

# Chapitre 8

## Exercices

### 8.1 Exercices corrigés

**Exercice 8.1** La société "Machin Machine" vient de publier un compte rendu d'activité de l'année 2002. Dans la rubrique "Part de marché(PM)", nous avons retenu le tableau suivant :

Période	1er semestre	3ème trimestre	octobre	novembre	décembre
Taux d'accroissement	9%	5%	1,5%	2%	-0,5%

Taux d'accroissement de la part de marché

1. Calculer le taux d'accroissement de la part de marché pour l'année 2002.
2. En déduire le taux d'accroissement mensuel moyen pendant l'année 2002.
3. Calculer le taux d'accroissement moyen pendant les six premiers mois.

#### Corrigé

1. Le taux 9% concerne un accroissement total sur les six premiers mois.  
Le taux 5% concerne un accroissement total sur les trois mois du troisième trimestre.

On a alors

$$PM_{2002} = PM_{2001} (1 + 0,09) (1 + 0,05) (1 + 0,015) (1 + 0,02) (1 - 0,005)$$

$$PM_{2002} = PM_{2001} \cdot 1,179$$

Ainsi, l'accroissement de la part de marché de la société en 2002 s'élève à 17,9%.

2. Le taux d'accroissement mensuel moyen est défini par

$$g_{ma} = \sqrt[12]{1,179} - 1 = 0,014 \quad \text{soit } 1,4\%.$$

3. Le taux d'accroissement mensuel moyen sur le premier semestre s'élève à  
 $g_{ms} = \sqrt[6]{1,09} - 1 = 0,014$  soit 1,4%.

**Exercice 8.2** Nous avons enquêté 912 ménages sur leurs revenus annuels de l'année 2002 ( $X$ ). Nous résumons les informations recueillies dans le tableau qui suit :

Revenu annuel (en dinars)	Effectifs	Fréquences
[3500, 5500[	88	0,096
[5500, 6500[	334	0,366
[6500, 8000[	213	0,234
[8000, 9500[	123	0,135
[9500, 11500[	96	0,105
[11500, 13500[	40	0,044
[13500, 16000[	18	0,020
Total	912	1

Calculer le mode.

1. Calculer la médiane.

2. En déduire une valeur approchée de la moyenne empirique.

3. On s'intéresse à l'inégalité de la répartition des revenus.

(a) Déterminer la classe de revenus associée aux 10% des ménages les plus pauvres et la classe de revenus associée aux 10% des ménages les plus riches.

(b) Déterminer l'intervalle interquartiles.

(c) Commenter.

4. Représenter l'histogramme.

5. Après avoir effectué le changement de variables qui vous semble adéquat, calculer la moyenne de la distribution ainsi que son écart-type.

6. Calculer le coefficient de variation.

### Corrigé

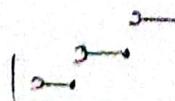
Le tableau qui suit sera complété au fur et à mesure de l'avancement de l'exercice.

classes	$n_i$	$f_i$	$n_i^c$	$F(x)$	$c_i$	$z_i$	$n_i z_i$
[3500, 5500[	88	0,096	44	0,096	4500	-17	-1496
[5500, 6500[	334	0,366	334	0,462	6000	-11	-3674
[6500, 8000[	213	0,234	142	0,696	7250	-6	-1278
[8000, 9500[	123	0,135	82	0,831	8750	0	0
[9500, 11500[	96	0,105	48	0,936	10500	7	672
[11500, 13500[	40	0,044	20	0,980	12500	15	600
[13500, 16000[	18	0,020	7,2	1	14750	24	432
Total	912	1	-	-	-	-	-4744

$$n_i^c = \frac{n_i}{\alpha_i}$$

$$[e_i, e_{i+1}]$$

$$\alpha_i = e_{i+1} - e_i$$



$$e_0 = e_1 + \frac{n^c - n^c_{0-1}}{(n^c_{0-1}) (n^c_{0-1} - n^c_{0+1})} \cdot (e_2 - e_1)$$

1. La classe modale est [5500, 6500[

$$M_o = 5500 + 1000 \cdot \frac{334 - 44}{(334 - 44)(334 - 142)} = 6101,66.$$

$M_o \cong 6102$  dinars.

2. La classe médiane est [6500, 8000[

$$M_e = 6500 + 1500 \cdot \frac{0,5 - 0,462}{0,696 - 0,462} = 6743,59.$$

$M_e \cong 6744$  dinars.

$$3. \bar{x} \cong \frac{2M_e - M_o}{2} = 7065 \text{ dinars.}$$

$$x = e_{i-1} + a_i \cdot \frac{(F(x) - F(e_{i-1}))}{f(e_i) - f(e_{i-1})}$$

(a) Il nous faut déterminer le premier et le dernier déciles.

$$10 \rightarrow 0,2. Q_{0,1} = 5500 + 1000 \cdot \frac{0,10 - 0,096}{0,462 - 0,096} = 5510,93 \cong 5511 \text{ dinars.}$$

$$Q_{0,9} = 9500 + 2000 \cdot \frac{0,90 - 0,831}{0,936 - 0,831} = 10814,29 \cong 10814 \text{ dinars.}$$

Ainsi, l'intervalle associé aux 10% les plus pauvres est [3500, 5511[. Celui associé aux 10% les plus riches est [10814, 16000[.

$$(b) Q_1 = 5500 + 1000 \cdot \frac{0,25 - 0,096}{0,462 - 0,096} = 5920,77 \cong 5921 \text{ dinars.}$$

$$Q_3 = 8000 + 1500 \cdot \frac{0,75 - 0,696}{0,831 - 0,696} = 8600 \text{ dinars.}$$

Donc  $[Q_1, Q_3] = [5921, 8600[$ .

(c) 50% des ménages ont un salaire inférieur à 6744 dinars (en dessous du salaire moyen).

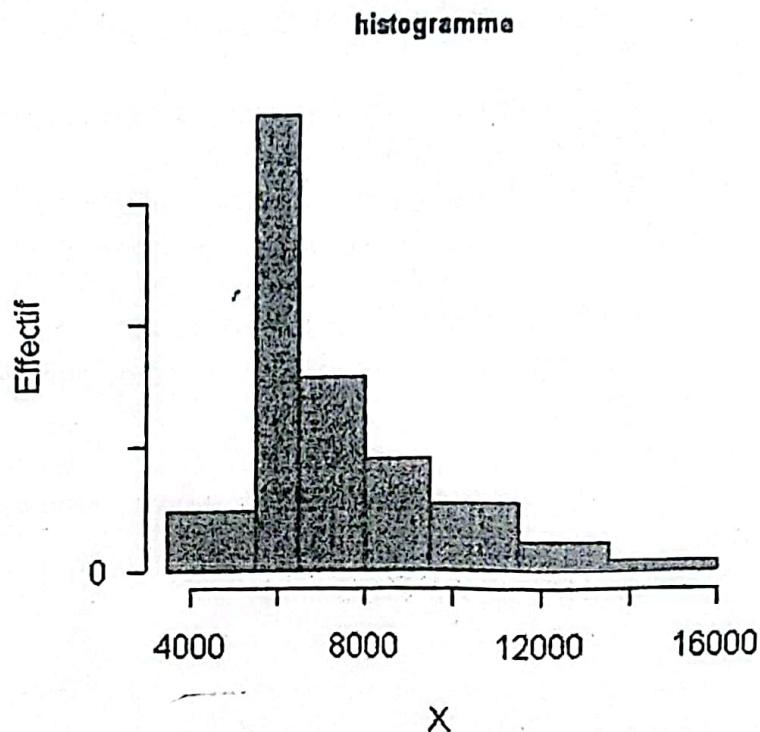
De plus, 10% des ménages ont un revenu inférieur à 5511 dinars soit 22% de moins que le salaire moyen; et 10% des ménages ont un revenu supérieur à 10814 dinars soit 50% de plus que le salaire moyen.

On peut aussi ajouter que 50% des ménages ont un revenu compris entre 5921 et 8600 dinars et inférieur à 8600 dinars.

Pour terminer, on peut conclure que dans cette population, il y a concentration des bas revenus alors que les revenus élevés ne concernent qu'une minorité ne dépassant pas 10% de la population.

## 4. Histogramme :

1996 - 20



5. On peut poser  $z_i = \frac{(c_i - 8750)}{250}$

$$\bar{z} = \frac{1}{912} \sum_i n_i z_i = -5,202.$$

Le revenu moyen est donc  $\bar{x} = 250 \cdot (-5,202) + 8750 = 7449$ .

$$s_z^2 = \frac{1}{912} \sum_i n_i z_i^2 - \bar{z}^2. \quad (\text{variance : } \frac{1}{n} \sum_{i=1}^n n_i x_i^2 - \bar{x}^2)$$

$$s_z^2 = 79,94.$$

$$\text{écart type} = \sqrt{\text{var}}$$

$$s_x^2 = 250^2 \cdot 79,94 = 4996250 \text{ donc } s_x = 2235 \text{ dinars.}$$

coeff de variation: 6.  $cv = \frac{s_x}{\bar{x}} = 0,30$

Exercice 8.3 Soit une variable statistique distribuée selon le tableau suivant :

Classe	[0, 4[	[4, 8[	[8, 12[	[12, $\alpha$ [	[ $\alpha$ , 22[	[22, 30[	[30, 42[	Total
Effectif	16	$n_2$	$n_3$	17	17	11	3	100
	0,16	$\frac{n_2}{100}$	$\frac{n_3}{100}$	0,17	0,17	0,11	0,03	

1. Sachant que la médiane  $M_e = \frac{35}{3}$ , démontrer que  $n_2 = 12$  et  $n_3 = 24$ .

2. Déterminer le premier quartile.

3. On donne  $\bar{x} = 12,99$ ; calculer la borne  $\alpha$ .

4. Tracer l'histogramme.
5. Déterminer par interpolation linéaire, la valeur du mode. La relation algébrique reliant mode, médiane et moyenne confirme-t-elle ce résultat ? Pourquoi ?
6. Déterminer un intervalle contenant 60% de la population répartis symétriquement autour de la médiane.
7. Calculer le coefficient de variation et l'écart interquartile relatif.
8. On suppose que la distribution correspond au revenu annuel (en milliers de dinars) de 100 ménages. On désire juger de l'inégalité de la répartition au sein de cette population.  
Représenter la courbe de Lorenz et calculer l'indice de Gini. Commenter.

**Corrigé**

Nous complèterons le tableau suivant à mesure que le besoin s'en ressentira.

Classe	effectif	fréquence	fréquence cumulée ↗	$c_i$	$(c_i - \bar{x})^2$	$n_i (c_i - \bar{x})^2$
[0, 4[	16	0,16	0,16	2	120,7801	1932,4816
[4, 8[	$n_2 = 12$	0,12	0,28	6	48,8601	586,3212
[8, 12[	$n_3 = 24$	0,24	0,52	10	8,9401	214,5624
[12, 16[ $\alpha = 16$	17	0,17	0,69	14	1,0201	17,3417
[16, 22[	17	0,17	0,86	19	36,1201	614,0417
[22, 30[	11	0,11	0,97	26	169,2601	1861,8611
[30, 42[	3	0,03	1	36	529,4601	1588,3803
Total	100	1				

(8.1)

$$1. n_2 + n_3 = 36$$

$$M_e = \frac{35}{3} \approx 11,667$$

Classe médiane : [8, 12[.

$$\frac{35}{3} = 8 + \frac{0,5 - 0,16 - \frac{n_2}{100}}{0,52 - 0,16 - \frac{n_2}{100}} \times 4 = 8 + \frac{0,34 - \frac{n_2}{100}}{0,36 - \frac{n_2}{100}} \times 4$$

Après calculs, on obtient  $n_2 = 12$  et donc  $n_3 = 24$ .

$$2. F(Q_1) = 0,25 \Rightarrow \in [4, 8[.$$

$$Q_1 = 4 + \frac{0,25 - 0,16}{0,28 - 0,16} \times 4 = 7$$

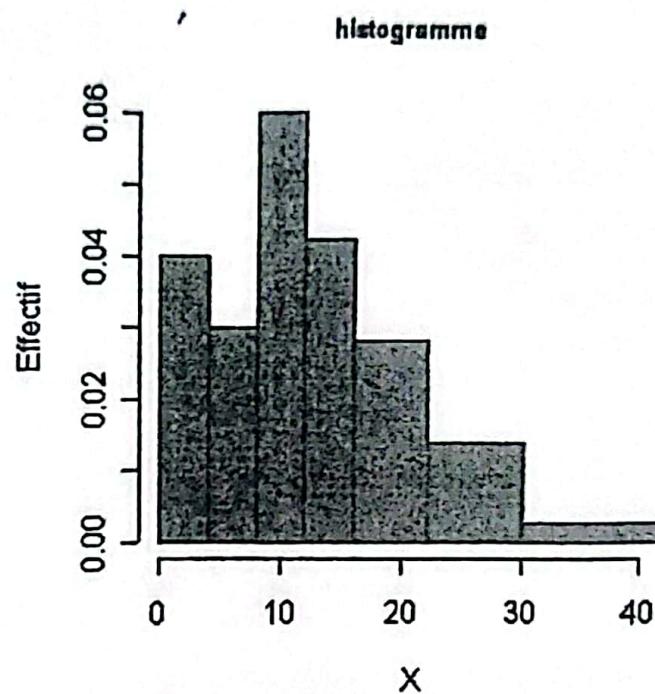
3.  $\bar{x} = 12,99$       *centrale  
de class.*

$$17 \frac{\alpha + 12}{2} + 17 \frac{\alpha + 22}{2} = 1299 - (32 + 72 + 240 + 286 + 108)$$

d'où

$$\alpha = \frac{272}{17} = 16$$

4. Représentation graphique :



5.

$$M_o = 8 + 4 \frac{12}{19} = 10,526$$

Ecrivons la relation algébrique reliant mode, médiane et moyenne :

$$M = \frac{3M_e - M_o}{2} \implies M_o = 3M_e - 2M = 9,02$$

L'égalité n'est pas vérifiée dans le cas présent car la distribution n'est pas parfaitement unimodale.

6.

$$P[M_e - a \leq X < M_e + b] = 0,6$$

$\Leftrightarrow$

$$P[X < M_e + b] - P[X < M_e - a] = 0,6$$

$\Leftrightarrow$

$$F(M_e + b) - F(M_e - a) = 0,6$$

De plus,

$$\begin{cases} F(M_e + b) = 0,8 \\ F(M_e - a) = 0,2 \end{cases}$$

La lecture du tableau (8.1) donne deux informations :

$$\begin{cases} (M_e - a) \in [4, 8[ \\ (M_e + b) \in [16, 22[ \end{cases}$$

d'où

$$\begin{cases} M_e + b = 16 + 6 \frac{0,8 - 0,69}{0,86 - 0,69} = 19,88 \\ M_e - a = 4 + 4 \frac{0,2 - 0,16}{0,28 - 0,16} = 5,33 \end{cases}$$

Finalement  $P[5,33 \leq X < 19,88] = 0,6$

7.  $cv = \frac{s_X}{\bar{x}}$ .

On a

$$s_X^2 = \frac{1}{n} \sum_i n_i (c_i - \bar{x})^2$$

A partir des données du tableau, on a donc

$$s_X = \left( \frac{1}{100} 6814,5247 \right)^{\frac{1}{2}} = 8,255$$

et  $cv = 0,635$ .

$$Q_3 = 16 + 6 \frac{0,75 - 0,69}{0,86 - 0,69} = 18,12$$

Finalement,

ecart interquartile

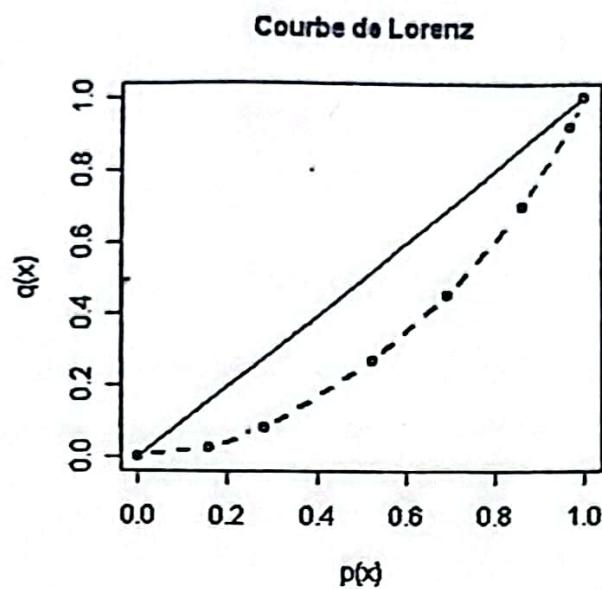
$$\frac{Q_3 - Q_1}{Q_2} = 0,95$$

8.

classe	$F(c_i)$	$n_i c_i$	$\frac{1}{n\bar{x}} \sum_i n_i c_i$
[0, 4[	0,16	32	0,0246
[4, 8[	0,28	72	0,0801
[8, 12[	0,52	240	0,2648
[12, 16[	0,69	238	0,4480
[16, 22[	0,86	323	0,6967
[22, 30[	0,97	286	0,9169
[30, 42[	1	108	1
Total		1299	

NB

*Représentation graphique :*



*On a*

$$G = 1 - 2 \sum_{i=1}^7 (p(e_i) - p(e_{i-1})) \frac{(q(e_i) + q(e_{i-1}))}{2}$$

*donc*

$$\begin{aligned} G = 1 - & (0,16 \cdot 0,0246 + 0,12 (0,0246 + 0,0801) + \\ & 0,24 (0,0801 + 0,2648) + 0,17 (0,2648 + 0,4480) + \\ & 0,17 (0,4480 + 0,6967) + 0,11 (0,6967 + 0,9169) + \\ & 0,03 \cdot 0,9169 + 1) \end{aligned}$$

*Finalement :*

$$G = 0,23687$$

*La distribution est par conséquent, moyennement inégalitaire G=20%.*

#### **Exercice 8.4 Examen de contrôle ESSAIT juin 2002**

*Un centre de loisirs pour enfants cherche à fixer les droits d'adhésion en fonction des revenus des parents sollicitant une inscription. Dans le cadre des travaux préliminaires, la Direction s'est intéressée aux revenus des parents des enfants déjà membres ( $x_i$ ).*

Ces revenus, exprimés en milliers de dinars, se répartissent comme suit :

$x_i$	$n_i$	$F(x_i)$
[5, 7[	-	0,04
[7, 11[	-	0,14
[11, 13[	-	0,44
[13, 15[	-	0,96
[15, 19[	-	1
Total	$n$	-

1. Sachant que  $s_X^2 = 4,93$ , que  $\sum_{i=1}^n f_i x_i^2 = 166,22$  et que

$$\sum_{i=1}^n n_i x_i = 1905, \text{ montrer que } n = 150 \text{ et compléter le tableau.}$$

2. Construire l'histogramme de la distribution.
3. Déterminer le mode et la médiane.
4. Calculer le coefficient d'asymétrie de Fisher (skewness). En déduire le sens d'obliquité de la distribution.
5. Tracer la courbe de Lorenz. et calculer l'indice de Gini.
6. Commenter.

### Corrigé

Comme à l'accoutumée, le tableau qui suit sera complété à mesure que les calculs seront effectués.

$x_i$	$c_i$	$n_i$	$F(x_i)$	$f_i$	$c_i - \bar{x}$	$(c_i - \bar{x})^3$	$f_i (c_i - \bar{x})^3$
[5, 7[	6	6	0,04	0,04	-6,7	-300,763	-12,031
[7, 11[	9	15	0,14	0,10	-3,7	-50,653	-5,065
[11, 13[	12	45	0,44	0,30	-0,7	-0,343	-0,108
[13, 15[	14	78	0,96	0,52	1,3	2,197	1,142
[15, 19[	17	6	1	0,04	4,3	79,507	3,180
Total	-	150	-	1			-12,909

$$1. n\bar{x} = \sum_{i=1}^n n_i x_i = 1905 \Rightarrow \bar{x} = \frac{1905}{n} \quad (\frac{\text{لـ}}{\text{X}} \text{ مـعـدـلـ})$$

$$s_X^2 = \sum_{i=1}^n f_i (x_i - \bar{x})^2 = \sum_{i=1}^n f_i x_i^2 - \bar{x}^2 = \sum_{i=1}^n f_i x_i^2 - \left( \frac{1}{n} \sum_{i=1}^n n_i x_i \right)^2$$

$$\Rightarrow \frac{1}{n^2} \left( \sum_{i=1}^n n_i x_i \right)^2 = \sum_{i=1}^n f_i x_i^2 - s_X^2$$

Conclusion

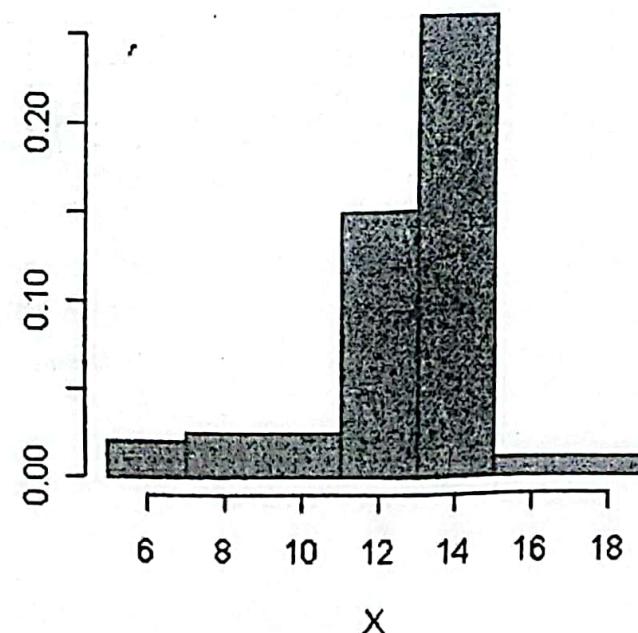
$$n^2 = \frac{\left( \sum_{i=1}^n n_i x_i \right)^2}{\sum_{i=1}^n f_i x_i^2 - s_X^2} = 22500 \quad \text{et} \quad n = 150$$

2. Les classes sont d'amplitudes inégales, on se doit alors de corriger les effectifs (ou les fréquences).

Amplitude choisie : 2. Les effectifs corrigés sont alors :

$$n_1 = 6 \quad n_2^c = 7,5 \quad n_3 = 45 \quad n_4 = 78 \quad n_5^c = 3$$

histogramme



3. La classe modale est [13, 15[. Déterminons la valeur du mode  $M_o$

$$M_o = l_1 + a \frac{d_1}{d_1 + d_2}$$

$$M_o = 13 + 2 \frac{78 - 45}{(78 - 45) + (78 - 3)} = 13,61$$

La classe médiane est aussi [13, 15[ et la médiane  $M_e$  est

$$M_e = 13 + 2 \frac{0,5 - 0,44}{0,96 - 0,44} = 13,23$$

4.

$$\gamma_1 = \frac{\frac{1}{n} \sum_{i=1}^n n_i (c_i - \bar{x})^3}{s_X^3} = \frac{\sum_{i=1}^n f_i (c_i - \bar{x})^3}{s_X^3}$$

$$= -\frac{12,909}{(4,93)^{\frac{3}{2}}} = -1,179 < 0$$

## Cours de Statistique Descriptive

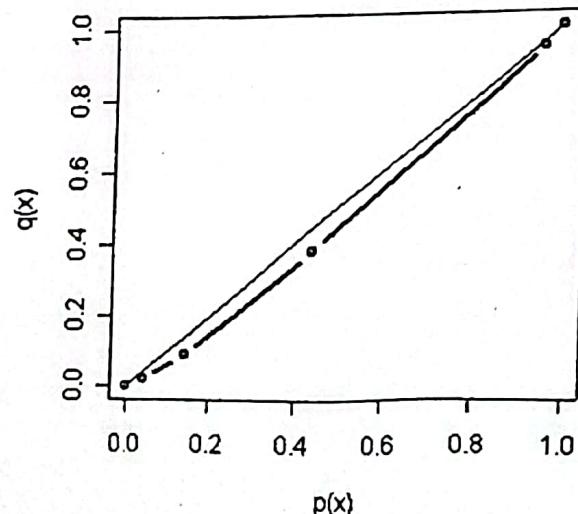
La distribution est donc oblique à droite (plus étalée sur la gauche); ceci est nettement visible sur l'histogramme. En effet, les petites valeurs sont plus dispersées.

5.

$x_i$	$c_i$	$n_i$	$F(x_i) = p(e_i)$	$n_i c_i$	$\sum_{i=1}^k n_i c_i$	$\frac{\sum_{i=1}^k n_i c_i}{n \bar{x}} = q(e_i)$
[5, 7[	6	6	0,04	36	36	0,019
[7, 11[	9	15	0,14	135	171	0,090
[11, 13[	12	45	0,44	540	711	0,373
[13, 15[	14	78	0,96	1092	1803	0,946
[15, 19[	17	6	1	102	1905	1
Total	-	150	-	1905	-	-

(a)

Courbe de Lorenz



$$(b) G = 2 \times \left( \frac{1}{2} - \sum S_i \right)$$

$S_i$  étant la surface du trapèze de la classe  $i$ .

$$G = 1 - 2 \sum S_i$$

$$= 1 - 2 \sum (p(e_i) - p(e_{i-1})) \frac{(q(e_i) + q(e_{i-1}))}{2}$$

$$= 1 - \left( (0,04 \cdot 0,019) + (0,1 \cdot 0,109) + (0,3 \cdot 0,463) \right)$$

$$= 0,086$$

6. La répartition des revenus des parents membres paraît peu inégalitaire, il semble alors inutile d'établir une grille tarifaire tenant compte des revenus.

**Exercice 8.5** Les données sur le chiffre d'affaires journalier en haute saison des hôtels de la région balnéaire Bella Playa ayant été mal archivées, certaines informations se sont effacées. On ne dispose que du tableau suivant :

Chiffre d'affaires en $10^3$ euros	$f_i\%$
[2,6]	8
[6,9]	19
[9, $e_3$ ]	$f_3$
[ $e_3$ , 16]	19
[16, 20]	16
[20, 80]	11

- La moyenne  $\bar{x}$  des chiffres d'affaires est égale à  $15,39$  ( $\text{en } 10^3 \text{ €}$ ), déterminer  $f_3$  et vérifier que  $e_3 = 11$ .
- Déterminer la valeur du chiffre d'affaires modal.
- Calculer le chiffre d'affaires médian et commenter le résultat.
- On s'intéresse à l'inégalité du chiffre d'affaires des hôteliers dans cette ville balnéaire. Après avoir représenté la courbe de Lorenz, Calculer l'indice de Gini et commenter les résultats.
- Proposer un indice de concentration et le calculer.

**Corrigé :**

$$1. f_3 = 100 - (8 + 19 + 19 + 16 + 11) = 27$$

$$\bar{x} = 15.39 \cdot 10^3 \text{ €}$$

$$\bar{x} = \frac{1}{100} \sum_{i=1}^{10} f_i \left( \frac{e_{i-1} + e_i}{2} \right) \Rightarrow e_3 = 11.$$

2. Tableau récapitulatif des calculs :

C.A $10^3$ €	$f_i\%$	$cif_i$	$f_i^c$	$p(e_i)$	$\sum cif_i$	$q(e_i)$	$\ln \frac{cif_i}{n\bar{x}}$	$\frac{cif_i}{n\bar{x}} \ln \frac{cif_i}{n\bar{x}}$
[2,6]	8	0.32		0.08	0.32	0.021	-3.87	-0.080
[6,9]	19	1.425	12.67	0.27	1.745	0.113	-2.98	-0.220
[9, 11]	27	2.7	27	0.54	4.445	0.289	-1.74	-0.307
[11, 16]	19	2.375	7.6	0.73	6.82	0.443	-1.87	-0.289
[16, 20]	16	2.88	8	0.89	9.7	0.630	-1.68	-0.314
[20, 80]	11	5.5	0.37	1	15.39	1	-1.03	-0.368
Total	100	15.39	-		-	-	-	-1.578

Calcul de la fréquence moyenne par unité de classe ou fréquence corrigée :

On prendra comme amplitude de référence  $a = 2$ .

La classe modale est  $[9, 11]$ .

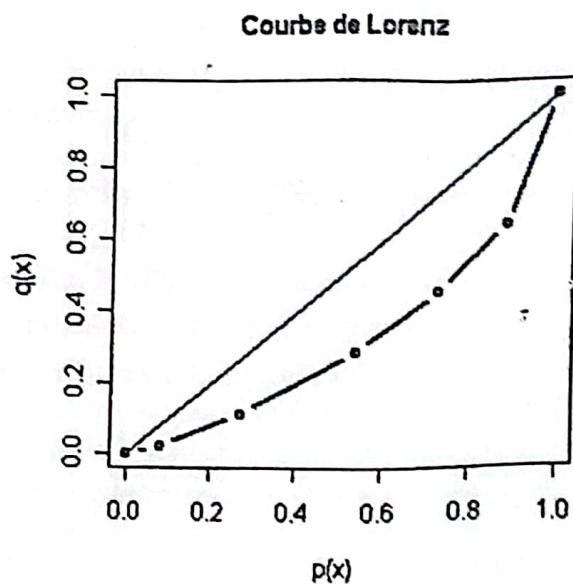
$$M_o = 9 + 2 \frac{27 - 12.67}{(27 - 12.67) + (27 - 7.6)} = 9.85 \cdot 10^3 \text{ €}$$

3. La classe médiane est [9, 11].

$$M_e = 9 + 2 \frac{0.5 - 0.27}{0.54 - 0.27} = 10.7 \text{ } 10^3 \text{ €}$$

50% des hôtels de Bella Playa ont un chiffre d'affaires inférieur à  $10.7 \text{ } 10^3 \text{ €}$ . comme le chiffre moyen vaut  $15.39 \text{ } 10^3 \text{ €}$  et que la distribution est unimodale, alors nécessairement, la distribution est étalée sur la droite.

4.



$$G = 1 - \sum_{i=1}^6 (p(e_i) - p(e_{i-1})) (q(e_i) - q(e_{i-1}))$$

$$G = 1 - \left( 0.08 \cdot 0.021 + 0.19 \cdot 0.134 + 0.27 \cdot 0.402 + 0.19 \cdot 0.732 + 0.16 \cdot 1.073 + 0.11 \cdot 1.63 \right)$$

$$G = 0.374$$

L'indice de Gini est assez élevé; il dénote d'une inégalité assez importante au niveau du chiffre d'affaires journalier. Ceci était d'ailleurs lisible sur la courbe de Lorenz.

5. On peut proposer par exemple, l'indice de Theil défini par

$$T = \sum_{i=1}^{10} \frac{c_i f_i}{\sum_{i=1}^{10} c_i f_i} \ln \frac{c_i f_i}{\sum_{i=1}^{10} c_i f_i}$$

A partir du tableau précédent, on obtient  $T = -1.578$ .

#### Exercice 8.6 Examen principal ESSAIT janvier 2002

Les dirigeants d'un groupe industriel ont décidé de réduire la dispersion des salaires. Au bout de cinq années de politique interne visant cet objectif,

les résultats sont les suivants :

salaire annuel	nombre de salariés à l'année t	nombre de salariés à l'année t+5
[2200, 2800[	1369	208
[2800, 4400[	4182	820
[4400, 5200[	1957	895
[5200, 5800[	1394	1325
[5800, 7200[	1650	2774
[7200, 8800[	806	2180
[8800, 14000[	1010	3426
[14000, 20000[	272	830
[20000, 35000[	160	542
Total	12.800	13.000

1. Tracer la courbe de Lorenz des salaires pour les deux années considérées (arrondir les calculs à  $10^{-2}$  près).
2. Calculer pour les deux années, les trois quartiles.
3. Interpréter les résultats obtenus.
4. Déterminer l'indice de Gini pour l'année t et l'année t+5.
5. Est-ce que l'objectif est atteint ?

#### Corrigé

1. On complétera le tableau de la page suivante.

Les courbes de Lorenz suivront.

- Pour l'année t, on a la classe médiane [4400, 5200[

$$M_e = 4400 + 800 \times \frac{0,5 - 0,43}{0,59 - 0,43} = 4750$$

les quartiles :

$$Q_{0,25} = 2800 + 1600 \times \frac{0,25 - 0,11}{0,43 - 0,11} = 3500$$

$$Q_{0,75} = 5800 + 1400 \times \frac{0,75 - 0,7}{0,82 - 0,7} = 6383$$

- Pour l'année t+5, la classe médiane est [4400, 5200[

$$M_e = 7200 + 1600 \times \frac{0,5 - 0,46}{0,63 - 0,46} = 7576$$

- les quartiles :

$$Q_{0,25} = 5800$$

$$Q_{0,75} = 8800 + 5200 \times \frac{0,75 - 0,63}{0,89 - 0,63} = 11200$$

Cours de Statistique Descriptive

Tableau des calculs intermédiaires pour la représentation  
des courbes de Lorenz

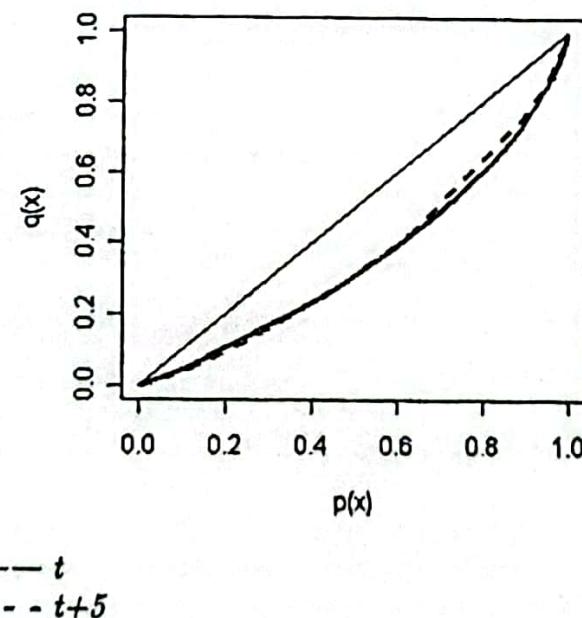
Salaire annuel	$c_i$	$n_{it}$	$\sum_i n_{it}$	$p_t$	$c_i p_t$	$\sum_i c_i p_t$	$q_t$	$n_{i+5}$	$p_{t+5}$	$c_i p_{t+5}$	$\sum_i c_i p_{t+5}$	$q_{t+5}$
[2200, 2800]	2500	1369	1369	0,11	34222500	34222500	0,05	208	0,02	520000	520000	0,004
[2800, 4400]	3600	4182	5551	0,43	15055200	18477700	0,25	820	0,08	2952000	3472000	0,03
[4400, 5200]	4800	1957	7508	0,59	9393600	27871900	0,38	895	0,15	4296000	7768000	0,07
[5200, 5800]	5500	1894	8902	0,70	7667000.	35538300	0,49	1325	0,25	7287500	15055500	0,13
[5800, 7200]	6500	1650	10552	0,82	10725000	46263300	0,63	2774	0,46	18031000	33086500	0,28
[7200, 8800]	8000	806	1358	0,89	6448000	52711300	0,72	2180	0,63	17440000	50525500	0,43
[8800, 14000]	11400	1010	12368	0,97	11514000	64225900	0,88	3426	0,89	39056400	89582900	0,76
[14000, 20000]	17000	272	12640	0,99	4624000	68849300	0,94	830	0,96	14110000	103692900	0,87
[20000, 35000]	27500	160	12800	1	4400000	73249300	1	542	1	14905000	118597900	1
<b>Total</b>	-	12800	-	-	73249300	-	-	13000	-	118597900	-	-

2. Durant l'année  $t$ , 50% des salariés recevaient un salaire inférieur à 4750; et donc 50% des salariés gagnaient moins que 83% du salaire moyen. Ce plafond atteint 7576 à l'année  $t+5$ , ce qui correspond encore à 83% du salaire moyen. Par ailleurs, 25% des salariés ne dépassaient pas 3500 soit 61% du salaire moyen ( $\frac{63149300}{12800} = 5723$ ) alors qu'à l'année  $t+5$ , ils atteignent 5800 ce montant correspondant à 64% du salaire moyen. Parallèlement, 25% des salariés gagnaient plus que 6383 soit 11,5% de plus que le salaire moyen alors qu'à  $t+5$ , ils gagnent 11200 soit 23% de plus que le salaire moyen.

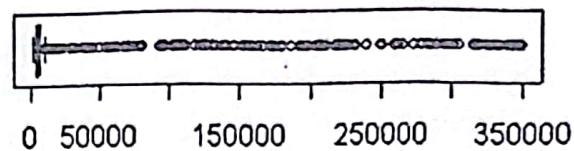
Ainsi, le niveau général des salaires a augmenté mais dans des proportions plus inégalitaires au niveau des bas salaires.

### Représentation des deux courbes de Lorenz

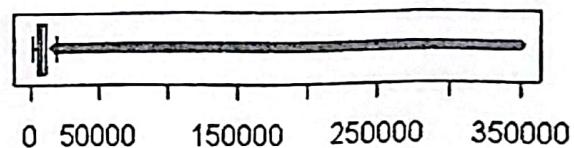
**Courbes de Lorenz des salaires annuels**



Distribution des salaires à l'année t



Distribution des salaires à l'année t+5



La comparaison des deux box-plot le confirme bien. La forte concentration des bas salaires demeuree toujours aussi nette : la "moustache" inférieure est bien plus courte que la moustache supérieure. et la longueur de la boîte reste dans les deux cas, très faible.

### 3. Indice de Gini pour les années t et t+5 :

$$\begin{aligned} G &= 1 - 2 \sum_i \frac{q_i + q_{i-1}}{2} (p_i - p_{i-1}) \\ &= 1 - \sum_i (q_i + q_{i-1}) (p_i - p_{i-1}) \end{aligned}$$

- Pour l'année t

$$\begin{aligned} G_t &= 1 - \left( (0,05 \times 0,11) + (0,30 \times 0,32) + (0,59 \times 0,16) \right. \\ &\quad \left. + (0,87 \times 0,11) + (0,12 \times 1,12) + (1,35 \times 0,07) \right. \\ &\quad \left. + (1,6 \times 0,08) + (1,82 \times 0,02) + (1,94 \times 0,01) \right) \\ G_t &= 0,2957 \end{aligned}$$

- Pour l'année t+5

$$\begin{aligned} G_{t+5} &= 1 - \left( (0,02 \times 0,004) + (0,06 \times 0,034) + (0,07 \times 0,1) \right. \\ &\quad \left. + (0,1 \times 0,2) + (0,21 \times 0,41) + (0,17 \times 0,71) \right. \\ &\quad \left. + (0,23 \times 1,19) + (0,07 \times 1,63) + (0,04 \times 1,87) \right) \\ G_{t+5} &= 0,2658 \end{aligned}$$

### 4. $G_{t+5} < G_t$

La nouvelle répartition des salaires n'est donc pas beaucoup moins inégalitaire que la première. La politique adoptée par les dirigeants de

*l'entreprise ne leur a pas permis d'atteindre leur objectif. Plus précisément,*

*Cette politique n'a pas réduit les inégalités au niveau des salaires les plus bas ; le peu de changement dans la répartition des salaires concerne les grands salaires. Nous avons une seule certitude : cette politique a augmenté le niveau général des salaires.*

### Exercice 8.7 Examen principal ESSAIT janvier 2002

Lors d'un concours d'entrée à une école de commerce, 55 étudiants ont obtenu en mathématiques ( $x$ ) et en statistique ( $y$ ) les notes résumées dans le tableau qui suit :

$x$	$y$	[9, 11[	[11, 13[	[13, 15[	[15, 17[	
[9, 11[		10	1			11
[11, 13[		3	12	1		16
[13, 15[			2	9	2	13
[15, 17[			1	3	11	16
		11				

1. Calculer les paramètres des deux droites exprimant une liaison linéaire entre les deux variables  $x$  et  $y$ .
2. Laquelle des deux droites choisir ? Expliquer.
3. Qualifier la liaison entre les deux variables.

### Corrigé

1.

$x$	$y$	[9, 11[	[11, 13[	[13, 15[	[15, 17[	$n_{i,j}$	$n_{i,j}c_i$	$n_{i,j}c_i^2$
[9, 11[	10	1				11	110	1100
[11, 13[	3	12	1			16	192	2304
[13, 15[		2	9	2		13	182	2548
[15, 17[		1	3	11		15	240	3840
$n_{i,j}$	13	16	13	13		55		
$n_{i,j}c_j$	130	192	182	208				
$n_{i,j}c_j^2$	1300	2304	2548	3328				

Le tableau sera complété au fur et à mesure que les calculs avanceront.

$$\sum_i n_{i,j} c_i = 724 \implies \bar{x} = 13,16 \quad \sum_i n_{i,j} c_i^2 = 9792$$

$$\sum_j n_{i,j} c_j = 712 \implies \bar{y} = 12,94 \quad \sum_j n_{i,j} c_j^2 = 9480$$

$\Delta_1$  : régression de  $y$  sur  $x$

$$y = \alpha x + \beta$$

$$\hat{y} = \hat{\alpha}x + \hat{\beta}$$

$\Delta_2$  : régression de  $x$  sur  $y$

$$x = \alpha'y + \beta'$$

$$\hat{x} = \hat{\alpha}'y + \hat{\beta}'$$

$$\text{On a } \hat{\alpha} = \frac{s_{XY}}{s_X^2} \quad \text{et} \quad \hat{\alpha}' = \frac{s_{XY}}{s_Y^2}$$

$$s_X^2 = \frac{1}{55} 9792 - (13, 16)^2 = 4,85$$

$$s_Y^2 = \frac{1}{55} 9480 - (12, 94)^2 = 4,92$$

A partir des données du tableau, on obtient :  $\sum_i \sum_j n_{ij} c_i c_j = 9604$   
d'où la covariance

$$s_{XY} = \frac{1}{55} 9604 - (13, 16)(12, 94) = 4,33$$

et finalement

$$\hat{\alpha} = \frac{4,33}{4,85} = 0,892 \quad \text{et} \quad \hat{\alpha}' = \frac{4,33}{4,92} = 0,874$$

On détermine alors les constantes :

$$\hat{\beta} = 12,94 - (0,892)(13, 16) = 1,201$$

$$\hat{\beta}' = 13,16 - (0,874)(12, 94) = 1,850$$

de là les équations des droites :

$$\hat{y} = 0,892 x + 1,201 \quad \hat{x} = 0,874 y + 1,850$$

## 2. Analyse de la variance du modèle de régression :

On choisira le modèle qui explique mieux la variable endogène (la part de variance expliquée dans la variance totale la plus élevée).

variance totale = variance expliquée + variance résiduelle

$$s_Y^2 = (0,892)^2 s_X^2 = 3,859$$

$$s_{\hat{X}}^2 = (0,874)^2 s_Y^2 = 3,758$$

$$\left. \begin{aligned} \frac{s_{\hat{Y}}^2}{s_Y^2} &= \frac{3,859}{4,920} = 0,784 \\ \frac{s_{\hat{X}}^2}{s_X^2} &= \frac{3,758}{4,850} = 0,775 < 0,784 \end{aligned} \right\}$$

Ainsi, la première régression permet de mieux expliquer la variance.

3.

$$r = \frac{s_{XY}}{s_X s_Y} = \frac{4,33}{\sqrt{(4,85)(4,92)}} = 0,886 > 0,8$$

Il y a donc une bonne corrélation linéaire entre les deux variables.

**Exercice 8.8 Examen de contrôle ESSAIT juin 2002**

*Les données sur la consommation et la production de l'Etat d'Imagineland (en millions de pièces d'or) se résument comme suit :*

	1995	1997	1998	1999	2001
PIB	197,41	192,78	203,33	177,46	174,31
CI	176,29	170,61	177,91	157,41	150,78

- Quel coefficient peut permettre d'affirmer l'existence d'une relation linéaire exprimant la consommation totale (CI) en fonction du PIB ?
- Donner l'expression de cette relation linéaire.
- L'indice élémentaire de la consommation totale pour l'année 2003, base 100 en 1999 étant estimé à 111,3, donner la valeur prévisible du PIB en 2003.

**Corrigé**

Nous désignerons par  $P$ , le PIB et par  $C$  la consommation finale.

Nous complèterons le tableau qui suit parallèlement à l'avancement du travail.

	1995	1997	1998	1999	2001	Total
PIB	197,41	192,78	203,33	177,46	174,31	945,25
CI	176,29	170,61	177,91	157,41	150,78	833,00
$P - \bar{P}$	8,352	3,722	14,272	-11,598	-14,748	-
$C - \bar{C}$	9,69	4,01	11,31	-9,19	-15,82	-
$(P - \bar{P}) \cdot (C - \bar{C})$	80,931	14,925	161,416	106,586	233,313	597,171
$(P - \bar{P})^2$	69,756	13,853	203,69	134,514	217,504	639,317
$(C - \bar{C})^2$	93,90	16,08	127,916	84,456	250,272	572,624

- Il s'agit de  $r^2$ , le coefficient de corrélation linéaire empirique défini par

$$r^2 = \frac{s_{XY}}{s_x s_y} = \frac{\sum_{i=1}^n (P_i - \bar{P}) \cdot (C_i - \bar{C})}{\sqrt{\sum_{i=1}^n (P_i - \bar{P})^2} \sqrt{\sum_{i=1}^n (C_i - \bar{C})^2}}$$

Si ce coefficient est proche de 1 (supérieur à 0,8) en valeur absolue, alors nous pouvons conclure quant à l'existence d'une relation linéaire.

Il faut toutefois préciser que cette conclusion n'est valide que lorsque le nuage de points laisse supposer une relation linéaire.

On a  $\bar{P} = 189,058$  et  $\bar{C} = 166,60$  et donc

$$r^2 = \frac{597,171}{\sqrt{639,317} \sqrt{572,624}} = 0,987$$

2. La droite de régression a pour équation

$$\hat{C} = \hat{\alpha}P + \hat{\beta}$$

avec

$$\hat{\alpha} = \frac{\sum_{i=1}^n (P_i - \bar{P}) \cdot (C_i - \bar{C})}{\sum_{i=1}^n (P_i - \bar{P})^2} = \frac{597,171}{639,317} = 0,934$$

et

$$\hat{\beta} = \bar{C} - \hat{\alpha}\bar{P} = 166,60 - 0,934 \times 189,058 = -9,98$$

$$3. I_{2003/1999}(CT) = 111,3 \implies \widetilde{CT}_{2003} = \frac{111,3 \times 157,41}{100} = 175,197$$

d'où

$$\widetilde{PIB}_{2003} = \frac{175,197 + 9,98}{0,934} = 199,09.$$

### Exercice 8.9 Examen principal ESSAIT janvier 2002

Le recensement de la population tunisienne sur la période 1966-1994 se résume comme suit :

Année	1966	1975	1980	1985	1994
Nombre d'habitants	4.533.351	5.588.209	6.369.000	6.966.173	8.785.700

1. Donner les taux d'accroissement annuels moyens pour chaque sous-période. En déduire le taux d'accroissement annuel moyen sur la période 1966-1994.
2. En supposant que le taux d'accroissement moyen de la population s'est stabilisé à partir de 1985, donner une valeur estimant la population en 2010.
3. Donner l'équation de la droite de régression de  $\ln y$  sur  $t$  où  $y$  et  $t$  désignent respectivement le nombre d'habitants et l'année. A quoi correspond la pente de cette droite ?

Corrigé :

1. Taux d'accroissement annuels moyens par sous-période

- 1966-1975

$$5.588.209 = (1 + g_1)^9 4.533.351 \text{ d'où } g_1 = 2,35\%$$

- 1975-1980

$$6.369.000 = (1 + g_2)^5 5.588.209 \text{ d'où } g_2 = 2,65\%$$

- 1980-1985

$$6.966.173 = (1 + g_3)^5 6.369.000 \text{ d'où } g_3 = 1,81\%$$

- 1985-1994

$$8.785.700 = (1 + g_4)^9 6.966.173 \text{ d'où } g_4 = 2,61\%$$

On en déduit alors, le taux d'accroissement annuel moyen :

$$g = \sqrt[28]{(1,0235)^9 (1,0265)^5 (1,0181)^5 (1,0261)^9} - 1 = 2,39\%$$

2.

$$P_{2010} = 8.785.700 (1,0261)^{16} = 13.268.181$$

3. La droite de régression par la méthode des moindres carrés s'obtient à partir du tableau suivant

$t$	1966	1975	1980	1985	1994
$\ln y$	15,327	15,536	15,667	15,757	15,989

Elle s'écrit :  $\widehat{\ln y} = \widehat{\alpha}t + \widehat{\beta}$

$$\text{avec } \widehat{\alpha} = \frac{s_{t \ln y}}{s_t^2} = \frac{154996,853 - 1980 \cdot 78,276}{19602442 - 19602000} = \frac{10,373}{442} = 0,0235$$

$$\text{et } \widehat{\beta} = \bar{y} - 0,0235\bar{t} = -30,875.$$

On obtient alors l'équation de la droite de régression de  $\ln y$  sur  $t$

$$\widehat{\ln y} = 0,0235t - 30,875$$

Partant de la droite  $\ln y = \alpha t + \beta$ , on a

$$\frac{\Delta \ln y}{\Delta t} = \alpha \Leftrightarrow \frac{\Delta y}{y} = \alpha \Delta t$$

$\alpha$  est donc le taux d'accroissement annuel moyen de  $y$ .

On aura remarqué que  $\widehat{\alpha} = 2,35\%$  alors que  $g = 2,39\%$ . Cette petite différence s'explique par le fait que la droite des moindres carrés ordinaires ne passe pas nécessairement par le premier et le dernier point de la série d'observations.

### Exercice 8.10 Examen de contrôle ESSAIT juin 2002

Le tableau ci-dessous donne la valeur en dinars courants ainsi que le prix de l'énergie consommée en 1970 et en 1990 dans une région Lambda.

Energie	Valeur en 1970	Valeur en 1990	Prix en 1970	Prix en 1990
charbon	8001	20088	105	310
gaz naturel	651	8512	62	320
pétrole	9522	73406	92	578
électricité	6148	23214	265	212

1. Calculer les indices élémentaires du prix des différentes sources d'énergie en 1990, base 100 en 1970.
2. Calculer pour 1990, base 100 en 1970, l'indice de prix de l'énergie
  - (a) selon Laspeyres

4. Déduire l'équation de la droite de régression de  $x$  sur  $y$ .
5. Laquelle des deux droites choisir ? (justifier)
6. Quel serait le pourcentage de cadres dont le salaire net mensuel est supérieur 740,6 dinars ?

**Corrigé****- Partie 1**

*Le tableau sera complété au fur et à mesure.*

salaires	[6,8[	[8,10[	[10,15[	[15,20[	[20,35[	[35,50[	[50,80[
$f$ en %	19.7	24.5	27.7	12.9	10.1	2.6	2.5
$X_i$	137.9	220.5	346.25	225.75	277.75	110.5	162.5
$f^c$	0.197	0.245	0.11	0.0516	0.0135	0.0035	0.0017
$p(e_i)$	0.197	0.442	0.719	0.848	0.949	0.975	1
$q(e_i)$	0.093	0.242	0.476	0.628	0.816	0.890	1

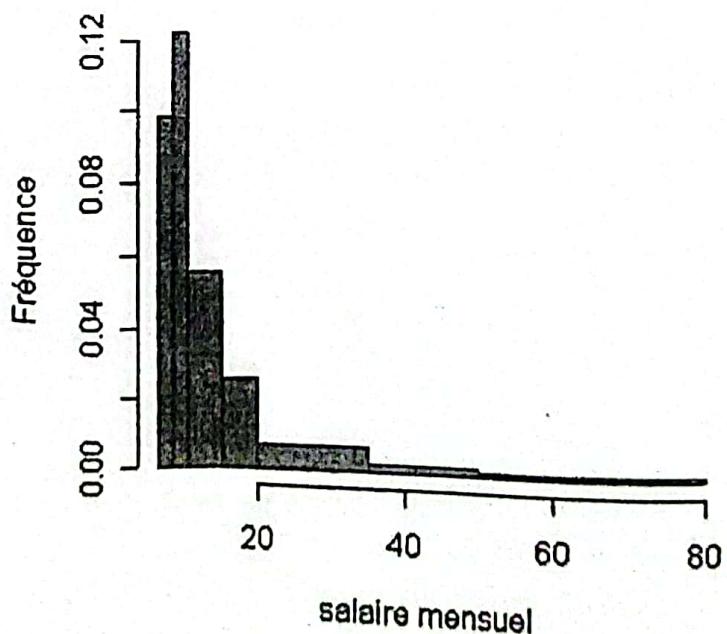
1. On choisit la valeur 2 pour l'amplitude de référence.

*La classe modale est [8, 10[.*

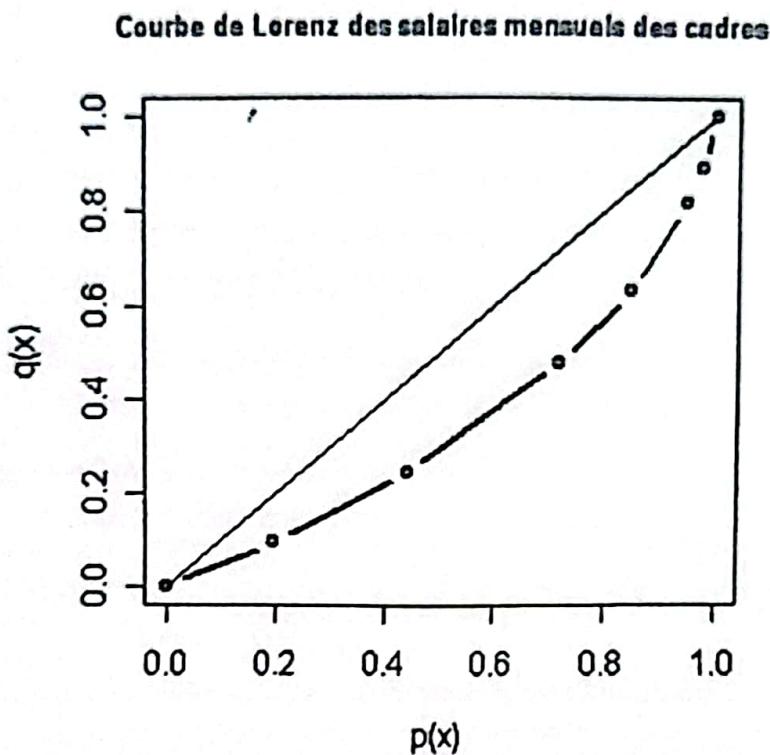
$$M_o = 8 + 2 \frac{0.245 - 0.197}{(0.245 - 0.197) + (0.245 - 0.11)} = 8.52 \cdot 10^2 \text{ dinars.}$$

2. Histogramme :

Répartition des cadres selon le salaire mensuel



## 3. Courbe de Lorenz



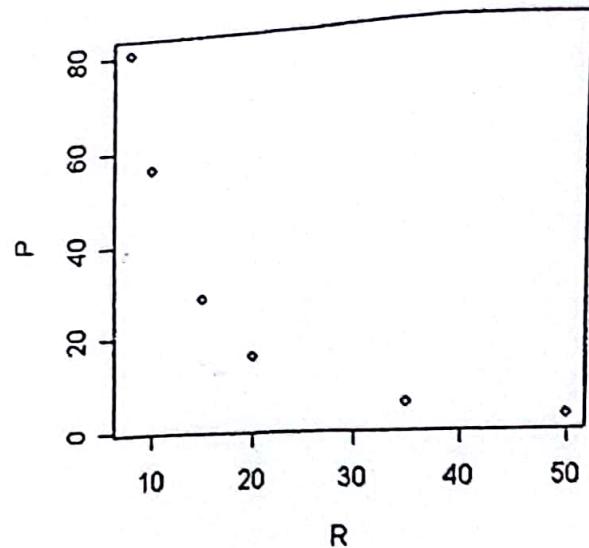
$$G = 1 - \sum_{i=1}^7 (p(e_i) - p(e_{i-1})) (q(e_i) + q(e_{i-1})) = 0.320$$

4. L'indice de Gini est assez élevé et ceci révèle une forte inégalité dans la répartition des salaires des cadres de l'entreprise, ce qui conforte l'opinion que nous fournit la courbe de Lorenz.

Cours de Statistique Descriptive

- Partie 2

1.



Le nuage de points dessine une évolution plutôt exponentielle que linéaire.

$$2. r_{XY} = \frac{s_{XY}}{s_X \cdot s_Y} = \frac{\frac{1}{n} \sum x_i y_i - \bar{x} \bar{y}}{\sqrt{\left( \frac{1}{n} \sum x_i^2 - \bar{x}^2 \right) \left( \frac{1}{n} \sum y_i^2 - \bar{y}^2 \right)}} = 0.999 \simeq 1$$

Le coefficient de corrélation linéaire est presque égal à 1. Il existe donc une relation quasi linéaire entre les variables  $x$  et  $y$ .

3. On a  $\hat{y} = \hat{\alpha}x + \hat{\beta}$

$$s_X^2 = \frac{1}{6} (53.878) - \frac{1}{36} (17.553)^2 = 0.421$$

$$s_Y^2 = \frac{1}{6} (57.436) - \frac{1}{36} (17.010)^2 = 1.535$$

$$\hat{\alpha} = \frac{s_{XY}}{s_X^2} = \frac{\frac{1}{n} \sum x_i y_i - \bar{x}}{\frac{1}{n} \sum x_i^2 - \bar{x}^2} = -1.908$$

$$\hat{\beta} = \bar{y} - \hat{\alpha}\bar{x} = 8.417$$

d'où l'équation de la droite des mco :

$$\hat{y} = -1.908x + 8.417$$

$$4. \hat{\alpha}\hat{\alpha}' = r_{XY}^2 \Rightarrow \hat{\alpha}' = \frac{0.999^2}{-1.908} = -0.523$$

$$\hat{\beta}' = \bar{x} - \hat{\alpha}'\bar{y} = 4.408$$

L'équation de la droite de régression de  $x$  sur  $y$  s'écrit alors :

$$\hat{x} = -0.523y + 4.408$$

5. On choisira la droite qui explique mieux le modèle (celle qui permet de mieux expliquer la variance).

Première droite :

$$\frac{\text{variance expliquée}}{\text{variance totale}} = \frac{s_{\hat{Y}}^2}{s_Y^2} = \frac{\hat{\alpha}^2 s_X^2}{s_Y^2} = 0.998$$

Seconde droite :

$$\frac{\text{variance expliquée}}{\text{variance totale}} = \frac{s_{\hat{X}}^2}{s_X^2} = \frac{\hat{\alpha}'^2 s_Y^2}{s_X^2} = 0.997$$

A peu de choses près, les deux modèles se valent. On peut donc aussi bien procéder à la régression de  $y$  sur  $x$  ou de  $x$  sur  $y$  et ceci s'explique par la relation presque linéaire entre  $x$  et  $y$ .

6.  $r = 7.406 \cdot 10^2$  dinars  $\Rightarrow x = 2.002 \Rightarrow \hat{y} = 4.597$

d'où  $\hat{p} = \exp \hat{y} = \exp(4.597) = 99.19$ .

### Exercice 8.13 Examen principal ESSAIT novembre 2003

La répartition d'un groupe d'entreprises selon la valeur de leurs exportations annuelles en milliers de dinars ( $X$ ) se résume comme suit :

$X$	[200, 400[	[400, 600[	[600, 800[	[800, 1200[	[1200, 1800[	Total
$n_i$	4	22	63	23	8	120
$f_i$	0,033	0,183	0,525	0,192	0,067	1

On donne  $\sum_i n_i c_i^2 = 77.730.000$  ( $c_i$  étant le centre de la  $i$ -ème classe).

Les deux parties peuvent être traitées indépendamment.

Première partie

1. Calculer la valeur moyenne des exportations par entreprise.
2. Calculer l'écart type et le coefficient de variation. Commenter.
3. Représenter la courbe de Lorenz.
4. Calculer l'indice de Gini.
5. Commenter.

Deuxième partie

On donne maintenant la répartition du groupe d'entreprises simultanément en fonction de la valeur des exportations ( $X$ ) et du nombre de pays destinataires ( $Y$ ). On dispose des informations supplémentaires suivantes :

$$\bar{y} = 3,325 \quad \sum_j n_{ij} y_j^2 = 1463$$

$$\sum_i \sum_j n_{ij} c_i y_j = 327.800 \quad \sum_i \sum_j \frac{n_{ij}^2}{n_{i+} n_{o+j}} = 2,4436 \quad (c_i \text{ étant le centre}$$

Cours de Statistique Descriptive  
de la  $i$ -ème classe de la variable  $X$ ).

$X$	$Y$	1	2	3	4	5	7	Total
[200, 400[	3	1						4
[400, 600[	1	14	6	1				22
[600, 800[	5		34	23	1			63
[800, 1200[		6		10	7			23
[1200, 1800[				3	3	2		8
Total		4	20	46	37	11	2	120

1. Peut-on conclure l'indépendance des variables  $X$  et  $Y$  ?
2. Peut-on conclure une liaison fonctionnelle ? Justifier.
3. Calculer la covariance de  $X$  et  $Y$ .
4. Calculer la distance du  $\chi^2$  entre  $X$  et  $Y$ . Interpréter.

### Corrigé

#### Première partie

$$1. \bar{x} = \frac{1}{n} \sum_i n_i c_i = \frac{1}{120} 91300 = 760,83$$

$$2. s_X^2 = \frac{1}{n} \sum_i n_i c_i^2 - \bar{x}^2 = \frac{77730000}{120} - (760,83)^2 = 68887,71$$

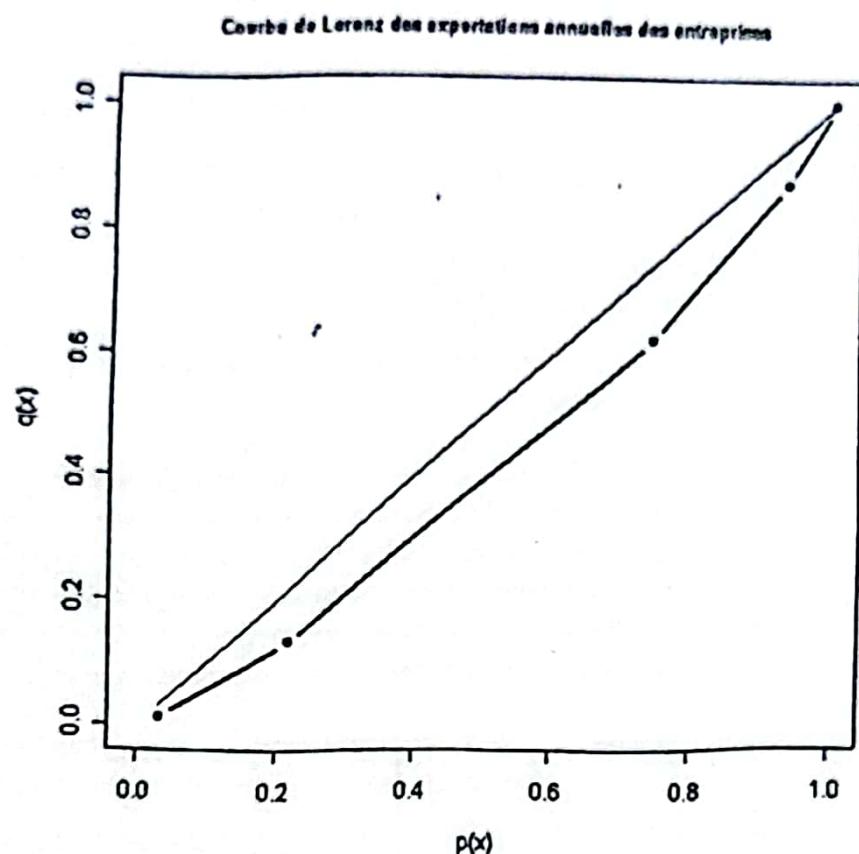
$$s_X = 262,465$$

$$cv = \frac{s_X}{\bar{x}} = 0,345.$$

Le coefficient de variation est assez faible. La distribution est donc peu dispersée et modérément inégalitaire.

3.

$p(x) = F(x)$	0,033	0,216	0,741	0,933	1
$n_i c_i$	1200	11000	44100	23000	12000
$\sum_i n_i c_i$	1200	12200	56300	79300	91300
$q(x) = \frac{\sum_i n_i c_i}{n \bar{x}}$	0,013	0,134	0,617	0,869	1



$$4. G = 1 - \sum_i (p(e_i) - p(e_{i-1})) (q(e_i) + q(e_{i-1}))$$

$$\begin{aligned} G &= 1 - (0,013 \cdot 0,033) + (0,216 - 0,033)(0,134 + 0,013) \\ &\quad + (0,741 - 0,216)(0,617 + 0,134) + (0,933 - 0,741)(0,869 + 0,617) \\ &\quad + (1 - 0,933)(1 + 0,869) \end{aligned}$$

$$G = 0,168$$

5. La courbe de Lorenz ainsi que l'indice de Gini confirment le résultat trouvé à travers le coefficient de variation, à savoir une inégalité peu importante.

### Seconde partie

1. Les deux variables ne sont pas indépendantes. En effet,

$$f_{11} = \frac{3}{120} \quad f_{1\bullet} = \frac{4}{120} \quad \text{et} \quad f_{\bullet 1} = \frac{4}{120} \quad f_{11} \neq \frac{16}{(120)^2}.$$

2. Il n'existe pas de liaison fonctionnelle entre les deux variables puisque dans le tableau de contingence, on a

$$n_{\bullet 1} = n_{11} + n_{21} \quad (\text{avec } n_{11} = 3 \neq 0 \text{ et } n_{21} = 1 \neq 0)$$

Y n'est donc pas lié fonctionnellement à X.

$$\text{Aussi, } n_{1\bullet} = n_{11} + n_{12} \quad (\text{avec } n_{11} = 3 \neq 0 \text{ et } n_{12} = 1 \neq 0)$$

X n'est donc pas lié fonctionnellement à Y.

$$3. s_{XY} = \frac{1}{n} \sum_i \sum_j n_{ij} c_i y_j - \bar{xy}$$

$$s_{XY} = \frac{1}{120} 327800 - 760,83 \cdot 3,325 = 201,907.$$

$$4. D^2 = n \left( \sum_i \sum_j \frac{n_{ij}^2}{n_{i\bullet} n_{\bullet j}} - 1 \right)$$

$$D^2 = 120(2,4436 - 1) = 173,23.$$

Calculons la borne supérieure de  $D^2$ .

$$D^2 \leq \inf(120(6-1), 120(5-1)) = 480.$$

Tout ce que l'on pourra dire avec certitude est que  $X$  et  $Y$  ne sont pas indépendantes.

#### Exercice 8.14 Examen de contrôle ESSAIT juin 2005

Avant de démarrer le projet d'un parc d'attraction dans une petite ville, le service marketing de la société "Manèges et Cie" a préféré procéder à une enquête sur les dépenses mensuelles des ménages pour les loisirs  $L$  (en dollars). L'enquête a été effectuée auprès de 200 ménages répartis selon leur taille  $T$ . Les résultats de l'enquête sont consignés dans le tableau qui suit :

$L$	$T$	1	2	3	4	5	6	total
[0, 20[	7	3	2					12
[20, 40[	4	6	12	2	1			25
[40, 80[	2	13	16	10	3	1		45
[80, 140[	1	9	14	21	10	2		57
[140, 200[		3	5	12	15	5		40
[200, 300[			1	3	10	7		21
total	14	34	50	48	39	15		200

Avant de commencer les interprétations des données collectées, le service marketing a calculé plusieurs grandeurs :

$$\sum_{i=1}^6 n_{i\bullet} c_i = 21890 \quad \sum_{j=1}^6 n_{\bullet j} t_j = 709 \quad \sum_{i=1}^6 n_{i\bullet} c_i^2 = 3343900$$

$$\sum_{i=1}^6 \sum_{j=1}^6 n_{ij} c_i t_j = 90920 \quad \sum_{i=1}^6 \sum_{j=1}^6 \frac{n_{ij}^2}{n_{i\bullet} n_{\bullet j}} = 1,702$$

$c_i$  désigne le centre de la  $i$ -ème classe de la variable  $L$ .

1. Calculer la variance marginale de  $L$  ainsi que le coefficient de variation associé à cette variable. Conclure.
2. Expliquer pourquoi il ne peut y avoir de liaison fonctionnelle entre les variables  $T$  et  $L$ .
3. Calculer la covariance empirique de ces deux variables. Conclure.

4. Calculer la distance du  $\chi^2$ . Quelle valeur aurait pris cette distance dans le cas d'une liaison fonctionnelle ? Commenter.

On s'intéresse maintenant à la distribution des ménages de taille 3 selon les dépenses en loisirs.

5. Calculer le mode et représenter l'histogramme correspondant à cette série.

6. Représenter la courbe de Lorenz et calculer l'indice de Gini. Conclure de l'inégalité dans le bien-être de cette catégorie de ménages.

### Corrigé

$$1. \bar{l} = \frac{1}{200} \sum_{i=1}^6 n_{i \bullet} c_i = 109,45 \text{ dinars.}$$

$$s_L^2 = \frac{1}{200} \sum_{i=1}^6 n_{i \bullet} c_i^2 - \bar{l}^2 = 4740,20.$$

$$cv_L = \frac{s_L}{\bar{l}} = 0,629$$

La série est fortement dispersée par rapport à sa moyenne. Ceci exprime une forte inégalité dans les dépenses en loisirs.

2. Il ne peut y avoir de liaison fonctionnelle entre les variables  $T$  et  $L$  car il existe au moins une ligne (une colonne) contenant plus d'une valeur non nulle.

$$3. \bar{t} = \frac{1}{200} \sum_{j=1}^6 n_{\bullet j} t_j = 3,545.$$

$$s_{TL} = \frac{1}{200} \sum_{i=1}^6 \sum_{j=1}^6 n_{ij} c_i t_j - \bar{l} \cdot \bar{t} = 66,60$$

$s_{TL} \neq 0$  :  $T$  et  $L$  ne sont donc pas indépendantes.

$$4. D^2 = n \left( \sum_{i=1}^6 \sum_{j=1}^6 \frac{n_{ij}^2}{n_{i \bullet} n_{\bullet j}} - 1 \right) = 200 (1,702 - 1) = 140,4.$$

On sait que  $D^2 \leq \min(n(k-1), n(p-1))$ .

On a  $k = p = 6$ . Aussi, dans le cas de liaison fonctionnelle, on aura  $D^2 = 200 \times 5 = 1000$ .

Il est vrai que  $D^2$  est bien éloignée de cette valeur, mais ceci nous autorise seulement à conclure à la non indépendance des deux variables car  $D^2$  est aussi bien loin du zéro !

5. Le tableau qui suit sera complété au fur et à mesure.

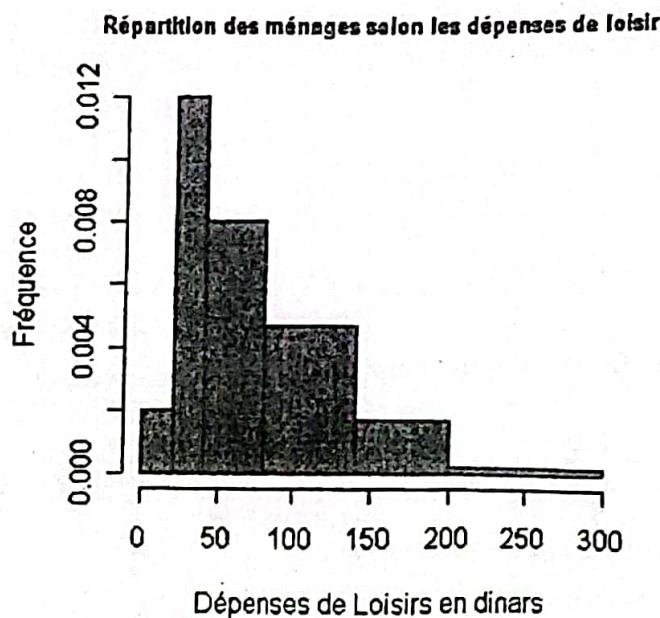
## Cours de Statistique Descriptive

$L$	[0, 20[	[20, 40[	[40, 80[	[80, 140[	[140, 200[	[200, 300[	total
$n_i$	2	12	16	14	5	1	50
$n_i^c$	2	12	8	4,67	0,83	0,2	-
$f_i$	0,04	0,24	0,32	0,28	0,1	0,02	1
$p_i$	0,04	0,28	0,6	0,88	0,98	1	-
$c_i f_i$	0,4	7,2	19,2	30,8	17	5	79,6
$\sum c_i f_i$	0,4	7,6	26,8	57,6	74,6	79,6	-
$q_i$	0,005	0,095	0,337	0,724	0,937	1	-

La classe modale est [20, 40[       $d_1 = 10$        $d_2 = 4$ .

$$M_o = 20 + \frac{d_1}{d_1+d_2} \cdot 2 = 20 + \frac{10}{14} \cdot 20 = 34,29.$$

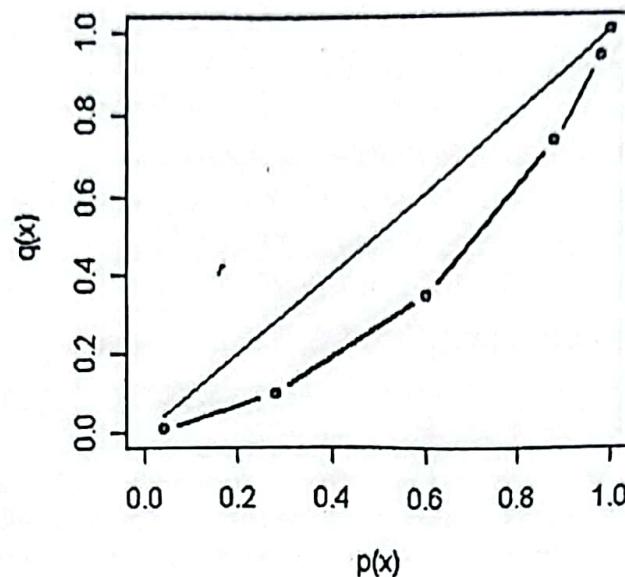
L'histogramme se présente comme suit :



6. La courbe de Lorenz représente les points  $(p_i, q_i)$ . Elle se présente comme

suit :

Courbe de Lorenz des dépenses en loisirs



$$\begin{aligned}
 G &= 1 - \sum_{i=1}^6 (p(e_i) - p(e_{i-1})) (q(e_i) + q(e_{i-1})) \\
 &= 1 - \left( (0,005 \cdot 0,04) + (0,1 \cdot 0,24) + (0,432 \cdot 0,32) \right. \\
 &\quad \left. + (1,061 \cdot 0,28) + (1,661 \cdot 0,1) + (1,937 \cdot 0,02) \right) \\
 &= 1 - 0,664 = 0,336 \approx \frac{1}{3}
 \end{aligned}$$

L'indice de Gini confirme la courbe de Lorenz en exprimant une assez importante inégalité dans les dépenses de loisirs et donc dans le bien-être.

### Exercice 8.15 Examen principal ESSAIT novembre 2005

Un bureau d'aide sociale a décidé d'accorder une "indemnité fournitures scolaires" aux personnes nécessiteuses. Dans le but de mieux chiffrer le montant à verser selon le nombre d'enfants, il a organisé une enquête sur les dépenses en fournitures auprès de 504 ménages. Les données recueillies sont consignées dans le tableau qui suit (X désigne le montant en dinars dépensé par le ménage pour les fournitures lors de l'année précédente et Y est le nombre d'enfants) :

X	Y = 1	Y = 2	Y = 3	Y = 4	Y = 5	Total
[0, 50[	88	49	8	4	0	149
[50, 100[	58	183	38	26	12	317
[100, 200[	1	3	12	2	0	18
[200, 400[	0	1	2	5	5	13
[400, 800[	0	0	1	2	4	7
Total	147	296	61	39	21	504

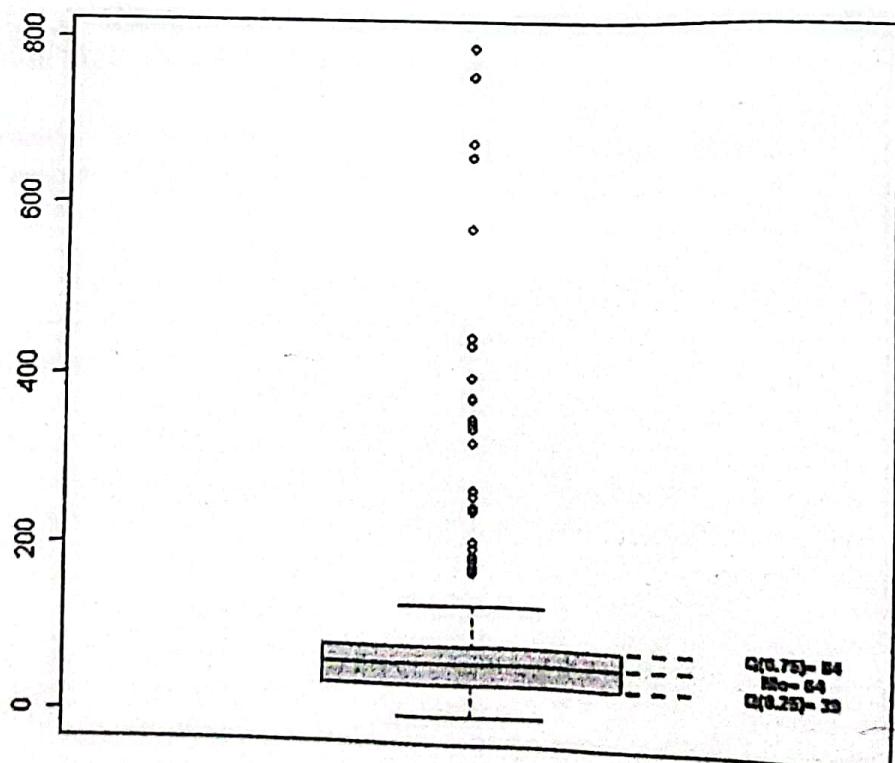
Certains calculs préliminaires ont fourni les résultats suivants :

$$\sum_{i=1}^5 n_{i \cdot} x_i = 38300 \quad \sum_{j=1}^5 n_{\cdot j} y_j = 1063$$

$$\sum_{i=1}^5 \sum_{j=1}^5 n_{ij} x_i y_j = 100450 \quad \sum_{i=1}^5 \sum_{j=1}^5 \frac{n_{ij}^2}{n_{i \cdot} n_{\cdot j}} = 1,5138$$

1. Donner la distribution conditionnelle des fréquences des dépenses en fournitures sachant que la famille possède 2 enfants.
2. Au vu de cette distribution, peut-on s'attendre à l'indépendance des deux variables  $X$  et  $Y$ ? Justifier.
3. Calculer la covariance de  $X$  et  $Y$ . Commenter.
4. Calculer la distance du  $\chi^2$  et en déduire l'absence de liaison fonctionnelle entre ces deux variables.
5. Le graphique qui suit résume la série des ménages selon les dépenses en fournitures scolaires. En donner une lecture détaillée : caractéristiques de position, de dispersion et de forme (on pourra s'appuyer sur les résultats déjà obtenus).

Nombre de ménages selon les dépenses en fournitures



**Corrigé**

1. Répartition des ménages possédant 2 enfants selon leur dépenses en fournitures :

montant dépensé	[0, 50[	[50, 100[	[100, 200[	[200, 400[	total
fréquence	0,208	0,775	0,013	0,004	1

2.  $X$  et  $Y$  sont indépendantes si  $\forall i, \forall j, f_{ij} = f_{i\bullet} \cdot f_{\bullet j}$ .

Vérifions pour  $i = 1$  et  $j = 2$  que cette égalité n'est pas valide. On sait que  $f_{ij} = f_i^j \cdot f_{\bullet j}$ .

$$f_{12} = \frac{49}{504} = 0,097 \quad f_1^2 = 0,208 \text{ et } f_{\bullet 2} = \frac{236}{504} = 0,468 \neq f_1^2.$$

Il n'y a donc pas d'indépendance.

$$3. s_{xy} = \frac{1}{504} \sum_{i=1}^5 \sum_{j=1}^5 n_{ij} x_i y_j - \bar{x}\bar{y} = \frac{100450}{504} - \frac{38300}{504} \frac{1063}{504} = 39,029 \neq 0.$$

Ceci confirme que les variables  $X$  et  $Y$  ne sont pas indépendantes et que de plus, elles varient dans le même sens.

$$4. D^2 = 504 \left( \sum_{i=1}^5 \sum_{j=1}^5 \frac{n_{ij}^2}{n_{i\bullet} n_{\bullet j}} - 1 \right) = 504 \cdot 0,5138 = 258,96.$$

Il y a liaison fonctionnelle si  $D^2 = \min(n(k-1), n(p-1)) = 2016$ , or  $D^2 << 2016$ .

5. La médiane vaut 64 Dinars  $< \bar{x} = \frac{38300}{504} = 75,99$ .

Comme la distribution est unimodale, on en conclut une asymétrie à droite ; celle-ci est confirmée par le grand nombre d'outliers au dessus de la moustache supérieure.

En dehors de l'intervalle interquartiles [39, 84] où les individus sont assez rapprochés, la série est fortement dispersée.

## 8.2 Autres exercices

**Exercice 8.16** Un syndicat de travailleurs cherche à démontrer les disparités de salaires entre les filiales d'une multinationale implantée dans deux régions différentes. Une enquête est effectuée dans ce cadre. Elle aboutit aux résultats suivants :

Salaire annuel en dinars	I $n_i$	I $f_i$	I $\sum f_i$	II $n_i$	II $f_i$	II $\sum f_i$
- de 3.500	314	0,074	0,074	266	0,222	0,222
[3.500; 4.000[	247	0,058	0,132	264	0,220	0,442
[4.000; 4.500[	468	0,110	0,242	220	0,183	0,625
[4.500; 5.500[	1.122	0,264	0,506	281	0,234	0,859
[5.500; 6.500[	918	0,216	0,722	100	0,089	0,942
[6.500; 8.500[	816	0,192	0,914	51	0,043	0,985
Plus de 8.500	365	0,086	1	18	0,015	1
Total	4.250	1		1200	1	

(avec pour la région I,  $\bar{x} = 5.742,5$  et  $\sum n_i c_i^2 = 15.382.843,75 \cdot 10^4$   
et pour la région II,  $\bar{x} = 4.290,4$  et  $\sum n_i c_i^2 = 24.467 \cdot 10^6$ .

La masse salariale correspondant à la première classe et la dernière classe est respectivement de 785 et 3.467,5 milliers de dinars pour la région I, et de 665 et 166,5 milliers de dinars pour la région II).

1. Déterminer pour les deux régions les extrémités des classes manquantes.  
Dans tout ce qui suit, on supposera que les extrémités des classes ne changent pas d'une région à l'autre et que  $e_0 = 1.500$  et  $e_7 = 10.500$ .
2. Représenter les deux histogrammes.
3. Déterminer pour chaque région, le mode et la médiane.
4. Calculer le coefficient de variation dans chacune des deux régions.
5. Commenter.

**Exercice 8.17** La nouvelle dirigeante d'une entreprise désire s'informer sur la distribution des rémunérations des femmes salariées de l'entreprise. Le tableau qui suit résume la répartition des salariées selon le salaire mensuel ( $X$  en dinars) et l'ancienneté ( $Y$  en nombre d'années).

$X_F$	$Y_F$	[0;4[	[4;8[	[8;12[	[12;20[	[20;28]	Total	Fréq.	cif.
[150;300[	12	10	10	8			40	0.36	81.82
[300;350[	8	14	n2 3	4	4	n2.	f2.	c2f2.	
[350;500[	3	6	n3 3	6	3	n3.	f3.	c3f3.	
[500;800[		2	2	2	3	9	0.08	53.18	
[800;1800]				1	2	3	0.03	35.45	
Total		23	32	n.3	21	12	110	1	
Répartition des femmes salariées selon le salaire et l'ancienneté									

1. On dispose du salaire médian :  $M_e = 322$ . Démontrer que  $n_2. = 35$  et  $n_3. = 23$ .
2. Déterminer le salaire modal et tracer l'histogramme des fréquences.
3. Tracer la courbe de Lorenz des salaires et calculer l'indice de Gini. Interpréter.
4. On donne maintenant,  $\bar{y} = 9.84$  et  $\sum n_j c_j^2 = 15732$  (ici,  $c_j$  désigne le centre de la  $j$ -ème classe de la variable  $Y$ ). Calculer le coefficient de variation associé à la variable nombre d'année d'ancienneté dans l'entreprise. Commenter.

**Exercice 8.18** On souhaite étudier l'évolution du niveau d'instruction de la population active en fonction de l'évolution des dépenses publiques dans le secteur de l'éducation. On observe les deux variables  $X$  (dépenses publiques pour l'éducation en pourcentage du PIB) et  $Y$  (niveau d'instruction de la

population active mesuré par le nombre moyen d'années d'étude). Les observations sont résumées dans le tableau qui suit :

Année	1960	1966	1975	1984	1989	1994	2000
X	3,5	5,82	4,92	6,16	6,29	7,14	9,21
Y	1,5	2,2	3,17	5,26	5,97	7,06	8,33

1. Justifier la validité d'un ajustement linéaire de Y sur X.
2. Déterminer la droite des moindres carrés.
3. Etudier la qualité de l'ajustement obtenu.
4. Calculer le taux d'accroissement annuel moyen du niveau d'instruction de la population active sur la totalité de la période 1960-2000.
5. En supposant que ce taux s'est stabilisé à partir de 1994, donner une prévision du niveau d'instruction en l'an 2005. En déduire une prévision de la part des dépenses publiques dans le PIB.

**Exercice 8.19** Un laboratoire pharmaceutique bénéficie du monopole de la fabrication d'un médicament A. Le département de recherche-développement est parvenu à élaborer un parfait substitut à ce médicament. Ce nouveau médicament B a des coûts de production beaucoup plus faibles.

Le médicament B a donc été introduit sur le marché. Voici le bilan des quantités vendues (en  $10^3$  unités de vente) sur dix mois pour A ( $y_i$ ) et B ( $x_i$ ).

$y_i$	16	18	17	14	15	11	8	12	9	6
$x_i$	2	4	6	6	9	11	14	16	18	20

1. Est-ce qu'une régression linéaire serait judicieuse ? Justifier.
2. Ecrire les équations des droites de régression de y par rapport à x et de x par rapport à y. Les représenter sur un même repère.
3. En déduire le coefficient de corrélation linéaire de ces deux variables. Pouvait-on s'y attendre ? Expliquer.

**Exercice 8.20** Une entreprise commerciale consacre une certaine somme ( $x$ ) en milliers de dinars à des opérations publicitaires au début de chaque mois. Les données qui suivent résument les douze observations de l'année écoulée (y étant le chiffre d'affaires mensuel en milliers de dinars).

$$\bar{x} = 3 \quad \bar{y} = 40$$

$$\sum x_i^2 = 114,58 \quad \sum x_i y_i = 1492,40 \quad \sum y_i^2 = 19708$$

On suppose que la distribution présente une progression linéaire.

1. Calculer le coefficient de corrélation linéaire entre ces deux variables. Commenter.

2. Déterminer l'équation de la droite d'estimation de  $y$  à partir de  $x$ .
3. Quel indice synthétique chaîne pour l'année 1990, base 100 en 1970 peut-on construire ? (justifier la réponse et calculer l'indice).

**Exercice 8.21** L'augmentation du niveau général des prix observée mensuellement au cours d'une année a donné les résultats suivants :

mois	01	02	03	04	05	06	07	08	09	10	11	12
augmentation en %	1	1,2	1	1,3	1,9	1,2	1,2	1,4	1,9	1,3	1,7	1,7

Quel est le taux d'inflation pour l'année ?

Quel est le taux d'inflation mensuel moyen ?

**Exercice 8.22** La structure de la consommation finale des ménages en 1990 se présente comme suit :

Groupe	Poids en %
Alimentation	41,22
Habitation	18,73
Hygiène et soins	9,11
Transport	8,78
Habillement	10,39
Loisirs	11,77
Total	100

Les indices élémentaires base 100 en 1990, des différents groupes sont donnés par le tableau suivant :

Groupe	1983	1991	1992
Alimentation	58,1	108,7	114,0
Habitation	67,2	105,3	111,8
Hygiène et soins	65,4	106,6	112,8
Transport	60,5	109,9	118,7
Habillement	53,2	108,0	116,9
Loisirs	65,7	111,3	116,9

1. Calculer l'indice de Laspeyres de la consommation finale pour les années 1991 et 1992, base 100 en 1990. Interpréter.
2. Quelle année a enregistré le taux d'inflation le plus faible ?
3. Proposer un indice synthétique pour 1990 base 100, en 1983. Justifier votre choix.
4. Quel est le groupe de produits qui a accusé la plus forte hausse des prix sur la totalité de la période 1983-1992 ?

**Exercice 8.23** On dispose de statistiques sur les produits manufacturiers. Le tableau qui suit en est le résumé.

	Pondération 1980	Indice 80/70	Pondération 1990	Indice 90/70
Meubles et tapis(I)	0,36	123,6	0,32	158
Articles de toilette(II)	0,15	129,3	0,14	174,8
Papeterie et librairie(III)	0,32	122	0,36	146
Autres équipements(IV)	0,17	134,5	0,18	182,4

Une pondération  $w_i$  représente la part dans les dépenses totales et dans le courant de l'année concernée, des dépenses allouées à la catégorie de biens  $i$ .

1. Déterminer les indices élémentaires de prix pour l'année 1990, base 100 en 1980.
2. Déterminer l'indice de Laspeyres relatif à l'ensemble des produits manufacturiers pour l'année 1990, base 100 en 1980.
3. Déterminer l'indice de Paasche de ces produits pour la même année 1990, base 100 en 1980.

**Exercice 8.24** On s'intéresse à la répartition d'un échantillon de 1636 ménages selon le revenu ( $R$ ) et la consommation totale ( $CT$ ).

R CT	4000	5000	6000	8000	10000
	5000	6000	8000	10000	12000
[4000, 5000[	10	148	30		
[5000, 6000[	42	320	98	43	
[6000, 7000[		52	567	33	
[7000, 8000[			18	162	15
[8000, 10000[			12	78	8

1. Calculer les moyennes et les variances marginales.
2. Calculer la covariance du revenu et de la consommation.
3. Existe-t-il une liaison fonctionnelle entre les deux variables ?
4. On s'intéresse à l'étude de l'inégalité de la répartition des revenus des ménages. Calculer le coefficient de variation.
5. Représenter la courbe de Lorenz.
6. Calculer l'indice de Gini. Interpréter.

**Exercice 8.25** Un importateur d'automobiles désire augmenter sa part de marché. Il étudie alors la répartition des  $n$  modèles commercialisés par ses

concurrents en fonction du prix  $y$ , exprimé en milliers de dinars, et de la puissance fiscale  $x$ , exprimée en chevaux :

$x$	$y$	[15, 30[	[30, 35[	[35, 40[	[40, 45[	[45, 50[	[50, 55[	total
3-4	22	8						30
5-6	36	41	20	13				110
7-8	12	36	n <sub>33</sub>	50	22	22		n <sub>3</sub>
9-10				12	26	32	30	100
11-12					6	14	20	40
13-15						2	8	10
total		70	85	n <sub>3</sub>	95	70	80	n

Les deux parties sont indépendantes.

1<sup>ère</sup> partie :

1. Déterminer n<sub>3</sub> et n<sub>3</sub> sachant que  $\bar{y} = 39,3$ .

2. Pour toute la suite, on donne  $\bar{x} = 7,67$  et  $s_y^2 = 87,26$ .

Calculer le moment d'ordre 2 de la puissance fiscale de l'ensemble des voitures étudiées. En déduire la variance de la variable x.

3. Calculer la covariance du prix et de la puissance fiscale.

(Indication :  $\sum_{i=1}^6 \sum_{j=1}^6 n_{ij} x_i y_j = 157.992,5$ ).

2<sup>ème</sup> partie :

L'importateur décide d'axer son action sur les véhicules de 7 à 8 chevaux.

1. Tracer l'histogramme de la distribution des prix des voitures de 7 à 8 chevaux.

2. Déterminer le mode et la médiane. En déduire la moyenne. Peut-on conclure la symétrie de la distribution ?

3. Après avoir effectué le changement de variables qui vous semble adéquat et recalculé la valeur exacte de la moyenne empirique, calculer le coefficient d'asymétrie de Fisher. Conclure.

**Exercice 8.26** Une entreprise commerciale consacre une certaine somme ( $x$ ) en milliers de dinars à des opérations publicitaires au début de chaque mois. Les données qui suivent résument les douze observations de l'année écoulée ( $y$  étant le chiffre d'affaires mensuel en milliers de dinars).

$$\bar{x} = 3 \quad \bar{y} = 40$$

$$\sum x_i^2 = 114,58 \quad \sum x_i y_i = 1492,40 \quad \sum y_i^2 = 19708$$

On suppose que la distribution présente une progression linéaire.

1. Calculer le coefficient de corrélation linéaire entre ces deux variables. Commenter.

2. Déterminer l'équation de la droite d'estimation de  $y$  à partir de  $x$ .
3. Quel indice synthétique chaine pour l'année 1990, base 100 en 1970 peut-on construire ? (justifier la réponse et calculer l'indice).

### 8.3 Exercices pratiques sous R

**Exercice 8.27** Simuler sous R un échantillon de taille 100 de la loi normale  $N(\mu = 50, \sigma^2 = 49)$ .

Soit  $y$  l'échantillon des 99 premières statistiques d'ordre des valeurs simulées.

Soit  $x$  le vecteur des centiles de la loi normale  $N(0, 1)$ .

On s'intéresse à la qualité de la corrélation linéaire de  $y$  sur  $x$  (ou de  $x$  sur  $y$ ) en fonction de  $\mu$  et  $\sigma^2$ .

- Calculer pour  $x$  et  $y$  la moyenne et la variance empiriques.
- Calculer la covariance empirique ainsi que le coefficient de corrélation linéaire empirique entre  $x$  et  $y$ .
- Donner les équations des droites de régression de  $y$  sur  $x$  et de  $x$  sur  $y$ .
- Représenter sur un même graphique le nuage de points  $(x, y)$  ainsi que les deux droites. Conclure.
- Refaire le processus en multipliant  $\mu$  par 0.1 ; 0.5 ; 5 ; 10 ; 20 et 50. Qu'en est-il de la qualité de la régression ?
- Est-ce que l'intensité de la relation entre  $x$  et  $y$  s'améliore lorsque  $\sigma^2$  diminue ?

**Exercice 8.28** Simuler deux échantillons de 100.000 valeurs chacun, le premier suivant une loi normale  $N(3, \frac{1}{4})$  et le second suivant une loi gamma  $\gamma(\frac{1}{6}, \frac{1}{2})$ .

Chacun de ces deux échantillons représentera la distribution des revenus annuels en milliers de dinars de 100.000 ménages.

Après avoir regroupé les données en 100 classes d'égales amplitudes, il s'agira de représenter la courbe de Lorenz pour les deux échantillons et de déterminer la répartition la plus inégalitaire.

Dans un second temps, les mêmes échantillons seront repris et regroupés en fonction des centiles. A partir du nouveau découpage, deux nouvelles courbes de Lorenz seront tracées et deux nouveaux indices de Gini seront calculés.

Pour chacun des deux échantillons,

- Peut-on comparer les deux courbes ?
- Est ce que  $G$  change ? Si oui, quelles sont les pistes à explorer pour en connaître les raisons ? Sinon, justifier de l'absence de changement.

**Exercice 8.29** Simuler un échantillon de  $N$  valeurs suivant une loi exponentielle  $E(\frac{1}{3})$ . Cet échantillon représentera la distribution des revenus annuels en milliers de dinars de  $N$  ménages.

Après avoir regroupé les données en  $n$  classes d'égales amplitudes, il s'agira d'étudier la répartition de ces revenus et de juger de l'inégalité du revenu dans l'échantillon (courbe de Lorenz et indice de Gini).

Dans un second temps, le même échantillon sera repris et il sera effectué un regroupement en fonction des centiles. A partir du nouveau découpage, une nouvelle courbe de Lorenz sera tracée et un nouvel indice de Gini sera calculé.

Donner la vraie valeur de l'indice de Gini. Peut-on comparer les deux courbes ? Est-ce que  $G$  change.

Le résultat dépend-il de la taille de l'échantillon ? (on prendra  $N$  compris entre 300 et 200.000)

Quelle procédure proposer pour déterminer la plus efficace de ces deux méthodes ?