L2 : Statistique Descriptive
Exemple

Enquête Dauphine

Fichier utilisé

C'est un extrait du fichier "enquête Dauphine 1992"

NUM	SEXE	BAC	LYCEE	VILLE-LYCEE	NOTE-BAC	NOTE-TERM	ETUDES-PER	REPAS	AGE-MERE	PRIERE	REVENUS
1	M	C	PUBL	ORLEANS	12	14	G.E	2	48	JAMA	158
2	F	D	PUBL	PARIS	14	14	B+4	8	44	1F/M	109
3	M	C	PUBL	ORLEANS	12	14	G.E	2	48	JAMA	210
4	F	D	PUBL	PARIS	14	14	B+4	8	44	1F/M	359
5	F	D	PUBL	PARIS	14	14	B+4	8	44	1F/M	169
6	F	D	PRIV	BORDEAUX	16	14	G.E	0	30		476
7	F	D	PUBL	ENGHIEN LE	13	14	G.E	8	53	1F/A	173
8		D		LA CELLE S	12	14	B+4	5	41	JAMA	471
9	M	D	PRIV	PARIS	14	14	BAC	7	51	1F/A	195
10	F	D	PUBL	PARIS	12	14	B+6	6	41	1F/A	123
11	M	B	PUBL	PARIS	12	14	B+6	7	52	1F/A	132
12		C	PRIV	RUEIL MALM	12	14	G.E	9	50	1F/S	137
13	F	D	PUBL	ST CYR LEC	13	14	BAC	9	55	1F/A	186
14	F	D	PRIV	PARIS	13	14		5	48	JAMA	344
15	M	C		NICE	14	14		9	53	JAMA	167
16	F	B	PUBL	MEUDON	14	14	SEC	9	46	JAMA	164
17	F	D	PUBL	PARIS	11	14	G.E	6	50	JAMA	106
18	F	D	PRIV	BOULOGNE	12	14		5	56	JAMA	357
19	F	C	PUBL	BOULOGNE	14	14	SEC	9		JAMA	260
20	F	C	PUBL	VERSAILLES	10	14	B+6	9	46	JAMA	163
21	F	C	PRIV	SAINT MAND	11	14	SEC	6	42	1F/A	240
22	F	D	PUBL	NOGENT SU	14	14	B+4	8	59	JAMA	291
23	F	C	PUBL	PARIS	11	14	B+2	9	57	JAMA	318
24	F	C	PUBL	SEVRES	13	14	G.E	5	46	1F/A	138

Accès au fichier DAU12.XLS

Ce que nous cherchons

- Résumer les distributions de quelques variables du fichier (liste diapo suivante)
- Pour chaque variable
  - Nous repérons le type de la variable
  - Nous établissons la table des fréquences
  - Nous représentons cette distribution par un graphique
  - Nous la résumons par un certain nombre de valeurs caractéristiques

DAU12 : Type desVariables à étudier

Variable	Type	Nbre modalités
Sexe		2 +NR
Bac		9 +NR
Ville Lycée		Ouverte : 193
Prière		5 + NR
Etudes père		7 + NR
Note-Bac		10
Repas		10 + NR
Age mère		29 + NR
Revenu parents		

Qualitative Nominale

Qualitative Ordinale

Quantitative Discrète

Quantitative Continue

DAU : Variables à étudier

Variable	Type	Nbre modalités
Sexe	nominale	2 +NR
Bac	nominale	9 +NR
Ville Lycée	nominale	Ouverte : 193
Prière	ordinale	5 + NR
Etudes père	ordinale	7 + NR
Note-Bac	quant. discrète	10
Repas	quant. discrète	10 + NR
Age mère	Quant. discrète	29 + NR
Revenu parents	Quant. continue	

Utilisation d'Excel

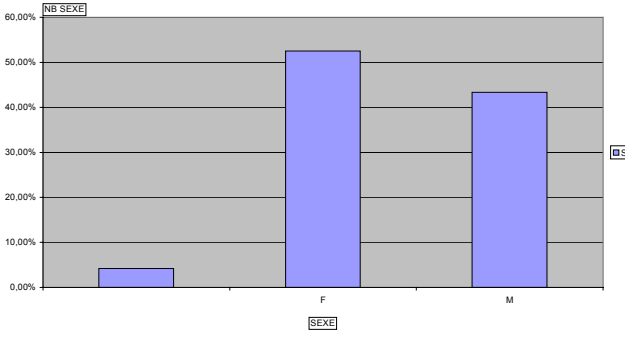
- Les fonctions statistiques utilisables sont :
  - Centile, Ecart moyen, Ecart-type, Var, Petite valeur, Max, Min, Médiane, Mode, Moyenne, Moyenne géom., Nbval, Quartile.
- L'utilitaire d'analyse "Statistiques descriptives" donne directement presque toutes les caractéristiques numériques mais il exige que la variable soit quantitative et ne comprenne aucune élément textuel. Il faut donc convertir les Non réponses "blanches" en cases "vides"
- L'utilitaire d'analyse "histogramme" ne fonctionne que dans les mêmes conditions et produit à la fois un classement et un pseudo-histogramme.
- Le "rapport de tableau croisé dynamique" permet d'obtenir des tables de fréquences et des représentations graphiques pour des variables qualitatives ou quantitatives.

Sexe

- C'est une variable nominale pour laquelle la table des fréquences suivante peut être obtenue avec "tableau croisé dynamique".
  - Utiliser l'assistant pour créer le premier tableau
  - Déposer-glisser la variable "sexe" en colonne et dans le centre du tableau"
  - Pour obtenir les fréquences relatives (second tableau) : Ouvrir le sous menu "champ pivot table"-"Options" -"Afficher les données"-"% par colonne"

NB SEXE	
SEXE	Somme
	26
F	326
M	269
Total	621

NB SEXE	
SEXE	Somme
	4,19%
F	52,50%
M	43,32%
Total	100,00%



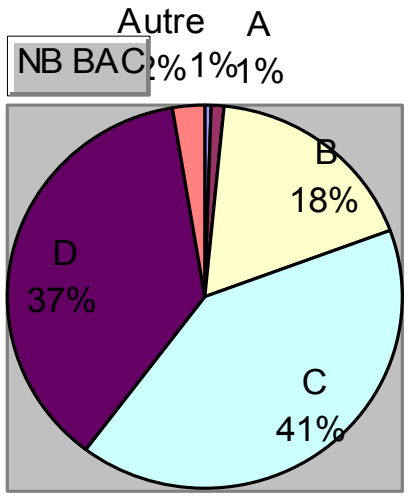
Type de Bac

Table de fréquence brute

Avec regroupements

Avec éclatement par sexe

Diagramme en secteurs



Mode : Bac C

Pas de médiane

Pas de moyenne

SEXE	(Tous)
NB BAC	
BAC	Somme
	4
A1	6
A2	1
B	109
C	255
D	231
EQ	11
F	1
G	3
Total	621

BAC2	BAC	Somme
		4
A		7
B		109
C		255
D		231
Autre		15
Total		621

BAC	SEXE	Somme
		4
A1		6
A2		1
B		4
	F	65
	M	40
Somme B		109
C		8
	F	126
	M	121
Somme C		255
D		12
	F	123
	M	96
Somme D		231
EQ		11
F		1
G		3
Total		621

Ville-Lycée

VILLE-LYCEE	Somme
	4
AIX EN PROVENCE	1
ALBI	1
ALLEMAGNE	1
ANTHONY	1
ANTONY	1
ARCACHON	1
ARMENTIERES	1
ARPAJON	3
ASNIERE	1
ATHIS MONS	1
BAGNEUX	1
BARBEZIEUX	1
BASTIA	1
BERLIN	1
BLOIS	1
BOIS COLOMBES	2
BORDEAUX	1
BOULOGNE	9
BOULOGNE BILLANCOURT	1
BOURG EN BRESSE	1
BOURG LA REINE	1
BRETIGNY SUR ORGE	1
BRUNOY	4
BRUXELLES	1
CASABLANCA	3

Cette variable nominale résulte d'une question ouverte. Les modalités n'ont donc pas été listées et codées avant le dépouillage du questionnaire. La table des fréquences (extrait ci-contre) qui comprend en fait 193 modalités (y compris erreurs de saisie) n'est plus un résumé très intéressant. On pourrait la recoder par département ou plus sommairement en Paris, Iles de France, Province, Etranger. Mais rien ne permet de le faire automati-quement ; les regroupements sont à faire à la main.

Paris	233
Ile de France	??
Province	??
Etranger	??

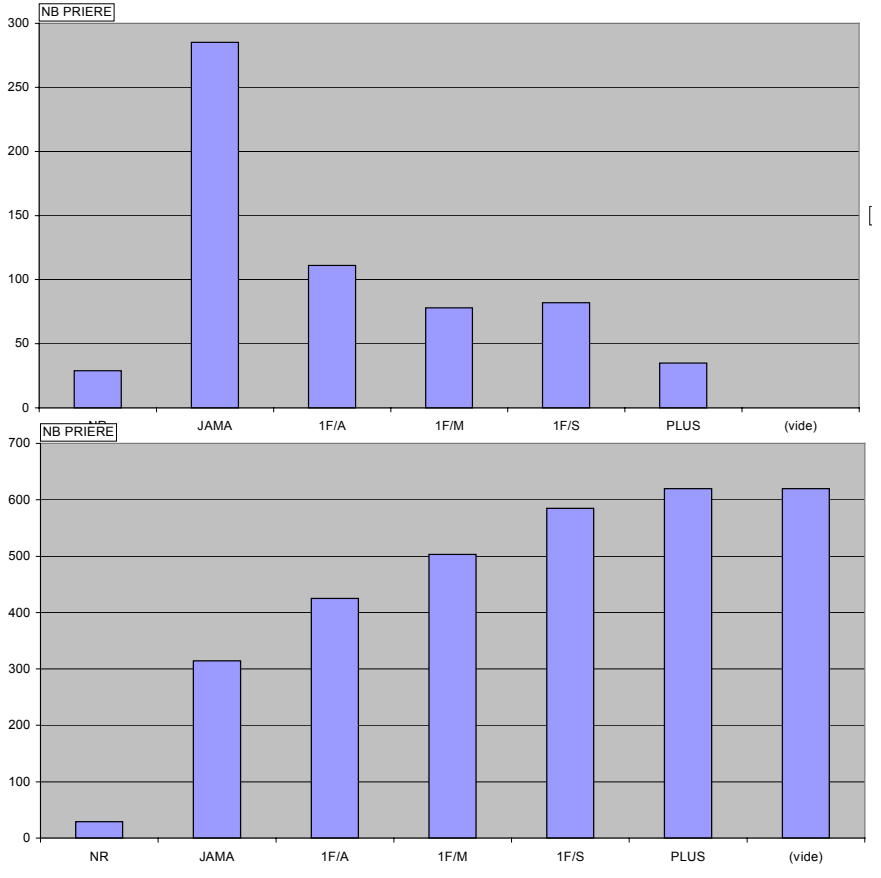
Prière

La variable étant ici ordinale (il y a un ordre sur les modalités), on peut compléter la table des fréquences par les fréquences cumulées (dans "afficher les données par". Attention à réordonner les modalités avec une liste personnalisée). Le second graphique correspond à ces cumuls.

PRIERE	Somme	Somme
NR	29	29
JAMA	285	314
1F/A	111	425
1F/M	78	503
1F/S	82	585
PLUS	35	620
(vide)		620
Total	620	

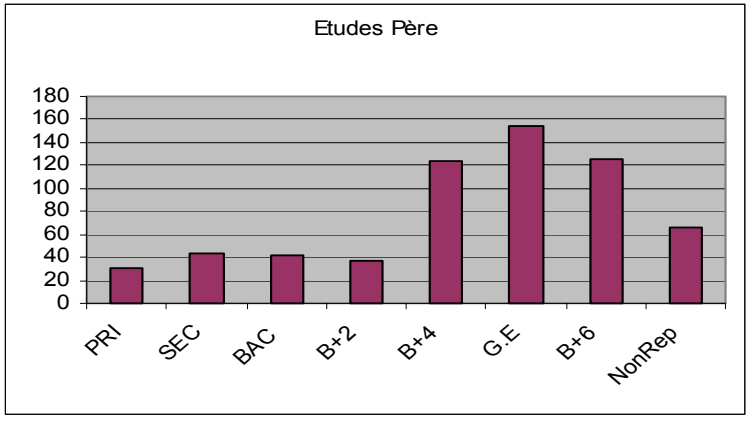
Le mode : "jamais"

L'individu médian est le 311è (ou le 296è si on élimine les NR). La médiane est : "jamais" dans les deux cas.

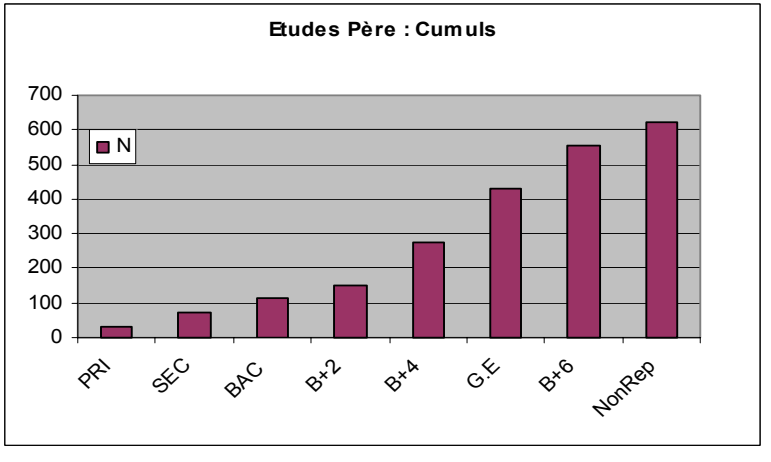


Etudes père

La variable est encore du type ordinale, et le cumul prend donc un sens (après avoir encore réordonné les modalités)
Le mode est G.E (Grandes Ecoles)
La médiane est (après avoir éliminé les 66 non-réponses) la modalité du 278è individu, soit "Bac + 4".



ETUDES-PERE	n	N
PRI	30	30
SEC	44	74
BAC	41	115
B+2	37	152
B+4	123	275
G.E	154	429
B+6	126	555
NonRep	66	621
Total	621	

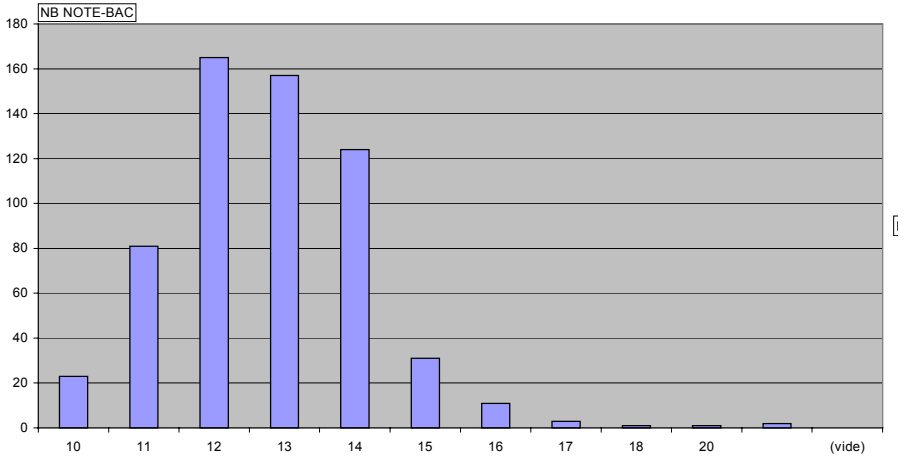


Note-Bac

Cette variable est quantitative discrète. Les cumuls ont un sens et on peut aussi calculer moyenne et écart-type. Comme la colonne Note-Bac contient des zones blanches de non réponse on ne peut pas utiliser l'utilitaire d'analyse "stat. descriptives". Les fonctions "Nbval", "moyenne","ecartypep", "Quartiles", et "Centiles" donnent les résultats ci-contre :

Nbre de val	599
Ecart-type	1,37
D1	11
Q1	12
Moyenne	12,74
Q2	13
Q3	14
Q9	14
C99	16

NOTE-BAC	Somme	Somme
10	23	23
11	81	104
12	165	269
13	157	426
14	124	550
15	31	581
16	11	592
17	3	595
18	1	596
20	1	597
	2	599
(vide)		599
Total	599	



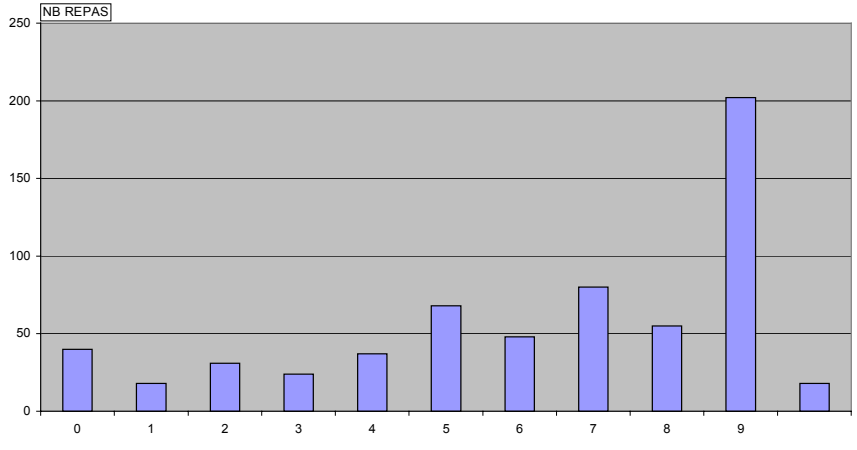
Nombre de repas pris à la maison

C'est encore une variable quantitative discrète. Le mode est aussi le maximum : 9 repas par semaine.

La médiane (=Q2) est de 7 repas/sem.

Nbre de val	605
Ecart-type	2,85
D1	2
Q1	5
Moyenne	6,21
Q2	7
Q3	9
D9	9
C99	9

REPAS	Somme	Somme	Somme
0	40	6,44%	40
1	18	2,90%	58
2	31	4,99%	89
3	24	3,86%	113
4	37	5,96%	150
5	68	10,95%	218
6	48	7,73%	266
7	80	12,88%	346
8	55	8,86%	401
9	202	32,53%	603
	18	2,90%	621
Total	621	100,00%	



Age Mère

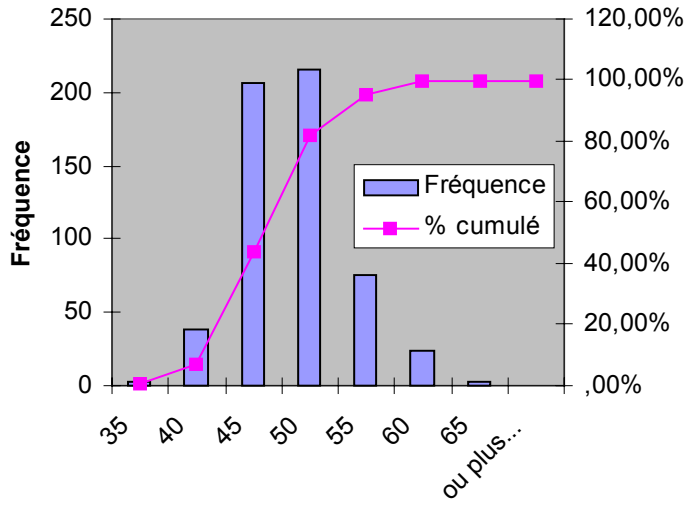
AGE-MERE	Somme
30	1
35	2
36	2
37	4
38	5
39	4
40	23
41	18
42	34
43	43
44	51
45	61
46	55
47	41
48	51
49	26
50	43
51	12
52	19
53	18
54	16
55	11
56	10
57	4
58	2
59	5
60	3
61	1
64	2
(vide)	
Total	567

La table des fréquences à gauche a été obtenue avec "tableaux croisés dynamiques". Cette variable quantitative discrète peut être traitée avec les utilitaires d'analyse "statistiques descriptives" et "histogramme" (bornes sup fixées par utilisateur)

D1	41
Q1	44
Q2	46
Q3	50
D9	53

	Fréquence	*% cumulé*
35	3	,53%
40	38	7,23%
45	207	43,74%
50	216	81,83%
55	76	95,24%
60	24	99,47%
65	3	100,00%
ou plus…	0	100,00%

Moyenne	46,755
Erreur-type	0,199
Médiane	46
Mode	45
Écart-type	4,740
Variance de l'	22,465
Kurstosis (Cc	0,660
Coefficient d'a	0,487
Plage	34
Minimum	30
Maximum	64
Somme	26510
Nombre d'éch	567



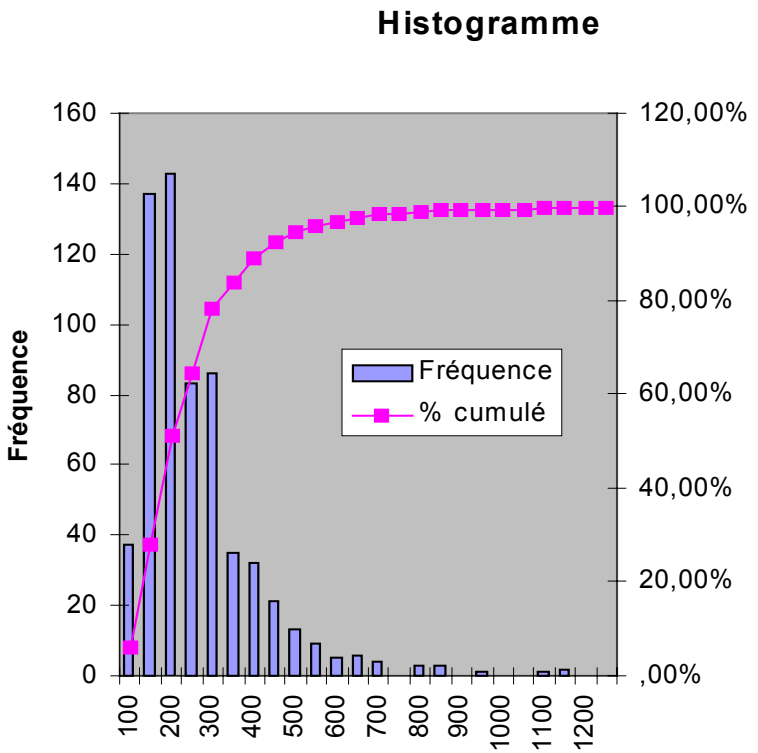
Revenus

Cette variable quantitative a presque autant de modalités que d'observations. Elle est quasi continue. Il faut donc la transformer en variable classée.

Moyenne	239,38
Erreur-type	5,87
Médiane	196,81
Mode	110,60
Écart-type	146,38
Variance de l'échantillon	21428,50
Kurstosis (Coefficient d'applat	7,54
Coefficient d'assymétrie	2,26
Plage	1056,77
Minimum	62,71
Maximum	1119,48
Somme	148656,45
Nombre d'échantillons	621

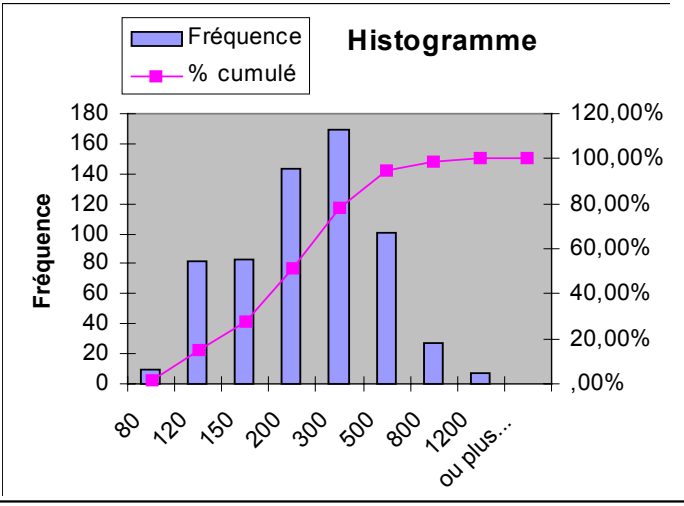
Ici nous avons choisi des classes d'égale amplitude (50) ce qui permet d'obtenir par l'utilitaire d'Excel un histogramme à peu près correct : les rectangles devraient être adjacents et les graduations de l'axe horizontal correspondent aux bornes supérieures des classes.
(à suivre)

	Fréquen	*% cumulé*
100	37	5,96%
150	137	28,02%
200	143	51,05%
250	83	64,41%
300	86	78,26%
350	35	83,90%
400	32	89,05%
450	21	92,43%
500	13	94,52%
550	9	95,97%
600	5	96,78%
650	6	97,75%
700	4	98,39%
750	0	98,39%
800	3	98,87%
850	3	99,36%
900	0	99,36%
950	1	99,52%
1000	0	99,52%
1050	0	99,52%
1100	1	99,68%
1150	2	100,00%
1200	0	100,00%

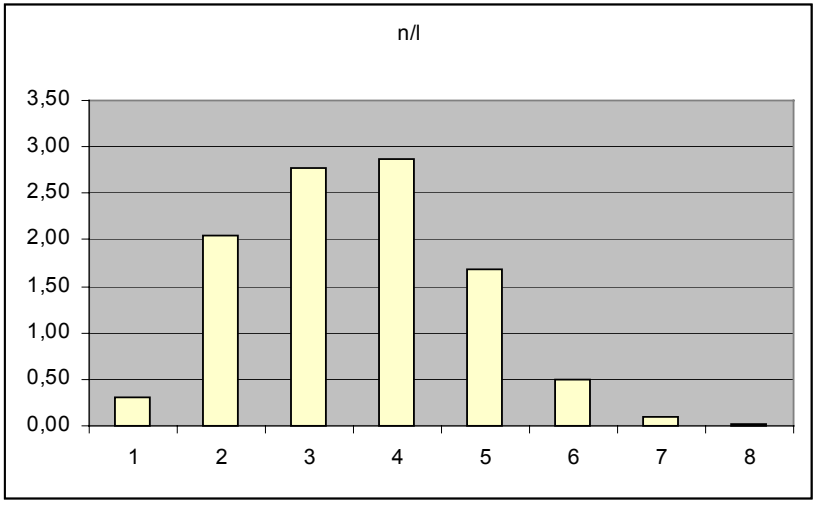


Revenus (2)

l	binf	bsup	n	f	F	N	n/l
30	50	80	9	0,014	0,014	9	0,30
40	80	120	82	0,132	0,147	91	2,05
30	120	150	83	0,134	0,28	174	2,77
50	150	200	143	0,23	0,51	317	2,86
100	200	300	169	0,272	0,783	486	1,69
200	300	500	101	0,163	0,945	587	0,51
300	500	800	27	0,043	0,989	614	0,09
400	800	1200	7	0,011	1	621	0,02
		TOTAL	621	1			

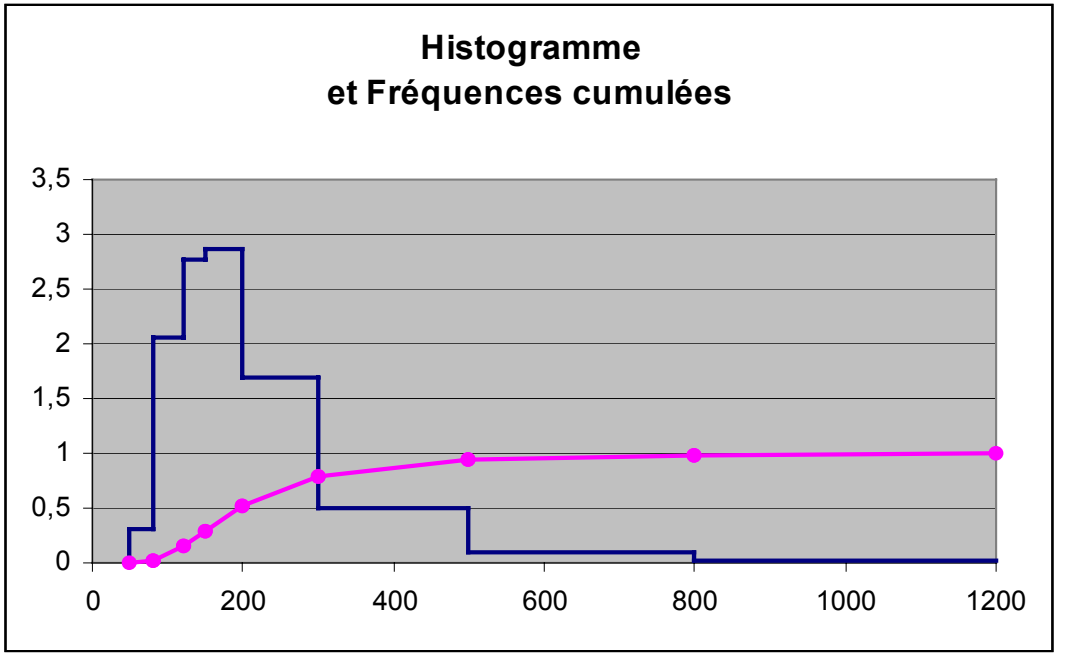


Si nous faisons pour les mêmes données des classes de largeurs inégales, alors l'histogramme fourni par Excel devient faux. Il faut faire des rectangles de hauteurs proportionnelles à n/l (2ème graphique) pour respecter le principe de conservation des aires (cf. cours) et il faudrait en plus que les rectangles reproduisent les largeurs de classe.



Revenus (3)

borne	n/l	F
50	0	0
50	0,3	0
80	0,3	0,014
80	2,05	0,014
120	2,05	0,147
120	2,767	0,147
150	2,767	0,280
150	2,86	0,280
200	2,86	0,510
200	1,69	0,510
300	1,69	0,783
300	0,505	0,783
500	0,505	0,945
500	0,09	0,945
800	0,09	0,989
800	0,018	0,989
1200	0,018	1



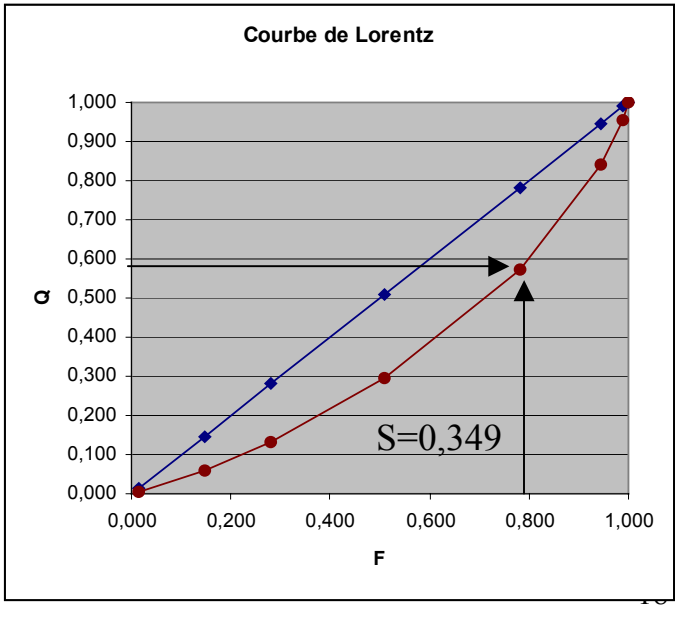
Un histogramme correct (qui respecte les échelles des classes en abscisse) est obtenu à partir du type "nuage de points" et en dédoublant les valeurs de n/l pour les bornes inf et sup de chaque classe. On a figuré sur le même graphique la courbe de répartition obtenue à partir des fréquences cumulées

Revenus (4)

On termine l'étude des revenus par la confection de la courbe de concentration ou courbe de Lorentz qui donne les revenus cumulés par classe Q en fonction des effectifs cumulés F. On lit par exemple sur cette courbe que 78% des familles les moins riches cumulent 57% de la masse des revenus (ou encore que 22% les plus riches cumulent 43% des revenus)

On peut calculer la surface S sous la courbe de Lorentz (somme des surfaces des trapèzes = 0, 349) , et en déduire un indice de concentration de Gini égal à : G = (0,5-S)/0,5 = 0,302.

I	binf	bsup	n	f	F	N	n/I	fx*	q	Q
30	50	80	9	0,014	0,014	9	0,30	0,94	0,004	0,004
40	80	120	82	0,132	0,147	91	2,05	13,20	0,054	0,058
30	120	150	83	0,134	0,280	174	2,77	18,04	0,074	0,131
50	150	200	143	0,23	0,510	317	2,86	40,30	0,164	0,296
100	200	300	169	0,272	0,783	486	1,69	68,04	0,278	0,573
200	300	500	101	0,163	0,945	587	0,51	65,06	0,265	0,839
300	500	800	27	0,043	0,989	614	0,09	28,26	0,115	0,954
400	800	1200	7	0,011	1,000	621	0,02	11,27	0,046	1,000
		TOTA	621	1				245,11	1	



Exercice

Produire, dans la mesure où ils sont définis, les résumés numériques et graphiques listés ci-contre…

Pour d'autres variables du fichier original.

- table des fréquences
- graphiques des fréquences (densité)
- graphique des fréquences cumulées
- mode
- médiane
- moyenne
- quartiles
- premier et dernier décile
- écart-type
- coefficient de variation
- indice de concentration