CAHIER DE RECHERCHE

CRÉDOC
L'ENTREPRISE DE RECHERCHE

CRÉDOC

Mai 1996 N°85

VISUALISATION GRAPHIQUE POUR L'ANALYSE EXPLORATOIRE DES DONNÉES

Christine Henriot-Olm
Marie-Odile Simon

Département
Évaluation des Politiques Sociales

SOMMAIRE

Page

INTRODUCTION 2

REPRÉSENTATIONS GRAPHIQUES MONOFACTORIELLES 7

I CAS D'UNE VARIABLE QUALITATIVE 8

II CAS D'UNE VARIABLE QUANTITATIVE 9

II-1. Représentation de l'ensemble des valeurs prises par la variable 12

II-1-1. Représentation unidimensionnelle des données 12

II-1-2. Représentation bidimensionnelle : fréquence et densité en fonction des valeurs de la variables 16

II-1-2.a Représentation de la fréquence et variable répartie en classe 16

II-1-2.b Représentation continue à deux dimensions : les estimations de densité 23

II-1-2.c Représentation de la fonction de distribution : « quantile plot » ou « Q-plot » 28

II-2. Représentation des résumés numériques de la variable 30

II-2-1. Les « box plots » ou « boîtes à pattes » 30

II-2-2. La symétrie 33

III. EXEMPLE D'ANALYSE EXPLORATOIRE À PARTIR DES GRAPHIQUES MONOFACTORIELS 34

III-1. Le revenu par u.c. 35

III-2. Le salaire par u.c. 37

ÉTUDE DES RELATIONS ENTRE DEUX VARIABLES CONTINUES 43

I. COMPARAISON DE DEUX DISTRIBUTIONS DANS LEUR ENSEMBLE 45

I-1. Comparaison des valeurs ordonnées 46

I-1-1. Les « QQ plots » ou Q plots de comparaison 46

I-1-2. Comparaison des diagrammes « stem and leaf » 50

I-1-3. Comparaison des courbes de densité 52

I-2. Comparaison des résumés numériques de chacune des deux variables 55

II. COMPARAISON DES VARIABLES POUR UN MÊME INDIVIDU 61

II-1. Représentation des individus en fonction des valeurs des deux variables 61

II-1-1. Représentation des individus en fonction des valeurs des deux variables 61

II-1-2. Représentation de la fréquence d'une variable à deux dimensions 64

II-1-3. La densité d'une variable à deux dimensions 66

II-2. Représentation des résumés numériques d'une variable en fonction de l'autre variable 70

II-2-1. Évolution des résumés numériques d'une des variables selon les intervalles de l'autre 71

II-2-2. Évolution de l'ensemble des valeurs d'une des variables en fonction de l'autre 87

ÉTUDE DES RELATIONS ENTRE VARIABLES QUANTITATIVES CONTINUES ET VARIABLES QUALITATIVES 92

I. LIEN ENTRE UNE VARIABLE QUANTITATIVE ET UNE OU PLUSIEURS VARIABLES QUALITATIVES 94

I-1. La répartition des valeurs d'une variable quantitative selon les modalités d'une variable qualitative 94

I-1-1. Représentation de l'ensemble des valeurs de la variable quantitative et modalités de la variable qualitative 94

I-1-2. Représentation de résumés numériques de la variable quantitative et modalités de la variable qualitative 98

I-2. Exemple d'analyse : le salaire en fonction de la situation 100

II. LIENS ENTRE DEUX VARIABLES QUANTITATIVES ET UNE OU PLUSIEURS VARIABLES QUALITATIVES 101

II-1. Influence d'une variable quantitative sur la liaison entre deux variables qualitatives 101

II-1-1. Influence d'une variable qualitative sur la représentation des individus en fonction des valeurs des deux variables 102

II-1-2. Influence de la variable qualitative sur la densité de deux variables quantitatives 105

II-1-3. Exemple d'analyse : le revenu hors salaire en fonction du salaire selon la situation familiale 108

II-2. Influence de deux variables qualitatives sur la liaison entre deux variables quantitatives 109

II-2-1. Représentation des valeurs des deux variables quantitatives : répartition par modalité d'une variable qualitative et symbolisation par les modalités de l'autre 110

II-2-2. Représentation des valeurs des deux variables qualitatives : répartition des individus selon les modalités des deux variables quantitatives 113

CONCLUSION 115

BIBLIOGRAPHIE 118

ANNEXES 120

LISTE DES GRAPHIQUES

		Page
Graphique n°1	Camembert : la situation matrimoniale des salariés	8
Graphique n°2	Diagramme en bâtons : le type de logement des salariés	9
Graphique n°3	Représentation unidimensionnelle du revenu par u.c.	12
Graphique n°4	Représentation unidimensionnelle du revenu par u.c. avec décalage vertical des points qui se superposent	13
Graphique n°5	Représentation unidimensionnelle du revenu avec décalage vertical des points qui se superposent	14
Graphique n°6	Histogramme du revenu par u.c. en 5 classes	17
Graphique n°7	Histogramme du revenu par u.c. en 14 classes	18
Graphique n°8	Diagramme stem and leaf du revenu par u.c.	20
Graphique n°9	Diagramme stem and leaf du revenu	21
Graphique n°10	Représentation de la densité du revenu par u.c. h=800	23
Graphique n°11	Estimation lissée de la densité du revenu par u.c. h=800	25
Graphique n°12	Estimation lissée de la densité du revenu par u.c. h=1600	26
Graphique n°13	Q-plot du revenu par u.c.	29
Graphique n°14	Box plot du revenu par u.c.	32
Graphique n°15	Représentation de la symétrie du revenu par u.c.	33
Graphique n°16	Diagramme stem and leaf du salaire par u.c.	37
Graphique n°17	Courbe de densité (h=1000) et représentation unidimensionnelle du salaire par u.c.	38
Graphique n°18	Box plot du salaire par u.c.	40
Graphique n°19	Tracé de symétrie du salaire par u.c. pour les valeurs proches du mode	41
Graphique n°20	QQ-plot du revenu par u.c. en fonction du salaire par u.c.	47
Graphique n°21	QQ-plot du revenu hors salaire par u.c. en fonction du revenu par u.c.	49
Graphique n°22	Comparaison des diagrammes stem and leaf pour le salaire et pour le revenu par u.c.	50
Graphique n°23	Superposition des tracés de densité du revenu, du salaire et du revenu hors salaire par u.c. h=1000	53
Graphique n°24	Box plot du revenu, du salaire et du revenu hors salaire par u.c.	56
Graphique n°25	Box plot du revenu, du salaire et du revenu hors salaire par u.c. Variables centrées	57
Graphique n°26	Le revenu hors salaire par u.c. en fonction du salaire par u.c.	62

Graphique n°27	Le revenu hors salaire en fonction du salaire par u.c. Le nombre d'observations par point est symbolisé par des étoiles	63
Graphique n°28	Cellulation du revenu hors salaire et du salaire par u.c.	64
Graphique n°29	Densité du revenu hors salaire et du salaire (méthode d'affûtage)	68
Graphique n°30	Box plot du revenu hors salaire en fonction du salaire réparti en 4 classes	71
Graphique n°31	Evolution de la médiane du revenu hors salaire en fonction du salaire réparti en 15 intervalles	74
Graphique n°32	Trace médiane du revenu hors salaire en fonction du salaire calculée à partir de 15 intervalles	77
Graphique n°33	Trace médiane du revenu hors salaire en fonction du salaire calculée à partir de 8 intervalles	78
Graphique n°34	Trace médiane empirique et traces médianes positive, négative, neutre pour le revenu hors salaire en fonction du salaire (calculée à partir de 15 intervalles)	80
Graphique n°35	Ratio de dépendance négative du revenu hors salaire en fonction du salaire calculé à partir de 15 intervalles	81
Graphique n°36	Traces quartiles et médiane du revenu hors salaire en fonction du salaire calculées à partir de 15 intervalles	84
Graphique n°37	Box plot à deux dimensions du revenu hors salaire en fonction du salaire calculée à partir de 15 intervalles	86
Graphique n°38	Estimation du centre de la distribution du revenu hors salaire en fonction du salaire	89
Graphique n°39	Représentation unidimensionnelle du salaire et situation familiale	95
Graphique n°40	Représentations unidimensionnelles du salaire pour chaque situation familiale	96
Graphique n°41	Diagramme stem and leaf selon les modalités de la situation familiale	97
Graphique n°42	Box plot du salaire par modalité de la situation familiale	99
Graphique n°43	Représentation du revenu hors salaire en fonction du salaire Individus identifiés par leur situation familiale	102
Graphique n°44	Représentation du revenu hors salaire en fonction du salaire selon la situation familiale	104
Graphique n°45	Densité du revenu hors salaire et du salaire représentée par la méthode d'affûtage Individus identifiés par leur situation familiale	106
Graphique n°46	Revenu hors salaire en fonction du salaire, par la situation familiale et symbolisé par le nombre d'enfants	110
Graphique n°47	Représentation en fenêtre du revenu hors salaire en fonction du salaire selon la situation familiale et le nombre d'enfants	112

Introduction

La statistique *exploratoire* est la première étape d'une analyse de données. Comme son nom l'indique, pendant ce temps de travail, le statisticien découvre les données, en synthétisant, résumant, structurant l'information qu'elles contiennent. Elle est aussi appelée statistique *descriptive*, justement parce qu'elle a pour objectif de décrire les données en mettant en évidence leurs principales propriétés. Elle précède souvent un deuxième temps de l'analyse, consacré à la statistique *inférentielle*. Il s'agit alors d'étendre les propriétés constatées sur l'échantillon à l'ensemble de la population par le biais de méthodes probabilistes, de tests, de modélisations...

Dans l'exploitation de données d'enquêtes, la statistique exploratoire constitue la part dominante de l'analyse. Pour cette étape, le statisticien utilise des représentations des données sous forme de tableaux, d'indicateurs numériques,... Un certain nombre de ces techniques font intervenir des représentations graphiques.

En France, les méthodes graphiques sont surtout mises en oeuvre pour l'analyse multidimensionnelle des données : analyse factorielle des correspondances, analyse canonique, analyse discriminante, classification... Ces méthodes permettent une approche globale des individus et des variables, représentés sur des espaces à deux ou trois dimensions. Cependant, les visualisations d'une seule variable ou d'un nombre restreint de variables (2 à 4) sont peu employées pour l'analyse descriptive. Pourtant, ces graphiques, souvent élaborés à partir d'outils statistiques ou mathématiques relativement simples, sont des outils d'analyse particulièrement puissants. Dans de nombreuses situations, quelques graphiques bien choisis remplacent les résumés numériques classiques, tout en apportant des informations supplémentaires sur l'allure des données. J.F. ANSCOMBE[1] a construit un exemple qui permet de bien comprendre l'intérêt de cette démarche : il a inventé quatre séries de deux variables (x,y), pour lesquelles les résultats de la régression sont identiques, alors que représentées sur un plan, les quatre séries de points sont en fait très différentes... Une analyse fondée seulement sur les résultats de la régression, sans confrontation avec la visualisation graphique, aurait entraîné des conclusions erronées.

C'est en particulier la comparaison de plusieurs graphiques qui permet d'enrichir l'analyse.L'amélioration des logiciels statistiques, l'augmentation de leur rapidité d'exécution et

---

[1] F.J. ANSCOMBE, Graphs in statistical analysis, The American Statistician, 1973.

de leur capacité de mémoire permet d'éditer en peu de temps un grand nombre de graphiques. Ainsi, un module de SAS, « SAS insight », entièrement consacré à l'analyse exploratoire des données, est maintenant disponible. Ces outils récents permettent de concevoir une nouvelle pratique plus interactive de l'analyse descriptive. Les temps de calcul courts, les nombreuses possibilités de manipulation sur les variables et sur les observations permettent de multiplier les graphiques.

Ce cahier de recherche a donc pour objectif de redécouvrir l'intérêt des graphiques en analyse exploratoire. Depuis les années 70, plusieurs recueils recensant les différentes méthodes graphiques uitlisables en cours d'analyse ont été publiés. En particulier, « Exploratory data analysis » de John W. TUKEY publié en 1977, et « Graphical methods for data analysis » écrit en collaboration par John M. CHAMBERS, William S. CLEVELAND, Beat KLEINER, Paul A. TUKEY en 1983 nous ont servi de référence dans nos recherches. Ces travaux sont très complets, et assez généraux, dans le sens où ils sont destinés à être utilisés par des publics très différents (statisticiens, ingénieurs, enseignants, médecins) et recensent des méthodes applicables sur toutes sortes de données. Rappelons que les variables d'une analyse sont soit qualitatives, les modalités de la variable sont alors en nombre fini et correspondent à des catégories (le sexe admet deux modalités homme et femme) ; soit quantitatives, les modalités sont alors des nombres, sur lesquels les opérations arithmétiques ont un sens. Pour les variables quantitatives, la variable est dite discrète si elle n'admet qu'un nombre fini ou dénombrable de valeurs (par exemple le nombre d'enfants dans une famille), la variable est dite continue si elle peut prendre toutes les valeurs d'un intervalle de R (par exemple le chiffre d'affaires des entreprises de la région parisienne).

Dans ce cahier de recherche, notre démarche est moins générale que dans les travaux cités précédemment. Il s'agit surtout de s'intéresser aux méthodes utilisables dans le cadre de données d'enquête semblables à celles exploitées au département Evaluation des Politiques Sociales du CRÉDOC, c'est à dire pour des données portant sur un nombre d'observations relativement restreint (inférieur à 800 le plus souvent), avec un grand nombre d'informations qualitatives et peu de variables quantitatives continues.

Afin d'améliorer la cohérence de ce travail, nous avons choisi de présenter l'ensemble des méthodes recensées à partir d'un nombre restreint de variables d'une même enquête. Il est alors plus facile de se rendre compte de ce qu'apporte comme informations supplémentaires un nouveau graphique par rapport au précédent.

L'enquête utilisée a été réalisée à la demande du Conseil Economique et Social auprès d'un échantillon de 754 personnes. Elle avait pour objectif essentiel de décrire les effets des politiques sociales sur des populations en situation de grande pauvreté. La méthode d'échantillonnage employée privilégiait un repérage de la population par sa localisation sur des sites présumés contenir une forte proportion de personnes pauvres. La liste des mesures et des politiques évaluées comprenait les principales politiques d'aide financière (RMI, prestations familiales, prestation chômage, allocation logement), complétées de mesures plus générales d'aide à la famille, de formation, d'accès à l'emploi, d'accès aux soins. Toutes ces données, ainsi que les principales caractéristiques socio-démographiques, ont été recueillies au cours d'entretiens en face-à-face.

Ces données ont déja fait l'objet d'une exploitation, utilisant les méthodes usuelles d'analyse descriptive des données[2]. Parmi les 754 personnes interrogées, un peu moins du tiers des foyers (constitués de la personne interrogée elle-même, son conjoint et des personnes à leur charge) perçevait un salaire au moment de l'enquête. Ce salaire est en moyenne inférieur au SMIC, de nombreux salariés exerçant une activité dans le cadre d'un contrat aidé (essentiellement un Contrat Emploi Solidarité) ou d'un emploi à temps partiel. Ces 193 foyers sont donc à la fois actifs et pauvres[3].

Dans ce cahier, ces foyers, qui sont en situation de pauvreté malgré un membre en activité, ont été isolés afin de constituer un nouvel échantillon d'étude. Une dizaine de variables socio-démographiques ont été selectionnées.

* Des variables quantitatives : le revenu et le salaire global ou par unité de consommation (u.c.). Le calcul des unités de consommation considéré est celui de l'echelle d'Oxford (1 pour la personne de référence, 0,7 pour le conjoint et les enfants de plus de 18 ans, 0,5 pour les enfants de moins de 18 ans). La variable revenu par u.c. a été calculée à 100 francs près. Elle ne prend donc que des valeurs multiples de 100. Rigoureusement, il s'agit d'une variable quantitative discrète. Cependant, toutes les autres valeurs ont un sens et peuvent potentiellement être prises par la variable. On parlera donc du revenu par u.c. comme étant une variable quantitative continue.

---

[2] Marie-Odile GILLES-SIMON, Michel LEGROS, *Politiques Sociales : l'épreuve de la pauvreté*, collection des rapports du CRÉDOC, n°159, avril 1995.

[3] Marie-Odile GILLES-SIMON, Michel LEGROS, *Pour certains, le travail n'empêche plus la pauvreté*, Consommation et modes de vie, n°100, septembre-octobre 1995.

* Des variables qualitatives : le logement, la situation matrimoniale, le nombre d'enfants. Ces variables ont été choisies parce qu'elles décrivent les caractéristiques du foyer et qu'elles sont apparues discriminantes sur le revenu.

Dans ce cahier, les méthodes décrites sont présentées selon le nombre et le type de variables qu'elles nécessitent pour être mises en oeuvre. Le premier chapitre est consacré aux graphiques représentant une seule variable à la fois. Les représentations de variables qualitatives sont peu nombreuses, celles de variables quantitatives sont plus variées. Il faut alors distinguer les graphiques qui représentent l'ensemble des valeurs, de ceux qui font intervenir des résumés numériques.

Le deuxième chapitre présente les méthodes de visualisation des relations entre deux variables quantitatives continues, en appréhendant les variables dans leur ensemble, ou en raisonnant individu par individu.

Enfin, le dernier chapitre présente les méthodes qui font intervenir sur un même graphique variables qualitatives et variables quantitatives.

Toutes les méthodes présentées dans ce cahier ont été mises en oeuvre sur l'échantillon des 193 foyers de l'enquête pauvreté selectionné. Les traitements statistiques ont été réalisés sur les logiciels SAS et Exel. Le détail des procédures utilisées pour la mise en oeuvre de ces méthodes, ainsi que le fichier complet des variables selectionnées, sont donnés en annexe.

Représentations graphiques monofactorielles

La première étape d'une étude statistique est monofactorielle : il s'agit d'observer le comportement de l'ensemble des individus de l'échantillon par rapport à chacune des variables étudiées. On décrit la distribution de la variable, c'est-à-dire la façon dont les individus se répartissent entre les différentes valeurs ou modalités de la variable.

I- CAS D'UNE VARIABLE QUALITATIVE

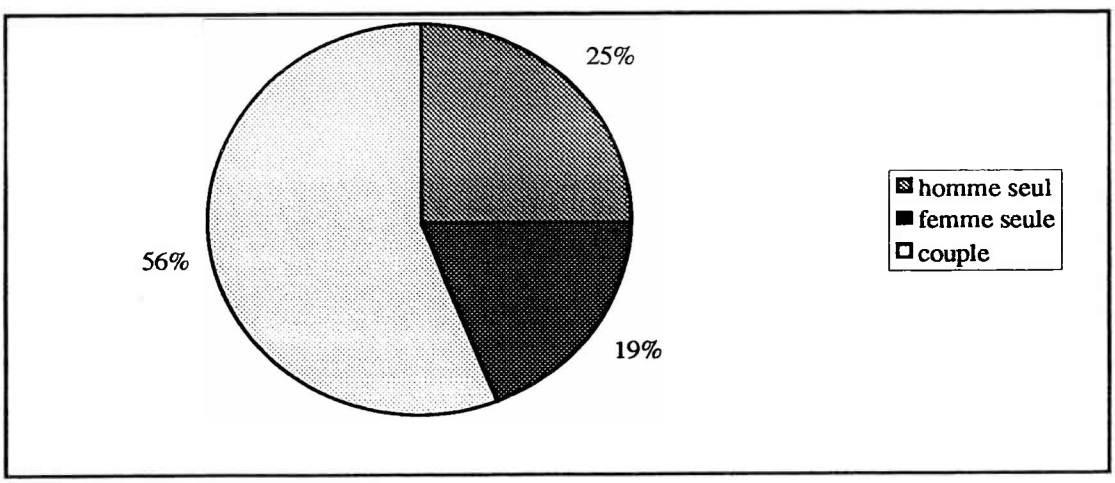
Pour chacune des modalités de la variable, on s'intéresse au nombre et à la proportion d'observations réalisant cette modalité (c'est-à-dire la fréquence de la modalité).

Graphiquement, on représente alors les modalités de la variable par des figures pouvant être très diverses (étoiles, figurines...), la règle générale étant que les dimensions de la figure soient proportionnelles à la fréquence ou au nombre d'observations. Les représentations les plus utilisées sont les diagrammes en bâtons (la fréquence des observations est proportionnelle à la hauteur des bâtons) ou les camemberts (la fréquence des observations est proportionnelle à l'aire des portions).

Les graphiques suivants permettent ainsi de répondre à la question « comment se répartissent les personnes en situation de grande pauvreté dans chacun des types de logement ? ».

Graphique n°1

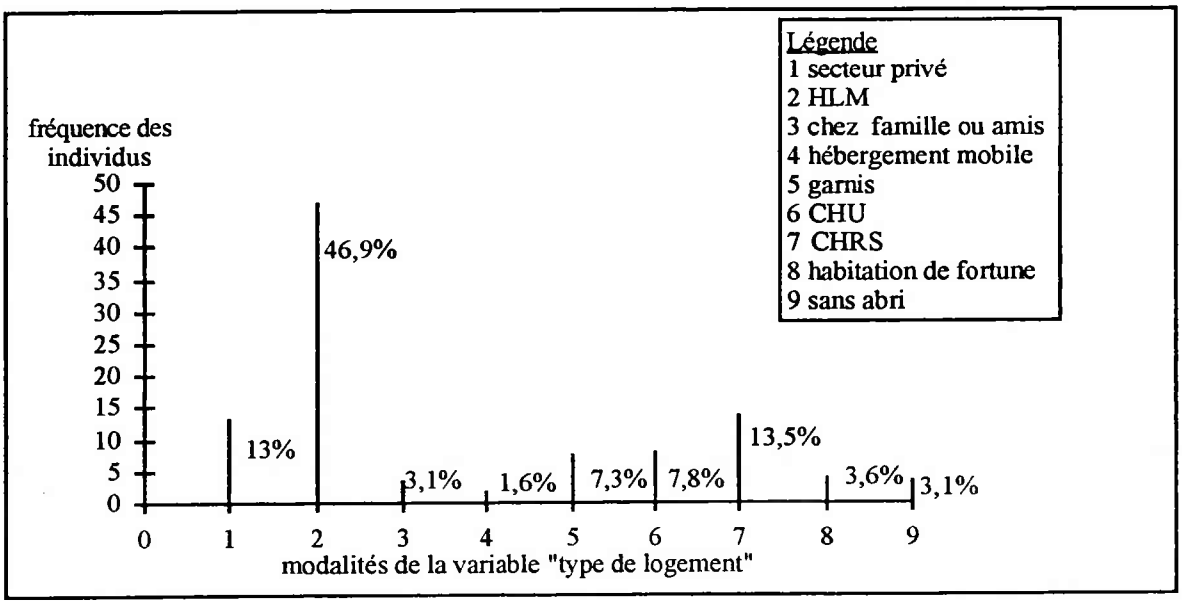
Camembert : la situation matrimoniale des salariés



Source : enquête Pauvreté CREDOC-CES 1995

Graphique n° 2

Diagramme en bâtons : le type de logement des salariés



Source : enquête Pauvreté CREDOC-CES 1995

II- CAS D'UNE VARIABLE QUANTITATIVE

La distribution est décrite par la position, la dispersion et la forme de la variable.

-la position de la variable : traduit le comportement général des individus. Son étude se fait par la recherche des valeurs prises le plus souvent par la variable ou représentant le mieux l'ensemble des individus (caractéristiques de tendance centrale : mode, moyenne, médiane) et par celle des valeurs extrêmes.

LES CARACTÉRISTIQUES DE TENDANCE CENTRALE

Leur but est de représenter au mieux l'ensemble des valeurs prises par la variable sur l'échantillon. Elles correspondent au centre de la distribution.

La moyenne : Soit x, une variable observée sur un échantillon de n individus. On note xi la valeur de x pour l'individu i. La moyenne est alors donnée par : $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} xi$. Toutes les observations entrent en ligne de

compte pour le calcul de cet indicateur. La présence de valeurs très éloignées de l'ensemble de l'échantillon le modifie donc considérablement.

La médiane : c'est la valeur telle qu'il y ait autant d'observations ayant une valeur inférieure que d'observations ayant une valeur supérieure. En notant x(i) les valeurs ordonnées de la variable x, si n est impair alors la médiane est M=x((n+1)/2), si n est pair, par convention on prend :

$$M = \frac{x(n/2) + x(n/2-1)}{2}$$

La médiane ne tient pas compte de la valeur exacte de la variable sur toutes les observations, mais uniquement de l'ordre de ces valeurs. Elle est robuste à l'éloignement des valeurs extrêmes mais par contre n'a pas de propriétés arithmétiques.

Le mode : c'est la valeur prise le plus souvent par une variable discrète. Pour une variable continue, il s'agit de l'intervalle de valeurs dans lequel il y a le plus d'observations. Cette définition pour une variable continue est peu satisfaisante, le mode dépendant alors de la largeur des intervalles considérés.

-<u>la dispersion de la variable</u> permet de savoir comment se répartissent les individus par rapport à la position de la variable. Son étude se fait par la recherche des valeurs aberrantes, par l'étude de résumés numériques (indicateurs de dispersion : écart-type, intervalle interquartile étendu par exemple), et par l'étude de la concentration et de la variation de ces concentrations des individus autour des différentes valeurs prises par la variable (fréquence, fonction de répartition, densité).

LES INDICATEURS DE DISPERSION

Ils fournissent des précisions sur les écarts entre les valeurs prises par la variable.

L'étendue est la différence entre la valeur maximale prise par la variable et la valeur minimale. Elle ne dépend que de ces deux valeurs et n'est pas robuste.

L'intervalle interquartile : le quartile inférieur (respectivement supérieur), traditionnellement noté Q1 (resp. Q3) est la valeur telle que pour 25% (resp 75%) des observations, la variable ait une valeur inférieure à Q1 (resp à Q3). L'intervalle interquartile est donné par Q3-Q1. Il est beaucoup plus robuste que l'étendue, mais ne donne pas d'information sur la dispersion des valeurs à l'extérieur de l'intervalle.

L'écart-type : il traduit la façon dont les données s'écartent en moyenne de la moyenne. Avec les mêmes notations que précédemment, c'est le nombre s tel que :

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(xi - \bar{x})^2}$$

Comme la moyenne, il est sensible aux valeurs extrêmes.

FONCTIONS DE DENSITÉ ET DE RÉPARTITION

Fonction de répartition : on note conventionnellement F cette fonction. F(xi) est alors le pourcentage d'observations pour lesquelles la valeur de la variable est inférieure ou égale à xi. C'est la fréquence de l'intervalle $]-\infty; xi]$.

Fonction de densité : elle est notée f. f(xi) est la fréquence d'un intervalle infiniment petit centré sur xi. C'est la dérivée de la fonction de répartition. Formellement, on a :

$$\int_{xi}^{xj} f(t)dt = F(xj) - F(xi)$$

La fonction de densité n'est donc pas accessible directement à partir d'une distribution empirique.

-la forme de la variable traduit la vitesse et la régularité avec laquelle le nombre d'individus décroît lorsque les valeurs de la variable s'éloignent des valeurs centrales, de part et d'autre de ces valeurs. Elle est de plus souvent utilisée pour comparer une distribution empirique à une distribution théorique caractérisée par sa forme. On observe en particulier la symétrie de la variable.

LA SYMETRIE

Une variable est symétrique si les valeurs de la variable de part et d'autre de son centre sont les mêmes. Dans ce cas, la médiane est égale à la moyenne et au mode.

En notant comme précédemment x(i) les valeurs ordonnées de la variable, on a alors, pour tout i compris entre 0 et la partie entière de (n+1)/2 : m-x(i)=x(n-i)-m, où m désigne indifféremment le mode, la moyenne ou la médiane. De plus, l'écart-type d'une telle distribution est nul.

Dans le cas d'une variable continue discrète, l'étude de sa position, de sa dispersion et de sa forme n'aura de sens que si elle a assez de modalités. La plupart des représentations graphiques présentées ci-dessous pourront alors être utilisées. Pour un faible nombre de modalités, la variable sera traitée comme une variable qualitative.

Chaque représentation graphique permet de visualiser, avec ses atouts et ses limites, une ou plusieurs de ces caractéristiques (position, dispersion, forme). L'étude de la variable se fait par une comparaison des différents éléments apportés par chacun des graphiques.

On distinguera les graphiques représentant l'ensemble des valeurs prises par la variable de ceux s'intéressant uniquement aux résumés numériques de ces valeurs.

II-1 Représentation de l'ensemble des valeurs prises par la variable

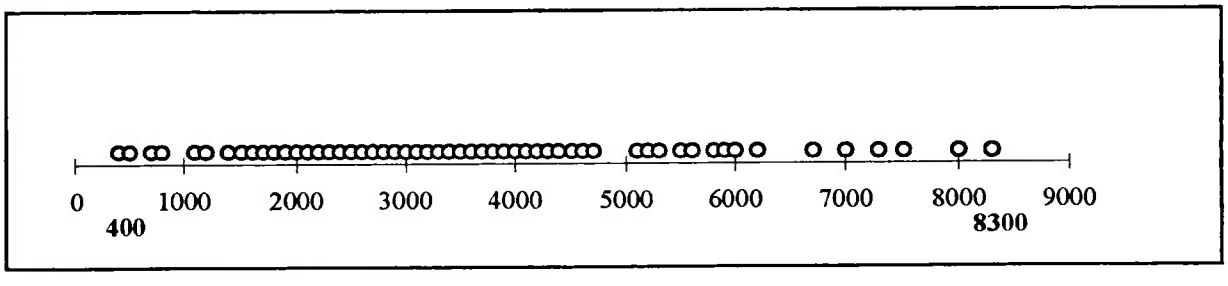
Les graphiques présentés dans ce paragraphe permettent de visualiser l'ensemble des données : une des coordonnées des points représentés sera la valeur effectivement prise par la variable. Les premiers graphiques sont unidimensionnels. Les suivants font intervenir une deuxième dimension proportionnelle à la fréquence des observations pour les différentes valeurs de la variable.

II-1-1 Représentation unidimensionnelle des données

Chacune des observations est représentée par un point porté sur un axe gradué par les valeurs de la variable.

Graphique n° 3

Représentation unidimensionnelle du revenu par unité de consommation



Source : enquête Pauvreté CREDOC-CES 1995

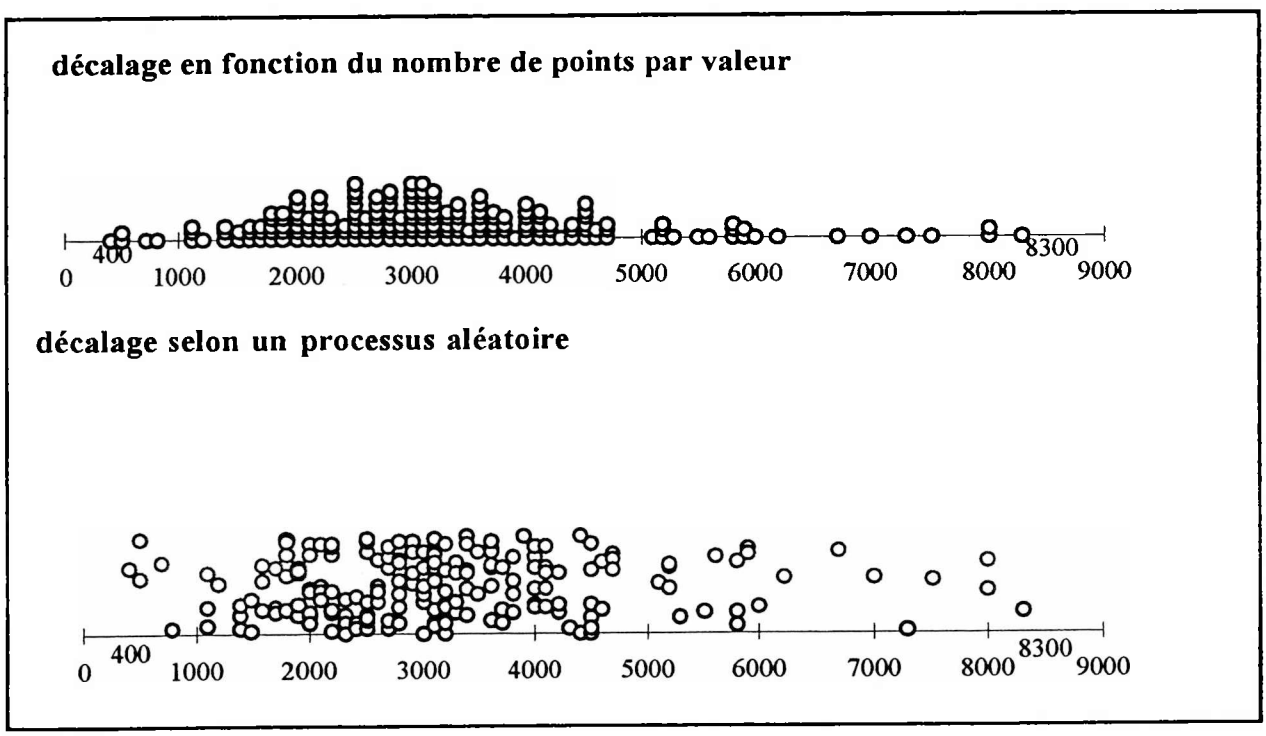
Le graphique montre les valeurs extrêmes prises par la variable. De plus, il permet de situer les intervalles de valeurs où se concentrent les individus. Cependant, les points qui prennent la même valeur se superposent et ne sont pas tous représentés.

Dans cet exemple, le revenu minimum par u.c. est de 400 francs, le revenu maximum est de 8300 francs. La plupart des personnes interrogées ont un revenu situé entre 1200 et 4800 francs. On ne peut voir sur ce graphique quelles sont précisément les différences de concentration des individus dans l'intervalle 1800-4400 francs et dans celui 5800-6000 francs.

CHAMBERS, CLEVELAND, KLEINER et TUKEY[4] proposent alors de décaler verticalement les points qui se superposent. Le décalage est ou bien proportionnel au nombre de points par valeur, ou bien aléatoire.

Graphique n° 4

Représentation unidimensionnelle du revenu par u.c. avec décalage vertical des points qui se superposent



Source : enquête Pauvreté CREDOC-CES 1995

---

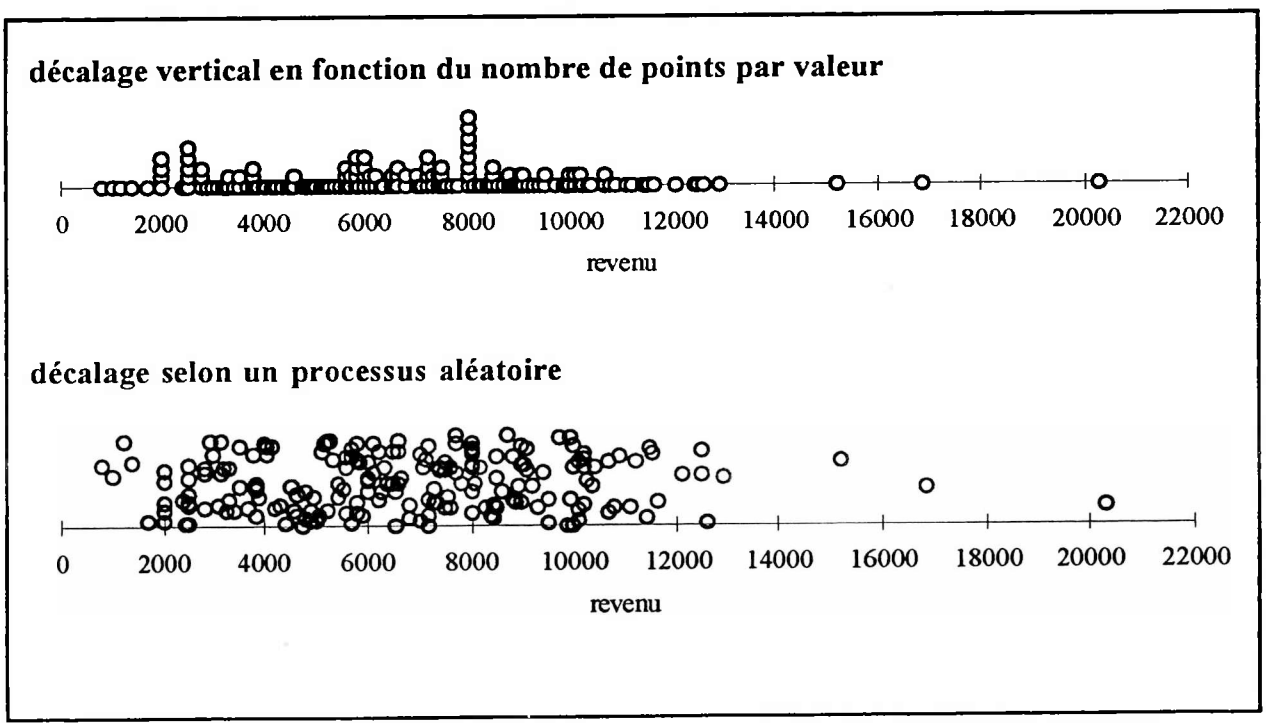
[4] J.M CHAMBERS, W.S.CLEVELAND, B. KLEINER, P.A. TUKEY, *Graphical methods for data analysis*, Boston Duxbury Press, 1983.

Dans ces graphiques, nous avons appelé représentation unidimensionnelle une représentation qui fait en fait appel à deux dimensions. Pour le deuxième graphique, la dimension verticale n'a pas de sens, car elle est issue d'un processus aléatoire. Pour le premier graphique, la dimension verticale représente le nombre de points par valeur de la variable. Ce graphique s'apparente donc à un histogramme (voir paragraphe suivant) et est donc strictement une représentation à deux dimensions.

Cette impression est renforcée dans le cas du revenu par u.c., qui est arrondi aux centaines et est en ce sens une variable discrète. Le nombre de modalités admissibles entre 0 et 8300 francs étant fini, il y a beaucoup de points superposés et on obtient clairement des « bâtons » sur chacune des centaines. Le graphique suivant montre les représentations unidimensionnelles dans le cas d'une variable continue non discrétisée. Il y a beaucoup moins de points qui se superposent que dans le cas du revenu par u.c., et le graphique tracé à partir du décalage vertical fonction du nombre de points ne donne plus une impression d'histogramme.

Graphique n°5

Représentations unidimensionnelles du revenu avec décalage vertical des points qui se superposent



Source : enquête Pauvreté CREDOC-CES 1995

Le but de ce tracé est de fournir une représentation rapide et concise de l'ensemble des valeurs prises par la variable. Le décalage vertical des points n'a d'importance que dans la mesure où il permet de dessiner toutes les observations. Sa détermination est fonction de la place verticale qu'il exige et non de la possibilité d'obtenir un comptage des points : on s'intéresse ici à la densité des points en fonction des valeurs de la variable et non au nombre de points par valeur (comptage qui n'aurait pas de sens dans le cas d'une variable réellement continue).

Pour un faible nombre d'observations (dans notre exemple moins de 200 observations) sur une variable discrète, le décalage en fonction du nombre de points par valeur donne une représentation plus claire que le décalage aléatoire. Le premier permet notamment de voir, au-delà de la concentration des observations dans certains intervalles, quelles sont les valeurs prises le plus souvent et donc de repérer le mode de la variable.

Ici, la première méthode montre dans le cas du revenu par u.c., que la concentration des points autour des revenus de 6000 francs est nettement plus faible que celle des revenus autour de 3000 francs. Cette différence n'apparaissait pas à la lecture du premier graphique unidimensionnel.

Pour un nombre d'observations plus grand, la première méthode se heurte à un problème de superposition horizontale des points et demande rapidement une place verticale importante. Le décalage aléatoire est alors préférable. Pour une variable réellement continue, la première méthode ne sépare pas clairement les points qui prennent deux valeurs différentes très voisines : la représentation des points sur l'axe horizontal est confuse.

La représentation unidimensionnelle de la variable permet de visualiser l'étendue de la distribution ainsi que les zones de forte concentration des observations. Elle fournit une première idée de la valeur du mode. Elle peut cependant conduire à des ambiguïtés dans l'interprétation, dues notamment au problème de superposition des points. Elle ne suffit pas à répondre quantitativement aux questions sur la distribution de la variable mais donne un bon aperçu qualitatif en particulier de la position et de la dispersion. Rapidement construite, elle sert à élaborer des premières hypothèses sur les caractéristiques de la variable. Ces hypothèses seront confirmées ou infirmées par l'étude d'autres graphiques.

Parce qu'elle est unidimensionnelle cette représentation tient peu de place. Elle peut donc être facilement tracée à côté d'un autre graphique pour le compléter.

II-1-2 Représentation bidimensionnelle : fréquence et densité en fonction des valeurs de la variable

II-1-2-a Représentation de la fréquence et variable répartie en classe

L'histogramme

La fréquence, pourcentage d'observations prenant une modalité donnée de la variable étudiée, n'a pas de sens pour une variable continue, ayant potentiellement une infinité de modalités. Pour représenter la fréquence, il faut donc « discrétiser » la variable en répartissant ses valeurs dans des intervalles (ou classes), puis calculer la fréquence de chacun des intervalles ainsi obtenus. Le graphique représentant ce calcul est l'histogramme. Il est construit en portant en abscisse les classes de la variable et en ordonnée la fréquence divisée par la largeur de l'intervalle. Par convention, l'intervalle considéré ne contient pas sa borne supérieure.

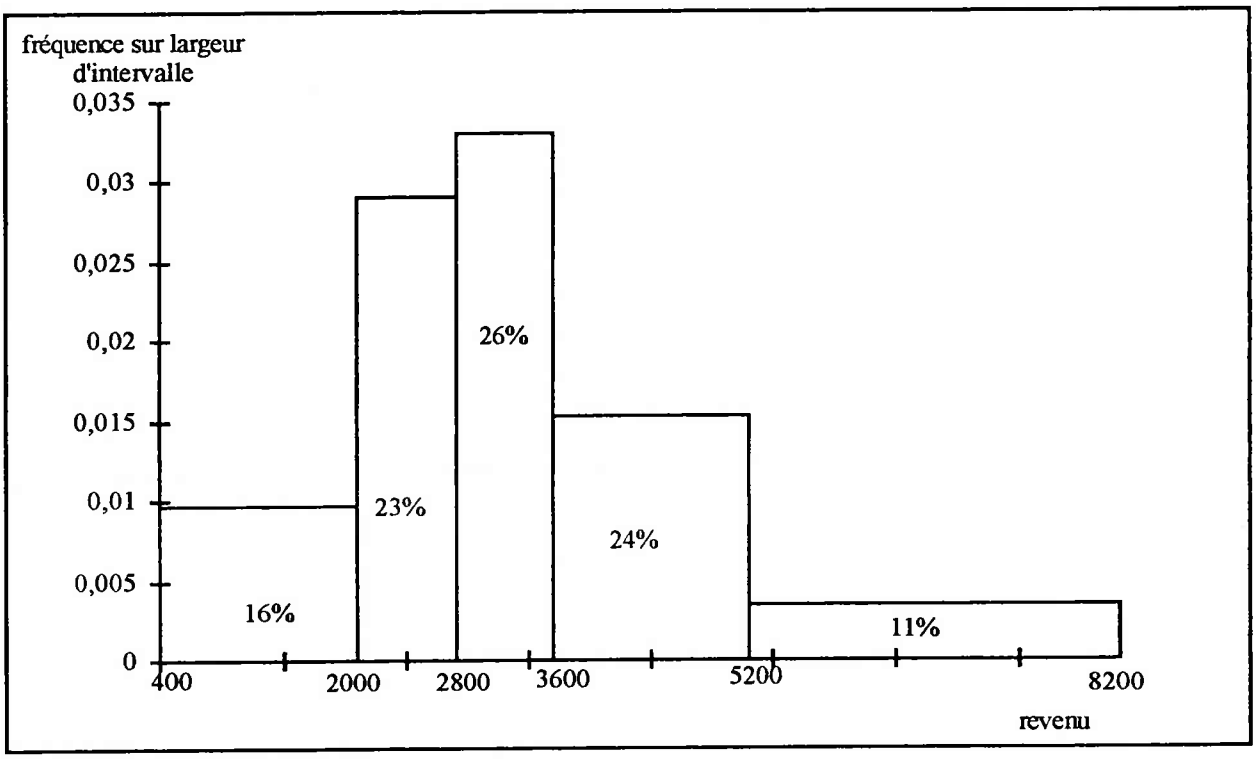
On trace sur chacun des intervalles des rectangles dont l'aire est proportionnelle à la fréquence de la classe. Ceci est cohérent avec le principe général qui veut que la taille de la figure dessinée soit proportionnelle à la valeur qu'elle représente. Dans le cas particulier d'intervalles d'amplitude égale, la hauteur des rectangles sera proportionnelle à la fréquence.

Cette représentation donne des indications sur la position de la variable : elle permet ainsi de situer le mode ainsi que les valeurs extrêmes. Elle fournit également une image de la dispersion de la variable. Elle montre les différences de concentration des individus selon les différentes valeurs du revenu. Elle donne des informations quantitatives précises en permettant de répondre à la question «quel est le pourcentage de personnes ayant un revenu situé dans chacun des intervalles considérés ? » et « comment varie ce pourcentage avec le revenu ? »

Enfin, l'observation de la variation des aires des rectangles de part et d'autre du mode donne une première idée de l'éventuelle symétrie de la distribution : si ces aires sont les mêmes de part et d'autre du mode, la distribution sera sans doute symétrique.

Graphique n°6

Histogramme du revenu par u.c. en 5 classes

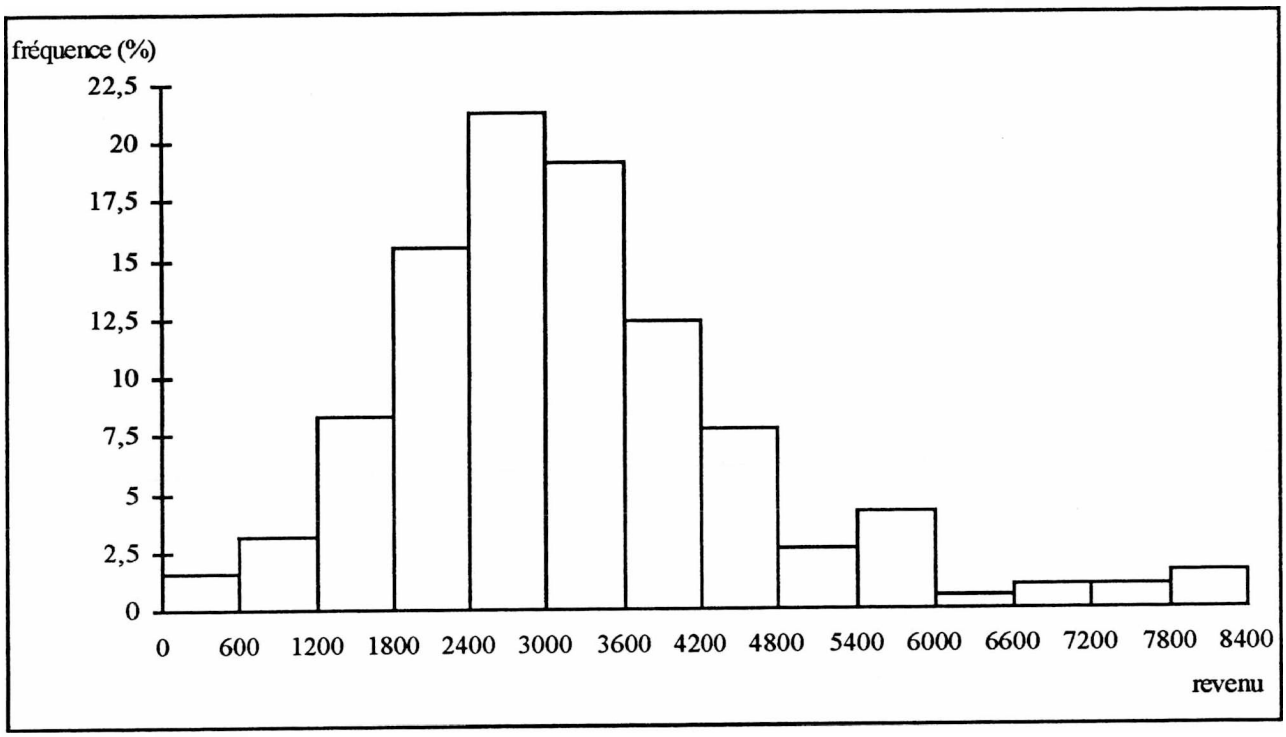


Source : enquête Pauvreté CREDOC-CES 1995

La forme de l'histogramme est très sensible au nombre et à la largeur des intervalles, qui sont choisis arbitrairement. Le graphique suivant montre un histogramme tracé sur la même variable, mais avec un choix d'intervalles différent. Les intervalles étant de même largeur, la fréquence a été portée directement en ordonnée.

Graphique n°7

Histogramme du revenu par u.c. en 14 classes



Source : enquête Pauvreté CREDOC-CES 1995

Pour le premier histogramme, le mode se situait dans l'intervalle 2800-3600, il est dans ce nouveau graphique entre 2400 et 3000. Le graphique 6 semble indiquer qu'il y a beaucoup plus d'individus ayant un revenu juste supérieur à 2000 francs que d'individus ayant un revenu juste inférieur à 2000 francs. La représentation unidimensionnelle du revenu montre que cette impression est erronée.

A la lecture de ces graphiques apparaissent des discontinuités qui ne proviennent que de la construction. Plus le nombre de classes est important et plus les discontinuités sont nombreuses. En revanche, un nombre de classe trop faible tend à effacer des variations significatives de la concentration des observations d'une valeur à l'autre de la variable. Pour que les différences observées soient significatives, il faut de plus que le nombre d'observations à l'intérieur de chacun des intervalles soit suffisant. Le choix du nombre de classes et de la largeur des intervalles doit tenir compte de ces contraintes.

S'il n'existe pas de règle générale pour la construction des histogrammes, il semble cependant pertinent de :

-choisir les valeurs extrêmes empiriques comme bornes respectivement inférieures et supérieures du premier et du dernier intervalle ;
-choisir des intervalles de largeur plus faible aux environs de la médiane, plus élevée lorsque la concentration des valeurs diminue ;
-tenir compte du nombre total d'observations dans le choix du nombre d'intervalles.

Le premier histogramme qui a été construit selon ces règles après observation des représentations unidimensionnelles, semble plus significatif que le deuxième, pour lequel beaucoup d'intervalles comptent moins de 5 observations.

L'histogramme est un graphique simple qui, sous réserve d'être construit de façon raisonnée, donne un aperçu de la position et notamment du mode, de la dispersion et en particulier des variations de la concentration des observations, de la forme et notamment de la symétrie. Il permet de répondre quantitativement aux questions portant sur la répartition des individus dans les intervalles de valeurs.

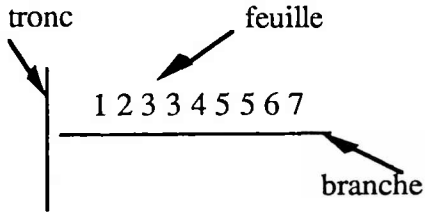
Son interprétation est cependant ambiguë. L'histogramme est en effet fortement dépendant du choix arbitraire de la largeur des intervalles. Il a le défaut de présenter sous forme discontinue une variable continue. Les précautions que demandent sa construction en font un instrument difficile à utiliser pour l'analyse, mais très utile pour la présentation des résultats.

Diagramme « stem and leaf »

Une forme particulière de l'histogramme est le diagramme « stem and leaf » (Tukey, 1977[5]). Il est constitué à partir des valeurs numériques de la variable. Le tronc est formé du premier chiffre des données (il correspond donc aux dizaines, centaines, milliers... suivant les cas). Les feuilles sont les enregistrements des chiffres restant pour chacune des données et forment les branches.

---

[5] J.W. Tukey *Exploratory data analysis*, Addison&Wesley, 1977.



Graphique n°8

Diagramme stem and leaf du revenu par u.c.

		nombre d'observations
8	003	3
7	5	1
7	03	2
6	71	2
6	02	2
5	5688899	7
5	12223	5
4	55555566777	11
4	000000111112223444	18
3	55666666677777888889	19
3	000000000111111111222222223333333444444	37
2	555555555666677777777888888889999	32
2	00000001111222222223333444	25
1	556667778888899999	18
1	1112444	7
0	5578	4
0	4	1
total		193
les données doivent être multipliées par 100		

Source : enquête Pauvreté CREDOC-CES 1995

Guide de lecture : chacune des feuilles des branches de droite complète le tronc de la colonne de gauche. Le chiffre obtenu est ensuite multiplié par 100. Les données les plus élevées sont ainsi 8300, 8000, 8000, 7500, 7300, 7000...

D'une manière générale, on n'inscrit dans le tronc qu'une seule fois le premier chiffre des données : les branches correspondront alors à des intervalles de 10, 100, 1000... selon la forme des données. Cependant, lorsque les branches construites ainsi sont trop longues (autour du mode notamment), elles compromettent la lisibilité du graphique. On construit alors le diagramme en inscrivant dans la colonne de gauche deux fois chaque premier chiffre des

données, toujours en les plaçant les uns en-dessous des autres. Le tronc comporte deux fois chacun des chiffres, les tiges recouvrent dans ce cas des intervalles de largeur 5, 50, 500..., les feuilles étant comprises entre 0 et 4 puis entre 5 et 9. C'est ce qui s'est produit dans l'exemple retenu.

Les données utilisées étant arrondies aux centaines, les feuilles sont constituées d'un seul chiffre. Lorsque les valeurs ont plus de deux chiffres significatifs, les feuilles peuvent comprendre 2 chiffres ou plus. Des astérisques sont dessinés à côté de chacune des bases du tronc, afin d'indiquer le nombre de chiffres par feuille (un astérisque par chiffre).

Graphique n°9

Diagramme stem and leaf du revenu

		nombre d'observations
20**	30	1
19**		
18**		
17**		
16**	87	1
15**	20	1
14**		
13**		
12**	0946506091	5
11**	102045505364	6
10**	000000091010192020232737407070799O	17
9**	00000210102030405050749093 95	14
8**	000000000000000001021304043475050506070808085909 7	24
7**	000005122020202030303140405050545960696970 80	22
6**	000000001010202028333750505255606060658080	21
5**	000211162023293039405060606066707080808080859O	23
4**	000206152028405056606075798088909 5	16
3**	000911122027303040505066788080808590	17
2**	000000000041414850505050808080 90	16
1**	00204070	4
0**	80	1
total		193

les données doivent être multipliées par 10

Source : enquête Pauvreté CREDOC-CES 1995

guide de lecture : le chiffre du tronc est complété cette fois par deux chiffres de la branche. Les données enregistrées à la tige 12 sont ainsi 12090, 12460, 12500, 12600, 12910.

La particularité du graphique stem and leaf par rapport aux histogrammes est de représenter les données non par des rectangles, mais par des branches constituées par les valeurs des données elles-mêmes et donc dont les longueurs respectives sont proportionnelles à la fréquence des intervalles (on admet le fait que les caractères utilisés ont tous la même largeur). Ce diagramme a ainsi les propriétés graphiques des histogrammes tout en permettant un enregistrement exact des données.

Il répond à des questions quantitatives précises sur :

-la position de la variable : le fait d'inscrire à droite du graphique le nombre d'observations par branche permet de retrouver facilement les quantiles et notamment la médiane.
-la dispersion de la variable : on retrouve les valeurs extrêmes. L'existence d'un certain nombre de branches vides entre les valeurs extrêmes de la variable et le reste des données permet de repérer les valeurs aberrantes ou déviantes.
-la forme de la variable : comme sur un histogramme, on peut voir apparaître les « pics » de données, l'éventuelle symétrie ou asymétrie.

Tout comme l'histogramme, le diagramme stem and leaf pose le problème de présenter de façon discontinue une variable continue. Il donne l'impression visuelle de « sauts » dans les valeurs de la variable, sauts qui ne proviennent que de la construction. Il ne peut de plus être utilisé que pour un faible nombre d'observations (de préférence inférieur à 200 observations). Pour un nombre d'observations supérieur, les branches sont trop longues pour être lisibles.

Enfin, pour certaines variables, sa construction demande d'arrondir les données ou de construire des feuilles à plus d'un chiffre. Augmenter le nombre de chiffres par feuille permet un enregistrement plus précis des valeurs de la variable mais se fait au détriment de la lisibilité. Ainsi, si on passe de feuilles à un chiffre à des feuilles à deux chiffres, les tiges sont deux fois plus longues et l'impression de discontinuités plus importante : une observation de plus par tige se traduit par l'inscription de deux caractères supplémentaires.

II-1-2-b Représentation continue à deux dimensions : les estimations de densité

La densité au point y est la fréquence d'un intervalle infinitésimal centré sur y. L'histogramme est une estimation grossière de la densité au centre de chacun des intervalles utilisés pour sa construction.

Une première méthode pour estimer la densité consiste à calculer la fréquence d'intervalles de largeur h centrés sur chacune des valeurs prises par la variable. Si y est une de ces valeurs, la densité estimée en y est alors :

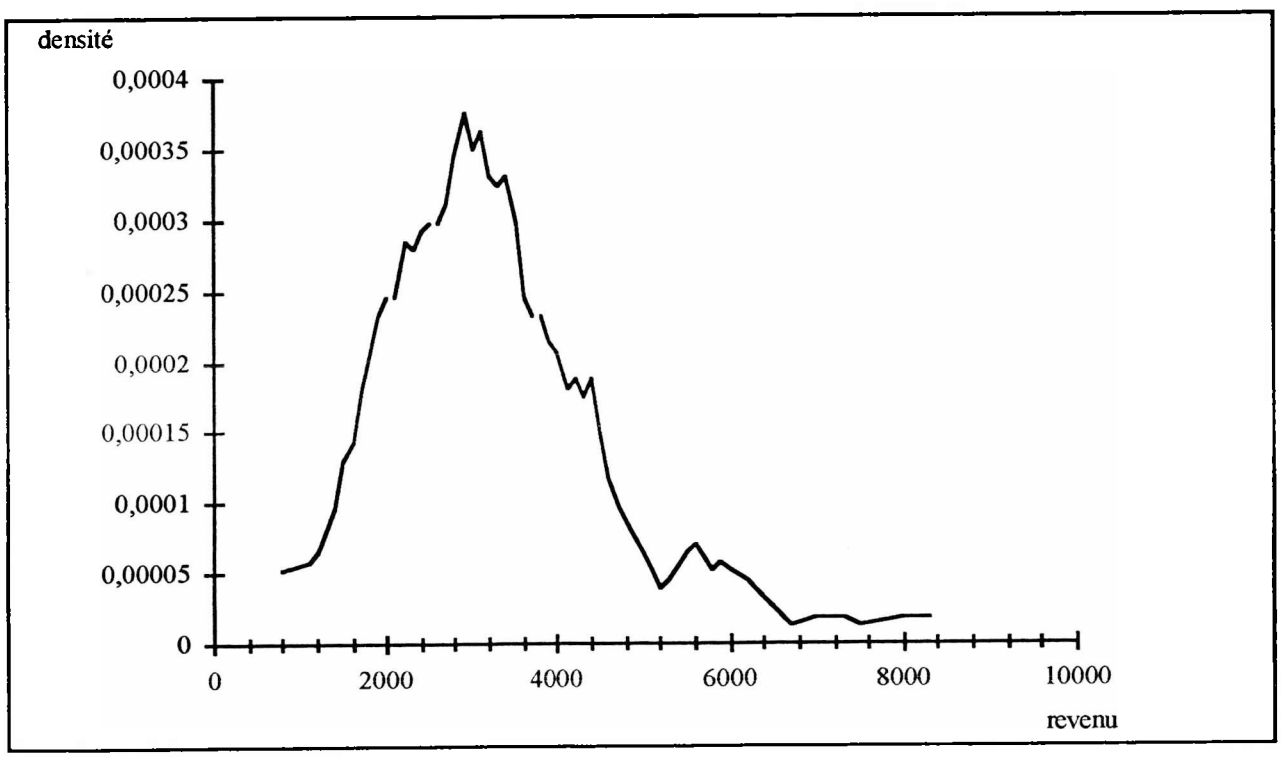
$$\hat{f}(y) = \frac{\text{nombre d' observations dans } [y - h/2; y + h/2[}{h \times \text{nombre total d' observations}}$$

h doit être choisi de façon à ce que le nombre d'observations par intervalle soit assez important pour être significatif. Il ne peut donc être considéré comme tendant vers 0.

La représentation graphique est dans ce cas la courbe $(y, \hat{f}(y))$.

Graphique n°10

Représentation de la densité du revenu par u.c. h=800



Source: enquête Pauvreté CREDOC-CES 1995

Dans cette estimation, le calcul de $\hat{f}(y)$ se fait en sommant les observations situées dans l'intervalle centré sur y de largeur h et en ne tenant pas compte des autres observations. Autrement dit, il s'agit de sommer les observations de façon pondérée, en accordant un poids de 1 aux observations proches de y et de 0 pour les autres. Là encore, on introduit des discontinuités arbitraires, puisque le poids accordé aux observations passe brusquement de 1 à 0 lorsque l'on sort de l'intervalle.

Cette estimation fait partie d'une méthode d'estimation dite du noyau. Elle consiste à accorder un poids K(u) à chacune des observations, u étant fonction de la distance des observations à y. Mathématiquement, cela s'écrit :

$$\hat{f}(y) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{y - y_i}{h}\right),$$

où n est le nombre total d'observations, h la largeur de l'intervalle centré sur y.

La fonction K utilisée pour le graphique précédent est définie par

$$\begin{cases} K(u)=1 \text{ pour } -0,5 \leq u < 0,5 \\ K(u)=0 \text{ pour } u < -0,5 \text{ ou } u \geq 0,5 \end{cases}$$

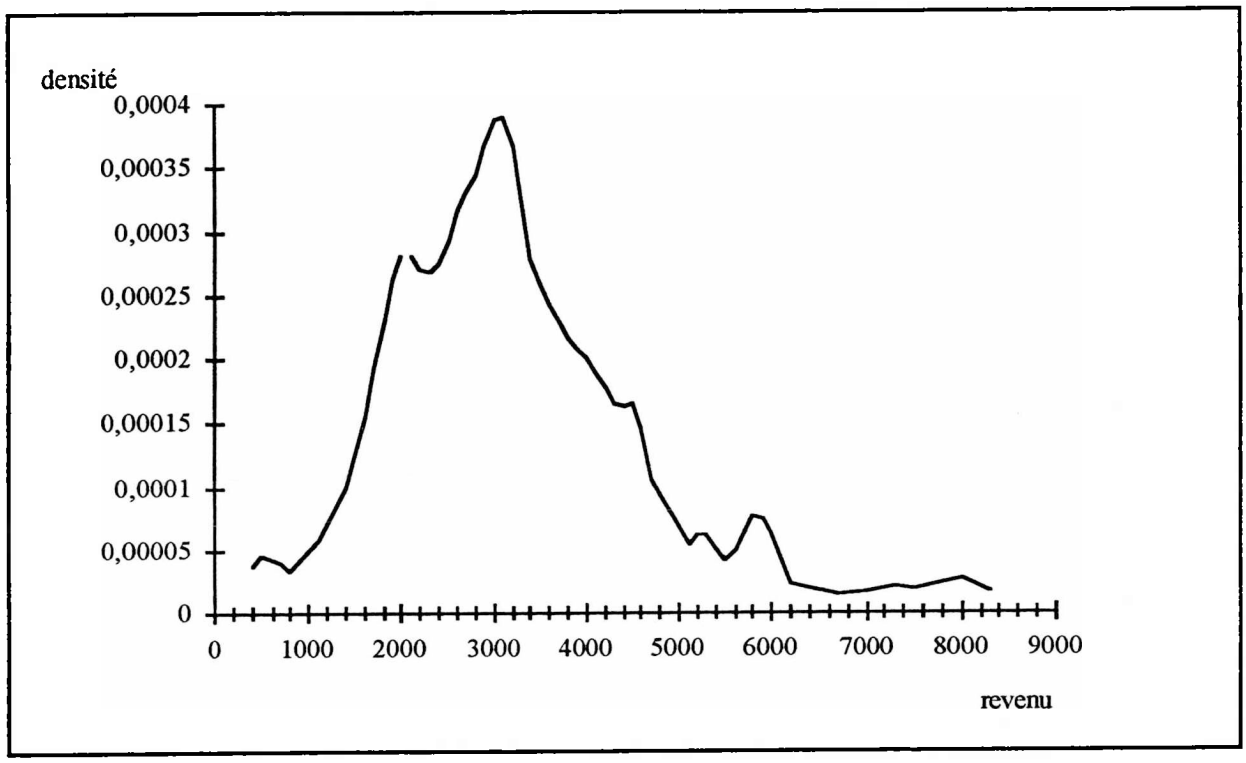
Une estimation meilleure car moins discontinue sera obtenue en prenant pour K une fonction continue accordant un poids décroissant aux observations lorsqu'elles s'éloignent de y. Le graphique suivant a été tracé en prenant :

$$\begin{cases} K(u)=1+\cos 2\pi u & \text{pour } -0,5 \leq u < 0,5 \\ K(u)=0 & \text{pour } u < -0,5 \text{ ou } u \geq 0,5 \end{cases}$$

On obtient ainsi une courbe proche de la précédente mais dont l'allure est moins heurtée. On peut dans ce cas parler d'une estimation lissée de la densité.

Graphique n°11

Estimation lissée de la densité du revenu par u.c. h=800



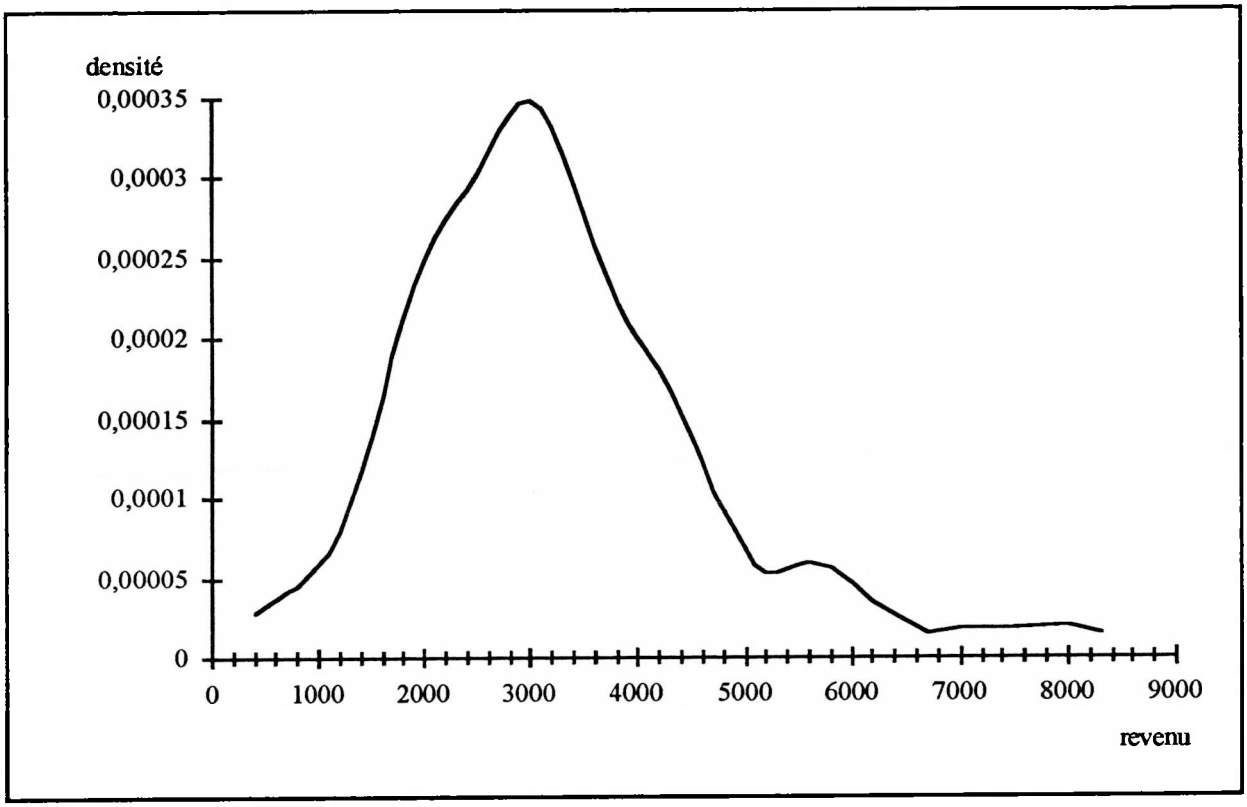
Source : enquête Pauvreté CREDOC-CES 1995

Le calcul de l'estimation dépend non seulement de la « fonction noyau » choisie, mais également de la largeur des intervalles (c'est-à-dire de h). Comme pour la construction de l'histogramme, l'arbitrage se fait entre un graphique à forte signification mais peu précis et un graphique précis mais dont les discontinuités ne sont pas toujours significatives.

Le graphique suivant montre le résultat obtenu pour un calcul de densité effectué avec le même noyau que le graphique précédent, mais en prenant h=1600. Le pic autour de 2000 francs est estompé et n'apparaît presque plus, par contre, les fluctuations de densité entre 6000 et 8000 francs persistent.

Graphique n°12

Estimation lissée de la densité du revenu par u.c. h=1600



Source : enquête Pauvreté CREDOC-CES 1995

La fréquence d'un intervalle se déduit du graphique en intégrant la fonction de densité sur cet intervalle. Elle correspond donc à l'aire située sous la courbe de densité pour un intervalle donné. La courbe de densité représente les variations de fréquence, ou de concentration des individus autour des valeurs du revenu lorsque celui-ci augmente.

La courbe de densité donne des réponses qualitatives aux questions : « avec quelle intensité les valeurs de la variable sont-elles prises ? » ; « Existent-ils plusieurs valeurs de la variable pour lesquelles on observe des sauts brusques de concentration ? »

Elle apporte des informations sur :

-la position de la variable : elle permet notamment de donner le mode plus précisément et de manière plus fiable que l'histogramme. En effet, la densité étant calculée pour chacune des valeurs de y, le mode est beaucoup moins sensible au choix des intervalles qu'il ne l'était pour l'histogramme.

-la dispersion de la variable : elle montre l'étendue des valeurs de la variable et la rapidité avec laquelle le nombre d'observations diminue lorsque l'on s'éloigne du mode. Elle permet de voir si toutes les observations sont concentrées autour du mode, si elles s'en éloignent progressivement, s'il existe d'autres valeurs à concentration relativement élevée par rapport aux valeurs voisines.

-la forme de la variable : parce qu'elle est continue, la courbe de densité en donne une meilleure impression que les histogrammes. La symétrie est appréciée en comparant la forme de la courbe de part et d'autre du mode. La rapidité avec laquelle elle s'incline respectivement vers les valeurs faibles et élevées permet de savoir si la distribution est étalée vers la droite ou vers la gauche.

Le mode de la variable revenu se situe dans l'intervalle 3000-3200. La courbe présente une certaine symétrie entre 1100 francs et 5100 francs, c'est à dire pour les valeurs de revenu les plus souvent prises. Cependant, même pour ces valeurs, la courbe sur les estimations de densité est plus pentue vers les bas revenus que vers les revenus élevés. Un pic apparaît autour de 2100 francs.

L'interprétation de la courbe de densité est malaisée en termes quantitatifs. La densité est un indicateur de la concentration des variables autour d'un point, autrement dit, c'est la dérivée de la fréquence des intervalles centrés sur ce point. Elle n'a pas quantitativement de signification par elle-même. Il s'agit nécessairement d'une estimation, qui dépend notamment de la valeur arbitraire choisie pour la largeur de l'intervalle. Le fait que le tracé soit continu rend cependant très bien compte de la forme de la variable et donne un bon aperçu visuel de la répartition des individus. C'est donc un bon instrument d'analyse, mais qui est difficilement utilisable pour la présentation des résultats.

II-1-2-c Représentation de la fonction de distribution : « quantile plot » ou « Q-plot »

Le but du quantile plot est de représenter les données en fonction de leur rang autrement dit de représenter les observations ordonnées par leur valeur pour la variable étudiée.

RANG D'UNE OBSERVATION ET QUANTILE

rang : soit y, une variable numérique, y(i), ses valeurs ordonnées. Si un individu j prend la valeur y(k), alors le rang de l'individu est k.

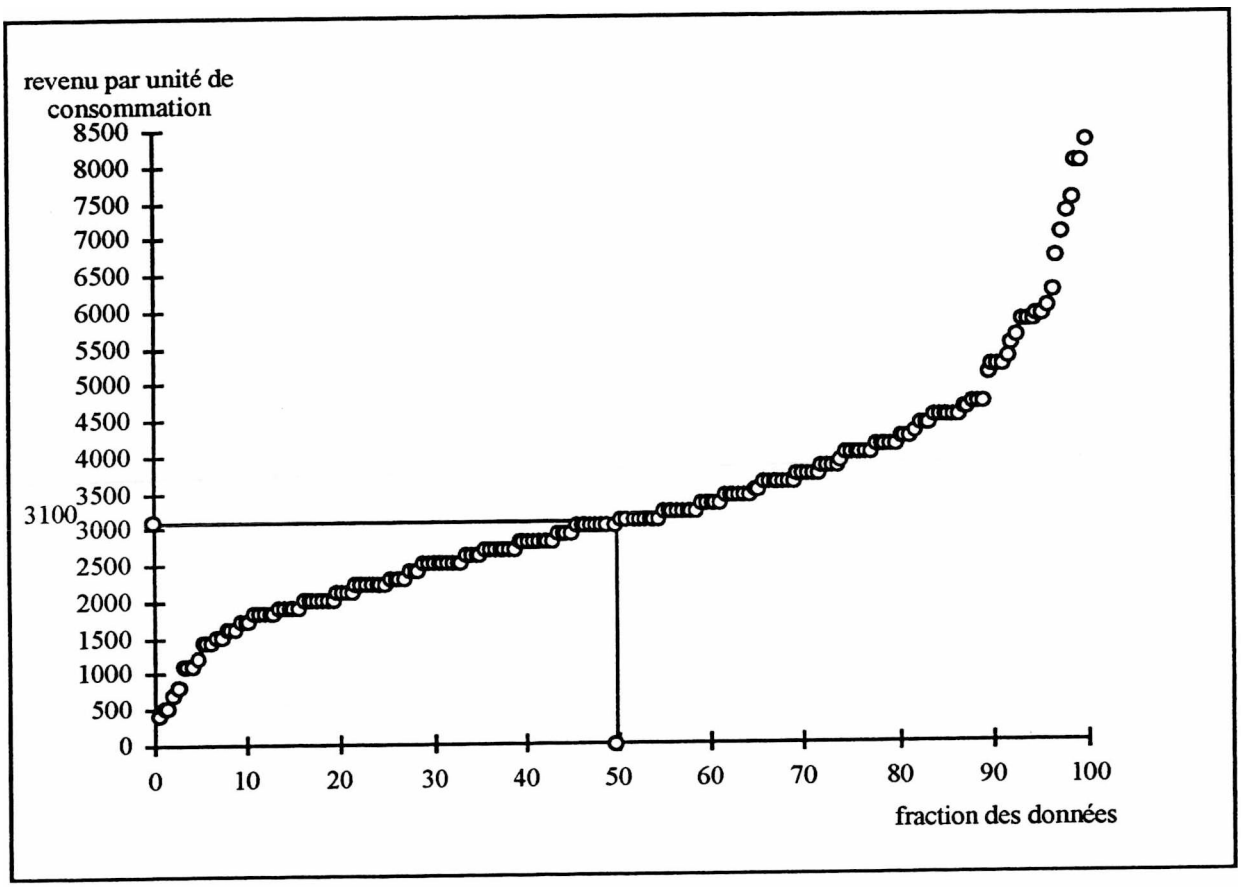
quantile : Soit p, un nombre compris entre 0 et 1. On appelle quantile d'ordre p le nombre Q(p) tel que pour une fraction p des observations, la variable ait une valeur inférieure ou égale à Q(p) et pour une fraction 1-p des observations, la variable ait une valeur supérieure ou égale à Q(p).
Si y est une variable numérique observée sur n individus, y(i) ses valeurs ordonnées, i compris entre 1 et n, si la proportion p des observations (pn) est de la forme i-0,5, alors Q(p)=y(i). Si ce n'est pas le cas, Q(p) est calculé par interpolation linéaire : si pn est à une distance f de i-0,5, soit pn=i-0,5+f alors Q(p)=(1-f)y(i)+fy(i+1).
Remarque : le quantile d'ordre 0,25 est le quartile inférieur, le quantile d'ordre 0,5 est la médiane, le quantile d'ordre 0,75 est le quartile supérieur.

Lors de la construction du Q-plot, on porte en fait en abscisse les quantiles des données qui correspondent aux valeurs effectivement prises par la variable : ils représentent le rang de la variable par rapport au nombre total d'observations[6]. L'axe des abscisses est ainsi indépendant du nombre d'observations, ce qui permet la comparaison des quantiles plots pour des nombres d'observations différents. A la lecture, on accède finalement à la fonction de répartition.

[6] Si on appelle x la variable x(i) ses valeurs ordonnées et que l'on suppose qu'elle a n observations, les quantiles représentés sont alors les quantiles d'ordre pi, i compris entre 1 et n, vérifiant : pi=i-0,5.

Graphique n°13

Q plot du revenu par u.c.



Source : enquête Pauvreté CREDOC-CES 1995

L'abscisse du point est le pourcentage des données ayant un revenu inférieur ou égal à son ordonnée. L'abscisse du graphique permet donc de lire la fonction de répartition.

Les Q plots apportent des informations sur la position de la variable, notamment parce qu'ils donnent rapidement accès aux quantiles et en particulier à la médiane et aux quartiles. Ils permettent de calculer et de mettre en évidence rapidement des indicateurs sur la dispersion de la variable, comme les intervalles interquantiles et en particulier interquartiles. Enfin, ils montrent le caractère symétrique ou asymétrique de la distribution. En effet, une distribution symétrique est telle que, pour tout p compris entre 0 et 0,5, en notant Q(p) le quantile d'ordre p, on a :

médiane-Q(p)=Q(1-p)-médiane

La courbe d'une distribution symétrique doit donc posséder le point correspondant à la médiane comme centre de symétrie.

L'impression visuelle donnée par les Q-Plots concerne surtout la forme de la variable et notamment son caractère symétrique ou non. Ces tracés permettent de retrouver facilement les quantiles, ce qui donne des indications chiffrées sur la position et la dispersion de la variable. Suivant le résultat recherché, on peut graphiquement mettre en évidence ces indications sur le tracé, comme cela a été fait pour la médiane sur le graphique précédent. Ils sont cependant peu utilisés pour la simple description de la variable. Ils permettent surtout la comparaison de deux variables, comme nous le verrons au cours du chapitre suivant.

II-2 Représentation des résumés numériques de la variable

La statistique exploratoire utilise un certain nombre de résumés numériques caractérisant la position (indicateurs de tendance centrale) et la dispersion (indicateurs de dispersion). Les graphiques précédents permettaient d'avoir un aperçu de certains de ces indicateurs (médiane, modes, quantiles, étendue, intervalles interquantiles, coefficients de forme). Les graphiques suivants s'intéressent spécifiquement à la représentation de ces résumés. Les premiers sont une visualisation des indicateurs de position et de dispersion. Les graphiques du deuxième paragraphe s'intéressent plus spécifiquement à la forme et notamment à la symétrie de la variable.

II-2-1 Les « box plots » ou « boîtes à pattes »

Les box plots sont construites après calcul de la médiane et des quartiles. Elles sont tout d'abord constituées d'une boîte rectangulaire, coupée en deux par un trait horizontal représentant la médiane, de longueur égale à l'intervalle interquartile et de largeur arbitraire. On cherche ensuite à représenter les observations s'éloignant le plus de l'ensemble des données. Le critère d'éloignement est défini à partir du pas : pas =1,5(Q3-Q1), où Q3 et Q1 représentent les quartiles respectivement supérieurs et inférieurs.

On définit alors les observations adjacentes supérieures et inférieures, les observations périphériques et les observations lointaines de la façon suivante :

-observation adjacente supérieure : la plus grande valeur inférieure à Q3+pas ;
-observation adjacente inférieure : la plus petite valeur supérieure à Q1-pas ;
-observations périphériques : ce sont les éventuelles observations comprises entre la valeur adjacente supérieure (respectivement inférieure) et Q3+2pas (respectivement Q1-2pas), appelée barrière supérieure (respectivement barrière inférieure) ;
- observations lointaines : éventuelles observations supérieures à Q3+2pas et inférieures à Q1-2pas.

Les valeurs périphériques et lointaines constituent les valeurs extérieures. Elles sont considérées comme étant des valeurs déviantes. Le corps de la boîte est alors muni de pattes venant se prolonger jusqu'aux valeurs adjacentes. Les valeurs extérieures sont ensuite représentées par des points (dont le graphisme diffère selon que la valeur est périphérique ou lointaine).

Ainsi, pour le revenu par unité de consommation, on obtient les chiffres suivants :

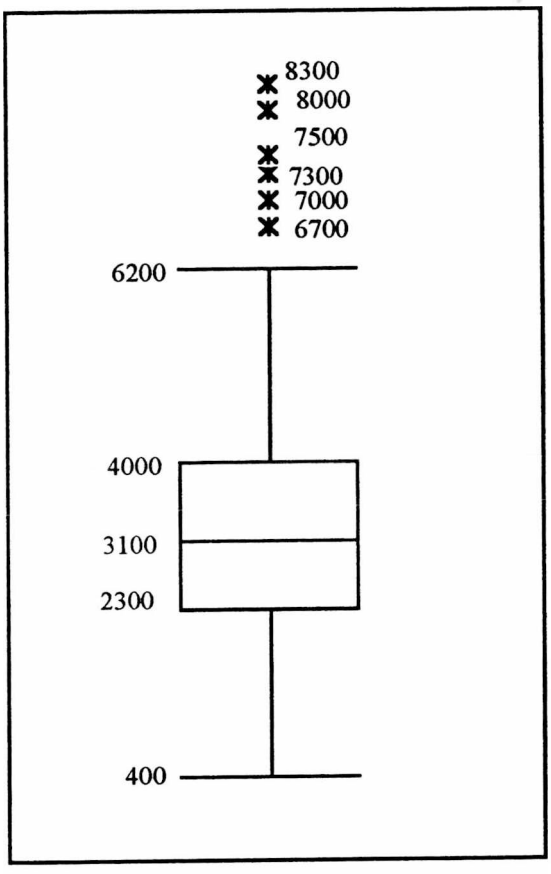
médiane=3200
Q1=2300
Q3=4000
pas=2550
valeur adjacente supérieure : 6200
valeur adjacente inférieure : 400
barrière supérieure : 9100

Il n'y a pas de valeurs lointaines, ni de valeurs périphériques inférieures.

Remarque : ces chiffres se déduisent rapidement du diagramme stem and leaf pour un nombre d'observations inférieur à 200.

Graphique n°14

Box plot du revenu par u.c.



Source : enquête Pauvreté CREDOC-CES 1995

La box plot fournit ainsi une représentation de la position de la variable en faisant apparaître sur un graphique la médiane, les quartiles et les valeurs adjacentes. Elle donne une bonne impression de la forme de la variable entre les valeurs adjacentes. On peut ainsi juger de la symétrie en étudiant la position de la médiane dans le corps de la boîte, la position des valeurs adjacentes. L'étendue de la boîte, la longueur des pattes, et la mise en évidence de valeurs déviantes (périphériques ou lointaines) apportent des renseignements sur la dispersion.

Cette représentation ne montre que des résumés des valeurs de la variable : elle indique comment se comportent les données en général, mais occulte les variations locales par rapport à ces tendances générales. Elle est très utilisée pour comparer plusieurs distributions ou pour étudier la réponse du comportement d'une variable continue aux variations d'une autre (discrète ou discrétisée), comme nous le verrons dans un chapitre suivant.

II-2-2 La symétrie

Lorsque l'on a résumé une variable par un indicateur de tendance centrale, l'étude de la symétrie permet de savoir si les individus réagissent de façon analogue lorsque l'on s'éloigne du centre de la distribution par valeurs supérieures ou par valeurs inférieures. La symétrie apporte des renseignements sur la façon dont cette position est déterminée : ainsi, on peut avoir une forte concentration des faibles valeurs dans un petit intervalle et une étendue peu concentrée des valeurs élevées dans un grand intervalle, ou au contraire, une répartition uniforme des données de part et d'autre de la caractéristique centrale de la distribution.

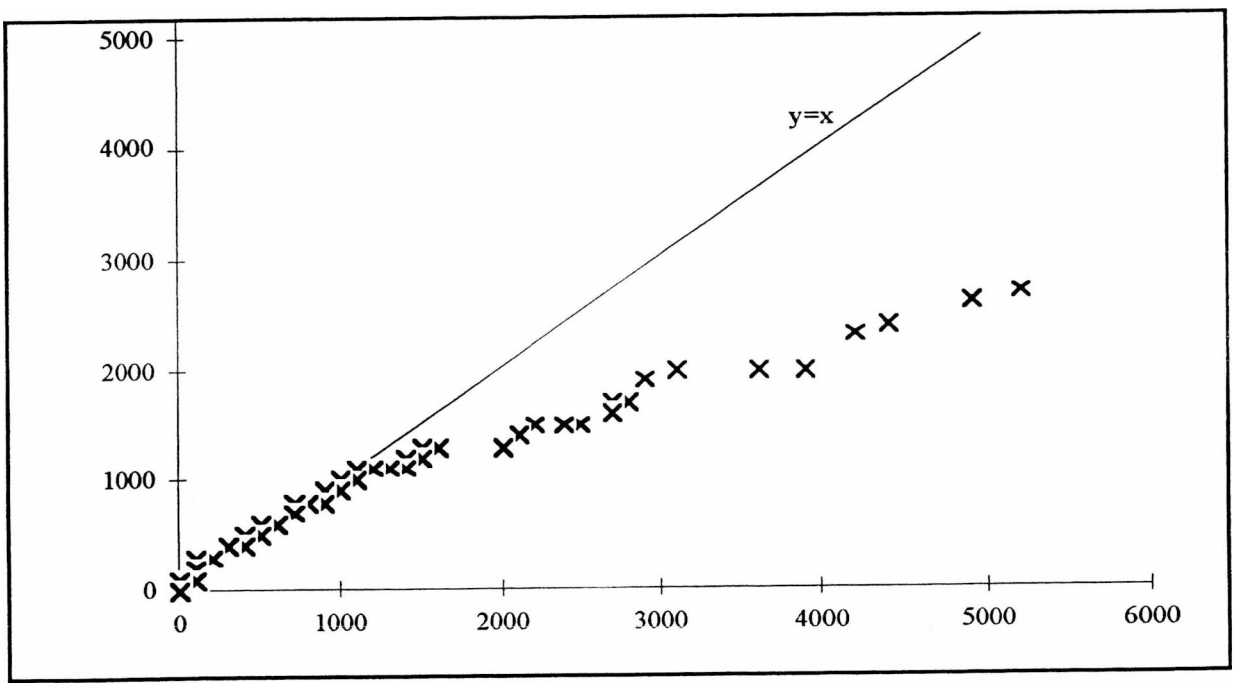
On considère une variable y sur n observations. On note y(i) les valeurs ordonnées de la variable. y(i) désigne donc la valeur de y pour l'observation de rang i. La distribution est symétrique si pour tout i entre 1 et la partie entière de (n+1)/2, on a :

médiane-y(i)=y(n+1-i)-médiane

On définit alors les variables u(i) et v(i) par : u(i)=y(n+1-i)-médiane et v(i)=médiane -y(i), et on trace la courbe (v(i), u(i)), que l'on compare à la droite y=x.

Graphique n°15

Représentation de la symétrie du revenu par u.c.



Source : enquête Pauvreté CREDOC-CES 1995

Pour une distribution symétrique, le tracé de symétrie doit coïncider avec la droite y=x. Les écarts du tracé avec cette droite permettent de voir pour quelles valeurs la distribution n'est pas symétrique et de quelle façon elle s'écarte de la symétrie.

Lorsque la courbe est au-dessous de la droite y=x, alors v(i)>u(i), autrement dit, pour deux observations équidistantes par rapport au rang de la médiane, la valeur supérieure est plus éloignée de la valeur de la médiane que la valeur inférieure. Il y a donc une concentration des observations sur un intervalle peu étendu pour les valeurs faibles tandis que la distribution s'étire vers les valeurs élevées.

Le tracé de symétrie donne un aperçu immédiat du caractère symétrique ou non de la variable. Il présente l'intérêt de fournir une réponse quantitative à la question « de combien deux observations de même profondeur[7] s'écartent-elles de la médiane ? »

III- EXEMPLE D'ANALYSE EXPLORATOIRE A PARTIR DES GRAPHIQUES MONOFACTORIELS

Les représentations graphiques décrites précédemment sont complémentaires. Chacune d'elles apporte des informations sur la position, la dispersion et la forme de la variable. Aucune ne peut suffire à décrire complètement la distribution. Analysées séparément, elles permettent de faire des hypothèses quant au comportement des individus par rapport à la variable. Leur comparaison permet de vérifier ces hypothèses. Il est en particulier souvent utile de comparer des graphiques présentant les valeurs de la variable à ceux construits à partir de résumés numériques. Enfin, l'utilisation de tous les graphiques n'est pas toujours nécessaire pour des variables dont les caractéristiques peuvent apparaître clairement.

Ce paragraphe présente une analyse des revenus des personnes en situation de pauvreté, à partir des différents graphiques qui ont été présentés tout au long de ce premier chapitre ainsi qu'une deuxième étude effectuée sur les salaires des personnes en situation de grande pauvreté.

---

[7] Profondeur : si la variable y est observée sur n individus et si une observation a pour rang i, alors sa profondeur est le plus petit nombre entre i et n-i (profondeur=Min(rang, nombre d'observations-rang)). La profondeur est donc le nombre d'individus qui sépare l'observation de la valeur extrême la plus proche.

III-1 Le revenu par u.c.

La position du revenu

La représentation unidimensionnelle et le diagramme stem and leaf montrent que les revenus mensuels des salariés en situation de grande pauvreté sont compris entre 400 et 9300 francs par mois. Le mode est dans l'intervalle 3000-3400 francs, la médiane est de 3100 francs.

Les estimations de densité fournissent une valeur du mode dans l'intervalle 3000-3200 francs.

La dispersion du revenu

L'étendue de la distribution est donc de 7900 francs. L'intervalle interquartile, calculé à partir du diagramme stem and leaf est de 1700 francs, ce qui est faible par rapport à l'étendue de la distribution. L'une au moins des valeurs extrêmes est donc éloignée de la plupart des autres observations. Si on prend de plus en compte la médiane de 3100 francs, on peut supposer que les revenus sont concentrés autour des faibles valeurs. Ceci est confirmé par l'étude de l'estimation lissée du revenu.

Le diagramme stem and leaf, les graphiques 10 et 11 montrent de plus que pour les revenus les plus élevés (supérieurs à 5000 francs), la concentration des individus est la plus forte aux alentours de 5800 francs. Elle s'étale ensuite faiblement jusqu'à 8300 francs. Pour les faibles valeurs de revenus, les plus basses (entre 400 et 1100 francs) sont peu représentées : la courbe de densité augmente doucement jusqu'à 1100 francs.

Le diagramme stem and leaf ne montre pas de valeurs aberrantes. La box plot met en évidence l'existence de valeurs extérieures, s'échelonnant régulièrement entre 6700 et 8300 francs. Il n'existe pas de telles observations pour les faibles revenus.

La forme du revenu

Le diagramme stem and leaf et les estimations de densité semblent indiquer que la courbe présente une certaine symétrie entre 1100 francs et 5100 francs, c'est à dire pour les valeurs du revenu le plus souvent prises. Cependant, même pour ces valeurs, la courbe sur les estimations de densité est plus pentue vers les bas revenus que vers les revenus élevés. Le graphique 11 met

d'ailleurs en évidence l'existence d'un pic autour de 2100 francs. En regardant les tiges du diagramme stem and leaf de part et d'autre du mode (c'est-à-dire de part et d'autre de la tige la plus longue, constituée des valeurs comprises entre 3000 et 3400 francs), on voit d'ailleurs que celles portant les valeurs inférieures au mode (entre 2000 et 2900 francs) sont plus longues que celles portant les valeurs supérieures au mode (entre 3500 et 4400 francs).

Le tracé de la symétrie confirme cette impression en donnant les valeurs exactes pour lesquelles la variable présente une certaine symétrie : le tracé se confond avec la droite y=x lorsque u et v sont compris entre 0 et 1400 francs, ce qui correspond à des valeurs du revenu comprises entre 1700 et 4500 francs. La courbe est ensuite d'autant plus éloignée de la première bissectrice que les valeurs du revenu sont loin de la médiane. Elle est de plus nettement en-dessous, ce qui traduit l'étirement vers les valeurs élevées du revenu. Cette impression persiste si on enlève les six points les plus à droite du graphique, qui font intervenir dans leur calcul les valeurs extérieures.

Conclusion : quel est le revenu des salariés en situation de grande pauvreté ?

Le revenu mensuel par u.c. des salariés en situation de grande pauvreté est compris entre 400 et 8300 francs. Il atteint donc des valeurs relativement élevées. Il faut relativiser cette remarque en notant que la tendance générale est à un revenu proche de 3000 francs.

La plupart des personnes interrogées ont un revenu faible, entre 400 et 5000 francs. Dans cet intervalle, on peut distinguer :

-les revenus compris entre 1700 et 3900 francs. Le revenu de la plupart des personnes ayant un revenu faible est dans cet intervalle. Les observations sont alors réparties de façon à peu près symétrique de part et d'autre du mode.

- les revenus compris entre 400 et 1700 francs : il n'y a que peu d'individus ayant un revenu très faible. Leur nombre augmente progressivement de 400 à 1100 francs, puis un peu plus rapidement entre 1100 et 1700 francs.

Les revenus les plus élevés, supérieurs à 5000 francs, se concentrent autour de 5700 francs. Six salariés ont un revenu que l'ont peut définir comme étant beaucoup plus élevé que les autres. Ce revenu est supérieur à 6200 francs. S'il y a peu de revenus très faibles, la distribution reste malgré tout concentrée sur les revenus peu élevés. Elle s'étire vers les revenus les plus hauts, ce qui ne concerne que peu d'individus.

Les résultats de cette analyse sont confirmés par le calcul des indicateurs de tendance centrale, de dispersion, de position : la moyenne est de 3250. Elle confirme donc la valeur calculée pour la médiane, ce qui est logique puisque qu'il n'y avait pas de points aberrants. L'écart-type est de 1400, qui est une valeur cohérente avec la concentration des observations entre 1200 et 4800 francs. Le coefficient d'asymétrie est de 1, ce qui correspond à une distribution relativement symétrique mais étalée vers la droite.

III-2 Le salaire par u.c.

La position des salaires

Graphique n°16

Diagramme stem and leaf du salaire par u.c.

		Nombre d'observations
8	03	2
7	5	1
7	00	2
6		
6	0	1
5	688	3
5	23	2
4	5	1
4	00001122	8
3	555677888	9
3	00113444	8
2	55555555555555566677788899999	28
2	00000000000011111111122222344444444	36
1	55555666666677777888888888888899	34
1	00000011111112223333344	25
0	55555556667777777888899999	27
0	234444	6
total		193

Les données doivent être multipliées par 100

Source : enquête Pauvreté CREDOC-CES 1995

Ce diagramme montre que les salaires sont situés entre 200 et 8300 francs par mois. Les observations sont concentrées dans l'intervalle 200-4200 francs. La médiane est de 2000 francs.

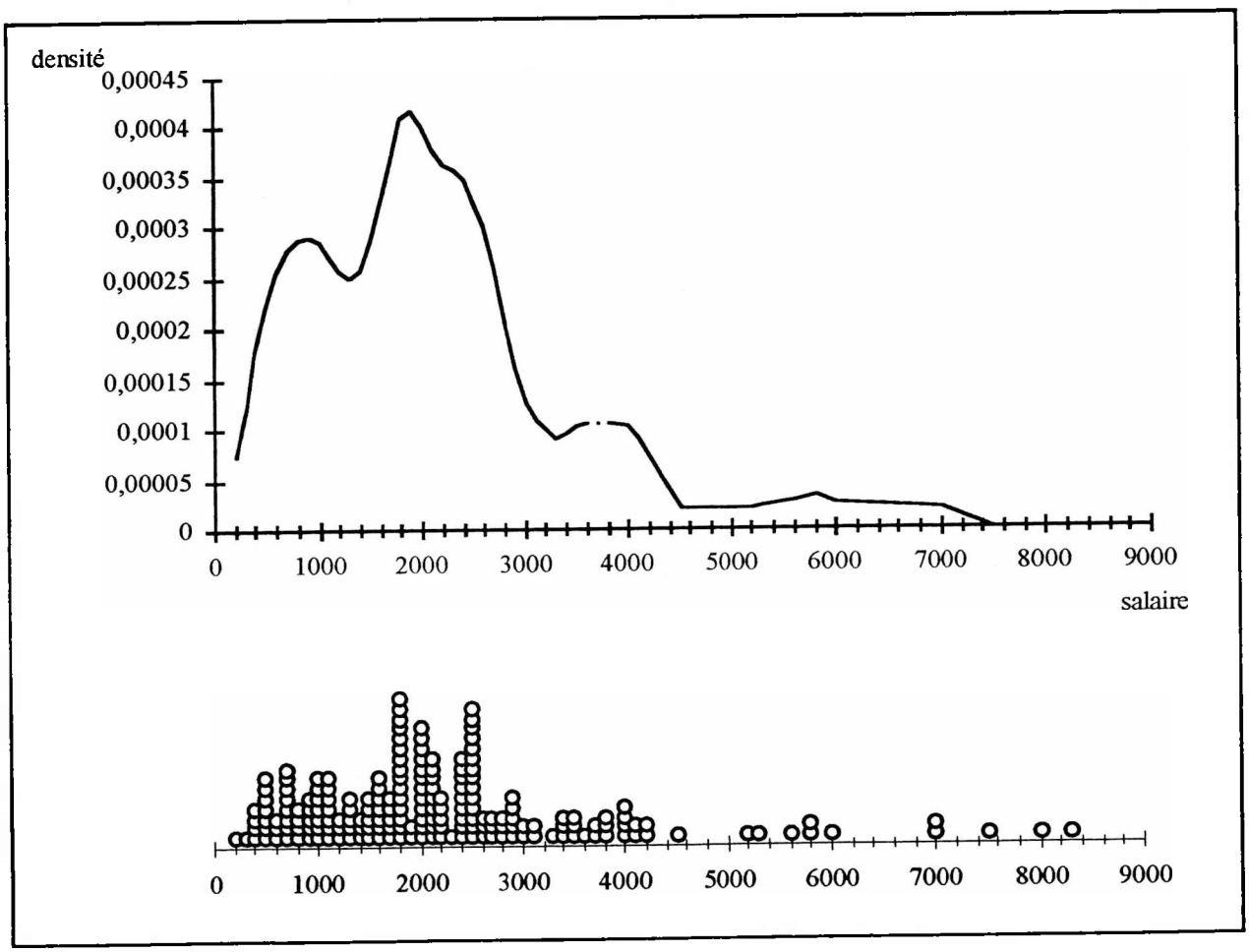
Le graphique suivant montre une estimation de la densité, tracée avec le noyau :

$$\begin{cases} K(u)=1+\cos 2\pi u & \text{pour } -0,5 \le u < 0,5 \\ K(u)=0 & \text{pour } u < -0,5 \text{ ou } u \ge 0,5 \end{cases}$$

et une largeur d'intervalles h=1000

Graphique N°17

Courbe de densité (h=1000) et représentation unidimensionnelle du salaire par u.c.



Source : enquête Pauvreté CREDOC-CES 1995

Le mode est dans l'intervalle 1800-2200 francs. Le graphique confirme que les salaires sont fortement concentrés entre 200 et 4200 francs.

La dispersion des salaires

L'étendue de la distribution est de 8100 francs, l'intervalle interquartile est de 1400 francs, ce qui est très faible par rapport à l'étendue. Ceci traduit le fait que la valeur supérieure est très éloignée de l'ensemble de la distribution, comme le montre la représentation unidimensionnelle.

Le nombre d'individus ayant un salaire inférieur au mode diminue brusquement lorsque l'on s'éloigne du mode en restant dans les valeurs supérieures à 1400 francs. le nombre d'individus augmente ensuite de nouveau et on observe un « pic » de la courbe de densité entre 500 et 1200 francs. Seuls quelques individus ont un salaire très faible, entre 200 et 500 francs.

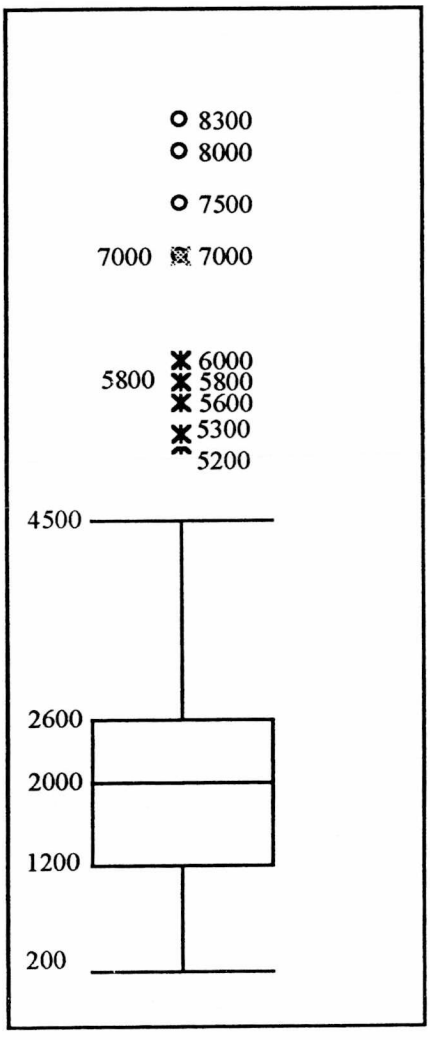
Le nombre d'individus ayant un salaire supérieur au mode diminue progressivement lorsque l'on reste proche du mode, jusqu'à 2500 francs puis chute brusquement jusqu'à 3200 francs.

Les revenus supérieurs à 3300 francs sont concentrés dans l'intervalle 3400-4500 francs. Quelques individus ont un salaire nettement supérieur à celui des autres. Ils sont alors dans l'intervalle 5200-7200 francs. Ceux ayant un salaire supérieur à 7200 francs sont très dispersés, puisque la densité est nulle pour ces valeurs. Le diagramme stem and leaf montre deux pôles de salaires supérieurs à 5000 francs : ceux compris entre 5200 et 6000 francs, puis ceux situés entre 7000 et 8300 francs.

L'étude de la box plot permet de mieux visualiser ces observations déviantes par rapport à l'ensemble de la distribution.

Graphique n° 18

Box plot du salaire par u.c.



Source : enquête Pauvreté CREDOC-CES 1995

Les valeurs périphériques sont représentées par des étoiles, les valeurs lointaines par des ronds. Lorsque deux points se superposent, leur valeur a été écrite deux fois. Six points extérieurs sont mis en évidence, entre 5200 et 6000 francs, et cinq points lointains, entre 7000 et 8300 francs.

La forme des salaires

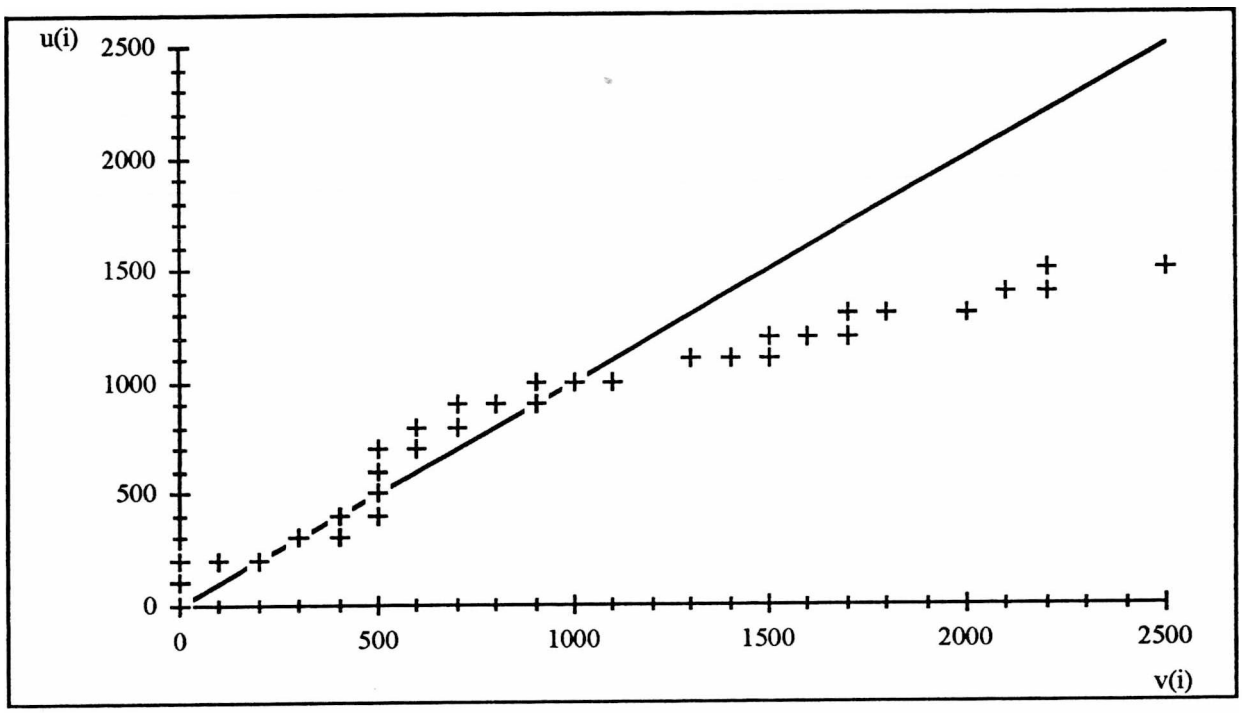
Le corps de la boîte montre que la médiane est plus proche du quartile supérieur que du quartile inférieur. Ceci indique que dans l'intervalle interquartile (entre 1200 et 2600 francs), les salaires supérieurs à la médiane restent en général plus proches de la médiane que les salaires

inférieurs : lorsque l'on s'éloigne de la médiane en restant dans l'intervalle interquartile, les salaires faibles chutent plus rapidement que les salaires élevés n'augmentent.

Le tracé de symétrie est alors effectué pour les observations restant dans le corps de la boîte. Il est donc construit pour des valeurs de u et v inférieures à 2500.

Graphique n° 19

Tracé de symétrie du salaire par u.c. pour les valeurs proches du mode



Source : enquête Pauvreté CREDOC-CES 1995

Ce tracé confirme les observations précédentes et permet de préciser dans quel intervalle la répartition des observations est asymétrique. Les points sont au-dessus de la première bissectrice pour u et v compris entre 500 et 900 francs, soit pour les salaires compris entre 1100 et 1500 francs d'une part et entre 2500 et 2900 francs d'autre part.

Lorsque l'on s'éloigne du mode, la courbe est par contre étalée vers la droite.

Conclusion : quel est le salaire des salariés en situation de grande pauvreté ?

Le salaire mensuel des salariés en situation de grande pauvreté s'étale en 400 et 8300 francs, avec une médiane de 2000 francs. Le salaire par u.c. des personnes interrogées est donc en général faible, même si certaines ont des salaires pouvant aller jusqu'à 8300 francs.

Les revenus faibles sont compris entre 500 et 3300 francs. Dans cet intervalle, les salaires élevés (supérieurs à la médiane) en restent généralement proches, alors que les salaires faibles s'en éloignent rapidement.

Les individus ayant un salaire très faible, entre 200 et 500 francs, sont peu nombreux. Les revenus élevés sont les revenus supérieurs à 3300 francs. Ils sont alors principalement dans l'intervalle 3400-4500 francs. Quelques unes des personnes interrogées ont un salaire nettement supérieur à celui des autres ; entre 5200 et 6000 francs d'une part et entre 7000 et 8300 francs d'autre part.

Les méthodes graphiques présentées ici ont permis de décrire les variable revenu et salaire par u.c. de manière exhaustive. Elles permettent de mettre en évidence visuellement les caractéristiques des variables étudiées.

Toutes ces méthodes sont fondées sur un comptage des données ordonnées et sur des résumés numériques. Ces résumés sont ceux qui sont accessibles directement par un tel comptage, c'est-à-dire la médiane, l'étendue, l'intervalle interquartile. Il y a donc une très grande cohérence entre les indicateurs et les graphiques utilisés.

Contrairement aux indicateurs fondés sur la notion de distance euclidienne (moyenne, écart-type), ces résumés n'ont pas de propriétés algébriques. Cela ne pose pas de problème pour une description monofactorielle. Ils sont tout aussi robustes, et même peut-être plus car moins sensibles aux valeurs extrêmes.

Les méthodes graphiques permettent d'étudier en détail l'ensemble de la distribution, y compris pour les valeurs qui s'éloignent des valeurs centrales. Elles requièrent par contre une grande attention dans leur construction et leur interprétation, notamment parce que les tracés présentés ne reposent pas toujours sur des choix objectifs et peuvent être très sensibles à des variations ne concernant qu'un faible nombre d'individus.

Etude des relations

entre

deux variables continues

Après avoir étudié les variables séparément, le deuxième temps de l'analyse exploratoire est consacré à l'étude des relations entre ces variables. Dans ce chapitre sont étudiées les relations entre deux variables quantitatives. La comparaison peut porter sur les variables dans leur ensemble, ou sur les individus.

* comparaison des distributions dans leur ensemble : on s'intéresse aux points communs et aux différences de la position, de la dispersion et de la forme de chacune des deux distributions. On mettra graphiquement en évidence les différences afin d'en avoir un aperçu qualitatif et quantitatif.

Ceci nous permettra de répondre aux questions suivantes :
-de combien diffèrent en général les deux distributions ?
-quelles sont les différences de concentration des individus d'une distribution à l'autre ?

* comparaison des distributions selon les individus : lorsque l'on effectuait une représentation monofactorielle, on étudiait les valeurs d'une variable sur les individus, et donc les points xi, valeurs de la variable x sur le ième individu. Ici, on étudie les valeurs de deux variables sur les individus, et donc les vecteurs (xi,yi) xi et yi étant respectivement les valeurs de la variable x et de la variable y pour le ième individu. Il s'agit donc de l'étude d'une variable à deux dimensions, de la forme (x,y). On s'intéresse alors à :

-l'évolution de la concentration des points en fonction à la fois des valeurs de x et de y. Il s'agit ici de décrire la distribution de la variable à deux dimensions.
- la liaison entre les deux variables. On cherche une relation yi=f(xi) pour tout individu i. Dans le cas d'une analyse numérique, on calcule en particulier des coefficients de corrélation.

On essaie de répondre aux questions suivantes :

-existe-t-il des intervalles de valeurs dans lesquels il est possible de mettre en évidence une telle relation ?
-quelle est alors l'influence du fait que l'individu i prenne la valeur xi pour la variable x sur la valeur yi qu'il prend pour la variable y ?

COEFFICIENTS DE CORRELATION

coefficient de corrélation linéaire : ce coefficient mesure le caractère plus ou moins linéaire du nuage de points formé par les variables x et y dans un repère carthésien.

$$r = \frac{1}{n}\frac{\sum_{i=1}^{n}(xi-\bar{x})(yi-\bar{y})}{s_x s_y}$$ où s_x et s_y sont les écarts types de x et de y.

On a $-1 \leq r \leq 1$

r=1 équivaut à l'existence d'une relation linéaire exacte ax+by+c=0

correlation des rangs : le coefficient de correlation est calculé sur les variables ordonnées.

Nous verrons comment les représentations graphiques permettent d'étudier chacune de ces deux relations. Dans les deux cas, nous essaierons alors de comprendre de quelle façon les salariés en situation de grande pauvreté complètent leur salaire par des revenus hors salaire.

Pour cela, on étudiera la relation globale entre le revenu, le salaire et le revenu hors salaire, puis la liaison entre le salaire et le revenu hors salaire. En effet, on a la relation suivante : revenus=salaires+revenus hors salaires. Le lien entre le revenu et le salaire et le revenu hors salaire est nécessairement élevé. Aucune liaison n'est par contre certaine entre le salaire et le revenu hors salaire.

Dans la suite du chapitre, nous appellerons x et y les deux variables, xi et yi les valeurs prises par ces variables pour un même individu i et x(i), y(i) les valeurs ordonnées des variables.

I- COMPARAISON DE DEUX DISTRIBUTIONS DANS LEUR ENSEMBLE

Pour comparer les deux distributions dans leur ensemble, on compare les valeurs ordonnées de chacune des deux variables et leurs résumés numériques. Dans les deux cas, les méthodes graphiques sont directement issues de celles utilisées pour la description monofactorielle. Elles sont fondées sur le fait que, pour deux variables prenant des valeurs

comparables, les tracés effectués sur chacune d'elles auront toujours au moins une dimension commune : les valeurs ordonnées prises par les variables.

I-1 Comparaison des valeurs ordonnées

I-1-1 Les « QQ plots » ou Q plots de comparaison

Le principe de ce graphique est de représenter les quantiles de la variable y en fonction des quantiles de même ordre de la variable x. Si x et y sont observées sur le même nombre n d'individus, on choisira n quantiles d'ordre pi de façon à ce que ces quantiles correspondent aux données ordonnées pour chacune des deux variables[8].

Si x et y ne sont pas observées sur le même nombre d'individus (à cause de la présence de valeurs manquantes notamment), on choisit un nombre de quantiles égal au nombre de valeurs prises par la variable ayant le moins d'observations. On effectue une extrapolation linéaire pour calculer les quantiles de même ordre de la deuxième variable[9].

Le graphique suivant représente le Q-Q plot des quantiles du revenu par u.c. en fonction des quantiles du salaire par u.c. Le nombre d'observations étant le même pour les deux variables, nous sommes dans le cas ou le tracé représente les valeurs ordonnées du revenu en fonction des valeurs ordonnées du salaire.

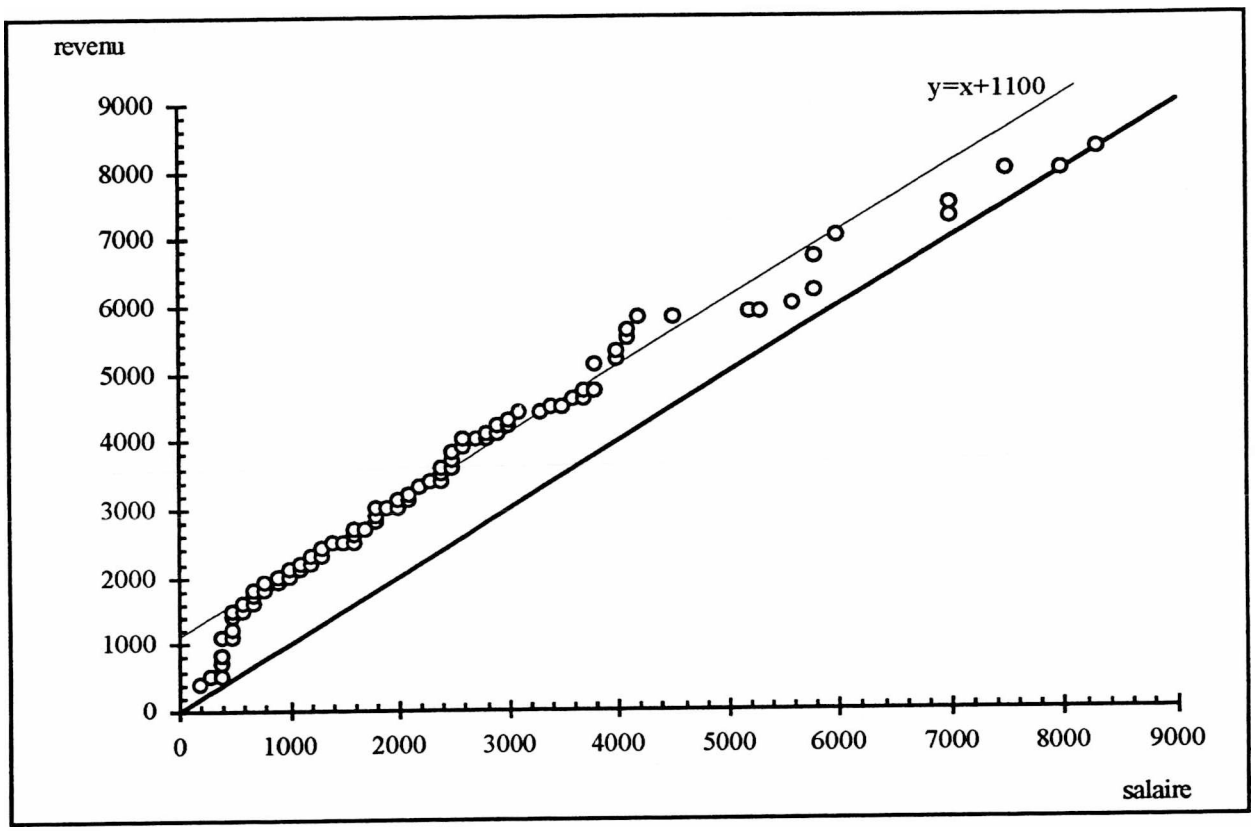
---

[8] En notant, pour tout i compris entre 1 et n, x(i) et y(i) les valeurs ordonnées de chacune des deux variables, $Q_x(pi)$ et $Q_y(pi)$ les quantiles d'ordre pi pour la variable x et la variable y respectivement, on a alors : $Q_x(pi)=x(i)$ et $Q_y(pi)=y(i)$ pour tout i, avec pi=(i-0,5)/n.

[9] On utilise les mêmes notations que pour la note précédente, mais on suppose cette fois que x à n observations et y, m observations, m<n. On choisit m quantiles d'ordre pi, i compris entre 1 et m, tels que pi=(i-0,5)/m. On alors Qy(pi)=y(i) pour tout i et Qx(pi)=(1-fi)Q(xi)+fiQ(x(i+1), fi=(i-0,5)(n-m)/m=pi(n-m)

Graphique n°20

QQ plot du revenu par u.c. en fonction du salaire par u.c.



Source : enquête Pauvreté CREDOC-CES 1995

Si les deux distributions prennent les mêmes valeurs, alors leurs quantiles de même ordre seront les mêmes et le tracé obtenu sera la droite y=x. La façon dont la courbe s'éloigne de cette droite donne une indication sur la relation entre les valeurs ordonnées des deux variables.

Ainsi, s'il est possible d'approcher la courbe par une droite d'équation y=ax+b, les valeurs des deux variables x et y seront telles que globalement, y=ax+b.

Si le nombre d'observations est le même pour les deux variables, cette relation signifie que y(i)=ax(i)+b, où y(i) et x(i) représentent la ième valeur ordonnée respectivement de y et de x et non la valeur observée pour l'individu i.

Si le nombre d'observations n'est pas le même pour les deux variables, l'équation y=ax+b se traduit par $Q_y(p)=aQ_x(p)+b$, , pour tout p compris entre 0 et 1, où $Q_x(p)$ et $Q_y(p)$ représentent les quantiles d'ordre p de x et y respectivement.

La relation obtenue est une relation globale entre les valeurs ordonnées des deux variables, et non une relation entre les valeurs prises sur chaque individu par les deux variables. Si on connaît la valeur xi de la variable x pour l'observation i, on ne peut rien déduire du graphique quant à la valeur de y sur le même individu.

Le graphique précédent a été complété par le tracé de la première bissectrice et par celui d'une droite passant « au mieux » parmi les points. Cette droite a pour équation y=x+1100. Elle représente bien le tracé pour des valeurs du salaire comprises entre 400 et 3700 francs, soit pour des valeurs du revenu comprises entre 1100 et 4800 francs. On peut donc en déduire que dans cet intervalle de valeurs, les revenus sont globalement plus élevés que les salaires de 1100 francs.

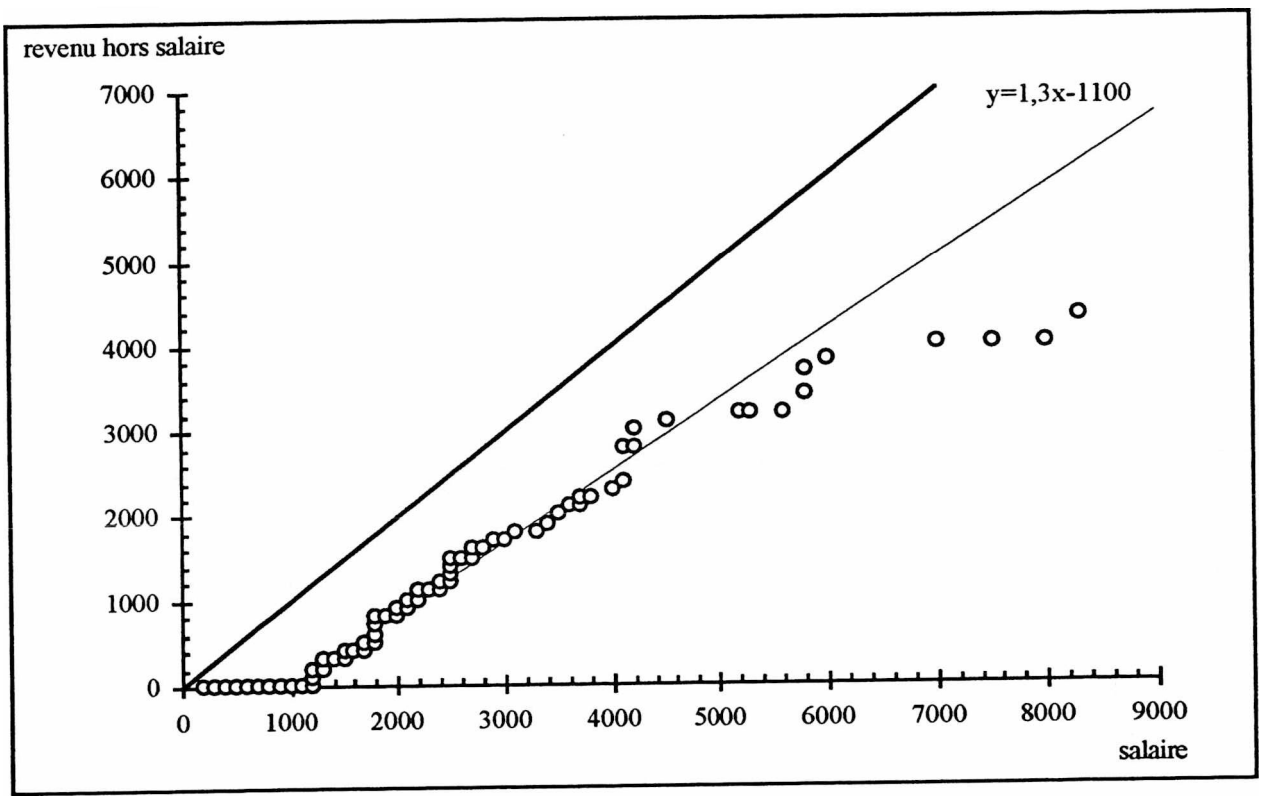
Le tracé ne permet pas de mettre en évidence une relation similaire sur d'autres intervalles de valeurs. En dehors du domaine de validité de la relation y=ax+b, il apprend que :

-les deux valeurs les plus élevées du revenu hors salaire et du salaire sont égales.

-les valeurs les plus basses du salaire sont plus faibles que les valeurs les plus basses du revenu.

Graphique n°21

QQ plot du revenu hors salaire par u.c. en fonction du revenu par u.c.



Source : enquête Pauvreté CREDOC-CES 1995

Le revenu hors salaire est calculé comme étant la différence du revenu et du salaire pour chaque individu. Il est donc normal que le « QQ-plot » du revenu hors salaire en fonction du salaire ne puisse pas se déduire du « QQ plot » du revenu en fonction du salaire. Par exemple, le premier graphique montre que les quantiles les plus élevés du salaire sont égaux aux quantiles les plus élevés du revenu. On ne retrouve pas cette information à la lecture du deuxième graphique, puisque les valeurs 0 du salaire sont placées, par construction même, en regard des quantiles les plus faibles du salaire.

Le graphique précédent permet de dire que pour un salaire compris entre 1100 et 3800 francs, les revenus hors salaire sont plus faibles globalement que les salaires selon la relation :

y=1,3x-1100.

Les tracés « QQ plot » permettent de répondre à la question « selon quelle relation diffère globalement la variable y de la variable x? ». Ils peuvent être utilisés pour des distributions n'ayant pas le même nombre d'observations. Ils ont de plus l'avantage de présenter la distribution y en fonction de la distribution x et donnent ainsi une bonne impression

visuelle de la relation globale entre les variables. Ces tracés se heurtent au problème de chevauchement des données : les valeurs de même rang pour les deux variables ne seront représentées que par un seul point.

I-1-2 Comparaison des diagrammes « stem and leaf »

Les graphiques stem and leaf de deux distributions prenant des valeurs comparables peuvent avoir le même tronc, à condition de choisir le même nombre de chiffres par feuilles et le même nombre de tiges. Un tel choix permet de construire côte à côte les diagrammes stem and leaf des deux variables.

Graphique n°22

Comparaison des diagrammes stem and leaf pour le salaire et pour le revenu par u.c.

	salaires	revenus
8	03	003
7	5	5
7	0	03
6		7
6	0	02
5	688	5688899
5	23	12223
4	5	55555566777
4	00001122	000000111112223444
3	555677888	5566666667777788889
3	00113444	000000000111111111222222223333344444
2	55555555555555566677788899999	555555555666677777777888888889999
2	000000000000111111111222223444444444	00000001111222222223333444
1	555556666666777778888888888888899	556667778888899999
1	00000001111111222333344	1112444
0	55555556667777777888899999	5578
0	234444	4

les données doivent être multipliées par 100

Source : enquête Pauvreté CREDOC-CES 1995

Nous avions vu qu'un tel diagramme donnait des informations sur la position (mode, médiane), la dispersion (valeurs extrêmes, valeurs déviantes ou aberrantes) et la forme des variables (symétrie). Mettre en regard les graphiques tracés pour les deux distributions permet

donc de comparer ces informations, dans la mesure où les chiffres du tronc sont bien les mêmes pour les deux variables. Comme précédemment, les données étant triées séparément sur chacune des variables, la comparaison est globale.

Dans le cas du revenu et du salaire par u.c., on voit ainsi que :

-le mode du revenu est de 1000 francs plus élevé que celui du salaire. D'une manière générale, les valeurs du salaire sont concentrées entre 200 et 4500 francs, alors que celles du revenu se trouvent surtout entre 1000 et 6000 francs ;

-les valeurs maximales du revenu et du salaire sont les mêmes, la valeur minimale du salaire est de 200 francs plus faible que celle du revenu ;

-les valeurs supérieures du salaire ont un caractère déviant plus prononcé que les valeurs extrêmes du revenu ;

-la distribution des revenus est plus symétrique que celle des salaires. Elle est notamment moins nettement étirée vers les valeurs supérieures.

La comparaison des médianes ou des quantiles n'est pas visuellement directement accessible. On ne se servira pas d'un tel graphique pour la faire.

Un des avantages du diagramme stem and leaf était d'enregistrer les valeurs exactes des données. Il est exploité dans la comparaison des valeurs extrêmes et déviantes et dans celle éventuelle de la médiane ou des quantiles. De plus, cette forme d'enregistrement assure que toutes les valeurs prises par les variables apparaissent sur le graphique. Ce n'était pas le cas pour le tracé de comparaison.

Le diagramme stem and leaf a l'inconvénient de présenter de façon discontinue des variables continues. Lors de la comparaison des variables, les chiffres du tronc étant les mêmes pour chacune d'elles, les discontinuités sont situées entre les mêmes valeurs pour les deux distributions. Cela peut fausser la comparaison dans la mesure où une variable peut présenter effectivement une certaine discontinuité pour ces valeurs et l'autre non. Dans un cas, la représentation montrera une discontinuité qui existe réellement, dans l'autre cas, elle montrera également une discontinuité qui en fait n'existe pas.

Par exemple, les salaires ont une réelle discontinuité entre les deux tiges « 4 » puisque leurs valeurs augmentent brusquement de 4200 à 4500 francs au même endroit. Les revenus par contre sont continus pour ces mêmes valeurs. Pour les deux variables l'impression visuelle est similaire.

Les revenus hors salaire auraient pu être tracés sur le même graphique. Le diagramme stem and leaf pour cette variable est cependant très peu lisible à cause du grand nombre de valeurs nulles prises par les revenus hors salaire. Il y aurait eu ainsi plus de 50 enregistrements correspondant au premier 0.

Le fait que les données soient enregistrées exactement permet de corriger ces discontinuités. Les histogrammes sont inadaptés pour des comparaisons. En effet, les discontinuités sont alors totalement arbitraires et ne peuvent être corrigées à la seule lecture du graphique.

Le diagramme stem and leaf permet de comparer qualitativement la position des deux variables et notamment l'intervalle contenant le mode, la dispersion des deux variables, et notamment les intervalles de concentration importante, la forme de la variable et en particulier la symétrie. Il apporte des informations quantitatives sur les valeurs extrêmes et les valeurs déviantes ou aberrantes. Pour que la comparaison soit valable, il faut que les deux variables soient observées sur le même nombre d'individus.

Il ne permet pas de donner une relation quantitative précise comme c'était le cas pour le Q-Plot de comparaison. Les discontinuités arbitraires inhérentes à sa construction peuvent de plus perturber la comparaison.

I-1-3 Comparaison des courbes de densité

Une courbe de densité a en abscisse les valeurs ordonnées d'une variable. L'axe des abscisses des courbes de densité de deux variables ayant des valeurs comparables est donc le même.

L'ordonnée de ces courbes est une fonction de densité estimée donnée par le nombre éventuellement pondéré, d'observations comprises dans un intervalle de largeur h et divisé par le nombre d'observations et la largeur h. L'axe des ordonnées de deux tracés de densité différents, pour des variables ayant ou non le même nombre d'observations, sont donc également les mêmes.

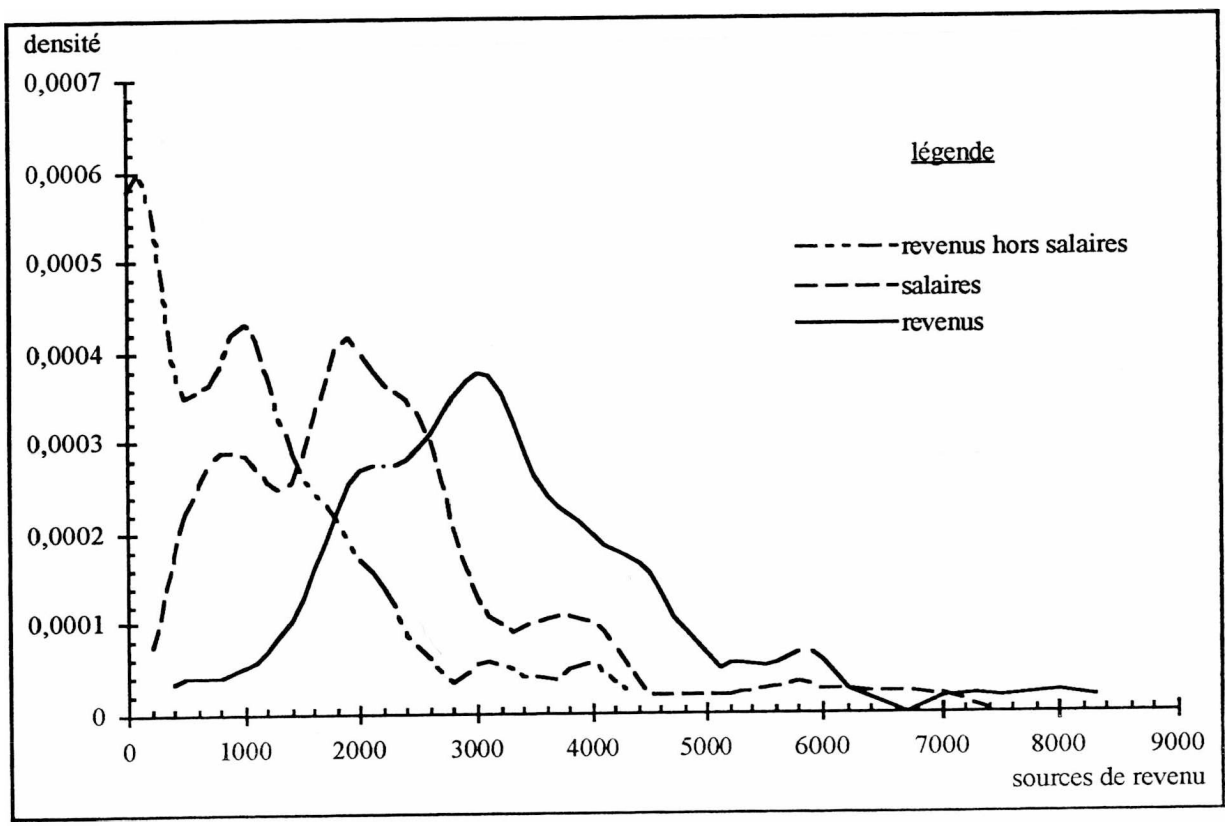
On peut ainsi comparer au moins deux tracés de densité de variables ayant des valeurs comparables en superposant les courbes obtenues lors de l'analyse monofactorielle.

Pour que la comparaison soit valable, il faut que la méthode d'estimation utilisée ainsi que la largeur d'intervalles h soient les mêmes. On limite ainsi les divergences dans les formes de tracés qui seraient dues aux méthodes utilisées et non aux disparités des distributions.

Le graphique suivant présente les tracés obtenus pour le revenu, les salaires et le revenu hors salaire. Dans les trois cas, on utilise une fonction de pondération $K(u)=1+\cos 2\pi u$ et une largeur d'intervalles h=1000.

Graphique n°23

Superposition des tracés de densité du revenu, du salaire et du revenu hors salaire par u.c. h=1000



Source : enquête Pauvreté CREDOC-CES 1995

La superposition des courbes de densité nous permet de retrouver éventuellement une relation globale entre les variables, analogue à celle mise en évidence sur le Q-Plot de comparaison. Une relation du type y=x+b se traduit sur le tracé de densité par l'existence d'une translation entre les courbes de x et de y. Une relation du type y=ax+b n'est pas par contre lisible.

Le graphique montre également les différences éventuelles dans :

-les positions des variables : on retrouve en effet le mode de chacune d'elles, mais aussi les valeurs extrêmes ;
-la dispersion des variables : le tracé permet de comparer l'étendue des deux variables. On s'intéressera de plus à la comparaison d'éventuels « pics » des courbes de densité, en dehors du mode. On peut également étudier les positions relatives et les points d'intersection des courbes ;
-la forme des variables : on compare la façon dont les variables s'étirent de part et d'autre du mode.

Le graphique précédent montre ainsi que :

-globalement, les revenus hors salaire sont plus faibles que les salaires (qui sont plus faibles que les revenus, ce qui est évident par définition même) : en effet, la courbe des revenus hors salaire est au-dessus de la courbe des salaires pour les valeurs faibles et en-dessous pour les valeurs élevées : il y a donc un pourcentage plus grand d'individus ayant un faible salaire que d'individus ayant un faible revenu.

-les valeurs du salaire sont globalement inférieures de 1100 francs aux valeurs du revenu lorsque le salaire est compris entre 600 et 3100 francs (ce qui confirme ce que l'on avait remarqué sur le Q-plot de comparaison). Le mode du salaire est notamment inférieur de 1100 francs au mode du revenu.

-on ne retrouve pas la relation entre le salaire et le revenu hors salaire observée sur le graphique 21. En effet, cette relation, de la forme $y=ax+b$, faisait intervenir une constante multiplicative sur la variable x. On ne peut pas la lire sur les courbes de densité, qui ne montrent que les translations d'une variable à l'autre (de la forme $y=x+b$).

-le revenu hors salaire est d'une manière générale plus faible que le salaire. Son mode est de 0, mais si on ne tient pas compte des valeurs nulles, on voit apparaître un « second mode » qui est inférieur de 1000 francs à celui des salaires. Les valeurs extrêmes supérieures des deux variables sont séparées par 4000 francs.

-la concentration des observations autour du mode est comparable pour le revenu et pour le salaire. L'étendue de la distribution l'est également. Dans les trois cas, les valeurs de la variable les plus élevées (supérieures à 3200 francs pour le salaire et pour le revenu hors salaire, à 5000 francs pour le revenu) présentent un autre « pôle » de concentration, quoique très faible, autour de 4000 francs pour le salaire et le revenu hors salaire et de 6000 francs pour les revenus. La concentration du revenu hors salaire diffère par contre pour les autres valeurs :

cette variable est très nettement concentrée sur les bas revenus. Son étendue est d'ailleurs beaucoup plus faible.

-salaires et revenus ont une forme comparable. Le graphique peut confirmer le fait que le revenu est un peu plus symétrique que le salaire, qui s'étire en particulier de façon plus marquée vers les valeurs élevées.

On ne peut pas parler de symétrie pour les revenus hors salaire car le mode est nul. Après le deuxième pic à 1200 francs, ils s'étirent vers les valeurs élevées plus que le salaire.

Le choix de la méthode d'estimation et de la largeur d'intervalle n'est pas neutre pour la comparaison, bien qu'il soit le même pour toutes les variables. En effet, nous avons vu que selon les valeurs de h et la fonction « noyau » choisie, des discontinuités pouvaient apparaître. Lorsque h augmente, les pics ou les creux de la courbe s'accentuent : des différences entre les courbes tracées sur deux variables différentes peuvent ainsi se former et s'accentuer. Lorsque h devient petit, ces différences peuvent provenir d'un faible nombre de variables et ne pas être significatives.

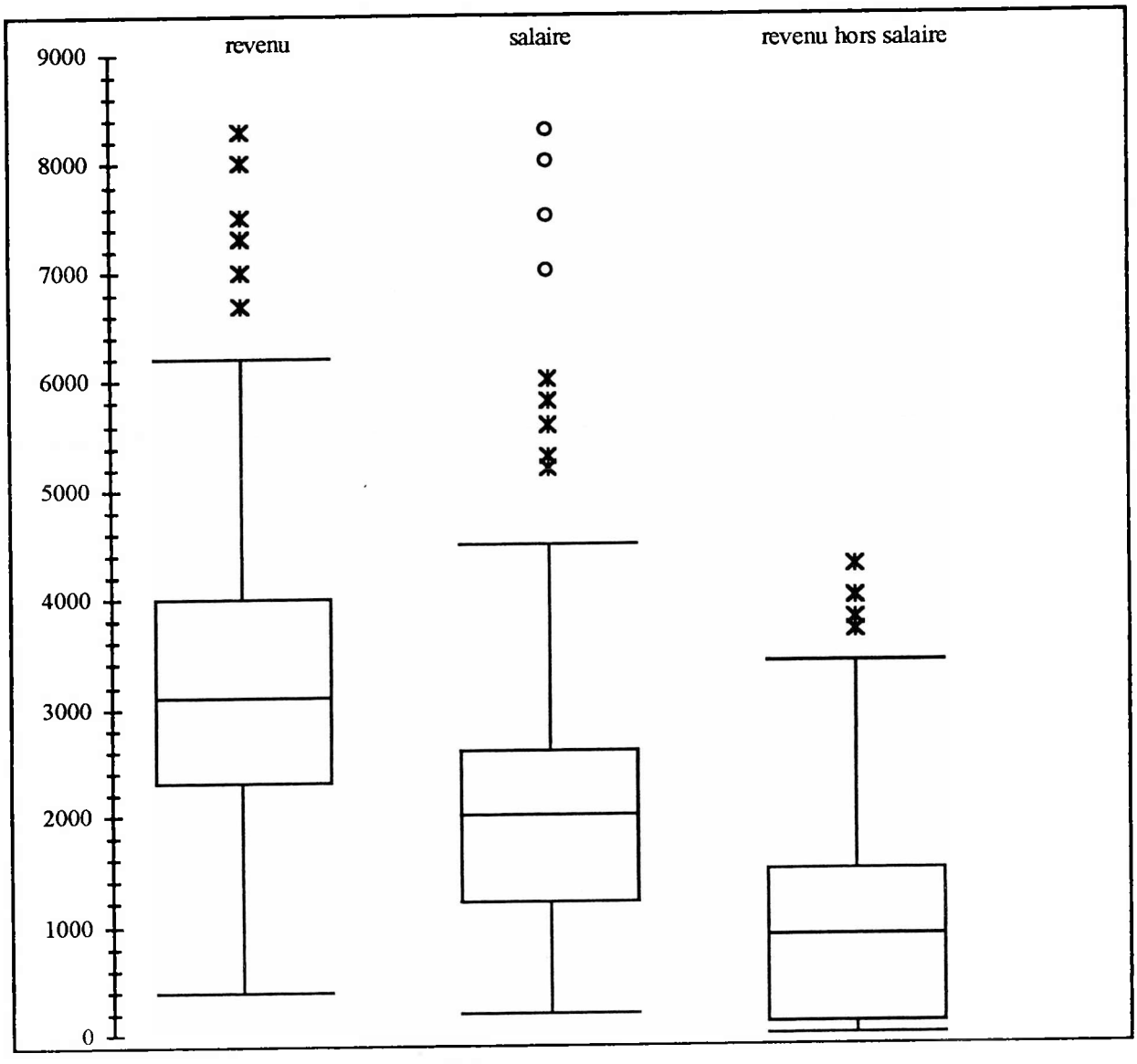
Les tracés de densité permettent de comparer la position, la dispersion et la forme des variables. Pour des tracés de densité relativement réguliers, ils montrent quelles sont les variables ayant globalement les valeurs les plus faibles. Ils peuvent être construits pour des variables n'ayant pas le même nombre d'observations.

I-2 Comparaison des résumés numériques de chacune des deux variables

La comparaison des résumés numériques se fait par l'intermédiaire des box plots. La dimension verticale est, comme pour les diagramme stem and leaf, les valeurs ordonnées de la variable. Pour des variables prenant des valeurs comparables, on peut donc placer les box-plots côte à côte.

Graphique n°24

Box plot du revenu, du salaire et du revenu hors salaire par u.c.



Source : enquête Pauvreté CREDOC-CES 1995

Le tracé permet la comparaison de :

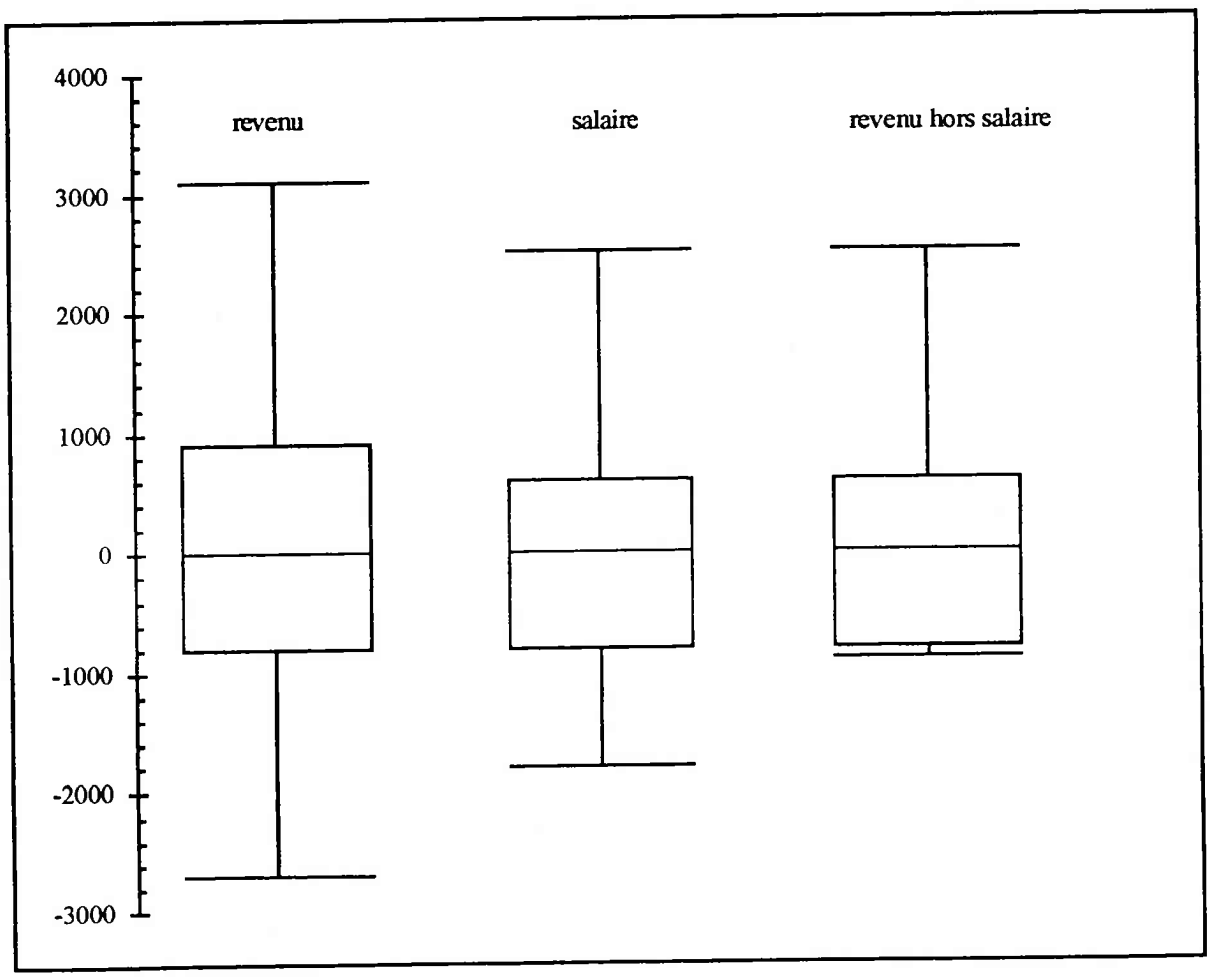
- la position des variables : on compare la position de la médiane, des quartiles, des valeurs extrêmes ;
-la dispersion des variables : on compare la taille du corps de la boîte, la taille des pattes, la présence de valeurs périphériques ou lointaines ;
-la forme des variables : on étudie la position de la médiane par rapport aux quartiles et aux pattes.

Au cours de l'étude comparative de la dispersion et de la forme, on s'intéresse uniquement aux différences de formes et de tailles des différents éléments du tracé, et non au niveau auquel est situé ce tracé.

Pratiquement, après avoir comparé la position des variables et les valeurs périphériques et lointaines, on refait un tracé en positionnant toutes les variables au même niveau. Cela permet de donner une meilleure perception des différences de forme et de dispersion. Le graphique suivant montre le résultat d'un tel tracé pour le revenu, le salaire et le revenu hors salaire. On a transformé les données de chacune des variables par soustraction de la médiane. On construit ainsi de nouvelles variables ayant l'origine pour centre de leur distribution : on dit qu'elles sont centrées.

Graphique n°25

Box plot du revenu, du salaire et du revenu hors salaire par u.c.
Variables centrées



Source : enquête Pauvreté CREDOC-CES 1995

Dans l'ensemble, les revenus sont nettement plus élevés que les salaires qui sont plus élevés que les revenus hors salaire. La médiane du revenu est de 1100 francs plus élevée que la médiane des salaires, elle-même plus élevée de 1100 francs que la médiane des revenus hors salaire. Le quantile supérieur du revenu est de 1400 francs plus élevé que celui du salaire, qui est de 1000 francs supérieur à celui du revenu hors salaire. Enfin, le quantile inférieur du revenu est de 1100 francs supérieur à celui du salaire, lui-même de 1100 francs plus élevé que celui du revenu hors salaire.

La valeur maximale du revenu est égale à celle du salaire. Il faut relativiser cette information en remarquant que si les valeurs les plus élevées du revenu sont périphériques, celles du salaire sont lointaines : les salaires les plus élevés sont plus éloignés du reste de la distribution que les revenus les plus élevés.

La valeur maximale du revenu et du salaire est beaucoup plus élevée que la valeur maximale du revenu hors salaire. Les valeurs périphériques du revenu hors salaire restent dans la patte du salaire, tout en s'approchant des valeurs adjacentes. Il en est de même pour les valeurs périphériques du salaire par rapport au revenu.

Le graphique précédent montre que la partie supérieure à la médiane des box plots du salaire et du revenu hors salaire est la même. Les valeurs supérieures à la médiane de chacune des distributions sont globalement réparties de la même façon jusqu'aux valeurs adjacentes. En fait, le tracé des courbes de densité montre que la concentration des individus chute plus brusquement pour les revenus hors salaire jusqu'à une valeur comprise entre le quartile et la frontière supérieure puis est un peu plus étirée.

L'intervalle interquartile est également à peu près le même, mais la patte inférieure du revenu hors salaire est beaucoup plus courte, ce qui traduit la forte asymétrie du revenu hors salaire et sa concentration sur les faibles valeurs de revenu.

La comparaison du revenu et du salaire confirme les résultats déjà obtenus quant à la dispersion et à la forme de ces deux variables : la taille totale de la boîte est plus grande pour les revenus que pour les salaires : les revenus sont plus dispersés. Ils sont également un peu plus symétriques : la médiane est plus centrée entre les deux quartiles et les pattes sont de même longueur, ce qui n'est pas le cas pour les salaires. Les salaires, plus que les revenus, sont concentrés sur les basses valeurs. Les revenus s'étalent plus vers les valeurs élevées, impression renforcée ici quant on ne tient pas compte des valeurs lointaines et périphériques.

Le tracé des box plots permet de comparer rapidement les deux variables. Il introduit la comparaison des valeurs lointaines et périphériques. On peut ainsi mettre en perspective la façon dont s'éloignent les valeurs extrêmes de chacune des variables du reste de leur distribution.

La comparaison globale de la position des variables se fait ici par l'intermédiaire de la médiane et non du mode, comme c'était le cas pour les autres tracés. Cela apporte des informations supplémentaires lorsque l'une des variables est fortement asymétrique et présente des valeurs du mode et de la médiane très différentes.

La comparaison permise par le tracé des box plots n'est fondée que sur les résumés numériques. Elle est donc moins précise que celle obtenue par les graphiques tracés à partir de l'ensemble des données. Ces tracés permettent en revanche de mettre en perspective un grand nombre de variables.

Les graphiques présentés ici permettent de montrer comment deux ou plusieurs distributions dont les valeurs sont comparables diffèrent dans leur position, leur dispersion et leur forme. Exceptés les diagrammes stem and leaf, ils sont utilisables pour des variables n'ayant pas le même nombre d'observations.

Comme c'était le cas pour l'analyse monofactorielle, ces graphiques sont complémentaires : chacun peut apporter des informations nouvelles, détailler ou valider des hypothèses faites à la lecture des autres représentations. Ainsi :

-le Q-plot de comparaison permet de définir une éventuelle relation globale entre les deux variables ;

- les diagrammes stem and leaf montrent les différences éventuelles entre les valeurs extrêmes et la présence de points aberrants. Ils permettent une première comparaison approximative des valeurs du mode et de la dispersion des données ;

-la superposition des tracés de densité précise les observations faites sur les diagrammes stem and leaf et les Q- plots de comparaison. Ils donnent une valeur plus exacte de la différence des modes. Ils permettent de retrouver une éventuelle relation du type y=x+b, ainsi que l'étude des différences entre les deux variables en dehors du domaine de validité de cette relation. Les disparités observées doivent alors être vérifiées sur les autres

graphiques, la construction des courbes de densité pouvant faire apparaître des différences qui ne sont en fait pas significatives ;

-la juxtaposition des box plots montre les différences dans les valeurs de la médiane et des quartiles. La comparaison de la taille des boîtes et des pattes donne un bon aperçu des disparités dans la dispersion des variables. La comparaison des valeurs périphériques et lointaines permet éventuellement de relativiser ou de renforcer les observations faites sur la comparaison des valeurs extrêmes.

Ces graphiques nous ont ainsi permis d'étudier les disparités et les similitudes des distributions du revenu, du salaire et du revenu hors salaire des salariés en situation de grande pauvreté.

Les revenus sont supérieurs de 1100 francs aux salaires quand on ne regarde que les valeurs du salaire comprises entre 400 et 3700 francs (Q-Plot de comparaison). Le mode, la médiane et les quartiles inférieurs des deux distributions diffèrent notamment de 1100 francs (Box plot et tracés de densité). Les quartiles supérieurs sont plus espacés, ce qui provient du fait que les revenus s'étalent plus que les salaires vers les valeurs élevées (tracés de densité et box plot). Les salaires sont ainsi plus concentrés sur les valeurs basses. L'étendue des deux distributions est à peu près la même, mais les valeurs les plus élevées du salaire sont plus éloignées de l'ensemble de la distribution que celles du revenus (Box plot, diagramme stem and leaf).

Les salaires sont globalement plus élevés de 1100 francs que les revenus hors salaire, si on s'en tient aux valeurs des revenus hors salaire inférieures à 2700 francs. La médiane et les quartiles supérieurs diffèrent ainsi de 1100 francs. Cette différence s'accentue lorsque les valeurs du revenu augmentent. Les revenus hors salaire sont beaucoup plus concentrés que les salaires sur les basses valeurs. L'étendue est également beaucoup plus faible.

II- COMPARAISON DES VARIABLES POUR UN MÊME INDIVIDU

Il s'agit cette fois de comparer les valeurs de chacune des variables individu par individu et non globalement. Les graphiques utilisent ici des points ayant pour coordonnées les valeurs de chacune des deux variables prises pour le même individu : les points représentent donc les individus, et non les valeurs ordonnées comme précédemment.

Les représentations ont pour but de montrer s'il existe ou non une relation entre les valeurs prises par chacune des variables. Pour cela, il existe deux types de tracés :

-ceux qui représentent les individus eux-mêmes ou la concentration de ces individus en fonction des valeurs des deux variables.
-ceux qui représentent, par l'utilisation de résumés numériques, l'évolution des valeurs d'une des variables en fonction des valeurs de l'autre variable.

II-1 Représentation des individus en fonction des valeurs des deux variables

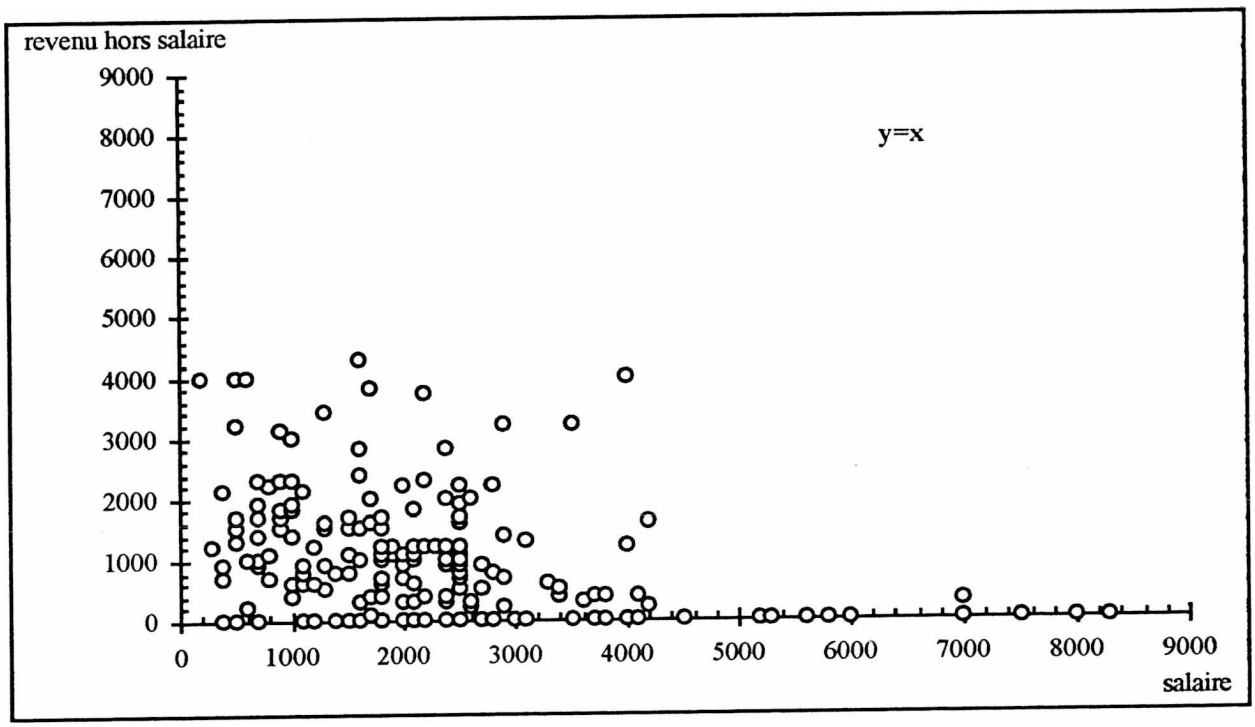
Les graphiques présentés dans ce paragraphe sont construits sur le même principe que ceux permettant de décrire une seule variable. On s'intéresse à la représentation des valeurs (xi, yi), à la fréquence de carrés correspondant à des intervalles de valeurs de x croisés par des intervalles de valeurs de y, à la densité de la variable à deux dimensions (x,y).

II-1-1 Représentation des individus en fonction des valeurs des deux variables

Le premier graphique représente directement les points (xi,yi).

Graphique n°26

Le revenu hors salaire par u.c. en fonction du salaire par u.c.



Source : enquête Pauvreté CREDOC-CES 1995

Le graphique permet de voir s'il existe une tendance à une évolution parallèle entre les deux variables et donc si effectivement, il peut être possible de trouver des domaines à l'intérieur desquels les deux variables sont fonctionnellement liées. Pour cela, on regarde la tendance générale dans la répartition des points et la façon dont la concentration des points évolue avec les valeurs des variables.

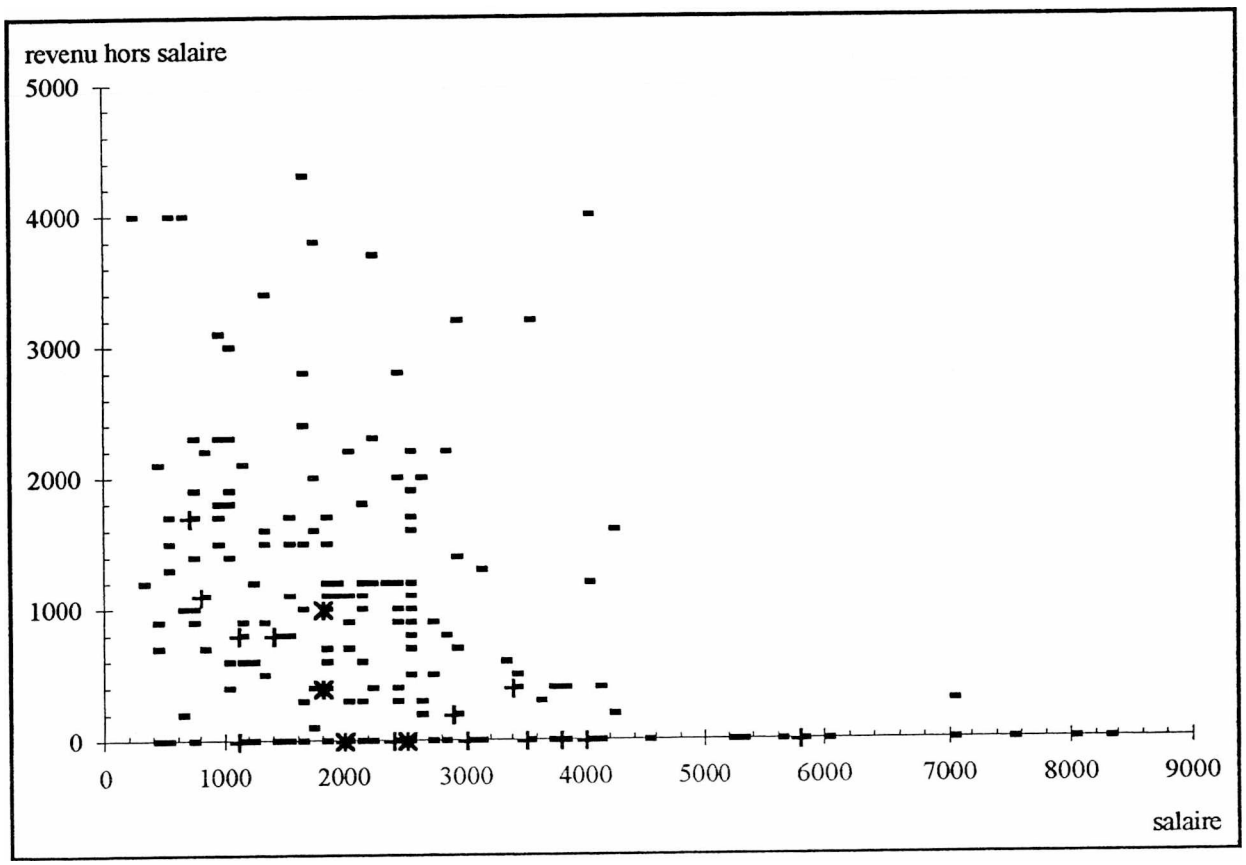
Les points de notre exemple se concentrent ici principalement dans l'intervalle correspondant aux revenus hors salaire compris entre 200 et 2000 francs et aux salaires situés entre 1500 et 3000 francs. D'une façon plus générale, il semble que la plupart des points se trouvent en dessous de la droite y=4000-x. Le revenu hors salaire a ainsi tendance à être d'autant plus faible que le salaire est élevé. Il est nul si le salaire est supérieur à 4500 francs (à une exception près). Il existe également un certain nombre de points ayant à la fois un salaire et un revenu hors salaire faibles.

Comme c'était le cas pour la représentation unidimensionnelle d'une variable, ce graphique se heurte au problème de superposition des points. Toutes les observations ne sont pas représentées, celles prenant la même valeur sur chacune des deux variables n'étant tracées

qu'une fois. On ne peut pas, comme on l'avait fait pour la représentation unidimensionnelle, décaler les points qui se superposent, les deux dimensions du plan ayant déja un sens. Une solution proposée par CLEVELAND ET MCGILL est d'utiliser des symboles différents pour les points en fonction du nombre d'observations qu'ils représentent. On utilise par exemple des étoiles, dont le nombre de branches indique le nombre d'observations : un tiret symbolise une observation, une croix deux...

Graphique n°27

Le revenu hors salaire en fonction du salaire par u.c.
Le nombre d'observations par point est symbolisé par des étoiles



Source : enquête Pauvreté CREDOC-CES 1995

Ces graphiques représentent tous les points de la distribution et permettent une première description de la variable à deux dimensions. Ils sont difficiles à exploiter pour une analyse plus approfondie. Ils donnent certes un premier aperçu de la tendance fonctionnelle éventuelle entre les deux variables, mais une telle relation est difficilement observable de façon précise. En effet, même dans le cas où il existe une relation entre les deux variables, elle n'est pas vérifiée par

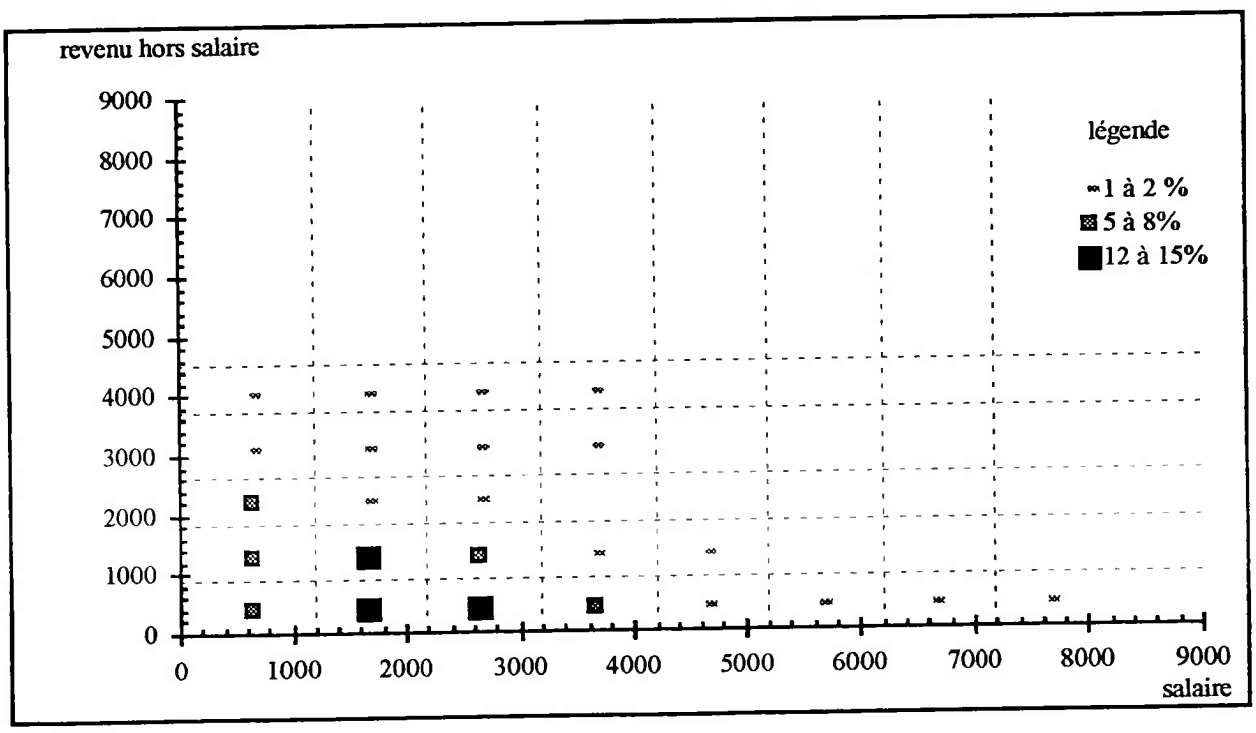
tous les points. Ils se répartissent alors de part et d'autre de la tendance générale. Cette dispersion peut empêcher l'observateur de voir la relation entre les deux variables.

II-1-2 Représentation de la fréquence d'une variable à deux dimensions

Comme pour une variable à une dimension, calculer la fréquence suppose de discrétiser chacune des variables. On calcule alors la fréquence des rectangles obtenus en croisant les intervalles de valeurs des deux variables. Il s'agit de reproduire sur un même graphique la fréquence et les intervalles de valeurs de deux variables. Pour cela, on représente la fréquence par des symboles graphiques. Après avoir défini des classes de fréquence, on convient d'un symbole par classe. On représente ce symbole au milieu des rectangles construits à partir des valeurs des deux variables : on a donc partagé le graphique précédent par des cellules et représenté la fréquence des cellules par un symbole. Un tel procédé s'appelle la cellulation.

Graphique n°28

Cellulation du revenu hors salaire et du salaire par u.c.



Source : enquête Pauvreté CREDOC-CES 1995

Les pointillés représentent les rectangles. Le fait d'utiliser des symboles pour la fréquence ne permet pas de prendre en compte la taille des rectangles. Celle-ci doit être égale pour tous

pour éviter les erreurs d'interprétation. Comme pour les histogrammes, le calcul de la fréquence des rectangles tient compte des bornes inférieures et non des bornes supérieures. On étudie la répartition des symboles représentant la fréquence.

Le graphique précédent permet de confirmer les remarques effectuées plus haut : les individus sont en général en-dessous de la droite y=4000-x. Les revenus hors salaire ont donc tendance à diminuer lorsque les salaires augmentent.

En-dessous de cette droite, on peut distinguer :

-les salaires les plus faibles (inférieurs à 1200 francs) pour lesquels les revenus hors salaire se répartissent uniformément jusqu'à 2700 francs ;

-les salaires moyens (entre 1200 et 3200 francs), pour lesquels les revenus hors salaire sont situés entre 0 et 1800 francs. Ils sont un peu moins nombreux entre 900 et 1800 francs qu'entre 0 et 900 francs ;

-les salaires élevés, supérieurs à 3200 francs, pour lesquels les revenus hors salaire sont inférieurs à 900 francs.

Les revenus hors salaire les plus élevés (supérieurs à 2700 francs), correspondent aux salaires inférieurs à 4200 francs.

Comme pour les histogrammes tracés sur une seule variable, ce graphique a l'inconvénient de représenter de façon discontinue des variables continues. Les discontinuités introduites sont sans fondement objectif. Ainsi dans l'interprétation, on a utilisé les bornes des intervalles arbitrairement choisis pour définir les catégories de salaire et de revenu hors salaire. Ces bornes doivent être considérées plus comme des indications qualitatives que réellement quantitatives.

Finalement la cellulation est un instrument pour répondre qualitativement à la question « comment évolue la concentration des individus par catégorie de valeurs d'une variable en fonction des catégories de valeurs de l'autre variable ? »

II-1-3 La densité d'une variable à deux dimensions

De la même façon que l'on avait défini la densité d'une variable x en un point xi comme étant le nombre d'observations situées dans un intervalle de largeur infinitésimale centré sur ce point, on définit la densité d'un vecteur de variables (x,y) en un point (xi, yi) comme étant le nombre d'observations comprises dans un cercle de rayon infinitésimal centré sur ce point.

Le principe utilisé pour l'estimation de la densité au point (xi, yi) est le même que celui utilisé pour l'estimation de la densité au point xi. Comme on ne peut pas, pour des variables observées empiriquement sur un nombre limité d'observations, compter le nombre de points dans un cercle dont le rayon tendrait vers 0, on choisit un cercle dont le rayon est assez grand pour englober un nombre de points significatif et on pondère les points situés dans ce cercle en fonction de leur éloignement du centre[10].

L'estimation de densité est alors donnée par :

$$\hat{f}(x,y) = \frac{1}{r^2 n} \sum_{i=1}^{n} W\left[\frac{x - xi}{r}, \frac{y - yi}{r}\right]$$

où n est le nombre total d'observations, r le rayon arbitraire choisi, W la fonction de poids.

La fonction de poids produira une estimation d'autant plus lisse qu'elle accorde un poids décroissant de façon continue lorsque l'on s'éloigne du centre du cercle.

Pour représenter la densité obtenue en des points faisant déjà appel à deux dimensions, on peut soit tracer des courbes d'isodensité, soit utiliser la méthode d'affûtage : il s'agit de représenter les points ayant une densité supérieure à une valeur donnée. Lorsque l'on diminue la valeur frontière de la densité, le nombre de points pris en compte s'étend. Etudier cette évolution d'un graphique à l'autre donne un bon aperçu de l'évolution des densités en fonction des valeurs des variables.

---

[10] la distance utilisée pour cette pondération est la distance euclidienne, ce qui est cohérent avec le fait d'effectuer un comptage dans des cercles. On aurait pu définir la pondération à partir de la distance valeur absolue. Dans ce cas, on aurait compté les points dans des rectangles.

On a calculé la densité sur le revenu hors salaire et le salaire en utilisant la fonction de poids :

$$W(u,v)= \begin{cases} \dfrac{1+\cos\pi(\sqrt{u^2+v^2})}{\pi} & \text{si } u^2+v^2 \leq 1 \\ 0 & \text{sinon} \end{cases}$$

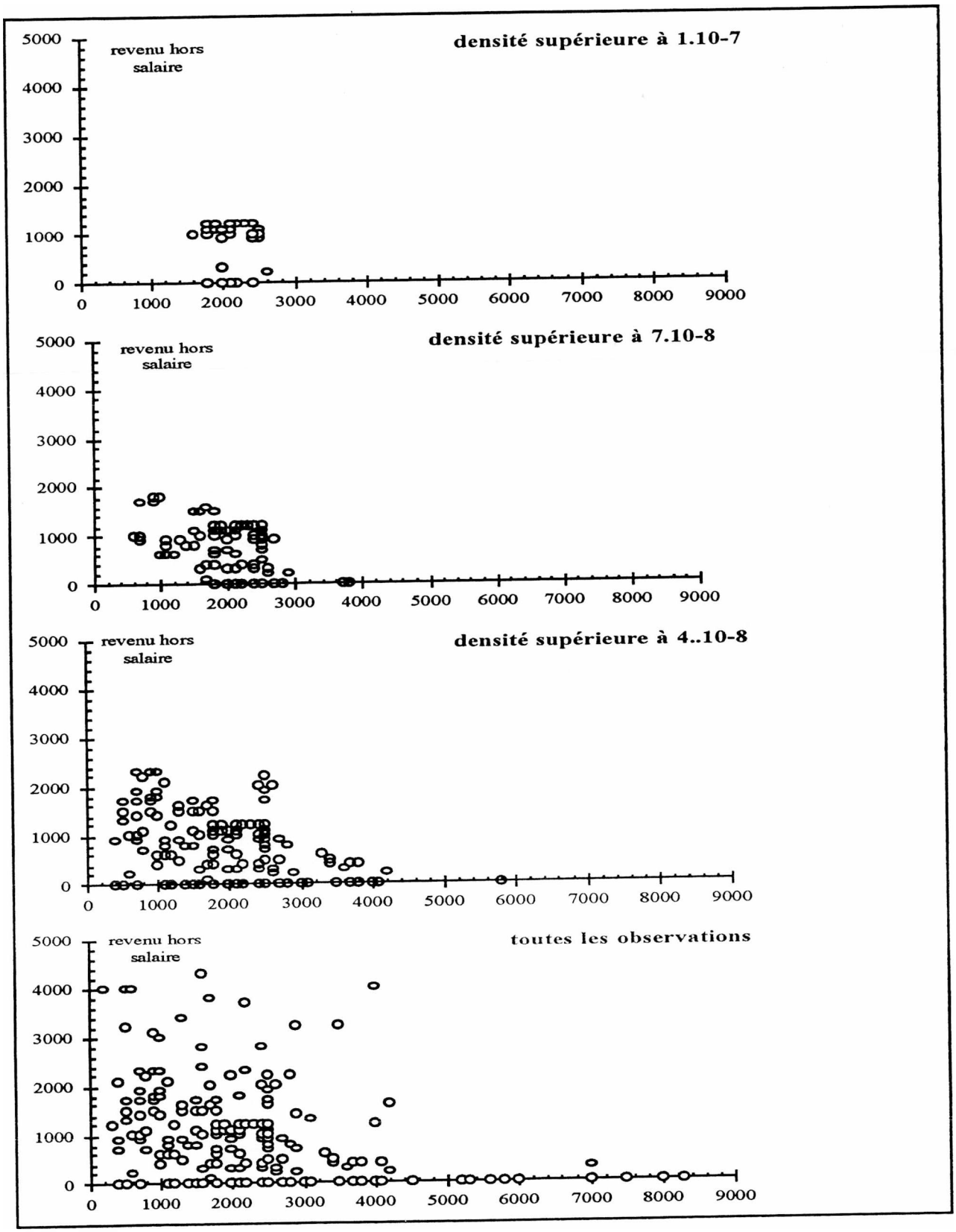
avec un rayon r=500

On ne présentera pas le graphique obtenu avec les courbes d'isodensité : les courbes ont trop de points d'intersection pour être lisibles.

Le graphique suivant montre la méthode d'affûtage appliquée à ce calcul de densité sur le revenu hors salaire et le salaire. Tous les axes des graphiques ont la même échelle, afin d'éviter les ambiguïtés dans la comparaison.

Graphique n°29

Densité du revenu hors salaire et du salaire (méthode d'affûtage)



Le graphique montre ici que :

-la zone où les points sont les plus nombreux correspond aux salaires compris entre 1600 francs et 1800 francs. Les revenus hors salaires sont alors ou bien nuls, ou bien sont situés entre 800 et 1200 francs ;

-la zone où la densité des points est un peu plus faible correspond aux valeurs du salaire comprises entre 1600 et 3800 francs. Les revenus hors salaire sont alors inférieurs à 1200 francs. A ces points s'ajoutent ceux pour lesquels le salaire est inférieur à 1000 francs et le revenu hors salaire situé ou bien autour de 1000 francs, ou bien autour de 1800 francs ;

-les points à faible densité correspondent aux valeurs du salaire inférieures à 1000 francs pour lesquelles le revenu hors salaire est inférieur à 2400 francs, aux valeurs du salaires supérieures à 3200 francs pour lesquelles le revenu hors salaire est inférieur à 800 francs ;

-les points à densité très faibles sont les points pour lesquels le salaire est très élevé (supérieur à 5200 francs) et les points situés au-dessus de la droite $y=-x+4000$.

Si on regarde l'évolution de la densité en fonction des valeurs du revenu hors salaire et du revenu, il apparaît donc que, en partant de la zone de forte concentration correspondant aux salaires compris entre 1800 et 2400 francs et aux revenus hors salaire compris entre 900 et 1200 francs, la densité augmente lorsque d'un côté les valeurs du salaire augmentent et les valeurs du revenu hors salaire diminuent et de l'autre côté lorsque les valeurs du salaire diminuent et les valeurs du revenu hors salaire augmentent.

Cette construction permet de préciser les remarques faites sur le tracé de cellulation. Il est plus précis car il fait appel à une estimation continue de la concentration des points. Il permet de répondre quantitativement à la question « comment évolue la concentration des points autour des valeurs d'une variable en fonction des valeurs de l'autre variable ? ».

Comme nous l'avions remarqué pour l'analyse monofactorielle, sa construction fait appel à des choix arbitraires. Ici le choix ayant le plus d'impact n'est pas celui du rayon du cercle ou de la fonction d'estimation, mais celui des valeurs frontières de la densité utilisées pour la méthode d'affûtage.

II-2 Représentation des résumés numériques d'une variable en fonction de l'autre variable

Les graphiques précédents montrent les variations de concentration des individus en fonction des valeurs de chacune des deux variables. Pour une variable continue, le résultat est peu lisible car les observations sont dispersées autour de l'éventuelle tendance générale.

Pour rendre le résultat plus visible, on étudie la variation d'une des variables (notée y) non plus en fonction de chacune des valeurs de l'autre variable (notée x), mais en fonction d'intervalles de ces valeurs.

Les graphiques suivants rendent compte de la distribution de y à l'intérieur de chacun des intervalles de x choisis et permettent la comparaison de ces distributions d'un intervalle à l'autre. Ils sont construits ou bien à partir de résumé numérique de y dans chaque intervalle de x, ou bien à partir de l'ensemble des valeurs de y.

Le choix des intervalles de x peut se faire de manière à obtenir des intervalles de largeur égale, ou bien en prenant des intervalles recouvrant le même nombre d'observations.

Dans le premier cas, la largeur des intervalles est déterminée par l'étendue de la distribution de x et par le nombre souhaité d'intervalles. Dans le deuxième cas, on choisit le nombre d'observations par intervalle en fonction du nombre total d'observations dans l'échantillon puis on en déduit le nombre d'intervalles. Les bornes des intervalles seront des quantiles d'ordre équidistant.

En fondant l'étude sur des intervalles de valeurs de x de même largeur, les tranches éloignées du mode de x compteront un nombre d'observations beaucoup plus faible que les intervalles proches du mode. Il est alors possible que ce nombre d'observations soit trop faible pour que les résumés de la distribution de y soient significatifs. Obtenir un nombre d'observations par intervalle significatif peut exiger, notamment pour les petits échantillons, que les intervalles soient peu nombreux.

Pour éviter cette ambiguïté et pour contrôler le nombre d'observations par intervalle, nous utiliserons dans la suite de l'étude des intervalles comprenant le même nombre d'observations et non des intervalles de même largeur.

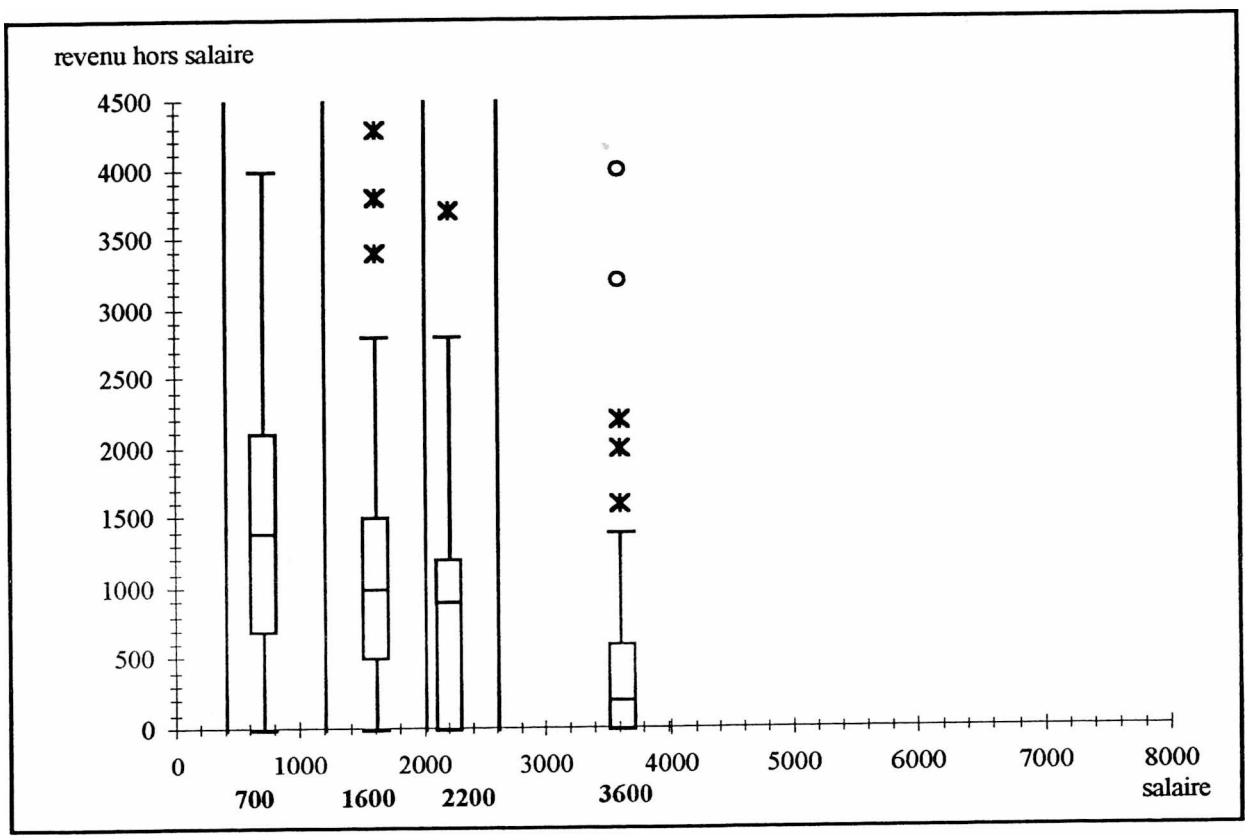
II-2-1 Evolution des résumés numériques d'une des variables selon les intervalles de l'autre

Utilisation des Box Plots

Les box plots de la variable y sont construites à l'intérieur de chacun des intervalles de x. On les place ensuite côte à côte. Contrairement aux graphiques utilisés pour la comparaison de deux variables, la dimension horizontale a cette fois une signification, puisqu'il s'agit des valeurs de x. Les box plot sont alors centrées sur la médiane de x à l'intérieur de chacun des intervalles. Leur largeur est arbitraire. Elle est la même pour toutes les boîtes et est telle qu'elle n'excède pas la largeur de l'intervalle de valeurs de x le plus étroit.

Graphique n°30

Box plot du revenu hors salaire en fonction du salaire réparti en 4 classes



Source : enquête Pauvreté CREDOC-CES 1995

On obtient un aperçu rapide de l'évolution d'une des variables en fonction de l'autre. Le fait de centrer les box plots dans chacun des intervalles sur la médiane de x permet de tenir compte de la répartition des valeurs de x à l'intérieur de chacun des intervalles.

La box plot des valeurs de y du premier intervalle correspond à des valeurs en général très faibles du salaire (700 francs), celle du second à des valeurs plutôt faibles (1600 francs), celle du troisième à des valeurs moyennes (2200 francs), celle du quatrième à des valeurs en général élevées (3600 francs). Le revenu hors salaire a tendance à diminuer lorsque le salaire augmente.

La box plot des valeurs de y correspondant aux très faibles valeurs de x est ainsi au-dessus des autres box plots. Sa médiane et son premier quartile sont légèrement plus élevés que ceux de la box plot des revenus hors salaire correspondant aux faibles valeurs de x. Son quartile supérieur et surtout sa patte supérieure sont beaucoup plus hauts. La distribution des valeurs du revenu hors salaire pour les très faibles valeurs du salaire est généralement un peu plus élevée que celles des revenus hors salaire des faibles valeurs du salaire. Elle est moins étirée vers les valeurs élevées du revenu hors salaire. Ces valeurs sont donc plus fréquentes dans le premier intervalle et plus proches de l'ensemble de la distribution. Dans le second intervalle, ces valeurs sont périphériques.

La comparaison des box plots du second et du troisième intervalle montre que, si dans les deux cas la médiane et la patte supérieure sont à peu près les mêmes, le quartile inférieur de la distribution correspondant aux valeurs moyennes du salaire est égale à la patte inférieure de la distribution correspondant aux faibles valeurs du salaire. Les revenus hors salaire sont beaucoup plus concentrés vers les basses valeurs pour les valeurs moyennes du salaire que pour les valeurs faibles.

La box plot des salaires élevés montre une distribution du revenu hors salaire globalement beaucoup plus faible que dans les autres intervalles, concentrée vers les basses valeurs du revenu hors salaire et s'étirant beaucoup vers les valeurs élevées. Deux points prennent des valeurs du revenu hors salaire très élevées mais ces points sont aberrants.

Le revenu hors salaire diminue globalement des valeurs très faibles aux valeurs moyennes du salaire, surtout parce que les valeurs élevées du revenu hors salaire sont de plus en plus éloignées de l'ensemble de la distribution et les valeurs faibles et très faibles de plus en plus fréquentes. Les revenus hors salaire correspondant aux salaires élevés sont très nettement concentrés sur les basses valeurs et globalement beaucoup plus faibles que ceux des autres intervalles.

Le tracé des box plots de y dans des intervalles de valeurs de x fournit une idée de l'évolution globale des valeurs de y en fonction de x ainsi que de l'évolution de la forme de la distribution de y en fonction des variations de x.

Parce qu'ils sont fondés sur des résumés numériques, ces tracés sont à la fois plus lisibles et moins précis que les tracés faisant appel à l'ensemble des données.

Leur construction nécessite de discrétiser une des variables. Comme toujours lorsque l'on discrétise une variable continue, l'interprétation du graphique repose alors sur des classes arbitraires et sur des discontinuités qui n'ont pas lieu d'être.

Evolution de la médiane de y en fonction d'intervalles de valeurs de x.

Le nombre de discontinuités est d'autant plus grand qu'il y a beaucoup d'intervalles. En augmentant le nombre d'intervalles de x, on peut obtenir une étude plus fine des variations de y en fonction de x. Mais on se heurte au risque d'avoir un trop faible nombre d'observations par classe pour que la représentation soit significative. Dans ce cas, on se contentera de tracer la médiane de y dans chacun des intervalles de x.

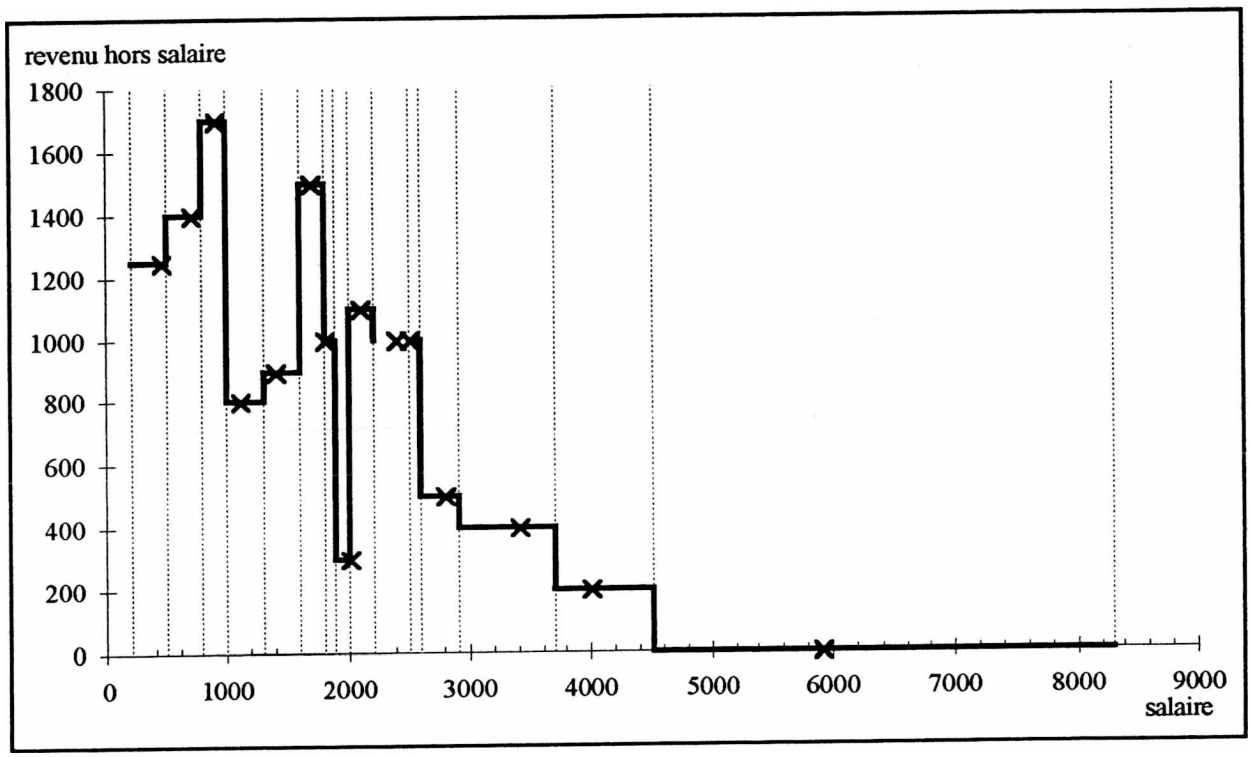
Le graphique suivant montre la représentation de la médiane du revenu hors salaire en fonction d'intervalles de valeurs de salaires. Le nombre d'intervalles est cette fois de 15. On a donc environ 13 observations par intervalle, ce qui est trop faible pour pouvoir tracer des box plots.

A l'intérieur de chacune des tranches de valeurs de x, on trace un trait horizontal ayant pour ordonnée la médiane de la distribution de y pour cette tranche. Pour mettre graphiquement en évidence l'évolution de la médiane de y, on joint ces traits horizontaux par des traits verticaux qui seront donc situés aux bornes des intervalles de valeurs de x.

La médiane de x à l'intérieur de chacun des intervalles de valeurs de x est représentée par une croix sur le trait de la médiane de y.

graphique n°31

Evolution de la médiane du revenu hors salaire en fonction du salaire réparti en 15 intervalles



Source : enquête Pauvreté CREDOC-CES 1995

Le graphique semble remettre en cause les observations faites précédemment sur la tendance générale du revenu hors salaire à diminuer lorsque le salaire augmente. En effet, la médiane du revenu hors salaire ne diminue constamment que pour un salaire supérieur à 2000 francs.

Le tracé obtenu ici est très irrégulier. Il fait apparaître un grand nombre de discontinuités qui ne sont dues qu'au choix arbitraire du nombre d'intervalles. Le faible nombre d'observations par intervalle rend par ailleurs peu significatives les variations des valeurs de la médiane observées.

Cependant, pour avoir une idée précise des variations de y en fonction de x, il est intéressant de s'en tenir à un tel nombre d'intervalles. Une solution pour conserver cette précision tout en ayant une courbe significative est de lisser les valeurs des médianes de x et de y pour chacun des intervalles.

Le principe du lissage est de remplacer chacune des valeurs de la série de données par une valeur qui tient compte de la valeur initiale et des valeurs voisines. Ce procédé permet d'obtenir des tracés plus réguliers que les tracés réels et permet donc à l'observateur de voir la tendance présentée éventuellement par la série. La courbe lissée est plus significative que la courbe de départ. En effet, les variations de la série de départ ne sont traduites dans la série lissée que si elles sont confirmées par plusieurs valeurs.

TUKEY propose comme méthode de lissage celle de la médiane de trois valeurs. On note xi les valeurs de la série. La méthode de la médiane peut être décomposée en cinq étapes.

* Première étape : si xi est une valeur non terminale ($i \neq 1$ et $i \neq n$) et si $xi \neq xi+1$ on remplace xi par la médiane de (xi, xi-1, xi+1). x1 et xn sont conservées inchangées. Soit x'i, cette valeur. A l'issue de la première étape, on obtient une série pour laquelle les valeurs terminales sont les mêmes que la série initiale, et qui présente un certain nombre de valeurs répétées dues au procédé employé : en effet, à chaque fois que le sens de la série change, l'utilisation de la médiane des trois valeurs sur deux valeurs consécutives les remplacera par le même nombre. La deuxième étape consiste à traiter les valeurs terminales ainsi que les valeurs répétées.

* Deuxième étape : Soit x1, la première valeur de la série. On considère que x1 est une valeur acceptable pour la série lissée si elle n'est pas éloignée de la valeur lissée la plus proche, x'2, de plus de deux fois l'écart entre ses deux valeurs les plus proches (x'2 et x'3). Si ce n'est pas le cas, on remplace x1 par une valeur x'1 acceptable, c'est à dire telle que :

$| x'1-x'2 | < 2 | x'2-x'3 |$ et assez proche de x1.

Pratiquement, on choisira x'2 si x'2 et plus proche de x1 que x'2 et x'2+2(x'2-x'3), et sinon on choisira x'2+2(x'2-x'3). De la même façon, xn est remplacée par la médiane de (xn,xn-1,xn-1-2(xn-2-xn-1)).

Les valeurs répétées sont traitées de la même façon : si x'i=x'i+1, alors on remplace x'i par la médiane de (x'i, x'i-1, x'i-1-2(x'i-2-x'i-1)) et x'i+1 par la médiane de (x'i+1, x'i+2, x'i+2-2(x'i+3-x'i+2)). Dans le cas où x'2=x'3, on ne cherche à remplacer que x'3.

Cette méthode permet de lisser d'une part les valeurs de la série initiale répétées deux fois et qui s'éloignent du reste de la distribution, d'autre part les valeurs de la série lissée répétée deux fois à cause du lissage et qui font apparaître des paliers éloignés du reste de la série.

* Troisième étape : on itère le processus (première et deuxième étapes), jusqu'à ce que le lissage ne modifie plus la série. Soit zi les valeurs ainsi obtenues.

* Quatrième étape : on remplace chacune des valeurs non terminales par la moyenne des deux valeurs les plus proches (celles situées juste avant et juste après). On obtient ainsi une valeur z'i=(zi-1+zi+1)/2. Les valeurs terminales sont inchangées.

* Cinquième étape : on remplace z'i par la moyenne de z'i et de zi.

Les quatrièmes et cinquièmes étapes permettent de lisser les discontinuités brusques dues à des valeurs trop écartées. Elles suppriment certaines répétitions de valeurs.

Tableau n°1

Lissage de la médiane du revenu hors salaire selon le salaire réparti en 15 intervalles

	Série initiale	Première étape médiane de trois	deuxième étape valeurs répétées et terminales	troisième étape Itération	quatrième étape Moyenne mobile	cinquième étape Série lissée
1	1250	1250	1250	1250	1250	1250
2	1400	1400	1400	1400	1325	1362,5
3	1700	1400	1400	1400	1150	1275
4	800	900	900	900	1150	1025
5	900	900	900	900	950	925
6	1500	1000	1000	1000	950	975
7	1000	1000	1000	1000	1000	1000
8	300	1000	1000	1000	1000	1000
9	1100	1000	1000	1000	1000	1000
10	1000	1000	1000	1000	850	925
11	1000	1000	700	700	750	725
12	500	500	500	500	550	525
13	400	400	400	400	350	375
14	200	200	200	200	200	200
15	0	0	0	0	0	0

Source : enquête Pauvreté CREDOC-CES 1995

Guide de lecture : au cours de la première étape par exemple, la 10ème et la 11ème observations prennent la même valeur 1000. La 8ème et la 9ème observations étant égales, la 10ème valeur reste nécessairement égale à 1000 (médiane de 1000, 1000 et 1000-2*(1000-1000)).

La 11ème observation est remplacée par la médiane de (1000, 500, 500-2(400-500)) donc par la médiane de (1000, 500, 700), qui est 700. Lors de l'itération, il n'y a par contre plus de changement. Les valeurs terminales sont également assez proches du reste de la série pour ne pas être modifiées.

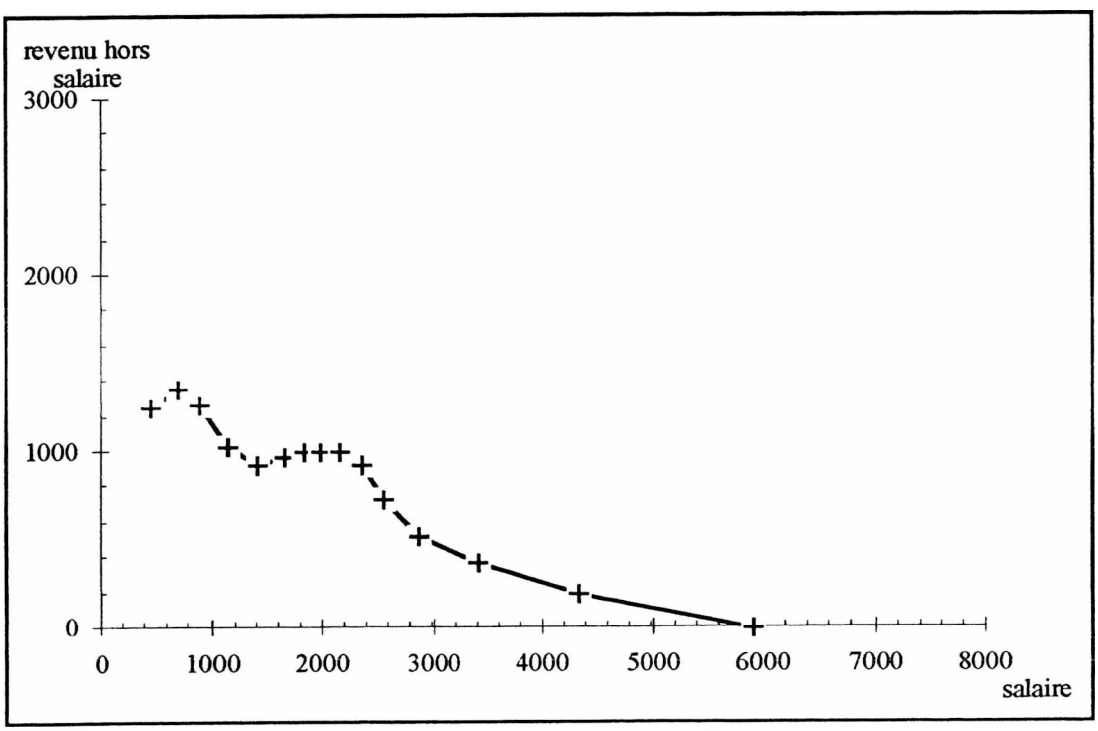
On obtient un grand nombre de valeurs répétées à la fin du lissage par la médiane. Ces valeurs répétées ne sont plus modifiées par l'itération (étape 3) mais sont estompées par la médiane mobile.

Pour les séries ordonnées (comme la médiane du salaire dans les 15 intervalles), le lissage par la médiane est sans effet : chacune des valeurs est elle-même la médiane des trois valeurs, puisqu'elles sont ordonnées. On effectue directement le lissage par la moyenne mobile.

On peut ainsi représenter la courbe obtenue en lissant la série constituée des médianes de la variable y à l'intérieur de chacun des intervalles de valeurs de la variable x en fonction de la série lissée des médianes de la variable x dans ces mêmes intervalles. On appelle trace médiane la courbe ainsi obtenue.

Graphique n°32

Trace médiane du revenu hors salaire en fonction du salaire (calculée à partir de 15 intervalles)



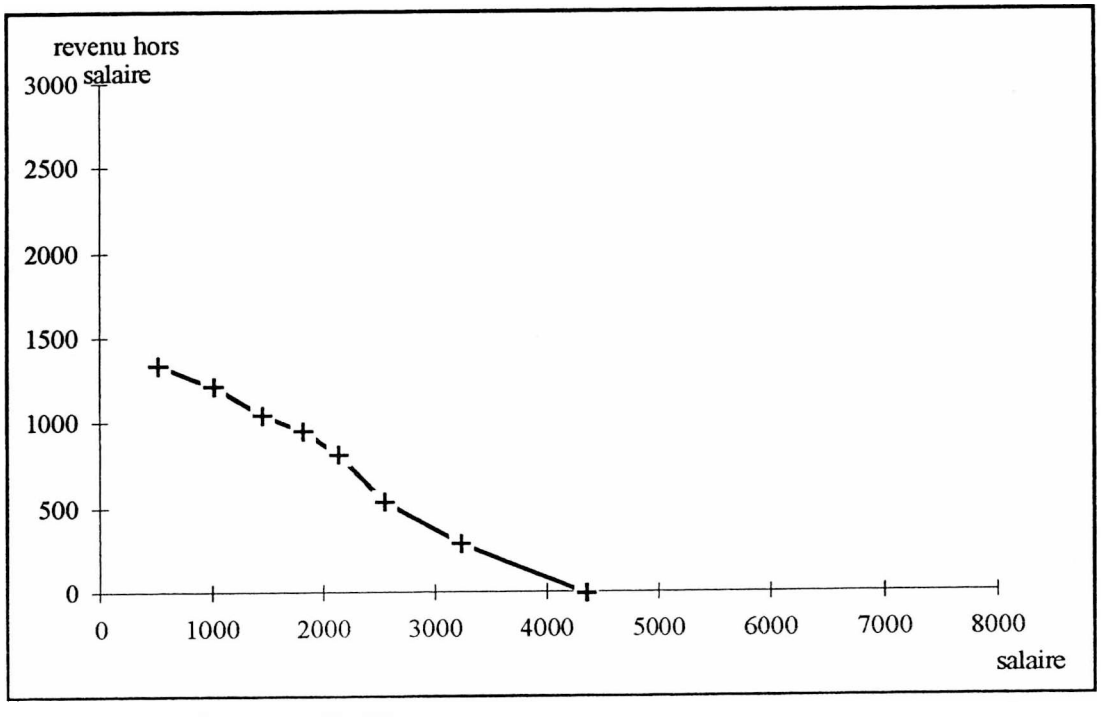
Source : enquête Pauvreté CREDOC-CES 1995

Le tracé ainsi obtenu est robuste dans la mesure où chaque point représente non seulement les valeurs des médianes dans un intervalle mais est également fonction des valeurs des médianes dans les intervalles voisins. Cette courbe peut être considérée comme une estimation grossière des médianes de la distribution de y et de x à l'intérieur d'intervalles infinitésimaux de valeurs de x.

Ce tracé dépend du nombre d'intervalles arbitrairement choisi. Lorsque l'on diminue le nombre d'intervalles, on diminue la précision de la courbe mais on augmente la robustesse du graphique.

Graphique n°33

Trace médiane du revenu hors salaire en fonction du salaire (calculée à partir de huit intervalles)



Source : enquête Pauvreté CREDOC-CES 1995

La trace médiane obtenue en utilisant huit intervalles est décroissante. Les revenus hors salaire montrent ainsi une nette tendance à décroître lorsque le salaire augmente. La courbe est la plus pentue pour les valeurs du salaire comprises entre 2200 et 2800 francs.

L'étude de la trace médiane obtenue en utilisant 15 intervalles conduit à relativiser ces observations. Il y a décroissance du revenu hors salaire médian en fonction du salaire lorsque

celui-ci est compris entre 800 et 1500 francs. Pour les valeurs du salaire comprises entre 1500 et 2800 francs, le revenu hors salaire reste stable. Dans cet intervalle de valeurs du salaire, il ne semble pas dépendre du salaire. Le revenu hors salaire diminue de nouveau pour les salaires supérieurs à 2800 francs, de façon assez régulière.

TUKEY propose de préciser le domaine de validité de telles relations en comparant la trace médiane réelle à celles qui correspondraient à une relation fortement positive entre les deux variables, à une relation fortement négative et enfin à une relation neutre entre les deux variables.

Les traces les plus positives et les plus négatives sont des traces théoriques définies en fonction des valeurs effectivement prises par les variables.

Etant données ces valeurs, les deux variables sont liées de la façon la plus positive possible si pour chaque individu, les valeurs de chacune des variables ont même rang. Avec les notations que nous avons utilisées jusqu'à présent, les couples (xi, yi) de valeurs des variables par individu et (x(j), y(j)) des valeurs ordonnées des variables sont les mêmes.

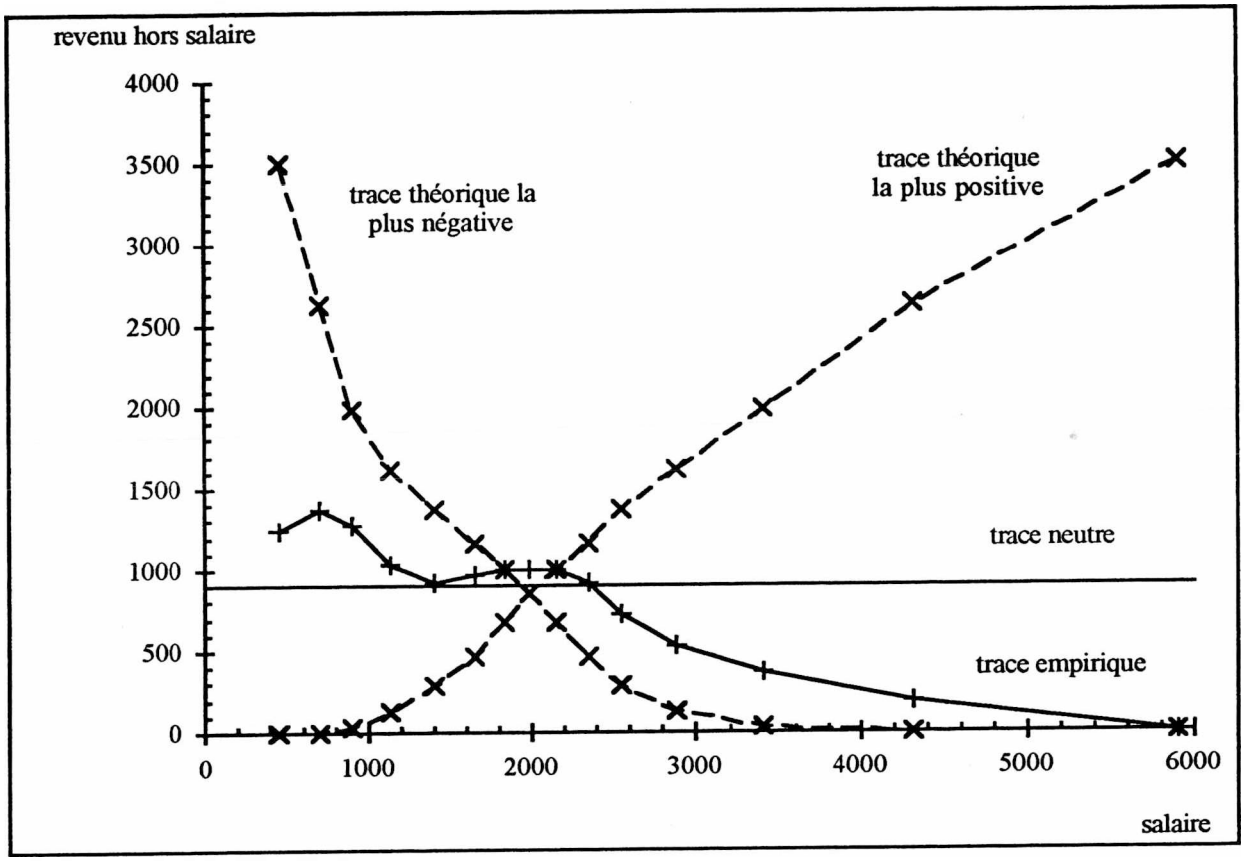
Les deux variables sont liées de la façon la plus négative possible si pour chaque individu, les valeurs de chaque variable ont un rang différent pour une même profondeur. Les couples (xi,yi) des valeurs de variables par individu correspondent aux couples (x(j),y(n-j)).

La trace médiane la plus positive et la trace médiane la plus négative sont calculées à partir de ces deux distributions théoriques de la même façon que la trace médiane empirique, et notamment en utilisant le même nombre d'intervalles pour pouvoir effectuer des comparaisons.

Le revenu hors salaire ne dépend pas du salaire si quel que soit l'intervalle de valeurs de la variable x, la médiane de la variable y est la même. La trace médiane de y en fonction de x est donc constante égale à la médiane de y.

Graphique n°34

Trace médiane empirique et traces médianes positive, négative, neutre pour le revenu hors salaire en fonction du salaire (calculées à partir de 15 intervalles)



Source : enquête Pauvreté CREDOC-CES 1995

Sur ce graphique, on lit alors quels sont les intervalles de valeurs de x dans lesquels la trace médiane se rapproche de la relation la plus positive, la plus négative ou de la neutralité.

Ici, on peut voir que la trace médiane est proche de la trace la plus négative pour les salaires inférieurs à 1400 francs et pour les salaires supérieurs à 2400 francs. Elle est proche de la neutralité pour les salaires compris entre 1400 francs et 2400 francs. On a donc retrouvé les observations faites précédemment, mais il est de plus intéressant de remarquer que la trace empirique diminue vers les valeurs élevées du salaire plus doucement que ne le fait la trace la plus négative. La relation négative entre le revenu hors salaire et le salaire pour les valeurs élevées du salaire est moins importante que l'étude de la trace médiane seule ne le laissait croire.

Pour préciser et quantifier ces observations, TUKEY propose de calculer des ratios de dépendance donnée par :

ratio de dépendance positive=(médiane empirique-médiane)/(médiane positive-médiane)

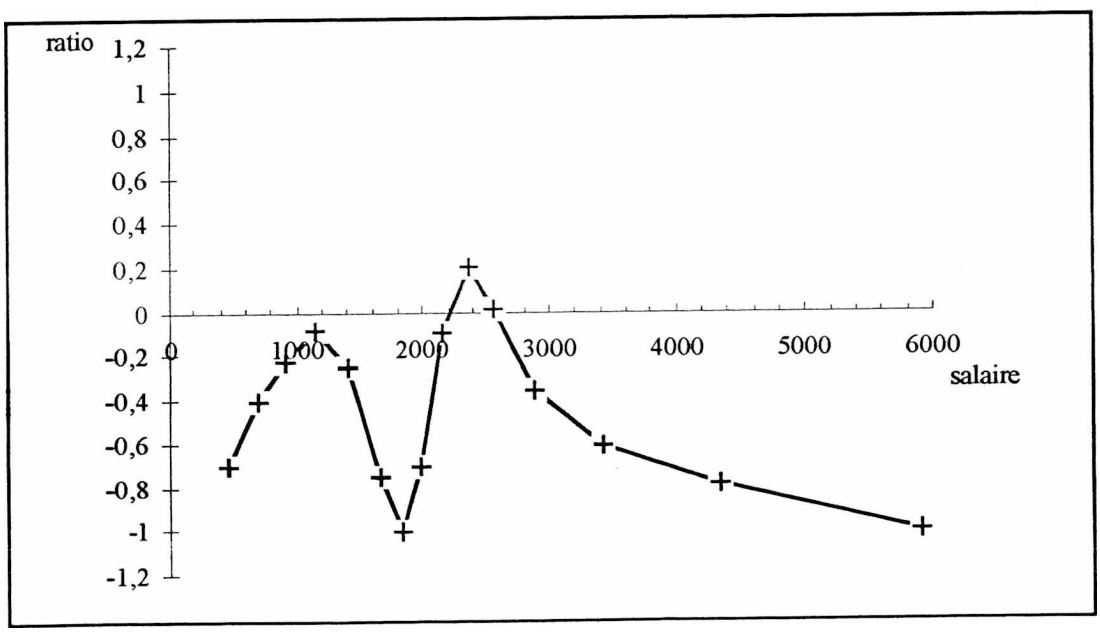
ratio de dépendance négative=(médiane empirique-médiane)/(médiane-médiane négative),

où les termes médiane empirique, médiane négative, médiane positive désignent respectivement la médiane de y à l'intérieur de chacun des intervalles de valeurs de x pour les distributions empirique, la plus négative, la plus positive. La médiane est ici la médiane de la distribution de y. Elle correspond à la trace neutre.

En pratique, on choisit le ratio que l'on calcule (positif ou négatif) en fonction de la proximité de la trace médiane empirique avec la trace médiane positive ou négative. Si elle est plutôt proche de la trace négative, on calcule le ratio négatif, si elle est plutôt proche de la trace positive, on calcule le ratio positif. On détermine le ratio ainsi choisi en fonction des valeurs réelles des distributions, c'est à dire avant lissage, puis on lisse les ratios obtenus. On représente finalement les ratios lissés en fonction de la série lissée des médianes de x.

Graphique n°35

Ratio de dépendance négative du revenu hors salaire en fonction du salaire (calculé à partir de 15 intervalles)



Source : enquête Pauvreté CREDOC-CES 1995

Dans les intervalles de valeurs de x dans lesquels les ratios sont proches de 0, la distribution médiane de y en fonction de x est proche de la neutralité. Dans les intervalles de valeurs pour lesquels les ratios sont proches de 1, la distribution médiane est proche de la distribution la plus positive. Enfin, lorsque les ratios sont proches de -1, la distribution est proche de la plus négative.

Sur le graphique précédent, on peut ainsi voir que la relation s'approche de la relation la plus négative pour les valeurs élevées du salaire (supérieures à 3000 francs). Elle tend vers la neutralité et est même légèrement positive pour les valeurs du salaire comprises entre 2000 et 2400 francs. Pour les salaires inférieurs à 2400 francs, elle est en général négative mais s'approche de la neutralité pour les valeurs du salaire voisines de 1000 francs.

Les tracés obtenus à partir de la trace médiane de y en fonction de x permettent de tracer une courbe continue rendant compte dans une certaine mesure de l'évolution d'une des variables en fonction de l'autre. L'utilisation d'une méthode de lissage est de plus un moyen de rendre plus robuste le tracé d'une telle courbe.

Ces graphiques peuvent cependant ne pas être significatifs. D'une part, parce qu'ils supposent en effet de discrétiser une des valeurs et donc d'effectuer un choix d'intervalles qui n'est pas neutre, d'autre part parce qu'ils ne rendent compte de l'évolution que d'un seul résumé numérique, ici la médiane. La dispersion des observations lorsque l'on s'éloigne de cette valeur centrale peut remettre en cause la tendance observée.

Les box plots à deux dimensions

Les traces médianes ne donnent des indications que sur l'évolution de la position de y en fonction de la position de x. Les box plots à deux dimensions ont pour but de donner des informations sur l'évolution de la forme et de la dispersion de la variable y en fonction de la position de x, et de confirmer les informations apportées par la trace médiane sur l'évolution de la position de y en fonction de x.

Pour l'étude monofactorielle d'une variable, on étudiait la position des quartiles par rapport à la médiane ainsi que celle des valeurs adjacentes afin d'obtenir des informations sur la forme et la dispersion de la variable. On va donc s'intéresser ici à l'étude de l'évolution des positions des quartiles et de la médiane de la variable y, calculés dans des intervalles de valeurs de la variable x. Le tracé de box plots parallèles permettait une première approche de cette évolution. Mais ils ne peuvent se faire que pour un nombre limité d'intervalles.

Comme pour la médiane, on calcule alors les quartiles de y à l'intérieur d'un nombre plus élevé d'intervalles. Pour obtenir un tracé robuste, chacune des valeurs des quartiles doit dépendre de valeurs voisines : nous allons utiliser une procédure de lissage pour obtenir des traces quartiles.

Dans la construction de ce graphique, on s'intéresse à l'évolution des positions relatives des quartiles par rapport à la médiane, plus qu'à leurs positions absolues. Nous ne lisserons donc pas les séries des quartiles elles-mêmes, mais les séries représentant l'écart des quartiles à la médiane.

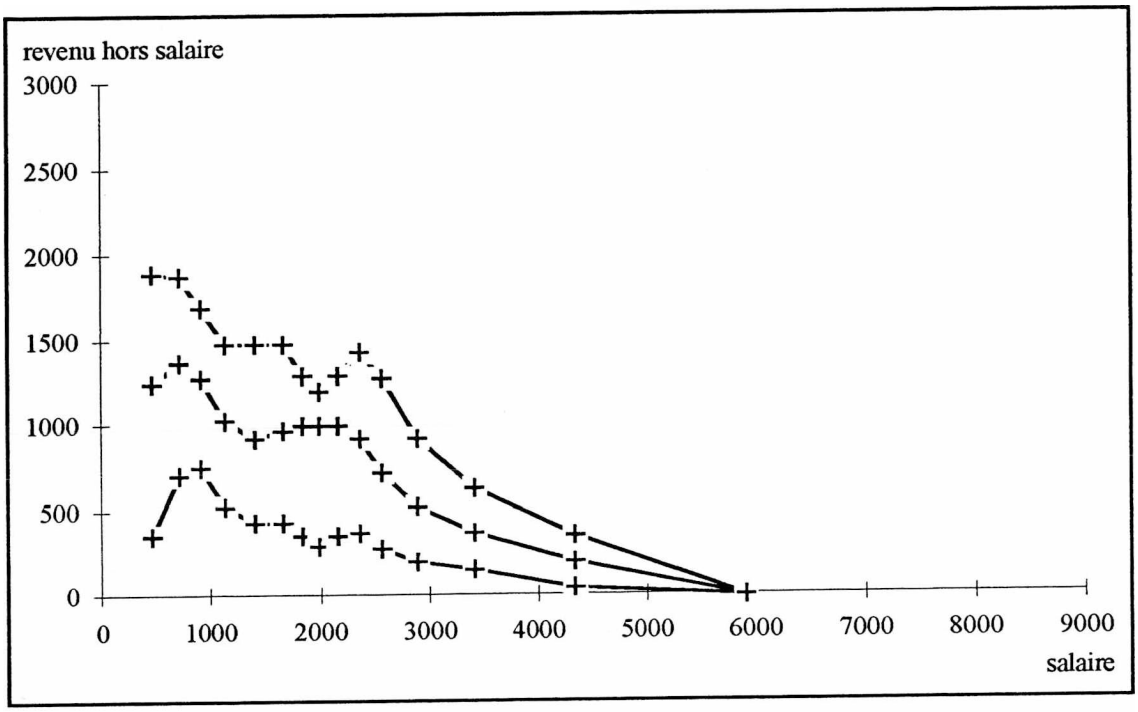
La trace quartile supérieure est ainsi obtenue en :

-définissant un certain nombre d'intervalles de valeurs de x, contenant le même nombre d'observations ;
-calculant le quartile supérieur et la médiane de la distribution de y à l'intérieur de chacun des intervalles de valeurs de x ;
-calculant la série des différences (quartile supérieur-médiane) pour chacun des intervalles ;
-lissant la série des différences obtenue précédemment, ainsi que la série des valeurs médianes de y ;
-calculant la valeur lissée des quartiles supérieurs, égale à la médiane lissée plus la valeur lissée des différences à l'intérieur de chacun des intervalles de x ;
-lissant la série des médianes de la variable x dans chacun des intervalles ;
-traçant la courbe des quartiles supérieurs lissés de la variable y en fonction de la médiane lissée de la variable x.

La trace quartile inférieure est obtenue de façon analogue.

Graphique n°36

Traces quartiles et médiane du revenu hors salaire en fonction du salaire calculées à partir de 15 intervalles



Source : enquête Pauvreté CREDOC-CES 1995

La comparaison, à partir du graphique obtenu, des traces quartiles et médiane permet de confirmer ou d'infirmer les hypothèses effectuées à partir de la trace médiane sur l'évolution de la position de y en fonction de x.

L'étude de l'évolution de la position relative des courbes donne des informations sur l'évolution de la dispersion de y en fonction de x.

Ainsi, le graphique précédent confirme les remarques faites sur l'évolution médiane de la distribution des revenus hors salaire en fonction des salaires mais permet de les nuancer. Pour les valeurs du salaire supérieures à 3000 francs, les trois courbes diminuent régulièrement et restent équidistantes. Pour ces valeurs, la distribution des revenus hors salaire est de plus symétrique. Le lien négatif entre les revenus hors salaire et les salaires concerne pour ces valeurs du salaire, non seulement les revenus hors salaire proches de la médiane, mais aussi ceux qui s'en écartent. Entre 1000 et 1400 francs, le lien médian reste négatif, mais avec une tendance des valeurs du revenu hors salaire à s'étirer de plus en plus vers les valeurs inférieures et à se concentrer vers les valeurs supérieures (restant inférieures au premier quartile). Entre

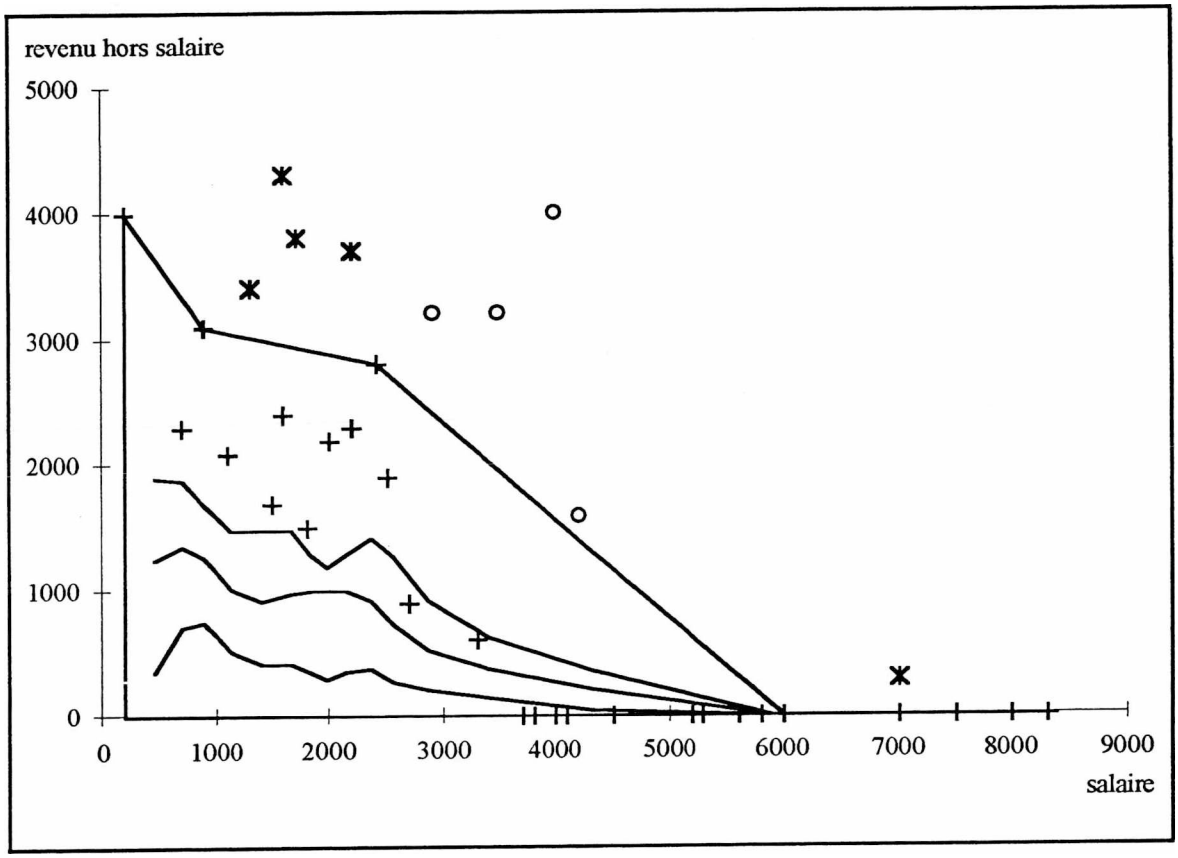
1400 et 2400 francs, nous avions vu que le lien médian est neutre, et parfois même s'approche d'un lien positif. L'étude des traces quartiles conduit à relativiser cette information. En effet, la courbe médiane est de plus en plus proche de la trace quartile supérieure, alors que la trace quartile inférieure chute visiblement et s'éloigne de la trace médiane. Pour ces valeurs du salaire, le lien médian est neutre mais les observations sont de plus en plus concentrées vers les basses valeurs du revenu hors salaire. Les valeurs supérieures à la médiane ne s'en écartent pas beaucoup.

Nous allons alors nous intéresser aux valeurs en dehors des intervalles interquartiles. Pour cela, on calcule dans chacun des intervalles de valeurs de x, les frontières supérieures et inférieures à partir des traces quartiles et médianes. On retrouve alors les observations adjacentes, périphériques et lointaines pour la distribution de y à l'intérieur de chacune des classes de x, telles qu'elles avaient été définies pour les box plots univariées. Pour chacune des valeurs de y adjacentes, on place sur le graphique précédent le point correspondant à l'observation qui la réalise. Si le ième individu prend une valeur yi adjacente, le point représenté sera le point (xi, yi).

On trace ensuite un polygone convexe joignant un certain nombre de points adjacents : ce polygone est le polygone convexe le plus petit englobant tous les points extérieurs. On peut alors représenter les points périphériques et lointains. Le résultat obtenu est une box plot à deux dimensions.

Graphique n°37

Box plot à deux dimensions du revenu hors salaire en fonction du salaire (calculée à partir de 15 intervalles)



Source : enquête Pauvreté CREDOC-CES 1995

Ce graphique apporte des informations supplémentaires sur la dispersion de la variable y en fonction de la variable x. La forme du polygone et son éloignement des traces quartiles, l'étude de la position des points adjacents et de celles des points périphériques et lointains permettent de mieux comprendre l'évolution de la distribution de y en fonction de x et de confirmer ou d'infirmer les hypothèses faites sur le lien entre les deux variables.

Il est possible, comme c'est le cas ici, que des points adjacents apparaissent entre les traces quartiles. Ils sont dus au fait que les tracés sont positionnés sur la médiane de x dans chaque intervalle, alors que les points adjacents correspondent aux valeurs empiriques de x.

Le graphique permet ici de voir que, pour les valeurs du salaire, entre 1400 et 2400 francs, pour lesquelles le revenu hors salaire médian diminue peu, le polygone adjacent, ainsi que les points adjacents eux-mêmes, se rapprochent des tracés : les valeurs du revenu hors

salaire supérieures à la médiane en restent proches, ce qui relativise cette augmentation du revenu hors salaire médian. Pour ces valeurs du salaire, les revenus hors salaire élevés, supérieurs à 3000 francs, sont alors soit périphériques, soit lointains, alors que pour les salaires les plus faibles, ces valeurs appartenaient au polygone adjacent. Pour les valeurs très élevées du salaire, supérieures à 6000 francs, les revenus hors salaire sont nuls sauf pour un point déviant.

Les box plots à deux dimensions permettent de valider ou de nuancer les hypothèses faites à la lecture de la trace médiane, notamment parce qu'elles apportent des informations sur l'évolution de la dispersion de la distribution de y en fonction de x. Alors que la trace médiane ne montrait que le lien pouvant exister entre les valeurs médianes de y et la variable x, les box plots parallèles montrent quel peut-être ce lien sur l'ensemble de la distribution de y.

Le polygone adjacent reste un tracé grossier car il est construit « au mieux » parmi les valeurs adjacentes. Obtenir un tracé plus précis se ferait au détriment de la significabilité des observations faites. Les box plots à deux dimensions sont construites à partir d'un choix arbitraire d'intervalles, choix dont il faut tenir compte dans l'interprétation. Même si ce tracé donne plus d'informations que ceux construits à partir de la seule médiane, il ne tient compte que de résumés numériques et non de l'ensemble des valeurs des deux variables.

II-2-2 Evolution de l'ensemble des valeurs d'une des variables en fonction de l'autre

La trace médiane était fondée sur l'étude du centre d'une des distributions (le milieu était représenté par la médiane), en fonction d'intervalles de valeurs de l'autre variable. Ces intervalles étaient fixes. Dans la prise en compte des valeurs de y pour le calcul du centre de la distribution par intervalle, on introduisait donc des discontinuités arbitraires, discontinuités que l'on atténuait par la méthode de lissage.

Un autre moyen pour éviter ces discontinuités est d'estimer le centre de la distribution de y non plus dans des intervalles de valeurs de x, mais en chaque point xi. Le nombre d'observations prenant la valeur xi n'est pas suffisant pour faire un calcul direct de la médiane de y en ce point. On fait donc appel à une méthode d'estimation. Pour toutes les valeurs xi prises par la variable x, on calcule le centre de la distribution de y en tenant compte des points ayant une valeur de x voisine de xi. Pour chaque valeur xi, on prend en compte le même nombre de points t, et on pondère ces points en fonction de leur éloignement de xi pour estimer le centre de y en cette valeur.

Plus précisément, l'estimation se fait de la façon suivante :

Soit xi, une valeur de la variable x. Soit (xi, yi) le point prenant la valeur xi pour la variable x et yi pour la variable y. Soit, $\hat{y}i$, la valeur estimée représentant le milieu de la distribution de y au point xi. On définit une fonction de poids sur les autres points de l'échantillon, de façon à ce qu'elle soit décroissante lorsque l'on s'éloigne de xi et non nulle uniquement pour les t points les plus proches.

Si di est la distance maximale entre xi et les t points les plus proches, cette fonction est donnée par

$$p_i(x)= \begin{cases} (1-\left|\frac{xi-x}{di}\right|^3)^3 & \text{si } \left|\frac{xi-x}{di}\right| < 1 \\ 0 & \text{sinon} \end{cases}$$

On trace alors une droite parmi ces points, estimée par la méthode des moindres carrés pondérés : on cherche les coefficients a et b minimisant :

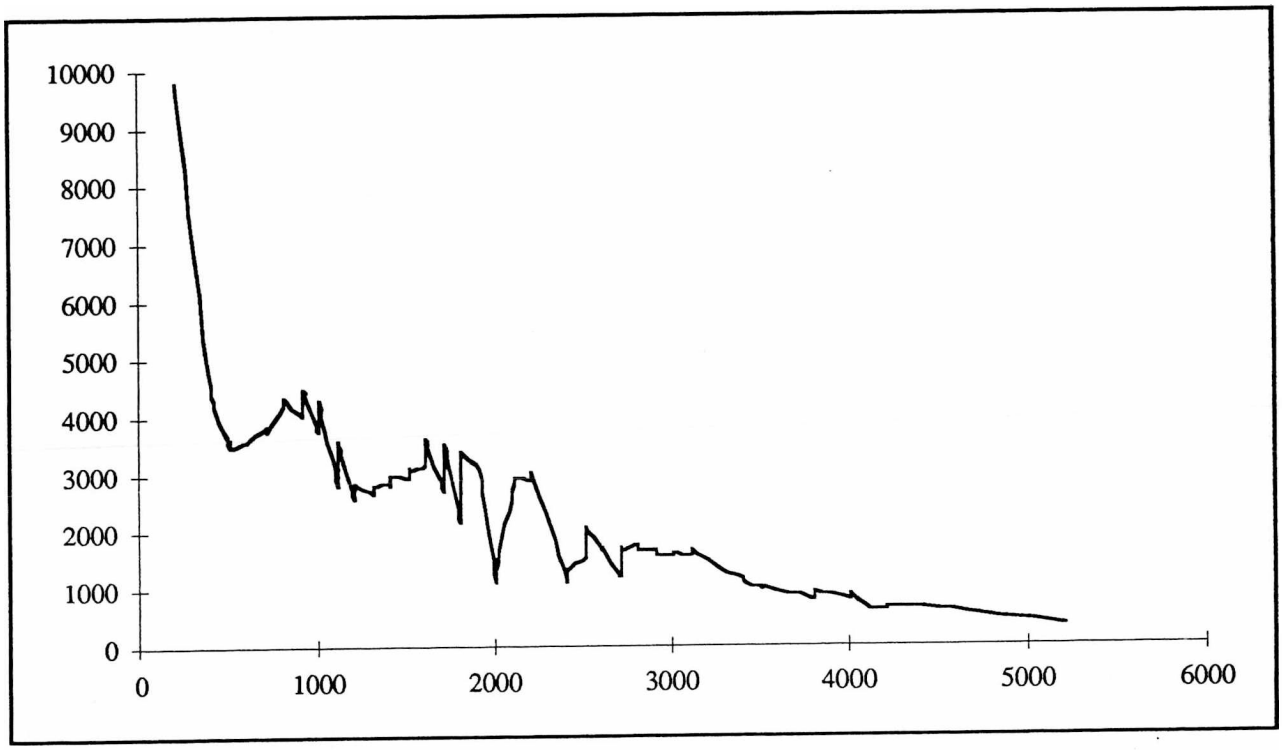
$$\sum_{j=1}^{nn} p_i(xj)(yj-a-bxj)^2$$

Le centre de la distribution de y au point xi est alors considéré comme étant le point se trouvant sur cette droite estimée ayant xi comme abscisse. Si $\hat{a}$ et $\hat{b}$ sont les coefficients de la droite estimée, la valeur $\hat{y}i$ est donc donnée par : $\hat{y}i = \hat{a} + \hat{b}xi$

On effectue ce calcul pour chacune des valeurs prises par la variable x.

Graphique n°38

Estimation du centre de la distribution du revenu hors salaire en fonction du salaire



Source : enquête Pauvreté CREDOC-CES 1995

Ce procédé a l'avantage de prendre en compte chacune des valeurs du salaire, ce qui n'était pas le cas pour la médiane. Le résultat obtenu est donc plus significatif que celui issu du lissage des médianes.

Cependant, la méthode utilisée fait que ce tracé est très sensible aux variations brusques des valeurs des deux variables, ce qui n'était pas le cas lorsque l'estimation utilisait la médiane.

Les graphiques utilisés pour étudier la liaison entre deux variables sont complémentaires. Les graphiques représentant l'ensemble des valeurs des deux variables sont souvent peu lisibles. Le tracé de la variable à deux dimensions ne permet pas de mettre en évidence une relation entre les deux distributions. Il a cependant pour avantage de montrer l'ensemble des valeurs prises par les deux variables. Il sera donc utilisé pour comparer les conclusions issues des autres graphiques avec les valeurs réelles des deux variables. Les tracés de cellulation et d'affûtage sont plus accessibles à l'analyse. Ils permettent l'étude de l'évolution des concentrations des points en fonction des valeurs de chacune des deux variables. Leur construction fait appel à des discontinuités arbitraires.

Les tracés fondés sur les résumés numériques ont pour but de montrer comment évolue une variable en fonction de l'autre. Les deux variables ne sont plus traitées de façon symétrique. Le tracé des box plots parallèles ne peut être construit qu'à partir d'un nombre d'intervalles assez faible, chacun d'eux contient alors suffisamment d'observations pour avoir une comparaison significative. Il est donc peu précis, mais simple à construire, il permet d'avoir un premier aperçu de l'évolution de la position, de la forme et de la dispersion d'une variable en fonction de l'autre.

Les box plots à deux dimensions font appel à des méthodes de lissage. Ces procédés permettent de rendre plus robustes les courbes obtenues : les variations des courbes ne sont conservées que si elles sont confirmées par plusieurs valeurs. Ils permettent d'étudier l'évolution du centre de la distribution d'une variable en fonction d'une autre ainsi que celle de la dispersion des valeurs autour de ce centre. Il est ainsi possible de retrouver une relation fonctionnelle qualitative, et les intervalles de valeurs de x dans lesquels cette relation est valable.

L'intérêt de ces graphiques est de montrer l'existence d'une dépendance entre les deux variables au delà d'une simple corrélation.

Tous les tracés utilisés ici, mais, notamment les tracés de cellulation les traces quartiles et médiane montrent que les revenus hors salaire ont tendance à diminuer lorsque les salaires augmentent.

La plupart des individus sont concentrés sous la droite « revenus hors salaire=4000-salaires » (cellulation). Pour les salaires inférieurs à 1000 francs, il est difficile de voir clairement une relation entre les deux variables. Il existe un certain nombre d'individus pour lesquels le revenu hors salaire et le salaire sont faibles. Cependant, la box plot à deux

dimensions de même que la cellulation, montre que c'est dans cet intervalle que les revenus hors salaire se répartissent le mieux jusqu'aux valeurs élevées du salaire. Le tracé de densité indique de plus que les individus ayant à la fois un faible salaire et un faible revenu hors salaire sont peu nombreux.

Pour les salaires compris entre 1000 et 1400 francs, le revenu hors salaire diminue rapidement, mais reste relativement bien réparti entre ses faibles valeurs et les valeurs les plus élevées.

Entre 1400 et 2800 francs, le revenu hors salaire médian diminue peu en fonction du salaire. Par contre, les valeurs supérieures à ce revenu hors salaire médian en sont de plus en plus proches, alors que les valeurs inférieures s'en éloignent.

Pour les salaires supérieurs à 2800 francs, les revenus hors salaire diminuent de plus en plus en fonction du salaire. Leur concentration sur les basses valeurs augmente. Les revenus hors salaire sont nuls lorsque le salaire est supérieur à 6000 francs.

Etude des relations entre variables quantitatives continues et variables qualitatives

Au cours des deux chapitres précédents, nous avons étudié :

- les valeurs de chacune des variables quantitatives qu'un individu donné a le plus de chances de prendre ;
- les valeurs d'une variable quantitative qu'un individu donné a le plus de chances de prendre sachant quelle valeur de l'autre variable quantitative il réalise.

Nous allons maintenant déterminer comment on peut expliquer et préciser ces probabilités en fonction des caractéristiques de l'individu observées sur des variables qualitatives. Plusieurs liens possibles vont être considérés :

* Lien entre une variable qualitative et une variable quantitative : l'étude a ici pour but de montrer si l'appartenance d'un individu à une catégorie donnée de la variable qualitative a une influence sur les valeurs de la variable quantitative qu'il est susceptible de prendre. Pour chacune des modalités de la variable qualitative, on montre s'il existe une réponse à la question : « quelles valeurs de la variable qualitative un individu possédant cette caractéristique a-t-il le plus de chances de prendre ? » Au cours de cette analyse, on étudie également si le fait qu'un individu prenne certaines valeurs caractéristiques de la variable quantitative peut être expliqué par son appartenance à une certaine catégorie de la variable quantitative. On cherche une réponse à la question « A quelle catégorie de la variable qualitative appartiennent en général les individus prenant ces valeurs ? »

* Lien entre plusieurs variables quantitatives et une ou plusieurs variables qualitatives : on cherche à préciser l'influence des variables qualitatives sur les variables quantitatives, en étudiant non plus les relations éventuelles séparément, mais l'interaction qui peut exister entre ces relations. On s'intéressera à :

- la liaison de deux variables quantitatives en fonction des modalités d'une qualitative ;
- la répartition des individus selon les valeurs des deux variables quantitatives en fonction des modalités de plusieurs variables qualitatives.

Dans la suite du chapitre, le terme « modalités » sera employé pour « modalités d'une variable qualitative », le terme « valeurs » sera employé pour « valeurs d'une variable quantitative ».

I- LIEN ENTRE UNE VARIABLE QUANTITATIVE ET UNE OU PLUSIEURS VARIABLES QUALITATIVES

On essaie de déterminer dans quelle mesure les modalités de la variable qualitative ont une influence sur les valeurs de la variable quantitative. On cherche donc à représenter les individus à la fois selon les valeurs et les modalités.

I-1 La répartition des valeurs d'une variable quantitative selon les modalités d'une variable qualitative

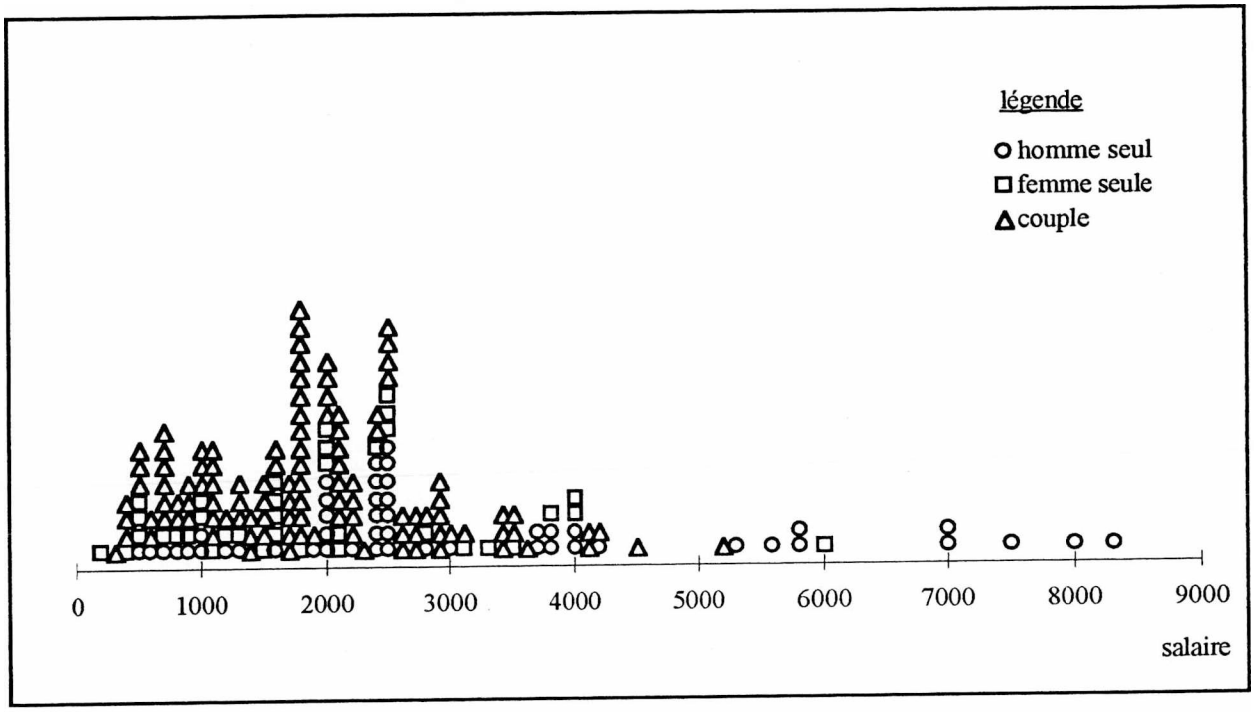
Comme lors des étapes précédentes de l'analyse, on distingue les graphiques représentant toutes les observations de ceux présentant des résumés numériques calculés à partir des variables étudiées.

I-1-1 Représentation de l'ensemble des valeurs de la variable quantitative et modalités de la variable qualitative

Pour représenter les modalités sur le graphique unidimensionnel, on remplace les points représentant les observations par des symboles différents selon les modalités. La représentation unidimensionnelle fait appel à un décalage vertical des points afin d'éviter les problèmes de superposition.

Graphique n°39

Représentation unidimensionnelle du salaire et situation familiale



Source : enquête Pauvreté CREDOC-CES 1995

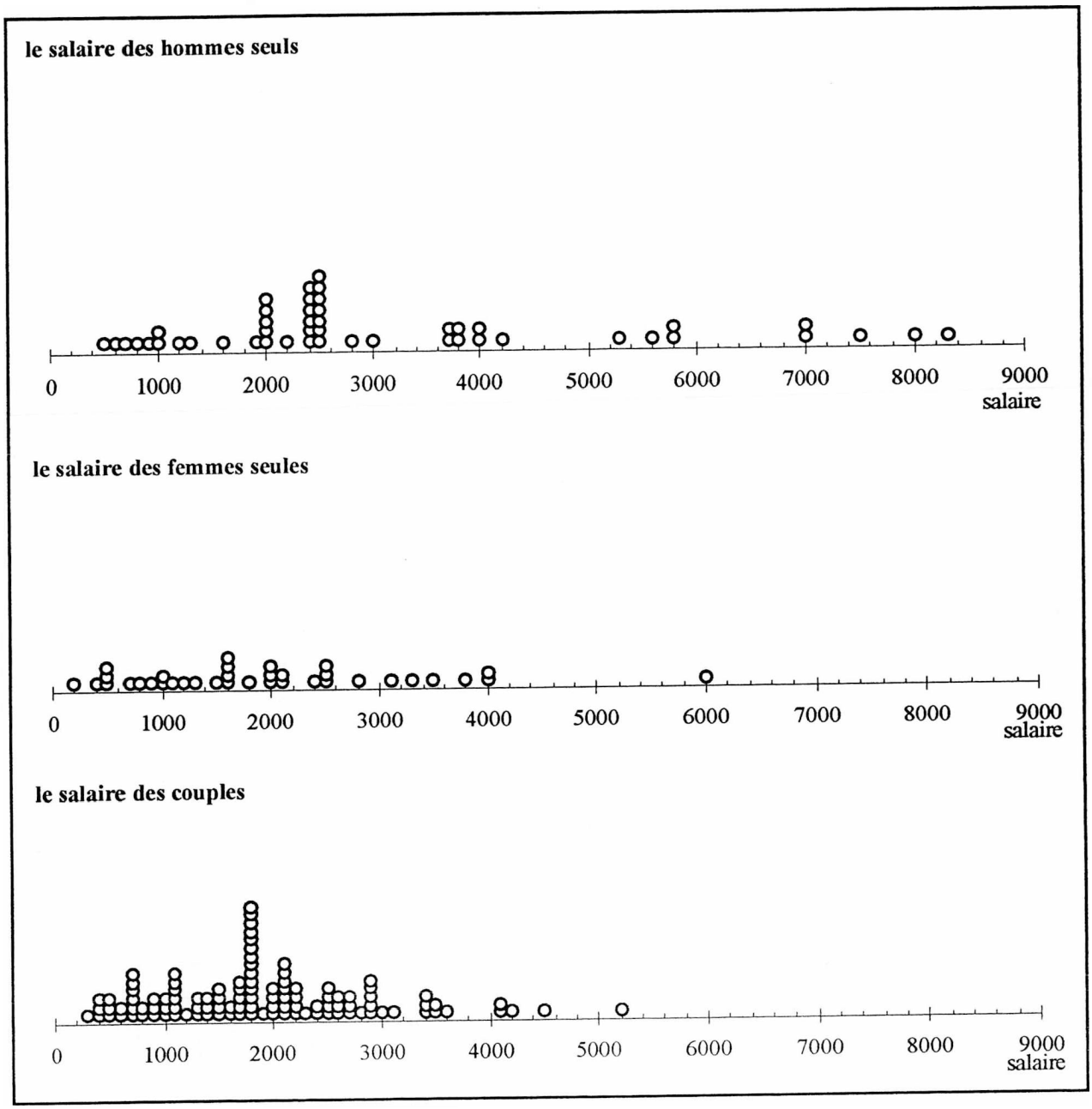
Le graphique obtenu ici a pour avantage de représenter toutes les observations mais est peu lisible. Sa construction exige un décalage vertical des points assez important pour que l'on puisse distinguer les symboles montrant les différentes modalités de la variable qualitative. Il nécessite donc une place verticale importante, plus que ce n'était le cas pour les représentations unidimensionnelles d'une variable qualitative.

Il est surtout lisible lorsque l'on s'écarte du mode. Il permet de savoir qui sont les individus prenant ces valeurs peu fréquentes. Le graphique précédent montre par exemple que ce sont les hommes seuls qui ont le salaire par unité de consommation le plus élevé (supérieur à 5000 francs).

Une possibilité pour étudier la répartition des valeurs de la variable selon les modalités pour les valeurs fréquentes est de tracer une représentation unidimensionnelle par modalité et de comparer ces représentations.

Graphique n°40

Représentations unidimensionnelles du salaire pour chaque situation familiale



Source : enquête Pauvreté CREDOC-CES 1995

Ce graphique montre la répartition des valeurs en fonction des modalités ainsi que la façon dont les modalités réalisent les différentes valeurs de la variable. Il permet d'étudier :

-les différences dans les positions des distributions de la variable quantitative et notamment l'évolution du mode de la variable quantitative selon les modalités, ainsi que celles les valeurs extrêmes ;
-les différences dans la dispersion et notamment dans l'étendue selon les modalités ;
-la nature des individus réalisant le mode et les valeurs extrêmes de la variable quantitative.

Cette représentation nécessite une place verticale importante. Elle ne peut donc être facilement utilisée en perspective d'autres graphiques comme c'était le cas pour la représentation unidimensionnelle d'une variable. Son interprétation peut donner lieu à des ambiguïtés, parce qu'elle conduit à comparer des distributions ayant un nombre d'observations différent.

Modalités et valeurs sur les diagrammes stem and leaf

Sur les diagrammes stem and leaf, les valeurs sont représentées par des symboles ayant déjà une signification, puisqu'il s'agit du chiffre complétant le tronc. On ne peut donc différencier les modalités de la variable qualitative en représentant les valeurs par des symboles différents. On construira les diagrammes stem and leaf pour chacune des modalités et on les placera côte à côte. Les chiffres du tronc sont les mêmes pour chacune des modalités, ce qui permet la comparaison.

Graphique n°41

Diagramme stem and leaf du salaire selon les modalités de la situation familiale

	Homme seul	Femme seule	couple
8	03		
7	005		
6		0	
5	3688		2
4	002	00	0024
3	07788	1358	0444446
2	00000244444455555558	0001145558	000000000002222244444466666688888
1	002369	00123566668	0000000002222444444466666668888888888888888
0	56789	24555789	24444466666666688888

les données doivent être multipliées par 100

Source : enquête Pauvreté CREDOC-CES 1995

Cette représentation permet de comparer la position (le mode et les valeurs extrêmes), la dispersion (l'étendue et les intervalles de valeurs où la concentration est élevée en dehors du mode), la forme (la répartition des observations lorsque l'on s'éloigne du mode), de la variable quantitative selon les modalités. Cette comparaison peut se faire en tenant compte des valeurs exactes des données.

Comme pour la représentation unidimensionnelle, elle montre également quelle est la modalité réalisant les valeurs extrêmes, le mode de la variable.

Son interprétation est rendue très ambiguë par le fait que ces diagrammes sont construits à partir d'un nombre d'observations différent pour chacune des modalités. De plus, comme cela a été le cas pour toutes les représentations de la classe des histogrammes, elle introduit des discontinuités arbitraires. L'impression visuelle qu'elle produit peut conduire à des erreurs d'interprétation.

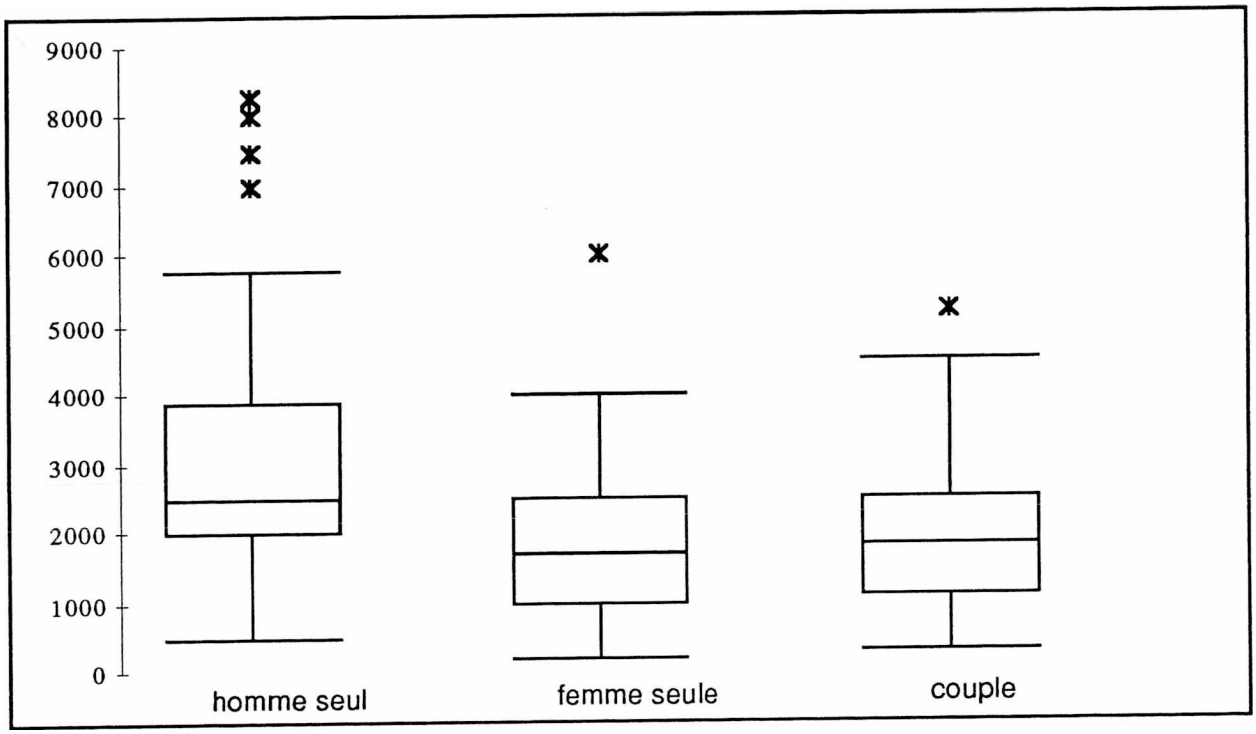
Finalement, les diagrammes stem and leaf sont ici plus utiles comme instrument d'enregistrement des valeurs de la variable selon les modalités que comme outil graphique. Ils ne peuvent être construits que pour des variables ayant un nombre de valeurs relativement faible, et bien réparties selon les modalités.

I-1-2 Représentation de résumés numériques de la variable quantitative et modalités de la variable qualitative

La méthode utilisée ici est très proche de celle appliquée à la comparaison de deux ou plusieurs distributions continues. Elle consiste à construire les box plots de la variable continue pour chacune des modalités de la variable qualitative et de les placer côte à côte.

Graphique n°42

Box plot du salaire par modalité de la situation familiale



Source : enquête Pauvreté CREDOC-CES 1995

Les box plots permettent de voir rapidement comment évoluent la position, la dispersion et la forme de la variable quantitative selon les modalités de la variable qualitative, par l'étude des positions relatives des boîtes, de leur taille, des observations déviantes. Elles ont pour avantage de permettre une comparaison fiable pour des tracés effectués sur des nombres d'observations différents.

Elles apportent une réponse graphique sur la façon dont les valeurs évoluent dans leur ensemble selon les catégories. Par contre, parce qu'elles n'utilisent que des résumés numériques, elles n'apportent pas d'information sur la qualification des valeurs en fonction des modalités.

Leur construction exige que le nombre d'observations pour chacune des modalités soit suffisant pour que les résumés numériques aient une signification.

I-2 Exemple d'analyse : le salaire en fonction de la situation familiale

On utilise les représentations précédentes pour répondre aux deux questions :

-comment évolue le salaire des personnes salariées en situation de grande pauvreté en fonction de leur situation familiale ?

-quelle est en général la situation familiale des salariés ayant le salaire par unité de consommation le plus élevé ? Des salariés ayant un salaire proche du mode (2000 francs) ?

Les box plots montrent qu'en général les salaires par u.c. des hommes seuls sont plus élevés que ceux des femmes seules et des couples.

La médiane des salaires des hommes seuls est de 2500 francs, celle des femmes seules de 1900 francs, celle des couples de 2000 francs. L'étude de la position des quartiles et de la longueur des pattes renforce la différence entre les salaires des hommes seuls et ceux des femmes seules et des couples. En effet, la médiane de la distribution des salaires pour les hommes seuls est très proche du quartile inférieur, alors que les distributions des salaires pour les femmes seules et les couples sont relativement symétriques. Non seulement la médiane des salaires des hommes seuls est ainsi plus élevée, mais de plus, dans l'intervalle interquartile, les valeurs inférieures à la médiane en restent très proches alors que les valeurs supérieures s'en éloignent rapidement.

Les pattes sont dans tous les cas symétriques, les différences observées entre la distribution des salaires pour les couples et celle pour les femmes seules (et notamment le fait que la patte supérieure de la distribution des couples est un peu plus grande que celle des femmes) ne sont pas significatives.

L'étude des représentations unidimensionnelles montre que les salaires des hommes sont beaucoup plus étendus et dispersés que les salaires des couples ou que ceux des femmes. Le fait que le nombre d'observations pour chacune des catégories est très différent ne permet pas de comparaison plus détaillée.

Les représentations unidimensionnelles indiquent que ce sont les hommes qui prennent en général les valeurs les plus élevées du salaire (supérieures à 4000 francs).

II- LIENS ENTRE DEUX VARIABLES QUANTITATIVES ET UNE OU PLUSIEURS VARIABLES QUALITATIVES

On s'intéresse à la façon dont la répartition des individus en fonction des valeurs des deux variables quantitatives est conditionnée par leur appartenance aux catégories des variables qualitatives.

On étudiera dans un premier temps l'influence d'une variable qualitative sur la répartition des individus selon deux variables quantitatives, puis celle de plusieurs variables qualitatives sur cette répartition.

II-1 Influence d'une variable quantitative sur la liaison entre deux variables qualitatives

L'étude peut s'appuyer sur :

-l'identification des individus sur les tracés représentant l'ensemble des valeurs des variables quantitatives et sur la comparaison de ces tracés effectués modalité par modalité ;

-la comparaison des tracés utilisant les résumés numériques d'une variable quantitative en fonction d'intervalles de valeurs de l'autre variable quantitative effectuée modalité par modalité.

Pour être significatifs, les résumés numériques des variables doivent être calculés à partir d'un nombre d'observations assez important. Tracer les résumés numériques d'une variable en fonction des valeurs de l'autre variable et des modalités suppose de ne retenir d'abord que les individus appartenant à une catégorie de la variable qualitative, puis de répartir ces individus dans des intervalles de valeurs d'une variable quantitative et enfin de calculer les résumés numériques de l'autre variable quantitative. Ces résumés ne seront donc construits que sur une fraction des données. Pour un échantillon de la taille de celui que nous utilisons ici, le nombre d'individus est alors trop faible pour qu'une telle étude soit possible.

Nous utiliserons d'abord la représentation de l'ensemble des valeurs d'une variable quantitative en fonction d'une autre, puis le tracé de densité et la méthode d'affûtage à deux dimensions.

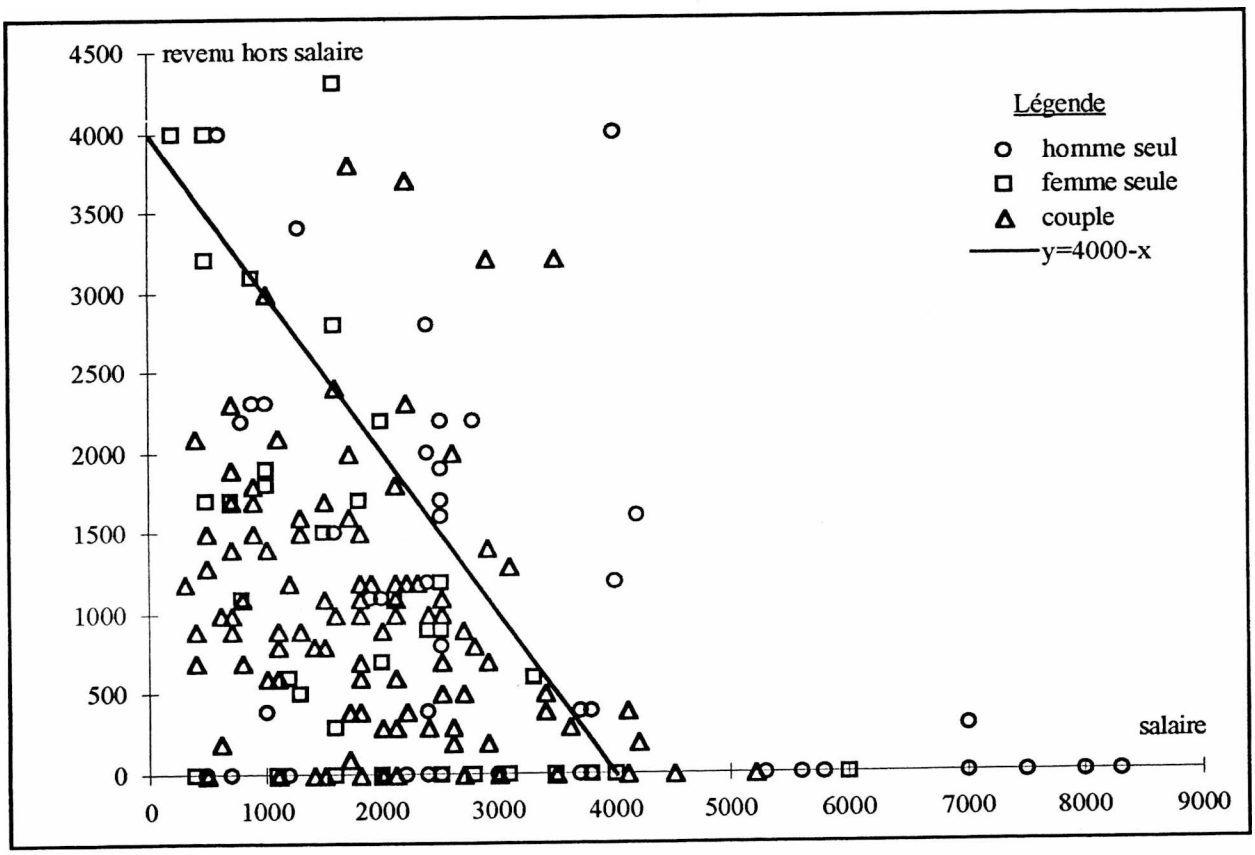
II-1-1 Influence d'une variable qualitative sur la représentation des individus en fonction des valeurs des deux variables

On observe la répartition des individus en fonction des valeurs des deux variables et de la modalité de la variable qualitative qu'ils réalisent.

Comme nous l'avons fait pour la représentation unidimensionnelle, nous allons remplacer les points du graphique d'une des variables qualitatives en fonction de l'autre par des symboles différents selon les modalités de la variable qualitative.

Graphique n°43

Représentation du revenu hors salaire en fonction du salaire
Individus identifiés par leur situation familiale



Source : enquête Pauvreté CREDOC-CES 1995

Dans les parties du graphique les plus denses, la superposition des points ne permet que difficilement l'identification des individus en fonction des modalités. Comme c'était le cas pour la représentation unidimensionnelle, ce graphique est surtout utilisable pour caractériser les individus qui prennent les valeurs les moins fréquentes et ceux qui ne respectent pas l'éventuelle relation observée lors de l'étude de la liaison des deux variables.

Dans le cas du revenu hors salaire en fonction du salaire, nous avions vu que la plupart des points étaient situés en-dessous de la droite y=4000-x. Nous avons donc tracé cette droite sur le graphique précédent afin de voir qui étaient les individus situés au-dessus. Il apparaît ici qu'il s'agit pour les salaires les plus faibles et les revenus hors salaire les plus élevés des femmes seules, pour les salaires les plus fréquents et les revenus hors salaire les plus élevés des couples et pour les salaires les plus élevés pour des revenus hors salaire moyens ou faibles des hommes seuls.

La représentation des valeurs d'une variable en fonction des valeurs d'une autre permettait de mettre en évidence une première relation fonctionnelle grossière entre les deux variables. Qualifier les individus sur ce graphique fait apparaître une éventuelle caractéristique des individus qui ne respectent pas la relation.

Ce graphique se heurte rapidement à un problème de lisibilité. On ne peut ainsi rien en déduire quant aux caractéristiques des individus respectant la relation. Pour augmenter la lisibilité, nous allons reconstruire le graphique modalité par modalité de la variable qualitative et comparer les tracés obtenus.

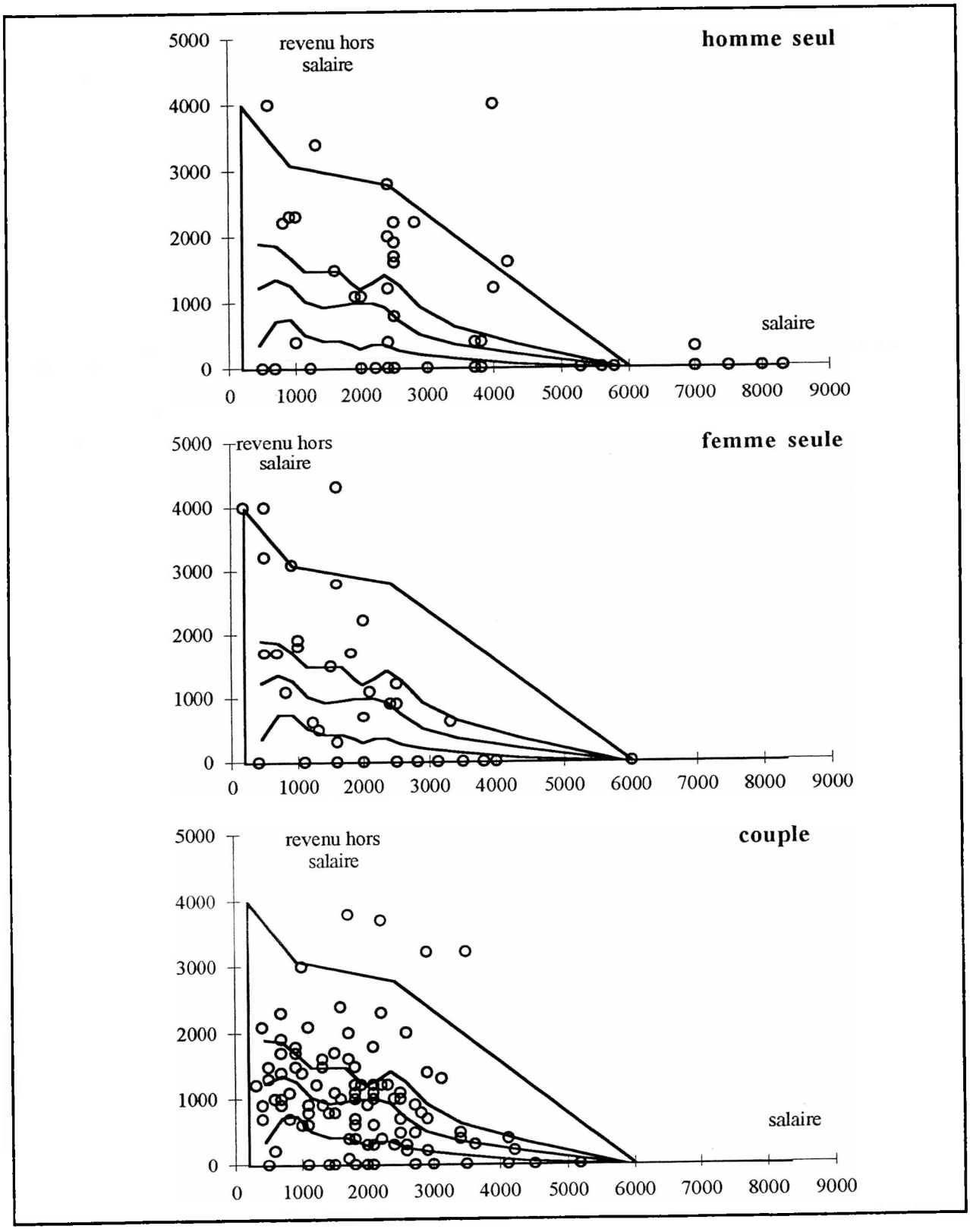
Pour que la comparaison soit possible, l'échelle utilisée est la même sur les différents graphiques. Afin de ne pas introduire de distorsion visuelle, les points seront également représentés de la même façon.

Comme précédemment, nous pourrions visualiser l'éventuelle relation grossière entre les deux variables observées lors du deuxième chapitre. Lorsque cela ne se fait pas au détriment de la lisibilité du graphique, on peut également sur chacun d'eux, tracer la box plot à deux dimensions. La box plot est celle calculée sur l'ensemble des individus, et non pour chacune des modalités de la variable. Il s'agit du même tracé d'une représentation à l'autre.

On étudie alors selon les modalités de la variable qualitative, la répartition des individus dans et autour de cette box plot.

Graphique n°44

Représentation du revenu hors salaire en fonction du salaire selon la situation familiale



Le graphique permet d'étudier la répartition des points d'une part à l'intérieur, d'autre part à l'extérieur du polygone adjacent, selon les modalités de la variable qualitative. Sa lecture se fait par la comparaison des différents graphiques et notamment celle :

-des individus extérieurs et adjacents. On étudie leur nombre et leur position d'un graphique à l'autre ;
-des individus proches de la relation schématisée par les tracés médians et quartiles. On essaie de déterminer si la relation entre les deux variables est plus ou moins bien respectée selon les modalités de la variable qualitative.

L'interprétation des points situés dans l'intervalle interquartile est cependant délicate. Elle se heurte en effet à deux problèmes. D'une part les points se superposent pour ces valeurs du salaire et du revenu puisque c'est là qu'ils sont en général les plus fréquents, d'autre part la différence de fréquence des modalités de la variable peut rendre la comparaison ambiguë. Dans l'exemple précédent, il semble ainsi que les couples respectent mieux la relation entre le revenu hors salaire et le salaire, mais cette impression provient en partie du fait que les couples sont beaucoup plus nombreux que les hommes et que les femmes seuls.

II-1-2 Influence de la variable qualitative sur la densité de deux variables quantitatives

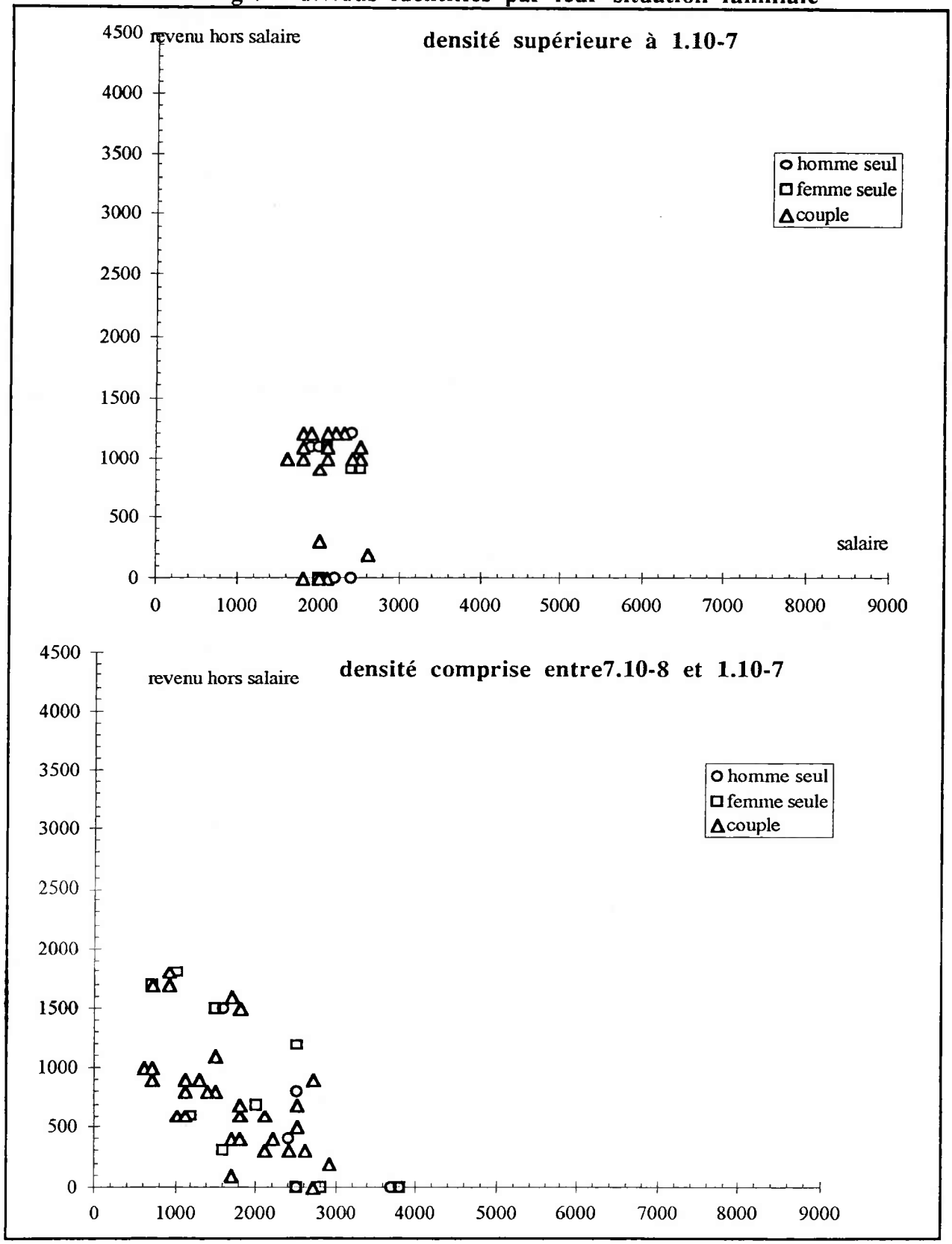
Le tracé de densité construit par la méthode d'affûtage sur deux variables quantitatives permettait de voir comment évoluait la concentration des points en fonction des valeurs des deux variables. Effectuer ce tracé en identifiant les points par les modalités d'une variable qualitative est donc un moyen de mettre en évidence l'influence de la variable qualitative sur la répartition des individus.

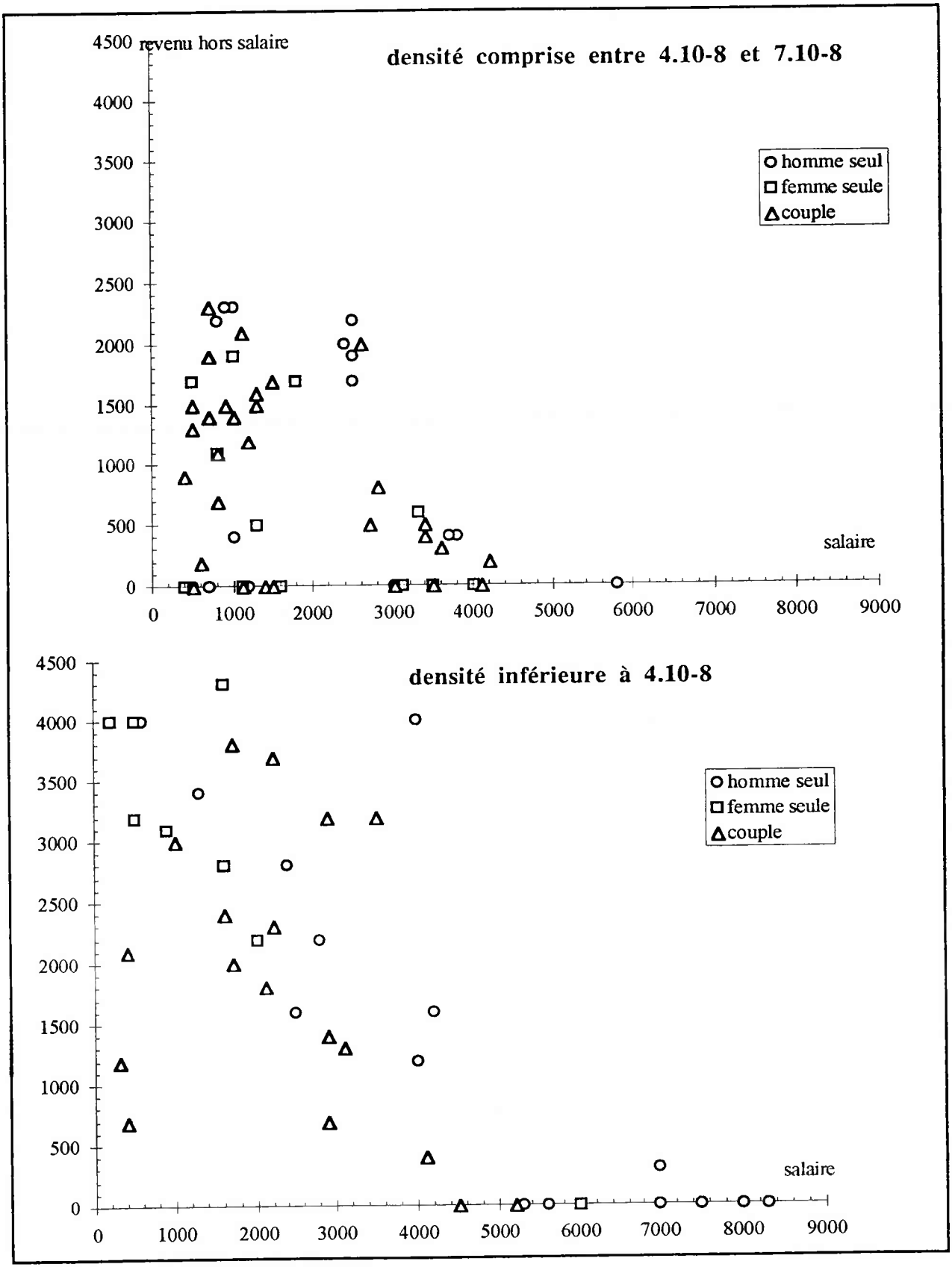
Dans un souci de lisibilité, les graphiques que nous construirons ne représenterons pas les points pour lesquels la densité est supérieure à un seuil fixé, mais ceux pour lesquels elle est comprise dans un intervalle.

Le graphique suivant montre les tracés obtenus pour le revenu hors salaire en fonction du salaire, avec la situation familiale comme identifiant. Les valeurs seuils et la fonction de densité estimée sont les mêmes que celles utilisées dans le deuxième chapitre.

graphique n°45

Densité du revenu hors salaire et du salaire représentée par la méthode d'affûtage. Individus identifiés par leur situation familiale




densité comprise entre 4.10-8 et 7.10-8
revenu hors salaire
salaire
homme seul
femme seule
couple
0
500
1000
1500
2000
2500
3000
3500
4000
4500
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
densité inférieure à 4.10-8
salaire
homme seul
femme seule
couple
0
500
1000
1500
2000
2500
3000
3500
4000
4500
0
1000
2000
3000
4000
5000
6000
7000
8000
9000

Les graphiques sont d'autant plus lisibles que la densité est faible, car les points seront alors plus espacés.

On étudie la répartition des points sur chaque graphique selon la modalité qu'ils représentent et la façon dont les modalités se répartissent d'un graphique à l'autre.

Ce graphique présente l'avantage de montrer des points ayant des modalités différentes tout en ne conservant qu'une partie des observations sur chacun des tracés. La comparaison est ainsi facilitée. Il reste peu lisible pour les points à densité élevée. Il permet surtout l'identification des individus ayant une faible densité.

Pour des variables qualitatives très discriminantes, il peut permettre de mettre en évidence la façon dont les individus se répartissent selon la modalité de la variable qualitative, et donc l'influence de la variable qualitative sur la dispersion des individus.

II-1-3 Exemple d'analyse : le revenu hors salaire en fonction du salaire selon la situation familiale

On s'intéresse d'abord aux individus ne respectant pas la relation entre le revenu hors salaire, que ce soit celle grossière mise en évidence sur le graphique représentant le revenu hors salaire en fonction du salaire, ou celle plus fine montrée par la box plot à deux dimensions.

Nous avions vu lors de l'étude de la représentation du revenu hors salaire en fonction du salaire, les points étant symbolisés par la situation familiale, que les points ne respectant pas la relation grossière était d'une part les femmes seules pour les salaires les plus faibles, les couples pour les salaires moyens, les hommes pour les salaires les plus élevés.

Lorsqu'on étudie la répartition des points par rapport aux box plots à deux dimensions, on voit qu'effectivement, les points adjacents ou extérieurs pour les femmes sont ceux pour lesquels le salaire est faible, les points adjacents ou extérieurs pour les couples sont ceux pour lesquels le salaire est moyen. Le résultat est moins évident pour les hommes. Les points qui étaient au-dessus de la droite y=4000-x sur le tracé précédent apparaissent maintenant comme étant sur le polygone adjacent. Ce sont par contre les hommes qui ont les valeurs particulièrement élevées du salaire pour un revenu hors salaire nul. L'étude des tracés de densité apporte peu de renseignements supplémentaires.

Lorsque l'on étudie les points respectant la relation, et notamment ceux situés dans l'intervalle interquartile, on peut remarquer que ce sont les couples qui semblent déterminer cette relation. Ce sont d'ailleurs les mieux répartis. En effet, ils apparaissent sur tous les tracés de densité de façon à peu près égale, alors que les femmes notamment prennent des valeurs du revenu hors salaire et du salaire moins fréquentes.

La situation familiale n'est donc pas très discriminante sur la relation du revenu hors salaire en fonction du salaire. La norme est déterminée par les couples, les hommes la respectent, mais ont légèrement tendance à avoir un salaire plus élevé pour un revenu hors salaire plus faible. Au contraire, les femmes seule tendent à avoir un salaire plus faible pour un revenu hors salaire plus élevé.

II-2 Influence de deux variables qualitatives sur la liaison entre deux variables quantitatives

Au cours du paragraphe précédent, nous avons vu comment on peut expliquer la répartition des individus selon les valeurs de deux variables quantitatives par les modalités d'une variable qualitative. Le but de ce paragraphe est d'étendre les méthodes précédentes à l'utilisation simultanée de deux variables qualitatives.

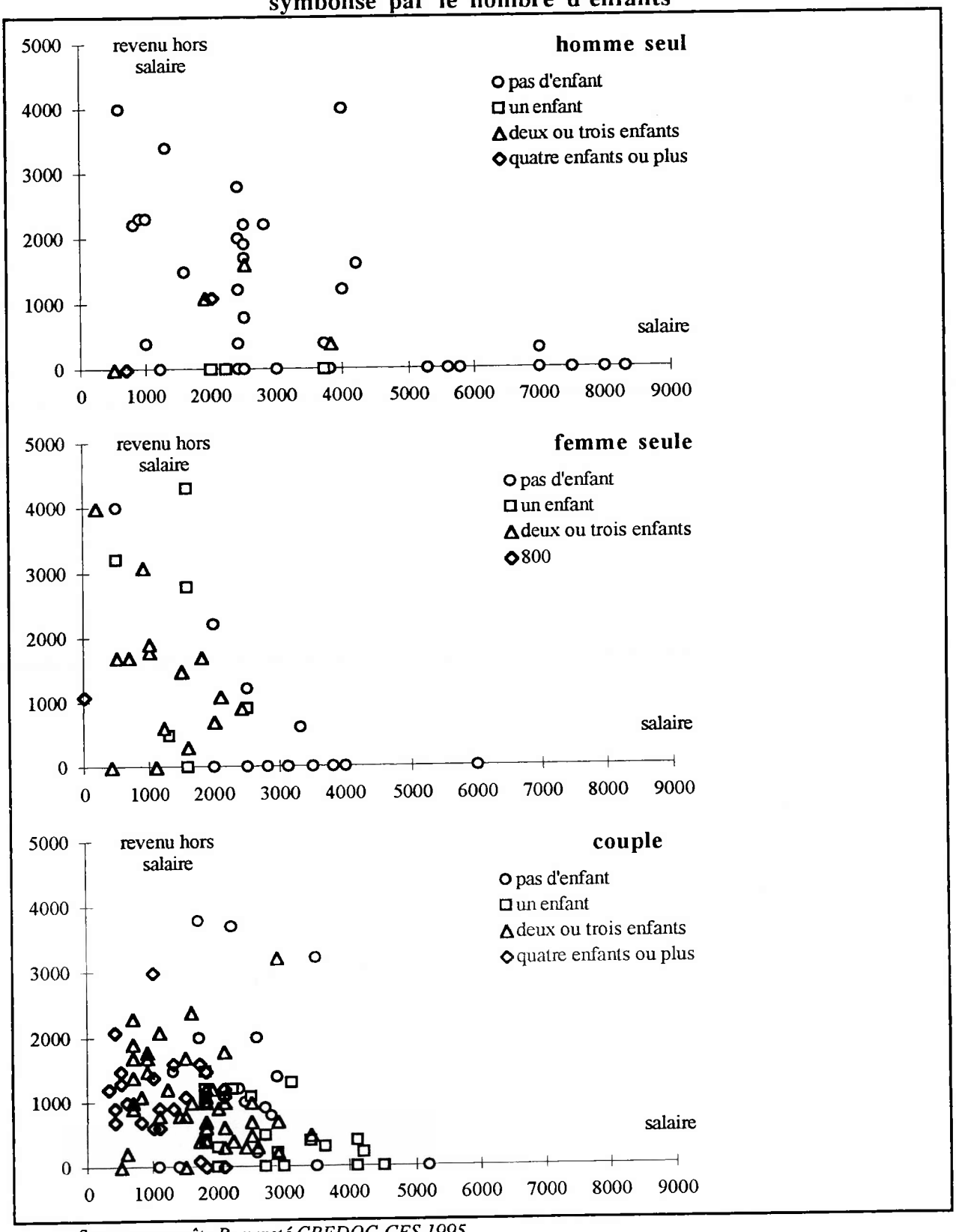
On va soit qualifier les individus selon les modalités d'une variable et les représenter modalité par modalité de l'autre variable, soit comparer des graphiques obtenus en croisant les modalités de deux variables.

II-2-1 Représentation des valeurs des deux variables quantitatives : répartition par modalité d'une variable qualitative et symbolisation par les modalités de l'autre

Le principe du graphique est de répartir les points selon les modalités d'une variable et de les qualifier en fonction des modalités de l'autre variable, puis de comparer les graphiques obtenus.

Graphique n°46

Revenu hors salaire en fonction du salaire, par la situation familiale et symbolisé par le nombre d'enfants



Source : enquête Pauvreté CREDOC-CES 1995

Pour chacune des modalités d'une variable qualitative, le graphique montre qui sont, en termes des modalités de l'autre variable qualitative, les individus qui respectent le mieux ou qui s'écartent de la relation entre les deux variables quantitatives.

On appelle A et B les deux variables qualitatives et X et Y les deux variables quantitatives. On utilise les modalités de A pour répartir les observations, celles de B pour les qualifier. Le graphique précédent permet alors de répondre à la question : « pour chacune des modalités de A, qui sont, en termes de modalités de B, les individus qui s'éloignent des valeurs les plus fréquentes de (X,Y) ? Qui sont ceux qui restent le plus proches de ces valeurs ? ».

Les graphiques sont d'autant moins lisibles que le nombre d'observations est grand. Ils sont peu lisibles pour les valeurs proches du mode des deux variables quantitatives. Leur interprétation doit tenir compte du fait que le nombre d'observations par modalité de chacune des deux variables qualitatives diffère en général.

Ils sont d'autre part asymétriques dans le traitement des variables qualitatives : les modalités de l'une des variables sont choisies comme identifiants des individus, les modalités de l'autre comme clé de partition des observations. Il y a deux choix possibles, qui fournissent deux impressions visuelles différentes.

II-2-2 Représentation des valeurs des deux variables qualitatives : répartition des individus selon les modalités des deux variables quantitatives

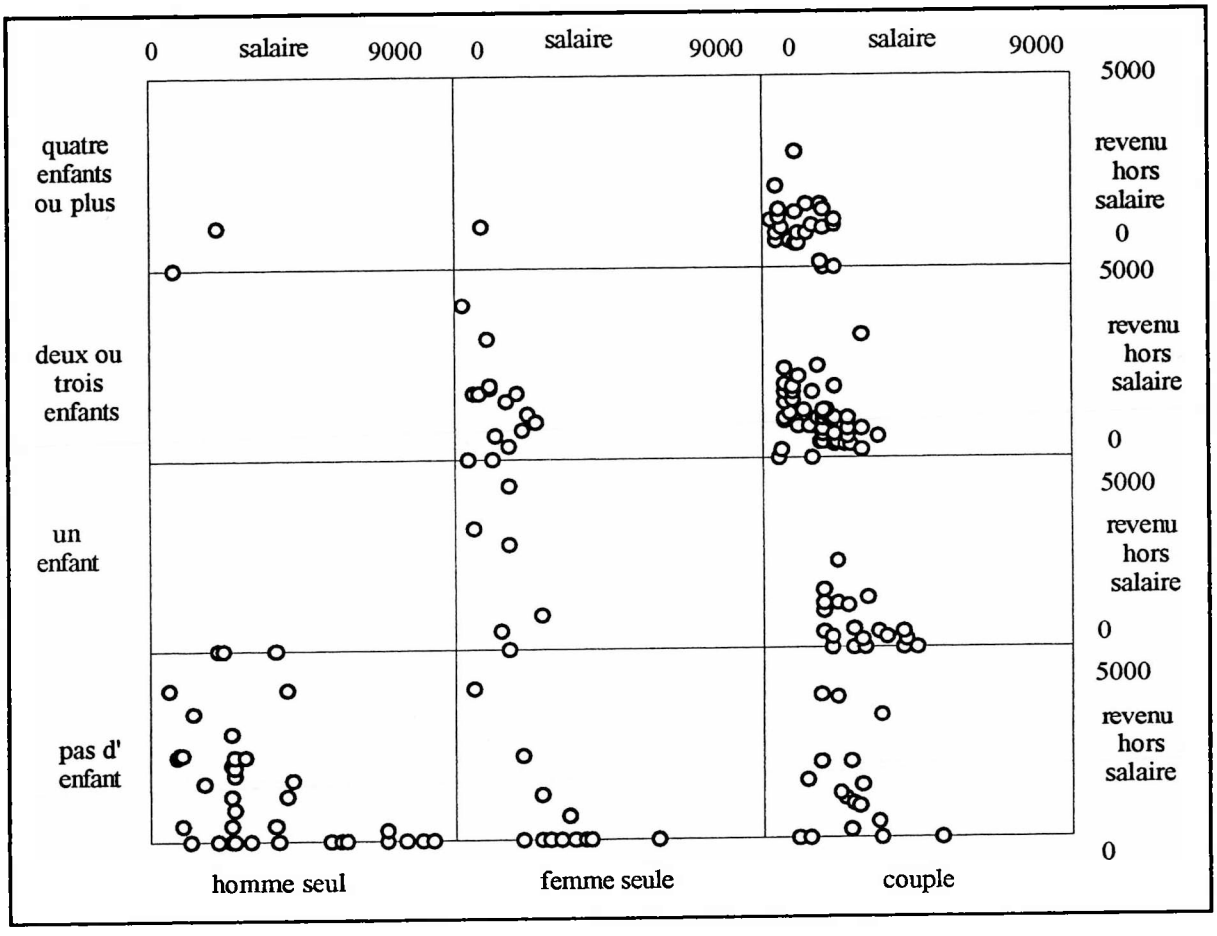
Le graphique utilisé ici est appelé représentation en fenêtres : les individus sont répartis en croisant les modalités des deux variables qualitatives. Si A a a modalités et B b modalités, on obtient a*b graphiques.

On les dispose alors en tableau : dans chaque colonne, on a les points qui respectent une modalité donnée de A, dans chaque ligne les points qui respectent une modalité donnée de B. Chacune des cases représente le graphique d'une des variables qualitatives en fonction de l'autre, les points respectant une modalité donnée de A et une modalité donnée de B.

A la fin de chaque ligne et de chaque colonne, on peut représenter la « distribution marginale » d'une des variables qualitatives, c'est-à-dire le graphique obtenu en ne tenant compte que des modalités d'une seule variable.

Graphique n°47

Représentation en fenêtre du revenu hors salaire en fonction du salaire selon la situation familiale et le nombre d'enfants



Source : enquête Pauvreté CREDOC-CES 1995

On n'a pas ici représenté les distributions marginales, qui ne sont pas lisibles.

Les deux variables qualitatives sont cette fois traitées de façon symétrique. Le graphique se lit en comparant chacune des cases à sa colonne et à sa ligne. Si les modalités de la variable A sont réparties en ligne, celles de la variables B en colonne, comparer en ligne les cases de la représentation permet d'étudier l'évolution du revenu hors salaire en fonction du salaire selon les modalités de A, les modalités de B étant fixées par ailleurs.

Le graphique permet ainsi de répondre à la question : «qui sont, en termes de A et pour chacune des modalités de B, les individus respectant ou s'éloignant de la relation entre X et Y ? » et « qui sont, en termes de modalités de B et pour chacune des modalités de A, les individus qui respectent ou qui s'éloignent de la relation entre X et Y ? ». On répond

finalement à la question : « Qui sont, en termes de A et de B, les individus qui s éloignent ou qui respectent la relation entre X et Y ?».

Ces réponses restent qualitatives, notamment parce que la taille de chacune des fenêtres ne permet pas de chiffrer les réponses apportées. Le graphique est également peu lisible dés que l'on a une forte densité de points ou que le nombre d'observations par case est élevé.

Une première vue de la représentation en fenêtre montre que quelle que soit la case ayant un nombre d'observations assez élevé, le fait que le revenu hors salaire diminue en fonction du salaire est vérifié.

Les individus qui respectent le moins la relation entre le revenu hors salaire et le salaire sont les hommes seuls sans enfants : les observations sont dispersées, notamment parce que ce sont eux qui ont des salaires élevés pour des revenus hors salaire faibles.

Une comparaison plus précise permet de détailler ces résultats :

les hommes seuls sont en général sans enfant. Les quelques hommes seuls avec enfants ont des salaires faibles et des revenus hors salaire nuls.

Les femmes seules avec enfants ont en général des salaires très faibles ou faibles, pour des revenus hors salaire se répartissant entre 0 et le maximum observé des revenus hors salaire. Les femmes seules sans enfants sont peu nombreuses. Elles ont, pour la plupart, un revenu hors salaire nul pour un salaire moyen à élevé.

Les couples avec enfants vérifient en général la relation entre le salaire et le revenu hors salaire. Le revenu hors salaire tend à être d'autant plus élevé et le salaire d'autant plus faible que le nombre d'enfants est grand. Lorsqu'ils n'ont pas d'enfants, ils ont un salaire très faible à élevé, pour un revenu hors salaire bien réparti.

Nous avions vu que les points les plus éloignés de la relation entre les revenus hors salaires et les salaire étaient les femmes pour les salaires très faibles, les couples pour les salaires moyens, les hommes pour les salaires élevés. Les deux représentations utilisées ici montrent qu'il s'agit plus précisément, pour les salaires moyens, des couples sans enfants, pour les salaires élevés des hommes sans enfants.

Nous avons ici présenté les graphiques permettant d'étudier deux variables quantitatives selon les modalités des variables qualitatives. On peut augmenter le nombre de variables qualitatives, et combiner les différentes méthodes utilisées pour faire apparaître leur rôle dans la répartition des points en fonction des valeurs des variables quantitatives.

L'étude demande alors de comparer de nombreux graphiques, utilisant des symboles différents, et représentant un nombre d'observations parfois trop faible pour être significatif.

De tels graphiques peuvent être utiles pour élaborer de premières hypothèses sur le rôle des variables qualitatives. Ils demandent cependant à être complétés par l'utilisation des méthodes de projection et d'analyse des données.

Conclusion

Ce cahier de recherche développe des méthodes graphiques qui demandent peu de calculs mathématiques sur les données au préalable. Elles s'appliquent directement aux données apurées ou après des transformations relativement simples. Ce sont en fait des méthodes qui sont à utiliser au cours de la première phase de l'analyse exploratoire, afin de se familiariser avec les données. En particulier, les regressions ainsi que toutes les méthodes d'analyse des données (analyse canonique, analyse discriminante, analyse factorielle) ont volontairement été exclues du champ de cette étude. Pour ces techniques aussi, la visualisation graphique apporte une meilleure lisibilité de l'information. Cependant, le recours aux graphiques pour l'analyse des résultats de ces méthodes est mieux connu et plus répandu.

Le deuxième parti pris de ce cahier est de s'intéresser surtout aux méthodes qui sont adaptées à des données d'enquête portant sur des effectifs restreints et contenant essentiellement des variables qualitatives. Afin de bien montrer l'interêt de la visualisation graphique au cours de l'analyse de ce type d'information, l'ensemble des méthodes ont été appliquées aux données d'une même enquête, l'enquête pauvreté réalisée en 1995 par le CREDOC pour le Conseil Economique et Social. L'étude des ressources des foyers bénéficiant d'un salaire selon plusieurs critères socio-démographiques a montré comment ces graphiques permettent de mener à bien une analyse complète.

Enfin, ce cahier développe des méthodes graphiques utiles à l'étape d'analyse. Mais le graphique peut aussi avoir un autre objectif : la communication. Cette deuxième utilisation a lieu dans une étape ultérieure à l'analyse. Il s'agit pour le statisticien de présenter par écrit ou oralement ses résultats.

Contrairement au graphique d'analyse, le graphique de communication n'est pas destiné au seul statisticien, mais à un public plus vaste, et notamment à des personnes moins expertes en statistique. Si le premier doit présenter les données en perdant le minimum d'informations, le deuxième doit au contraire être le plus simple possible afin d'être lisible par le plus grand nombre. Il faut toujours privilégier la clarté du graphique au détriment de la masse d'informations qu'il peut contenir. Un graphique clair limite les erreurs d'interprétation.

De plus, le graphique de communication n'a pas de public assuré. Il doit être suffisamment attrayant pour inciter les personnes à le déchiffrer. C'est pourquoi les critères

d'esthétisme et de variété sont plus importants pour ce type de graphique que pour un graphique d'analyse, qui est un instrument de travail. Cependant, la recherche de l'attrait du graphique ne doit pas occulter son objectif : aider le lecteur à bien comprendre, et si possible retenir, un résultat issu d'une analyse statistique.

Malgré l'enjeu des graphiques de communication, peu de chercheurs ont travaillé sur ce thème. C.F. SCMID et E.R. TUFTE[11] ont tous les deux essayé de sélectionner les critères à prendre en compte pour réaliser des "bons graphiques" de communication. D'autres recherches ont été menées par W.S. CLEVELAND et R. MACGILL[12] afin de déterminer l'influence des différents paramètres d'un graphique sur la perception du lecteur. Ils prouvent en particulier que marquer des écarts de proportion par des différences de surface (comme dans un camembert) ou par des différences de droite (comme dans les diagrammes en bâtons), n'est pas un choix de représentation neutre pour les futurs lecteurs. Ces différents travaux permettent d'isoler un certain nombre de consignes simples à suivre pour élaborer un "bon graphique de communication". Cependant, la subjectivité de la perception ne permet pas de déterminer des règles absolues. Certains aspects sont particulièrement difficiles à prendre en compte, comme par exemple le temps que le lecteur va garder l'information en mémoire selon la visualisation utilisée.

Parmi les graphiques détaillés dans ce cahier de recherche, un certain nombre peut aussi être utilisé comme graphique de communication. Cependant, la plupart sont un peu complexes et sont sujets à des erreurs d'interprétation. Il est donc préférable d'en limiter l'accès aux statisticiens.

---

[11] E. R. TUFTE, *The visual display of quantitative information*, 1983

C.F. SCMID, *Statisticals graphics*, 1983

[12] W.S. CLEVELAND et R. MCGILL, *Graphical perception : Theory, experimentation and application to the development of graphicals methods*, journal of the American Statistical Association, 1984

Bibliographie

BERTIN J. "Sémiologie graphique. Les diagrammes, les réseaux, les cartes" Mouthon Gauthier Villars, 1973

CHAMBERS J., CLEVELAND W.S., KLEINER B., TUKEY P.A. "Graphical methods for Data Analysis", Wadsworth & Brooks, 1983

CLEVELAND W.S, MC GILL R. "Graphical Perception : then visual decoding of qauntitive information on Graphical displays of data", Royal Statistical Society, 1987

FOX J., LONG S.J. " Modern methods of Data Analysis", Newsbury Park, Sage, 1990

HORBER E. "Visualisation semi-graphique et graphique" Journées de la méthodologie statistique, INSEE, 1995

TOMASINI M. "Etude de délais de la construction : aspects exploratoires" Journées de la méthodologie statistique, INSEE, 1995

TUFTE E.R. "Envisioning information", Cheshire Graphics Press, 1990

TUKEY J.W. "Exploratory Data Analysis" Addison-Wesley, Reading, Massachusetts, 1977

Annexes

Annexe N° 1 - Valeurs prises par les variables quantitatives ..... p. 3

Annexe N° 2 - Les valeurs de deux variables quantitatives par individu ..... p. 6

Annexe N° 3 - Les valeurs des variables quantitatives par modalité des qualitatives..... p. 8

Annexe N° 4 - Construction des représentations unidimensionnelles..... p. 11

Annexe N° 5 - Les estimations de densité ..... p. 13

Annexe N° 6 - Les Q-Plots..... p. 22

Annexe N° 7 - Les box plots..... p. 24

Annexe N° 8 - Box plot à deux dimensions ..... p. 27

Annexe N° 9 - Estimation du centre d'une variable quantitative en fonction d'une autre... p. 30

Organisation : les annexes N°1 à N°3 donnent les valeurs des variables utilisées pour construire les graphiques, les annexes N°4 à N°9 donnent les coordonnées des points des graphiques lorsqu'elles ne se déduisent pas immédiatement des valeurs des variables.

Annexe N°1

Valeurs prises par les variables quantitatives

Le revenu par unité de consommation : données triées par ordre croissant

400	2000	2700	3200	4000	5800
500	2100	2700	3200	4000	5800
500	2100	2700	3200	4000	5900
700	2100	2800	3200	4000	5900
800	2100	2800	3200	4000	6000
1100	2200	2800	3300	4100	6200
1100	2200	2800	3300	4100	6700
1100	2200	2800	3300	4100	7000
1200	2200	2800	3300	4100	7300
1400	2200	2800	3300	4100	7500
1400	2200	2800	3400	4200	8000
1400	2200	2900	3400	4200	8000
1500	2300	2900	3400	4200	8300
1500	2300	2900	3400	4300	
1600	2300	2900	3400	4400	
1600	2300	3000	3400	4400	
1600	2400	3000	3500	4400	
1700	2400	3000	3500	4500	
1700	2400	3000	3600	4500	
1700	2500	3000	3600	4500	
1800	2500	3000	3600	4500	
1800	2500	3000	3600	4500	
1800	2500	3000	3600	4500	
1800	2500	3000	3600	4600	
1800	2500	3100	3600	4600	
1900	2500	3100	3700	4700	
1900	2500	3100	3700	4700	
1900	2500	3100	3700	4700	
1900	2600	3100	3700	5100	
1900	2600	3100	3700	5200	
2000	2600	3100	3800	5200	
2000	2600	3100	3800	5200	
2000	2700	3100	3800	5300	
2000	2700	3200	3800	5500	
2000	2700	3200	3900	5600	
2000	2700	3200	4000	5800	

Source : enquête Pauvreté CREDOC - CES 1995

Le salaire par unité de consommation : données triées par ordre croissant

200	1000	1700	2100	2600	4200
300	1000	1700	2100	2700	4500
400	1000	1700	2100	2700	5200
400	1000	1800	2100	2700	5300
400	1100	1800	2100	2800	5600
400	1100	1800	2200	2800	5800
500	1100	1800	2200	2800	5800
500	1100	1800	2200	2900	6000
500	1100	1800	2200	2900	7000
500	1100	1800	2200	2900	7000
500	1100	1800	2300	2900	7500
500	1200	1800	2400	2900	8000
500	1200	1800	2400	3000	8300
600	1200	1800	2400	3000	
600	1300	1800	2400	3100	
600	1300	1800	2400	3100	
700	1300	1800	2400	3300	
700	1300	1800	2400	3400	
700	1300	1900	2400	3400	
700	1400	1900	2400	3400	
700	1400	2000	2500	3500	
700	1400	2000	2500	3500	
700	1500	2000	2500	3500	
700	1500	2000	2500	3600	
800	1500	2000	2500	3700	
800	1500	2000	2500	3700	
800	1500	2000	2500	3800	
800	1600	2000	2500	3800	
900	1600	2000	2500	3800	
900	1600	2000	2500	4000	
900	1600	2000	2500	4000	
900	1600	2000	2500	4000	
900	1600	2100	2500	4000	
1000	1600	2100	2500	4100	
1000	1700	2100	2600	4100	
1000	1700	2100	2600	4200	

Source : enquête Pauvreté CREDOC - CES 1995

Le revenu hors salaire par unité de consommation : données triées par ordre croissant

0	0	400	1000	1500	3000
0	0	500	1000	1500	3100
0	0	500	1000	1500	3200
0	0	500	1000	1600	3200
0	0	500	1000	1600	3200
0	0	600	1000	1600	3400
0	0	600	1100	1600	3700
0	0	600	1100	1700	3800
0	0	600	1100	1700	4000
0	0	600	1100	1700	4000
0	0	600	1100	1700	4000
0	0	700	1100	1700	4000
0	100	700	1100	1700	4300
0	200	700	1100	1700	
0	200	700	1100	1800	
0	200	700	1100	1800	
0	200	700	1200	1800	
0	200	800	1200	1900	
0	300	800	1200	1900	
0	300	800	1200	1900	
0	300	800	1200	2000	
0	300	800	1200	2000	
0	300	800	1200	2000	
0	300	800	1200	2100	
0	300	900	1200	2100	
0	400	900	1200	2200	
0	400	900	1200	2200	
0	400	900	1300	2200	
0	400	900	1300	2200	
0	400	900	1400	2300	
0	400	900	1400	2300	
0	400	900	1400	2300	
0	400	900	1500	2300	
0	400	1000	1500	2400	
0	400	1000	1500	2800	
0	400	1000	1500	2800	

Source : enquête Pauvreté CREDOC - CES 1995

Annexe N°2

Les valeurs de deux variables quantitatives sur chaque individu

Le revenu par unité de consommation et le salaire par unité de consommation : données triées sur le revenu

revenu	*salaire*	*revenu*	*salaire*	*revenu*	*salaire*	*revenu*	*salaire*	*revenu*	*salaire*	*revenu*	*salaire*
400	400	2000	800	2700	1600	3200	1100	4000	1000	5800	5800
500	500	2100	1700	2700	2000	3200	2900	4000	4000	5800	4200
500	500	2100	2100	2700	900	3200	900	4000	4000	5900	1600
700	700	2100	500	2800	1800	3200	2100	4000	3400	5900	2200
800	600	2100	1100	2800	2100	3200	2100	4000	3600	6000	6000
1100	1100	2200	1400	2800	2800	3300	1000	4100	2500	6200	2900
1100	400	2200	700	2800	1300	3300	2100	4100	3700	6700	3500
1100	1100	2200	1800	2800	1000	3300	2500	4100	1600	7000	7000
1200	1200	2200	1300	2800	2400	3300	1700	4100	900	7300	7000
1400	1000	2200	2200	2800	1800	3300	1500	4100	4100	7500	7500
1400	1400	2200	1800	2800	1800	3400	1800	4200	2500	8000	4000
1400	400	2200	1800	2900	2600	3400	2500	4200	2000	8000	8000
1500	800	2300	1500	2900	1000	3400	1800	4200	200	8300	8300
1500	1500	2300	1400	2900	2000	3400	2400	4300	3800		
1600	300	2300	2000	2900	1300	3400	2400	4400	2400		
1600	700	2300	500	3000	1900	3400	2200	4400	1600		
1600	1600	2400	1800	3000	2600	3500	3500	4400	2900		
1700	1000	2400	2400	3000	3000	3500	3500	4500	3100		
1700	700	2400	2400	3000	800	3600	1800	4500	500		
1700	600	2500	1000	3000	1800	3600	2500	4500	2500		
1800	1300	2500	2500	3000	1500	3600	2700	4500	4200		
1800	1800	2500	2500	3000	3000	3600	2800	4500	4100		
1800	1700	2500	700	3000	1800	3600	2300	4500	4500		
1800	1100	2500	700	3000	2000	3600	2500	4600	600		
1800	500	2500	900	3100	3100	3600	2400	4600	2200		
1900	1100	2500	1200	3100	2500	3700	500	4700	2500		
1900	1200	2500	2500	3100	2100	3700	3700	4700	2600		
1900	1100	2500	2100	3100	1600	3700	2900	4700	1300		
1900	1600	2600	1800	3100	2900	3700	1700	5100	2800		
1900	800	2600	2200	3100	1900	3700	2500	5200	5200		
2000	2000	2600	700	3100	2000	3800	3800	5200	4000		
2000	2000	2600	400	3100	1800	3800	3800	5200	2400		
2000	2000	2700	1500	3100	700	3800	3400	5300	5300		
2000	2000	2700	2700	3200	2500	3800	3400	5500	1700		
2000	2000	2700	900	3200	2700	3900	3300	5600	5600		
2000	2000	2700	2400	3200	2100	4000	2100	5800	5800		

Source : enquête Pauvreté CREDOC - CES 1995

Le salaire par unité de consommation et le revenu hors salaire par unité de consommation : données triées sur le salaire.

salaire	*revenu hors salaire*	*salaire*	*revenu hors salaire*	*salaire*	*revenu hors salaire*	*salaire*	*revenu hors salaire*	*salaire*	*revenu hors salaire*	*salaire*	*revenu hors salaire*
200	4000	1000	1400	1700	1600	2100	1200	2600	200	4200	200
300	1200	1000	1900	1700	400	2100	1000	2700	500	4500	0
400	2100	1000	1800	1700	100	2100	1100	2700	900	5200	0
400	700	1000	600	1800	1700	2100	1100	2700	0	5300	0
400	900	1100	2100	1800	1500	2100	1100	2800	2200	5600	0
400	0	1100	0	1800	1500	2200	3700	2800	800	5800	0
500	3200	1100	0	1800	600	2200	2300	2800	0	5800	0
500	4000	1100	900	1800	700	2200	400	2900	3200	6000	0
500	1700	1100	600	1800	400	2200	0	2900	1400	7000	0
500	1300	1100	800	1800	400	2200	1200	2900	700	7000	300
500	1500	1100	800	1800	400	2300	1200	2900	200	7500	0
500	0	1200	1200	1800	0	2400	2800	2900	200	8000	0
500	0	1200	0	1800	1200	2400	2000	3000	0	8300	0
600	4000	1200	600	1800	1200	2400	400	3000	0		
600	200	1300	3400	1800	1000	2400	300	3100	1300		
600	1000	1300	1600	1800	1000	2400	0	3100	0		
700	2300	1300	1500	1800	1000	2400	0	3300	600		
700	0	1300	500	1800	1100	2400	1200	3400	500		
700	1400	1300	900	1900	1200	2400	900	3400	400		
700	1900	1400	0	1900	1100	2400	1000	3400	400		
700	900	1400	800	2000	2200	2500	1600	3500	3200		
700	1000	1400	800	2000	700	2500	2200	3500	0		
700	1700	1500	0	2000	300	2500	1700	3500	0		
700	1700	1500	1700	2000	0	2500	1900	3600	300		
800	2200	1500	1500	2000	0	2500	500	3700	400		
800	700	1500	1100	2000	0	2500	700	3700	0		
800	1100	1500	800	2000	0	2500	1200	3800	400		
800	1100	1600	4300	2000	0	2500	800	3800	0		
900	3100	1600	2800	2000	0	2500	0	3800	0		
900	2300	1600	2400	2000	900	2500	0	4000	4000		
900	1500	1600	0	2000	900	2500	0	4000	1200		
900	1800	1600	300	2000	1100	2500	1100	4000	0		
900	1700	1600	1500	2100	1800	2500	900	4000	0		
1000	3000	1600	1000	2100	600	2500	1000	4100	400		
1000	400	1700	3800	2100	300	2600	2000	4100	0		
1000	2300	1700	2000	2100	0	2600	300	4200	1600		

Source : enquête Pauvreté CREDOC - CES 1995

Annexe N°3

Les valeurs des variables quantitatives par modalité des variables qualitatives

Le salaire en fonction de la situation familiale : données triées par modalité sur le salaire

homme seul				femme seule		couple					
salaire	*revenu hors salaire*	*salaire*	*revenu hors salaire*	*salaire*	*revenu hors salaire*	*salaire*	*revenu hors salaire*	*salaire*	*revenu hors salaire*	*salaire*	*revenu hors salaire*
500	0	3800	400	200	4000	300	1200	1500	0	2100	1800
600	4000	4000	1200	400	0	400	700	1500	800	2200	400
700	0	4000	4000	500	1700	400	900	1500	1100	2200	1200
800	2200	4200	1600	500	3200	400	2100	1500	1700	2200	2300
900	2300	5300	0	500	4000	500	0	1600	1000	2200	3700
1000	400	5600	0	700	1700	500	1300	1600	2400	2300	1200
1000	2300	5800	0	800	1100	500	1500	1700	100	2400	300
1200	0	5800	0	900	3100	600	200	1700	400	2400	1000
1300	3400	7000	0	1000	1800	600	1000	1700	1600	2500	500
1600	1500	7000	300	1000	1900	700	900	1700	2000	2500	700
1900	1100	7500	0	1100	0	700	1000	1700	3800	2500	1100
2000	0	8000	0	1200	600	700	1400	1800	0	2500	1000
2000	0	8300	0	1300	500	700	1700	1800	400	2600	200
2000	0			1500	1500	700	1900	1800	400	2600	300
2000	0			1600	0	700	2300	1800	400	2600	2000
2000	1100			1600	300	800	700	1800	600	2700	0
2200	0			1600	2800	800	1100	1800	700	2700	500
2400	0			1600	4300	900	1500	1800	1000	2700	900
2400	0			1800	1700	900	1700	1800	1000	2800	800
2400	400			2000	0	900	1800	1800	1000	2900	200
2400	1200			2000	700	1000	600	1800	1100	2900	200
2400	2000			2000	2200	1000	1400	1800	1200	2900	700
2400	2800			2100	1100	1000	3000	1800	1200	2900	1400
2500	0			2100	1100	1100	0	1800	1500	2900	3200
2500	0			2400	900	1100	600	1800	1500	3000	0
2500	800			2500	0	1100	800	1900	1200	3100	1300
2500	1600			2500	900	1100	800	2000	0	3400	400
2500	1700			2500	1200	1100	900	2000	300	3400	400
2500	1900			2800	0	1100	2100	2000	900	3400	500
2500	2200			3100	0	1200	1200	2000	900	3500	0
2800	2200			3300	600	1300	900	2100	0	3500	3200
3000	0			3500	0	1300	1500	2100	300	3600	300
3700	0			3800	0	1300	1600	2100	600	4100	0
3700	400			4000	0	1400	0	2100	1000	4100	400
3800	0			4000	0	1400	800	2100	1100	4200	200
				6000	0	1400	800	2100	1200	4500	0
										5200	0

Source : enquête Pauvreté CREDOC - CES 1995

Le salaire et le revenu hors salaire en fonction du nombre d'enfants : données triées par modalité sur le salaire

pas d'enfant						un enfant					
homme seul		femme seule		couple		homme seul		femme seule		couple	
salaire	*revenu hsal*	*salaire*	*revenu hsal*	*salaire*	*revenu hsal*	*salaire*	*revenu hsal*	*salaire*	*revenu hsal*	*salaire*	*revenu hsal*
1200	0	2000	0	1100	0	2000	0	1600	0	1100	0
1000	400	2500	0	1400	0	2200	0	1300	500	1400	0
2000	0	2800	0	1300	1500	3700	0	2500	900	1300	1500
2000	0	3100	0	2600	200			500	3200	2600	200
2000	0	3500	0	2400	1000			1600	2800	2400	1000
2400	0	2500	1200	3500	0			1600	4300	3500	0
2400	0	3800	0	2700	900					2700	900
2500	0	3300	600	2800	800					2800	800
2500	0	4000	0	2300	1200					2300	1200
2400	400	4000	0	1700	2000					1700	2000
3000	0	2000	2200	3400	400					3400	400
800	2200	500	4000	2900	1400					2900	1400
1600	1500	6000	0	2600	2000					2600	2000
900	2300			5200	0					5200	0
1000	2300			1700	3800					1700	3800
2500	800			2200	3700					2200	3700
2400	1200			3500	3200					3500	3200
3800	0										
3700	400										
2500	1700										
2400	2000										
2500	1900										
600	4000										
2500	2200										
1300	3400										
2800	2200										
4000	1200										
2400	2800										
5300	0										
5600	0										
5800	0										
5800	0										
4200	1600										
7000	0										
7000	300										
7500	0										
4000	4000										
8000	0										
8300	0										

Source : enquête Pauvreté CREDOC - CES 1995

pas d'enfant						un enfant					
homme seul		femme seule		couple		homme seul		femme seule		couple	
salaire	*revenu hors salaire*	*salaire*	*revenu hors salaire*	*salaire*	*revenu hors salaire*	*salaire*	*revenu hors salaire*	*salaire*	*revenu hors salaire*	*salaire*	*revenu hors salaire*
500	0	400	0	2000	0	700	0	800	1100	400	700
1900	1100	1100	0	1800	400	2000	1100			400	900
2500	1600	1200	600	1800	400					800	700
3800	400	1600	300	2000	300					300	1200
		500	1700	2700	0					1000	600
		700	1700	1800	1000					600	1000
		2000	700	3000	0					1800	0
		1000	1800	1800	1200					1700	100
		1000	1900	2700	500					1100	600
		1500	1500	2900	200					500	1300
		2100	1100	1800	1500					2100	0
		2100	1100	2200	1200					500	1500
		2400	900	2500	1100					1100	900
		1800	1700	3400	400					1300	900
		900	3100	3600	300					1000	1400
		200	4000	4100	0					400	2100
				3100	1300					1500	1100
				4200	200					1800	1000
				4100	400					1300	1600
				4500	0					2100	1100
				2200	2300					2100	1200
										1700	1600
										1800	1500
										1000	3000

Source : enquête Pauvreté CREDOC - CES 1995

Annexe N° 4
Construction de la représentation unidimensionnelle : décalage vertical des points selon un processus aléatoire

Le tirage aléatoire a été fait à partir de SAS, selon une loi uniforme. Le premier nombre a été généré d'après l'horloge de l'ordinateur. Le choix de la loi pour effectuer le tirage n'a pas d'importance, on aurait pu choisir un autre processus.

Le revenu par unité de consommation et l'ordonnée aléatoire de chaque valeur - Données triées sur le revenu par ordre croissant (graphique N° 4)

revenu par uc	*décalage*	*revenu par uc*	*décalage*	*revenu par uc*	*décalage*	*revenu par uc*	*décalage*	*revenu par uc*	*décalage*	*revenu par uc*	*décalage*
400	129	2000	21	2700	26	3200	172	3800	105	5200	135
500	110	2000	84	2700	150	3200	0	3800	77	5300	29
500	183	2100	90	2700	168	3200	57	3900	192	5500	42
700	139	2100	82	2700	8	3200	22	4000	149	5600	151
800	9	2100	65	2700	27	3200	52	4000	58	5800	41
1100	120	2100	176	2800	177	3200	80	4000	51	5800	136
1100	14	2200	152	2800	53	3200	50	4000	88	5800	17
1100	53	2200	170	2800	142	3200	123	4000	171	5900	163
1200	98	2200	173	2800	139	3300	120	4000	130	5900	152
1400	34	2200	39	2800	21	3300	139	4100	49	6000	52
1400	56	2200	47	2800	105	3300	119	4100	130	6200	108
1400	8	2200	5	2800	77	3300	41	4100	120	6700	160
1500	69	2200	84	2800	137	3300	64	4100	168	7000	107
1500	3	2300	35	2900	96	3400	125	4100	87	7300	4
1600	45	2300	21	2900	117	3400	38	4200	41	7500	102
1600	102	2300	67	2900	179	3400	116	4200	117	8000	140
1600	134	2300	1	2900	159	3400	192	4200	55	8000	82
1700	51	2400	20	3000	60	3400	176	4300	8	8300	41
1700	128	2400	74	3000	76	3400	89	4400	190	8000	82
1700	40	2400	10	3000	126	3500	157	4400	1	8300	41
1800	114	2500	35	3000	93	3500	77	4400	191		
1800	185	2500	61	3000	159	3600	163	4500	57		
1800	154	2500	11	3000	90	3600	27	4500	173		
1800	44	2500	24	3000	2	3600	127	4500	124		
1800	181	2500	160	3000	77	3600	158	4500	1		
1900	117	2500	180	3000	90	3600	131	4500	33		
1900	127	2500	181	3100	185	3600	179	4500	10		
1900	35	2500	32	3100	158	3600	90	4600	138		
1900	124	2500	183	3100	105	3700	127	4600	46		
1900	56	2600	145	3100	45	3700	42	4700	153		
2000	36	2600	94	3100	18	3700	21	4700	142		
2000	85	2600	63	3100	152	3700	19	4700	123		
2000	173	2600	80	3100	13	3700	45	5100	97		
2000	40	2700	149	3100	128	3800	42	5200	87		
2000	152	2700	128	3100	141	3800	151	5200	126		

Source : enquête Pauvreté CREDOC - CES 1995

Le revenu par unité de consommation et l'ordonnée aléatoire de chaque valeur : tirage selon une loi uniforme. Données triées sur le revenu par ordre croissant (Graphique N° 4)

revenu	*décalage*	*revenu*	*décalage*	*revenu*	*décalage*	*revenu*	*décalage*	*revenu*	*décalage*	*revenu*	*décalage*
800	129	3800	84	5700	168	7200	22	8800	58	11200	136
1000	110	3800	90	5700	8	7300	52	8850	51	11449	17
1200	183	3852	82	5800	27	7300	80	8907	88	11500	163
1400	139	3900	65	5800	177	7310	50	8973	171	11530	152
1700	9	4000	176	5800	53	7400	123	9000	130	11640	52
2000	120	4024	152	5800	142	7400	120	9000	49	12090	108
2000	14	4055	170	5850	139	7500	139	9023	130	12468	160
2000	53	4150	173	5900	21	7500	119	9100	120	12500	107
2000	98	4200	39	6000	105	7500	41	9100	168	12600	4
2002	34	4280	47	6000	77	7544	64	9200	87	12911	102
2407	56	4400	5	6000	137	7590	125	9300	41	15200	140
2450	8	4500	84	6000	96	7600	38	9400	117	16869	82
2478	69	4559	35	6100	117	7685	116	9500	55	20300	41
2500	3	4600	21	6100	179	7688	192	9500	8		
2500	45	4600	67	6200	159	7700	176	9744	190		
2500	102	4750	1	6200	60	7800	89	9900	1		
2500	134	4794	20	6282	76	8000	157	9933	191		
2500	51	4798	74	6325	126	8000	77	9950	57		
2800	128	4880	10	6371	93	8000	163	9997	173		
2800	40	4900	35	6500	159	8000	27	9997	124		
2800	114	4950	61	6500	90	8000	127	10000	1		
2900	185	5000	11	6516	2	8000	158	10085	33		
3000	154	5029	24	6550	77	8000	131	10100	10		
3090	44	5112	160	6600	90	8000	179	10100	138		
3111	181	5164	180	6600	185	8100	90	10192	46		
3119	117	5200	181	6600	158	8210	127	10200	153		
3200	127	5230	32	6645	105	8300	42	10200	142		
3266	35	5286	183	6800	45	8400	21	10228	123		
3300	124	5300	145	6800	18	8425	19	10270	97		
3300	56	5394	94	7000	152	8475	45	10370	87		
3400	36	5400	63	7000	13	8500	42	10400	126		
3500	85	5500	80	7050	128	8500	151	10700	135		
3500	173	5600	149	7124	141	8500	105	10700	29		
3662	40	5600	128	7200	172	8600	77	10797	42		
3776	152	5600	26	7200	0	8700	192	10900	151		
3800	21	5661	150	7200	57	8800	149	11100	41		

Source : enquête Pauvreté CREDOC - CES 1995

Annexe N°5

Les estimations de densité

La densité a été calculée à l'aide du logiciel SAS.

Soit y, la variable considérée, p le nombre d'observations. On appelle yi la valeur de la variable sur l'individu i.

Afin de pouvoir compter, pour chacune des valeurs de la variable, le nombre d'observations dont la valeur reste dans l'intervalle choisi, nous avons utilisé les matrices suivantes :

-la matrice (a_{ij}) de dimension p telle que : $\left| \begin{array}{l} a_{ij}=0 \text{ si } i \neq j \\ a_{ii}=yi \end{array} \right.$

-la matrice (bij) de dimension p dont les p lignes sont les mêmes, telles que $b_{ij}=yj$, pour tout i variant de 1 à p.

Chacune des valeurs de y a ainsi pu être comparée à toutes les autres valeurs.

Pour le calcul de la densité d'une variable à deux dimensions, on utilise les mêmes matrices que précédemment, en les construisant séparément à partir de chacune des variables du vecteur.

Densité du revenu par unité de consommation, estimée grossièrement : h=800 (graphique N°9)

$$\hat{f}(y) = \frac{\text{nbre d' obs ds } [y - 400; y + 400[}{800 * 193}$$

revenu	*densité*	*revenu*	*densité*	*revenu*	*densité*	*revenu*	*densité*	*revenu*	*densité*	*revenu*	*densité*
400	25,9E-6	2000	246,1E-6	2700	310,9E-6	3200	330,3E-6	4000	207,3E-6	5800	51,8E-6
500	32,4E-6	2100	246,1E-6	2700	310,9E-6	3200	330,3E-6	4000	207,3E-6	5800	51,8E-6
500	32,4E-6	2100	246,1E-6	2700	310,9E-6	3200	330,3E-6	4000	207,3E-6	5900	58,3E-6
700	32,4E-6	2100	246,1E-6	2800	343,3E-6	3200	330,3E-6	4000	207,3E-6	5900	58,3E-6
800	51,8E-6	2100	246,1E-6	2800	343,3E-6	3200	330,3E-6	4000	207,3E-6	6000	51,8E-6
1100	58,3E-6	2200	285,0E-6	2800	343,3E-6	3300	323,8E-6	4100	181,3E-6	6200	45,3E-6
1100	58,3E-6	2200	285,0E-6	2800	343,3E-6	3300	323,8E-6	4100	181,3E-6	6700	13,0E-6
1100	58,3E-6	2200	285,0E-6	2800	343,3E-6	3300	323,8E-6	4100	181,3E-6	7000	19,4E-6
1200	64,8E-6	2200	285,0E-6	2800	343,3E-6	3300	323,8E-6	4100	181,3E-6	7300	19,4E-6
1400	97,2E-6	2200	285,0E-6	2800	343,3E-6	3300	323,8E-6	4100	181,3E-6	7500	13,0E-6
1400	97,2E-6	2200	285,0E-6	2800	343,3E-6	3400	330,3E-6	4200	187,8E-6	8000	19,4E-6
1400	97,2E-6	2200	285,0E-6	2900	375,6E-6	3400	330,3E-6	4200	187,8E-6	8000	19,4E-6
1500	129,5E-6	2300	278,5E-6	2900	375,6E-6	3400	330,3E-6	4200	187,8E-6	8300	19,4E-6
1500	129,5E-6	2300	278,5E-6	2900	375,6E-6	3400	330,3E-6	4300	174,9E-6		
1600	142,5E-6	2300	278,5E-6	2900	375,6E-6	3400	330,3E-6	4400	187,8E-6		
1600	142,5E-6	2300	278,5E-6	3000	349,7E-6	3400	330,3E-6	4400	187,8E-6		
1600	142,5E-6	2400	291,5E-6	3000	349,7E-6	3500	297,9E-6	4400	187,8E-6		
1700	181,3E-6	2400	291,5E-6	3000	349,7E-6	3500	297,9E-6	4500	149,0E-6		
1700	181,3E-6	2400	291,5E-6	3000	349,7E-6	3600	246,1E-6	4500	149,0E-6		
1700	181,3E-6	2500	297,9E-6	3000	349,7E-6	3600	246,1E-6	4500	149,0E-6		
1800	207,3E-6	2500	297,9E-6	3000	349,7E-6	3600	246,1E-6	4500	149,0E-6		
1800	207,3E-6	2500	297,9E-6	3000	349,7E-6	3600	246,1E-6	4500	149,0E-6		
1800	207,3E-6	2500	297,9E-6	3000	349,7E-6	3600	246,1E-6	4500	149,0E-6		
1800	207,3E-6	2500	297,9E-6	3000	349,7E-6	3600	246,1E-6	4600	116,6E-6		
1800	207,3E-6	2500	297,9E-6	3100	362,7E-6	3600	246,1E-6	4600	116,6E-6		
1900	233,2E-6	2500	297,9E-6	3100	362,7E-6	3700	233,2E-6	4700	97,2E-6		
1900	233,2E-6	2500	297,9E-6	3100	362,7E-6	3700	233,2E-6	4700	97,2E-6		
1900	233,2E-6	2500	297,9E-6	3100	362,7E-6	3700	233,2E-6	4700	97,2E-6		
1900	233,2E-6	2600	297,9E-6	3100	362,7E-6	3700	233,2E-6	5100	51,8E-6		
1900	233,2E-6	2600	297,9E-6	3100	362,7E-6	3700	233,2E-6	5200	38,9E-6		
2000	246,1E-6	2600	297,9E-6	3100	362,7E-6	3800	233,2E-6	5200	38,9E-6		
2000	246,1E-6	2600	297,9E-6	3100	362,7E-6	3800	233,2E-6	5200	38,9E-6		
2000	246,1E-6	2700	310,9E-6	3100	362,7E-6	3800	233,2E-6	5300	45,3E-6		
2000	246,1E-6	2700	310,9E-6	3200	330,3E-6	3800	233,2E-6	5500	64,8E-6		
2000	246,1E-6	2700	310,9E-6	3200	330,3E-6	3900	213,7E-6	5600	71,2E-6		
2000	246,1E-6	2700	310,9E-6	3200	330,3E-6	4000	207,3E-6	5800	51,8E-6		

Source : enquête Pauvreté CREDOC - CES 1995

Densité du revenu par unité de consommation, estimation lissée : h=800 (graphique N° 10)

$$\hat{f}(y) = \frac{1}{193 * 800} \sum_{i=1}^{193} K(\frac{y - y_i}{800}), \text{ avec}$$

$K(u)=1+\cos 2\pi u$ si $-0,5 \leq u < 0,5$

$K(u)=0$ sinon

revenu	*densité*	*revenu*	*densité*	*revenu*	*densité*	*revenu*	*densité*	*revenu*	*densité*	*revenu*	*densité*
400	3,70E-05	2000	2,81E-04	2700	3,30E-04	3200	3,67E-04	4000	2,01E-04	5800	7,58E-05
500	4,53E-05	2100	2,80E-04	2700	3,30E-04	3200	3,67E-04	4000	2,01E-04	5800	7,58E-05
500	4,53E-05	2100	2,80E-04	2700	3,30E-04	3200	3,67E-04	4000	2,01E-04	5900	7,39E-05
700	3,89E-05	2100	2,80E-04	2800	3,44E-04	3200	3,67E-04	4000	2,01E-04	5900	7,39E-05
800	3,35E-05	2100	2,80E-04	2800	3,44E-04	3200	3,67E-04	4000	2,01E-04	6000	6,10E-05
1100	5,75E-05	2200	2,70E-04	2800	3,44E-04	3300	3,21E-04	4100	1,91E-04	6200	2,32E-05
1100	5,75E-05	2200	2,70E-04	2800	3,44E-04	3300	3,21E-04	4100	1,91E-04	6700	1,49E-05
1100	5,75E-05	2200	2,70E-04	2800	3,44E-04	3300	3,21E-04	4100	1,91E-04	7000	1,68E-05
1200	6,94E-05	2200	2,70E-04	2800	3,44E-04	3300	3,21E-04	4100	1,91E-04	7300	2,13E-05
1400	9,83E-05	2200	2,70E-04	2800	3,44E-04	3300	3,21E-04	4100	1,91E-04	7500	1,94E-05
1400	9,83E-05	2200	2,70E-04	2800	3,44E-04	3400	2,79E-04	4200	1,77E-04	8000	2,78E-05
1400	9,83E-05	2200	2,70E-04	2900	3,66E-04	3400	2,79E-04	4200	1,77E-04	8000	2,78E-05
1500	1,23E-04	2300	2,67E-04	2900	3,66E-04	3400	2,79E-04	4200	1,77E-04	8300	1,68E-05
1500	1,23E-04	2300	2,67E-04	2900	3,66E-04	3400	2,79E-04	4300	1,66E-04		
1600	1,55E-04	2300	2,67E-04	2900	3,66E-04	3400	2,79E-04	4400	1,64E-04		
1600	1,55E-04	2300	2,67E-04	3000	3,87E-04	3400	2,79E-04	4400	1,64E-04		
1600	1,55E-04	2400	2,75E-04	3000	3,87E-04	3500	2,57E-04	4400	1,64E-04		
1700	1,92E-04	2400	2,75E-04	3000	3,87E-04	3500	2,57E-04	4500	1,65E-04		
1700	1,92E-04	2400	2,75E-04	3000	3,87E-04	3600	2,44E-04	4500	1,65E-04		
1700	1,92E-04	2500	2,94E-04	3000	3,87E-04	3600	2,44E-04	4500	1,65E-04		
1800	2,29E-04	2500	2,94E-04	3000	3,87E-04	3600	2,44E-04	4500	1,65E-04		
1800	2,29E-04	2500	2,94E-04	3000	3,87E-04	3600	2,44E-04	4500	1,65E-04		
1800	2,29E-04	2500	2,94E-04	3000	3,87E-04	3600	2,44E-04	4500	1,65E-04		
1800	2,29E-04	2500	2,94E-04	3000	3,87E-04	3600	2,44E-04	4600	1,47E-04		
1800	2,29E-04	2500	2,94E-04	3100	3,89E-04	3600	2,44E-04	4600	1,47E-04		
1900	2,62E-04	2500	2,94E-04	3100	3,89E-04	3700	2,29E-04	4700	1,06E-04		
1900	2,62E-04	2500	2,94E-04	3100	3,89E-04	3700	2,29E-04	4700	1,06E-04		
1900	2,62E-04	2500	2,94E-04	3100	3,89E-04	3700	2,29E-04	4700	1,06E-04		
1900	2,62E-04	2600	3,15E-04	3100	3,89E-04	3700	2,29E-04	5100	5,26E-05		
1900	2,62E-04	2600	3,15E-04	3100	3,89E-04	3700	2,29E-04	5200	6,29E-05		
2000	2,81E-04	2600	3,15E-04	3100	3,89E-04	3800	2,16E-04	5200	6,29E-05		
2000	2,81E-04	2600	3,15E-04	3100	3,89E-04	3800	2,16E-04	5200	6,29E-05		
2000	2,81E-04	2700	3,30E-04	3100	3,89E-04	3800	2,16E-04	5300	6,10E-05		
2000	2,81E-04	2700	3,30E-04	3200	3,67E-04	3800	2,16E-04	5500	4,19E-05		
2000	2,81E-04	2700	3,30E-04	3200	3,67E-04	3900	2,07E-04	5600	4,91E-05		
2000	2,81E-04	2700	3,30E-04	3200	3,67E-04	4000	2,01E-04	5800	7,58E-05		

Source : enquête Pauvreté CREDOC - CES 1995

Densité du revenu par unité de consommation, estimation lissée : h=1600 (graphique N° 11)

$$\hat{f}(y) = \frac{1}{193 * 1600} \sum_{i=1}^{193} K\left(\frac{y - y_i}{1600}\right), \text{ avec}$$

$K(u)=1+\cos 2\pi u$ si $-0,5 \leq u < 0,5$

$K(u)=0$ sinon

revenu	*densité*	*revenu*	*densité*	*revenu*	*densité*	*revenu*	*densité*	*revenu*	*densité*	*revenu*	*densité*
400	2,74E-05	2000	2,49E-04	2700	3,30E-04	3200	3,31E-04	4000	2,00E-04	5800	5,58E-05
500	3,23E-05	2100	2,63E-04	2700	3,30E-04	3200	3,31E-04	4000	2,00E-04	5800	5,58E-05
500	3,23E-05	2100	2,63E-04	2700	3,30E-04	3200	3,31E-04	4000	2,00E-04	5900	5,18E-05
700	4,07E-05	2100	2,63E-04	2800	3,40E-04	3200	3,31E-04	4000	2,00E-04	5900	5,18E-05
800	4,49E-05	2100	2,63E-04	2800	3,40E-04	3200	3,31E-04	4000	2,00E-04	6000	4,68E-05
1100	6,55E-05	2200	2,74E-04	2800	3,40E-04	3300	3,15E-04	4100	1,91E-04	6200	3,39E-05
1100	6,55E-05	2200	2,74E-04	2800	3,40E-04	3300	3,15E-04	4100	1,91E-04	6700	1,42E-05
1100	6,55E-05	2200	2,74E-04	2800	3,40E-04	3300	3,15E-04	4100	1,91E-04	7000	1,74E-05
1200	7,81E-05	2200	2,74E-04	2800	3,40E-04	3300	3,15E-04	4100	1,91E-04	7300	1,79E-05
1400	1,16E-04	2200	2,74E-04	2800	3,40E-04	3300	3,15E-04	4100	1,91E-04	7500	1,80E-05
1400	1,16E-04	2200	2,74E-04	2800	3,40E-04	3400	2,96E-04	4200	1,80E-04	8000	1,97E-05
1400	1,16E-04	2200	2,74E-04	2900	3,46E-04	3400	2,96E-04	4200	1,80E-04	8000	1,97E-05
1500	1,39E-04	2300	2,84E-04	2900	3,46E-04	3400	2,96E-04	4200	1,80E-04	8300	1,54E-05
1500	1,39E-04	2300	2,84E-04	2900	3,46E-04	3400	2,96E-04	4300	1,68E-04		
1600	1,65E-04	2300	2,84E-04	2900	3,46E-04	3400	2,96E-04	4400	1,53E-04		
1600	1,65E-04	2300	2,84E-04	3000	3,48E-04	3400	2,96E-04	4400	1,53E-04		
1600	1,65E-04	2400	2,93E-04	3000	3,48E-04	3500	2,77E-04	4400	1,53E-04		
1700	1,89E-04	2400	2,93E-04	3000	3,48E-04	3500	2,77E-04	4500	1,37E-04		
1700	1,89E-04	2400	2,93E-04	3000	3,48E-04	3600	2,57E-04	4500	1,37E-04		
1700	1,89E-04	2500	3,04E-04	3000	3,48E-04	3600	2,57E-04	4500	1,37E-04		
1800	2,12E-04	2500	3,04E-04	3000	3,48E-04	3600	2,57E-04	4500	1,37E-04		
1800	2,12E-04	2500	3,04E-04	3000	3,48E-04	3600	2,57E-04	4500	1,37E-04		
1800	2,12E-04	2500	3,04E-04	3000	3,48E-04	3600	2,57E-04	4500	1,37E-04		
1800	2,12E-04	2500	3,04E-04	3000	3,48E-04	3600	2,57E-04	4600	1,21E-04		
1800	2,12E-04	2500	3,04E-04	3100	3,43E-04	3600	2,57E-04	4600	1,21E-04		
1900	2,32E-04	2500	3,04E-04	3100	3,43E-04	3700	2,38E-04	4700	1,04E-04		
1900	2,32E-04	2500	3,04E-04	3100	3,43E-04	3700	2,38E-04	4700	1,04E-04		
1900	2,32E-04	2500	3,04E-04	3100	3,43E-04	3700	2,38E-04	4700	1,04E-04		
1900	2,32E-04	2600	3,17E-04	3100	3,43E-04	3700	2,38E-04	5100	5,68E-05		
1900	2,32E-04	2600	3,17E-04	3100	3,43E-04	3700	2,38E-04	5200	5,23E-05		
2000	2,49E-04	2600	3,17E-04	3100	3,43E-04	3800	2,22E-04	5200	5,23E-05		
2000	2,49E-04	2600	3,17E-04	3100	3,43E-04	3800	2,22E-04	5200	5,23E-05		
2000	2,49E-04	2700	3,30E-04	3100	3,43E-04	3800	2,22E-04	5300	5,22E-05		
2000	2,49E-04	2700	3,30E-04	3200	3,31E-04	3800	2,22E-04	5500	5,71E-05		
2000	2,49E-04	2700	3,30E-04	3200	3,31E-04	3900	2,09E-04	5600	5,86E-05		
2000	2,49E-04	2700	3,30E-04	3200	3,31E-04	4000	2,00E-04	5800	5,58E-05		

Source : enquête Pauvreté CREDOC - CES 1995

Densité du revenu par unité de consommation, estimation lissée : h=1000 (graphique N° 22)

$$\hat{f}(y) = \frac{1}{193 * 1000} \sum_{i=1}^{193} K(\frac{y - y_i}{1000}), \text{ avec}$$

$K(u)=1+\cos 2\pi u$ si $-0,5 \leq u < 0,5$

$K(u)=0$ sinon

revenu	*densité*	*revenu*	*densité*	*revenu*	*densité*	*revenu*	*densité*	*revenu*	*densité*	*revenu*	*densité*
400	3,37E-05	2000	2,69E-04	2700	3,29E-04	3200	3,52E-04	4000	1,97E-04	5800	6,80E-05
500	4,05E-05	2100	2,74E-04	2700	3,29E-04	3200	3,52E-04	4000	1,97E-04	5800	6,80E-05
500	4,05E-05	2100	2,74E-04	2700	3,29E-04	3200	3,52E-04	4000	1,97E-04	5900	6,64E-05
700	3,99E-05	2100	2,74E-04	2800	3,50E-04	3200	3,52E-04	4000	1,97E-04	5900	6,64E-05
800	3,96E-05	2100	2,74E-04	2800	3,50E-04	3200	3,52E-04	4000	1,97E-04	6000	5,72E-05
1100	5,78E-05	2200	2,74E-04	2800	3,50E-04	3300	3,24E-04	4100	1,86E-04	6200	2,73E-05
1100	5,78E-05	2200	2,74E-04	2800	3,50E-04	3300	3,24E-04	4100	1,86E-04	6700	0
1100	5,78E-05	2200	2,74E-04	2800	3,50E-04	3300	3,24E-04	4100	1,86E-04	7000	1,75E-05
1200	7,00E-05	2200	2,74E-04	2800	3,50E-04	3300	3,24E-04	4100	1,86E-04	7300	2,07E-05
1400	1,03E-04	2200	2,74E-04	2800	3,50E-04	3300	3,24E-04	4100	1,86E-04	7500	1,72E-05
1400	1,03E-04	2200	2,74E-04	2800	3,50E-04	3400	2,93E-04	4200	1,79E-04	8000	2,43E-05
1400	1,03E-04	2200	2,74E-04	2900	3,66E-04	3400	2,93E-04	4200	1,79E-04	8000	2,43E-05
1500	1,27E-04	2300	2,75E-04	2900	3,66E-04	3400	2,93E-04	4200	1,79E-04	8300	1,75E-05
1500	1,27E-04	2300	2,75E-04	2900	3,66E-04	3400	2,93E-04	4300	1,74E-04		
1600	1,58E-04	2300	2,75E-04	2900	3,66E-04	3400	2,93E-04	4400	1,65E-04		
1600	1,58E-04	2300	2,75E-04	3000	3,76E-04	3400	2,93E-04	4400	1,65E-04		
1600	1,58E-04	2400	2,82E-04	3000	3,76E-04	3500	2,63E-04	4400	1,65E-04		
1700	1,93E-04	2400	2,82E-04	3000	3,76E-04	3500	2,63E-04	4500	1,52E-04		
1700	1,93E-04	2400	2,82E-04	3000	3,76E-04	3600	2,41E-04	4500	1,52E-04		
1700	1,93E-04	2500	2,95E-04	3000	3,76E-04	3600	2,41E-04	4500	1,52E-04		
1800	2,26E-04	2500	2,95E-04	3000	3,76E-04	3600	2,41E-04	4500	1,52E-04		
1800	2,26E-04	2500	2,95E-04	3000	3,76E-04	3600	2,41E-04	4500	1,52E-04		
1800	2,26E-04	2500	2,95E-04	3000	3,76E-04	3600	2,41E-04	4500	1,52E-04		
1800	2,26E-04	2500	2,95E-04	3000	3,76E-04	3600	2,41E-04	4600	1,32E-04		
1800	2,26E-04	2500	2,95E-04	3100	3,73E-04	3600	2,41E-04	4600	1,32E-04		
1900	2,54E-04	2500	2,95E-04	3100	3,73E-04	3700	2,28E-04	4700	1,03E-04		
1900	2,54E-04	2500	2,95E-04	3100	3,73E-04	3700	2,28E-04	4700	1,03E-04		
1900	2,54E-04	2500	2,95E-04	3100	3,73E-04	3700	2,28E-04	4700	1,03E-04		
1900	2,54E-04	2600	3,11E-04	3100	3,73E-04	3700	2,28E-04	5100	4,92E-05		
1900	2,54E-04	2600	3,11E-04	3100	3,73E-04	3700	2,28E-04	5200	5,44E-05		
2000	2,69E-04	2600	3,11E-04	3100	3,73E-04	3800	2,20E-04	5200	5,44E-05		
2000	2,69E-04	2600	3,11E-04	3100	3,73E-04	3800	2,20E-04	5200	5,44E-05		
2000	2,69E-04	2700	3,29E-04	3100	3,73E-04	3800	2,20E-04	5300	5,56E-05		
2000	2,69E-04	2700	3,29E-04	3200	3,52E-04	3800	2,20E-04	5500	5,10E-05		
2000	2,69E-04	2700	3,29E-04	3200	3,52E-04	3900	2,11E-04	5600	5,48E-05		
2000	2,69E-04	2700	3,29E-04	3200	3,52E-04	4000	1,97E-04	5800	6,80E-05		

Source : enquête Pauvreté CREDOC - CES 1995

Densité du salaire par unité de consommation, estimation lissée : h=1000
(graphiques N° 17 et 22)

$$\hat{f}(y) = \frac{1}{193 * 1000} \sum_{i=1}^{193} K\left(\frac{y - y_i}{1000}\right), \text{ avec}$$

$K(u)=1+\cos 2\pi u$ si $-0,5 \leq u < 0,5$

$K(u)=0$ sinon

sal aire	*densité*	*sal aire*	*densité*	*sal aire*	*densité*	*sal aire*	*densité*	*sal aire*	*densité*	*sal aire*	*densité*
200	7,49E-05	1000	2,85E-04	1700	3,73E-04	2100	3,78E-04	2600	3,00E-04	4200	7,32E-05
300	1,23E-04	1000	2,85E-04	1700	3,73E-04	2100	3,78E-04	2700	2,59E-04	4500	1,95E-05
400	1,76E-04	1000	2,85E-04	1700	3,73E-04	2100	3,78E-04	2700	2,59E-04	5200	2,07E-05
400	1,76E-04	1000	2,85E-04	1800	4,08E-04	2100	3,78E-04	2700	2,59E-04	5300	2,33E-05
400	1,76E-04	1100	2,72E-04	1800	4,08E-04	2100	3,78E-04	2800	2,06E-04	5600	2,95E-05
400	1,76E-04	1100	2,72E-04	1800	4,08E-04	2200	3,63E-04	2800	2,06E-04	5800	3,43E-05
500	2,22E-04	1100	2,72E-04	1800	4,08E-04	2200	3,63E-04	2800	2,06E-04	5800	3,43E-05
500	2,22E-04	1100	2,72E-04	1800	4,08E-04	2200	3,63E-04	2900	1,58E-04	6000	2,49E-05
500	2,22E-04	1100	2,72E-04	1800	4,08E-04	2200	3,63E-04	2900	1,58E-04	7000	2,07E-05
500	2,22E-04	1100	2,72E-04	1800	4,08E-04	2200	3,63E-04	2900	1,58E-04	7000	2,07E-05
500	2,22E-04	1100	2,72E-04	1800	4,08E-04	2300	3,56E-04	2900	1,58E-04	7500	0
500	2,22E-04	1200	2,58E-04	1800	4,08E-04	2400	3,46E-04	2900	1,58E-04	8000	0
500	2,22E-04	1200	2,58E-04	1800	4,08E-04	2400	3,46E-04	3000	1,27E-04	8300	0
600	2,55E-04	1200	2,58E-04	1800	4,08E-04	2400	3,46E-04	3000	1,27E-04		
600	2,55E-04	1300	2,49E-04	1800	4,08E-04	2400	3,46E-04	3100	1,08E-04		
600	2,55E-04	1300	2,49E-04	1800	4,08E-04	2400	3,46E-04	3100	1,08E-04		
700	2,77E-04	1300	2,49E-04	1800	4,08E-04	2400	3,46E-04	3300	9,01E-05		
700	2,77E-04	1300	2,49E-04	1800	4,08E-04	2400	3,46E-04	3400	9,46E-05		
700	2,77E-04	1300	2,49E-04	1900	4,18E-04	2400	3,46E-04	3400	9,46E-05		
700	2,77E-04	1400	2,57E-04	1900	4,18E-04	2400	3,46E-04	3400	9,46E-05		
700	2,77E-04	1400	2,57E-04	2000	4,00E-04	2500	3,27E-04	3500	1,02E-04		
700	2,77E-04	1400	2,57E-04	2000	4,00E-04	2500	3,27E-04	3500	1,02E-04		
700	2,77E-04	1500	2,87E-04	2000	4,00E-04	2500	3,27E-04	3500	1,02E-04		
700	2,77E-04	1500	2,87E-04	2000	4,00E-04	2500	3,27E-04	3600	1,05E-04		
800	2,88E-04	1500	2,87E-04	2000	4,00E-04	2500	3,27E-04	3700	1,07E-04		
800	2,88E-04	1500	2,87E-04	2000	4,00E-04	2500	3,27E-04	3700	1,07E-04		
800	2,88E-04	1500	2,87E-04	2000	4,00E-04	2500	3,27E-04	3800	1,07E-04		
800	2,88E-04	1600	3,28E-04	2000	4,00E-04	2500	3,27E-04	3800	1,07E-04		
900	2,90E-04	1600	3,28E-04	2000	4,00E-04	2500	3,27E-04	3800	1,07E-04		
900	2,90E-04	1600	3,28E-04	2000	4,00E-04	2500	3,27E-04	4000	1,02E-04		
900	2,90E-04	1600	3,28E-04	2000	4,00E-04	2500	3,27E-04	4000	1,02E-04		
900	2,90E-04	1600	3,28E-04	2000	4,00E-04	2500	3,27E-04	4000	1,02E-04		
900	2,90E-04	1600	3,28E-04	2100	3,78E-04	2500	3,27E-04	4000	1,02E-04		
1000	2,85E-04	1600	3,28E-04	2100	3,78E-04	2500	3,27E-04	4100	9,07E-05		
1000	2,85E-04	1700	3,73E-04	2100	3,78E-04	2600	3,00E-04	4100	9,07E-05		
1000	2,85E-04	1700	3,73E-04	2100	3,78E-04	2600	3,00E-04	4200	7,32E-05		

Source : enquête Pauvreté CREDOC - CES 1995

Densité du revenu hors salaire par unité de consommation, estimation lissée - h=1000 (graphique N°22)

$$\hat{f}(y) = \frac{1}{193 * 1000} \sum_{i=1}^{193} K(\frac{y - y_i}{1000}), \text{ avec}$$

$K(u)=1+\cos 2\pi u$ si $-0,5 \leq u < 0,5$

$K(u)=0$ sinon

rev hsal	*densité*	*rev hsal*	*densité*	*rev hsal*	*densité*	*rev hsal*	*densité*	*rev hsal*	*densité*	*rev hsal*	*densité*
0	5,78E-04	0	5,78E-04	400	3,82E-04	1000	4,31E-04	1500	2,62E-04	3000	5,46E-05
0	5,78E-04	0	5,78E-04	500	3,51E-04	1000	4,31E-04	1500	2,62E-04	3100	5,86E-05
0	5,78E-04	0	5,78E-04	500	3,51E-04	1000	4,31E-04	1500	2,62E-04	3200	5,60E-05
0	5,78E-04	0	5,78E-04	500	3,51E-04	1000	4,31E-04	1600	2,45E-04	3200	5,60E-05
0	5,78E-04	0	5,78E-04	500	3,51E-04	1000	4,31E-04	1600	2,45E-04	3200	5,60E-05
0	5,78E-04	0	5,78E-04	600	3,56E-04	1000	4,31E-04	1600	2,45E-04	3400	3,99E-05
0	5,78E-04	0	5,78E-04	600	3,56E-04	1100	4,14E-04	1600	2,45E-04	3700	3,76E-05
0	5,78E-04	0	5,78E-04	600	3,56E-04	1100	4,14E-04	1700	2,31E-04	3800	4,79E-05
0	5,78E-04	0	5,78E-04	600	3,56E-04	1100	4,14E-04	1700	2,31E-04	4000	5,54E-05
0	5,78E-04	0	5,78E-04	600	3,56E-04	1100	4,14E-04	1700	2,31E-04	4000	5,54E-05
0	5,78E-04	0	5,78E-04	600	3,56E-04	1100	4,14E-04	1700	2,31E-04	4000	5,54E-05
0	5,78E-04	0	5,78E-04	700	3,64E-04	1100	4,14E-04	1700	2,31E-04	4000	5,54E-05
0	5,78E-04	100	6,02E-04	700	3,64E-04	1100	4,14E-04	1700	2,31E-04	4300	2,47E-05
0	5,78E-04	200	5,54E-04	700	3,64E-04	1100	4,14E-04	1700	2,31E-04		
0	5,78E-04	200	5,54E-04	700	3,64E-04	1100	4,14E-04	1800	2,11E-04		
0	5,78E-04	200	5,54E-04	700	3,64E-04	1100	4,14E-04	1800	2,11E-04		
0	5,78E-04	200	5,54E-04	700	3,64E-04	1200	3,76E-04	1800	2,11E-04		
0	5,78E-04	200	5,54E-04	800	3,88E-04	1200	3,76E-04	1900	1,88E-04		
0	5,78E-04	300	4,65E-04	800	3,88E-04	1200	3,76E-04	1900	1,88E-04		
0	5,78E-04	300	4,65E-04	800	3,88E-04	1200	3,76E-04	1900	1,88E-04		
0	5,78E-04	300	4,65E-04	800	3,88E-04	1200	3,76E-04	2000	1,70E-04		
0	5,78E-04	300	4,65E-04	800	3,88E-04	1200	3,76E-04	2000	1,70E-04		
0	5,78E-04	300	4,65E-04	800	3,88E-04	1200	3,76E-04	2000	1,70E-04		
0	5,78E-04	300	4,65E-04	800	3,88E-04	1200	3,76E-04	2100	1,55E-04		
0	5,78E-04	300	4,65E-04	900	4,19E-04	1200	3,76E-04	2100	1,55E-04		
0	5,78E-04	400	3,82E-04	900	4,19E-04	1200	3,76E-04	2200	1,39E-04		
0	5,78E-04	400	3,82E-04	900	4,19E-04	1200	3,76E-04	2200	1,39E-04		
0	5,78E-04	400	3,82E-04	900	4,19E-04	1300	3,30E-04	2200	1,39E-04		
0	5,78E-04	400	3,82E-04	900	4,19E-04	1300	3,30E-04	2200	1,39E-04		
0	5,78E-04	400	3,82E-04	900	4,19E-04	1400	2,90E-04	2300	1,16E-04		
0	5,78E-04	400	3,82E-04	900	4,19E-04	1400	2,90E-04	2300	1,16E-04		
0	5,78E-04	400	3,82E-04	900	4,19E-04	1400	2,90E-04	2300	1,16E-04		
0	5,78E-04	400	3,82E-04	900	4,19E-04	1500	2,62E-04	2300	1,16E-04		
0	5,78E-04	400	3,82E-04	1000	4,31E-04	1500	2,62E-04	2400	8,71E-05		
0	5,78E-04	400	3,82E-04	1000	4,31E-04	1500	2,62E-04	2800	3,51E-05		
0	5,78E-04	400	3,82E-04	1000	4,31E-04	1500	2,62E-04	2800	3,51E-05		

Source : enquête Pauvreté CREDOC - CES 1995

Densité à deux dimensions : la densité du salaire et du revenu hors salaire par unité de consommation. Les données sont triées sur la densité. (Graphique N°28)

$$\hat{f}(x,y) = \frac{1}{500^2 * 193} \sum_{i=1}^{193} W\left[\frac{x - x_i}{500}, \frac{y - y_i}{500}\right], \text{ avec } \begin{cases} W(u,v) = \dfrac{1 + \cos \pi\sqrt{u^2 + v^2}}{\pi} & \text{si } u2+v2 \leq 1 \\ W(u,v)=0 & \text{sinon} \end{cases}$$

salaire	*revenu hors salaire*	*densité*	*salaire*	*revenu hors salaire*	*densité*	*salaire*	*revenu hors salaire*	*densité*
1300	3400	1,32E-08	300	1200	3,03E-08	500	0	5,46E-08
1600	4300	1,32E-08	6000	0	3,17E-08	500	0	5,46E-08
1700	3800	1,32E-08	2500	1600	3,50E-08	800	2200	5,49E-08
2200	3700	1,32E-08	5600	0	3,75E-08	1300	1500	5,51E-08
2400	2800	1,32E-08	2900	700	3,84E-08	3100	0	5,59E-08
2900	3200	1,32E-08	700	2300	4,11E-08	2800	800	5,62E-08
3500	3200	1,32E-08	3300	600	4,14E-08	800	700	5,65E-08
4000	4000	1,32E-08	1200	1200	4,16E-08	4100	0	5,76E-08
7500	0	1,32E-08	2500	2200	4,16E-08	3400	500	5,80E-08
4000	1200	1,36E-08	600	200	4,19E-08	1400	0	5,91E-08
4200	1600	1,36E-08	4200	200	4,25E-08	1300	500	5,95E-08
500	3200	1,42E-08	2500	1700	4,26E-08	1500	0	6,07E-08
1600	2800	1,45E-08	5800	0	4,37E-08	1800	1700	6,13E-08
1600	2400	1,58E-08	5800	0	4,37E-08	3400	400	6,15E-08
4500	0	1,68E-08	400	900	4,40E-08	3400	400	6,15E-08
7000	0	1,78E-08	1100	2100	4,51E-08	1500	1700	6,18E-08
7000	300	1,78E-08	700	0	4,52E-08	2700	500	6,29E-08
8000	0	1,78E-08	3800	400	4,54E-08	700	1400	6,55E-08
8300	0	1,78E-08	1000	400	4,57E-08	700	1900	6,61E-08
200	4000	1,90E-08	1000	2300	4,63E-08	4000	0	6,67E-08
400	2100	1,99E-08	1100	0	4,64E-08	4000	0	6,67E-08
3100	1300	2,09E-08	1100	0	4,64E-08	1000	1900	6,69E-08
2900	1400	2,16E-08	400	0	4,69E-08	3000	0	6,72E-08
1000	3000	2,40E-08	2600	2000	5,04E-08	3000	0	6,72E-08
2800	2200	2,41E-08	2400	2000	5,06E-08	900	1500	6,75E-08
1700	2000	2,45E-08	900	2300	5,11E-08	1600	0	6,86E-08
900	3100	2,49E-08	500	1700	5,16E-08	800	1100	6,86E-08
2000	2200	2,50E-08	3500	0	5,16E-08	800	1100	6,86E-08
2100	1800	2,50E-08	3500	0	5,16E-08	3600	300	6,95E-08
600	4000	2,64E-08	1200	0	5,19E-08	700	900	7,02E-08
5200	0	2,64E-08	500	1300	5,19E-08	3700	0	7,09E-08
2200	2300	2,72E-08	1000	1400	5,24E-08	600	1000	7,10E-08
400	700	2,86E-08	500	1500	5,30E-08	2100	600	7,14E-08
5300	0	2,97E-08	2500	1900	5,42E-08	2900	200	7,19E-08
500	4000	2,97E-08	1300	1600	5,45E-08	2900	200	7,19E-08
4100	400	2,99E-08	3700	400	5,45E-08	1000	1800	7,22E-08

La densité du salaire et du revenu hors salaire - Suite du tableau

salaire	*revenu hors salaire*	*densité*	*salaire*	*revenu hors salaire*	*densité*	*salaire*	*revenu hors salaire*	*densité*
1500	1500	7,27E-08	1800	700	9,18E-08	2000	900	1,44E-07
700	1000	7,29E-08	2500	800	9,31E-08	2100	1200	1,45E-07
2200	400	7,33E-08	2700	0	9,34E-08	1800	1000	1,53E-07
2500	500	7,40E-08	1700	100	9,77E-08	1800	1000	1,53E-07
3800	0	7,42E-08	1800	400	9,77E-08	1800	1000	1,53E-07
3800	0	7,42E-08	1800	400	9,77E-08	1800	1100	1,55E-07
1600	300	7,43E-08	1800	400	9,77E-08	1900	1200	1,56E-07
1500	1100	7,49E-08	2100	300	9,84E-08	2100	1000	1,59E-07
1000	600	7,55E-08	2500	0	9,97E-08	2100	1100	1,65E-07
2700	900	7,69E-08	2500	0	9,97E-08	2100	1100	1,65E-07
1700	1600	7,78E-08	2500	0	9,97E-08	2100	1100	1,65E-07
1800	1500	7,85E-08	1600	1000	1,02E-07	1900	1100	1,73E-07
1800	1500	7,85E-08	2600	200	1,03E-07	2000	1100	1,75E-07
2400	400	7,94E-08	2400	0	1,03E-07			
1500	800	8,01E-08	2400	0	1,03E-07			
1100	900	8,10E-08	2400	1200	1,04E-07			
1600	1500	8,14E-08	2500	1100	1,04E-07			
900	1800	8,15E-08	2500	900	1,05E-07			
700	1700	8,24E-08	1800	0	1,06E-07			
700	1700	8,24E-08	2000	300	1,10E-07			
2500	700	8,31E-08	2400	900	1,11E-07			
900	1700	8,43E-08	2500	1000	1,12E-07			
2500	1200	8,49E-08	2300	1200	1,18E-07			
1700	400	8,56E-08	2200	0	1,20E-07			
1100	600	8,57E-08	2100	0	1,25E-07			
1400	800	8,60E-08	2400	1000	1,25E-07			
1400	800	8,60E-08	2000	0	1,26E-07			
2800	0	8,61E-08	2000	0	1,26E-07			
1200	600	8,68E-08	2000	0	1,26E-07			
2000	700	8,69E-08	2000	0	1,26E-07			
1300	900	8,78E-08	2000	0	1,26E-07			
1800	600	8,88E-08	2000	0	1,26E-07			
2600	300	8,95E-08	2200	1200	1,30E-07			
1100	800	9,15E-08	1800	1200	1,42E-07			
1100	800	9,15E-08	1800	1200	1,42E-07			
2400	300	9,15E-08	2000	900	1,44E-07			

Source : enquête Pauvreté CREDOC - CES 1995

Annexe N° 6

Les Q-Plots

Pour une variable ayant n valeurs, le quantile plot représente les valeurs des n quantiles en fonction de leur ordre. L'ordonnée du graphique est donc les valeurs ordonnées de la variable, l'abscisse le rang des observations divisé par n.

Afon de représenter tous les points, le rang augmente de 1 à chaque valeur de la variable, même dans le cas où elle et égale à la précédente.

Dans le cas des QQ-plots, les graphiques étant ici tracés sur des variables ayant le même nombre de valeurs, l'abscisse est constituée des valeurs ordonnées d'une des variables et l'ordonnée des valeurs ordonnées de l'autre variable. Nous avons donné ces valeurs dans l'annexe N°1.

Ordre et valeurs des 193 quantiles du revenu par unité de consommation (Graphique N° 12) : données triées sur l'ordre des quantiles

ordre du quantile	*revenu*	*ordre du quantile*	*revenu*	*ordre du quantile*	*revenu*	*ordre du quantile*	*revenu*	*ordre du quantile*	*revenu*	*ordre du quantile*	*revenu*
0,5	400	19,2	2000	37,8	2700	56,5	3200	75,1	4000	93,8	5800
1,0	500	19,7	2100	38,3	2700	57,0	3200	75,6	4000	94,3	5800
1,6	500	20,2	2100	38,9	2700	57,5	3200	76,2	4000	94,8	5900
2,1	700	20,7	2100	39,4	2800	58,0	3200	76,7	4000	95,3	5900
2,6	800	21,2	2100	39,9	2800	58,5	3200	77,2	4000	95,9	6000
3,1	1100	21,8	2200	40,4	2800	59,1	3300	77,7	4100	96,4	6200
3,6	1100	22,3	2200	40,9	2800	59,6	3300	78,2	4100	96,9	6700
4,1	1100	22,8	2200	41,5	2800	60,1	3300	78,8	4100	97,4	7000
4,7	1200	23,3	2200	42,0	2800	60,6	3300	79,3	4100	97,9	7300
5,2	1400	23,8	2200	42,5	2800	61,1	3300	79,8	4100	98,4	7500
5,7	1400	24,4	2200	43,0	2800	61,7	3400	80,3	4200	99,0	8000
6,2	1400	24,9	2200	43,5	2900	62,2	3400	80,8	4200	99,5	8000
6,7	1500	25,4	2300	44,0	2900	62,7	3400	81,3	4200	100,0	8300
7,3	1500	25,9	2300	44,6	2900	63,2	3400	81,9	4300		
7,8	1600	26,4	2300	45,1	2900	63,7	3400	82,4	4400		
8,3	1600	26,9	2300	45,6	3000	64,2	3400	82,9	4400		
8,8	1600	27,5	2400	46,1	3000	64,8	3500	83,4	4400		
9,3	1700	28,0	2400	46,6	3000	65,3	3500	83,9	4500		
9,8	1700	28,5	2400	47,2	3000	65,8	3600	84,5	4500		
10,4	1700	29,0	2500	47,7	3000	66,3	3600	85,0	4500		
10,9	1800	29,5	2500	48,2	3000	66,8	3600	85,5	4500		
11,4	1800	30,1	2500	48,7	3000	67,4	3600	86,0	4500		
11,9	1800	30,6	2500	49,8	3000	67,9	3600	86,5	4500		
12,4	1800	31,1	2500	49,7	3000	68,4	3600	87,0	4600		
13,0	1800	31,6	2500	50,3	3100	68,9	3600	87,6	4600		
13,5	1900	32,1	2500	50,8	3100	69,4	3700	88,1	4700		
14,0	1900	32,6	2500	0,0	3100	69,9	3700	88,6	4700		
14,5	1900	33,2	2500	51,8	3100	70,5	3700	89,1	4700		
15,0	1900	33,7	2600	52,3	3100	71,0	3700	89,6	5100		
15,5	1900	34,2	2600	52,8	3100	71,5	3700	90,2	5200		
16,1	2000	34,7	2600	53,4	3100	72,0	3800	90,7	5200		
16,6	2000	35,2	2600	53,9	3100	72,5	3800	91,2	5200		
17,1	2000	35,8	2700	54,4	3100	73,1	3800	91,7	5300		
17,6	2000	36,3	2700	54,9	3200	73,6	3800	92,2	5500		
18,1	2000	36,8	2700	55,4	3200	74,1	3900	92,7	5600		
18,7	2000	37,3	2700	56,0	3200	74,6	4000	93,3	5800		

Source : enquête Pauvreté CREDOC - CES 1995

Annexe N° 7

Les box plots

Les résumés numériques utilisés pour calculer la position, la hauteur de la boîte, la taille des pattes ont été calculés en utilisant la procédure univariate de SAS. Le calcul des valeurs lointaines et périphériques et le tracé de la boîte ont ensuite été effectués sous EXCEL. La médiane, les quartiles, les frontières, les valeurs lointaines et périphériques ont été entrés en ordonnée. Les abscisses utilisées n'ont aucune signification. Elles ont simplement permis de positionner la boîte (ou les boîtes) dans le cas de comparaison.

Les box plots du revenu, du salaire et du revenu hors salaire par unité de consommation (graphiques N°13, 18 et 23)

Le revenu	**Le salaire**	**Le revenu hors salaire**
Valeurs calculées	**Valeurs calculées**	**Valeurs calculées**
La patte supérieure	*La patte supérieure*	*La patte supérieure*
6200	4500	3400
le corps de la boîte	*Le corps de la boîte*	*le corps de la boîte*
4000	2600	1500
3100	2000	900
2300	1200	100
la patte inférieure	*La patte inférieure*	*La patte inférieure*
400	400	0
les valeurs péripériques	*Les valeurs périphériques*	*les valeurs périphériques*
6700	5200	3700
7000	5300	3800
7300	5600	4000
7500	5800	4300
8000	5800	
8000	6000	
8300	6000	
	Les valeurs lointaines	
	7000	
	7000	
	7500	
	8000	
	8300	

Source : enquête Pauvreté CREDOC - CES 1995

Box plot du revenu hors salaire en fonction du salaire réparti en 4 classes (graphique N°30)

400≤sal<1200 médiane=700	**1200≤sal<2000 médiane=1600**	**2000≤sal<2600 médiane=2200**	**2600≤sal médiane=3600**
Box plot revenu hors salaire	**Box plot revenu hors salaire**	**Box plot revenu hors salaire**	**Box plot revenu hors salaire**
La patte supérieure	*La patte supérieure*	*La patte supérieure*	*La patte supérieure*
4000	2800	2800	1400
le corps de la boîte	*Le corps de la boîte*	*Le corps de la boîte*	*le corps de la boîte*
2100	1500	1200	600
1400	1000	900	200
700	500	0	0
la patte inférieure	*La patte inférieure*	*La patte inférieure*	*La patte inférieure*
0	0	0	0
	Les valeurs périphériques	*Les valeurs périphériques*	*les valeurs périphériques*
	3400	3700	1600
	3800		2000
	4300		2200
			les valeurs lointaines
			3200
			4000

Source : enquête Pauvreté CREDOC - CES 1995

Box plot du salaire par modalité de la situation familiale (Graphique N°42)

homme seul	femme seule	couple
Box plot salaire	**Box plot salaire**	**Box plot salaire**
La patte supérieure	*La patte supérieure*	*La patte supérieure*
5800	4000	4500
le corps de la boîte	*Le corps de la boîte*	*le corps de la boîte*
3900	2500	2500
2500	1700	1800
2000	1000	1100
la patte inférieure	*La patte inférieure*	*La patte inférieure*
500	200	300
les valeurs péripériques	*Les valeurs périphériques*	*les valeurs périphériques*
7000	6000	5200
7500		
8000		
8300		

Annexe N°8

Box plot à deux dimensions

Trace médiane du revenu hors salaire en fonction du salaire calculée à partir de 15 intervalles (Graphiques N°32, N°34 N°36 et N°37)

Avant lissage		**Après lissage**	
médiane des salaires	*médiane des revenus hors salaire*	*médiane des salaires*	*médiane des revenus hors salaire*
450	1250	450	1250
700	1400	690	1365
900	1700	900	1275
1100	800	1125	1025
1400	900	1400	925
1700	1500	1650	975
1800	1000	1825	1000
2000	300	1975	1000
2100	1100	2150	1000
2400	1000	2350	925
2500	1000	2550	725
2800	500	2875	525
3400	400	3400	375
4000	200	4325	200
5900	0	5900	0

Source : enquête Pauvreté CREDOC - CES 1995

Traces quartiles du revenu hors salaire calculée à partir du salaire réparti en quinze intervalles (graphiques N° 36 et N° 37)

Les deux tableaux ne donnent que les valeurs pour le revenu hors salaire. L'abscisse du graphique est la série lissée des médianes du salaire en 15 intervalles, donnée dans le tableau précédent.

Q1 (resp. Q3) désigne le quartile inférieur (resp. supérieur) des revenus hors salaire, série non lissée. M désigne la médiane des revenus hors salaire, série non lissée.

			trace quartile inférieure
Q1	M-Q1	(M-Q1) Lissée	(M)lissée-(M-Q1)lissée
350	650	900	350
900	500	650	715
1100	200	475	750
500	400	350	525
800	600	300	425
400	900	300	425
700	200	300	350
0	600	400	300
400	100	600	350
300	1000	600	375
700	600	400	275
200	400	300	200
0	200	275	150
0	200	175	50
0	0	0	0
			Trace quartile supérieure
Q3	Q3-M	(Q3-M)Lissée	(M)lissée+(Q3-M)lissée
1900	650	650	1900
1900	500	513	1875
1900	200	425	1700
1200	400	450	1475
1500	600	550	1475
2400	900	600	1475
1200	200	600	1300
900	600	600	1200
1200	100	600	1300
2000	1000	600	1425
1600	600	550	1275
900	400	400	925
600	200	250	625
400	200	150	350
0	0	0	0

Source : enquête Pauvreté CREDOC - CES 1995

Traces médianes empiriques et ratios de dépendance du revenu hors salaire en fonction du salaire réparti en 15 intervalles (graphiques N° 34 et N° 35)

Comme pour les tableaux précédents, on ne donne que les valeurs pour le revenu hors salaire, les valeurs utilisées pour le salaire sont toujours les médianes lissées.

M désigne toujours la série des médianes empiriques avant lissage.

Avant lissage				**Après lissage**			
Médiane théorique positive	*Médiane théorique négative*	*Médiane théorique Neutre*	*ratio = (M-900) /(900-méd neg)*	*trace médiane positive*	*trace médiane négative*	*trace médiaze neutre*	*ratios*
0	3500	900	-0,1	0	3500	900	-0,7
0	2550	900	-0,3	0	2625	900	-0,4
0	1900	900	-0,8	25	1990	900	-0,2
100	1600	900	0,1	125	1615	900	-0,1
300	1350	900	0	290	1365	900	-0,3
450	1150	900	-2,4	475	1165	900	-0,8
700	1000	900	-1	675	1000	900	-1
850	850	900	-12	850	850	900	-0,7
900	700	900	1	1000	675	900	-0,1
900	450	900	0,2	1165	475	900	0,2
900	300	900	0,2	1365	290	900	0,01
900	100	900	-0,5	1615	125	900	-0,4
900	0	900	-0,6	1990	25	900	-0,6
900	0	900	-0,8	2625	0	900	-0,8
900	0	900	-1	3500	0	900	-1

Source : enquête Pauvreté CREDOC - CES 1995

Annexe N°9
Estimation du centre d'une distribution quantitative en fonction des valeurs d'une autre distribution quantitative

L'estimation a été calculée à partir du logiciel SAS. Nous avons effectué les 193 régressions nécessaires en utilisant des Macros.

Estimation du centre de la distribution du revenu hors salaire en fonction du salaire, 20 points pris en compte : données triées sur le salaire. (Graphique N°38)

salaire	*centre estimé*	*salaire*	*centre estimé*	*salaire*	*centre estimé*	*salaire*	*centre estimé*	*salaire*	*centre estimé*	*salaire*	*centre estimé*
200	3592	1000	1514	1700	1299	2100	1071	2600	1037	4200	571
300	2060	1000	1514	1700	1580	2100	1120	2700	724	4500	519
400	1113	1000	1514	1700	1580	2100	1135	2700	950	5200	327
400	1113	1000	1557	1800	709	2100	1143	2700	953		
400	1113	1100	908	1800	750	2100	1154	2800	851		
400	1113	1100	908	1800	807	2200	1368	2800	851		
500	1451	1100	908	1800	857	2200	1368	2800	868		
500	1427	1100	908	1800	913	2200	1388	2900	765		
500	1412	1100	948	1800	913	2200	1406	2900	762		
500	1408	1100	1218	1800	913	2200	1421	2900	774		
500	1410	1100	1227	1800	1071	2300	1233	2900	798		
500	1535	1200	1005	1800	1081	2400	833	2900	759		
500	1535	1200	1032	1800	1079	2400	804	3000	726		
600	1566	1200	1002	1800	913	2400	815	3000	774		
600	1566	1300	1128	1800	978	2400	834	3100	737		
600	1566	1300	1128	1800	1023	2400	917	3100	777		
700	1358	1300	1128	1800	1075	2400	924	3300	618		
700	1414	1300	1062	1800	1136	2400	933	3400	605		
700	1414	1300	1072	1900	1150	2400	939	3400	604		
700	1414	1400	1019	1900	1150	2400	945	3400	571		
700	1414	1400	1037	2000	354	2500	709	3500	524		
700	1414	1400	1062	2000	508	2500	791	3500	529		
700	1414	1500	1115	2000	516	2500	876	3500	564		
700	1468	1500	1291	2000	534	2500	971	3600	565		
800	1575	1500	1086	2000	554	2500	971	3700	602		
800	1575	1500	1086	2000	578	2500	971	3700	599		
800	1575	1500	1086	2000	597	2500	971	3800	608		
800	1634	1600	1340	2000	610	2500	1070	3800	642		
900	1662	1600	1483	2000	508	2500	1091	3800	636		
900	1717	1600	1533	2000	508	2500	1077	4000	637		
900	1717	1600	1533	2000	508	2500	977	4000	619		
900	1717	1600	1533	2000	554	2500	1050	4000	644		
900	1717	1600	1533	2100	911	2500	1135	4000	691		
1000	1333	1600	1605	2100	911	2500	1207	4100	524		
1000	1448	1700	1267	2100	911	2600	981	4100	529		
1000	1459	1700	1285	2100	911	2600	1009	4200	546		

Source : enquête Pauvreté CREDOC - CES 1995

Dépôt légal : Mai 1996

ISSN : 1257-9807

ISBN : 2-84104-060-7