Préparation à l'agrégation externe de Sciences Sociales

Statistique descriptive

2022-2023

Exercice 1 (Question 2019)

1. Pour calculer la moyenne on utilise la formule :

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{k} n_i c_i,$$

où c_i est le centre de la classe. Alors,

$$\begin{aligned}\bar{x} &= \frac{1}{400}[300 \times 1250 + 55 \times 1750 + 35 \times 2750 + 10 \times 4000] \\ &= 1518,75\end{aligned}$$

2. Pour calculer l'écart type on utilise la formule :

$$\sigma_x = \sqrt{\frac{1}{N}\left(\sum_{i=1}^{k} n_i c_i^2\right) - \bar{x}^2},$$

où c_i est le centre de la classe. Alors,

$$\begin{aligned}\sigma_x &= \sqrt{\frac{1}{400}[300 \times 1250^2 + 55 \times 1750^2 + 35 \times 2750^2 + 10 \times 4000^2] - 1518,75^2} \\ &= 590\end{aligned}$$

Le coefficient de variation est égal à :

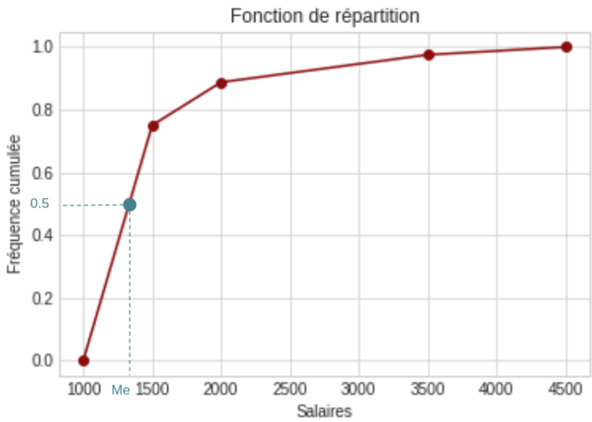
$$cv = \frac{\sigma_x}{\bar{x}} = \frac{590}{1518,75} = 0,388.$$

L'écart-type est de 38,8% de la moyenne.

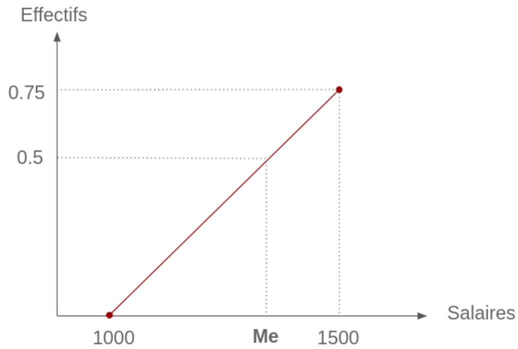
3. Vu qu'on ne dispose pas des données initiales, pour trouver une valeur ponctuelle de la médiane, on fait une interpolation linéaire dans la classe médiane, c'est-à-dire qu'on calcule par proportionnalité le point de cette classe pour lequel la fréquence cumulée est 0,5.

Salaires	[1000 ; 1500[	[1500 ; 2000 [	[2000 ; 3500[	[3500 ; 4500]
Effectifs	300	55	35	10
Fréquence cumulée	**0,75**	**0,8875**	**0,975**	**1**

La fonction de répartition est illustrée dans la figure suivante.



La classe médiane est la classe [1000 ; 1500[. On assume que la distribution des effectifs est uniforme entre les deux extrêmes de la classe, donc pour trouver la valeur de la médiane il faut construire la droite que passe par les points (1000 ; 0) et (1500 ; 0,75) et après l'utiliser pour trouver la valeur de x pour laquelle (x ; 0,5) appartient à cette droite. Cette situation est illustrée dans la figure ci-dessous :



L'équation de la droite qui passe par $(1000; 0)$ et $(1500; 0.75)$ est :

$$y = \frac{(x - 1000)}{500} 0.75.$$

En remplaçant y par $0,5$ on arrive à l'équation qu'on doit résoudre pour trouver la médiane :

$$\frac{0.5}{0.75} = \frac{Me - 1000}{500} \Longrightarrow Me = 1333,33.$$

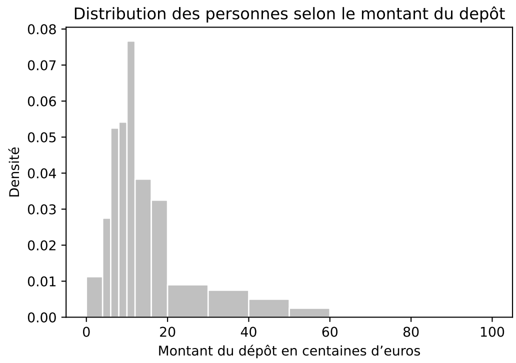
50% des effectifs ont un salaire inférieure ou égal à 1333,33 et 50% des effectifs ont un salaire supérieure ou égal à 1333,33.

Exercice 2 (2014)

1. On représente graphiquement cette distribution en utilisant un histogramme. Pour le construire, il faut trouver la fréquence (f_i) de chaque classe et sa densité ($d_i = f_i/a_i$) :

x_i	n_i	f_i	$d_i = f_i/a_i$
$[0;4[$	270	0.045	0.01125
$[4;6[$	330	0.055	0.0275
$[6;8[$	630	0.105	0.0525
$[8;10[$	650	0.108	0.054
$[10;12[$	920	0.153	0.0765
$[12;16[$	920	0.153	0.038
$[16;20[$	780	0.13	0.0325
$[20;30[$	540	0.09	0.009
$[30;40[$	450	0.075	0.0075
$[40;50[$	300	0.05	0.0075
$[50;60[$	150	0.025	0.0025
$[60;100[$	60	0.01	0.00025

L'histogramme qu'on obtient est le suivant :



La classe modale est la classe avec la plus grande densité, dans ce cas il s'agit de la classe $[10;12[$.

2. Pour calculer la moyenne on utilise la formule :

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{k} n_i c_i,$$

où c_i est le centre de la classe. Alors,

$$\begin{aligned}\bar{x} &= \frac{1}{6000}[2 \times 270 + 5 \times 330 + 7 \times 630 + \cdots + 55 \times 150 + 80 \times 60] \\ &= 17.55\end{aligned}$$

Pour calculer l'écart type on utilise la formule :

$$\sigma_x = \sqrt{\frac{1}{N}\left(\sum_{i=1}^{k} n_i c_i^2\right) - \bar{x}^2},$$

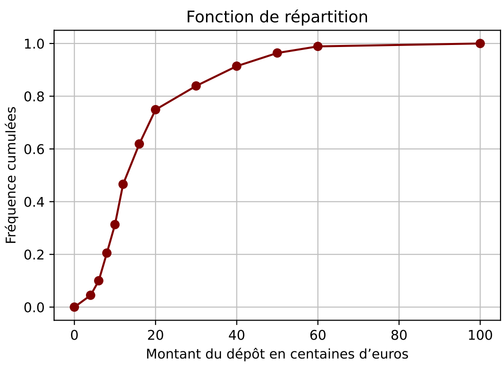
où c_i est le centre de la classe. Alors,

$$\begin{aligned}\sigma_x &= \sqrt{\frac{1}{6000}[2^2 \times 270 + 5^2 \times 330 + 7^2 \times 630 + \cdots + 55^2 \times 150 + 80^2 \times 60] - 17.55^2} \\ &= 13.7\end{aligned}$$

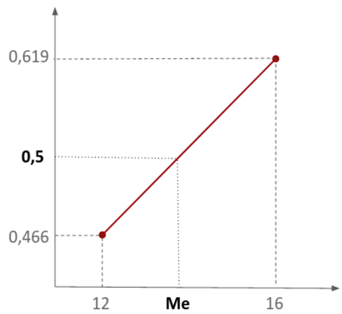
3. Pour tracer la fonction de répartition il faut trouver les fréquences cumulées (F_i) :

x_i	n_i	f_i	F_i
$[0;4[$	270	0.045	0.045
$[4;6[$	330	0.055	0.1
$[6;8[$	630	0.105	0.205
$[8;10[$	650	0.108	0.313
$[10;12[$	920	0.153	0.466
$[12;16[$	920	0.153	0.619
$[16;20[$	780	0.13	0.749
$[20;30[$	540	0.09	0.839
$[30;40[$	450	0.075	0.914
$[40;50[$	300	0.05	0.964
$[50;60[$	150	0.025	0.989
$[60;100[$	60	0.01	1

Alors, le graphique de la fonction de répartition est :



Pour trouver la médiane il faut noter qu'elle se trouve dans la classe $[12;16[$. On assume que la fonction de répartition est linéaire entre les deux extrêmes de la classe, donc pour trouver la valeur de la médiane il faut construire la droite que passe par les points $(12;0.466)$ et $(16;0.619)$ et après l'utiliser pour trouver la valeur de x pour laquelle $(x;0.5)$ appartient à cette droite. Cette situation est illustrée dans la figure ci-dessous :



L'équation de la droite qui passe par $(12;0.466)$ et $(16;0.619)$ est :

$$(y - 0.466) = \frac{(x-12)}{(16-12)}(0.619 - 0.466).$$

En remplaçant y par 0.5 on arrive à l'équation qu'on doit résoudre pour trouver la médiane :

$$\frac{Me - 12}{16 - 12} = \frac{0.5 - 0.466}{0.619 - 0.466} \Longrightarrow Me = 12.89.$$

C'est aussi possible de trouver la droite qui passe par $(12;0.466)$ et $(16;0.619)$ en utilisant une droite d'équation générique :

$$y = ax + b$$

et en construisant le système linéaire :

$$\begin{cases} 0.466 = 12a + b \\ 0.619 = 16a + b \end{cases}$$

d'où vous trouverez que l'équation de cette droite est :

$$y = 0.03825x + 0.007.$$

En remplaçant y par 0.5, on arrive au même résultat :

$$Me = \frac{0.5 - 0.007}{0.03825} = 12.89.$$

L'intervalle inter décile est l'intervalle $[D_1; D_9]$. La valeur de D_1 est 6. Pour trouver D_9 il faut tout d'abord noter que D_9 se trouve dans la classe $[30; 40[$. On le calcule par interpolation linéaire a l'intérieure de la classe :

$$\frac{D_9 - 30}{40 - 30} = \frac{0.9 - 0.839}{0.914 - 0.839} \Longrightarrow D_9 = 38.13.$$

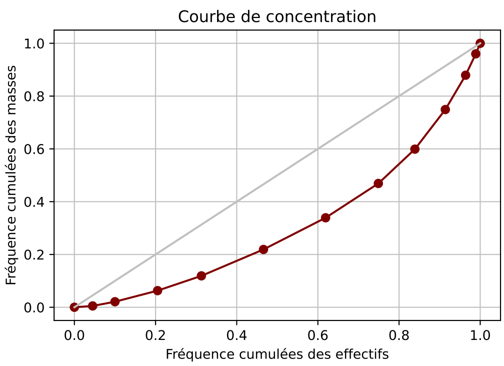
L'intervalle inter décile est alors $[6; 38.13]$. Cela veut dire que le 80% des individus ont fait des dépôts d'un montant d'entre 6 et 38.13 centaines d'euros.

4. Pour construire la courbe de concentration il faut calculer la masse de chaque classe ($n_i c_i$), la fréquence des masses ($n_i c_i / m_{\text{totale}}$) et les fréquences cumulées des masses (Q_i) :

x_i	n_i	$n_i c_i$	$n_i c_i / m_{\text{totale}}$	Q_i
$[0; 4[$	270	540	0.005	0.005
$[4; 6[$	330	1650	0.016	0.021
$[6; 8[$	630	4410	0.042	0.063
$[8; 10[$	650	5850	0.056	0.119
$[10; 12[$	920	10120	0.1	0.219
$[12; 16[$	920	12880	0.12	0.339
$[16; 20[$	780	14040	0.13	0.469
$[20; 30[$	540	13500	0.13	0.599
$[30; 40[$	450	15750	0.15	0.749
$[40; 50[$	300	13500	0.13	0.879
$[50; 60[$	150	8250	0.08	0.96
$[60; 100[$	60	4800	0.05	1

$m_{\text{totale}} = 105290$

La courbe de concentration ou courbe de Lorentz est alors :



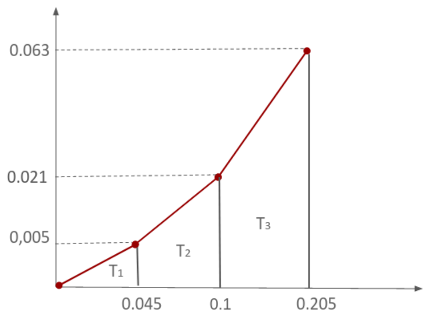
On va calculer l'indice de concentration de Gini comme :

$$G = \frac{\int_0^1 l(t)dt}{1/2},$$

où l désigne la courbe de Lorentz. Pour calculer l'intégral $\int_0^1 l(t)dt$ on va décomposer l'aire au dessous de la courbe comme une addition d'aires de trapèzes :

$$\int_0^1 l(t)dt = A(T_1) + A(T_2) + A(T_3) + \cdots + A(T_{11}).$$

Les premiers trois trapèzes sont montrés dans la figure ci-dessous.



En rappelant la formule pour calculer l'aire d'un trapèzeon calcule l'aire de chacun de ces trapèzes :

$$\begin{aligned}
A(T_1) &= (0.045 - 0)\frac{(0.005 + 0)}{2} = 0.0001125, \\
A(T_2) &= (0.1 - 0.045)\frac{(0.005 + 0.021)}{2} = 0.000715, \\
A(T_3) &= (0.313 - 0.205)\frac{(0.021 + 0.063)}{2} = 0.00441, \\
&\vdots \\
A(T_{11}) &= (1 - 0.989)\frac{(1 + 0.96)}{2} = 0.01078.
\end{aligned}$$

En additionnant l'aire de tous les trapèzes, on obtient $\int_0^1 l(t)dt = 0.309$, donc $G = 0.618$.
Pour trouver la médiale il faut noter qu'elle se trouve dans la classe $[20; 30[$. On calcule la médiale par interpolation linéaire a l'intérieure de la classe :

$$\frac{Ml - 20}{30 - 20} = \frac{0.5 - 0.469}{0.599 - 0.469} \Longrightarrow Ml = 22.38.$$

Le 50% de la masse totale est obtenue à partir des individus qui ont fait des dépôts d'un montant inférieure à 22.38 centaines d'euros.