Publications de
l'Institut de Mathématiques
de Toulouse

Statistique Descriptive Élémentaire

(version de mai 2010)

Alain Baccini

Institut de Mathématiques de Toulouse — UMR CNRS 5219
Université Paul Sabatier — 31062 – Toulouse cedex 9.

Table des matières

Chapitre 1

Introduction générale à la statistique

Pour fixer les idées, on commence par donner une définition très générale de la statistique. On fait ensuite un bref historique et l'on précise la différence entre statistique descriptive (objet de ce document) et statistique inférentielle (non traitée ici). On termine par la définition d'une dizaine de termes indispensables à la bonne compréhension de la suite de ce cours.

1.1 Généralités sur la statistique

1.1.1 Définition

Il n'est pas commode, dans une introduction, de donner une définition précise du concept de statistique, alors que son contenu sera en partie élaboré dans la suite de ce cours. Nous nous contenterons donc, pour fixer les idées, d'en donner une définition volontairement assez vague.

Definition 1 *On appelle Statistique l'ensemble des méthodes (ou encore des techniques) permettant d'analyser (on dira plutôt de* traiter*) des ensembles d'observations (nous parlerons de* données*).*

Les méthodes en question relèvent essentiellement des mathématiques et font largement appel à l'outil informatique pour leur mise en œuvre.

Pour éviter toute confusion, on notera la distinction entre *la* Statistique, au sens défini ci-dessus, et *une* statistique, terme parfois utilisé pour désigner des "données statistiques" (voir ce terme plus loin) ; par exemple, on parle de la statistique du commerce extérieur de la France. Dans la suite de ce cours, nous n'utiliserons pas le terme de statistique dans ce dernier sens.

1.1.2 Bref historique

De façon un peu grossière, on peut distinguer trois phases essentielles dans l'évolution de la statistique.

- Depuis l'antiquité (on fait remonter les premières manifestations de la statistique à la haute Égypte, avec l'enregistrement des crues du Nil) et jusqu'à la fin du $19^{\text{ième}}$ siècle, la statistique est restée principalement un ensemble de techniques de dénombrement : comptage d'une population (ou recensement, voir ce mot plus loin), des effectifs d'une armée (on en imagine aisément l'objectif !), etc. Les techniques étaient très rudimentaires et leur mise en œuvre restait l'apanage du pouvoir politique.
- Entre la fin du $19^{\text{ième}}$ siècle et les années 1960, s'est construit, notamment à la suite de l'école anglaise, la statistique mathématique (ou statistique inférentielle, voir plus loin). Le développement de la statistique au cours de cette période a, en fait, suivi le mouvement général de développement des sciences, notamment des mathématiques, de la physique et de la théorie des probabilités.

– Depuis les années 1960, avec le développement et la banalisation des outils informatiques et graphiques, la statistique, et surtout la statistique descriptive multidimensionnelle, a connu une expansion considérable.
À titre d'illustration, penser à un graphique de type "camembert" (voir le chapitre 2) : il est immédiat de le réaliser aujourd'hui avec n'importe quel tableur tel qu'*Excel* (même si le résultat peut parfois être contestable...) ; jusque dans les années 1985/90, réaliser un tel graphique de façon rigoureuse se faisait "à la main" et nécessitait des outils tels que rapporteur, compas..., une bonne dose d'application et de patience, et un temps non négligeable.

1.1.3 Statistique descriptive et statistique inférentielle

De manière un peu approximative, il est possible de classer les méthodes statistiques en deux groupes : celui des méthodes descriptives et celui des méthodes inférentielles.

– La statistique **descriptive**.
On regroupe sous ce terme les méthodes dont l'objectif principal est la *description* des données étudiées. Cette description des données se fait à travers leur **présentation** (la plus commode et la plus synthétique possible), leur **représentation graphique** et le calcul de **résumés numériques** (ou caractéristiques numériques). Dans cette optique, aucune hypothèse de type probabiliste n'est faite sur les données considérées (par exemple, il n'est pas nécessaire de supposer que les données sont les observations d'une loi normale). Les trois directions signalées ci-dessus (présentation, représentations graphiques et calcul de résumés numériques) serviront de guide dans la suite de ce cours.
On notera que les termes de statistique descriptive, *statistique exploratoire* et *analyse des données* sont quasiment synonymes.
– La statistique **inférentielle**.
Ce terme regroupe les méthodes dont l'objectif principal est de préciser un phénomène sur une population globale, à partir de son observation sur une partie restreinte de cette population (penser aux sondages). D'une certaine manière, il s'agit donc d'induire (ou encore d'inférer) du particulier au général. Le plus souvent, ce passage ne pourra se faire que moyennant des hypothèses de type probabiliste.
Les termes de statistique inférentielle, *statistique mathématique* et *statistique inductive* sont eux aussi quasiment synonymes.
Cette partie de la statistique, plus délicate, n'est pas traitée dans ce document.

D'un point de vue méthodologique, on notera que la statistique descriptive précède en général la statistique inférentielle dans une démarche de traitement de données : ces deux aspects de la statistique se complètent bien plus qu'ils ne s'opposent.

1.2 Terminologie de base

On précise ici un certain nombre de termes statistiques très courants qui seront régulièrement utilisés par la suite et qu'il convient de bien connaître.

Population (ou population statistique) : ensemble concerné par une étude statistique. On parle aussi de *champ de l'étude*.

Si l'on s'intéresse aux notes d'un groupe d'étudiants, ce groupe constitue la population. À noter que si l'on s'intéresse maintenant à la circulation automobile dans Toulouse, la population est alors constituée de l'ensemble des véhicules susceptibles de circuler dans Toulouse à une date donnée. Le terme de population et donc plus large en statistique que dans le langage courant.

Individu (ou unité statistique) : on désigne ainsi tout élément de la population considérée.

Dans les exemples indiqués ci-dessus, un individu est tout étudiant du groupe dans le premier cas et tout véhicule susceptible de circuler dans Toulouse dans le second. Là encore, on constate que le terme d'individu est plus large en statistique que dans le langage courant.

Échantillon : dans une étude statistique, il est fréquent que l'on n'observe pas la population tout entière (par exemple, on n'observe pas tous les véhicules ayant circulé un jour donné dans

Toulouse, mais seulement ceux étant passés dans certains points particuliers). Les observations du phénomène considéré sont donc réalisées sur une partie restreinte de la population, appelée échantillon.

On appelle donc échantillon le sous-ensemble de la population sur lequel sont effectivement réalisées les observations.

Taille de l'échantillon : c'est le cardinal de l'échantillon, autrement dit c'est le nombre d'individus qu'il contient (échantillon de taille 800, de taille 1000...).

En général, on note n la taille de l'échantillon considéré.

Enquête (statistique) : c'est l' opération consistant à observer (ou mesurer, ou questionner...) l'ensemble des individus d'un échantillon (ou, éventuellement, de la population complète).

Recensement : enquête dans laquelle l'échantillon observé est en fait la population tout entière (on parle aussi d'enquête *exhaustive*).

En France, on organise ainsi, de façon plus ou moins régulière, le recensement général de la population, le recensement général agricole...

Sondage : c'est, au contraire, une enquête dans laquelle l'échantillon observé est un sous-ensemble strict de la population (on parle, dans ce cas, d'enquête *non exhaustive*).

Les exemples de sondages dans la vie courante sont, de nos jours, légion.

Variable (statistique) : c'est une caractéristique (âge, salaire, sexe...), définie sur la population et observée sur l'échantillon.

D'un point de vue mathématique, une variable est une application définie sur l'échantillon. Si cette application est à valeurs dans $\mathbb{R}$ (ensemble des nombres réels), ou dans une partie de $\mathbb{R}$, elle est dite *quantitative* (âge, salaire, taille...) ; sinon elle est dite *qualitative* (sexe, catégorie socio-professionnelle...).

On retiendra que les **variables quantitatives** sont celles prenant des valeurs numériques et que les **variables qualitatives** sont celles prenant des valeurs non numériques (en faisant bien attention au fait qu'un codage ne représente pas une valeur : même si on code 1 les hommes et 2 les femmes, la variable "sexe" demeure qualitative).

Données (statistiques) : le terme de *données* est très utilisé en statistique. Il désigne l'ensemble des individus observés (ceux de l'échantillon), l'ensemble des variables considérées et les observations de ces variables sur ces individus.

Les données sont en général présentées sous forme de *tableaux* (individus en lignes et variables en colonnes) et stockées dans un fichier informatique.

Voici un tout petit exemple de données :

	sexe	âge	revenu mensuel net
individu 1	1	55	2068
individu 2	1	41	4687
individu 3	1	28	1235
individu 4	2	64	1941
individu 5	2	32	2456

Chapitre 2

Statistique descriptive unidimensionnelle

On considère ici une variable statistique unique, notée X. L'objectif est d'exposer les outils élémentaires, adaptés à la nature de X, permettant de présenter cette variable de façon synthétique, d'en faire une représentation graphique appropriée et d'en résumer les principales caractéristiques. Nous présenterons successivement le cas d'une variable quantitative discrète, puis celui d'une variable quantitative continue, enfin le cas d'une variable qualitative.

2.1 Cas d'une variable quantitative discrète

On introduit tout d'abord la notion de tableau statistique, façon synthétique de présenter les données après leur rangement par ordre croissant. Ce tableau fait intervenir les notions assez élémentaires d'effectif, de fréquence (ou pourcentage), d'effectif cumulé et de fréquence cumulée. Les représentations graphiques usuelles de ces variables sont le diagramme en bâtons (pour positionner les observations) et le diagramme cumulatif (pour les quantités cumulées). Enfin, les caractéristiques numériques permettant de résumer une variable quantitative discrète sont soit de tendance centrale (médiane et moyenne), soit de dispersion (variance et écart-type).

2.1.1 Introduction

En général, on appelle variable quantitative discrète une variable quantitative ne prenant que des valeurs entières (plus rarement décimales). Le nombre de valeurs distinctes d'une telle variable est habituellement assez faible (sauf exception, moins d'une vingtaine). Citons, par exemple, le nombre d'enfants dans une population de familles, le nombre d'années d'études après le bac dans une population d'étudiants...

Exemple 1 *On a noté l'âge (arrondi à l'année près) des 48 salariés d'une entreprise. Les données sont listées ci-dessous (il s'agit de données fictives).*

43 29 57 45 50 29 37 59 46 31 46 24 33 38 49 31
62 60 52 38 38 26 41 52 60 49 52 41 38 26 37 59
57 41 29 33 33 43 46 57 46 33 46 49 57 57 46 43

2.1.2 Présentation des données

Le tableau statistique

Les observations ci-dessus ne sont pas présentées de façon commode. Ainsi, la première d'entre elles, 43, figure au total 3 fois dans la liste. L'idée est de ne la faire figurer qu'une seule fois, en précisant qu'elle y est répliquée 3 fois. Si, en plus de n'être pas répétées, les différentes observations sont rangées par ordre croissant, les résultats seront bien plus commodes à lire. C'est ce que l'on fait lorsqu'on présente les données sous forme de *tableau statistique*.

On appelle donc tableau statistique un tableau dont la première colonne comporte l'ensemble des r observations distinctes de la variable X. Ces observations sont rangées par ordre croissant et non répétées; nous les noterons $\{x_i \ ; \ i = 1, \ldots, r\}$. Dans une seconde colonne, on dispose, en face de chaque valeur x_i, le nombre de réplications qui lui sont associées. Ces réplications sont appelées **effectifs** et notées n_i (ainsi, $n_i = 3$ lorsque $x_i = 43$: à l'observation 43 est associé l'effectif 3, autrement dit la valeur 43 a été observée 3 fois). Les effectifs n_i sont souvent remplacés par les quantités $f_i = \dfrac{n_i}{n}$, appelées **fréquences**, souvent exprimées en **pourcentages**, c'est-à-dire multipliées par 100 (ici, n désigne le nombre total d'observations : $n = \sum_{i=1}^{r} n_i = 48$; toujours pour $x_i = 43$, $f_i = \dfrac{3}{48} = 0,0625 = 6,25\%$).

Remarque 1 : Le symbole sigma. *Nous avons utilisé ci-dessus le symbole Σ (sigma majuscule). Il s'agit tout simplement d'une notation permettant de raccourcir certaines écritures. Ainsi, lorsqu'on fait la somme des valeurs indicées n_i (les effectifs de la série), au lieu d'écrire $n = n_1 + \cdots + n_i + \cdots + n_r$, il est plus commode d'écrire $n = \sum_{i=1}^{r} n_i$. On peut également écrire :*

$$n = \sum_{i=1}^{r} n_i = \sum_{i=1}^{i=r} n_i = \sum_{1}^{r} n_i.$$

Bien entendu, toutes ces écritures représentent la même quantité.

Les effectifs cumulés et les fréquences cumulées

Il peut être utile de compléter le tableau statistique en y rajoutant soit les effectifs cumulés, soit les fréquences cumulées. Ces quantités sont respectivement définies de la façon suivante :

$$N_i = \sum_{j=1}^{i} n_j \ ; \ F_i = \sum_{j=1}^{i} f_j.$$

Autrement dit, N_i représente le nombre d'observations inférieures ou égales à x_i et F_i leur fréquence (ou leur pourcentage si l'on considère 100 F_i). On notera que $N_r = n$ et $F_r = 1$ (bien comprendre pourquoi en se reportant au Tableau 2.1).

Illustration

Dans le tableau 2.1, on a calculé, sur les données présentées dans l'Exemple 1, les effectifs, les effectifs cumulés, les pourcentages et les pourcentages cumulés. Il est conseillé au lecteur de reprendre les calculs pour bien en comprendre le principe.

Remarque 2 *Dans la pratique, on utilise plutôt les pourcentages que les fréquences. Ensuite, il est rare de présenter à la fois les effectifs et les pourcentages (qui fournissent, pratiquement, la même information). On choisit donc entre les deux ensembles de quantités. Si l'on souhaite disposer des cumuls, on choisit de même entre effectifs cumulés et pourcentages cumulés.*

Le choix entre effectifs (resp. effectifs cumulés) et pourcentages (resp. pourcentages cumulés) est très empirique. Il semble naturel de choisir les effectifs lorsque l'effectif total n est faible et les pourcentages lorsqu'il est plus important. La limite approximative de 100 *paraît, dans ces conditions, assez raisonnable.*

2.1.3 Représentations graphiques usuelles

Pour une variable discrète, on rencontre essentiellement deux sortes de représentations graphiques qui sont, en fait, complémentaires : le diagramme en bâtons et le diagramme cumulatif (en escaliers).

x_i	n_i	N_i	$f_i(\%)$	$F_i(\%)$
24	1	1	2,08	2,08
26	2	3	4,17	6,25
29	3	6	6,25	12,50
31	2	8	4,17	16,67
33	4	12	8,33	25,00
37	2	14	4,17	29,17
38	4	18	8,33	37,50
41	3	21	6,25	43,75
43	3	24	6,25	50,00
45	1	25	2,08	52,08
46	6	31	12,50	64,58
49	3	34	6,25	70,83
50	1	35	2,08	72,91
52	3	38	6,25	79,16
57	5	43	10,42	89,58
59	2	45	4,17	93,75
60	2	47	4,17	97,92
62	1	48	2,08	100,00
Total	48	—	100,00	—

TAB. 2.1 – *Tableau statistique avec valeurs observées, effectifs, effectifs cumulés, fréquences et fréquences cumulées.*

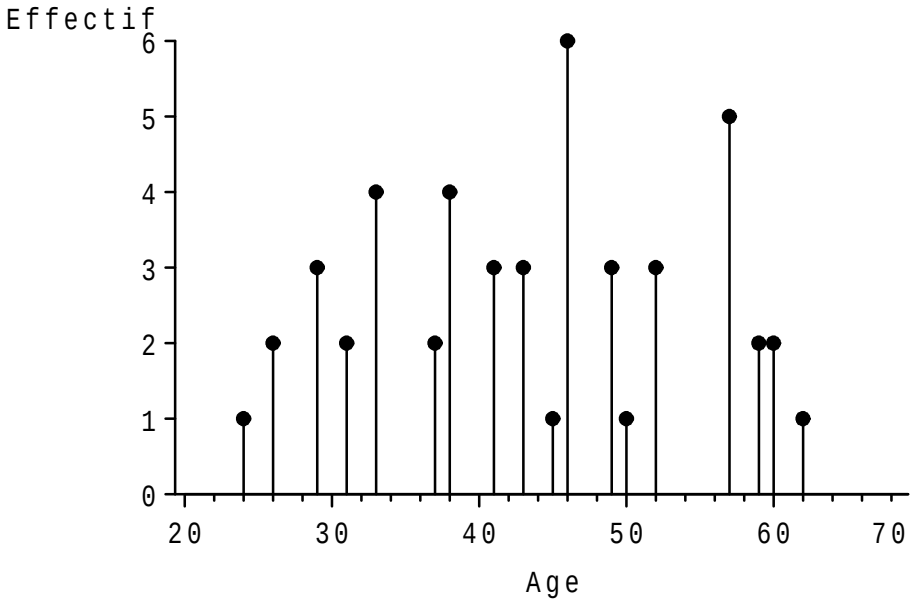

FIG. 2.1 – *Diagramme en bâtons*

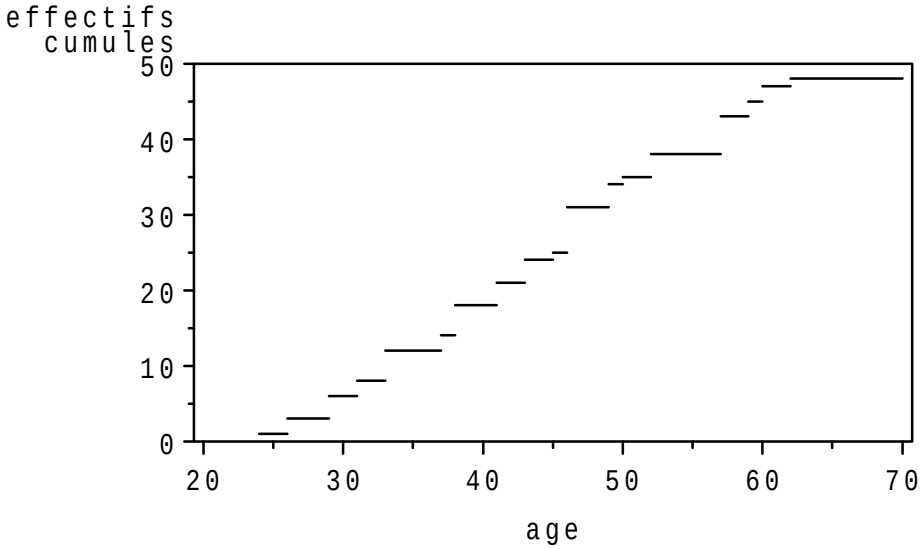

FIG. 2.2 – *Diagramme cumulatif*

Le diagramme en bâtons

Il permet de donner une vision d'ensemble des observations réalisées. La Figure 2.1 donne le diagramme en bâtons des données de l'Exemple 1.

Ce diagramme comporte donc un axe horizontal (l'abscisse), sur lequel figurent les observations de la variable considérée (ici les âges), et un axe vertical (l'ordonnée), sur lequel figurent les effectifs (mais on aurait pu, tout aussi bien, y faire figurer les fréquences ou les pourcentages ; cela n'aurait rien changé à l'allure du graphique, puisque les fréquences sont, par définition, proportionnelles aux effectifs). En face de chaque observation figure un trait vertical (un bâton), dont la hauteur est proportionnelle à l'effectif (ou à la fréquence, ou au pourcentage) correspondant.

Le diagramme cumulatif

Ce second type de graphique sert à visualiser les effectifs cumulés, ou encore les fréquences ou les pourcentages cumulés. Il permet ainsi de déterminer simplement le nombre, ou la proportion, d'observations inférieures ou égales à une valeur donnée de la série.

La Figure 2.2 donne le diagramme cumulatif relatif à l'Exemple 1.

On voit que c'est ce qu'on appelle une fonction en escaliers (la raison en est évidente !). En abscisse figurent, encore une fois, les observations de la variable considérée, tandis qu'en ordonnée figurent maintenant les effectifs cumulés, les fréquences cumulées ou les pourcentages cumulés (là encore, l'allure générale du graphique est la même, quel que soit le choix effectué). Dans un premier temps, en face de chaque observation figure un point dont l'ordonnée est égale à l'effectif cumulé correspondant. Ensuite, pour compléter le graphique, les différents points sont joints par des segments horizontaux puisque, par définition, le cumul reste constant entre deux observations (la variable considérée est *discrète*, ce qui signifie qu'entre deux entiers il n'y a pas d'observation possible).

2.1.4 Notion de quantile et applications

Définition

On a vu que la fréquence cumulée F_i ($0 \leq F_i \leq 1$) donne la proportion d'observations inférieures ou égales à x_i. Une approche complémentaire consiste à se donner, a priori, une valeur α, comprise

entre 0 et 1, et à rechercher x_α, valeur telle qu'une proportion α des observations lui sont inférieures ou égales (autrement dit, x_α vérifie $F(x_\alpha) \simeq \alpha$). La valeur x_α (qui n'est pas nécessairement unique) est appelée quantile (ou *fractile*) d'ordre α de la série. Les quantiles les plus utilisés sont associés à certaines valeurs particulières de α.

Autrement dit, le quantile d'ordre α, noté x_α, est tel que la proportion des observations qui lui sont inférieures ou égales vaut α, tandis que la proportion des observations qui lui sont supérieures vaut $1 - \alpha$.

La médiane et les quartiles

La médiane est le quantile d'ordre $\frac{1}{2}$. Elle partage donc la série des observations en deux ensembles d'effectifs égaux.

Reprenons les données de l'Exemple 1 et la colonne des pourcentages cumulés du Tableau 2.1. Le pourcentage cumulé associé à la valeur 43 est 50. Cela signifie que 50 % des observations sont inférieures ou égales à 43 (et donc 50 % sont supérieures à cette valeur). On pourrait choisir 43 pour médiane, mais on préfèrera 44 : toute valeur supérieure à 43 et inférieure à 45 est en effet telle que la moitié (exactement) des observations lui sont inférieures et la moitié lui sont supérieures. La valeur intermédiaire 44 sera choisie ici pour médiane.

Le premier quartile est le quantile d'ordre $\frac{1}{4}$, le troisième quartile celui d'ordre $\frac{3}{4}$ (le second quartile est donc confondu avec la médiane). On voit donc que 25 % des observations sont inférieures ou égales au premier quartile, tandis que 75 % lui sont supérieures. Pour le troisième quartile, les proportions s'inversent : 75 % des valeurs lui sont inférieures ou égales, tandis que 25 % lui sont supérieures.

Dans l'exemple, on a obtenu $x_{\frac{1}{4}} = 35$ avec un raisonnement analogue à celui fait pour déterminer la médiane : le pourcentage cumulé associé à 33 est 25 ; toute valeur strictement comprise entre 33 et 37 (la valeur observée suivante) partage donc la série en deux parties représentant exactement 25 % et 75 % des observations, respectivement. La valeur intermédiaire 35 est alors très naturelle pour $x_{\frac{1}{4}}$.

Par contre, comme il n'y a pas de valeurs dont le pourcentage cumulé soit exactement égal à 75, on prend la valeur pour laquelle il est juste supérieur (52, de pourcentage cumulé 79,16) pour valeur de $x_{\frac{3}{4}}$.

En fait, avec une variable discrète, la détermination des quantiles est souvent approximative comme on peut le constater avec cet exemple.

Les autres quantiles

Les *déciles* et les *centiles* sont également d'usage relativement courant. Il existe 9 déciles qui partagent l'ensemble des observations en 10 parties d'égale importance (chacune contient 10 % des observations) et 99 centiles qui la partagent de même en 100 parties d'effectifs égaux. Il va de soi qu'un grand nombre d'observations est nécessaire pour que les déciles (et, à plus forte raison, les centiles) aient un sens. Ainsi, lorsqu'on étudie l'écart des revenus entre les plus riches et les plus pauvres au sein d'une population, on utilise en général le rapport du neuvième décile au premier décile de la série des revenus.

Le diagramme en boîte (ou boîte-à-moustaches, ou "box-and-whisker plot")

Il s'agit d'un graphique très simple qui résume la série à partir de ses valeurs extrêmes, de ses quartiles et de sa médiane. La Figure 2.3 donne le diagramme en boîte de l'Exemple 1.

Dans ce graphique, il n'y a en fait qu'une échelle verticale, pour les observations de la variable (aucune échelle ne figure sur l'axe horizontal). La "boîte" est la partie du graphique comprise entre les premier et troisième quartiles. La médiane, nécessairement située à l'intérieur de la boîte, est représentée par un trait horizontal. Dans les parties basse et haute du graphique figurent les "moustaches", joignant le minimum au premier quartile (en bas) et le troisième quartile au maximum (en haut). Ce graphique fournit donc très simplement quelques valeurs caractéristiques de la variable considérée.

On notera qu'il est encore possible de réaliser le diagramme en boîte de façon horizontale (on permutte les axes horizontaux et verticaux), ce que font certains logiciels.

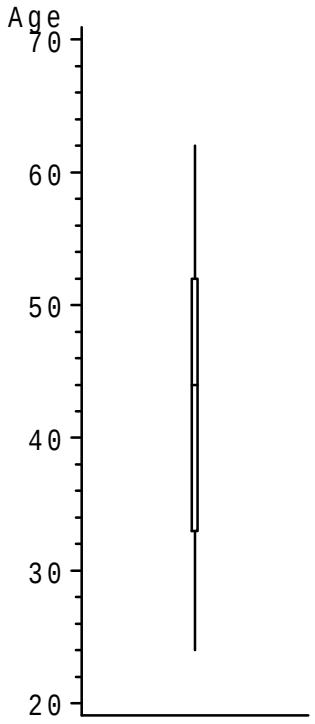

FIG. 2.3 – *Diagramme en boîte*

2.1.5 Caractéristiques numériques

Les caractéristiques (ou résumés) numériques introduites ici servent à synthétiser la série étudiée au moyen d'un petit nombre de valeurs. On distingue essentiellement des caractéristiques de **tendance centrale** (ou encore de *position* ou de *localisation*) et des caractéristiques de **dispersion**.

Caractéristiques de tendance centrale

Leur objectif est de fournir un ordre de grandeur de la série étudiée, c'est-à-dire d'en situer le centre, le milieu. Les deux caractéristiques les plus usuelles sont :

– la *médiane*, déjà définie ;
– la *moyenne* (ou moyenne arithmétique).

Tout le monde a déjà calculé et utilisé, à diverses reprises, une moyenne. La signification de cette caractéristique de tendance centrale est donc à peu près évidente pour tout un chacun. Nous en donnons ci-dessous la définition générale, mathématique (il s'agit de l'écriture formelle représentant le calcul que l'on est habitué à faire lorsqu'on détermine une moyenne).
Formule de la moyenne pour une variable quantitative discrète :

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{r} n_i x_i = \sum_{i=1}^{r} f_i x_i.$$

Chaque observation x_i est donc multipliée par l'effectif correspondant ($n_i x_i = x_i + \cdots + x_i$, n_i fois) ; on somme ensuite les quantités obtenues et l'on divise le résultat par n (l'effectif total).

Le calcul de la moyenne des données de l'Exemple 1 est précisé plus loin dans ce paragraphe.

Remarque 3 *Si l'on doit donner un indicateur de tendance centrale unique pour caractériser le centre d'une série d'observations, on doit choisir entre la moyenne et la médiane. On retiendra que la moyenne est l'indicateur le plus naturel, le plus connu, et donc le plus utilisé dans la pratique. De plus, la moyenne est définie par une formule mathématique donnant un résultat sans ambiguïté, ce qui n'est pas le cas de la médiane. Par contre, il faut noter que la signification de la*

moyenne peut être faussée par quelques valeurs très grandes ou très petites par rapport à la plupart des observations, ce qui n'est pas le cas de la médiane. On préfère donc cette dernière lorsqu'on rencontre la situation évoquée ci-dessus, en particulier dans le cas de séries très dissymétriques.

Caractéristiques de dispersion

Elles servent à préciser la variabilité de la série, c'est-à-dire à résumer l'éloignement de l'ensemble des observations par rapport à leur tendance centrale. On trouve diverses caractéristiques de dispersion, certaines étant plus courantes que d'autres.

Citons, tout d'abord, quelques caractéristiques assez peu utilisées, mais qu'il est néanmoins bon de connaître.

- L'*étendue* : $x_r - x_1$. Écart entre la plus grande et la plus petite des observations, cette caractéristique est totalement liée à ces 2 valeurs extrêmes et donc peu "fiable". Néanmois, elle donne une première idée de la dispersion des observations.
- L'*intervalle interquartiles* : $x_{\frac{3}{4}} - x_{\frac{1}{4}}$. Écart entre le troisième et le premier quartiles, c'est donc la longueur de la boîte dans le diagramme en boîte. Cette caractéristique est intéressante, car totalement indépendante des valeurs extrêmes (et donc très fiable).

Comme indiqué, les deux caractéristiques ci-dessus sont intéressantes mais, en pratique, la caractéristique de loin la plus utilisée est l'**écart-type**, racine carrée positive de la **variance**.

Formules donnant la variance :

$$\begin{aligned} \mathrm{var}(X) = s_X^2 &= \frac{1}{n}\sum_{i=1}^{r} n_i (x_i - \overline{x})^2 \\ &= [\frac{1}{n}\sum_{i=1}^{r} n_i (x_i)^2] - (\overline{x})^2. \end{aligned}$$

La variance est, par définition, la moyenne des carrés des écarts à la moyenne (première formule ci-dessus). On peut aussi retenir son expression sous la forme suivante : c'est la moyenne des carrés moins le carré de la moyenne (pour le vérifier, il faut développer le carré, puis la sommation : c'est la seconde formule ci-dessus). L'écart-type de X, racine carrée positive de la variance, sera donc noté s_X.

Quelques commentaires. Malgré les apparences, la formule de définition de la variance est très *naturelle*. En effet, un indicateur de dispersion doit résumer l'"écartement" des observations autour de leur tendance centrale. Si l'on prend la moyenne comme caractéristique de tendance centrale (c'est la plus naturelle et la plus utilisée), on pense immédiatement à la moyenne des écarts à la moyenne $x_i - \overline{x}$ comme indicateur de dispersion :

$$\frac{1}{n}\sum_{i=1}^{r} n_i (x_i - \overline{x}).$$

Il se trouve que cette quantité est toujours nulle (quelle que soit la série étudiée), les écarts $x_i - \overline{x}$ négatifs compensant exactement les écarts positifs. Pour contourner cette difficulté, il n'y a que deux solutions : soit prendre ces écarts en valeur absolue, et l'on obtient ce que l'on appelle l'écart-moyen à la moyenne, peu commode dans les calculs ; soit les prendre au carré, et l'on obtient alors la variance. Il faut ensuite prendre la racine carrée de la variance pour retrouver un indicateur s'exprimant dans la même unité que les observations : c'est l'écart-type.

On retiendra qu'une variance est toujours positive ou nulle puisque c'est une moyenne de carrés. Elle ne peut être nulle que si tous ces carrés sont eux-même nuls, c'est-à-dire si tous les x_i sont égaux à $\overline{x}$, autrement dit si toutes les observations sont égales (il n'y a plus alors de variabilité et la variance est nulle). Par ailleurs, on choisit pour l'écart-type la valeur positive de la racine carrée de la variance, car un indicateur de dispersion est, par principe, toujours positif.

Illustration

En utilisant toujours l'exemple 1, on a calculé :

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{r} n_i x_i = \frac{2094}{48} = 43,625 \simeq 43,6 \text{ ans} ;$$

SAU (en ha)	pourcentages	pourc. cumulés	x_i
moins de 5	24,0	24,0	2,5
de 5 à 10	10,9	34,9	7,5
de 10 à 20	17,8	52,7	15,0
de 20 à 35	20,3	73,0	27,5
de 35 à 50	10,2	83,2	42,5
plus de 50	16,8	100,0	100,0
Total	100,0	—	—

TAB. 2.2 – *Répartition des exploitations agricoles selon la S.A.U.*

$$s_X^2 = \frac{1}{n}\sum_{i=1}^{r} n_i (x_i)^2 - (\overline{x})^2 = \frac{96620}{48} - (43,625)^2 \simeq 109,7760 ;$$
$$s_X = \sqrt{s_X^2} \simeq 10,5 \text{ ans.}$$

Nous invitons le lecteur à refaire ces calculs (en s'aidant d'une calculette).

2.2 Cas d'une variable quantitative continue

Comme dans le cas discret, le tableau statistique permet de présenter de manière synthétique les observations d'une variable quantitative continue. Par contre, les graphiques changent dans ce cas : la répartition des observations est représentée au moyen d'un histogramme, tandis que leur cumul est maintenant représenté au moyen de la courbe cumulative. Enfin, les caractéristiques numériques qui résument ces variables sont les mêmes que dans le cas discret, mais leur calcul nécessite quelques adaptations.

2.2.1 Généralités

Une **variable quantitative** est dite **continue** lorsque les observations qui lui sont associées ne sont pas des valeurs précises, mais des intervalles. Cela signifie que, dans ce cas, l'ensemble des valeurs possibles de la variable étudiée a été divisé en r intervalles contigus appelés **classes**.

En général, les deux raisons principales qui peuvent amener à considérer comme continue une variable quantitative sont le grand nombre d'observations distinctes (un traitement en discret serait, dans ce cas, peu commode) et le caractère "sensible" d'une variable (lors d'une enquête, il est moins gênant de demander à des individus leur classe de salaire que leur salaire précis ; même chose pour l'âge). Deux exemples de variables quantitatives fréquemment considérées comme continues sont ainsi le revenu et l'âge (pour un groupe d'individus).

Nous noterons $(b_0 \ ; \ b_1) \cdots (b_{r-1} \ ; \ b_r)$ les classes considérées. Les quantités b_{i-1} et b_i sont appelés les **bornes** de la $i^{\text{ième}}$ classe ; $\dfrac{b_{i-1}+b_i}{2}$ est le **centre** de cette classe, en général noté x_i ; enfin, $(b_i - b_{i-1})$ en est l'**amplitude**, en général notée a_i.

2.2.2 Présentation des données

On utilise encore un tableau statistique analogue à celui vu dans la section précédente, en disposant maintenant dans la première colonne les classes rangées par ordre croissant. Les notions d'effectifs, de fréquences (ou pourcentages), d'effectifs cumulés et de fréquences (ou pourcentages) cumulées sont définies de la même façon que dans le cas discret.

Exemple 2 *Le Tableau 2.2 donne, pour l'année 1987, la répartition des exploitations agricoles françaises selon la S.A.U. (surface agricole utilisée) exprimée en hectares (Tableaux Economiques de Midi-Pyrénées, INSEE, 1989, p. 77). La S.A.U. est ici une variable quantitative continue comportant 6 classes.*

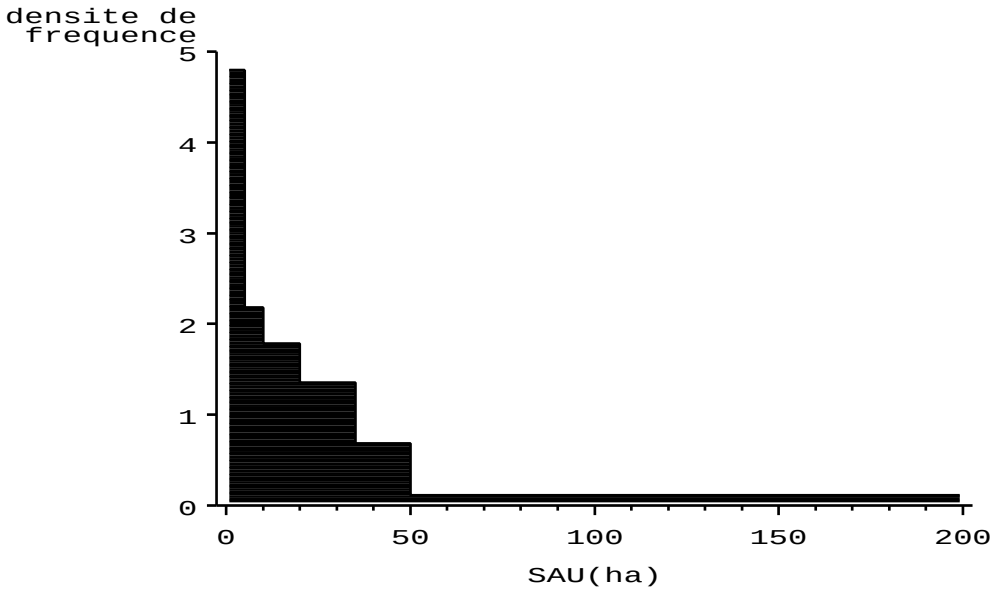

FIG. 2.4 – *Histogramme*

2.2.3 Représentations graphiques

Les deux graphiques usuels remplaçant respectivement dans ce cas le diagramme en bâtons et le diagramme cumulatif sont l'histogramme et la courbe cumulative.

L'histogramme

De façon élémentaire, on peut dire que l'histogramme est un graphique qui juxtapose divers rectangles, un pour chaque classe. Un axe horizontal sert à représenter les bornes des classes de la variable considérée. Chaque classe est alors représentée par un rectangle dont la base est délimitée par les bornes correspondantes et dont la hauteur est ce que l'on appelle la **densité d'effectif** (ou de fréquence, ou de pourcentage). La densité d'effectif de la classe $(b_{i-1}\ ;\ b_i)$ est égale au rapport entre son effectif n_i et son amplitude $a_i = b_i - b_{i-1}$. Par conséquent, la surface du rectangle représentant la classe $(b_{i-1}\ ;\ b_i)$ est égale à son effectif n_i (le but de l'histogramme est en effet de mettre graphiquement en évidence la répartition des effectifs des classes considérées). Ainsi, dans l'Exemple 2, la densité de fréquence de la seconde classe vaut : $\dfrac{10,9}{10-5} = \dfrac{10,9}{5} = 2,18$. La surface du rectangle associé vaut : $(10-5) \times 2,18 = 10,9$.

La Figure 2.4 donne l'histogramme correspondant aux données de l'Exemple 2.

La courbe cumulative

Donnons en encore une description élémentaire. Pour la variable quantitative continue étudiée, chaque classe considérée doit d'abord être représentée par un point unique dont l'abscisse est la borne supérieure de la classe et l'ordonnée est l'effectif (ou la fréquence, ou le pourcentage) cumulé de cette classe.

Remarque 4 *En fait, comme on ne connaît dans ce cas que l'appartenance des observations aux classes, on ne dispose pas d'information sur l'évolution des effectifs cumulés à l'intérieur des classes. On connaît seulement le niveau de l'effectif cumulé en la borne supérieure de chaque classe. Pour chacune d'elles, c'est cette information qui est représentée par le point décrit ci-dessus.*

La courbe cumulative est alors la courbe joignant les points en question. Elle représente donc l'évolution des effectifs (ou fréquences, ou pourcentages) cumulés, comme le faisait le diagramme cumulatif dans le cas discret. La modification de l'allure de la courbe correspond à la modification de la nature des données.

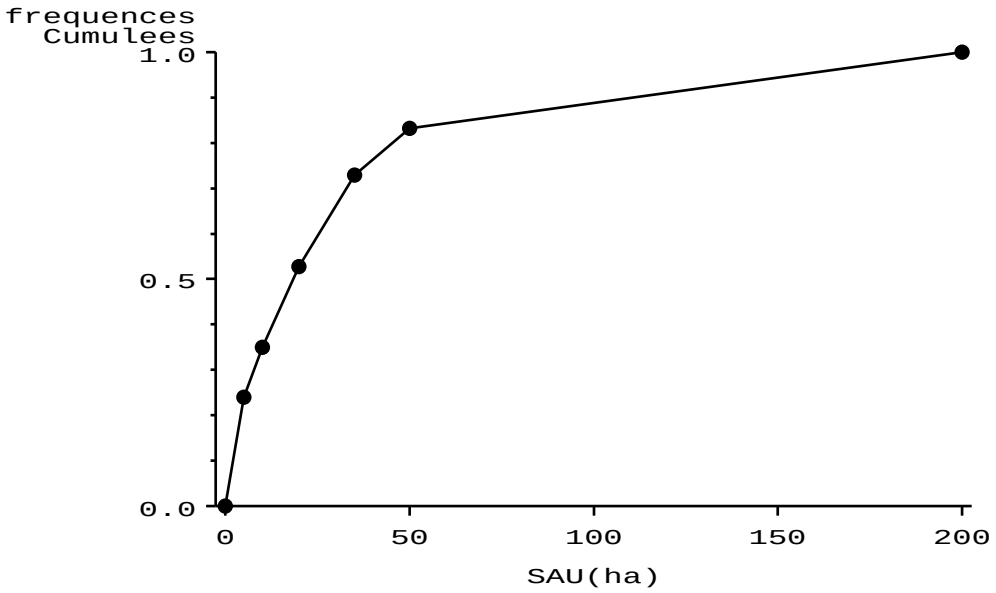

FIG. 2.5 – *Courbe cumulative*

La Figure 2.5 donne la courbe cumulative relative à l'Exemple 2. On notera que, dans cet exemple, comme c'est souvent le cas avec une variable quantitative continue, il a fallu fixer arbitrairement la borne inférieure de la première classe (il était naturel ici de prendre $b_0 = 0$) ainsi que la borne supérieure de la dernière classe (on a choisi $b_6 = 200$, mais d'autres choix étaient possibles).

2.2.4 Détermination des quantiles

Les quantiles x_α d'une variable continue peuvent être déterminés de façon directe à partir de la courbe cumulative. Cela signifie que, par le calcul, on doit commencer par déterminer la classe dans laquelle se trouve le quantile cherché, puis le déterminer dans cette classe par interpolation linéaire (voir l'illustration plus loin), à moins qu'on ne se contente du centre de classe.

On peut également réaliser, dans ce cas, le diagramme en boîte à partir de la médiane, des quartiles, de b_0 et de b_r.

2.2.5 Détermination des autres caractéristiques numériques

La moyenne, la variance et l'écart-type d'une variable continue se déterminent de la même manière que dans le cas discret ; dans les formules, on doit prendre pour valeur x_i les centres de classes au lieu des observations (qui ne sont pas connues). Les valeurs obtenues pour ces caractéristiques sont donc assez approximatives ; cela n'est pas gênant dans la mesure où le choix de traiter une variable quantitative comme continue correspond à l'acceptation d'une certaine imprécision dans le traitement statistique.

2.2.6 Illustration

La médiane de la variable présentée dans l'exemple 2 se situe dans la classe (10 ; 20), puisque le pourcentage cumulé de cette classe (52,7) est le premier à dépasser 50 (le pourcentage cumulé 50 étant celui qui, par définition, correspond à la médiane). On peut alors se contenter de prendre le centre de cette classe (15) pour médiane, mais on peut aussi la déterminer de façon plus précise en faisant l'interpolation linéaire suivante au sein de cette classe (l'indice i ci-dessous désigne en fait la troisième classe) :

$$x_{\frac{1}{2}} \quad = \quad b_{i-1} + a_i \frac{50 - F_{i-1}}{F_i - F_{i-1}} = b_{i-1} + a_i \frac{50 - F_{i-1}}{f_i}$$

$$\begin{aligned} &= 10 + 10\frac{50 - 34,9}{52,7 - 34,9} = 10 + 10\frac{15,1}{17,8} \\ &\simeq 10 + 10 \times 0,85 = 18,5 \text{ ha.} \end{aligned}$$

La moyenne vaut :

$$\overline{x} = \sum_{i=1}^{6} f_i x_i = \frac{3080,5}{100} \simeq 30,8 \text{ ha.}$$

Remarque 5 *Dans cet exemple, il convient de noter trois choses :*
- *tout d'abord, pour le calcul de la moyenne, nous avons choisi* $x_6 = 100$*, plutôt que* 125*, car cette valeur nous a semblé plus proche de la réalité ;*
- *ensuite, il se trouve que, dans ce cas, on peut calculer la* vraie *valeur de la moyenne, connaissant la SAU totale en France (31 285 400 ha) et le nombre total d'exploitations agricoles (981 720) ; on obtient 31,9 ha, ce qui signifie que l'approximation obtenue ici est très correcte ;*
- *enfin, le fait que la médiane soit sensiblement plus faible que la moyenne caractérise les séries fortement concentrées sur les petites valeurs, comme c'est le cas dans l'exemple 2.*

2.3 Cas d'une variable qualitative

C'est encore le tableau statistique qui permet de présenter les données dans ce cas, même si les quantités cumulées n'ont souvent plus de signification. Les graphiques relatifs aux variables qualitatives sont assez particuliers, à cause de la nature (non numérique) de ces variables. Toujours pour la même raison, il n'existe pas de résumé numérique pour une variable qualitative (qui n'est pas numérique, par définition).

2.3.1 Variables nominales et variables ordinales

Par définition, les observations d'une variable qualitative ne sont pas des valeurs numériques, mais des caractéristiques, appelées **modalités**. Lorsque ces modalités sont naturellement ordonnées (par exemple, la mention au bac dans une population d'étudiants), la variable est dite **ordinale**. Dans le cas contraire (par exemple, la profession dans une population de personnes actives) la variable est dite **nominale**.

2.3.2 Traitements statistiques

Il est clair qu'on ne peut pas envisager de calculer des caractéristiques numériques avec une variable qualitative (qu'elle soit nominale ou ordinale). Dans l'étude statistique d'une telle variable, on se contentera donc de faire des tableaux statistiques et des représentations graphiques. Encore faut-il noter que les notions d'effectifs cumulés et de fréquences cumulées n'ont de sens que pour des variables ordinales (elles ne sont pas définies pour les variables nominales).

Exemple 3 *Le tableau ci-dessous donne la répartition de la population active occupée (ayant effectivement un emploi) selon la CSP (catégorie socioprofessionnelle), en France, en mars 1988 (Tableaux de l'Économie Française, INSEE, 1989, p. 59).*

CSP	effectifs en milliers	fréquences (%)
1. agriculteurs exploitants	1312	6,1
2. artisans, commerçants, chefs d'entreprises	1739	8,1
3. cadres, professions intellectuelles supérieures	2267	10,6
4. professions intermédiaires	4327	20,1
5. employés	5815	27,0
6. ouvriers	6049	28,1
Total	21509	100,0

2.3.3 Représentations graphiques

Les représentations graphiques que l'on rencontre avec les variables qualitatives sont assez nombreuses. Les trois plus courantes, qui sont aussi les plus appropriées, sont :

- le *diagramme en colonnes,*
- le *diagramme en barre,*
- le *diagramme en secteurs.*

Les Figures 2.6, 2.7 et 2.8 présentent chacun de ces trois graphiques sur les données de l'Exemple 3. Le principe général de ces trois graphiques est le même : les différentes modalités de la variable qualitative sont représentées par des parties du graphique dont la surface est proportionnelle à l'effectif (ou la fréquence, ou le pourcentage) correspondant. Cette proportionnalité est très claire à voire dans le diagramme en colonnes, elle l'est moins dans les deux autres diagrammes.

On notera ici que ces graphiques (en particulier le diagramme en colonnes et le diagramme en secteurs) sont conçus pour être des graphiques plans, sans épaisseur pour les colonnes ou les secteurs. Nous ne pouvons donc que déconseiller vivement les graphiques réalisés par certains logiciels (tels qu'Excel) qui, au nom d'un parti pris esthétique erroné, utilisent de façon plus ou moins systématique une épaisseur en perspective, donc une impression de volume, faussant ainsi la vision du graphique, les volumes des secteurs n'étant plus proportionnels aux effectifs. **Attention : de tels graphiques sont faux.**

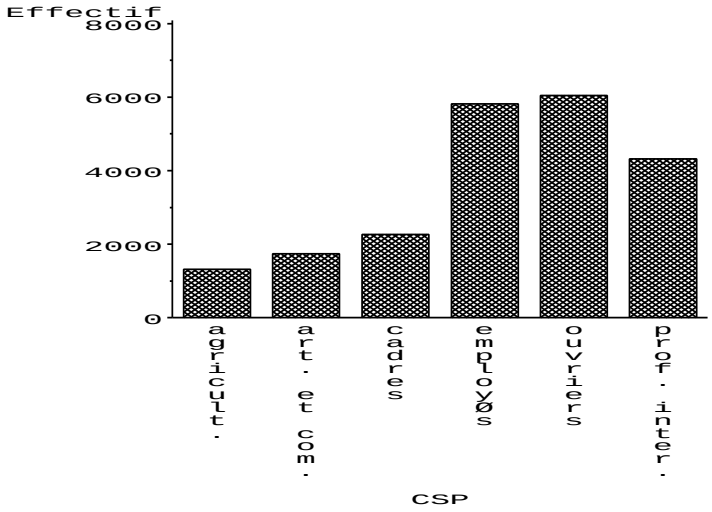

FIG. 2.6 – *Diagramme en colonnes*

FIG. 2.7 – *Diagramme en barre*

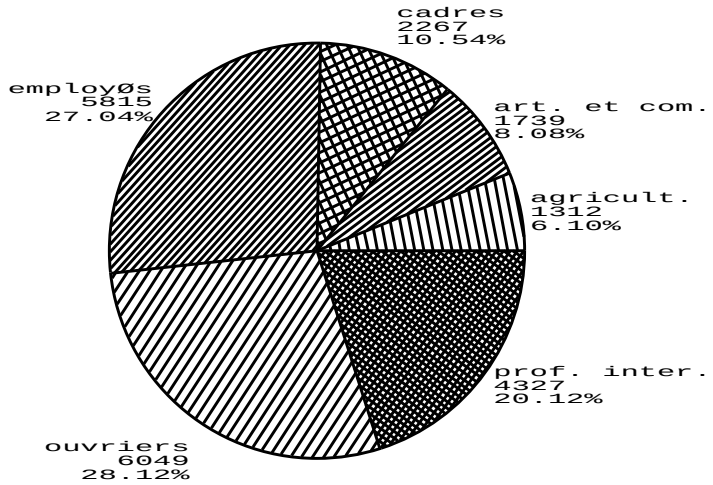

FIG. 2.8 – *Diagramme en secteurs*

Chapitre 3

Statistique descriptive bidimensionnelle

On s'intéresse maintenant à l'étude simultanée de deux variables, notées X et Y, observées sur le même échantillon (le même ensemble d'individus). L'objectif essentiel des méthodes présentées dans ce chapitre est de mettre en évidence une éventuelle variation simultanée des deux variables, que nous appellerons alors liaison. *C'est en effet l'un des objectifs fondamentaux de la statistique que de mettre en évidence des liaisons entre variables, ces liaisons exprimant certaines relations entre les phénomènes représentés par ces variables. Par exemple, dans un groupe d'hommes adultes, on peut penser qu'il existe une liaison entre la taille et le poids.*

Dans certains cas, une liaison peut être considérée a priori comme causale, *une variable expliquant l'autre (c'est la cas dans l'exemple ci-dessus, puisqu'on peut penser que la taille explique, au moins partiellement, le poids) ; dans d'autres, ce n'est pas le cas et les deux variables jouent alors des rôles symétriques. Dans la pratique, il conviendra de bien différencier les deux situations.*

Pour étudier la liaison entre deux variables, nous allons introduire d'une part des graphiques spécifiques, d'autre part des caractéristiques numériques exprimant cette liaison (la notion de présentation des données, autre élément de la description statistique, est moins fondamentale ici et n'interviendra que dans le cas de deux variables qualitatives). Les notions introduites dépendant de la nature des variables considérées (quantitatives ou qualitatives), ce chapitre se décompose en trois sections. On notera que les variables quantitatives intervenant ici seront obligatoirement discrètes, car les valeurs exactes de ces variables sont nécessaires pour les calculs que l'on va introduire.

Dans la première section, consacrée à l'étude simultanée de deux variables quantitatives (discrètes), nous allons ainsi introduire le graphique appelé nuage de points et les notions de covariance, de coefficient de corrélation linéaire et de régression linéaire. Dans la deuxième section, consacrée au cas d'une variable quantitative et d'une autre qualitative, nous utiliserons les diagrammes en boîtes parallèles et nous introduirons le rapport de corrélation. Enfin, dans la troisième section, pour l'étude simultanée de deux variables qualitatives, outre la présentation particulière des données sous forme de table de contingence, nous utiliserons les diagrammes de profils et nous introduirons l'indicateur de liaison khi-deux et quelques-uns de ses dérivés.

3.1 Deux variables quantitatives

Pour étudier la liaison entre deux variables quantitatives (discrètes), on commence par faire un graphique du type nuage de points. La forme générale de ce graphique indique s'il existe ou non une liaison entre les deux variables.

Pour préciser les choses, on calcule ensuite un indicateur de liaison. Pour cela, il faut d'abord introduire la covariance, généralisation bidimensionnelle de la variance. Comme elle dépend des unités de mesure des deux variables considérées, on doit la rendre intrinsèque en la divisant par le produit des écarts-types. On définit ainsi le coefficient de corrélation linéaire, indicateur de liaison cherché. Il est toujours compris entre -1 et $+1$, son signe indique le sens de la liaison, tandis que sa valeur absolue en indique l'intensité.

En complément, on explique ce qu'est la régression linéaire d'une variable sur une autre. Lorsqu'il existe une liaison causale entre les deux variables considérées, la régression linéaire permet d'approcher la variable réponse par une fonction de la variable causale.

3.1.1 Les données

Redisons ici que les variables quantitatives considérées dans ce chapitre seront toujours discrètes. En effet, cela n'a pas d'intérêt de regrouper les valeurs identiques de l'une des deux variables, puisqu'elles ne correspondent pas nécessairement à des valeurs identiques de l'autre. Dans cette première section, où l'on considère deux variables quantitatives, la donnée de base est donc constituée par la série statistique brute qui correspond au fichier informatique contenant les données : n lignes pour les n individus et 2 colonnes pour les 2 variables X et Y.

Considérons tout de suite un exemple de ce type de données.

Exemple 4 *Les données ci-dessous proviennent des "Tableaux de l'Economie Française" (INSEE, 1989, p.109). Pour 51 secteurs d'activité (numérotés de 04 à 54, selon la nomenclature NAP 100), on a considéré (au 01.01.1986, en France) le nombre total d'entreprises (variable notée NB), l'effectif salarié (notée EF), et le chiffre d'affaires hors-taxes en millions de francs (notée CA). Dans le tableau ci-dessous, nous donnons les trois premières et les trois dernières lignes de l'ensemble des données.*

code	secteur	NB	EF	CA
04	production de combustibles minéraux solides et cokéfaction	19	49251	14111
05	production de pétrole et de gaz naturel	120	46594	306293
06	production et distribution d'electricité	731	129723	138389
⋮	⋮	⋮	⋮	⋮
52	industrie du caoutchouc	746	87121	37502
53	transformation des matières plastiques	3232	102437	58122
54	industries diverses	10171	84012	38071

Pour mémoire, nous indiquons ci-dessous quelques caractéristiques numériques des trois variables considérées.

variable	minimum	maximum	moyenne	écart-type
NB	11	41866	4135	7435
EF	1701	425082	92591	83832
CA	992	306293	68010	64532

3.1.2 Représentation graphique : le nuage de points

Il s'agit d'un graphique très commode pour représenter les observations simultanées de deux variables quantitatives. Il consiste à considérer deux axes perpendiculaires, l'axe horizontal représentant la variable X et l'axe vertical la variable Y, puis à représenter chaque individu observé i par le point d'abscisse x_i et d'ordonnée y_i. L'ensemble de ces points donne en général une idée assez bonne de la variation conjointe des deux variables et est appelé *nuage de points*. On notera qu'on rencontre parfois la terminologie de *diagramme de dispersion*, traduction littérale du terme anglais *scatter-plot*.

La Figure 3.1 présente le nuage de points réalisé, avec les données de l'exemple 4, en utilisant les variables CA en ordonnées et EF en abscisses. De plus, on a tracé la droite de régression de CA sur EF (voir le paragraphe 3.1.5).

Remarque 6 *Le choix des échelles à retenir pour réaliser un nuage de points peut s'avérer délicat. D'une façon générale, on distinguera le cas de variables* homogènes *(représentant la même grandeur et exprimées dans la même unité) de celui des variables* hétérogènes. *Dans le premier cas, on*

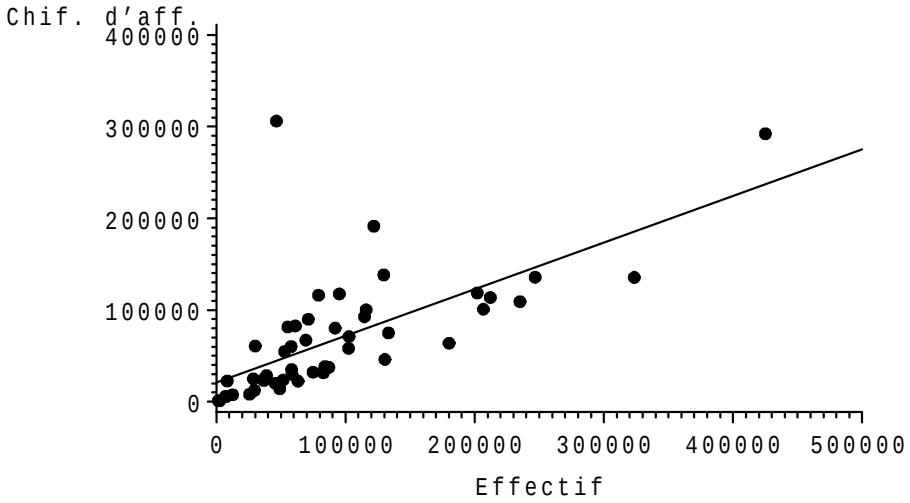

FIG. 3.1 – *Nuage de points*

choisira la même échelle sur les deux axes (qui seront donc orthonormés) ; dans le second cas, il est recommandé soit de représenter les variables centrées et réduites (voir ci-dessous) sur des axes orthonormés, soit de choisir des échelles telles que ce soit sensiblement ces variables là que l'on représente (c'est en général cette seconde solution qu'utilisent, de façon automatique, les logiciels statistiques, comme l'a fait le logiciel SAS pour la Figure 3.1).

Variables centrées et variables réduites

Si X est une variable quantitative de moyenne $\overline{x}$ et d'écart-type s_X, on appelle variable centrée associée à X la variable $X - \overline{x}$ (elle est de moyenne nulle et d'écart-type s_X), et variable centrée et réduite (ou tout simplement variable réduite) associée à X la variable $\dfrac{X - \overline{x}}{s_X}$ (elle est de moyenne nulle et d'écart-type égal à un). Une variable centrée et réduite s'exprime toujours sans unité.

À propos de la forme du nuage

Lorsque le nuage est nettement allongé (grosso-modo le long de la droite de régression), les évolutions de X et de Y sont très proches et, donc, les deux variables sont très liées (dans ce cas, il est facile de positionner un stylo le long du nuage ; c'est presque le cas dans la Figure 3.1). Au contraire, lorque le nuage est assez arrondi, il n'y a pas de relation nette entre les évolutions des deux variables et elles sont donc peu liées (il est alors difficile de positionner un stylo le long du nuage car on hésite sur la direction à choisir).

Pour plus de précisions sur cette interprétation concrète d'un nuage de points (qui est très importante dans la pratique), on se reportera au point 3.1.4.

3.1.3 La covariance et le coefficient de corrélation linéaire

L'objectif de ce paragraphe est de définir un indicateur (ou un indice) rendant compte numériquement de la manière dont les deux variables considérées varient simultanément. Autrement dit, il s'agit de définir un indice de liaison entre les deux variables considérées. Cet indice est le coefficient de corrélation linéaire ; il nécessite la définition préalable de la covariance.

La covariance : définition

La covariance généralise à deux variables la notion de variance. Sa formule de définition est la suivante :

$$\begin{aligned} \mathrm{cov}(X,Y) = s_{XY} &= \frac{1}{n}\sum_{i=1}^{n}[x_i - \overline{x}][y_i - \overline{y}] \\ &= [\frac{1}{n}\sum_{i=1}^{n} x_i y_i] - \overline{x}\ \overline{y}. \end{aligned}$$

La covariance est donc la moyenne des produits des écarts aux moyennes (dans chaque produit, chacun des deux écarts est relatif à l'une des deux variables considérées). On peut, là encore, retenir son expression sous la forme suivante : c'est la moyenne des produits moins le produit des moyennes. Comme la variance, la covariance n'a pas de signification concrète. Dans le cas de la variance, on doit passer à l'écart-type pour avoir un indicateur interprétable ; dans celui de la covariance, il faudra passer au coefficient de corrélation linéaire.

La covariance : propriétés

- La covariance est un indice *symétrique*. De façon évidente, on a $s_{XY} = s_{YX}$ (les deux variables jouent donc le même rôle dans la définition de la covariance).
- La covariance peut prendre *toute valeur réelle* (négative, nulle ou positive ; "petite" ou "grande" en valeur absolue).
- La covariance possède par ailleurs d'intéressantes propriétés mathématiques que nous ne développerons pas ici. Signalons toutefois une propriété fondamentale liant variances et covariance, appellée inégalité de Cauchy-Schwarz, et permettant de définir le coefficient de corrélation linéaire :
$$[\mathrm{cov}(X,Y)]^2 \leq \mathrm{var}(X)\ \mathrm{var}(Y).$$

Le coefficient de corrélation linéaire : définition

Il est clair que la covariance dépend des unités de mesure dans lesquelles sont exprimées les variables considérées. En ce sens, ce n'est pas un indice de liaison "intrinsèque". C'est la raison pour laquelle on définit le coefficient de corrélation linéaire (souvent appelé coefficient de Pearson, plus rarement de Bravais-Pearson), rapport entre la covariance et le produit des écarts-types. Ce coefficient caractérise, de façon intrinsèque, la liaison linéaire entre les deux variables considérées. En particulier, il ne dépend pas des unités de mesure des deux variables.

Sa définition est donc la suivante :

$$\mathrm{corr}(X,Y) = r_{XY} = \frac{s_{XY}}{s_X s_Y}.$$

Le coefficient de corrélation linéaire : propriétés

- Le coefficient de corrélation est égal à la covariance des variables centrées et réduites respectivement associées à X et Y, puisque les écarts-types de ces variables sont tous deux égaux à 1 :
$$r_{XY} = \mathrm{cov}(\frac{X - \overline{x}}{s_X}, \frac{Y - \overline{y}}{s_Y}).$$
On retrouve ainsi le fait que r_{XY} est indépendant des unités de mesure de X et de Y.
- Le coefficient de corrélation est *symétrique* : $r_{XY} = r_{YX}$ (cette propriété est évidente au vu de la définition de r_{XY}). On notera que ce coefficient étant symétrique, il ne peut pas permettre de conclure à la causalité d'une liaison, même si elle existe réellement.
- $-1 \leq r_{XY} \leq +1$. Cet encadrement de r_{XY} découle de l'inégalité de Cauchy-Schwarz. Nous indiquons ci-dessous l'importance pratique de cette propriété.

Le coefficient de corrélation linéaire : interprétation

- Le signe du coefficient indique le sens de la liaison.
Ainsi, une valeur positive indique que les deux variables ont tendance à varier dans le même sens (sur une population de ménages, penser aux revenus, variable X, et aux dépenses vestimentaires, variable Y). Au contraire, une valeur négative du coefficient de corrélation linéaire

indique que les deux variables ont tendance à varier en sens opposés (toujours sur une population de ménages, penser maintenant aux dépenses totales, variable X, et à l'épargne, variable Y).

- La valeur absolue du coefficient indique l'intensité de la liaison.
  Plus cette valeur absolue est proche de 1, plus la liaison est forte ; au contraire, plus elle est proche de 0 et plus la liaison est faible. Ainsi, un coefficient de 0,9 indique une liaison très forte ; un coefficient de 0,5 indique une liaison moyenne ; un coefficient de 0,1 indique une liaison très faible.
- Les valeurs -1 et $+1$ correspondent à une liaison linéaire parfaite entre X et Y. Dans ce cas, il existe des nombres a, b, c et d tels que : $Y = aX + b$ et $X = cY + d$. Bien entendu, un tel cas ne se rencontre en général pas avec des données réelles.

Illustration

En reprenant les données de l'Exemple 4, nous avons calculé la covariance et le coefficient de corrélation linéaire entre les variables EF et CA. On a obtenu :

$$\text{cov}(EF, CA) = 35769 \times 10^5 \ ; \ \text{corr}(EF, CA) = \frac{35769 \times 10^5}{83832 \times 64532} \simeq 0,66 \ .$$

La liaison linéaire entre les deux variables est positive et moyenne. On peut donc dire, qu'en moyenne, plus l'effectif salarié d'un secteur est important, plus le chiffre d'affaire de ce secteur l'est aussi (ce qui est logique). Mais, la liaison n'étant que moyenne, il est clair que ce phénomène n'est pas vraiment systématique. On peut donc aussi penser que des facteurs autres que l'effectif salarié influent sur le chiffre d'affaire du secteur, ce qui semble assez évident.

3.1.4 Quelques exemples

Un exemple de calcul du coefficient de corrélation linéaire

Précisons tout d'abord qu'on ne calcule plus, aujourd'hui, des variances, covariances et coefficients de corrélation linéaire "à la main", mais avec un ordinateur (et un outil tel qu'Excel, par exemple) ou, à défaut, avec une calculette. Le tableau de calcul ci-dessous n'est donc là que pour faciliter, pour le lecteur, la compréhension du mécanisme de calcul et, peut-être, celle des notions correspondantes. Notons néanmoins que la pratique des calculs "à la main", avec des feuilles du type de celle ci-dessous, était encore très courante autour de 1980.

Les données ci-dessous sont fictives (les calculs "s'arrangent bien") et correspondent à deux variables quantitatives (discrètes) X et Y observées sur 10 individus et pour lesquelles on a calculé :

$\overline{x} = 30$; $XX = X - \overline{x} = X - 30$ (variable centrée associée à X) ;

$\overline{y} = 60$; $YY = Y - \overline{y} = Y - 60$ (variable centrée associée à Y) ;

les produits $XX \times YY$, pour le calcul de la covariance, ainsi que les carrés XX^2 et YY^2, pour le calcul des variances.

On a également introduit une troisième variable quantitative Z, définie par $Z = 100 - Y$, sur laquelle on a encore calculé

$\overline{z} = 40$; $ZZ = Z - \overline{z} = Z - 40$ (variable centrée associée à Z),

ainsi que les carrés ZZ^2, pour le calcul de la variance de Z.

Voici la feuille de calculs correspondante.

X	XX	Y	YY	$XX \times YY$	XX^2	YY^2	Z	ZZ	ZZ^2
20	-10	47	-13	130	100	169	53	13	169
22	-8	50	-10	80	64	100	50	10	100
23	-7	52	-8	56	49	64	48	8	64
27	-3	53	-7	21	9	49	47	7	49
29	-1	59	-1	1	1	1	41	1	1
31	1	60	0	0	1	0	40	0	0
34	4	64	4	16	16	16	36	-4	16
37	7	67	7	49	49	49	33	-7	49
38	8	72	12	96	64	144	28	-12	144
39	9	76	16	144	81	256	24	-16	256
300	0	600	0	593	434	848	400	0	848

Pour les calculs de variances et covariances, on notera qu'on a utilisé la formule de définition ($s_X^2 = \frac{1}{10}\sum_{i=1}^{10}(x_i - \overline{x})^2 = \frac{1}{10}\sum_{i=1}^{10}(xx_i)^2$), très commode lorsque la moyenne est une valeur entière (mais pas dans les autres cas).

On a ainsi obtenu :

$$s_{XY} = \frac{593}{10} = 59,3 \ ; \quad s_X^2 = \frac{434}{10} = 43,4 \Longrightarrow s_X = 6,6 \ ; \quad s_Y^2 = \frac{848}{10} = 84,8 \Longrightarrow s_Y = 9,2$$

(Y est donc plus dispersée que X). On en déduit le coefficient de corrélation linéaire entre X et Y :

$$r_{XY} = \frac{59,3}{\sqrt{43,4 \times 84,8}} = 0,98,$$

très proche de +1, ce qui signifie que les deux variables X et Y sont positivement et très fortement corrélées.

Venons en maintenant à la variable Z, dont on aura remarqué qu'elle est liée à Y par une liaison linéaire, ce qui permet d'obtenir toutes les caractéristiques relatives à Z sans calcul (certains calculs ont néanmoins été faits ci-dessus afin de contrôler). Sa moyenne vaut $\overline{z} = 100 - \overline{y} = 100 - 60 = 40$, ce que l'on peut aussi vérifier par le calcul (précisons que la moyenne de $Y + a$ vaut $\overline{y} + a$ pour tout a). La variance de Z est la même que celle de Y, car un indicateur de dispersion ne change pas lorsqu'on ajoute une constante aux données (réfléchissez, c'est normal !) ; on retrouve donc 84,8. Enfin, nous laissons le soin au lecteur de vérifier que

$$s_{XZ} = -59,3 \ ; \quad r_{XZ} = -0,98 \ ; \quad s_{YZ} = -s_Y^2 = -84,8 \ ; \quad r_{YZ} = -1 \ ,$$

et de comprendre pourquoi tous ces résultats peuvent s'obtenir sans calcul, en utilisant les propriétés de la covariance et de la corrélation...

Quatre exemples de nuages de points

Illustrons maintenant la relation entre la forme d'un nuage de points et la liaison entre les deux variables associées au moyen de quatre exemples très simples.

On considère ci-dessous quatre jeux de données fictives, relatifs à deux variables quantitatives X et Y, sur lesquels on a d'une part calculé le coefficient de corrélation linéaire et d'autre part réalisé le graphique de type nuage de points. L'objectif est d'illustrer différents cas de figure donnant lieu à des corrélations positives ou négatives et plus ou moins importantes.

Voici tout d'abord les jeux de données. Ils ont respectivement 10, 10, 11 et 12 observations (en lignes) et systématiquement 2 variables (en colonnes), notées X et Y.

$$\text{jeu 1} = \begin{pmatrix} 5 & 6 \\ 6 & 5 \\ 6 & 7 \\ 7 & 6 \\ 7 & 7 \\ 7 & 8 \\ 8 & 7 \\ 8 & 8 \\ 9 & 8 \\ 9 & 9 \end{pmatrix} \quad \text{jeu 2} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 2 & 1 \\ 2 & 4 \\ 3 & 2 \\ 3 & 3 \\ 3 & 4 \\ 4 & 1 \\ 4 & 2 \end{pmatrix} \quad \text{jeu 3} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 2 & 1 \\ 2 & 2 \\ 2 & 3 \\ 3 & 1 \\ 3 & 2 \\ 3 & 3 \\ 8 & 8 \\ 9 & 9 \end{pmatrix} \quad \text{jeu 4} = \begin{pmatrix} 1 & 6 \\ 2 & 5 \\ 2 & 6 \\ 3 & 4 \\ 3 & 5 \\ 4 & 3 \\ 4 & 4 \\ 5 & 2 \\ 5 & 3 \\ 6 & 1 \\ 6 & 2 \\ 7 & 1 \end{pmatrix}$$

Avec tout logiciel de statistique (voire avec le tableur `Excel`), il est posible de faire calculer le coefficient de corrélation linéaire (coefficient de Pearson) entre les deux variables considérées, X et Y, sur chacun de ces jeux. Nous avons ici calculé également les deux moyennes et les deux écarts-types.

Voici les résultats :

jeu	$\overline{x}$	$\overline{y}$	σ_x	σ_y	r_{xy}
jeu 1	7.2	7.1	1.25	1.14	0.76
jeu 2	2.4	2.3	1.11	1.10	- 0.02
jeu 3	3.2	3.2	2.62	2.62	0.92
jeu 4	4.0	3.5	1.78	1.71	- 0.96

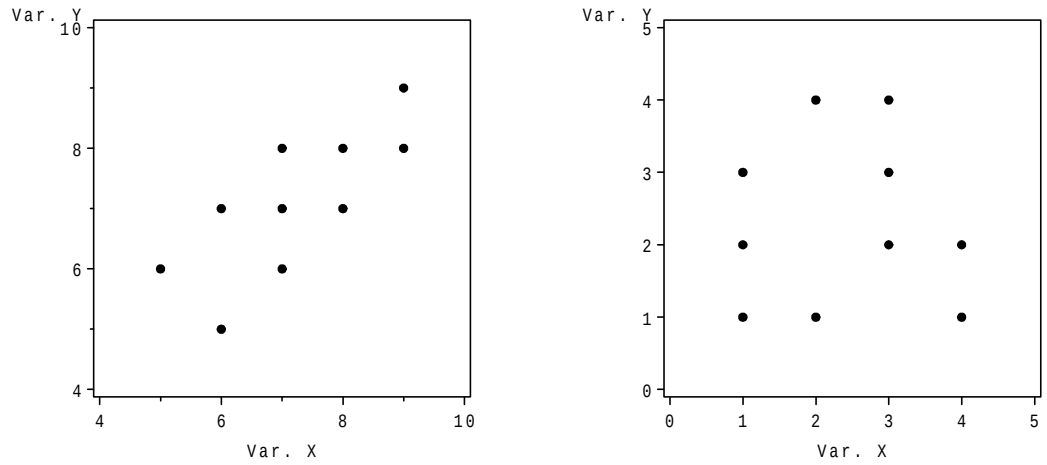

FIG. 3.2 – *jeu 1 et jeu 2*

Nous avons ensuite réalisé les graphiques *nuages de points* dans chacun des quatre cas, en utilisant le logiciel SAS. Ils sont donnés dans les Figures 3.2 et 3.3

Commentaires.

- Avec le jeu 1, on voit qu'il est assez facile de poser un stylo dans la direction du nuage de points, et que celui-ci sera "ascendant" ; il existe donc une liaison linéaire positive et assez forte entre les deux variables ; cela est confirmé par la valeur du coefficient qui vaut $+0.76$.
- Notons tout de suite qu'avec le jeu 4, le phénomène est encore plus net, donc la liaison plus forte, mais cette fois-ci "descendante" ; logiquement, le coefficient vaut -0.96.
- Par contre, avec le jeu 2, il semble très difficile de "couvrir" le nuage avec un stylo, auquel on ne saurait pas trop quelle direction donner ; il n'existe pratiquement pas de liaison linéaire entre ces deux variables et le coefficient est proche de 0, sans qu'on puisse, a priori, préciser son signe ; en fait, il vaut -0.02, sans que son signe ait ici la moindre signification.
- Enfin, avec le jeu 3, on dispose d'un cas très particulier qui montre pourquoi la construction et l'examen du nuage de points est indispensable, en plus du calcul du coefficient de corrélation linéaire ; ce coefficient est ici très fort ($+0.92$), mais la liaison est artificielle, car seulement due à deux points sur 11 (il s'agit de deux individus atypiques).

3.1.5 Régression linéaire entre deux variables

Introduction

Lorsque deux variables quantitatives sont correctement correlées ($|r_{XY}|$ voisin de 1) et que l'on peut considérer, a priori, que l'une (nous supposerons qu'il s'agit de X) est cause de l'autre (il s'agira donc de Y), il est alors assez naturel de chercher une fonction de X approchant Y, "le mieux possible" en un certain sens. La méthode statistique permettant de trouver une telle fonction s'appelle la régression de Y sur X.

Pour pouvoir mettre en œuvre une régression, il est au préalable nécessaire d'une part de définir un ensemble de fonctions dans lequel on va chercher "la meilleure", d'autre part de préciser le sens (mathématique) que l'on donne aux expressions telles que "le mieux possible" ou encore "la meilleure".

Si l'on choisit pour ensemble de fonctions celui des fonctions affines (du type $f(X) = aX + b$), on parle alors de régression linéaire (parce que le graphe d'une telle fonction est une droite). C'est le choix que l'on fait le plus fréquemment dans la pratique et c'est celui que nous ferons ici. Pour donner un sens mathématique à l'expression "le mieux possible", on utilise en général le critère appelé *des moindres carrés* car il consiste à minimiser une somme de carrés. Nous explicitons ce critère ci-dessous.

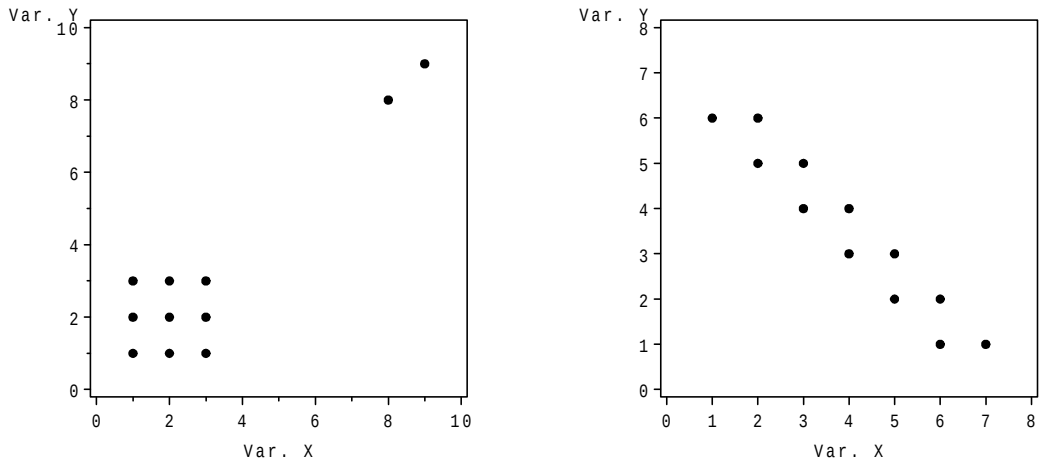

FIG. 3.3 – *jeu 3 et jeu 4*

Le critère des moindres carrés

Il consiste à minimiser la quantité suivante :

$$F(a,b) = \sum_{i=1}^{n} \{y_i - [ax_i + b]\}^2.$$

Explicitons sa signification. Si l'on applique la fonction $aX + b$ à la valeur x_i de la variable X observée sur l'individu i, on obtient $ax_i + b$. La différence entre cette valeur et celle qu'elle est censée approcher, y_i, vaut $y_i - [ax_i + b]$. Elle représente l'erreur commise en approchant y_i par $[ax_i + b]$. Pour obtenir l'erreur globale commise sur l'ensemble de l'échantillon, il faut ensuite faire la somme de l'ensemble de ces quantités. Comme dans la définition de la variance, il est nécessaire au préalable de les prendre soit en valeur absolue soit au carré, pour éviter que les erreurs positives ne compensent les erreurs négatives. L'utilisation des carrés étant nettement plus commode au niveau des calculs, c'est eux que l'on utilise en général et c'est ainsi que l'on obtient le critère ci-dessus.

Pour mémoire, on notera que $|y_i - [ax_i + b]|$ représente, dans le nuage de points associé aux observations, la distance verticale du point figurant i à la droite d'équation $Y = aX + b$.

Solution

La minimisation de F en a et b fournit la solution unique suivante :

$$\widehat{a} = \frac{s_{XY}}{s_X^2} \;\; ; \;\; \widehat{b} = \overline{y} - \widehat{a}\overline{x}.$$

Propriétés

- La droite d'équation $y = \widehat{a}x + \widehat{b}$ est appelée *droite de régression* de Y sur X ; elle passe par le *barycentre* du nuage des points, qui est le point de coordonnées $(\overline{x}, \overline{y})$ (c'est le "point moyen" du nuage).
- L'orientation de la droite de régression (ascendante ou descendante) indique le sens de la liaison entre les deux variables. En effet :
  $r_{xy} > 0 \Longleftrightarrow s_{xy} > 0 \Longleftrightarrow \hat{a} > 0$: la droite est ascendante ;
  $r_{xy} < 0 \Longleftrightarrow s_{xy} < 0 \Longleftrightarrow \hat{a} < 0$: la droite est descendante.
- Les valeurs $\widehat{y_i} = \widehat{a}x_i + \widehat{b}$ sont appelées les *valeurs prédites* ; elles ont la même moyenne $\overline{y}$ que Y.

– Les valeurs $\widehat{e_i} = y_i - \widehat{y_i}$ sont appelées les *résidus*. Ils sont de moyenne nulle et de variance $\frac{1}{n}F(\widehat{a},\widehat{b})$.

Illustration

En considérant toujours l'Exemple 4, nous avons réalisé la régression linéaire de la variable CA sur la variable EF. On a obtenu :

$$\hat{a} = \frac{cov(EF,CA)}{var(EF)} = \frac{35769 \times 10^5}{70278 \times 10^5} \simeq 0,509;$$

$$\hat{b} = \overline{CA} \ - \ \hat{a}\ \overline{EF} = 68010 - 0,509 \times 92591 \simeq 20884.$$

La droite de régression correspondante a été tracée sur la Figure 3.1.

On notera, pour terminer, que la droite de régression permet de préciser la direction générale d'un nuage de points, plus précisemment la direction dans laquelle il est le plus "allongé" (ou, mieux, le plus dispersé).

3.1.6 Généralisation : cas de plus de deux variables

Si l'on a observé simultanément plusieurs variables quantitatives sur le même échantillon (p variables, $p \geq 3$), il est possible de calculer d'une part les variances de chacune de ces variables, d'autre part les $\frac{p(p-1)}{2}$ covariances des variables prises deux à deux. L'ensemble de ces quantités peut alors être disposé dans un tableau carré ($p \times p$), symétrique, comportant les variances sur la diagonale et les covariances à l'extérieur de la diagonale (un tel tableau est appelé *matrice* en mathématiques). Ce tableau, appelé **matrice des variances-covariances**, sera notée **S**. Il est utilisé en statistique multidimensionnelle, mais n'a pas d'interprétation concrète.

De la même manière, on peut construire la matrice symétrique $p \times p$, comportant des 1 sur toute la diagonale et, en dehors de la diagonale, les coefficients de corrélation linéaire entre les variables prises deux à deux. Cette matrice est appelée **matrice des corrélations** et sera notée **R**. Elle est de lecture commode et indique quelle est la structure de corrélation des variables étudiées.

Les matrices **S** et **R** sont à la base de l'Analyse en Composantes Principales, méthode de base de la statistique multidimensionnelle.

Illustration

Nous avons repris ici encore l'Exemple 4 et calculé les matrices des variances-covariances et des corrélations entre les trois variables intervenant.

$$\mathbf{S} = \begin{bmatrix} 55284 \times 10^3 & 25084 \times 10^4 & 25156 \times 10^3 \\ 25084 \times 10^4 & 70278 \times 10^5 & 35769 \times 10^5 \\ 25156 \times 10^3 & 35769 \times 10^5 & 41644 \times 10^5 \end{bmatrix};$$

$$\mathbf{R} = \begin{bmatrix} 1 & 0,402 & 0,052 \\ 0,402 & 1 & 0,661 \\ 0,052 & 0,661 & 1 \end{bmatrix}.$$

3.2 Une variable quantitative et une qualitative

Si X est la variable qualitative à r modalités, elle définit une partition de l'ensemble des observations en r "classes". La classe courante, notée $\mathcal{C}_\ell$ ($\ell = 1, \ldots, r$), contient les individus ayant présenté la modalité x_ℓ de X. On peut alors définir moyenne et variance partielles de la variable quantitative Y au sein de chaque classe $\mathcal{C}_\ell$. La façon dont les moyennes partielles varient donne une première idée de la liaison entre X et Y.

On peut ensuite représenter, sur le même graphique, la boîte-à-moustache de Y dans chaque classe $\mathcal{C}_\ell$: on obtient le diagramme en boîtes parallèles qui précise les choses concernant la liaison entre X et Y.

Enfin, une idée encore plus précise sur cette liaison est donnée par le rapport de corrélation, indicateur compris entre 0 et 1 et d'autant plus grand que la liaison est forte.

3.2.1 Les données

Nous disposons toujours ici de deux variables mais, maintenant, l'une est quantitative et l'autre qualitative. La variable qualitative est X, supposée à r modalités notées

$$x_1, \ldots, x_\ell, \ldots, x_r.$$

La variable quantitative est Y, de moyenne $\overline{y}$ et de variance s_Y^2. On peut ainsi répartir l'ensemble des individus observés en r parties, ou sous-ensembles, en fonction de la modalité de X présentée par chaque individu. Ainsi, nous noterons $\mathcal{C}_\ell$ l'ensemble des individus de l'échantillon ayant présenté la modalité x_ℓ de X ; on obtient ainsi ce que l'on appelle une *partition* en r classes (on parle de partition lorsque chaque individu présente une modalité et une seule de la variable X). Nous noterons $n_1, \ldots, n_r$ les effectifs des différentes classes (avec toujours $\sum_{\ell=1}^{r} n_\ell = n$, où n est le nombre total d'individus observés). Par exemple, avec la variable sexe, on définit deux classes : $\mathcal{C}_1$ pour les hommes et $\mathcal{C}_2$ pour les femmmes.

On peut alors définir la moyenne et la variance partielles de Y sur chaque classe $\mathcal{C}_\ell$ de la partition ; nous les noterons respectivement $\overline{y}_\ell$ et s_ℓ^2 :

$$\overline{y}_\ell = \frac{1}{n_\ell} \sum_{i \in \mathcal{C}_\ell} y_i \, ;$$

$$s_\ell^2 = \frac{1}{n_\ell} \sum_{i \in \mathcal{C}_\ell} (y_i - \overline{y}_\ell)^2.$$

Exemple 5 *Cet exemple est extrait de l'ouvrage* Applied Multivariate Statistical Analysis*, de R.A. Johnson & D.W. Wichern (2007 pour la dernière édition). Les individus sont 19 chiens prémédiqués au pentobarbital et l'on étudie l'effet sur leur rythme cardiaque de deux facteurs (variables qualitatives explicatives) croisés. L'effet est mesuré par le temps entre deux battements de cœur successifs (variable quantitative Y, mesurée en millisecondes) et les deux facteurs, à deux niveaux chacun, sont la pression d'administration de dioxyde de carbone (CO_2), qui peut être élevée (E) ou faible (F), et la présence (1) ou l'abscence (0) d'halothane ; la variable X est celle obtenue par le croisement de ces deux facteurs ; elle est donc qualitative à 4 modalités : $x_1 = E0, x_2 = F0, x_3 = E1, x_4 = F1$. L'expérience ayant été répétée 4 fois sur chaque chien (une fois dans chacune des 4 conditions ainsi définies), on dispose donc de $n = 4 \times 19 = 76$ individus. Les données se trouvent dans le tableau ci-dessous.*

	modalité de X			
numéro du chien	x_1	x_2	x_3	x_4
1	426	609	556	600
2	253	236	392	395
3	359	433	349	357
4	432	431	522	600
5	405	426	513	513
6	324	438	507	539
7	310	312	410	456
8	326	326	350	504
9	375	447	547	548
10	286	286	403	422
11	349	382	473	497
12	429	410	488	547
13	348	377	447	514
14	412	473	472	446
15	347	326	455	468
16	434	458	637	524
17	364	367	432	469
18	420	395	508	531
19	397	556	645	625

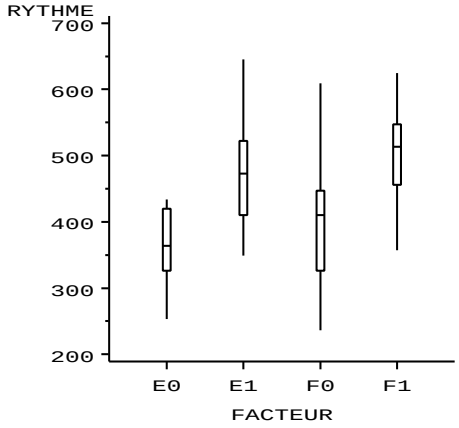

FIG. 3.4 – *Boîtes parallèles*

Nous indiquons ci-dessous les moyennes et les écarts-types partiels des 4 classes, ainsi que la moyenne et l'écart-type de la population globale.

	$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_3$	$\mathcal{C}_4$	pop. globale
moyennes	368, 2	404, 6	479, 3	502, 9	438, 8
écarts-types	51, 7	86, 9	80, 6	68, 0	91, 1

Remarque 7 *Ces données sont un peu particulières, dans la mesure où d'une part les individus sont dupliqués 4 fois et d'autre part les 4 modalités de la variable X correspondent au croisement de deux facteurs. Ainsi, d'autres traitements statistiques que ceux indiqués ici sont envisageables.*

3.2.2 Représentation graphique : les boîtes parallèles

Une façon commode de représenter les données dans le cas de l'étude simultanée d'une variable quantitative et d'une variable qualitative consiste à réaliser des boîtes parallèles. Il s'agit, sur un même graphique doté d'une échelle verticale unique, de représenter pour Y un diagramme en boîte (c'est-à-dire une boîte-à-moustaches) pour chacune des sous-populations (chacune des classes) définies par X. La comparaison de ces boîtes donne une idée assez claire de l'influence de X sur les valeurs de Y : plus les boîtes sont positionnées différemment, plus les valeurs de Y sont fonction de X, donc plus les deux variables sont liées. La Figure 3.4 donne les boîtes parallèles de l'Exemple 5. Elle indique une liaison relativement importante.

3.2.3 Formules de décomposition

Ces formules sont nécessaires pour définir un indice de liaison entre les deux variables. Elles indiquent comment se décomposent la moyenne et la variance de Y sur la partition définie par X (c'est-a-dire comment s'écrivent les caractéristiques globales en fonction de leurs valeurs partielles). Ces formules sont les suivantes :

$$\begin{aligned} \overline{y} &= \frac{1}{n}\sum_{\ell=1}^{r} n_\ell \overline{y}_\ell \,; \\ s_Y^2 &= \frac{1}{n}\sum_{\ell=1}^{r} n_\ell (\overline{y}_\ell - \overline{y})^2 \;+\; \frac{1}{n}\sum_{\ell=1}^{r} n_\ell s_\ell^2 = s_E^2 + s_R^2 \,. \end{aligned}$$

La décomposition de $\overline{y}$ est très naturelle. Le premier terme de la décomposition de s_Y^2, noté s_E^2, est appelé *variance expliquée* par la partition, c'est-à-dire par X (d'où la notation) ; on l'appelle aussi variance inter-classes, ou entre les classes. Le second terme, noté s_R^2, est appelé variance résiduelle (d'où la notation) ; on parle encore de variance intra-classes, ou à l'intérieur des classes.

On notera qu'une formule de décomposition analogue existe pour la covariance entre deux variables quantitatives.

Interprétation

La variance expliquée, s_E^2, représente ce que serait la variance de Y si, dans chaque classe $\mathcal{C}_\ell$ de la partition définie par X, Y était constante et valait $\overline{y}_\ell$. De son côté, la variance résiduelle s_R^2 représente ce qu'il reste comme variation de Y, en moyenne, dans chaque classe. Ainsi, plus s_E^2 est grande par rapport à s_R^2, plus les deux variables X et Y sont liées.

3.2.4 Le rapport de correlation

Définition

Il s'agit d'un indice de liaison entre les deux variables X et Y. Il est défini de la façon suivante :

$$c_{Y/X} = \sqrt{\frac{s_E^2}{s_Y^2}} = \frac{s_E}{s_Y} \ .$$

C'est donc la racine carrée positive de la part de variance expliquée par X.

Propriétés

- $c_{Y/X}$ n'est pas symétrique. Cette propriété est évidente, compte-tenu que X et Y ne sont pas de même nature.
- $0 \leq c_{Y/X} \leq 1$. Cet encadrement de $c_{Y/X}$ découle directement de la formule de décomposition de la variance. Les valeurs 0 et 1 ont encore une signification intéressante.
  - $c_{Y/X} = 1 \Leftrightarrow s_R^2 = 0$; dans ce cas, $s_\ell^2 = 0$ pour tout ℓ, d'après la définition de s_R^2 (la somme des carrés de ces quantités est nulle, donc chacune de ces quantités est nulle) ; par conséquent, Y est constante sur chaque $\mathcal{C}_\ell$ (puisque sa variance est nulle sur chacune de ces classes) ; dans un tel cas, la connaissance de X (donc de la classe $\mathcal{C}_\ell$ à laquelle appartient chaque individu) est suffisante pour connaître Y (qui vaut $\overline{y}_\ell$) : il y a liaison totale entre X et Y.
  - $c_{Y/X} = 0 \Leftrightarrow s_E^2 = 0 \Leftrightarrow \overline{y}_\ell = \overline{y}, \forall \ell = 1, \ldots, r$; en moyenne, X n'a aucune influence sur Y (puisque la valeur moyenne de Y est la même, quelle que soit la modalité de X) : il n'y a pas de liaison entre les deux variables.
  - On retiendra que plus $c_{Y/X}$ est grand, plus la liaison entre X et Y est forte.

Illustration

Sur l'Exemple 5, la variance totale vaut 8305,90 ; la variance inter-groupes, ou expliquée, vaut 2973,94 ; la variance intra-groupes, ou résiduelle, vaut 5331,96. On en déduit que le rapport de corrélation vaut :

$$c_{Y/X} = \sqrt{\frac{2973,94}{8305,90}} \simeq 0,60.$$

La liaison entre X et Y est donc un peu supérieure à la moyenne.

3.2.5 Un autre exemple

En fait, nous reprenons ici la série des observations de la variable X introduite dans l'exemple du point 3.1.4. On suppose que les individus 1, 2, 3, 5, et 6 appartiennent à un premier groupe (par exemple, ce sont des hommes), tandis que les autres individus 4, 7, 8, 9 et 10 appartiennent à un autre groupe (par exemple, ce sont des femmes). Nous refaisons les calculs de moyenne et de variance sur chacun des deux groupes ainsi définis, afin d'illustrer les notions présentées dans cette section.

Voici tout d'abord le tableau de calcul :

individu	X	$X-25$	$(X-25)^2$	individu	X	$X-35$	$(X-35)^2$
1	20	- 5	25	4	27	- 8	64
2	22	- 3	9	7	34	- 1	1
3	23	- 2	4	8	37	2	4
5	29	4	16	9	38	3	9
6	31	6	36	10	39	4	16
total	125	0	90	total	175	0	94

La moyenne de la variable X est égale à 25 sur le premier groupe (125/5) et à 35 sur le second groupe (175/5). On a calculé les écarts $X-25$ dans le premier groupe (de somme forcément nulle), ainsi que leurs carrés ; la somme de ces carrés vaut 90, ce qui donne une variance de 18 (90/5). Dans le second groupe, la somme correspondante des carrés vaut 94 et la variance 18,8 (94/5). Ainsi, dans le groupe 1, la moyenne partielle vaut 25 et la variance partielle 18 . Dans le groupe 2, la moyenne partielle vaut 35 et la variance partielle 18,8.

Utilisons maintenant les formules de décomposition de la moyenne et de la variance pour retrouver moyenne et variance de la variable X sur l'ensemble des 10 individus. Calcul de la moyenne globale :

$$\frac{(5\times 25)+(5\times 35)}{10}=\frac{125+175}{10}=\frac{300}{10}=30.$$

Pour retrouver la variance globale (43,4) il faut maintenant calculer deux quantités.

– La variance résiduelle, non expliquée par le découpage en deux groupes, celle que l'on trouve à l'intérieur des groupes et que nous pouvons déterminer à partir des variances partielles :

$$S_R^2=\frac{(5\times 18)+(5\times 18,8)}{10}=\frac{90+94}{10}=\frac{184}{10}=18,4\ .$$

– La variance expliquée par la partition (le découpage en deux groupes), qui se calcule à partir des moyennes (partielles et globale) de la façon suivante :

$$S_E^2=\frac{5\times(25-30)^2+5\times(35-30)^2}{10}=\frac{(5\times 25)+(5\times 25)}{10}=\frac{250}{10}=25.$$

Ajoutons maintenant $18,4$ et 25, cela nous donne $43,4$ qui est bien la variance de X sur l'ensemble des 10 individus.

On peut enfin déduire le rapport de corrélation entre la variable X et le découpage en groupes (par exemple, le sexe) : il vaut $\sqrt{\frac{25}{43,4}}\simeq 0,76$. Il y a donc une bonne liaison entre ces deux variables.

3.3 Deux variables qualitatives

Lorsqu'on étudie simultanément deux variables qualitatives, il est commode de présenter les données sous forme d'une table de contingence, synthèse des observations selon les modalités des variables qu'elles ont présentées.

À partir de cette table, on définit la notion de profil, dont on se sert pour réaliser un diagramme de profils faisant bien apparaître la liaison entre les deux variables, lorsqu'il en existe une.

Pour quantifier cette liaison, l'indicateur fondamental est le khi-deux. Toutefois, comme il n'est pas d'usage commode dans la pratique, on introduit encore les indicateurs phi-deux, T de Tschuprow et C de Cramér, liés au khi-deux. Les deux derniers sont compris entre 0 et 1, et sont d'autant plus grands que la liaison est forte, ce qui facilite leur interprétation.

3.3.1 Les données et leur présentation

On considère dans cette section deux variables qualitatives observées simultanément sur n individus. La première variable, notée X, possède r modalités notées $x_1,\ldots,x_\ell,\ldots,x_r$; la seconde, notée Y, possède c modalités notées $y_1,\ldots,y_h,\ldots,y_c$.

Le plus souvent, ces données sont présentées dans un tableau à double entrée, appelé *table de contingence*, dans lequel on dispose les modalités de X en lignes et celles de Y en colonnes.

Ce tableau est donc de dimension $r \times c$ et a pour élément général le nombre $n_{\ell h}$ d'observations conjointes des modalités x_ℓ de X et y_h de Y. Les quantités $n_{\ell h}$ sont appelées les *effectifs conjoints*.

Une table de contingence se présente donc sous la forme suivante :

	y_1	$\cdots$	y_h	$\cdots$	y_c	sommes
x_1	n_{11}	$\cdots$	n_{1h}	$\cdots$	n_{1c}	n_{1+}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_ℓ	$n_{\ell 1}$	$\cdots$	$n_{\ell h}$	$\cdots$	$n_{\ell c}$	$n_{\ell +}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_r	n_{r1}	$\cdots$	n_{rh}	$\cdots$	n_{rc}	n_{r+}
sommes	n_{+1}	$\cdots$	n_{+h}	$\cdots$	n_{+c}	n

Les quantités $n_{\ell +}(\ell = 1, \ldots, r)$ et $n_{+h}(h = 1, \ldots, c)$, appelées *effectifs marginaux*, sont définies de la façon suivante : $n_{\ell +} = \sum_{h=1}^{c} n_{\ell h}$; $n_{+h} = \sum_{\ell=1}^{r} n_{\ell h}$. Elles vérifient $\sum_{\ell=1}^{r} n_{\ell +} = \sum_{h=1}^{c} n_{+h} = n$.

De façon analogue, on définit les notions de fréquences conjointes ($f_{\ell h} = \dfrac{n_{\ell h}}{n}$) et de fréquences marginales ($f_{\ell +} = \dfrac{n_{\ell +}}{n} = \sum_{h=1}^{c} f_{\ell h}$; $f_{+h} = \dfrac{n_{+h}}{n} = \sum_{\ell=1}^{r} f_{\ell h}$) ; ces dernières vérifient : $\sum_{\ell=1}^{r} f_{\ell +} = \sum_{h=1}^{c} f_{+h} = 1$.

Exemple 6 *Dans cet exemple, on a considéré un échantillon de 797 étudiants de l'Université Paul Sabatier (Toulouse III) ayant obtenu soit le DEUG A soit le DEUG B (diplômes scientifiques de premier cycle), et uniquement ce diplôme, durant la période 1971-1983. Quatre variables ont été prises en compte, toutes qualitatives : la série de bac, à 2 modalités (C ou E, D), la mention au bac, à 4 modalités (très bien ou bien, assez bien, passable, inconnue), l'âge d'obtention du bac, à 4 modalités (moins de 18 ans, 18 ans, 19 ans, plus de 19 ans), et la durée d'obtention du DEUG, à 3 modalités (2 ans, 3 ans, 4 ans).*

Dans la table de contingence ci-dessous, on a croisé en lignes la durée d'obtention du DEUG (variable X, à $r = 3$ modalités), et en colonnes l'âge d'obtention du bac (variable Y, à $c = 4$ modalités).

	$<$ 18 ans	18 ans	19 ans	$>$ 19 ans	sommes
2 ans	84	224	73	19	400
3 ans	35	137	75	27	274
4 ans	14	59	34	16	123
sommes	133	420	182	62	797

3.3.2 Les représentations graphiques

On peut envisager, dans le cas de l'étude simultanée de deux variables qualitatives, d'*adapter* les graphiques présentés dans le cas unidimensionnel : on découpe chaque partie (colonne, partie de barre ou secteur) représentant une modalité de l'une des variables selon les effectifs des modalités de l'autre. Mais, de façon générale, il est plus approprié de réaliser des graphiques représentant des quantités très utiles dans ce cas, que l'on appelle les *profils*.

Définition des profils

On appelle $\ell^{\text{ième}}$ profil-ligne l'ensemble des fréquences de la variable Y conditionnelles à la modalité x_ℓ de X (c'est-à-dire définies au sein de la sous-population $\mathcal{C}_\ell$ de $\mathcal{C}$ associée à cette modalité). Il s'agit donc des quantités :

$$\{\frac{n_{\ell 1}}{n_{\ell +}}, \ldots, \frac{n_{\ell h}}{n_{\ell +}}, \ldots, \frac{n_{\ell c}}{n_{\ell +}}\}.$$

On définit de façon analogue le $h^{\text{ième}}$ profil-colonne :

$$\{\frac{n_{1h}}{n_{+h}}, \ldots, \frac{n_{\ell h}}{n_{+h}}, \ldots, \frac{n_{rh}}{n_{+h}}\}.$$

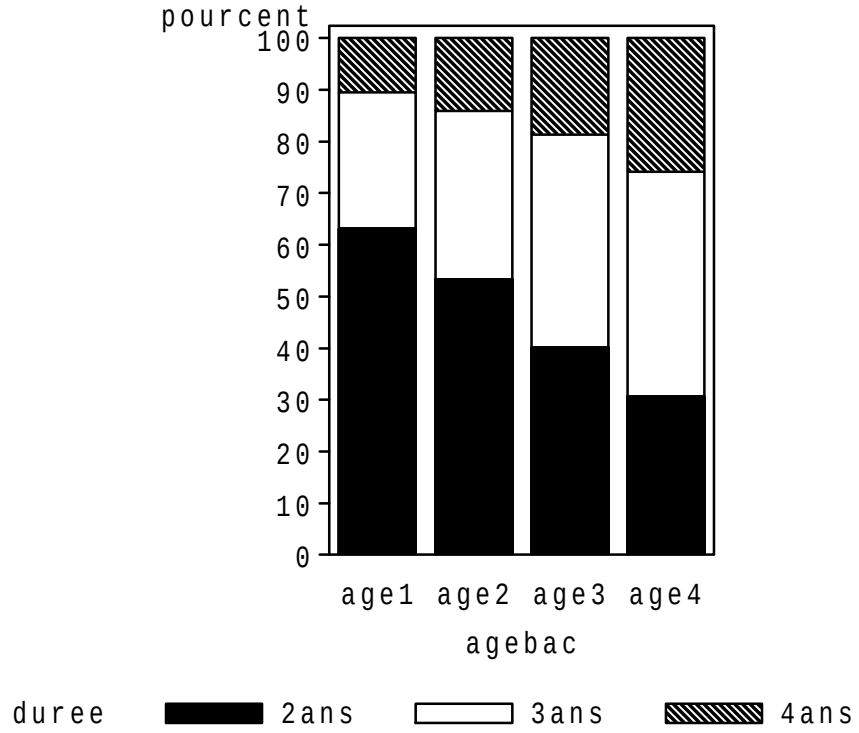

FIG. 3.5 – *Diagramme en barres des profils-colonnes*

La représentation graphique des profils-lignes ou des profils-colonnes au moyen, par exemple, d'un diagramme en barres parallèles, donne une idée assez précise de la variation conjointe des deux variables.

Illustration

Nous avons déterminé les profils-colonnes (en pourcentages) relatifs à la table de contingence présentée dans l'Exemple 6 et nous les donnons ci-dessous.

	<18 ans	18 ans	19 ans	>19 ans	profil moyen
2 ans	63,2	53,3	40,1	30,6	50,2
3 ans	26,3	32,6	41,2	43,6	34,4
4 ans	10,5	14,1	18,7	25,8	15,4
sommes	100,0	100,0	100,0	100,0	100,0

La Figure 3.5 donne le diagramme en barres pour les profils-colonnes ci-dessus, et une liaison entre les deux variables étudiées apparaît très clairement.

3.3.3 Les indices de liaison : le khi-deux et ses dérivés

Propriété préliminaire

On peut établir l'équivalence des trois propriétés suivantes :

(i) tous les profils-lignes sont égaux ;

(ii) tous les profils-colonnes sont égaux ;

(iii) pour tout couple d'indices (ℓ, h), on a : $n_{\ell h} = \dfrac{n_{\ell +} n_{+h}}{n}$.

Cette propriété, de nature mathématique, est très importante au niveau pratique. En effet, si une table de contingence vérifie ces trois propriétés, on peut alors dire qu'il n'existe aucune forme de liaison entre les deux variables considérées X et Y.

Pour s'en rendre compte, considérons, par exemple, l'égalité des profils-lignes : elle signifie que la répartition des individus selon les modalités de Y est la même, quelle que soit la modalité de X considérée. Autrement dit, X n'a pas d'influence sur la répartition selon Y, donc X n'a pas d'influence sur Y : les deux variables ne sont pas liées. De même en cas d'égalité des profils-colonnes.

Pour la construction d'un indicateur de liaison sur une table de contingence, c'est toutefois la troisième propriété qui va être utilisée. D'une part, elle est symétrique selon les lignes et les colonnes de la table, ce qui est très commode, d'autre part, elle se prête bien à la construction d'un tel indice : on va évaluer l'écart entre la situation observée (la table de contingence dont on dispose) et l'état de non liaison défini par (iii).

Définition du khi-deux

Il est courant, en statistique, de comparer une table de contingence observée, dont les effectifs conjoints sont notés $n_{\ell h}$, à une table de contingence donnée a priori (et appelée *standard*), dont les effectifs conjoints sont notés $s_{\ell h}$, en calculant la quantité

$$\sum_{\ell=1}^{r}\sum_{h=1}^{c}\frac{(n_{\ell h}-s_{\ell h})^2}{s_{\ell h}}$$

(il s'agit de la somme de tous les carrés des écarts rapportés aux effectifs standards). La somme de tous les écarts élevés au carré rappelle la définition de la variance ; la division de chaque carré par l'effectif standard correspondant permet de relativiser les carrés considérés.

De façon naturelle, pour mesurer la liaison sur une table de contingence, on utilise l'expression ci-dessus en choisissant pour effectif standard ($s_{\ell h}$) l'effectif correspondant à l'absence de liaison ($\frac{n_{\ell+}n_{+h}}{n}$). On mesure de la sorte l'écart à la non liaison, autrement dit l'importance de la liaison. On appelle donc khi-deux (en anglais : *chi-square*), l'indicateur défini comme suit :

$$\chi^2=\sum_{\ell=1}^{r}\sum_{h=1}^{c}\frac{(n_{\ell h}-\frac{n_{\ell+}n_{+h}}{n})^2}{\frac{n_{\ell+}n_{+h}}{n}}=n\left[(\sum_{\ell=1}^{r}\sum_{h=1}^{c}\frac{n_{\ell h}^2}{n_{\ell+}n_{+h}})-1\right].$$

La première égalité correspond donc à la définition de l'indicateur khi-deux, tandis que la seconde est obtenue en développant le carré, puis en sommant sur les deux indices. En général, la seconde formule est plus commode dans les calculs pratiques de cet indicateur de liaison.

En ce qui concerne ses propriétés, le coefficient χ^2 est toujours positif ou nul (par construction) et il est d'autant plus grand que la liaison entre les deux variables considérées est forte (il est construit pour cela). Malheureusement, il dépend aussi des dimensions r et c de la table étudiée, ainsi que de la taille n de l'échantillon observé (ce n'est pas un coefficient "intrinsèque"). En particulier, il n'est pas majoré ; autrement dit, on peut trouver des coefficients χ^2 aussi grand qu'on le souhaite, ce qui est gênant pour l'interprétation concrète de ce coefficient. C'est pour cette raison qu'on a défini d'autres indices, liés au khi-deux, et dont l'objectif est de corriger ces défauts.

Autres indicateurs liés au khi-deux

On en trouve un certain nombre dans la littérature statistique. Nous citerons les trois plus importants.

– *Le phi-deux* :

$$\Phi^2=\frac{\chi^2}{n}.$$

Il ne dépend plus de n, mais dépend encore de r et de c.

– *Le coefficient T de Tschuprow* :

$$T=\sqrt{\frac{\Phi^2}{\sqrt{(r-1)(c-1)}}}.$$

On peut vérifier : $0\leq T\leq 1$.

– *Le coefficient C de Cramér* :

$$C = \sqrt{\frac{\Phi^2}{d-1}},$$

avec $d = \inf(r, c)$. On vérifie maintenant : $0 \leq T \leq C \leq 1$.

Le coefficient Φ^2 est peu utilisé dans la pratique, mais il joue un rôle important en Analyse Factorielle des Correspondances. On utilise beaucoup plus les coefficients de Tschuprow et de Cramér (T et C) car, comme la valeur absolue du coefficient de corrélation linéaire et comme le rapport de corrélation, ils sont compris entre 0 et 1 et sont d'autant plus grands que la liaison entre les deux variables considérées est forte. Toutefois, on notera que T et C sont rarement supérieurs à 0,5 dans la pratique ; sur des exemples réels, ils sont le plus souvent compris entre 0,1 et 0,3 et sont donc difficiles à interpréter dans l'absolu. Ils sont plus utiles lorqu'on recherche, dans une liste de variables qualitatives, celle qui est le plus liée à une autre variable qualitative.

Illustration

En utilisant les données de l'Exemple 6, on a obtenu :

$$\chi^2 = 28,7\,;\ \Phi^2 = 0,036\,;\ T = 0,12\,;\ C = 0,13.$$

On a donc affaire à une liaison peu importante (alors qu'on se doute, a priori, que les deux phénomènes sont liés et que l'étude des profils a montré l'existence effective d'une liaison).

Un autre exemple

Considérons la table de contingence suivante, avec ses marges :

10	15	15	40
20	5	35	60
30	20	50	100

Pour calculer le khi-deux correspondant, on peut utiliser la formule de définition (première solution). Pour cela, on a besoin de déterminer la table donnant le quantités $\frac{n_{\ell+}n_{+h}}{n}$, c'est-à-dire les quantités que l'on aurait (à la place des $n_{\ell h}$) s'il n'y avait aucune liaison dans la table. Voici cette table (on laisse le soin au lecteur de contrôler) :

12	8	20	40
18	12	30	60
30	20	50	100

On remarquera tout d'abord que cette table a les mêmes marges que la table initiale (c'est toujours le cas). Ensuite, la formule de définition donne :

$$\chi^2 = \frac{(10-12)^2}{12} + \frac{(15-8)^2}{8} + \frac{(15-20)^2}{20} + \frac{(20-18)^2}{18} + \frac{(5-12)^2}{12} + \frac{(35-30)^2}{30} = 12,85.$$

Cela étant, il est aussi possible de calculer le khi-deux par la formule développée (deuxième solution). Dans chaque cellule (chaque case) de la table, on doit calculer les carrés des effectifs conjoints $n_{\ell h}$, puis les diviser par le produit des deux effectifs marginaux correspondants $n_{\ell+}n_{+h}$. On additionne alors les 6 quantités ainsi obtenues, on leur retranche 1, et on multiplie le tout par $n = 100$. On laisse encore le soin au lecteur de vérifier que cela conduit au même résultat (12,85).

On peut maintenant calculer le phi-deux, qui vaut $\frac{12,85}{100} = 0,1285$. Quant au coefficient de Cramér, il est égal à la racine carrée de $\frac{0,1285}{1}$, soit 0,36. Comme d'habitude, cette valeur est faible par rapport à son intervalle de variation $[0, 1]$, mais la liaison n'est certainement pas négligeable.

3.3.4 Généralisation : le tableau de Burt

Terminons cette section sur les variables qualitatives en faisant une incursion dans l'étude simultanée de plusieurs variables qualitatives (trois ou plus). Du point de vue de la présentation des données, on utilise en général dans ce cas un tableau particulier, appelé tableau de Burt, que nous présentons maintenant.

	bacC	bacD	< 18	18ans	19ans	> 19	2ans	3ans	4ans
bacC	583	0	108	323	114	38	324	192	67
bacD	0	214	25	97	68	24	76	82	56
< 18	108	25	133	0	0	0	84	35	14
18ans	323	97	0	420	0	0	224	137	59
19ans	114	68	0	0	182	0	73	75	34
> 19	38	24	0	0	0	62	19	27	16
2ans	324	76	84	224	73	19	400	0	0
3ans	192	82	35	137	75	27	0	274	0
4ans	67	56	14	59	34	16	0	0	123

TAB. 3.1 – *Tableau de Burt*

Principe

Le tableau de Burt est une généralisation particulière de la table de contingence dans le cas où l'on étudie simultanément p variables qualitatives. Notons $X^1, \ldots, X^p$ ces variables, appelons c_j le nombre de modalités de X^j, $j = 1, \ldots, p$, et posons $c = \sum_{j=1}^{p} c_j$. Le tableau de Burt est en fait une matrice (un tableau) carrée $c \times c$, constituée de p^2 sous-matrices. Chacune des p sous-matrices diagonales est relative à l'une des p variables ; la $j^{\text{ième}}$ d'entre elles est carrée d'ordre c_j, diagonale, et comporte sur la diagonale les effectifs marginaux de X^j. La sous-matrice figurant dans le bloc d'indice $(j, j'), j \neq j'$, est la table de contingence construite en mettant X^j en lignes et $X^{j'}$ en colonnes ; le tableau de Burt est donc symétrique. Le tableau de Burt est à la base de l'Analyse des Correspondances Multiples, méthode importante en statistique multidimensionnelle.

Illustration

Toujours avec les données de l'Exemple 6, nous avons déterminé le tableau de Burt pour les trois variables série de bac, âge au bac, et durée du DEUG. Il est donné dans le Tableau 3.1.