évolutive de l'estimation (pour n = 100)



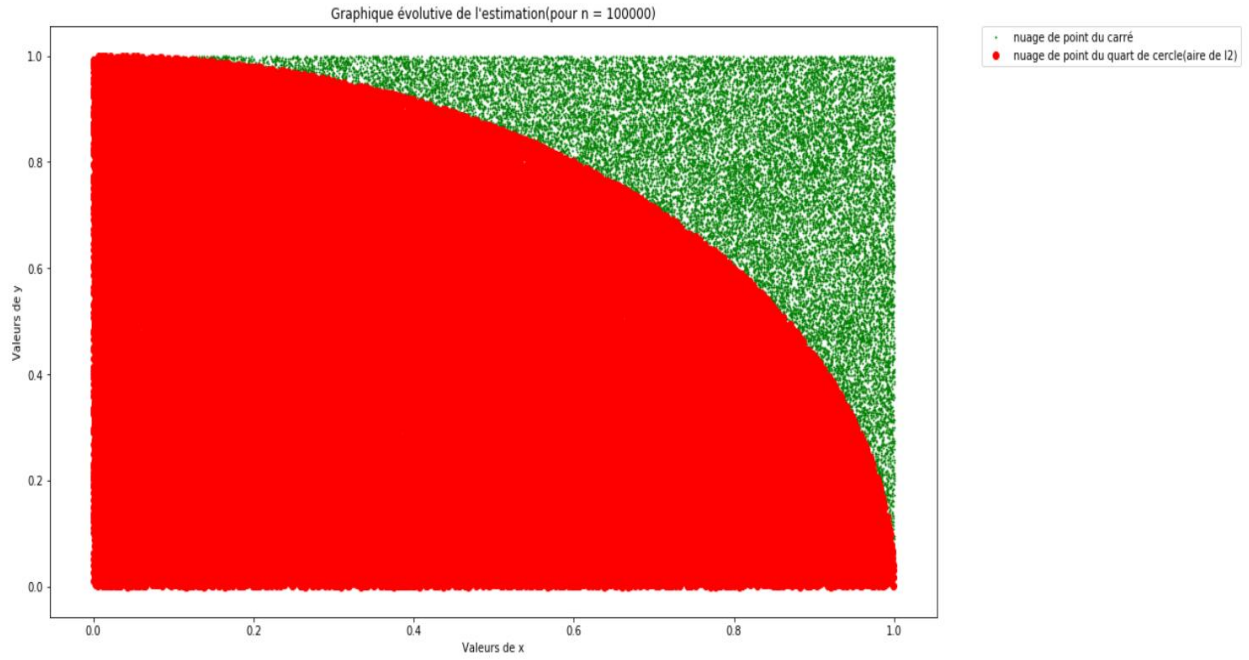
Graphique évolutive de l’estimation (pour n = 1000)



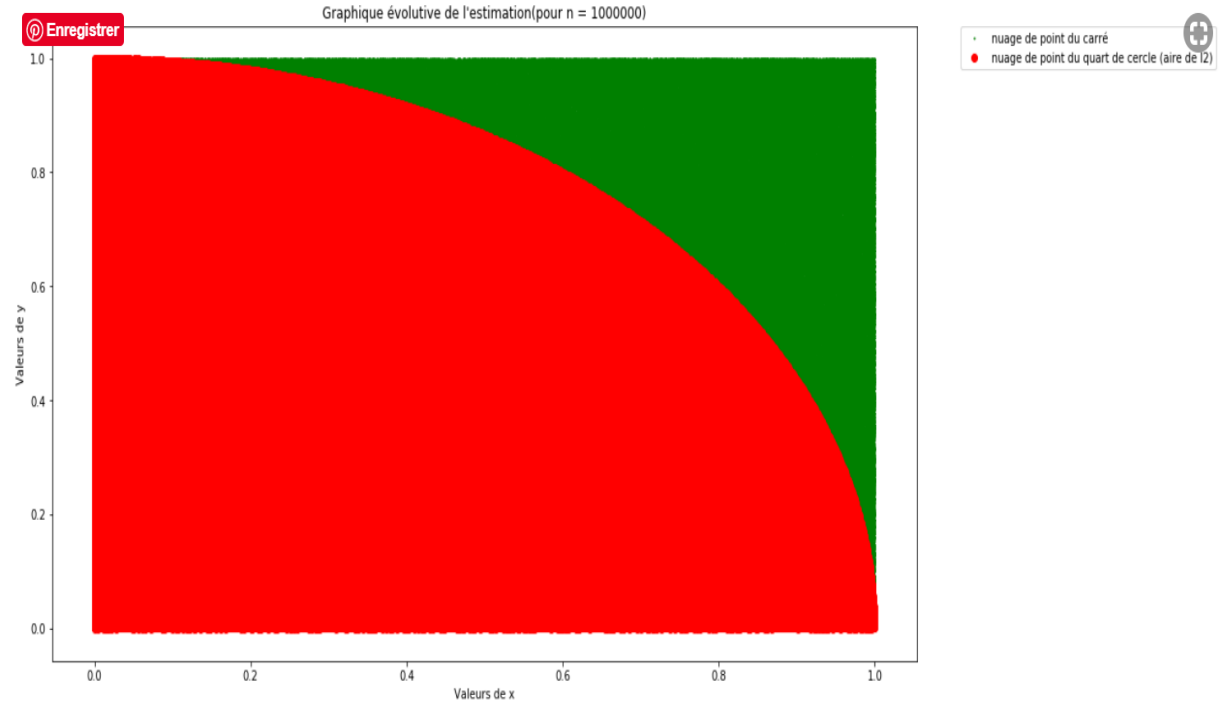
Graphique évolutive de l’estimation (pour n = 10000)



Graphique évolutive de l’estimation (pour n = 100000)



Graphique évolutive de l’estimation (pour n = 1000000)



D'après la fonction d’estimation (au niveau de 1), le calcul de la valeur de l'intégral de I2 (voire notebook question 4) et les observations graphiques: on voit que lorsque n augmente on obtient une délimitation exacte de l'aire de I2 et une approximation sensible ment égale à la valeur de I2 tandis que l'on note l'effet contraire lorsque la valeur de n diminue.