Statistique descriptive unidimensionnelle

Résumé

Les objectifs et la démarche d'un première exploration d'un jeu de données, les outils de la description statistique d'une variable quantitative (indicateur de tendance centrale, de dispersion, histogramme, diagramme-boîte), puis d'une variable qualitative (fréquences).

Retour au plan.

1 Introduction

l'objectif des outils de Statistique descriptive élémentaire est de fournir des résumés synthétique de séries de valeurs, adaptés à leur type (qualitatives ou quantitatives), et observées sur une population ou un échantillon.

Dans le cas d'une seule variable, Les notions les plus classiques sont celles de médiane, quantile, moyenne, fréquence, variance, écart-type définies parallèlement à des représentations graphiques : diagramme en bâton, histogramme, diagramme-boîte, graphiques cumulatifs, diagrammes en colonnes, en barre ou en secteurs.

Dans le cas de deux variables, on s'intéresse à la corrélation, au rapport de corrélation ou encore à la statistique d'un test du χ^2 associé à une table de contingence. Ces notions sont associées à différents graphiques comme le nuage de points (scatterplot), les diagrammes-boîtes parallèles, les diagrammes de profils ou encore en mosaïque.

Les définitions de ces différentes notions se trouvent dans n'importe quel ouvrage élémentaire de Statistique, nous nous proposons simplement de rappeler dans ce chapitre certains outils moins classiques mais efficaces et présents dans la plupart des logiciels statistiques. Cela nous permettra également d'illustrer les premières étapes descriptives à réaliser sur un jeu de données.

1.1 Démarche

Toute étude sophistiquée d'un corpus de données doit être précédée d'une étude *exploratoire* à l'aide d'outils, certes rudimentaires mais robustes, en privilégiant les représentations graphiques. C'est la seule façon de se familiariser avec des données et de dépister les sources de problèmes :

- valeurs manquantes, erronées ou atypiques, biais expérimentaux,
- modalités trop rares,
- distributions "anormales" (dissymétrie, multimodalité, épaisseur des queues),
- incohérences, liaisons non linéaires.
- ...

C'est ensuite la recherche de prétraitements des données afin de corriger les sources de problèmes et les rendre exploitables par des techniques plus sophistiquées :

- transformation : logarithme, puissance, réduction, rangs... des variables,
- codage en classe ou recodage de classes,
- imputations ou non des données manquantes,
- lissage, décompositions (ondelettes, Fourier) de courbes,

Ensuite, les techniques exploratoires multidimensionnelles permettent des

- représentations graphiques synthétiques,
- réductions de dimension pour la compression ou le résumé des données,
- recherches et représentations de typologies des observations.

1.2 Avertissement

Attention le côté rudimentaire voire trivial des outils de statistique descriptive uni et bidimensionnelle ne doit pas conduire à les négliger au profit d'une mise en œuvre immédiate de méthodes beaucoup plus sophistiquées, donc beaucoup plus sensibles aux problèmes cités ci-dessus. S'ils ne sont pas pris en compte, ils réapparaîtront alors comme autant d'*artefacts* susceptibles de dénaturer voire de fausser toute tentative de modélisation.

Plus précisément, les méthodes descriptives ne supposent, *a priori*, aucun modèle sous-jacent, de type probabiliste. Ainsi, lorsque l'on considère un ensemble de variables quantitatives sur lesquelles on souhaite réaliser une Analyse en Composantes Principales, il n'est pas nécessaire de supposer que ces variables sont distribuées selon des lois normales. Néanmoins, l'absence de

données atypiques, la symétrie des distributions sont des propriétés importantes des séries observées pour s'assurer de la qualité et de la validité des résultats.

Le déroulement pédagogique linéaire ne doit pas faire perdre de vue que la réalité d'une analyse est plus complexe et nécessite différentes étapes en boucle afin, par exemple, de contrôler l'influence possible des choix parfois très subjectifs opérés dans les étapes de normalisation ou transformation des données pour éventuellement les remettre en cause.

2 Variable quantitative

2.1 Variable quantitative discrète

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...

On a noté l'âge (arrondi à l'année près) des 48 salariés d'une entreprise ; la série statistique brute est donnée 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

Présentation des données

Le tableau statistique C'est 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 $\{x_l\ ;\ l = 1, \ldots, r\}$. Dans une seconde colonne, on dispose, en face de chaque valeur x_l, le nombre de réplications qui lui sont associées ; ces réplications sont appelées *effectifs* et notées n_l. Les effectifs n_l sont souvent remplacés par les quantités $f_l = \frac{n_l}{n}$, appelées *fréquences* (rappelons que n désigne le nombre total d'observations, c'est–à–dire le cardinal de Ω : $n = \sum_{l=1}^{r} n_l$).

x_l	n_l	N_l	$f_l(\%)$	$F_l(\%)$
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

TABLE 1 – *Effectifs, effectifs cumulés, fréquences et fréquences cumulées.*

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 :

$$N_l = \sum_{j=1}^{l} n_j \text{ et } F_l = \sum_{j=1}^{l} f_j.$$

On notera que $N_r = n$ et $F_r = 1$.

Illustration Dans le tableau statistique (1), on a calculé, sur les données présentées dans l'exemple 2.1, les effectifs, effectifs cumulés, fréquences et fréquences cumulées.

Remarque. —

- Comme c'est le cas ci-dessus, les fréquences sont souvent exprimées en pourcentages.
- Le choix entre effectifs (resp. effectifs cumulés) et fréquences (resp. fréquences cumulées) est très empirique ; il semble naturel de choisir les effectifs lorsque l'effectif total n est faible et les fréquences lorsqu'il est plus important ; la limite approximative de 100 paraît, dans ces conditions, assez raisonnable.

La présentation tige–et–feuille (ou "stem–and–leaf") Cette façon particulière de présenter les données est assez commode, dans la mesure où elle préfigure déjà un graphique. Elle est illustrée ci–dessous sur le même exemple que précédemment.

2	4 6 6 9 9 9
3	1 1 3 3 3 3 7 7 8 8 8 8
4	1 1 1 3 3 3 5 6 6 6 6 6 6 9 9 9
5	0 2 2 2 7 7 7 7 7 9 9
6	0 0 2

Elle consiste donc, dans la présentation des données, à séparer la partie des dizaines de celle des unités. En face de la partie des dizaines, chaque unité est répétée autant de fois qu'il y a d'observations de la valeur correspondante. Bien entendu, cette présentation doit être adaptée de façon appropriée lorsque les données sont d'un autre ordre de grandeur.

Représentations graphiques

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).

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

Le diagramme cumulatif Il figure les effectifs cumulés (resp. les fréquences cumulées) et permet de déterminer simplement le nombre (resp. la proportion) d'observations inférieures ou égales à une valeur donnée de la série. Lorsqu'il est relatif aux fréquences, c'est en fait le graphe de la *fonction de répartition empirique* F_X définie de la façon suivante :

$$F_X(x) = \begin{cases} 0 & \text{si } x < x_1, \\ F_l & \text{si } x_l \leq x < x_{l+1}, \quad l = 1, \ldots, r-1, \\ 1 & \text{si } x \geq x_r. \end{cases}$$

Le diagramme cumulatif relatif à l'exemple 2.1 est donné par la figure 2.

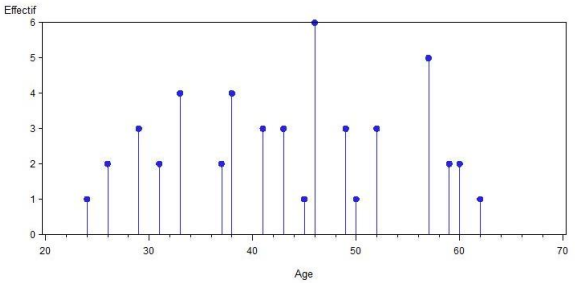

FIGURE 1 – *Diagramme en bâtons*

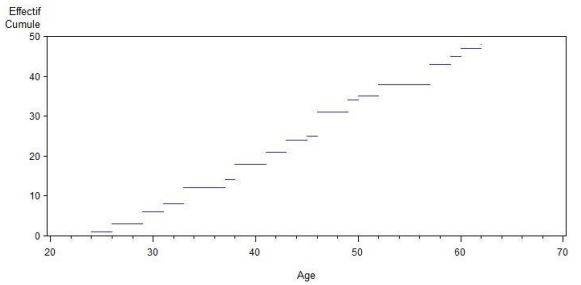

FIGURE 2 – *Diagramme cumulatif*

Notion de quantile

Définition La fréquence cumulée F_l ($0 \leq F_l \leq 1$) donne la proportion d'observations inférieures ou égales à x_l. Une approche complémentaire consiste à se donner a priori une valeur α, comprise entre 0 et 1, et à rechercher x_α vérifiant $F_X(x_\alpha) \simeq \alpha$. 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.

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. 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).

Les autres quantiles Les *quintiles*, *déciles* et *centiles* sont également d'usage assez courant.

Le diagramme-boîte (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 3 donne le diagramme–boîte de l'exemple 2.1. Dans cet exemple, on a obtenu $x_{\frac{1}{4}} = 35$, $x_{\frac{1}{2}} = 44$ et $x_{\frac{3}{4}} = 52$; on notera que l'obtention, d'une part de $x_{\frac{1}{4}}$ et $x_{\frac{1}{2}}$, d'autre part de $x_{\frac{3}{4}}$, ne s'est pas faite de la même façon (en fait, avec une variable discrète, la détermination des quantiles est souvent approximative comme on peut le constater avec cet exemple).

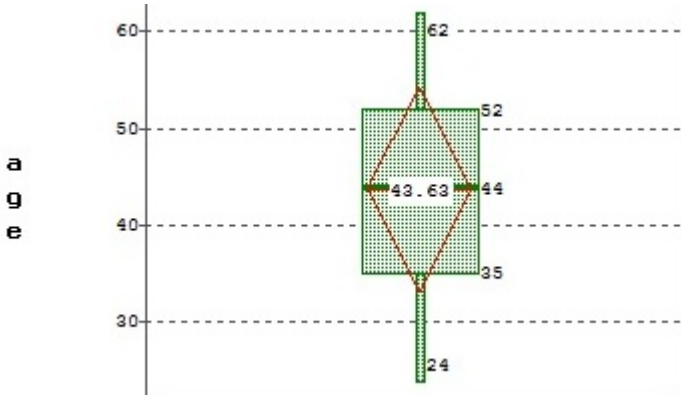

FIGURE 3 – *Diagramme-boîte et moyenne en rouge*

Caractéristiques numériques

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

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*,
- la *moyenne* (ou moyenne arithmétique).

Formule de la moyenne pour une variable quantitative discrète :

$$\overline{x} = \frac{1}{n}\sum_{l=1}^{r} n_l x_l = \sum_{l=1}^{r} f_l x_l.$$

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.

- L'*étendue* ($x_r - x_1$),
- l'*intervalle inter-quartiles* ($x_{\frac{3}{4}} - x_{\frac{1}{4}}$),
- l'*écart-moyen à la médiane* ($\frac{1}{n}\sum_{l=1}^{r} n_l \left|x_l - x_{\frac{1}{2}}\right|$),
- l'*écart-moyen à la moyenne* ($\frac{1}{n}\sum_{l=1}^{r} n_l |x_l - \overline{x}|$),

sont des caractéristiques de dispersion que l'on rencontre parfois.

Mais, la caractéristique de loin la plus utilisée est l'*écart–type*, racine carrée positive de la *variance*. Formules de la variance :

$$\begin{aligned} var(X) = \sigma_X^2 &= \frac{1}{n}\sum_{l=1}^{r} n_l (x_l - \overline{x})^2 \\ &= \frac{1}{n}\sum_{l=1}^{r} n_l (x_l)^2 - (\overline{x})^2. \end{aligned}$$

L'écart–type de X sera donc noté σ_X.

Illustration En utilisant toujours l'exemple 2.1, on a calculé :

$$\begin{aligned} \overline{x} &= \frac{1}{n}\sum_{l=1}^{r} n_l x_l = \frac{2094}{48} = 43,625 \simeq 43,6 \text{ ans} ; \\ \sigma_X^2 &= \frac{1}{n}\sum_{l=1}^{r} n_l (x_l)^2 - (\overline{x})^2 = \frac{96620}{48} - (43,625)^2 \simeq 109,7760 ; \\ \sigma_X &= \sqrt{\sigma_X^2} \simeq 10,5 \text{ ans}. \end{aligned}$$

Remarque. — Toutes les caractéristiques numériques introduites ici (médiane, moyenne, variance, écart–type...) sont dites *empiriques*, c'est–à–dire calculées sur un échantillon Ω ; par opposition, on parle, par exemple, de moyenne *théorique* (ou espérance mathématique) pour désigner le concept de moyenne relatif à une variable aléatoire réelle.

2.2 Variable quantitative continue

Généralités

Une variable quantitative est dite continue lorsque les observations qui lui sont associées ne sont pas des valeurs précises mais des intervalles réels. Cela signifie que, dans ce cas, le sous–ensemble de $\mathbb{R}$ 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 (il est moins gênant de demander à des individus leur classe de salaire que leur salaire précis). Deux exemples de variables quantitatives fréquemment considérées comme continues sont l'âge et le revenu (pour un groupe d'individus).

Nous noterons $(b_0\,;b_1)$,...,$(b_{r-1}\,;b_r)$ les classes considérées. Les nombres b_{l-1} et b_l sont appelés les *bornes* de la $l^{\text{ième}}$ classe ; $\frac{b_{l-1}+b_l}{2}$ est le *centre* de cette classe et $(b_l - b_{l-1})$ en est l'*amplitude* (en général notée a_l).

Présentation des données

On utilise encore un tableau statistique analogue à celui vu au paragraphe précédent, en disposant dans la première colonne les classes rangées par ordre croissant. Les notions d'effectifs, de fréquences, d'effectifs cumulés et de fréquences cumulées sont définies de la même façon que dans le cas discret. On notera que l'on n'utilise pas dans ce cas la présentation tige–et–feuille car les valeurs exactes de la série sont inconnues.

Le tableau ci-dessous donne, pour l'année 1987, la répartition des exploitations agricoles françaises selon la SAU (surface agricole utilisée) exprimée en hectares (Tableaux Économiques de Midi–Pyrénées, INSEE, 1989, p. 77) ; la SAU est ici une variable quantitative continue comportant 6 classes.

SAU (en ha)	fréquences (%)
moins de 5	24,0
de 5 à 10	10,9
de 10 à 20	17,8
de 20 à 35	20,3
de 35 à 50	10,2
plus de 50	16,8

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.

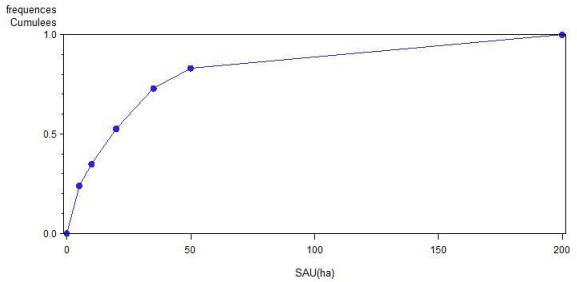

FIGURE 4 – *Courbe cumulative*

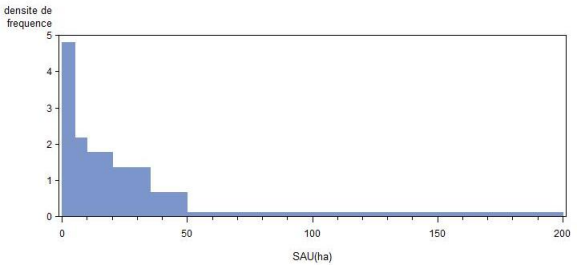

FIGURE 5 – *Histogramme (classes d'effectifs égaux) des répartitions des SAU*

Courbe cumulative C'est encore une fois le graphe de la *fonction de répartition empirique*, cette dernière devant maintenant être précisée au moyen d'*interpolations linéaires*.

On appelle fonction de répartition empirique de la variable continue X la fonction F_X définie par :

$$F_X(x) = \begin{cases} 0 & \text{si } x < b_0, \\ F_{l-1} + \frac{f_l}{b_l - b_{l-1}}(x - b_{l-1}) & \text{si } b_{l-1} \leq x < b_l, \quad l = 1, \ldots, r, \\ 1 & \text{si } x \geq b_r \end{cases}$$

(on a supposé $F_0 = 0$).

La courbe cumulative relative à l'exemple 2.2 est donnée par la figure 4. 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).

Histogramme La fonction de répartition empirique est, dans le cas continu, une fonction dérivable sauf, éventuellement, aux points d'abscisses $b_0, b_1, \ldots, b_r$. Sa fonction dérivée, éventuellement non définie en ces points, est appelée *densité empirique* de X et notée f_X. On obtient :

$$f_X(x) = \begin{cases} 0 & \text{si } x < b_0, \\ \frac{f_l}{b_l - b_{l-1}} & \text{si } b_{l-1} < x < b_l, \quad l = 1, \ldots, r, \\ 0 & \text{si } x \geq b_r. \end{cases}$$

Le graphe de f_X est alors appelé histogramme de la variable X. Un histogramme est donc la juxtaposition de rectangles dont les bases sont les amplitudes des classes considérées ($a_l = b_l - b_{l-1}$) et dont les hauteurs sont les quantités $\frac{f_l}{b_l - b_{l-1}}$, appelées *densités de fréquence*. L'aire du $l^{\text{ième}}$ rectangle vaut donc f_l, fréquence de la classe correspondante.

L'histogramme correspondant aux données de l'exemple 2.2 est présenté dans la figure 5.

Estimation fonctionnelle La qualité de l'estimation d'une distribution par un histogramme dépend beaucoup du découpage en classe. Malheureusement, plutôt que de fournir des classes d'effectifs égaux et donc de mieux répartir l'imprécision, les logiciels utilisent des classes d'amplitudes égales et tracent donc des histogrammes parfois peu représentatifs. Ces 20 dernières années, à la suite du développement des moyens de calcul, sont apparues des méthodes d'estimation dites *fonctionnelles* ou *non-paramétriques* qui proposent d'estimer la distribution d'une variable ou la relation entre deux variables par une fonction construite point par point (noyaux) ou dans une base de fonctions *splines*. Ces estimations sont simples à calculer (pour l'ordinateur) mais nécessitent le choix d'un paramètre dit de *lissage*. Les démonstrations du ca-

ractère optimal de ces estimations fonctionnelles, liée à l'optimalité du choix de la valeur du paramètre de lissage, font appel à des outils théoriques plus sophistiquées sortant du cadre de ce cours (Eubank, 1988, Silverman, 1986).

L'estimation de la densité par la méthode du noyau se met sous la forme générale :

$$\widehat{g}_\lambda(x) = \frac{1}{n\lambda}\sum_{i=1}^{n} K\left(\frac{x - x_i}{\lambda}\right)$$

où λ est le paramètre de lissage optimisée par une procédure automatique qui minimise une approximation de l'erreur quadratique moyenne intégrée (norme de l'espace L^2) ; K est une fonction symétrique, positive, concave, appelée *noyau* dont la forme précise importe peu. C'est souvent la fonction densité de la loi gaussienne :

$$K(t) = \frac{1}{\sqrt{2\pi}}\exp(-t^2/2)$$

qui possède de bonnes propriétés de régularité. Le principe consiste simplement à associer à chaque observation un "élément de densité" de la forme du noyau K et à sommer tous ces éléments. Un histogramme est une version particulière d'estimation dans laquelle l'"'élément de densité" est un "petit rectangle" dans la classe de l'observation.

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).

Moyenne et écart-type

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 x_l 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.

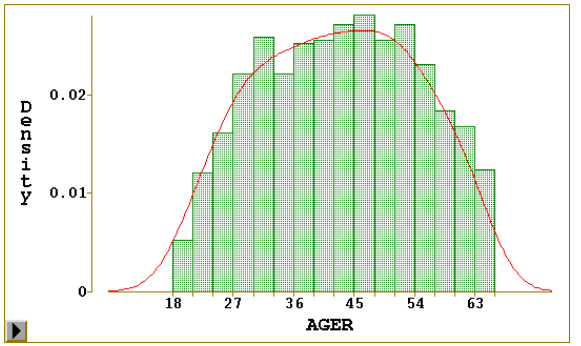

FIGURE 6 – *Histogramme (classes amplitudes égales) des répartitions des âges et estimation non paramétrique de la densité par la méthode du noyau (en rouge).*

Illustration

La médiane de la variable présentée dans l'exemple 2.2 se situe dans la classe (10 ; 20), puisque la fréquence cumulée de cette classe (52,7) est la première à dépasser 50. On détermine la médiane en faisant l'interpolation linéaire suivante (l'indice l ci–dessous désigne en fait la troisième classe) :

$$\begin{aligned} x_{\frac{1}{2}} &= b_{l-1} + a_l \frac{50 - F_{l-1}}{F_l - F_{l-1}} \\ &= 10 + 10\frac{15,1}{17,8} \\ &\simeq 18,5 \text{ ha.} \end{aligned}$$

La moyenne vaut :

$$\overline{x} = \sum_{l=1}^{r} f_l x_l = \frac{3080,5}{100} \simeq 30,8 \text{ ha.}$$

Remarque. —

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.

2.3 Variables quantitatives et logiciels

Le volume des données et la pratique généralisée des logiciels statistiques induit une prise en compte particulière des notions précédentes. Par principe, le codage des valeurs, mêmes réelles, est toujours discret, et la précision fonction du nombre de chiffres significatifs pris en compte En conséquences, tous les calculs des indicateurs (moyenne, variance, quantile...) sont traités avec les formules considérant les valeurs comme connues et discrètes, sans pour autant s'intéresser aux fréquences des valeurs car ces dernières sont généralement distinctes les unes des autres. En revanche, les graphiques produits (histogramme, courbe cumulative mais pas l'estimation fonctionnelle) sont issus de découpages automatiques en classes d'amplitudes égales, pas toujours très judicieux, selon les principes des variables continues.

3 Variable qualitative

3.1 Variables nominales et 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*.

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).

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 8, 7 et 9 présentent chacun de ces trois graphiques sur les données de l'exemple 3.3.

> 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

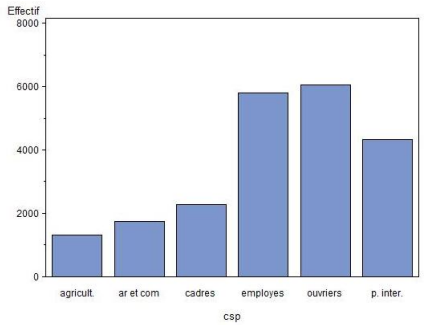

FIGURE 7 – *Diagramme en colonnes*

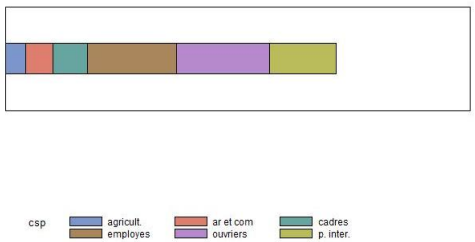

FIGURE 8 – *Diagramme en barre*

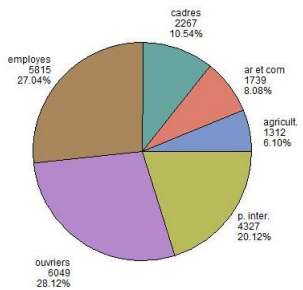

FIGURE 9 – *Diagramme en secteurs*

4 Détection de problèmes

Les quelques outils de ce chapitre permettent déjà de se faire une première idée d'un jeu de données mais surtout, en préalable à toute analyse, ils permettent de s'assurer de la fiabilité des données, de repérer des valeurs extrêmes atypiques, éventuellement des erreurs de mesures ou de saisie, des incohérences de codage ou d'unité.

Les erreurs, lorsqu'elle sont décelées, conduisent naturellement et nécessairement à leur correction ou à l'élimination des données douteuses mais d'autres problèmes pouvant apparaître n'ont pas toujours de solutions évidentes.

- Le mitage de l'ensemble des données ou absence de certaines valeurs en fait partie. Faut-il supprimer les individus incriminés ou les variables ? Faut-il compléter, par une modélisation et prévision partielles, les valeurs manquantes ? Les solutions dépendent du taux de valeurs manquantes, de leur répartition (sont-elles aléatoires) et du niveau de tolérance des méthodes qui vont être utilisées.
- La présence de valeurs atypiques peut influencer sévèrement des estimations de méthodes peu robustes car basées sur le carré d'une distance. Ces valeurs sont-elles des erreurs ? Sinon faut-il les conserver en transformant les variables ou en adoptant des méthodes robustes basées sur des écarts absolus ?
- Même sans hypothèse explicite de normalité des distributions, il est préférable d'avoir à faire à des distributions relativement symétriques. Une transformation des variables par une fonction monotone (log, puissance) est hautement recommandée afin d'améliorer la symétrie de leur distribution ou encore pour linéariser (nuage de points) la nature d'une liaison.

4.1 Marketing bancaire

Les données de patrimoine, de revenu, comme également celles de concentration présente des distributions très disymétriques (figure 10 accompagnées de nombres importants de valeurs atypiques. Le diagramme boîte est un outil efficace pour identifier ce problème avant d'y remédier par une transformation appropriée, ici le logarithme.

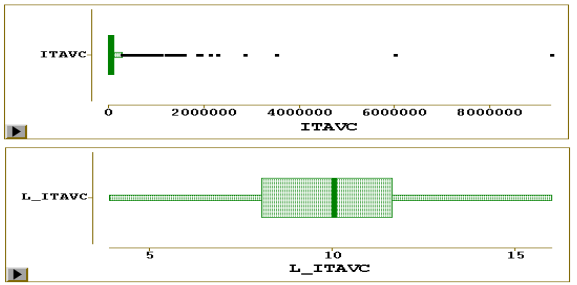

FIGURE 10 – *Banque : La simple transformation* ($\log(50+x)$)*, de la variable cumulants les avoirs, résout bien les problèmes posés par l'allure "log-normale" de sa distribution avec son cortège de valeurs atypiques.*

4.2 Données génomiques

Le diagramme boîte parallèle est également très efficace pour visualiser simultanément les distributions d'un grand nombre de variables, par exemple de centaines voire de milliers de gènes, dont l'expression a été observée dans différentes conditions expérimentales. Dans cet exemple, la représentation des diagrammes en boîtes pour les souris, ordonnées selon le génotype et le régime suivi (Fig. 11) ne donne *a* priori aucune tendance spécifique sur le comportement de l'ensemble des gènes. Cette représentation atteste de la qualité de la production et de prétraitement des données. En effet, celles-ci ont été recueillies en utilisant une membrane par souris ; ainsi, une quelconque anomalie sur un support, affectant l'ensemble des mesures relatives à une souris particulière, apparaîtrait nécessairement sur cette représentation. Notons seulement que quelques gènes atypiques, facilement repérables sur la figure 12 comme les plus sur-exprimés, se retrouvent dans les valeurs extrêmes pour chaque souris sur la figure 11.

Les diagrammes en boîtes pour chaque gène (Fig. 12) révèlent des gènes dont l'expression est, sur l'ensemble des souris, nettement différentes des autres (par exemple, `16SR`, `apoA.I`, `apoE`). Les gènes des ARN ribosomiques comme le `16SR` (ARN 16s ribosomique mitochondrial), présentent,

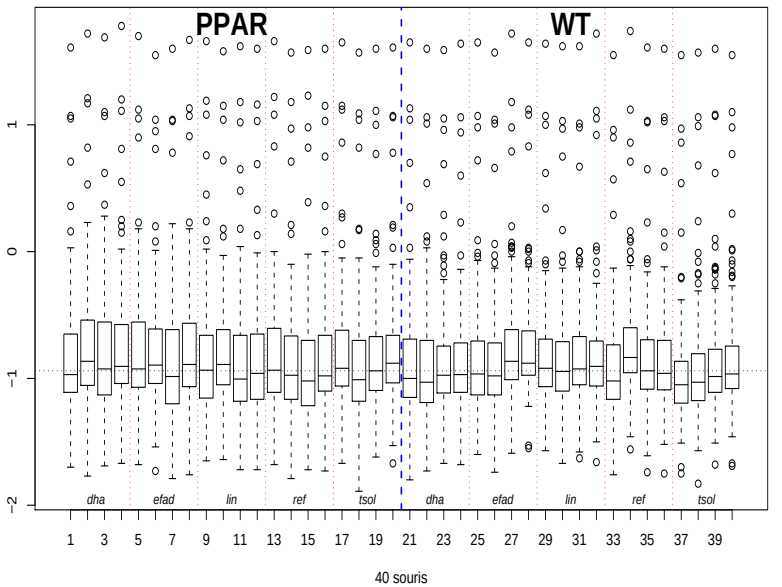

FIGURE 11 – *Souris : diagrammes en boîtes pour les 40 souris. La ligne verticale et épaisse sépare les souris selon leur génotype. Les lignes verticales et fines séparent les souris selon le régime qu'elles ont suivi. La ligne horizontale représente la médiane de l'ensemble des valeurs.*

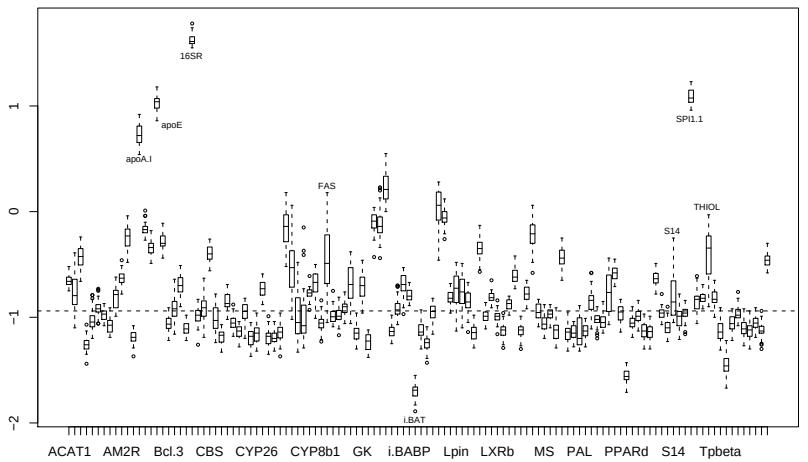

FIGURE 12 – *Souris : Diagrammes-boîtes parallèles représentant simultanément les distributions des logarithmes des expressions des gènes.*

dans toutes les cellules de l'organisme, des niveaux d'expression plus élevés que tous les gènes codant des ARN messagers. Ces ARN servent en effet à la traduction des ARN messagers en protéines. Par ailleurs, on peut constater que les expressions de certains gènes varient beaucoup plus que d'autres sur l'ensemble des souris (par exemple, `FAS`, `S14` et `THIOL`). Pour ces derniers gènes, on peut supposer qu'une part de cette variabilité est due aux facteurs considérés, ce que nous essaierons de confirmer par la suite au moyen de techniques de modélisation.

L'intérêt de ces représentations réside davantage dans la vision synthétique qu'elles offrent que dans l'information biologique que l'on peut en extraire. Elles nous orientent également dans les premiers choix méthodologiques à établir avant de poursuivre l'analyse. En effet, les boîtes relatives à la distribution des gènes mettent clairement en évidence un certain nombre de gènes dont l'expression est systématiquement supérieure à celle des autres, quelles que soient les conditions expérimentales. De plus, la variabilité de ces expressions est, le plus souvent, très faible. Ce constat nous conduit à effectuer un centrage des gènes (en colonnes), afin d'éviter un effet taille lors de la mise en œuvre de techniques factorielles.