Lycée Carnot
ECS 4 – Mathématiques
A. Troesch

Décembre 2005

Correction du Devoir Maison n° 7 – Statistiques

Problème –

On regroupe toutes les informations déduites des données salariales dans le tableau de la figure 1.

1. (a) Histogramme et polygône des effectifs : Voir Figure 2.
   Les hauteurs des rectangles sont indiquées dans le tableau de la figure 1. La classe maximale a été centrée sur sa moyenne 3320 : il s'agit donc de la classe $]3000, 3640]$.
   Polygône des effetifs cumulés : voir Figure 3

   (b) Voir figures 1 et 3 (avec l'échelle représentée entre parenthèses sur l'axe des ordonnées). Contrairement qu polygône des effectifs cumulés, la fonction de répartition est prolongée par la fonction constante égale à 0 en dessous de la valeur minimale, et par la fonction constante égale à 1 au dessus de la valeur maximale (tracé en traits interrompus)

   (c) La classe modale est la classe dont la hauteur du rectangle de l'histogramme est maximale (donc de concentration maximale). Il s'agit donc de la classe $]1300, 1500]$.

   (d) Moyenne : $\overline{x} = \sum f_i x_i = 1463.08$.
   Moyenne quadratique : $q = \sqrt{\sum f_i x_i^2} = 1498.68$.
   Variance : on utilise la formule de König-Hyughens, qui nous évite de répercuter trop de fois l'erreur d'approximation faite sur la moyenne. Ainsi : $V = \sigma^2 = q^2 - m^2 = 105552$.
   Écart-type : $\sigma = \sqrt{V} = 324.71$.

   (e) La moitié de l'effectif est atteint dans la classe $]1300, 1500]$. La médiane m est donc obtenue par la règle de 3 suivante :

$$\frac{m - 1300}{1500 - 1300} = \frac{500 - 251}{674 - 261}, \qquad \text{soit :} \qquad m = \frac{500 - 261}{674 - 251} \cdot 200 + 1300 = 1417.73.$$

   La médiane est donc inférieure à la moyenne. Les effectifs sont donc plus concentrés sur les bas salaires que sur les hauts ; seul un petit nombre de salariés ont droit à des gros salaires, qui sont vraiment plus gros que les autres. Un peu comme les notes aux DS...

   (f) Déterminons les deux quartiles :
   Premier quartile : il se situe dans la classe $]1000, 1300]$:

$$\frac{q_1 - 1000}{1300 - 1000} = \frac{250 - 6}{251 - 6} \qquad \text{soit :} \qquad q_1 = 1298.78.$$

   Remarquez que cette valeur est proche de 1300, ce qui est normal puisque la valeur en 1300 dépasse le quart de l'effectif de seulement 1.
   Troisième quartile : il se situe dans la classe $]1500, 1700]$:

$$\frac{q_3 - 1500}{1700 - 1500} = \frac{750 - 674}{885 - 674} \qquad \text{soit :} \qquad q_3 = 1572.04.$$

   Ainsi, l'intervalle interquartile est $[1298.78, 1572.04]$, et l'espace interquartile est $q_3 - q_1 = 273.26$. Ainsi, la moitié centrale de l'effectif se situe dans une tranche de salaire assez réduite, d'une largeur de moins de 300 euros. Il y a une assez forte concentration des salaires intermédiaires.
   1er décile : il se situe dans la tranche $]1000, 1300]$:

$$\frac{d_1 - 1000}{1300 - 1000} = \frac{100 - 6}{251 - 6} \qquad \text{soit :} \qquad d_1 = 1115.10.$$

salaires (a_i)	500	1000	1300	1500	1700	2000	2500	3000	(3640)
effectifs (n_i)	6	245	423	211	65	26	20	4	
fréquences (f_i)	0.006	0.245	0.423	0.211	0.065	0.026	0.02	0.004	
eff. cumulés	0	6	251	674	885	950	976	996	1000
fréq. cumulées (u_i)	0	0.006	0.251	0.674	0.885	0.950	0.976	0.996	1
$u_i = u'_i = u''_i$	0.006	0.251	0.674	0.885	0.950	0.976	0.996	1	
v_i	0.003	0.196	0.600	0.831	0.913	0.953	0.991	1	
Discrétisation (x_i)	750	1150	1400	1600	1850	2250	2750	3320	
Hauteurs (1/euro)	0.012	0.817	2.115	1.055	0.217	0.052	0.04	0.006	
$x_i^2/10000$	56.25	132.25	196	256	342.25	506.25	756.25	1102.24	
Salaires doublés (a'_i)	1000	2000	2600	3000	3400	4000	5000	6000	(7280)
Discrétisation (x'_i)	1500	2300	2800	3200	3700	4500	5500	6640	
${x'_i}^2/10000$	225	529	784	1024	1369	2025	3025	4409	
u'_i	0.006	0.251	0.674	0.885	0.950	0.976	0.996	1	
v'_i	0.003	0.196	0.600	0.831	0.913	0.953	0.991	1	
$(u_i - u_{i-1})(v_i + v_{i-1})$	0.000	0.049	0.337	0.302	0.113	0.049	0.039	0.008	
Discrétisation (x''_i)	2213	2613	2863	3063	3313	3713	4213	4783	
v''_i	0.005	0.223	0.637	0.858	0.932	0.965	0.993	1	
$(u''_i - u''_{i-1})(v''_i + v''_{i-1})$	0.000	0.056	0.364	0.315	0.116	0.049	0.039	0.008	

FIG. 1 – Tableau des salaires, et données déduites

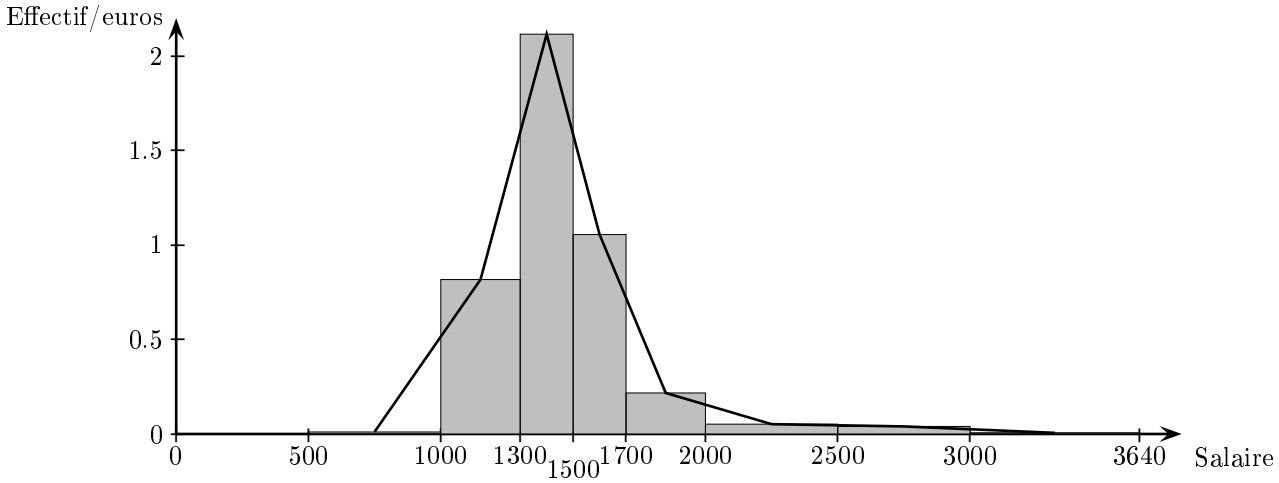


FIG. 2 – Histogramme et polygône des effectifs

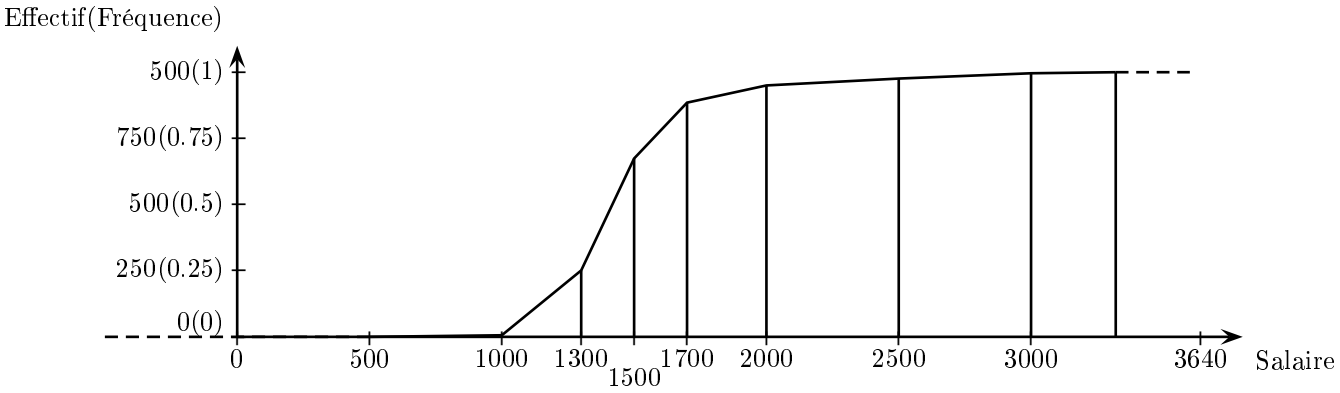


FIG. 3 – Diagramme en bâton et polygône des effectifs (fréquences) cumulé(e)s, fonction de répartition

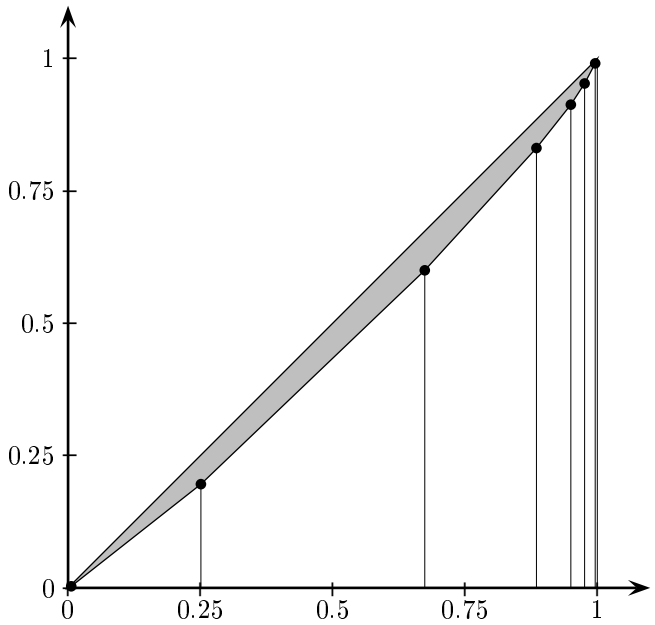


FIG. 4 – Courbe de Lorentz de la répartition de salaires donnée

9e décile : il se situe dans la tranche $]1700, 2000]$:

$$\frac{d_9 - 1700}{2000 - 1700} = \frac{900 - 885}{950 - 885} \qquad \text{soit :} \qquad d_9 = 1769.23.$$

Le dixième inférieur de l'effectif a un très petit salaire ; l'entreprise aime le personnel sous-payé semble-t-il (stagiaires...)

Les 9 dixièmes de l'effectif ont un salaire inférieur à 1769, c'est-à-dire loin d'être exhorbitant. Il y a donc peu d'employés privilégiés avec de gros salaires. Ce qui n'exclut pas le fait que ces quelques employés soient vraiment beaucoup privilégiés.

Affinons l'étude du resserrement médian en calculant les 4e et 6e déciles : **4e décile :** il se situe dans la tranche $]1300, 1500]$:

$$\frac{d_4 - 1300}{1500 - 1300} = \frac{400 - 251}{674 - 251} \qquad \text{soit :} \qquad d_4 = 1370.45.$$

6e décile : il se situe aussi dans la tranche $]1300, 1500]$:

$$\frac{d_6 - 1300}{1500 - 1300} = \frac{600 - 251}{674 - 251} \qquad \text{soit :} \qquad d_6 = 1465.23.$$

Cela confirme la concontration médiane des salaires : le cinquième médian de l'effectif (200 personnes) se retrouve dans une tranche de salaire inférieure à 100 euros.

3e centile : il se situe dans la tranche $]1000, 1300]$:

$$\frac{c_3 - 1000}{1300 - 1000} = \frac{30 - 6}{251 - 6} \qquad \text{soit :} \qquad c_3 = 1029.39.$$

97e centile : il se situe dans la tranche $]2000, 2500]$:

$$\frac{c_{97} - 2000}{2500 - 2000} = \frac{970 - 950}{976 - 950} \qquad \text{soit :} \qquad c_{97} = 2384.62.$$

Les différences entre c_3, d_1 et d_4 sont nettement inférieures aux différences entre d_6, d_9 et c_{97}. Cela confirme le fait que les salaires sont beaucoup plus étalés (et donc inégaux) sur les tranches supérieures.

2. Soit I tel que $(x_i)_{i\in I}$ représente l'ensemble des classes discrétisées de la série statistique initiale. Soit $(x'_i)_{i\in I}$ les nouvelles tranches de salaire discrétisées. Alors pour tout $i \in I$, $x'_i = 2x_i$. Les effectifs quant à eux ne changent pas. Ainsi, la nouvelle moyenne est :

$$\overline{x}' = \sum_{i\in I} f_i x'_i = \sum_{i\in I} 2 f_i x_i = 2\overline{x} = 2926.16.$$

De même, le moment d'ordre 2 est :

$$m'_2 = {q'}^2 = \sum_{i\in I} f_i {x'_i}^2 = 4 \sum_{i\in I} f_i {x_i}^2 = 4q^2.$$

Ainsi, la nouvelle variance est donc : ${\sigma'}^2 = 4q^2 - (2\overline{x})^2 = 4\sigma^2$.

Donc, le nouvel écart-type est : $\sigma' = 2\sigma = 649.4$.

Ce résultat est très logique : tous les salaires ayant doublé, on obtient une répartition des salaires identique, à un facteur multiplicatif 2 près. Ainsi, l'écart moyen par rapport à la moyenne est le double du précédent.

La répartition des salaires n'en est pas pour autant moins équitable, puisqu'elle est la même, à une autre échelle. Ainsi, l'écart-type n'est pas un paramètre satisfaisant pour mesurer l'équité d'une répartition de salaires.

3. La question étant ambiguë, on va tracer la courbe de Lorentz des deux séries avant et après multiplication des salaires. Après calcul des coordonnées $(u_i, v_i)_{i\in I}$ de la courbe de Lorentz avant multiplication, et des coordonnées $(u'_i, v'_i)_{i\in I}$ après multiplication, on se rend compte que la courbe de Lorentz est la même avant et après multiplication. On peut donc se contenter d'en tracer une, ce qu'on fait dans la figure 4

4. Par définition, une des extrémités est le point $(0,0)$. De plus, l'autre extrémité est le point (u_p, v_p) défini par :

$$u_p = \sum_{j=1}^{p} f_j = 1 \quad \text{(fréquence totale)} \qquad \text{et} \qquad v_p = \frac{\displaystyle\sum_{j=1}^{p} f_j x_j}{\overline{x}} = \frac{\overline{x}}{\overline{x}} = 1,$$

par définition de la moyenne. Ainsi, la courbe de Lorentz est d'extrémités $(0,0)$ et $(1,1)$.
De plus, la courbe de Lorentz est toujours en dessous de la diagonale $x = y$. Pour montrer cela, il suffit de vérifier que pour tout $i \in [\![1,p]\!]$, $v_i \leqslant u_i$. Posons, pour tout $i \in [\![i,p]\!]$, $w_i = v_i - u_i$, et $w_0 = 0$. Alors :

$$\forall i \in [\![1,p]\!],\ w_i = \frac{1}{\overline{x}} \sum_{j=1}^{i} f_i(x_i - \overline{x}) \qquad \text{donc:} \qquad \forall i \in [\![1,p]\!],\ w_i - w_{i-1} = \frac{f_i}{\overline{x}}(x_i - \overline{x}).$$

Or $(x_i)_{i\in[\![1,p]\!]}$ est croissante, donc aussi $(x_i - \overline{x})_{i\in[\![1,p]\!]}$. Ainsi, l'ensemble des valeurs négatives cette suite est concentré en début de l'ensemble $[\![1,p]\!]$ et l'ensemble des valeurs positives en fin d'intervalle : il existe $k \in [\![1,p+1]\!]$ tel que :

$$\forall i \in [\![1,k-1]\!],\ x_i - \overline{x} < 0 \qquad \text{et} \qquad \forall i \in [\![k,p-1]\!],\ x_i - \overline{x} \geqslant 0.$$

(si $k = 1$, toutes les valeurs sont positives ; si $k = p+1$, elles sont toutes négatives). Comme pour tout $i \in [\![1,p]\!]$, $f_i \geqslant 0$ et $\overline{x} > 0$, $(w_i)_{i\in[\![0,p]\!]}$ est décroissante jusqu'au rang $k-1$, puis croissante. Comme $w_0 = 0$ et $w_p = 1 - 1 = 0$, on en déduit que pour tout $i \in [\![0,p]\!]$, $w_i \leqslant 0$, c'est-à-dire $v_i \leqslant u_i$.
Remarquez qu'il manque une hypothèse dans l'énoncé : il faut supposer que les valeurs x_i sont toutes positives (ou au moins que la moyenne $\overline{x}$ soit positive) pour pouvoir déterminer le signe de $w_i - w_{i-1}$. Il est évident que si les valeurs x_i sont toutes négatives, le résultat est inversé, donc la courbe se retrouvera au-dessus de la droite $y = x$.

5. Si les salaires sont tous égaux, la série statistique est de la forme $\{(x_1, n_1)\}$ (une seule valeur). Alors les extrémités de la courbe de Lorentz sont $M_0 = O$ et M_1, donc $M_1 = (1,1)$ d'après le début de la question 4. Nous traçons la courbe de Lorentz dans la figure 5(a).

6. La série statistique discrète d'une entreprise escalagiste est $\{(0, n-1), (x, 1)\}$, où n est l'effectif total (y compris le patron), et x est le salaire du patron. Alors, la courbe de Lorentz est la ligne polygonale reliant les trois points :

$$M_0 = O, \qquad M_1 = \left(1 - \frac{1}{n}, 0\right), \qquad M_2 = (1,1).$$

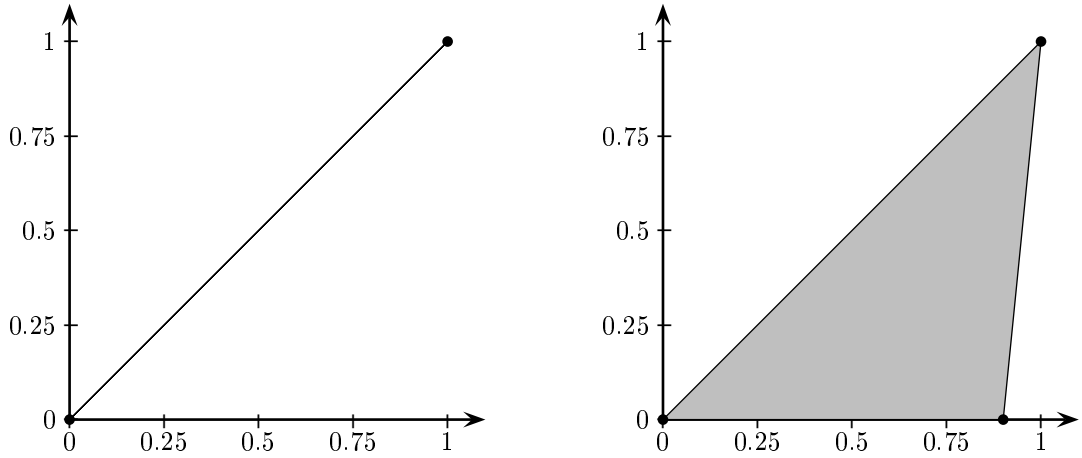
On la trace dans la figure 5(b).

7. On a montré que la courbe de Lorentz est sous la droite d'équation $y = x$. Elle est également au dessus de l'axe des abscisses, toutes les valeurs étant positives. Ainsi, elle se situe dans le triangle de sommets $(0,0)$, $(0,1)$ et $(1,1)$. Par conséquent, l'aire comprise entre cette courbe et la droite $y = x$, c'est-à-dire un des côtés de ce triangle, est inférieure à l'aire du triangle, à savoir $\frac{1}{2}$. Elle est bien sûr positive. Après doublement, on obtient donc que l'indice de concentration est compris entre 0 et 1.

8. L'indice de concentration de l'entreprise égalitaire est $2 \cdot 0 = 0$.
L'indice de concentration de l'entreprise esclavagiste est $1 - \dfrac{1}{n}$ (double de l'aire d'un triangle de hauteur 1 de base $1 - \dfrac{1}{n}$).
La répartition des salaires de l'entreprise égalitaire est de faible concentration (la somme totale n'est pas concentrée sur peu de personnes) ; la repartition de l'entreprise esclavagiste est de forte concentration (la somme totale est concentrée sur une seule personne).

9. On calcule d'abord l'aire sous la courbe, en découpant cette aire en trapèzes de hauteur $u_{i+1} - u_i$, de bases v_i et v_{i+1}, pour tout $i \in [\![0,p-1]\!]$ (avec $u_0 = v_0 = 0$) (cf figure 4 Ainsi, cette aire vaut :

$$A = \sum_{i=0}^{p-1} \frac{(u_{i+1} - u_i)(v_i + v_{i+1})}{2}.$$



(a) Entreprise égalitaire

(b) Entreprise esclavagiste ($n = 10$)

FIG. 5 – Exemples de courbes de Lorentz

L'indice de concentration vaut donc :

$$c = 1 - \sum_{i=1}^{p} (u_i - u_{i-1})(v_{i-1} + v_i) = 1 - 0.896 = 0.104.$$

Nous avons regroupé dans le tableau de la figure 1 les valeurs approchées à 10^{-3} des termes de la somme.

10. Comme on l'a fait remarquer, la courbe de Lorentz est la même avant et après dédoublement, donc l'indice de concentration est le même. Cela n'est pas étonnant : un doublement de tous les salaires ne crée pas d'inégalité supplémentaire ; la répartition est en effet la même, à une échelle près. L'indice de concentration ne peut pas dépendre d'une telle échelle, sinon il dépendrait de l'unité de mesure choisi, ce qui n'aurait pas de sens.

11. Il faut bien sûr supposer dans cette question que les salaires sont uniformément répartis dans chaque classe. Cela revient, pour le calcul de moyennes et de sommes totales, à considérer la discrétisation. Ainsi, le coût total pour l'entreprise (si on oublie toutes les cotisations patronales) est la somme

$$\sum_{i=1}^{p} x_i n_i = 1000\overline{x} = 1463080.$$

Il faut donc répartir cette somme équitablement entre les différents salariés, donc augmenter de 1456.58 chaque salaire (cela revient bien sûr à ajouter le salaire moyen à tout le monde). Cela fait une translation de toutes les classes, donc aussi de la discrétisation, d'une valeur de 1456.58.

12. On note $\{(x''i, n_i), i \in [\![1, p]\!]$ la nouvelle série statistique obtenue. On note pour tout $i \in [\![1, p]\!]$ (u''_i, v''_i) les coordonnées des points de la courbe de Lorentz associée à cette nouvelle série. On regroupe toutes ces données dans le tableau 1 (on a tronqué x''_i à l'euro). La valeur u''_i (effectif cumulé) correspond bien sûr à u_i. La nouvelle moyenne est le double de l'ancienne, donc $\overline{x''} = 2926.16$.

On trouve alors le nouvel indice de concentration : $c'' = 1 - 0.948 = 0.052$.

Ainsi, l'indice de concentration est plus faible, donc les salaires sont mieux répartis ainsi. Cela semble logique : en doublant les salaires, on double aussi les disparités entre les salaires. En revanche, en les translatant, les disparités restent les mêmes (l'échelle ayant, elle doublé). Ainsi, les disparités vont être deux fois moins importante par la méthode proposée par les syndicats.