ENSAI École nationale de la statistique et de l'analyse de l'information

Insee Mesurer pour comprendre

INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES

ÉCOLE NATIONALE DE LA STATISTIQUE ET DE L'ANALYSE DE L'INFORMATION

Concours interne d'attaché statisticien de l'Insee

AVRIL 2018

ÉPREUVE DE MATHÉMATIQUES ET STATISTIQUES

Durée 4 heures

Coefficient 3

Sans documents – L'usage de la calculatrice est interdit

Le sujet comprend 8 pages (y compris celle-ci)

Ce sujet se compose de 3 exercices et un problème.

Exercice 1

On note $\mathcal{C} = (e_1, e_2, e_3)$ la base canonique de $\mathbb{R}^3$, $M_3(\mathbb{R})$ l'espace des matrices carrées de taille 3. Soit f l'application linéaire de $\mathbb{R}^3$ dans lui-même dont la matrice dans la base $\mathcal{C}$ est

$$A = \begin{pmatrix} 0 & 1 & -2 \\ 5 & 13 & -21 \\ 3 & 8 & -13 \end{pmatrix}.$$

1. Calculer le noyau $\ker(f)$ de f. On note u le vecteur de $\ker(f)$ dont la 1ère coordonnée dans la base $\mathcal{C}$ est -1.
2. La matrice A est-elle inversible ? Quelle est la dimension de l'image de A ?
3. Déterminer le vecteur v de $\mathbb{R}^3$ tel que $f(v) = u$ et tel que la 1ère coordonnée de v dans la base $\mathcal{C}$ soit 2.
4. Soit $w = \begin{pmatrix} 0 \\ -3 \\ -2 \end{pmatrix}$. Montrer que $\mathcal{B} = (u, v, w)$ est une base de $\mathbb{R}^3$.
5. Calculer les coordonnées de $f(w)$ dans la base $\mathcal{B}$. En déduire la matrice N de f dans la base $\mathcal{B}$.
6. (a) On note P la matrice dont les colonnes sont u, v, w. Calculer P^{-1}.
   (b) Calculer PNP^{-1}.
   (c) Quelles sont les valeurs propres de A ? La matrice est-elle trigonalisable ? Diagonalisable ?
   (d) Calculer A^k pour $k \in \mathbb{N}$.

Exercice 2

On dispose de deux urnes $\mathcal{A}$ et $\mathcal{B}$, d'un dé à six faces équilibré, et de six boules numérotées de 1 à 6. On place au début la boule 1 dans l'urne $\mathcal{A}$ et les autres boules dans l'urne $\mathcal{B}$. On répète l'expérience suivante : on jette le dé, puis on change d'urne la boule dont le numéro est celui obtenu par le jet de dé.
Pour $n \in \mathbb{N}^*$ on note D_n la variable aléatoire représentant la valeur obtenue au n-ième lancer du dé, A_n le nombre de boules dans l'urne $\mathcal{A}$ après le n-ième jet de dé (après qu'on ait effectué le n-ième échange).

1. Calculer la loi de A_1, puis son espérance $\mathbb{E}(A_1)$.
2. On souhaite déterminer la loi du couple (A_1, A_2).
   (a) Déterminer les valeurs possibles pour le couple (A_1, A_2).
   (b) Pour chaque valeur possible (n_1, n_2), écrire l'évènement $\{(A_1, A_2) = (n_1, n_2)\}$ à l'aide d'évènements faisant intervenir les variables aléatoires D_n.
   (c) En déduire la loi du couple (A_1, A_2) puis celle de A_2.
   (d) Les variables aléatoires A_1 et A_2 sont-elles indépendantes ?
3. Soit $n \geq 1$. On souhaite relier $\mathbb{E}(A_{n+1})$ et $\mathbb{E}(A_n)$.
   (a) Relier $P(A_{n+1} = 0)$ et $P(A_n = 1)$.
   (b) Relier $P(A_{n+1} = 6)$ et $P(A_n = 5)$.
   (c) On fixe k entier entre 1 et 5. Relier $P(A_{n+1} = k)$ aux probabilités $P(A_n = k-1)$, $P(A_n = k+1)$.
   (d) Montrer que $\mathbb{E}(A_{n+1}) - \frac{2}{3}\mathbb{E}(A_n)$ est constant (indépendant de n).
4. Montrer que la suite $(\mathbb{E}(A_n)-3)_{n\in\mathbb{N}^*}$ est géométrique. La suite $(\mathbb{E}(A_n))_{n\in\mathbb{N}^*}$ converge-t-elle ?

Exercice 3

Questions préliminaires :

1. Montrer que la matrice symétrique $\begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}$ n'est ni positive ni négative, c'est à dire montrer qu'il existe $(h,k) \in \mathbb{R}^2$ tel que

$$(h \quad k) \cdot \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix} \cdot \begin{pmatrix} h \\ k \end{pmatrix} < 0$$

et qu'il existe $(h',k') \in \mathbb{R}^2$ tel que

$$(h' \quad k') \cdot \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix} \cdot \begin{pmatrix} h' \\ k' \end{pmatrix} > 0.$$

2. Pour $(h,k) \in \mathbb{R}^2$, écrire $(h \quad k) \cdot \begin{pmatrix} 3 & -1 \\ -1 & 3 \end{pmatrix} \cdot \begin{pmatrix} h \\ k \end{pmatrix}$ comme une somme de carrés. En déduire que la matrice symétrique $\begin{pmatrix} 3 & -1 \\ -1 & 3 \end{pmatrix}$ est positive.

On considère la fonction $f : \mathbb{R}^2 \to \mathbb{R}$ définie par $f(x,y) = x^4 + y^4 - 4xy$.
On note $\nabla_f(a) = \left(\frac{\partial f}{\partial x}(a), \frac{\partial f}{\partial y}(a)\right)$ le gradient de f en un point $a \in U = \mathbb{R}^2$ et $H_f(a) = \begin{pmatrix} \frac{\partial^2 f}{\partial x^2}(a) & \frac{\partial^2 f}{\partial x \partial y}(a) \\ \frac{\partial^2 f}{\partial y \partial x}(a) & \frac{\partial^2 f}{\partial y^2}(a) \end{pmatrix}$ la Hessienne de f en a.

1. (a) Calculer le gradient $\nabla_f(a)$ pour $a \in U$.
   (b) Calculer les points critiques de f, c'est à dire les points en lequels le gradient de f s'annule.
   (c) Calculer la Hessienne de f sur U.
   (d) Etudier les extrema locaux de f.
2. On étudie maintenant les extrema éventuels de f sur $V_g = \{a \in U; g(a) = 0\}$ avec $g(x,y) = xy+1$. On suppose que $a \in U$ est un extremum local de f sous la contrainte $g = 0$.
   (a) Justifier qu'il existe une constante $\lambda \in \mathbb{R}$ (appelée multiplicateur de Lagrange) telle que $\nabla_f(a) = \lambda.\nabla_g(a)$.
   (b) En déduire les deux possibilités pour a, notées a_1 et a_2.
   (c) On note G_{a_1} la fonction définie sur U par $G_{a_1}(x,y) = f(x,y) - f(a_1)$.
      i. Montrer que $x^4.G_{a_1}(x,y) \geq 0$ pour $(x,y) \in V_g$. En déduire le signe de $G_{a_1}(x,y)$ pour $(x,y) \in V_g$.
      ii. Le point a_1 est-il un extremum global de f sur V_g ? Si oui est-ce un minimum, un maximum ?
      iii. Faire l'étude en a_2.

Problème

En 1897, l'économiste italien Vilfredo Pareto (1848-1923), professeur d'économie politique à l'université de Lausanne, observa que 20% de la population italienne possèdait 80% de la richesse nationale, d'où le nom de loi 80-20 ou 20-80. Plus précisement, on observa que la répartition cumulative des revenus sur un graphique log-log, est approximativement une droite. De ce constat,une modélisation de la loi des revenus, baptisée « loi de Pareto » par Joseph Juran (économiste américain d'origine roumaine 1904-2008) fut adoptée et peut être mathématiquement définie comme suit :
une variable aléatoire X, absolument continue, suit une loi de Pareto de paramètres $\alpha, x_0 \in \mathbb{R}_+^*$ si sa densité est donnée par :

$$\begin{cases} f(x) = \dfrac{k}{x^{\alpha+1}} & \text{si } x \in [x_0; +\infty[\\ f(x) = 0 & \text{sinon} \end{cases}$$

où k est défini de manière à ce que f soit une densité. On écrit alors que X suit $\mathcal{LP}(\alpha, x_0)$.

Dans tout le problème α et x_0 désigne deux nombres réels strictement positifs.

Partie A: Quelques propriétés de la loi de Pareto

Dans toute cette partie, X est une variable aléatoire suivant $\mathcal{LP}(\alpha, x_0)$.

1. (a) Calculer $\lim\limits_{x\to+\infty} \int_{x_0}^{x} f(t)\,\mathrm{d}t$.
   En déduire k pour que f soit une densité de probabilité.
   (b) Déterminer la fonction de répartition F de X.
   On rappelle que $F(x) = P(X \leq x) = \int_{x_0}^{x} f(t)\,\mathrm{d}t$ pour $x \in \mathbb{R}$.
   (c) Calculer, pour $x \geq x_0$, $\ln(1 - F(x))$ en fonction de $\ln(x)$.
   On rappelle que ln désigne la fonction logarithme népérien.
   (d) Déterminer la médiane de la distribution.
   On appelle médiane la valeur qui partage la distribution en deux parties égales.
2. (a) Calculer $M_x = \int_{x_0}^{x} tf(t)\,\mathrm{d}t$ pour $x \geq x_0$. Discuter selon les valeurs de α de la valeur de $\lim_{x\to+\infty} M_x$.
   (b) On note M_α la limite de M_x lorsque celle-ci est finie. Que représente cette valeur pour la variable aléatoire X ?
3. On note $Y = \ln\left(\dfrac{X}{x_0}\right)$. Déterminer la fonction de répartition de Y, puis la densité de Y.
   Reconnaître la loi de Y.
4. **Loi de la queue de distribution**
   Soit $x_1 \in \mathbb{R}$ tel que $x_1 > x_0 > 0$.
   (a) Calculer $P(X > z | X > x_1)$ pour $z \geq x_1$.
   (b) On note $H(z) = 1 - P(X > z | X > x_1)$ pour $z \geq x_1$ et $H(z) = 0$ pour $z < x_1$.
   Montrer que H est une fonction de répartition.
   (c) Reconnaître la loi correspondante à cette fonction de répartition.

Partie B: Une mesure des inégalités

La courbe de concentration est une courbe statistique introduite par Lorenz (économiste américain 1876-1959) et développée par Gini (statisticien italien 1884-1965) pour rendre compte de l'inégalité de la distribution des revenus.

Dans toute la suite, X est une variable aléatoire qui représente le revenu d'un individu de cette population.

On suppose que X suit $\mathcal{LP}(\alpha, x_0)$ avec $\alpha > 1$.

On note F la fonction de répartition de X, $E(X)$ son espérance et on pose

$$Q(x) = \frac{1}{E(X)} \int_{x_0}^{x} t\, f(t)\, \mathrm{d}t \text{ pour } x \geq x_0.$$

Ainsi $Q(x)$ représente le quotient de la masse des revenus des individus ayant un salaire inférieur ou égal à x par la masse totale des revenus de la population.

1. Calculer $Q(x)$ pour $x \geq x_0$.
2. On note $F(x)$ la fonction de répartition de X restreinte à $[x_0, +\infty[$. On rappelle que $F(x) = \int_{x_0}^{x} f(t)\, \mathrm{d}t$.
   Montrer que F établit une bijection de $[x_0, +\infty[$ dans $[0, 1[$.
   On note F^{-1} l'application réciproque de F de $[0, 1[$ dans $[x_0, +\infty[$.
   On note $C = Q \circ F^{-1}$ et on prolonge en 1 avec $C(1) = 1$. La courbe représentative de C est appelée courbe de concentration de X. Ainsi, elle donne $Q(x)$ en fonction de $F(x)$.
3. On pose $D(t) = 1 - (1-t)^{\frac{\alpha-1}{\alpha}}$ pour $t \in [0, 1[$ et $D(1) = 1$.
   Vérifier que $(D \circ F)(x) = Q(x)$ pour $x \geq x_0$.
   En déduire que $C = D$.
4. Pour quelle valeur de α, 20% des salariés concentre 80% de la masse salariale ?
   (On ne demande pas de calculer une valeur décimale de α).
5. On appelle indice d'inégalité de Gini de la variable X le réel $I(X)$ qui est égal à deux fois l'aire située entre la courbe de concentration de X et la première bissectrice.
   C'est à dire : $I(X) = 2 \int_0^1 (t - C(t))\, dt$.
   On estime que plus $I(X)$ est grand, plus l'inégalité des revenus est grande.
   Montrer que $I(X) = \dfrac{1}{2\alpha - 1}$.

Partie C: La distribution des revenus français

Les données fiscales françaises fournissent une répartition des salariés selon leur rémunération rapportée en SMIC horaire. On compte 913 milliers de salariés ayant des salaires horaires au-delà de 4 SMIC.

x représente le revenu mesuré en SMIC horaire ;

$N(x)$ est le nombre de milliers de salariés ayant un revenu de x.

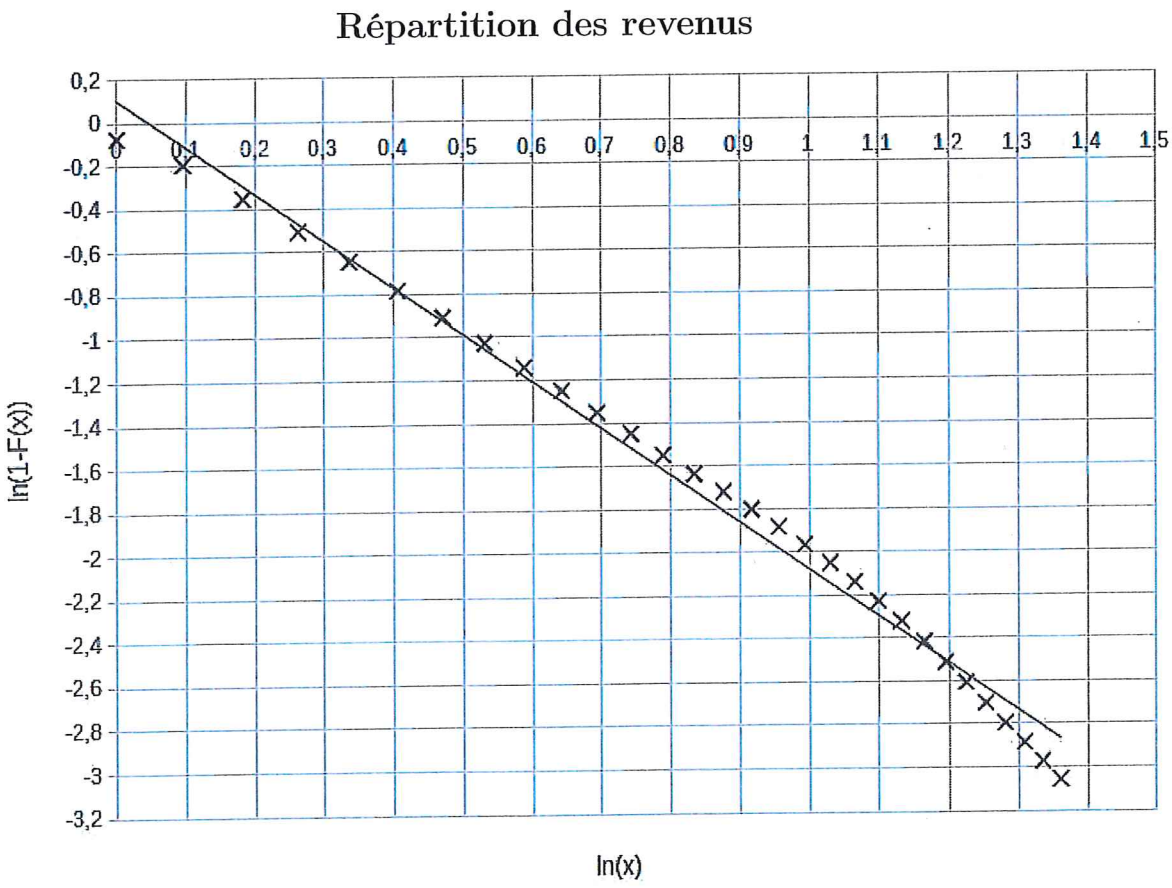
x	N(x)	x	N(x)	x	N(x)
1	1 381	2	543	3	197
1,1	2 072	2,1	466	3,1	186
1,2	2 334	2,2	437	3,2	172
1,3	1 913	2,3	342	3,3	158
1,4	1 544	2,4	302	3,4	142
1,5	1 262	2,5	271	3,5	129
1,6	1 055	2,6	252	3,6	115
1,7	881	2,7	233	3,7	102
1,8	745	2,8	220	3,8	91
1,9	638	2,9	208	3,9	81

Note de lecture : *2 334 salariés ont une rémunération horaire comprise entre 1,2 et 1,3 SMIC.*

Ces données ont été représentées sur le graphique intitulé « Répartition des revenus » de l'énoncé. Chacune des données a été représentée par une croix. En abscisse, on a reporté $\ln(x)$ où x est le revenu en SMIC horaire. En ordonnées, on a reporté $\ln(1 - F(x))$ où F est une estimation de la fonction de répartition.

Une étude statistique sur ces données a permis de calculer un ajustement linéaire par la méthode des moindres carrés de $\ln(1 - F(x))$ sur $\ln(x)$. Cette droite a été représentée sur le graphique.

Par ailleurs, l'étude a également fournit les éléments suivants : la moyenne est de 1,60 et la médiane s'élève à 1,44. Pour les revenus supérieurs à 4 SMIC, l'étude a réalisé l'estimation en prenant un revenu égal à 4 SMIC.



Dans cette partie, on discute des méthodes utilisées pour fournir ces indicateurs statistiques et on propose une autre méthode pour calculer la moyenne des revenus.

1. Discussion des grandeurs calculées

(a) Expliquer pourquoi la valeur proposée pour la moyenne est biaisée.

(b) L'estimation de la médiane possède-t-elle le même biais ?

2. **Estimations basées sur la droite de régression linéaire**

On suppose que les observations enregistrées précédemment suivent une loi de Pareto de paramètres $\mathcal{LP}(\alpha, x_0)$. On suppose également que personne n'a un salaire horaire inférieur au SMIC.

Dans cette question, on peut s'appuyer sur les données fournies dans le tableau ou par le graphique pour établir les raisonnements.

(a) Donner la valeur de x_0.

(b) Par lecture graphique, donner le coefficient directeur de la droite représentée (on demande une valeur approximative).

(c) En exploitant le graphique, discuter de la modélisation d'une distribution des revenus par une loi de Pareto.

(d) Donner une valeur approchée de la valeur de α. On notera $\hat{\alpha}$ cette valeur.

3. **Proposition d'une meilleure estimation de la moyenne**

(a) En formulant clairement quelques hypothèses et en utilisant la question A.4, expliquer pourquoi on peut supposer que la répartition des revenus supérieurs à 4 SMIC est celle d'une loi de Pareto dont on donnera les paramètres.

(b) Proposer une estimation de la moyenne des revenus supérieurs à 4 SMIC et en donner une valeur numérique.