

Ébauche de parties de sujets HEC-ESSEC Mathématiques appliquées

19 janvier 2022

ÉNONCÉ 1

On considère un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$.

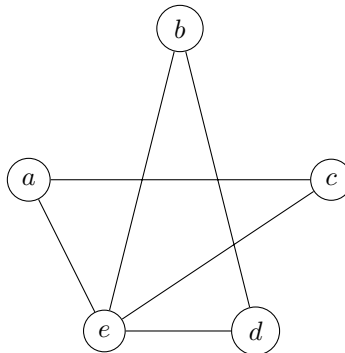
Graphes aléatoires d'Erdős-Renyi

Pour définir un graphe aléatoire non orienté G on se donne :

- $S = \{s_1, \dots, s_n\}$, un ensemble fini de $n \geq 2$ sommets ;
- pour toute paire de sommets $\{s_i, s_j\}$ avec $i < j$, $T_{i,j}$ est une variable de Bernoulli de paramètre p sur $(\Omega, \mathcal{A}, \mathbb{P})$;

Les arêtes du graphe sont les paires $\{s_i, s_j\}$ telles que $T_{i,j} = 1$ avec $i < j$. Les variables $T_{i,j}$ sont supposées indépendantes.

Voici un exemple de graphe aléatoire avec $S = \{a, b, c, d, e\}$ et $p = 0,4$:



On peut considérer qu'un graphe aléatoire est un modèle très simplifié de réseau social à un instant donné.

1. Quel est le nombre maximal d'arêtes de G ?
2. Écrire une fonction Python `listAdj(S,p)` qui génère la liste des listes d'adjacence d'un tel graphe aléatoire ayant S pour liste de sommets.

Le graphe dessiné correspond à la liste des listes d'adjacence suivante, $[[\text{'c'}, \text{'e'}], [\text{'d'}, \text{'e'}], [\text{'a'}, \text{'e'}], [\text{'b'}, \text{'e'}], [\text{'a'}, \text{'b'}, \text{'c'}, \text{'d'}]]$, avec la liste des sommets qui est $S = \text{'abcde'}$.

3. Pour tout $k \in \llbracket 1, n \rrbracket$, on considère la variable aléatoire D_k égale au degré du sommet s_k . Déterminer la loi de D_k .
4. On dit que s_k est isolé si $D_k = 0$ est réalisé. On note Z_n la variable aléatoire comptant les sommets isolés de G et X_k la variable aléatoire de Bernoulli qui vaut 1 si s_k est isolé et 0 sinon.

a) Montrer que : $Z_n = \sum_{k=1}^n X_k$. En déduire que $\mathbb{E}(Z_n) = n(1-p)^{n-1}$.

b) Montrer que : $Z_n^2 = \sum_{k=1}^n X_k + 2 \sum_{1 \leq i < j \leq n} X_i X_j$.

c) Justifier que $\mathbb{P}([X_i = 1] \cap [X_j = 1]) = (1-p)^{2n-3}$. En déduire que $\mathbb{E}(Z_n^2) = n(1-p)^{n-1} + n(n-1)(1-p)^{2n-3}$.

On suppose désormais que $p = p_n = c \frac{\ln(n)}{n}$, avec $c > 0$, $c \neq 1$.

5.a) Écrire une fonction Python `Z` qui renvoie le nombre de sommets isolés d'un graphe donné par sa liste de listes d'adjacence `lst`. On importera les bibliothèques `numpy` as `np` et `numpy.random` as `rd`.

b) On souhaite estimer l'influence de la valeur de c sur le nombre de sommets isolés. En exécutant le script suivant

```
list2c=[0.3,0.5,0.7,1.3,1.5,1.7];n=1000;res=[]
```

```
for c in list2c:
    s=0
    for k in range(200):
        if Z(listAdj(range(1,n+1),c*np.log(n)/n))==0:
            s+=1
    res.append(s/200)
print(res)
```

on obtient `[0.0, 0.0, 0.0, 0.91, 0.975, 0.99]` après de longues minutes.

Quelle conjecture sur la valeur d'une probabilité pouvez-vous faire lorsque $c < 1$ et $c > 1$? Justifier.

6.a) Montrer que $(1 - p_n)^{n-1} \underset{n \rightarrow +\infty}{\sim} (1 - p_n)^n$ puis que $(1 - p_n)^{n-1} \underset{n \rightarrow +\infty}{\sim} n^{-c}$.

b) On rappelle que l'inégalité de Markov affirme que si X est une variable aléatoire positive admettant une espérance et $a > 0$,

$$\mathbb{P}([X \geq a]) \leq \frac{\mathbb{E}(X)}{a}$$

Si $c > 1$, en déduire la limite de $\mathbb{P}([Z_n \geq 1])$ puis de $\mathbb{P}([Z_n = 0])$, lorsque $n \rightarrow +\infty$.

c) En utilisant l'inégalité de Bienaymé-Tchébichev, montrer que $\mathbb{P}([Z_n = 0]) \leq \frac{\mathbb{V}(Z_n)}{\mathbb{E}(Z_n)^2}$. En déduire que si $c < 1$, $\mathbb{P}([Z_n = 0]) \rightarrow 0$ quand $n \rightarrow +\infty$.

d) Votre conjecture est-elle correcte?

Solution

1. Le nombre maximal d'arêtes correspond au nombre de paires de sommets soit $\binom{n}{2} = \frac{n(n-1)}{2}$.

2. Un exemple de fonction qui convient :

```
def listAdj(S,p):
    l=[[[]for k in range(len(S))]
    for i in range(len(S)-1):
        for j in range(i+1,n):
            if rd.random()<p:
                l[i].append(S[j])
                l[j].append(S[i])
    return l
```

3. On a $D_k = \sum_{i=1}^{k-1} T_{i,k} + \sum_{i=k+1}^n T_{k,i}$. D_k est la somme de $(n-1)$ variables de Bernoulli indépendantes de paramètre p d'où D_k suit la loi binomiale de paramètres $n-1$ et p .

4.a) Classique. Or pour tout k , $\mathbb{E}(X_k) = \mathbb{P}([X_k = 1]) = \mathbb{P}([D_k = 0])$ qui vaut $(1-p)^{n-1}$ d'après la loi de D_k , d'où par linéarité de l'espérance, $\mathbb{E}(Z_n) = n(1-p)^{n-1}$.

b) On a :

$$Z_n^2 = \left(\sum_{k=1}^n X_k \right)^2 = \sum_{(i,j) \in [1,n]^2} X_i X_j = \sum_{1 \leq i=j \leq n} X_i X_j + \sum_{1 \leq i < j \leq n} X_i X_j + \sum_{1 \leq j < i \leq n} X_i X_j$$

Or comme toute variable de Bernoulli $X_i^2 = X_i$ et on a aussi $\sum_{1 \leq j < i \leq n} X_i X_j = \sum_{1 \leq i < j \leq n} X_i X_j$ d'où on obtient bien que :

$$Z_n^2 = \sum_{k=1}^n X_k + 2 \sum_{1 \leq i < j \leq n} X_i X_j$$

c) On a pour $i < j$, $[X_i = 1] \cap [X_j = 1]$ est réalisé ssi les événements $[T_{k,i} = 0]$ pour $k < i$, $[T_{i,k} = 0]$ pour $k > i$, $[T_{k,j} = 0]$ pour $k < j$ et $k \neq i$, $[T_{j,k} = 0]$ pour $k > j$ sont réalisés.

Or ces événements sont au nombre de $(n-1) + (n-2) = 2n-3$, indépendants et de probabilité p . D'où $\mathbb{P}([X_i = 1] \cap [X_j = 1]) = (1-p)^{2n-3}$.

De plus par linéarité de l'espérance :

$$\mathbb{E}(Z_n^2) = n(1-p)^{n-1} + 2 \sum_{1 \leq i < j \leq n} \mathbb{E}(X_i X_j)$$

Or $\mathbb{E}(X_i X_j) = \mathbb{P}([X_i X_j = 1]) = \mathbb{P}([X_i = 1] \cap [X_j = 1]) = (1-p)^{2n-3}$.

et il y a $\frac{n(n-1)}{2}$ couples (i, j) tels que $1 \leq i < j \leq n$, alors on a bien :

$$\mathbb{E}(Z_n^2) = n(1-p)^{n-1} + n(n-1)(1-p)^{2n-3}$$

5.a) def Z(lst):

```
isolés=0
for k in range(len(lst)):
    isolés+=(lst[k]==[])
return isolés
```

b) Ce script utilise la loi faible des grands nombres qui permet d'affirmer que, la fréquence empirique de réalisation d'un événement, ici $[Z_n = 0]$, lors d'une répétition d'un grand nombre d'expériences identiques liées à la réalisation de cet événement et indépendantes, est proche de sa probabilité. On peut conjecturer que, lorsque n est grand, $\mathbb{P}([Z_n = 0])$ est proche de 1 pour $c > 1$ et de 0 pour $c < 1$.

6.a) On remarque que $(1-p_n)^{n-1} \sim (1-p_n)^n$ car $1-p_n \rightarrow 1$ quand $n \rightarrow +\infty$.

De plus $(1-p_n)^n = \exp\left(n \ln\left(1 - c \frac{\ln(n)}{n}\right)\right)$. Or, $n \ln\left(1 - c \frac{\ln(n)}{n}\right) = n\left(-c \frac{\ln(n)}{n} - c^2 \frac{\ln(n)^2}{2n^2} + o\left(\frac{\ln(n)^2}{n^2}\right)\right)$,

d'où $n \ln\left(1 - c \frac{\ln(n)}{n}\right) = -c \ln(n) + o(1)$ donc $(1-p_n)^n \sim n^{-c}$.

b) Si $c > 1$, $\mathbb{E}(Z_n) \rightarrow 0$ car $\mathbb{E}(Z_n) \sim n^{1-c}$, quand $n \rightarrow +\infty$. D'après Markov, $\mathbb{P}([Z_n \geq 1]) \leq \mathbb{E}(Z_n)$, d'où $\mathbb{P}([Z_n \geq 1]) \rightarrow 0$ quand $n \rightarrow +\infty$ et ainsi $\mathbb{P}([Z_n = 0]) \rightarrow 1$ quand $n \rightarrow +\infty$.

c) Appliquons l'inégalité de BT à Z_n pour $\varepsilon = \mathbb{E}(Z_n)$:

$$\mathbb{P}(|Z_n - \mathbb{E}(Z_n)| \geq \mathbb{E}(Z_n)) \leq \frac{\mathbb{V}(Z_n)}{\mathbb{E}(Z_n)^2}$$

Or si $Z_n = 0$ alors $|Z_n - \mathbb{E}(Z_n)| \geq \mathbb{E}(Z_n)$, donc $\mathbb{P}([Z_n = 0]) \leq \mathbb{P}(|Z_n - \mathbb{E}(Z_n)| \geq \mathbb{E}(Z_n))$ et par transitivité, $\mathbb{P}([Z_n = 0]) \leq \frac{\mathbb{V}(Z_n)}{\mathbb{E}(Z_n)^2}$.

D'après les questions précédentes, $\frac{\mathbb{V}(Z_n)}{\mathbb{E}(Z_n)^2} = \frac{1}{n(1-p_n)^{n-1}} + \frac{n-1}{n(1-p_n)} - 1$.

Or $\frac{n-1}{n(1-p_n)} \rightarrow 1$ quand $n \rightarrow +\infty$ et si $c < 1$, $\frac{1}{n(1-p_n)^{n-1}} \rightarrow 0$, d'où $\mathbb{P}([Z_n = 0]) \rightarrow 0$ quand $n \rightarrow +\infty$.

d) On a confirmation de la conjecture.

N.B :

Les considérations qui suivent pourraient faire l'objet d'une première partie.

La définition d'un graphe aléatoire à N_n arêtes, sur un ensemble de n sommets, se fait a priori en choisissant N_n paires au hasard parmi $\binom{n}{2}$ paires de sommets.

Ainsi, si on considère les variables de Bernoulli $X_{i,j}$, leur paramètre vaut $\frac{N_n}{\binom{n}{2}}$.

Plus généralement pour $\{i_1, j_1\}, \dots, \{i_k, j_k\}$ k arêtes :

$$\mathbb{P}([X_{i_1, j_1} = 1] \cap \dots \cap [X_{i_k, j_k} = 1]) = \frac{N_n(N_n - 1) \dots (N_n - k + 1)}{\binom{n}{2} \left(\binom{n}{2} - 1 \right) \dots \left(\binom{n}{2} - k + 1 \right)}$$

Si $n \rightarrow +\infty$ et $N_n \rightarrow +\infty$ alors

$$\mathbb{P}([X_{i_1, j_1} = 1] \cap \dots \cap [X_{i_k, j_k} = 1]) \sim \left(\frac{N_n}{\binom{n}{2}} \right)^k = \mathbb{P}([X_{i_1, j_1} = 1]) \times \dots \times \mathbb{P}([X_{i_k, j_k} = 1])$$

On peut donc lorsque n grand, dans ces conditions, supposer que les $X_{i,j}$ sont indépendantes de loi de Bernoulli de paramètre $p_n = \frac{N_n}{\binom{n}{2}}$. On retrouve ainsi les conditions de l'exercice dans lequel $N_n \sim \frac{c}{2} n \ln(n)$.

ENONCÉ 2

Loi de Pareto-Zipf

On souhaite modéliser la loi de probabilité de la variable aléatoire T qui à une ville, choisie au hasard parmi les villes françaises, associe l'effectif de sa population. On note n le nombre de villes.

1.a) Pour 2018, on dispose d'un fichier Data1.csv de données provenant de l'INSEE et on utilise la bibliothèque Pandas. On exécute les instructions suivantes :

```
import pandas as pd; import numpy.random as rd; import numpy as np; import matplotlib.pyplot as plt

dataset=pd.read_csv('Data1.csv',sep=";")

données=dataset[["Libellé","Population" ]]

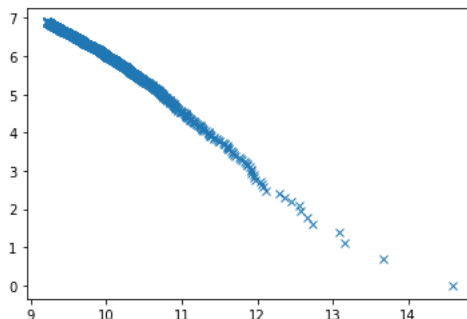
données=données.sort_values(by=["Population","Libellé"], ascending=False, ignore_index=True)

logPopu=[np.log(t) for t in données["Population"]]

logRang=[np.log(k) for k in range (1,len(logPopu)+1)]

plt.plot(logPopu,logRang,'*')
```

ce qui produit le graphique suivant :



Expliquer pourquoi le programme et le graphique précédent justifient qu'il existe deux réels a et b tels que, pour toute ville, le réel $\frac{a}{t^b}$, où t est l'effectif de celle-ci, est une approximation raisonnable du rang de celle-ci dans la liste des villes classées par ordre décroissant de population ?

Quelle grandeur peut-on calculer pour confirmer ce que l'on constate graphiquement ? Quelle méthode peut-on utiliser pour obtenir le meilleur couple (a, b) en un certain sens ?

b) On suppose que l'on a pour toutes les villes, $r = \frac{a}{t^b}$, r étant le rang de cette ville, t son effectif, a et b deux réels identiques pour toutes les villes. Si x est l'effectif d'une des villes françaises, quelle est, en fonction de a , b , x et n , la proportion de villes dont la population urbaine est supérieure ou égale à x ?

• On suppose désormais que T suit la loi de Pareto de paramètre $\theta > 1$ et $x_0 > 0$, c'est à dire qu'elle admet pour densité la fonction f définie par :

$$f(x) = \begin{cases} \theta \frac{x_0^\theta}{x^{\theta+1}} & \text{si } x \geq x_0 \\ 0 & \text{sinon.} \end{cases}$$

On note $\text{Par}(\theta, x_0)$ cette loi.

2.a) Déterminer la fonction de répartition F de T .

- b) En calculant $\mathbb{P}([T > x])$, montrer que ce résultat est cohérent avec le résultat de la question 1 et exprimer x_0 et θ en fonction de a , b et n .
- c) Montrer que $\mathbb{E}(T)$ existe et vaut $\frac{\theta}{\theta-1}x_0$.
3. Soit $x \geq x_0$, on note \mathbb{P}_x la probabilité conditionnelle sachant $[T > x]$.
- a) Montrer que pour tout $t \geq x \geq x_0$, $\mathbb{P}_x(T > t) = \left(\frac{x}{t}\right)^\theta$.
- b) On pose pour tout $t \in \mathbb{R}$, $F_x(t) = \mathbb{P}_x([T \leq t])$. De quelle loi F_x est-elle la fonction de répartition? Quelle est alors l'espérance de cette loi?
4. Soit $\delta \in]1, +\infty[$.
- On suppose que Y est une variable aléatoire à valeurs dans $[x_0, +\infty[$ dont f_Y est une densité, continue sur $[x_0, +\infty[$, F_Y la fonction de répartition. On suppose que $\forall x \geq x_0$, $\mathbb{P}([Y \geq x]) > 0$.
- a) Soit $x \geq x_0$. Montrer que la fonction G_x définie sur \mathbb{R} par $G_x : t \mapsto \mathbb{P}_{[Y \geq x]}(Y \leq t)$ est la fonction de répartition d'une variable aléatoire à densité dont on déterminera une densité.
- On suppose que pour tout $x \geq x_0$, l'espérance d'une variable aléatoire de fonction de répartition G_x existe et vaut δx .
- b) En déduire que pour tout $x \geq x_0$, $(\delta - 1)xf_Y(x) = \delta(1 - F_Y(x))$.
- c) Résoudre l'équation différentielle $(1 - \delta)xy' - \delta y = 0$ sur $[x_0, +\infty[$.
- d) En conclure que Y suit une loi de Pareto dont on précisera les paramètres.

Solution

1.a) Cette relation peut aussi s'écrire $\ln(r) = \ln(a) - b \ln(t)$. On a représenté le nuage de points $(\ln(t_i), \ln(r_i))_{i \in [1, n]}$ on constate un très bon alignement des points, donc on peut considérer qu'une telle relation est raisonnable.

On peut par ailleurs calculer le coefficient de corrélation linéaire entre les $\ln(t_i)$ et les $\ln(r_i)$. S'il est proche de 1 en valeur absolue, on peut alors valider la relation.

On trouve $-0,9947$ ce qui confirme l'utilisation de ce modèle.

Pour obtenir (a, b) on peut réaliser un ajustement aux moindres carrés du nuage de points $(\ln(t_i), \ln(r_i))$.

b) Notons $r(x)$ le rang de la ville d'effectif x , cette proportion vaut $\frac{r(x)}{n} = \frac{a}{nx^b}$.

2.a) Si $x < x_0$, alors $F(x) = 0$. Si $x \geq x_0$, $F(x) = \int_{x_0}^x \theta \frac{x_0^\theta}{t^{\theta+1}} dt = x_0^\theta \left[-\frac{1}{t^\theta} \right]_{x_0}^x = 1 - \left(\frac{x_0}{x}\right)^\theta$.

b) On a donc que si x est l'effectif d'une ville, $\mathbb{P}([T > x]) = \frac{x_0^\theta}{x^\theta}$. Si l'on rapproche ce résultat du résultat de la question précédente on peut alors considérer que cette probabilité correspond à la valeur trouvée dans cette question soit $\frac{a}{nx^b}$. Ces deux résultats sont donc cohérents en identifiant θ à b et x_0^θ à $\frac{a}{n}$ donc x_0 à $\left(\frac{a}{n}\right)^{\frac{1}{\theta}}$.

c) On s'intéresse à la convergence de $\int_{-\infty}^{+\infty} tf(t)dt$ i.e. de $\int_{x_0}^{+\infty} \theta \frac{x_0^\theta}{t^{\theta+1}} dt$. La fonction $H : t \mapsto \frac{\theta}{-\theta+1} \frac{x_0^\theta}{t^{\theta-1}}$ est une primitive de la fonction dans l'intégrale qui admet une limite finie égale à 0 en $+\infty$. Donc cette intégrale est convergente et vaut $-H(x_0) = \frac{\theta}{\theta-1}x_0$.

3.a) Par définition, $\mathbb{P}_x(T > t) = \frac{\mathbb{P}(T > t) \cap [T > x]}{\mathbb{P}([T > x])} = \frac{1 - F(t)}{1 - F(x)}$ puisque $t \geq x$. Finalement $\mathbb{P}_x(T > t) = \frac{\left(\frac{x_0}{t}\right)^\theta}{\left(\frac{x_0}{x}\right)^\theta} = \left(\frac{x}{t}\right)^\theta$.

b) D'après la question précédente, $F_x(t) = \begin{cases} 0 & \text{si } t \leq x \\ 1 - \left(\frac{x}{t}\right)^\theta & \text{sinon.} \end{cases}$ On reconnaît la fonction de répartition de la loi

$\text{Par}(\theta, x)$. D'où l'espérance de cette loi existe et vaut $\frac{\theta}{\theta-1}x$.

4.a) En reprenant la méthode utilisée dans le 4.a), on obtient que :

$$G_x(t) = \begin{cases} \frac{F_Y(t) - F_Y(x)}{1 - F_Y(x)} & \text{si } t > x \\ 0 & \text{sinon.} \end{cases}$$

On vérifie que G_x est bien la fonction de répartition d'une variable à densité. Elle est de classe C^1 sur $\mathbb{R} \setminus \{x\}$ et

$$G'_x(t) = \begin{cases} \frac{f_Y(t)}{1 - F_Y(x)} & \text{si } t > x \\ 0 & \text{si } t < x. \end{cases}$$

d'où une densité...

b) On peut donc en déduire que pour tout $x \geq x_0$, $\int_x^{+\infty} t f_Y(t) dt$ converge et vérifie : $\int_x^{+\infty} t f_Y(t) dt = \delta x (1 - F_Y(x))$.

En utilisant une primitive, on voit que la fonction $K : x \mapsto \int_x^{+\infty} t f_Y(t) dt$ est de classe C^1 sur $[x_0, +\infty[$ et vérifie $K'(x) = -x f_Y(x)$. Donc, pour tout $x \geq x_0$, $-x f_Y(x) = \delta(1 - F_Y(x)) - x f_Y(x)$ i.e $(\delta - 1)x f_Y(x) = \delta(1 - F_Y(x))$.

c) Sur $[x_0, +\infty[$, $(1 - \delta)xy' - \delta y = 0 \iff y' = \frac{\delta}{(1 - \delta)x}y$. Posons $\beta = \frac{\delta}{\delta - 1}$. Les solutions sont de la forme $y = \lambda \exp(-\beta(\ln(x))) = \frac{\lambda}{x^\beta}$ avec λ réel.

d) Si l'on pose pour tout $x \geq x_0$, $\varphi(x) = 1 - F_Y(x)$, φ vérifie l'équation différentielle précédente et $\varphi(x_0) = 1$ d'où $\lambda = x_0^\beta$ et pour tout $x \geq x_0$, $F_Y(x) = \left(\frac{x_0}{x}\right)^\beta$. On en conclut que Y suit la loi Par $\left(\frac{\delta}{\delta - 1}, x_0\right)$.