Étude de Cas

Corrigé du TD 1
Janvier 2012

Exercice 1 :

1. La distribution des dividendes versés aux actionnaires d'une entreprise pour l'exercice 2005 est donnée par le tableau suivant :

Dividendes versés en milliers d'euros	**[0 ; 2[**	**[2 ; 10[**	**[10 ; 20[**	**[20 ; 60[**	**[60 ; 100[**	**100 ≥**	**Ensemble**
Pourcentage d'actionnaires	**16**	p_2	p_3	**17**	**10**	**3**	**100**

Les proportions p_2 , p_3 des classes [2 ; 10[et [10 ; 20[sont inconnues.

1.1. Déterminez p_2 et p_3 sachant que le 4ème décile est égal à 12500 euros.

Pour déterminer les fréquences manquantes, nous avons besoin de trouver deux équations qui nous permettront de les identifier. La première équation que l'ont peut écrire est

$$0,16+p_2+p_3+0,17+0,1+0,03=1$$

parce que la somme des fréquences vaut toujours 1, soit $p_2+p_3=0,54$.
Ensuite, l'hypothèse d'équirépartition au sein des classes nous permet de déterminer une équation supplémentaire. En effet, la formule de l'interpolation linéaire nous permet d'écrire:

$$12,5 = 10 + (20-10) \times \frac{0,4-(p_2+0,16)}{p_3}$$

$$\Leftrightarrow \frac{p_3}{10} = \frac{0,24-p_2}{2,5}$$

Nous avons ainsi un système linéaire à deux équations et deux inconnus

$$\begin{cases} p_2 + p_3 = 0,54 \\ p_3 = 0,96 - 4p_2 \end{cases}$$

qui possède une unique solution égale à $p_3=0,4$ et $p_2=0,14$.

1.2. Calculez l'écart-type de la distribution, sachant que le dividende moyen versé par l'entreprise à ses actionnaires est 60 kiloeuros.

En faisant l'hypothèse d'équirépartition au sein de chaque classe, nous pouvons identifier la moyenne de chaque classe au milieu de l'intervalle. Cette hypothèse permet d'avoir une estimation de la moyenne des 5 premières classes, mais ne permet pas d'en déduire la moyenne de la dernière classe $(X \geq 100)$. Nous avons cependant à notre disposition la moyenne de la population totale $\overline{x} = 60$ keuros. Nous pouvons utiliser la relation générale (et exacte) suivante qui relie moyenne des classes et moyenne globale :

$$p_1\overline{x}_1 + p_2\overline{x}_2 + \ldots + p_5\overline{x}_5 + p_6\overline{x}_6 = \overline{x}$$

Grâce à la question 1, nous n'avons qu'une seule inconnue $\overline{x}_6$, qui est égale à

$$0,03\overline{x}_6 = 60 - (1\times 0,16 + 6\times 0,14 + 15\times 0,4 + 40\times 0,17 + 80\times 0,1)$$

soit $\overline{x}_6 = 1273,33$ keuros.
Nous pouvons maintenant calculer l'écart-type par la formule :. $\sqrt{\sum_{i=1}^{n} f_i(x_i - \bar{x})^2}$

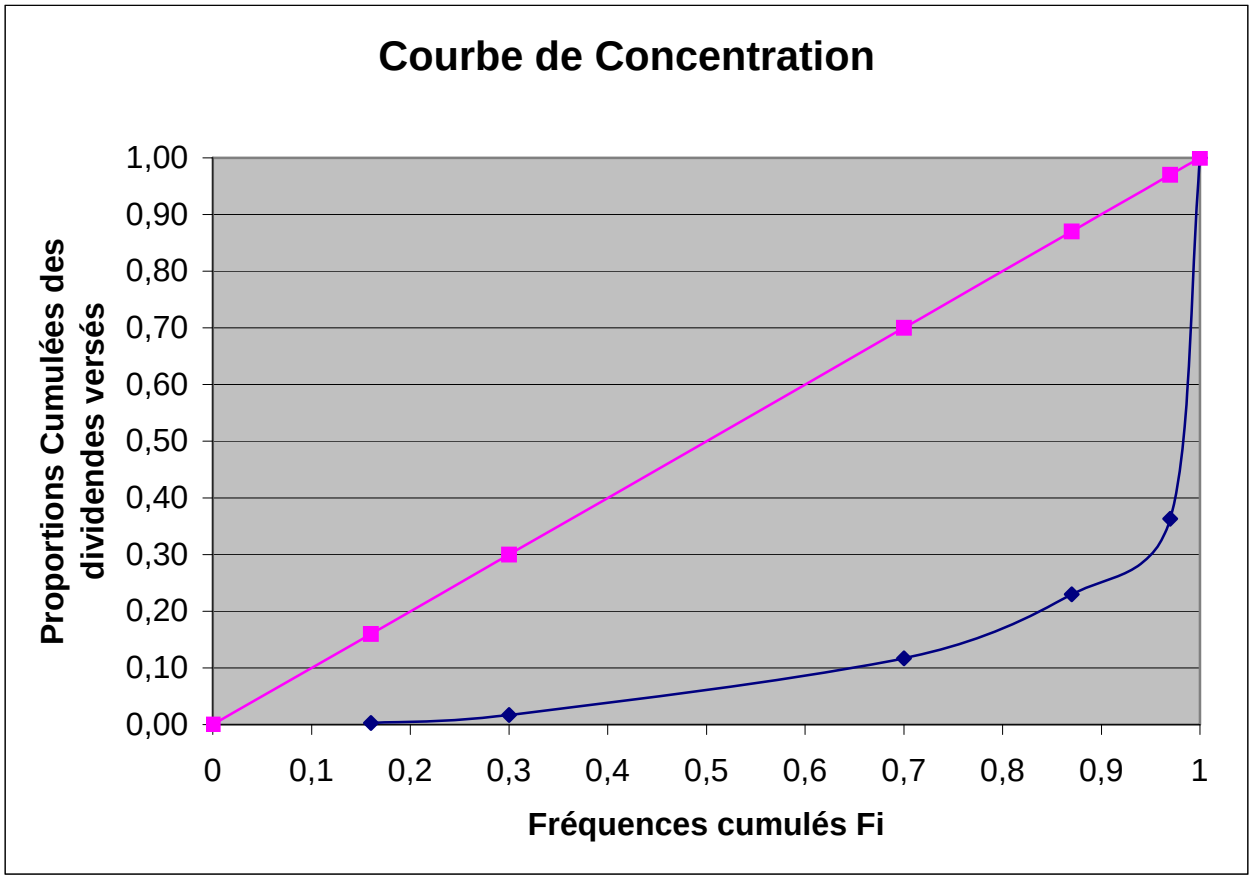
L'application numérique nous donne $s_X = 214,589$ keuros.

1.3 Tracez la courbe de concentration.

Nous avons besoin de calculer les fréquences cumulées et les dividendes cumulées (et leur proportion), il était donc indispensable de calculer la moyenne de la dernière classe. Pour calculer la proportion, il suffit de diviser

Classe	[0 ; 2[	[2 ; 10[	[10 ; 20[	[20 ; 60[	[60 ; 100[	100 ≥
Fréquence %	16	14	40	17	10	3
Moyenne	1	6	15	40	80	1273,33
Cumul dividendes	0,16	1	7	13,8	21,8	60
Proportion Dividendes	0,00	0,02	0,12	0,23	0,36	1
Fréquences cumulées	0,16	0,3	0,7	0,87	0,97	1

par la moyenne totale égale à 60 keuros et donner à la question précédente. La courbe de Lorentz (ou de concentration) est :



La courbe met en évidence l'importance écrasante des gros actionnaires par rapport aux petits porteurs. Nous pourrions mesurer l'aire entre la courbe de concentration et la 1ère bissectrice à l'aide de l'indice de Gini, ce qui représente un indice d'inégalité (disparité) de la population des actionnaires.

2. Dans la Creuse, la répartition du nombre d'employés dans les entreprises de moins de 250 employés (TPME et PME) et de la consommation électrique annuelle (en MégaWattheure, MWh) est donnée dans le tableau suivant :

Conso (MWh) : *Y* Nbre employés : *X*	**< 10**	**[10 ; 50[**	**[50 ; 100[**	**≥100**	**Total**
[0 ; 10[	**2345**	**301**	**24**	**0**	**2670**
[10 ; 50[	**1345**	**2006**	**960**	**389**	**4700**
[50 ; 150[	**570**	**1560**	**3076**	**3250**	**8456**
[150 ; 250]	**0**	**35**	**656**	**1097**	**1788**
Total	**4260**	**3902**	**4716**	**4736**	**17614**

2.1. Définissez la population, l'unité statistique, les caractères étudiés et leur nature.

La population est l'ensemble des entreprises de moins de 250 employés dans la Creuse.
Unité statistique : une PME ou une TPME.
Caractères étudiés : le nombre d'employés = variable quantitative discrète, la consommation électrique annuelle = variable quantitative continue exprimée en MWh.

Nous allons donc faire une étude bivariée sur la population des PME, TPME.

2.2. Tracez en parallèle les boîtes à pattes de la distribution de la consommation électrique des entreprises de moins de 49 employés et de plus de 49 employés. On ne représentera pas les points extrêmes, ni les points éloignés, mais on prendra comme extrémités des « pattes » les déciles D_1 et D_9 convenablement estimés. Commentez.

Nous calculons uniquement les quartiles et les déciles, sans tracer les boîtes à moustache. Dans cette question, tout repose sur l'habituelle **hypothèse d'équirépartition** au sein de chaque classe, ce qui nous permet d'utiliser la formule d'interpolation linéaire pour les quantiles.

Population des entreprises <50 employés

Effectifs	3690	2307	984	389
fi	0,50	0,31	0,13	0,05
Fi	0,50	0,81	0,94	1,00
Y	**< 10**	**[10 ; 50[**	**[50 ; 100[**	**≥ 100**

Quantiles	D1	Q1	Me	Q3	D9
En MWH	2	5	10	42,26	84,62

Remarques :

On trouve directement la médiane, sans passer par interpolation.

Pour calculer les quantiles de cette distribution, il n'est pas nécessaire de déterminer la borne supérieure de la classe extrême (car ils n'appartiennent pas à cette classe).

Pour les entreprises >49 employés :

Effectifs	570,00	1595,00	3732,00	4347,00
fi	0,06	0,16	0,36	0,42
Fi	0,06	0,22	0,58	1,00
Y en MWh	< **10**	**[10 ; 50[**	**[50 ; 100[**	**≥100**

Dans cette configuration, pour calculer par interpolation les deux quantiles Q_3 et D_9 qui sont dans la dernière classe, nous avons besoin de déterminer une borne supérieure à la classe $(Y \geq 100)$. Nous n'avons pas l'information de la moyenne de Y dans cette classe ou de la moyenne de la consommation sur toute la population des entreprises de plus de 49 employés (de laquelle nous aurions pu déduire la moyenne de la classe $(Y \geq 100)$, cf. question 1.2.). Nous avons vu qu'en faisant l'hypothèse d'équirépartition au sein de cette classe, il aurait été possible de construire une borne supérieure en symétrisant l'intervalle autour de la moyenne.
Ici ce n'est pas possible : nous faisons une hypothèse différente et **arbitraire** en supposant que $(Y \geq 100) = [100; 200]$.

Quantiles	D1	Q1	Me	Q3	D9
En MWh	20,00	54,17	88,89	140,48	176,19

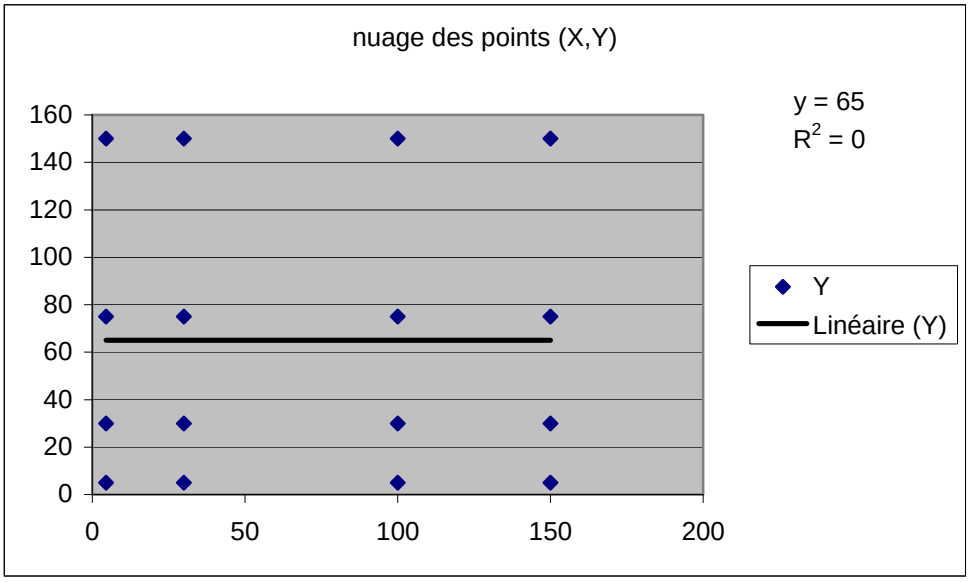
2.3. Calculez la distribution conditionnelle (en pourcentage) de la consommation électrique sachant que le nombre d'employés est compris entre 10 et 150. Tracez l'histogramme de cette distribution.

Pour calculer cette distribution conditionnelle, il faut additionner les deux lignes $[10;50[$ et $[50;100[$ et « normaliser » par rapport au nombre total d'entreprises ayant entre 10 et 150 employés, soit 13156 employés.
Pour tracer l'histogramme, nous avons besoin de calculer les hauteurs de chaque rectangle car les amplitudes des classes sont différentes, voir le tableau suivant :

	[0,10]	[10,50]	[50,100]	[100,200]
effectif	1915	3566	4036	3639
frequence conditionnelle	0,15	0,27	0,31	0,28
amplitude	10,00	40,00	50,00	100,00
hauteur (facteur 10)	0,15	0,07	0,06	0,03

2.4. Représentez le nuage de points d'abscisse le nombre moyen d'employés par classe et d'ordonnée la consommation moyenne d'électricité correspondante par classe. Expliquez, au moyen d'une relation que l'on précisera, les variations de la consommation moyenne d'électricité par le nombre moyen d'employés. Quelle est la qualité explicative de ce modèle ?

Une première méthode est de tracer les 16 points d'abscisse et ordonnée les centres des classes, et à construire la droite des moindres carrés (et donc un modèle explicatif) à partir de ces points-là, que l'on voit dans le graphe suivant.



Lorsque l'on calcule la droite des moindres carrées correspondantes, le R^2 est nulle, ce qui montre que nous n'expliquons pas du tout la consommation d'électricité par le nombre d'employés ce qui pose donc un problème, car le tableau des données nous indiquent clairement une corrélation positive entre ces deux grandeurs (les cases diagonales ont des effectifs beaucoup plus grand que les autres), ce qui est paradoxal Le problème provient du fait que nous considérons que tous les points ont le même poids, ce qui n'est clairement pas le cas. Il faut donc en prendre compte les effectifs dans le calcul de la covariance et des variances, et ensuite pour le calcul de la droite.

X	Y	EFFECTIF	POIDS (P)	X*P	Y*P
4,50	5,00	2345,00	0,133	0,60	0,67
4,50	30,00	301,00	0,017	0,08	0,51
4,50	75,00	24,00	0,001	0,01	0,10
4,50	150,00	0,00	0,000	0,00	0,00
30,00	5,00	1345,00	0,076	2,29	0,38
30,00	30,00	2006,00	0,114	3,42	3,42
30,00	75,00	960,00	0,055	1,64	4,09
30,00	150,00	389,00	0,022	0,66	3,31
100,00	5,00	570,00	0,032	3,24	0,16
100,00	30,00	1560,00	0,089	8,86	2,66
100,00	75,00	3076,00	0,175	17,46	13,10
100,00	150,00	3250,00	0,185	18,45	27,68
150,00	5,00	0,00	0,000	0,00	0,00
150,00	30,00	35,00	0,002	0,30	0,06
150,00	75,00	656,00	0,037	5,59	2,79
150,00	150,00	1097,00	0,062	9,34	9,34

On trouve alors :

Moyenne X	71,92
Moyenne Y	68,27
Sx	46,43
Sy	55,68
r(X,Y)	0,63
a	0,52
b	30,68

Avec R^2=0,4, ce qui est maintenant plus explicatif comme modèle. La corrélation est égale à 0,63 ce qui permet aussi de mettre en évidence la liaison linéaire que l'on pouvait discerner dans le tableau.

Une deuxième méthode permet de faire beaucoup moins de calcul en faisant un peu plus d'hypothèse. Comme nous voulons expliquer la consommation d'électricité des entreprises par le nombre d'employés, nous calculons

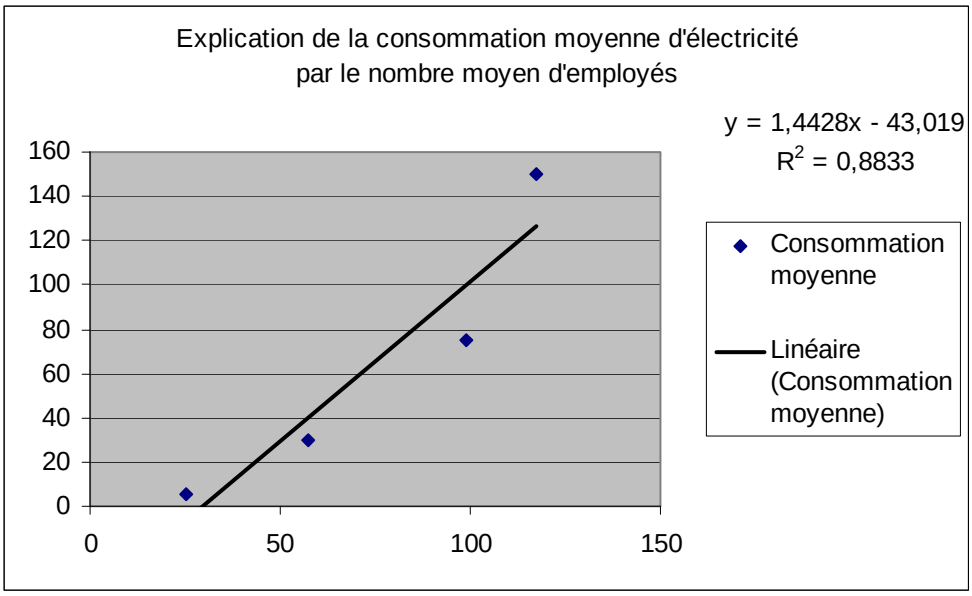
(estimons) tout d'abord le nombre moyens d'employés par entreprise pour chaque tranche de consommation.

Nbre moyen d'employés par entreprise	[0;10]	[10;50]	[50;100]	[100;200]
4,5	2345	301	24	0
30	1345	2006	960	389
100	570	1560	3076	3250
200	0	35	656	1097
Nombre moyen d'employés par classe de consommation d'élec.	25,33	57,54	99,17	117,41

En supposant que la répartition est uniforme au sein de chaque classe de consommation, nous pouvons extraire 4 couples de valeurs

Nombre moyen d'employés	Consommation Moyenne MWh
25,33	5
57,54	30
99,17	75
117,41	150

Nous pouvons alors calculer très facilement la droite de régression (donnée sur le graphique ci-dessous) pour ces 4 valeurs ainsi que l'autocorrélation qui est égale à 0, 938 (lorsque l'on suppose que tous les points ont le même poids, ce qui permet d'utiliser la calculatrice pour calculer directement la droite et ses coef).
L'hypothèse d'équirépartition a tendance à surestimer la force du lien entre les deux grandeurs, ainsi que la capacité d'explication (le R^2 est bcp plus grand que dans le cas précédent), mais cela permet d'avoir une première approximation du lien entre ces 2 grandeurs.



Exercice 3.

3.1. Population: formée de 6 sous-population constituées de l'ensemble des journées en agglomération parisienne de 1998 à 2003
Unité statistique: une journée.
Caractère étudié: niveau de l'indice de la qualité de l'air, caractère qualitatif, mais aussi ordinal (les modalités du caractère sont ordonnées).