REPERES - IREM . N° 35 - avril 1999

HEURS ET MALHEURS DU SU ET DU PERÇU EN STATISTIQUE

Des données à leurs représentations graphiques

Bernard PARZYSZ
IREM de Lorraine et Equipe DIDIREM

Le problème fondamental de la statistique descriptive est la résolution du dilemme résultant de la transformation de l'information "brute" recueillie en une synthèse qui parvienne à concilier au mieux ces deux pôles antagonistes que sont la fidélité et la clarté. Prenons, pour fixer les idées, le cas d'une variable numérique dont l'observation conduit à la réalisation d'un *tableau* donnant, pour chacune des valeurs de la variable, la fréquence correspondante. Ce tableau n'est pas toujours parlant, surtout si les valeurs observées sont nombreuses : il permet certes de repérer les valeurs les plus fréquentes et les valeurs extrêmes (ce qu'un *graphique* permettra en principe de faire d'un seul coup d'œil, au lieu de parcourir la liste des effectifs), mais on n'a souvent pas vraiment besoin, pour se faire une idée suffisante de la distribution, de la totalité des données : c'est l'intérêt des *paramètres* que l'on calcule à partir de ces données, et qui permettent de définir des valeurs "moyennes" (dans un sens qui reste à définir), une "dispersion" des valeurs observées autour de l'une de ces moyennes, etc. La distribution sera donc, en quelque sorte, résumée grâce à un nombre réduit de valeurs "typiques" : plus ce nombre sera faible, plus on aura une vue synthétique, mais aussi plus on aura perdu d'information. D'où le dilemme signalé au début. Les représentations graphiques constituent une autre façon de représenter l'information initiale, ou une information déjà partiellement synthétisée, cette fois par l'intermédiaire de la vision. C'est cette question que je voudrais évoquer ici, d'autant plus :

– que nos élèves sont de plus en plus confrontés à des représentations graphiques variées d'origine statistique, tant en classe (géographie, économie, biologie...) qu'en dehors (presse, télévision...) ;

– que la lecture des manuels de mathématiques du secondaire (collège et lycée) ne permet pas toujours – comme nous le verrons – de se faire une vision claire de la question des représentations graphiques utilisées dans le domaine statistique.

Qu'est-ce qu'une variable statistique ?

Je crois utile, avant d'entrer dans le vif du sujet, de rappeler quelques notions relatives aux variables statistiques qui nous seront utiles pour la suite, et je m'appuierai pour cela à la fois sur un article de Jean-François Pichard (**[4]**) et sur un ouvrage de Sabin Lessard et Monga (**[3]**).

Une *variable (ou caractère) statistique*, définie sur une population déterminée, est une *"caractéristique susceptible de variations observables"* (**[3]** p. 2). Une variable statistique présente donc *a priori* plusieurs *modalités*. Par exemple, dans la population adulte résidant en France métropolitaine le 1er janvier 1998 à 0 heure, le caractère S "groupe sanguin" présente 4 modalités : O, A, B, AB ; le caractère D "département de la résidence principale" présente 95 modalités ; le caractère P "nombre de parents décédés avant l'âge de 50 ans" présente 3 modalités : 0, 1, 2 ; le caractère T "taille exprimée en cm" présente *théoriquement* une infinité de modalités (qui sont des réels strictement positifs) ; etc.

Une variable est dite quantitative *"lorsque* [ses] *valeurs possibles* [...] *sont mesurables (c'est-à-dire qu'on peut les évaluer numériquement en utilisant une unité de mesure bien définie servant de référence et permettant de faire des comparaisons précises entre deux valeurs distinctes"* (**[3]** p. 3) ; sinon on dit qu'elle est qualitative. Parmi les exemples ci-dessus, les variables P et T sont quantitatives, tandis que S et D sont qualitatives.

N.B. : pour une variable qualitative, on parle plutôt de *modalités*, alors que pour une variable quantitative on préfère parler de *valeurs*, puisqu'il s'agit de réels.

Une variable qualitative est dite *ordinale* lorsque *"les modalités sont* naturellement *ordonnées bien que les valeurs du caractère ne soient pas numériques"* (**[4]** p. 84) (c'est en particulier le cas lorsqu'il existe une variable quantitative sous-jacente) ; sinon, on dit qu'elle est *nominale*. Remarquons que, dans le cas d'une variable qualitative ordinale, l'existence de cet ordre peut dépendre du contexte. Soient par exemple (d'après **[3]** p. 4) les quatre couleurs bleu, vert, rouge, jaune, considérées comme ensemble des modalités d'une variable statistique (qualitative) :

a) si l'on s'intéresse à la couleur préférée des Français, il n'y a *a priori* aucune raison d'ordonner ces quatre couleurs : la variable est nominale ;
b) si l'on s'intéresse à la réfraction de ces quatre couleurs, on sera amené à les ordonner en fonction de leurs longueurs d'onde : rouge < jaune < vert < bleu ; la variable est ordinale.

Ainsi, dans les deux exemples de variables qualitatives cités plus haut, S est nominale et D est ordinale (on peut par exemple ordonner les départements selon leurs numéros de code postal).

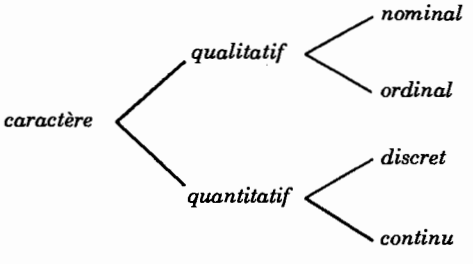
Une variable quantitative est dite *discrète* lorsque "[ses] *valeurs possibles a priori* [...] *sont toutes des nombres représentant des quantités, et ces nombres sont isolés les uns des autres"* (**[3]** p. 6) ; sinon, elle est dite *continue*.

En langage symbolique, si V est l'ensemble des valeurs de la variable *"possibles a priori"* ($V \subset \mathbf{R}$), la variable est discrète si, et seulement si :

$$\forall x \in V \quad \exists\, \alpha(x) > 0$$
$$]x - \alpha(x);\ x + \alpha(x)[\cap V = \{x\}.$$

N.B. : Ceci est vrai en particulier si V = **N**, ou une partie de **N**. Remarquons en outre que *"dans la pratique, les valeurs observées* [d'une] *variable quantitative continue sont toujours arrondies"* (**[3]** p. 8). Par exemple, la variable T définie plus haut est une variable continue *par nature*, même si les valeurs observées sont nécessairement discrètes (par ex. taille exprimée par un nombre *entier* de cm, ou même de mm). Cette distinction entre le phénomène observé et la *mesure* de ce phénomène – qui peut sembler oiseuse étant donné qu'on n'a toujours qu'un nombre fini de valeurs – s'avère pourtant utile, comme on le verra plus loin.

Finalement, on peut représenter les différents types de caractères statistiques sous la forme de l'arbre suivant :



Le recueil et le tableau des données

Dans le cas d'une statistique simple (où n'intervient qu'un seul caractère), le recueil des données statistiques consiste, pour chacun des individus de la population considérée, à relever la modalité de la variable qu'il présente ; on obtient ainsi un ensemble de couples (individu ; modalité). Puis, étant donné qu'il s'agit de présenter en définitive une vue synthétique de la situation, et de *"révéler l'essentiel masqué sous l'abondance des données"* (**[4]** p.76) :

a) si l'ensemble des modalités effectivement réalisées comporte "peu" d'éléments (*i.e.* de l'ordre de quelques unités), on construit un *tableau* indiquant, pour chaque modalité, l'effectif (ou la fréquence) correspondant(e) ; on obtient ainsi un ensemble de couples (modalité ; effectif) ou (modalité ; fréquence). C'est le cas, parmi les exemples du début, pour les variables S et P. Remarquons enfin que la variable D, quoique qualitative, présente 95 valeurs ; on aura sans doute intérêt ici à regrouper certaines modalités, par exemple en considérant la région de résidence plutôt que le département. Notons qu'en outre, dans le cas d'une variable quantitative ou d'une variable qualitative ordinale, le tableau est *ordonné* par les modalités.

b) si l'ensemble des modalités effectivement réalisées comporte "beaucoup" d'éléments – ce cas comprend les variables continues, comme on vient de le voir, mais ne s'y réduit pas – on regroupe plusieurs modalités pour en former une seule. C'est le cas des variables D et T : les modalités de D (qualitative) pourront être, comme on l'a dit, regroupées par régions administratives (mais on changera alors de variable, et on perdra le caractère ordinal) ; quant aux valeurs de T (quantitative), on les regroupera en *classes* réalisant une partition d'un intervalle contenant toutes les valeurs possibles du caractère, classes qui peuvent ou non être de même amplitude. Dans ce cas on obtient une suite (ordonnée)

de couples (classe ; effectif) ou (classe ; fréquence). Notons que ces regroupements s'effectuent au prix d'une *perte d'information* sur les données initiales.

En outre, le regroupement en classes des valeurs d'une variable quantitative est fréquemment suivi, en vue d'une utilisation ultérieure pour des calculs de paramètres, de l'adjonction – voire de la substitution – à chacune des classes du tableau, d'une *valeur unique* associée à cette classe (en général, le *centre* de la classe) ; ce qui revient à remplacer la variable initiale par une variable discrète (même si la première est continue). En cas de substitution, les limites de classes disparaîtront, avec pour conséquence une nouvelle perte d'information.

Considérons par exemple, dans une population d'escargots d'une espèce donnée, la variable (quantitative continue) "diamètre de la coquille (mesuré en mm)", et supposons que le tableau disponible pour l'étude statistique ne comporte que les valeurs suivantes (classées dans l'ordre croissant) : 18, 19, 20, 21, 22, etc. Peut-on, à partir de ces valeurs, retrouver les classes initiales ? La réponse est évidemment non, car on ignore comment ces valeurs ont été obtenues à partir des classes. Ainsi :

– s'il y a eu *troncature*, la classe correspondant à la valeur 21 est [21 ; 22[;
– s'il y a eu *arrondi*, la classe correspondant à la valeur 21 est [20,5 ; 21,5[(ou]20,5 ; 21,5] (sans préjudice d'autres procédés d'obtention).

Supposons maintenant que nous voulions, par exemple, déterminer la moyenne (arithmétique) du diamètre ; pour ce faire, on affecte à tous les individus d'une même classe la valeur centrale de la classe. Ici, dans le premier cas (troncature), le centre de la classe "21" est 21,5, tandis que dans le second (arrondi), ce centre est 21. D'où un biais systématique de 0,5 dans le calcul de la moyenne (ce qui, entre parenthèses, montre l'importance de conserver l'information "limites des classes").

Pour résumer ce qui précède, on peut considérer que l'on se trouve maintenant en présence d'un ensemble réduit de couples :

– soit (classe ; fréquence), et l'ensemble est alors ordonné selon les classes ;
– soit (modalité ; fréquence), cet ensemble étant éventuellement ordonné, lorsque c'est possible, selon les modalités (*i.e.* lorsque le caractère est quantitatif, ou qualitatif ordinal).

Du su au perçu

Dans la plupart des cas, le tableau ainsi obtenu sert de base à la réalisation d'une (voire de plusieurs) représentation(s) graphique(s). Dans la pratique, ces représentations présentent une grande variété : diagramme en bâtons, en barres, circulaire, semi-circulaire, histogramme, polygone, etc. De plus, à côté de ces graphiques "classiques", on trouve également des diagrammes plus spécifiques, comme celui [1] de la *fig. 1*, relatif aux taux de SO_2 relevés le 13 janvier 1997 par les 10 capteurs parisiens. Dans ce domaine, l'imagination des graphistes est apparemment sans bornes, comme le montre [2] la *fig. 2* qui traite de la rentrée scolaire 1997.

(1) *Le Nouvel Observateur* du 23 janvier 1997.
(2) *Le Figaro* du 2 septembre 1997.

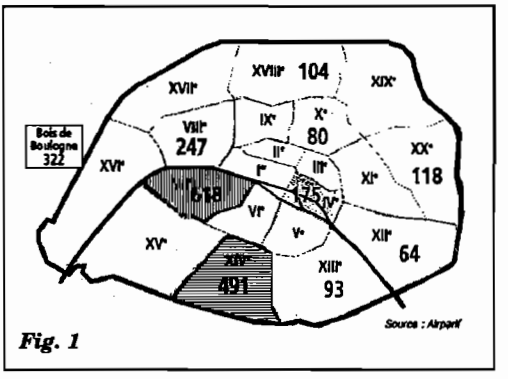


Fig. 1

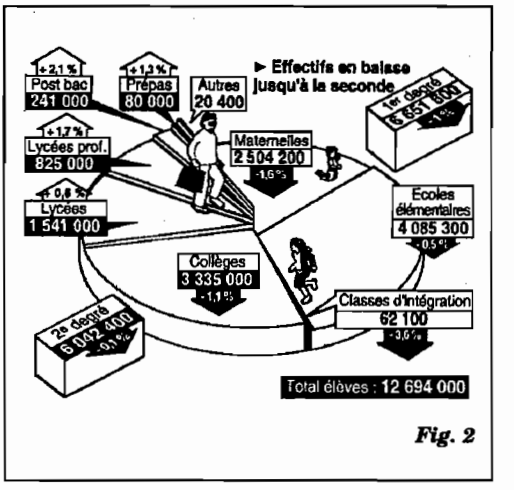


Fig. 2

Et, comme le dit un manuel [3] : *"le souci de présenter des documents de façon attrayante et le développement des moyens informatiques conduisent à une très grande variété de représentations graphiques"*.

On peut, à propos de ces représentations diverses et variées, se poser à juste titre les questions du pourquoi et du comment.

a) le pourquoi : on retrouve ici certaines fonctions bien connues des dessins en mathématiques, et en particulier celles de synthèse de l'information et de contrôle des résultats. De ce point de vue, un graphique remplira son office s'il préserve au mieux l'information recueillie et mise en forme, et s'il permet éventuellement un contrôle ultérieur (au moins de vraisemblance) de certains paramètres qui seront calculés par la suite, ainsi qu'une comparaison avec d'autres séries. Cette question du pourquoi débouche bien sûr sur celle des moyens utilisés, c'est-à-dire du comment.

b) le comment : il s'agit en fait de passer du *su* (le tableau de données, en général) au *perçu* (la représentation graphique qu'on lui associe), c'est-à-dire de changer de registre de représentation ; de nouvelles questions s'ensuivent, liées à celles du pourquoi, et parmi celles-ci : l'information initiale est-elle préservée ? le graphique est-il adapté au type de variable en jeu ? est-il compréhensible ? etc. Pour reprendre la terminologie utilisée par Raymond Duval ([1]), la question est d'étudier la plus ou moins grande *congruence sémantique* entre deux registres : d'une part le tableau de données, et d'autre part le graphique utilisé pour le représenter.

Je ne tenterai certes pas ici une étude exhaustive des divers types de représentations que l'on peut associer à un tableau de données statistiques, car elle dépasserait largement le cadre d'un article. Je me contenterai simplement de donner quelques idées générales, permettant une bonne adéquation du graphique aux données, et partant une bonne efficacité de cet outil. Mais il faut pour cela définir ce que peut être une "bonne adéquation". J.-F. Pichard ([4] pp. 76-77) distingue en particulier :

(3) *Mathématiques Seconde* (p. 148), par B. Joppin, I. Houdart et D. Guinin, Ed. Bréal 1990.

– la *lisibilité* : "*un graphique doit être plus directement et rapidement lisible que les données chiffrées*[...]"
– la *fidélité* : "*un graphique doit respecter les données et rendre fidèlement la réalité. L'impression visuelle ne doit pas conduire à déformer cette réalité* [...]"
– l'*auto-suffisance* : "*il doit pouvoir être compris, indépendamment de la série qu'il représente, par :*
– son titre qui désigne le phénomène de façon précise,
– le libellé des axes avec l'échelle retenue,
– l'indication des sources".

La dernière condition ne demande, somme toute, qu'un peu de vigilance de la part de l'auteur ; mais le problème est moins aisé pour les deux autres, et ce sont elles que je vais maintenant examiner d'un peu plus près.

La préservation de l'information

On a vu que, du fait de regroupements possibles, l'information recueillie ne se retrouve pas toujours en totalité après la mise en forme (tableau). Mais, dans la plupart des cas courants, la question est celle de la représentation d'un nombre fini – et en général faible – de couples (modalité ; fréquence) ou (classe ; fréquence), la fréquence (l'effectif) pouvant être :

– soit celle (celui) de la modalité considérée du caractère (statistique simple)
– soit celle (celui) d'un second caractère au sein de la modalité considérée du premier caractère (cas particulier de statistique double).

On pourrait dire, en reprenant la terminologie de J.-F. Pichard ([4] pp. 77-80), que le premier cas correspond à des *situations de partition* (les diverses modalités déterminent une partition de la population), tandis que le second correspond à des *situations de fonction* (la fréquence, ou l'effectif, du second caractère est fonction des modalités du premier) (4).

Puisqu'il s'agit dans tous les cas de couples, leur représentation dans le plan ne présente pas de problème, et on peut distinguer, dans les pratiques, trois modes principaux :

– selon le couple de directions horizontal / vertical : diagramme en bâtons, diagramme en barres, histogramme, polygone ;
– selon un arrangement circulaire : diagramme circulaire ("camembert") ou semi-circulaire ("demi-camembert") ;
– selon un procédé de type analogique (du point de vue des modalités) : c'est le cas de la *fig. 1* ci-dessus et de la *fig. 3*, relative à l'origine des pièces de la Ford Escort – fabriquée en Grande-Bretagne et en Allemagne –, et dans laquelle la situation géographique des pays est suggérée par la disposition relative des disques représentant le nombre de pièces produites par chacun (5).

N.B. : Je me bornerai ici – sauf exception – aux deux premiers modes, qui restent malgré tout d'une plus grande généralité, et sont plus largement utilisés.

(4) En fait, Pichard parle de "graphique de partition" et de "graphique de fonction", mais je préfère distinguer la situation de sa représentation graphique.
(5) *Géographie Terminales* (p. 57) par M. Hagnerelle *et al.*, Ed. Magnard 1995.

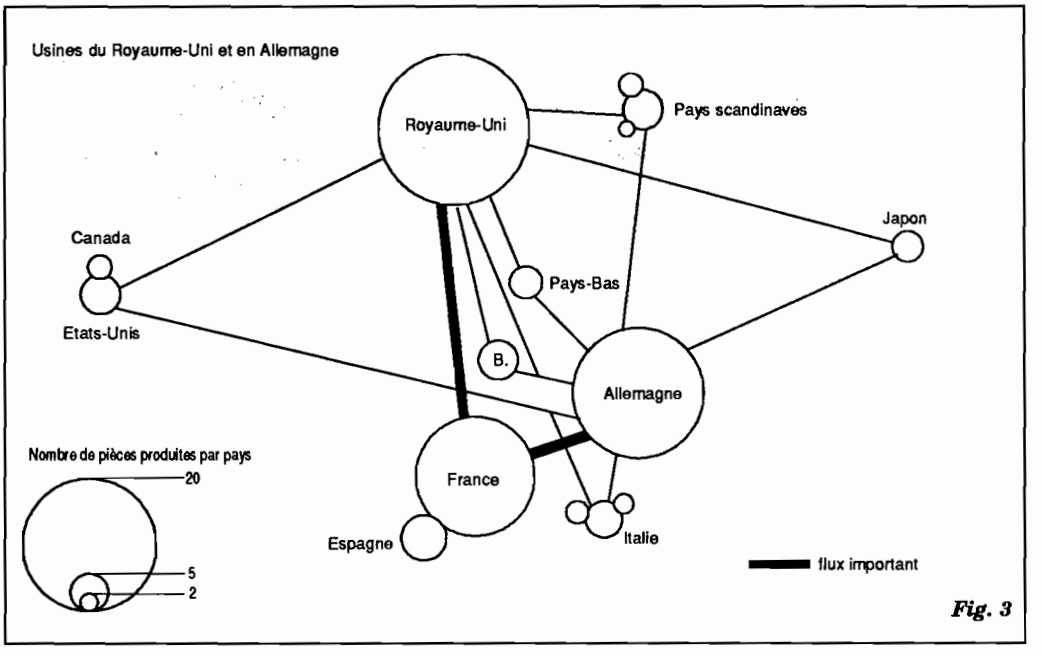


Fig. 3

L'adaptation du graphique à la variable

La nature de la variable prise comme base (celle qui correspond aux modalités considérées) rend plus ou moins pertinente une interprétation de certains éléments de la représentation graphique. Ainsi, par exemple, un *diagramme en barres* (non jointives) – encore appelé *diagramme à bandes*, voire *diagramme en tuyaux d'orgue* – conviendra pour une variable qualitative, telle que l'"énergie primaire" (*fig. 4*, où l'on s'intéresse à la consommation française en 1985 (6)) : il y a bien un axe vertical (celui des fréquences), mais il ne saurait y avoir d'axe (horizontal) des modalités de la variable ; par contre, dans le cas d'une variable quantitative discrète, on pourra donner un sens à la *position sur un axe* des différentes valeurs, et on préférera, pour cette raison, les *bâtons* aux barres (*fig. 5*, relative à la répartition des

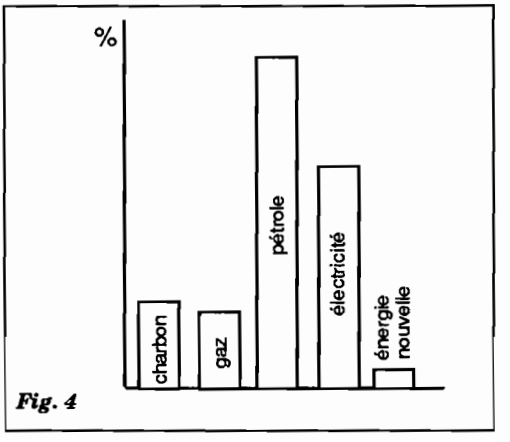


Fig. 4

(6) *Mathématiques Cinquième*. Nouvelle collection Durrande, Ed. Bordas 1987.

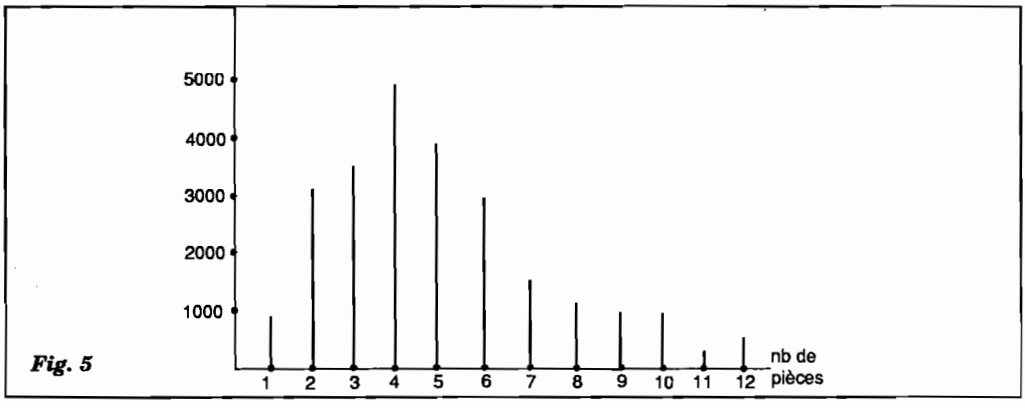


Fig. 5

logements d'une ville moyenne selon le nombre de pièces qu'ils comportent (7)). On peut remarquer que les bâtons ne sont utiles que par leur extrémité supérieure, et ne servent en fait qu'à "faire voir" cette extrémité. Dans le même ordre d'idées, les barres sont souvent préférées aux bâtons, même dans le cas d'un caractère ordinal ; c'est sans doute également pour des raisons d'impact visuel.

L'*histogramme*, de par la continuité "horizontale" qu'il visualise, correspond bien à une variable continue représentée sur l'axe des abscisses (*fig. 6* (8), relative à la répartition des femmes françaises dans la vie active, selon leur âge, en 1975, selon l'INSEE. *N.B. :* l'effectif de la classe [15 ; 25[est de 1 817 000). Un tel graphique présente l'avantage d'indiquer les limites de classes (on a vu plus haut l'importance de leur préservation), mais il faudra se souvenir que la hauteur des rectangles correspond, non à la fréquence, mais à la *densité* des classes (rapport effectif / amplitude) Nous verrons plus loin pourquoi. Cependant, dans le cas particulier assez fréquent où les classes sont toutes de même amplitude, cette hauteur est également proportionnelle à la fréquence. Notons enfin que la ligne supérieure de l'histogramme peut s'interpréter comme la représentation graphique de la densité, *supposée constante dans chaque classe.*

Autre représentation fréquente, le *polygone* s'obtient, dans le cas d'une variable quantitative discrète, d'une façon

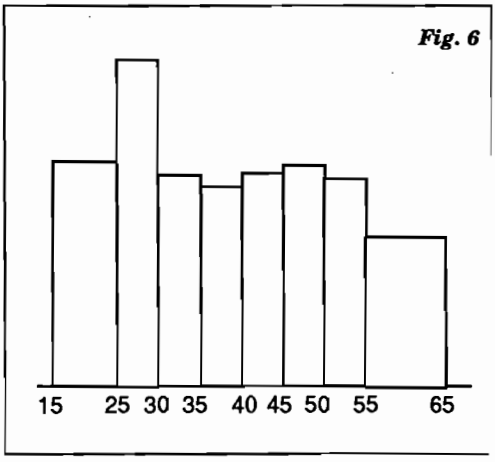


Fig. 6

(7) D'après *Statistiques Bac Pro*, par J. Enel et F. Leiritz, Ed. IREM de Lorraine 1990.

(8) *Mathématiques Seconde* (p. 374). Coll. Fractale, Ed. Bordas 1990.

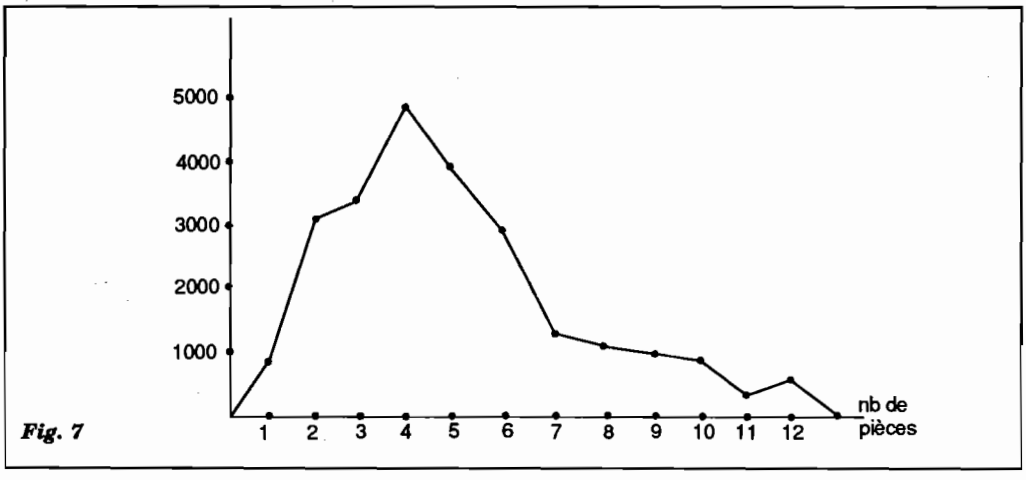


Fig. 7

analogue au diagramme en bâtons (*fig. 7*, obtenue à partir de la fig. 5). Remarquons que l'on peut également y avoir recours dans le cas d'une variable continue, à condition de la "discrétiser" (mais il est pour cela nécessaire que les classes aient toutes la même amplitude) ; dans ce cas, on substitue à chaque classe son centre, affecté de la fréquence de la classe.

La question se pose de l'interprétation des segments de la ligne brisée. Si la variable est continue, cela revient à effectuer une interpolation "linéaire" (9) entre deux centres de classes consécutifs, et donc à remplacer les rectangles de l'histogramme par des trapèzes (dont l'aire totale est identique à celle des rectangles). Mais, si la variable est discrète, ces segments n'ont pas de sens, et leur seule fonction (comme plus haut celle des bâtons) est de permettre une meilleure visualisation des sommets des barres "virtuelles" ; il vaut donc mieux ne pas l'utiliser, tout au moins avec des élèves jeunes, afin d'éviter une éventuelle utilisation non pertinente (non congruence sémantique).

(9) Il s'agit en fait d'une interpolation *affine*, mais je me conforme à la terminologie usuelle.

Histogramme et polygone

Plaçons-nous maintenant dans le cas d'une variable quantitative X, pour laquelle le tableau de données donne la fréquence de chaque classe :

Classe	[3;5[	[5;7[	[7;9[	[9;11[	[11;13[	[13;15[	Total
Centre	4	6	8	10	12	14	
Fréquence (%)	4	14	28	33	16	5	100

(On remarquera qu'ici toutes les classes ont la même amplitude.) L'histogramme correspondant est celui de la *fig. 8* , et le polygone celui de la *fig. 9*. On a déjà vu que, dans la représentation d'une série de fréquences par un histogramme, donner un sens à la ligne en escalier supérieure revient à considérer que la distribution est *uniforme* à l'intérieur de chaque classe. Pour ce qui est du polygone, l'hypothèse est

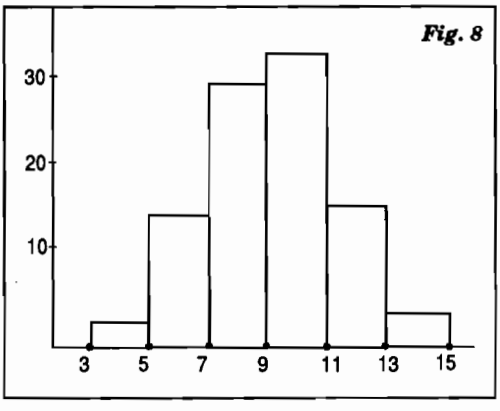


Fig. 8

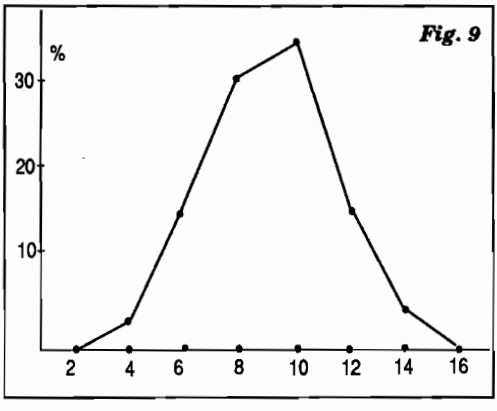


Fig. 9

cette fois que la distribution est *affine* du centre d'une classe à celui de la suivante. Par exemple, dans le cas représenté, la fréquence des individus pour lesquels on a $7 \leq X < 8$ est *(fig. 10)* :

– dans le cas de l'histogramme (aire du rectangle) : $\frac{28}{2} = 14\ \%$

– dans le cas du polygone (aire du trapèze) :

$\frac{1}{2}\ [(7 + \frac{14-7}{2}) + 14] = 12{,}25\ \%.$

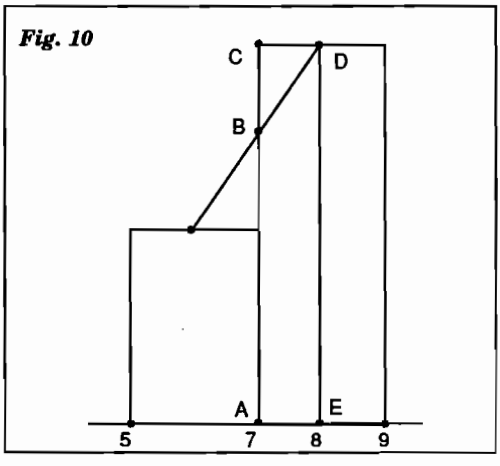


Fig. 10

On peut remarquer que, pour une distribution de type courant (c'est-à-dire unimodale "en cloche"), l'hypothèse sur la répartition faite implicitement dans le cas du polygone est certainement plus proche de la réalité que l'autre. Mais, si l'on se pose la question du "retour inverse", c'est-à-dire celle de la reconstruction du tableau de fréquences à partir de la représentation graphique, on voit que, si elle est immédiate dans le cas de l'histogramme, le polygone demande quelques calculs pour la détermination des limites, ainsi que celle des effectifs des classes si elles sont d'amplitudes différentes. Finalement, on voit donc que l'histogramme est plus facile à encoder et à décoder que le polygone, car il est plus congruent que celui-ci au tableau qui a servi à le créer. C'est sans doute ce qui explique son succès.

Des fréquences aux fréquences cumulées

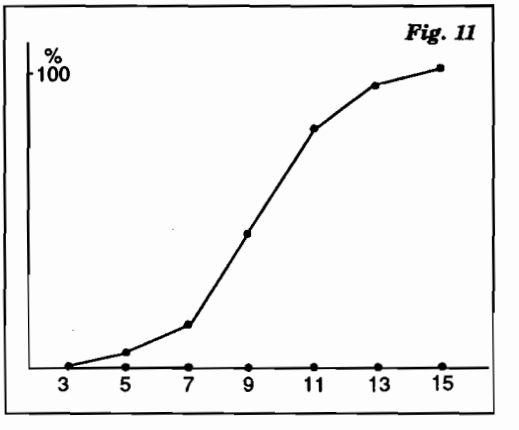
Dans la pratique, une autre question se pose à propos de l'histogramme : celle du passage des fréquences aux fréquences cumulées (croissantes, par exemple). A partir du tableau initial, seules les fréquences cumulées correspondant aux

valeurs extrêmes des classes peuvent être déterminées. On obtient ainsi le tableau suivant :

Valeur	3	5	7	9	11	13	15
Fréq. cum. %	0	4	18	46	79	95	100

Or, la fréquence cumulée est une fonction continue (c'est l'intégrale de la densité) ; il s'agit donc de l'interpoler, à partir des valeurs connues de cette fonction. Le moyen le plus courant (et le plus simple) consiste à interpoler linéairement, c'est-à-dire à supposer de nouveau que la densité est constante à l'intérieur de chaque classe. On obtient alors une courbe de type "polygone" *(fig. 11)*, et non pas un "histogramme cumulé", comme on le voit parfois.

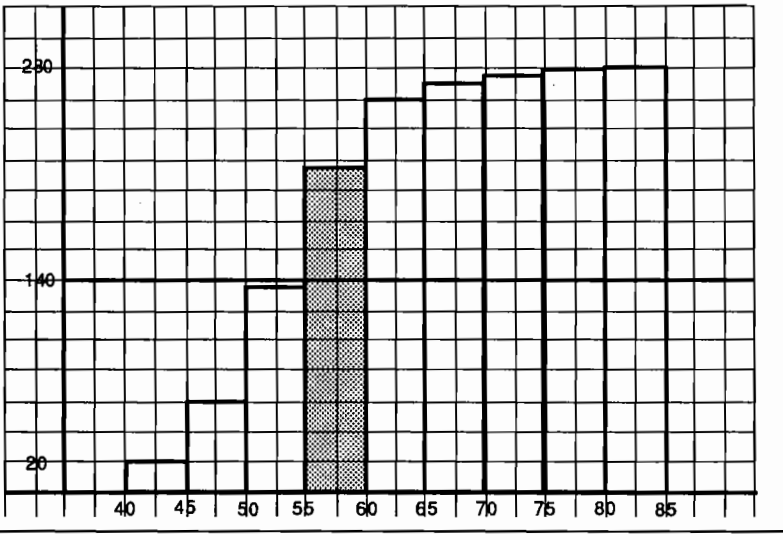
Fig. 11



Voici par exemple un graphique *(fig. 12)*, extrait d'un manuel de Seconde de 1990, dans lequel un tel "histogramme" (d'effectifs, en l'occurrence) est utilisé dans la recherche graphique de la médiane d'une série statistique. Les intentions – louables – des auteurs sont :

– d'une part, de donner du sens à la notion de médiane, à l'aide d'une conversion de registres (symbolique → graphique) ;

Fig. 12



– d'autre part, de montrer que l'information dont on dispose ne permet pas de déterminer la médiane, mais seulement une classe médiane, dans laquelle on est sûr que se trouve la médiane.

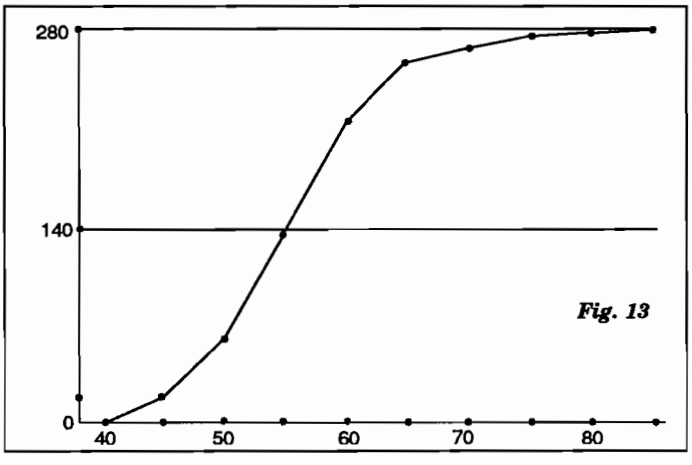


Fig. 13

Mais, à la lecture, des questions viennent à l'esprit : Quelle signification peut-on donner à l'aire du rectangle grisé ? Que représente-t-elle ? Dans le cas d'un histogramme d'effectifs (non cumulés) à classes de même amplitude, cette aire est proportionnelle à l'effectif de la classe considérée, mais ici ? En fait, dans le cas présent seul le sommet de ce rectangle situé "en haut à droite" peut être interprété : il représente l'effectif cumulé de la *valeur* 60 (et non celui de la *classe* 55-60).

En ce qui concerne la question de la médiane, le recours au polygone (cumulé) conduit, par le même procédé graphique *(fig. 13)*, à la détermination d'une valeur médiane, moyennant l'hypothèse supplémentaire que la densité est constante à l'intérieur des classes (10). Pour présenter la notion de classe médiane aux élèves, on peut leur faire prendre conscience du fait que, si l'on se refuse à faire cette hypothèse, on est contraint de se rabattre sur un résultat moins fort : la valeur médiane (inconnue) se trouve certainement à l'intérieur d'une classe : celle qui contient la valeur fictive trouvée précédemment, et c'est cette classe que l'on appelle la classe médiane.

Diagramme circulaire ou semi-circulaire ?

Le diagramme circulaire et le diagramme semi-circulaire s'appliquent tous deux aux variables qualitatives, et on peut constater que la plupart des manuels du secondaire ne les différencient pas du point de vue des utilisations. Ainsi : *"la population est représentée par un cercle ou un demi-cercle et à chaque valeur on fait correspondre un secteur, d'aire proportionnelle à la fréquence"* (11), ou encore : *"On peut découper un disque (ou un demi-disque) en secteurs circulaires dont les mesures sont proportionnelles aux pourcentages associés à chaque modalité"* (12). Ou

(10) En fait, la plupart des manuels utilisent l'intersection des polygones des fréquences cumulées croissantes et décroissantes pour déterminer la médiane. On peut se poser la question de l'intérêt pédagogique de ce procédé, par rapport à l'intersection avec la droite horizontale dont l'ordonnée correspond à la moitié de l'effectif.

(11) *Mathématiques Seconde*, par L. Corrieu *et al.*, Ed. Delagrave 1990.

(12) *Mathématiques Seconde* (p. 168), Coll. Fredon, Ed. A. Colin 1990.

Fig. 14

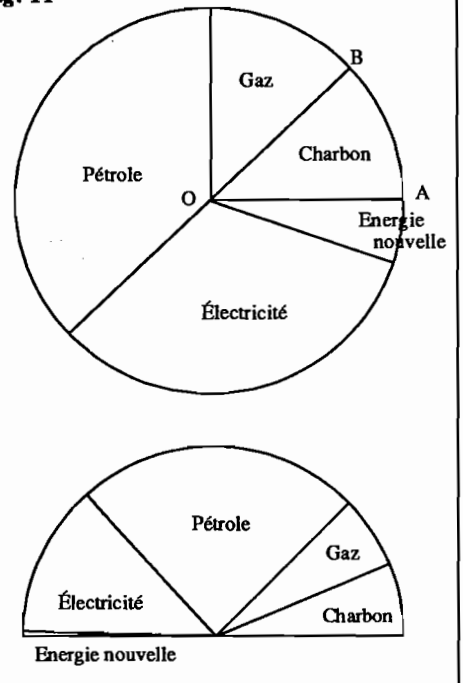
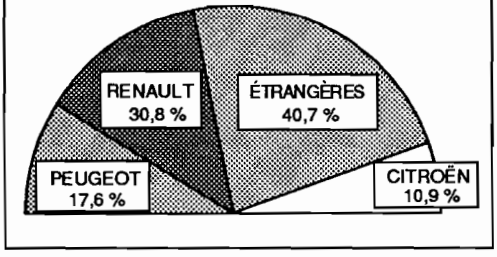


Fig. 15



bien, joignant l'image à la parole, on place côte à côte les deux représentations, correspondant à une même série statistique, en l'occurrence les énergies primaires consommées en France en 1985 *(fig. 14)*. situation déjà rencontrée à la *fig. 4*. Mais ces deux types de représentations sont-ils *vraiment* interchangeables ?

Considérons un disque divisé en secteurs ; ceux-ci ne présentent pas *a priori* de structure d'ordre, sauf à préciser un plus petit élément (c'est-à-dire un rayon origine) et le sens croissant. Par contre, sur un demi-disque divisé en secteurs, deux structures d'ordre (opposées) se lisent immédiatement. C'est pourquoi il apparaît préférable de réserver plutôt les diagrammes circulaires à l'étude des variables nominales (comme c'est le cas sur l'exemple de la *fig.* 14), et d'utiliser plutôt un diagramme semi-circulaire dans le cas d'une variable ordinale. Cependant, dans un cas comme dans l'autre, la *contiguïté* – ou la non-contiguïté – de deux modalités peut poser problème : faut-il la considérer comme signifiante ou non ? Prenons l'exemple de la *fig. 15* (13), relative aux immatriculations d'automobiles en France en septembre 1989 ; deux questions – sans réponse – viennent en effet immédiatement à l'esprit lorsqu'on regarde ce graphique :

– Pourquoi avoir séparé Citroën des autres marques françaises ?
– Pourquoi ne pas l'avoir mise à côté de Peugeot (les deux marques font en effet partie du groupe PSA) ?

Outre le fait qu'un diagramme circulaire aurait été plus indiqué, on aurait pu ici prendre en compte ces deux critères de regroupement (nationalité, appartenance à un même groupe économique), ce qui aurait accru la congruence sémantique du graphique.

(13) *Mathématiques Sixième* (p. 231), Coll. Transmath, Ed. Nathan 1990.

40%	15%	20%	7%	11%	7%
Nourriture	Transport	Logement	Vêtements	Energie	Loisirs

Fig. 16

La question n'en subsiste pas moins de savoir si, dans un diagramme circulaire, toutes les places se valent (visuellement, s'entend), ou si l'on peut mettre en relief certaines modalités en les situant à une place particulière. Nous y reviendrons plus loin. En attendant, nous pouvons regrouper dans un tableau les modes de représentation graphique *a priori* les mieux adaptés aux divers types de caractères statistiques que nous avons identifiés au début :

caractère qualitatif	nominal	*en barres, circulaire*
	ordinal	*en barres, semi-circulaire*
caractère quantitatif	discret	*en bâtons, polygones*
	continu	*histogramme (polygone)*

L'adaptation du graphique à la situation

Considérons de nouveau la distinction entre *situations de fonction* et *situations de partition*. On a vu que, dans les premières, on étudie, en fonction des modalités d'un caractère, l'évolution de la fréquence (ou de l'effectif) d'un autre caractère ; ainsi, les situations correspondant à la figure 1 et aux figures 20 à 23 ci-après, sont des situations de fonction. Les secondes se caractérisent par le fait qu'on s'y intéresse à la répartition d'un "tout" (la population de référence) selon les diverses modalités d'un caractère ; à chaque modalité correspond donc une sous-population (celle qui présente la modalité concernée), et par suite l'effectif et la fréquence de cette sous-population. C'est le cas, parmi les exemples précédents, des situations qui correspondent aux figures 4 à 6.

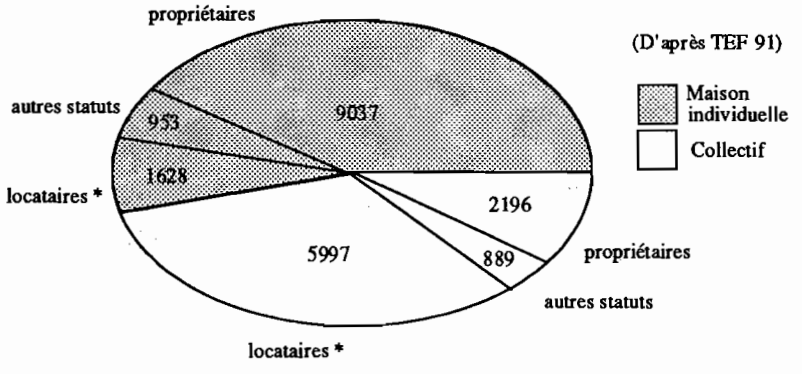
A la lumière de cette distinction, il apparaît alors intéressant – afin d'optimiser la congruence sémantique entre situation et représentation – de réserver plus particulièrement aux situations de fonction des représentations graphiques dans laquelle l'idée d'un "tout" est visuellement présente, de par la *connexité* de la partie signifiante. C'est le cas de l'histogramme, du diagramme circulaire et semi-circulaire, ainsi que du *rectangle subdivisé* (*fig. 16*, d'après **[4]** (14), relative à la ventilation d'un budget familial suivant divers postes).

Le "point de vue" visuel

Les considérations précédentes, relatives aux diagrammes circulaires et semi-circulaires, nous amènent à prendre en compte le rôle fondamental de la perception dans l'utilisation des graphiques. Une représentation graphique étant un outil *visuel*, il semble aller de soi que c'est cet aspect qui doit présider à l'encodage comme au décodage des données statistiques. C'est ce qui explique que, comme nous l'avons vu, dans le cas d'une variable quantitative continue représentée par un histogramme ce sont les *aires* des rectangles qui doivent être proportionnelles aux fréquences, et non leur hauteur. Et c'est bien ce qui en est dit aux élèves. Par exemple :

(14) p. 80.

Fig. 17



* locataires d'un local loué vide

"Chaque couple (classe, effectif) (ou (classe, fréquence)) est représenté par un rectangle [...] *dont l'aire est proportionnelle à l'effectif de la classe"* (15)

"Chaque modalité est représentée par un rectangle dont l'aire est proportionnelle à l'effectif ou à la fréquence" (16), etc.

Il en est de même dans le cas d'un diagramme circulaire : ce sont les aires des secteurs qui doivent être proportionnelles aux fréquences. Mais ici la justification "visuelle" – qui seule pourrait donner du sens à la règle de construction – est fréquemment escamotée par les manuels (et par les enseignants ?). Cependant, si le secteur a une *"aire proportionnelle à la fréquence"*, c'est tout bonnement parce que, étant donné un disque, l'aire d'un secteur circulaire est proportionnelle à son angle !

Autre problème qui mérite qu'on s'y intéresse : le diagramme circulaire est de plus en plus fréquemment représenté sous la forme d'un cylindre "en perspective". Faisons pour l'instant abstraction de l'épaisseur du cylindre, pour ne considérer que la représentation de sa face supérieure, c'est-à-dire une ellipse *(fig. 17* (17), relative à la répartition du parc des résidences principales entre locataires et propriétaires selon le type d'habitat, en France, en 1988). Le passage du disque à l'ellipse s'effectue alors sans problème grâce à une affinité orthogonale ayant pour axe un diamètre du cercle *(fig. 18)* ; de plus on sait que, dans une affinité, l'aire de l'image d'un secteur est proportionnelle à celle de ce secteur (dans le rapport d'affinité) : l'aspect visuel est donc préservé dans le "passage en perspective".

(15) *Mathématiques Seconde*, par L. Corrieu *et al.*, Ed. Delagrave 1990.

(16) *Mathématiques Seconde*, Coll. Dimathème, Ed. Didier 1990.

(17) D'après *Mathématiques Terminale B* (p. 181), Coll. Déclic, Ed. Hachette 1992. J'ai "gommé" l'épaisseur du graphique, pour plus de clarté. D'autre part, remarquons qu'ici sont simultanément étudiés deux caractères : la nature du logement (2 modalités) et le type d'occupant (3 modalités).

Fig. 18

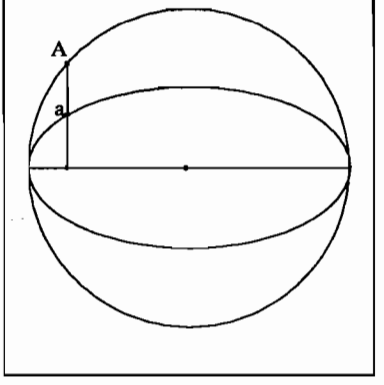
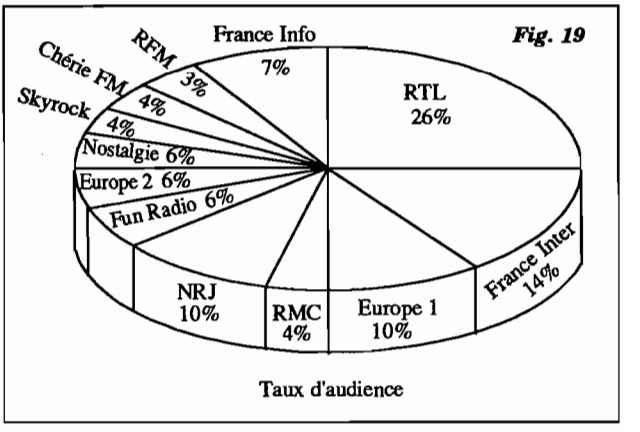


Fig. 19



Prenons maintenant en compte l'"épaisseur" du cylindre représenté *(fig. 19* (18), concernant le taux d'audience des stations de radio françaises (19)). La différence de traitement est alors nette entre les secteurs situés "en arrière" et ceux qui se trouvent "en avant" : pour ceux-ci, en effet, l'impact visuel est renforcé par la "tranche" visible du cylindre, et ils se trouvent de la sorte favorisés par rapport aux autres. Sur la *fig.19*, l'aire "réelle" de la surface correspondant à France Inter (14% d'audience) est – à cause de la "tranche" – pratiquement la même que celle qui correspond à RTL (26% d'audience). De même – et c'est sans doute encore plus net – la part revenant à RMC *est la même* (4%) que celle de Chérie FM, mais sur le diagramme elle *apparaît* bien plus importante... On voit le parti que pourraient en tirer des gens peu scrupuleux, tablant sur une lecture superficielle du graphique.

(18) *Mathématiques Seconde* (p. 186), par D. Guinin et B. Joppin, Ed. Bréal 1997.

(19) De septembre 1995 à juin 1996, du lundi au vendredi de 5h à 24h. Enquête Médiamétrie.

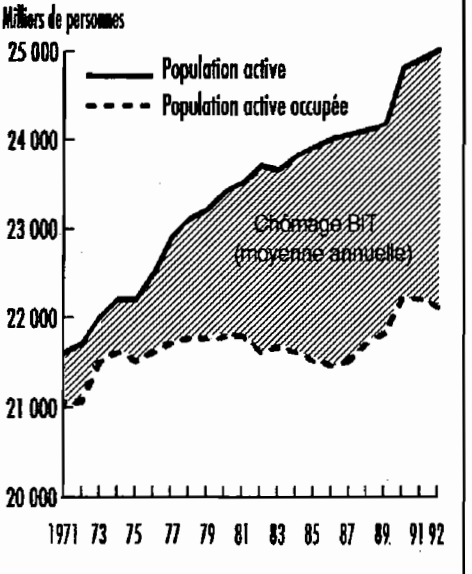
Ayons l'œil

Ne faisons pourtant pas de procès d'intention : la plupart des procédés utilisés par les auteurs de graphiques statistiques sont sous-tendus par une intention louable : mettre en évidence certains phénomènes, en les amplifiant. Cependant, si l'on n'y prend garde, on risque de prendre pour argent comptant les caractéristiques perçues, et être ainsi induit en erreur. Je n'envisagerai ici que deux types de procédés, parmi bien d'autres :

a) le changement d'origine sur un axe : par exemple, lorsque les variations du caractère considéré sont relativement faibles, on peut être tenté de ne faire apparaître que leur partie "utile", c'est-à-dire le "haut" du graphique habituel. Pour illustrer ceci, je me contenterai de deux exemples :

– la première représentation *(fig 20)*, extraite d'un manuel de sciences économiques (20), est une série chronologique

(20) *Economie générale Terminale STT*, par A. Serdeczny *et al.*, Ed. Bréal 1994.



TEF 1993/1994, INSEE.

Fig. 20

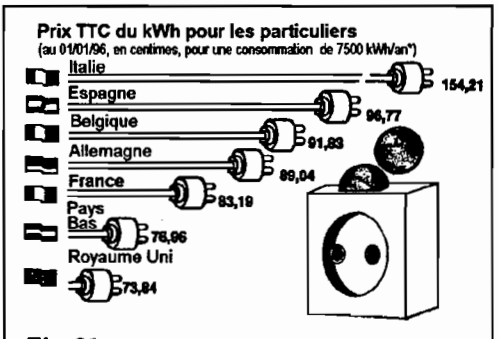


Fig. 21

concernant l'évolution des effectifs de la population active et de la population active occupée en France, entre 1971 et 1992. Une "lecture" rapide du graphique pourrait laisser croire, par comparaison des ordonnées, que le nombre de chômeurs est au moins aussi grand que le nombre de personnes occupées... Cependant, un coup d'œil à l'échelle verticale rectifie cette interprétation erronée.

– la seconde représentation *(fig. 21)* est relative au prix du kilowatt-heure d'électricité, en France et dans les pays voisins [21]. Tout d'abord – et pour agrémenter le graphique – les barres du diagramme ont été représentées sous la forme de cordons électriques munis d'une prise de courant (identique pour tous), et on s'aperçoit immédiatement que la barre de l'Italie n'est pas représentée en entier (l'électricité coûte donc beaucoup plus cher dans ce pays qu'ailleurs). En outre, un examen détaillé de la représentation (avec triple décimètre et calculatrice, dans le but de rechercher une proportionnalité prix/longueur) fait apparaître les faits suivants :

– seul le cordon est à prendre en compte dans la longueur de la barre, à l'exclusion de la prise de courant ;
– toutes les barres ont été "tronquées" au niveau de celle du Royaume-Uni (qui n'apparaît donc plus), pays correspondant au tarif le plus bas.

En fait, pour respecter l'échelle du diagramme, il faudrait allonger toutes les barres d'environ 18 cm [22] (et ajouter environ 14 cm à la barre de l'Italie). Ce qui, étant donné la place disponible dans le journal, se traduirait en fait par une diminution d'échelle de l'ensemble, rédui-

(21) *Le Républicain Lorrain* du 13 février 1997.

(22) A titre de comparaison, la barre de l'Espagne mesure 5,5 cm (hors prise).

sant nettement, visuellement parlant, les différences entre pays. Il n'en reste pas moins qu'un lecteur qui ne prendrait pas en compte les valeurs portées en bout de barre pourrait en déduire :

– que le prix de l'électricité domestique est beaucoup plus faible au Royaume-Uni qu'en France (alors qu'il n'est que 11 % moins cher),
– qu'il est beaucoup plus élevé en Espagne qu'en France (alors qu'il n'est que 16 % plus cher).

Ces différences, quoique non négligeables, apparaissent fort exagérées sur le diagramme, et l'adjonction de caractères iconiques (le cordon et la prise) n'est pas fait pour en faciliter la lecture. Bien sûr, il ne faut pas se contenter de regarder ces graphiques à la va-vite. Mais il n'en reste pas moins que ces exemples montrent qu'il nous faut apprendre à nos élèves à se montrer vigilants, et à ne pas conclure trop vite ; en outre, on voit sur ces exemples que la suppression des données chiffrées pourrait laisser la porte ouverte à toutes les manipulations.

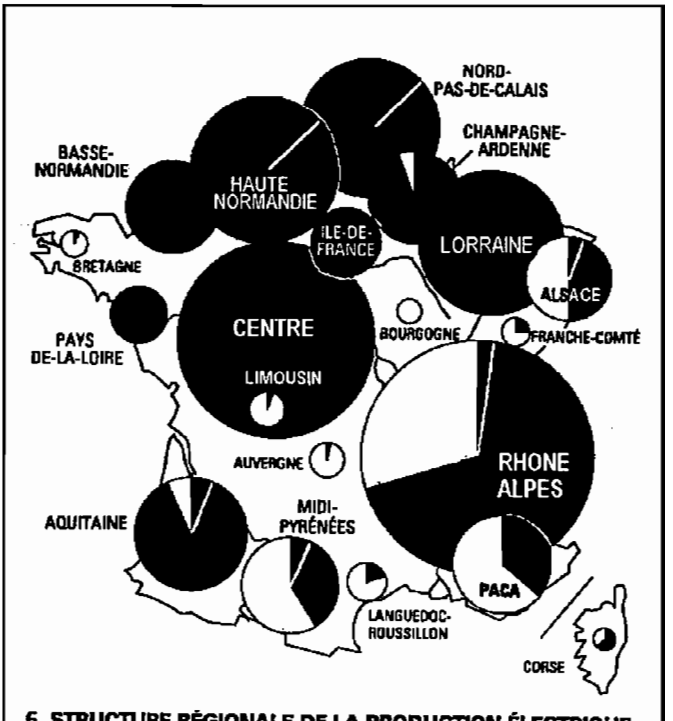
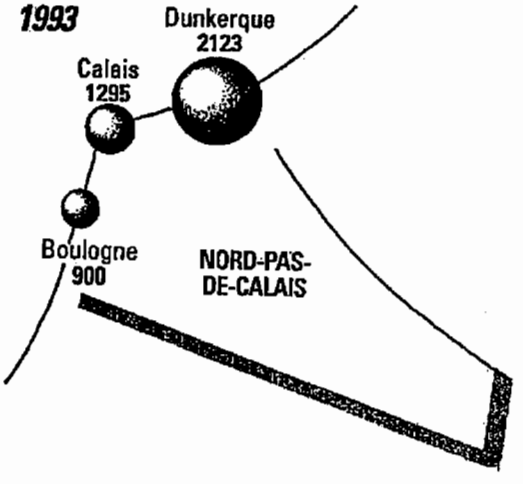


Fig. 22

b) le passage en 2 ou 3 D : il a en fait pour but de donner à voir d'un seul coup d'œil, et/ou de donner – littéralement – du "relief" au graphique, mais, comme nous en avons déjà vu un exemple ci-dessus avec les "camemberts", la lecture risque d'en être faussée. En voici deux autres exemples encore plus criants, si possible :

– le premier... n'est autre que celui de la *fig. 3* : chacun des pays fabricants est représenté par un disque, dont la taille varie en fonction de l'effectif fabriqué.



Source : université du Littoral.

Fig. 23

du Littoral en 1993 *(fig. 23)*. L'effectif de chaque site est indiqué par un nombre, accompagné d'un symbole le représentant, qui apparaît comme une "boule" grâce à un effet d'ombrage. Or, si l'on examine les diamètres de ces boules, on peut constater qu'ils sont *grosso modo* proportionnels aux effectifs. Cependant, visuellement le volume de la grosse boule apparaît énorme en comparaison de celui de la boule moyenne ; et pourtant elle correspond à un effectif qui n'en atteint pas le double ! Il y a, ici aussi, contradiction entre la proportionnalité attendue (effectif / volume) et celle qui est présentée (effectif / diamètre), donc non-congruence sémantique.

L'échelle des tailles est indiquée dans le coin inférieur gauche, et l'on peut s'apercevoir que c'est le diamètre – et non l'aire – qui est proportionnel à l'effectif. Le résultat en est qu'un disque qui, visuellement, apparaît 10 fois plus gros qu'un autre, ne correspond en fait qu'à un effectif guère plus de 3 fois plus grand (comparer par exemple l'Espagne à la France sur le diagramme) ! Notons toutefois que certains auteurs représentent des disques pour lesquels c'est bien l'aire qui est proportionnelle à l'effectif ou la fréquence (*fig. 22*, relative à la production d'électricité par région en France en 1992 [23]).

– le second [24] est relatif à l'effectif d'étudiants des trois sites de l'université

(23) *Géographie Premières L*, ES, S (p. 149), par R. Knafrou, Ed. Belin 1994.

(24) *Le Monde de l'Education* n° 209 (novembre 1993).

La lisibilité du graphique

Outre la question de la détermination des éléments pertinents du graphique, déjà évoquée, se pose celle de la plus ou moins grande facilité de "décodage" par le lecteur. A propos des histogrammes, J.-F. Pichard fait remarquer, dans son article déjà cité, que *"des études de perception visuelle ont prouvé que l'œil juge mal une surface vide (blanche)"* ([4] p. 92). Il conseille en conséquence de tramer les rectangles de l'histogramme, ce qui permet de différencier clairement – si l'on peut dire – l'intérieur de l'extérieur.

D'autre part, Aline Jelinski ([2]) s'est intéressée à la comparaison, dans ce domaine, des diagrammes "orthogonaux" (du type histogramme) et des diagrammes circulaires (du type "camembert"). Elle constate, en particulier :

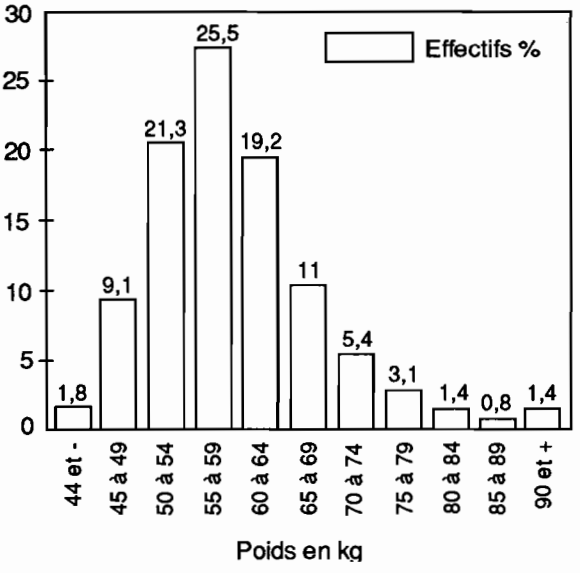


Fig. 24

– que la relation d'ordre sur les fréquences est nettement mieux perçue sur les colonnes d'un diagramme orthogonal que sur les secteurs d'un diagramme circulaire (et surtout pour les secteurs dont l'angle est "petit"),
– que l'évaluation quantitative est nettement plus difficile sur les secteurs que sur les colonnes.

Et elle conclut qu'il est préférable de ne pas avoir recours au diagramme circulaire lorsqu'il y a plus de 4 ou 5 classes. Cependant, elle ajoute que *"le graphique circulaire est plus utile et même irremplaçable dans le cas des données représentant une entité, un tout"* ([2] p. 56), le disque présentant visuellement, de par sa forme, un aspect clos et complet que ne possède pas – tout au moins au même degré – l'histogramme. C'est pourquoi il ne paraît pas judicieux d'amoindrir encore la congruence de ce dernier en le parcellisant, c'est-à-dire en séparant artificiellement les rectangles, le faisant ainsi ressembler à un diagramme en barres (*fig. 24* [25], représentant la distribution des poids des clientes d'une mutuelle d'assurance complémentaire).

Conclusion

Cette petite étude est certes loin d'être exhaustive, mais j'espère néanmoins avoir pu montrer que, en matière de représentations graphiques statistiques, *tout ne se vaut pas*, et qu'on a intérêt à rechercher un type de représentation approprié au type de situation, à la (aux) variable(s) en jeu (pour ne pas induire des propriétés inexi-

(25) *Mathématiques Seconde* (p. 201), par D. Guinin et B. Joppin, Ed. Bréal 1997.

stantes), et aussi lisible que possible (par le choix des classes éventuelles, des unités, le coloriage des zones signifiantes, etc.). Pourtant, il n'est pas question ici d'établir un dogme, d'autant plus qu'en général plusieurs possibilités sont tout à fait acceptables. Mon but est plutôt d'inciter enseignants et élèves à réfléchir un peu sur le sens de cet outil souvent marginalisé, et traité – comme on a pu s'en rendre compte – sans grand intérêt au sein d'un enseignement de la statistique lui-même souvent marginalisé. Et plus précisément, à se rendre compte qu'il convient d'être doublement vigilant :

1° en tant que "producteur" de graphiques statistiques, en essayant de ne pas induire (involontairement) le lecteur potentiel en erreur ;
2° en tant que "consommateur", en ne se contentant pas d'une lecture superficielle, mais en cherchant à savoir si le graphique est bien ce qu'il a l'air d'être ; et dans le doute, en s'abstenant de conclure.

Au delà de cette application particulière des représentations graphiques, c'est aussi la prise de conscience de la puissance des images dans la communication de l'information, avec tous les dérapages et les manipulations possibles, qui est en jeu.

Et enfin, pour ne pas terminer sur une note grave, je livre à votre méditation cette citation du regretté Alexandre Vialatte : "*Je recommande beaucoup* [le] *schéma statistique* [de M. Price], *où trois rectangles hachurés, soit obliquement, soit en chevrons, représentent comparativement le nombre d'Américains mâles atteints de schizophrénie en 1950, le même en 1960, et un échantillon du tissu que M. Price choisit pour sa veste de sport. De tels schémas font voir les choses*" (26).

Commentaire personnel (applicable à tous les graphiques) : **Certes, mais sont-ce bien celles que l'on était en droit d'attendre ?**

(26) *Antiquité du Grand Chosier* (1984), Chroniques choisies par Ferny Besson. Ed. Julliard, Paris.

RÉFÉRENCES BIBLIOGRAPHIQUES

[1] DUVAL Raymond (1996) : *Sémiosis et pensée humaine*. Ed. Peter Lang, Berne.

[2] JELINSKI Aline (1993) : "Diagramme circulaire ou orthogonal ? Une efficacité différente des images graphiques dans la transmission de l'information", *in Les Sciences de l'Education* n° 1-3, 39-56.

[3] LESSARD Sabin et MONGA (1993) : *Statistique. Concepts et méthodes,* Presses Universitaires du Québec. Diffusé par Masson, Paris.

[4] PICHARD Jean-François (1992) : "Représentations graphiques en statistiques", *in Bulletin inter-IREM* "Des chiffres et des lettres" (ed. IREM de Rouen), 75-101.

On pourra également consulter, dans *Repères-IREM*

ARNAUD René (1992) : "Peut-on commencer le premier trimestre de Seconde par les statistiques ?", *in Repères-IREM* n° 6, 20-26.

GIRARD Jean-Claude (1996) : "Pourquoi il ne faut pas laisser de côté les chapitres de statistiques au collège", *in Repères-IREM* n° 23, 5-18.

Le comité de rédaction de *Repères-Irem* aux anciens auteurs d'articles

Che(è)r(e) collègue,

Au cours des années passées, vous avez publié un ou plusieurs articles dans *Repères-Irem*. Il est souhaitable qu'ils entrent dans ***la base de données PUBLIMATH*** que l'APMEP et l'ADIREM s'efforcent de tenir à jour. Pour faciliter le travail des responsables de cette entreprise considérable, nous vous demandons de leur donner les informations suivantes, sur disquette (ou par courrier électronique) ***pour chaque article publié :***

a) L'auteur ou les auteurs (variable AUT1).

Indiquez le nom, suivi du prénom, séparés par une virgule. S'il y a plusieurs auteurs, séparez ces informations par un # (Exemple : AUT1 : **César, Jules # Einstein, Albert # Fourier, Joseph**).

b) Titre(s) (variables TIT1, TIT2).

Indiquez le titre (dans TIT1) et, éventuellement le sous-titre de l'article (dans TIT2).

c) Matériel utilisé dans l'article (variable MAT).

Indiquez le cas échéant le matériel mis en œuvre dans l'article (calculatrice, logiciel, tablette de rétroprojection, etc.)

d) Le public concerné (variables, AGE, NIV).

Indiquez éventuellement l'âge (AGE) et le niveau scolaire ou universitaire (NIV) des élèves ou étudiants concernés.

e) Résumé.

Proposez un résumé de 5 à 10 lignes *en*

caractères de taille 10. Ce résumé est capital pour ceux qui consultent PUBLIMATH. Il permet de vérifier l'adéquation de l'article avec leur requête. Chaque mot du résumé joue le rôle de MOT-CLE.

f) Notes.

Ce champ reçoit des informations difficilement classables ailleurs : par exemple la référence aux programmes officiels (avec possibilité de les lire intégralement), le rôle respectif des auteurs de l'article, les documents annexes, l'origine du document, le fait qu'il s'agit d'une conférence (son lieu et sa date) etc.

h) Bibliographie (variable BBL)

Il suffit de répondre OUI ou NON dans BBL, sans plus.

i) Mots-clés. (variable MCL)

Proposez une liste large, décrivant ***tous les aspects de l'article (entre 10 et 30 mots-clés).*** Signalez les mots-clés principaux et les secondaires.

Ce travail ne vous demandera pas beaucoup de temps (vous connaissez parfaitement vos textes) et il en évitera beaucoup aux responsables de PUBLIMATH (qui ne les connaissent pas...). Vous donnerez ainsi une seconde vie (qui peut être très longue) à vos articles : il seront consultés en France et à l'étranger par des collègues, des chercheurs, des étudiants *(près de 12500 consultations mensuelles actuellement...)*

Merci d'envoyer l'ensemble de ces informations sur disquette en WORD 5, 6 ou 97, ou plus simplement encore en courrier électronique (fichier joint) à :

Michèle Pécal,
260 Chemin des Cerisiers,
06740 Chateauneuf-de-Grasse
(Tél et fax 04 93 42 53 43)
email : pecal@unice.fr