Exercice 1

Dans une ville, deux voyageurs, Albert et Béatrice, savent que les bus passent à intervalle de temps régulier. Ils souhaiteraient connaître cet intervalle de temps, mais ils n'observent que le temps d'attente du bus. Chacun propose sa méthode d'estimation.

Le premier voyageur, Albert, pense qu'en moyenne, il arrive à l'arrêt de bus à mi-temps entre deux bus. Il calcule la moyenne de ses observations et la double.

Comme Béatrice trouve que parfois elle attend vraiment longtemps, elle propose de prendre le temps maximal qu'elle a pu observer.

On note X_i le $i^{\text{ème}}$ temps d'attente à l'arrêt de bus. On suppose que X_i suit une loi uniforme sur $[0; a]$ où a est un réel strictement positif, a est donc le paramètre que les voyageurs souhaiteraient connaître.

Numéro de l'observation	1	2	3	4	5	6	7	8	9	10
temps d'attente en min	7	9	7	2	8	3	6	0	4	5
Numéro de l'observation	11	12	13	14	15	16	17	18	19	20
temps d'attente en min	3	6	8	11	9	5	8	2	7	10

Ils disposent d'un jeu de n observations indépendantes X_1, ... X_n identiquement distribuées selon la loi uniforme. On rappelle que la densité de cette loi est $f_a(x) = \frac{1}{a}$ pour $x \in [0, a]$ et $f_a(x) = 0$ pour $x \notin [0, a]$.

On rappelle que le biais d'un estimateur $\hat{\theta}$ cherchant à estimer la valeur du paramètre θ est : $\text{Biais}(\hat{\theta}) = E(\hat{\theta}) - \theta$.

1. On note $A_n = \frac{2}{n}(X_1 + X_2 + \ldots + X_n)$ l'estimateur d'Albert.
   (a) Calculer l'espérance et la variance de X_i.
   (b) Calculer l'espérance et la variance de l'estimateur A_n. Est-ce un estimateur sans biais de a (c'est-à-dire que le biais est nul) ?
2. Soit $B_n = \max(X_1, X_2, \ldots, X_n)$ l'estimateur de Béatrice.
   (a) Calculer la fonction de répartition de la variable aléatoire X_i. et en déduire la probabilité $P(B_n \leq t)$ pour $t > 0$.
   (b) Calculer la dérivée, notée f_{B_n}, de la fonction de répartition de B_n sur $]0, a[$.
   On admet que la fonction f_{B_n} calculée précédemment sur $]0, a[$ et prolongée par 0 en dehors de $]0, a[$ est la densité de B_n sur $\mathbb{R}$.
   (c) Calculer le biais de B_n. Pourriez-vous proposer à Béatrice l'expression d'un autre estimateur B_n^* sans biais ?
   (d) Calculer la variance de B_n^*.
3. Quel estimateur conseillez-vous d'utiliser ? Justifier votre choix.
4. On donne $\sum_{i=1}^{20} X_i = 120$ et $\sum_{i=1}^{20} X_i^2 = 886$.
   Donner une estimation de la valeur de a ainsi qu'une estimation de la variance pour chacun des deux estimateurs.
   Les valeurs ainsi calculées confirment-elles les arguments théoriques avancés précédement ?

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 : $I(X) = 2 \displaystyle\int_0^1 (t - C(t))\, dt$.
   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.

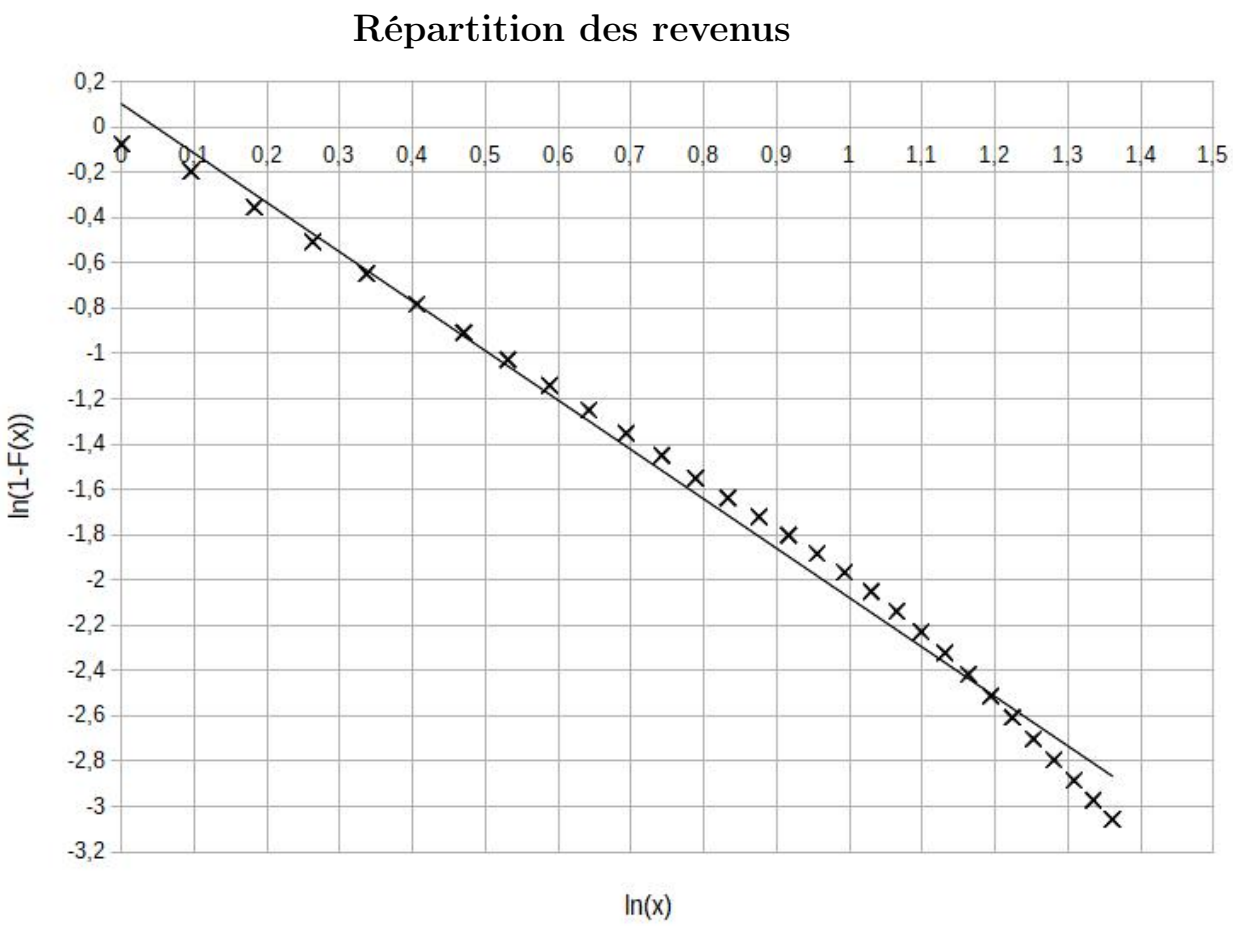
x	N(x)	x	N(x)	x	N(x)
1	1 381	2	543	3	197
1,1	2 072	2,1	466	3,1	186
1,2	2 334	2,2	437	3,2	172
1,3	1 913	2,3	342	3,3	158
1,4	1 544	2,4	302	3,4	142
1,5	1 262	2,5	271	3,5	129
1,6	1 055	2,6	252	3,6	115
1,7	881	2,7	233	3,7	102
1,8	745	2,8	220	3,8	91
1,9	638	2,9	208	3,9	81

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.