Chapitre 3

Graphiques, diagrammes et autres messages visuels

Ce sont les **figures très colorées** sous forme de rectangles, de cercles ou de lignes brisées montantes ou descendantes, que l'on voit dans la presse, spécialisée ou non, dans les émissions de télévision, dans les rapports professionnels et même dans certaines publicités ou encore sur des boîtes de produits de consommation courante.

On ne peut pas y échapper : il y en a un peu partout, et il est vrai qu'on les « comprend » assez bien généralement **au premier coup d'œil.** Du moins on comprend **assez bien le** « **message** » qu'elles veulent faire passer : telle proportion de réussite au baccalauréat, telle part de cadres dans les entreprises du bâtiment, telle progression d'intentions de vote pour chaque candidat à l'élection présidentielle, telle croissance espérée de nos actions en bourse, etc.

En ce sens, **l'effort de communication** est indéniable. À la différence d'un tableau statistique (chapitre précédent), terne et monochrome en général, rempli de nombres plus ou moins longs, placés systématiquement les uns au-dessous des autres sans grande possibilité d'originalité, un peu « lourd », *un peu rébarbatif* pour tout dire, le diagramme ou le graphique[1] apparaît lumineux, éclatant, *attractif*.

Cependant, rien n'étant jamais parfait (même en statistique), il faut bien reconnaître que le graphique donne, de façon visuelle, plutôt des **ordres de grandeur** que des chiffres précis, contrairement au tableau. Le graphique intégrera, bien sûr, quelques chiffres exacts à la virgule près, placés dans une partie de cercle, ou sur le haut d'un bâtonnet, ou sur une échelle verticale, mais jamais on ne pourra donner tous les chiffres, au risque de le charger à tel point qu'il en devienne incompréhensible et dès lors « non communicant ».

Les deux instruments (*tableaux et graphiques*) sont donc complémentaires.

Chacun conserve sa fonction, *faire passer un message* ; l'un est complet et un peu triste, l'autre plus direct et plus gai, mais tous les deux sont *parfaitement rigoureux* notamment dans leur construction.

1. Les termes de *graphique* et de *diagramme* pourront, ici, être substitués sans trop de problèmes de sémantique. Cependant, on qualifiera plutôt de diagrammes les représentations visuelles générales et de graphiques celles qui sont plus particulièrement construites sur un repère normé en au moins deux axes d'abscisses et d'ordonnées.

Tout comme pour un tableau, le diagramme est évidemment dépendant de la *consigne de présentation* décrite en fin de chapitre précédent, respectant les principes : ***source-titre-intitulés-unités.***

Mais ici, il faut rajouter un principe : c'est la « **fameuse règle d'or** » des diagrammes, qui s'applique à toutes les figures statistiques que l'on peut voir, construire ou imaginer.

Elle est basée sur la **vitesse de compréhension** du lecteur (celui à qui s'adresse le message). Une fois que le lecteur a lu le titre du diagramme, repéré la source et déchiffré la légende, il ne faut pas qu'il dépasse **quelques secondes supplémentaires** pour comprendre le sens du message. Il peut rester béat d'admiration pendant l'heure qui suit s'il estime que le diagramme est magnifique, mais il aura **compris le message au tout début**, sinon ce n'est pas de la communication.

Règle d'or en matière de messages statistiques visuels

UN DIAGRAMME DOIT ÊTRE COMPRIS EN DIX SECONDES, UNE FOIS LE TITRE LU.

Précisons bien : si le lecteur dépasse la dizaine de secondes pour cette compréhension, cela ne veut surtout pas dire qu'il est idiot… bien au contraire, c'est que le graphe est mal fait (volontairement ou non). Cela arrive quelquefois quand le concepteur ne se met pas à la place de son lecteur. C'est donc dans ce cas **l'auteur du diagramme (et non l'utilisateur) qui est « limité » au point de ne pas avoir bien réalisé son devoir de « communicateur »** en la matière.

On ne construit pas un diagramme pour se faire plaisir ou pour rester la seule personne au monde à pouvoir le déchiffrer.

§ 1 – LES REPRÉSENTATIONS GRAPHIQUES À TRADUCTION DIRECTE

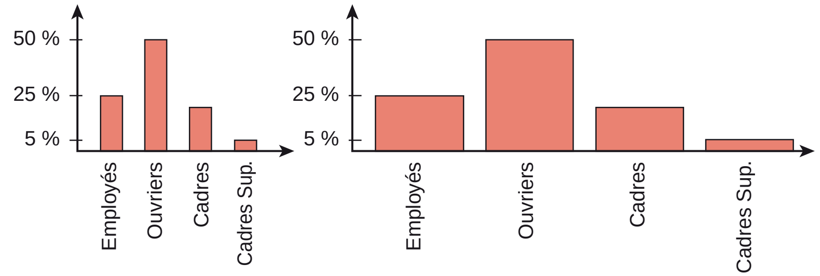
Ce sont celles, les plus classiques, qui n'utilisent **pas d'autres instruments de transformation que la géométrie traditionnelle pour leur construction** : on utilise ici des rectangles, des cercles, ou des courbes, mais pas des illustrations (pictogrammes ou figurines), ou de la cartographie, ou des échelles spéciales comme les échelles logarithmiques.

A – Les diagrammes de hauteur

L'effet visuel est ici rendu par l'appréciation de différentes *hauteurs* (et non pas, par exemple, de surfaces) des éléments qui composent le diagramme. Si un étroit rectangle vertical représente la proportion d'employés dans une entreprise, ce sera *sa hauteur* et non *sa base* qui exprimera la mesure, comme le montre l'exemple ci-après :

Diagramme en tuyaux d'orgues

Figure 3.1 – Répartition du personnel salarié de l'entreprise X par type d'activité au 1er janvier 2007

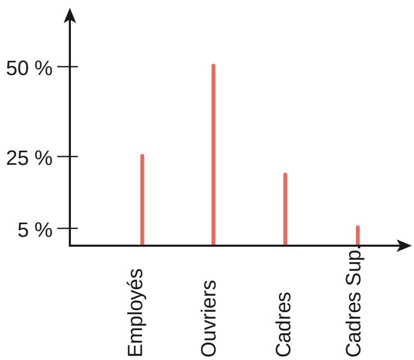


Plusieurs remarques peuvent être faites grâce à cet exemple :

- Ce que jusqu'à présent nous avons appelé « rectangles » va maintenant prendre sa terminologie réelle de « tuyaux d'orgues ». Il s'agit donc d'un **diagramme en tuyaux d'orgues**.
- On voit bien que la *hauteur* des tuyaux d'orgues est la même dans les deux cas de figure. Seule change la **dimension de la base**, et son choix ne dépend que du rendu esthétique que l'on cherche : quel diagramme préférons-nous ? Celui de droite ou celui de gauche ? Ou un qui soit entre les deux ? Tant que les hauteurs sont conservées **et les bases constantes**, nous pourrons faire comme il nous plaira.
- Dans cet exemple, on a affaire à un **diagramme à une dimension**, comme pour les tableaux. Il montre un seul caractère (la PCS) face à sa fréquence. On aurait pu le dessiner de la manière suivante :

Diagramme en bâtons

Figure 3.2 – Répartition du personnel salarié de l'entreprise X par type d'activité au 1er janvier 2007

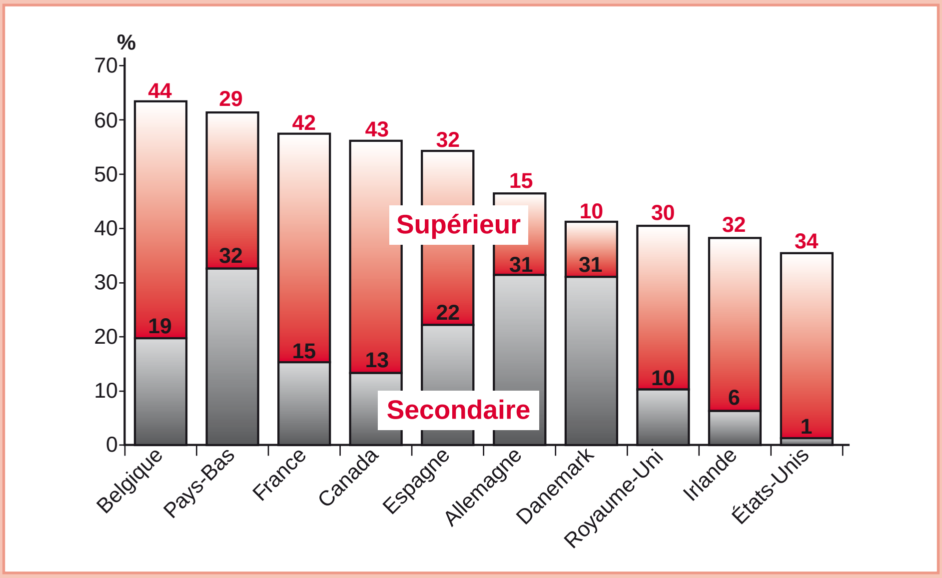


Ce diagramme-là, qui a exactement la même fonction que le précédent, est appelé « **diagramme en bâtons** ». On a simplement réduit les bases à un point[1] et les hauteurs sont les mêmes.

Le diagramme en bâtons peut être jugé moins esthétique, mais c'est surtout un problème de fonctionnalité qui nous fait souvent préférer l'autre : en effet, avec les tuyaux d'orgues, on peut insérer **deux dimensions** et, ce faisant, *on double la richesse de l'information* :

Diagramme en tuyaux d'orgues à 2 dimensions

Figure 3.3 – Taux de scolarisation à 20 ans, selon les pays et les niveaux d'enseignement, en 1996



Source : « Les grands chiffres de l'Éducation nationale », Ministère de l'Éducation nationale, de la recherche et de la technologie, septembre 1999.

Ce diagramme en tuyaux d'orgues est bien à deux dimensions : la dimension quantitative « taux de scolarisation selon les pays » et la dimension qualitative « taux de scolarisation selon l'enseignement », *supérieur* ou *secondaire*.

Un autre moyen d'insérer deux dimensions est **de juxtaposer deux modalités**, soit de façon verticale, soit de façon horizontale, comme dans le diagramme ci-après.

Notons que ce diagramme s'appelle *diagramme en tuyaux d'orgues couchés*, ou plus généralement « **diagramme en barres** ». La logique de construction est la même que précédemment ; on a simplement « renversé » les axes du graphique.

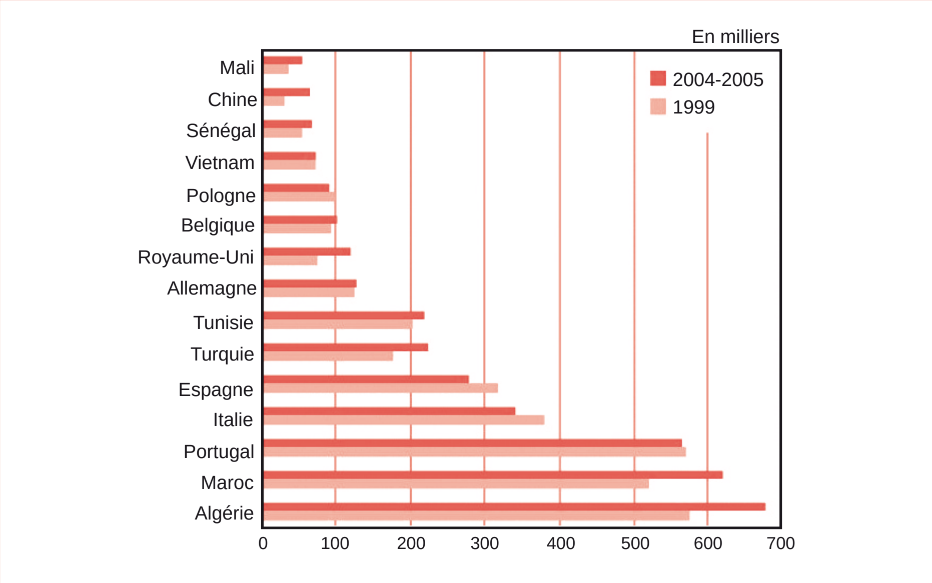
Pour bien comprendre la diversité des possibilités de choix, et pour se mettre en situation pratique, le lecteur peut se reporter aux deux exercices intitulés « **Choisir le message visuel adapté** », page 100 et page 105.

Il est fréquent de trouver de tels diagrammes (*barres* ou *tuyaux d'orgues* à deux dimensions). À plus de deux ou trois dimensions, le diagramme est difficilement lisible.

1. En anglais, on fait la différence entre *Rod Graph*, les bâtons, et *Column Chart*, les tuyaux d'orgues.

Diagramme en barres

Figure 3.4 – Les immigrés selon le pays de naissance, en 1999 et en 2004-2005



Source : INSEE, enquêtes de recensement de 2004 et 2005. Champ : France métropolitaine. Référence : INSEE Première, n° 1098, août 2006

Au-delà de l'effet visuel, il faut noter que les caractères de ces tableaux sont soit **qualitatifs** (comme ici, la nationalité ou la date), soit **quantitatifs discrets** (par exemple le nombre de pièces d'un appartement) ou **quantitatifs regroupés par blocs** (par exemple des normes de surface d'appartement).

En aucun cas, ces diagrammes ne s'adaptent au cas continu. Nous verrons que le cas des variables *quantitatives continues* est traité par un autre type de graphe : *l'histogramme*, dont l'effet visuel n'est pas la hauteur mais **la surface**. C'est d'ailleurs une erreur que commettent bien des gens et même parfois certains logiciels : confondre les deux types de diagrammes et tomber dans le piège de la mauvaise interprétation en cas de classes inégales (voir plus loin, page 79 et page 80).

B – Les diagrammes à axes multiples : toile d'araignée et radar

Ce sont encore des diagrammes de hauteur, mais leur ambition est de schématiser **des comparaisons** visuelles sur deux – mais souvent beaucoup plus – modalités du caractère étudié. Il y a **autant d'axes que de modalités.**

Les axes, **gradués depuis un même centre**, donnent une image d'étoile ou de *toile d'araignée*, selon leur nombre[1]. Certains logiciels utilisent le terme de « **radar** » ; en anglais, on parle de *spider chart* ou de *radar chart*.

---

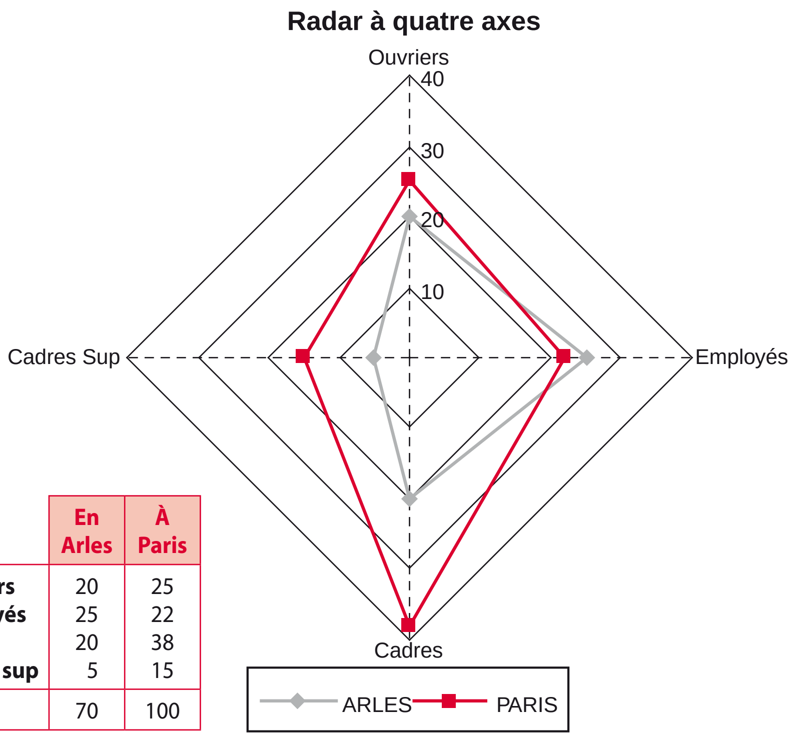
1. Ces graphiques ressemblent aux ***graphiques en coordonnées polaires***, dans lesquels la lecture est identique, mais pas la construction, celle-ci étant angulaire (ou trigonométrique) dans ces derniers cas.

Les comparaisons peuvent être *temporelles* (comparer des chiffres d'affaires à des dates différentes), *spatiales* (comparer des équipements sur des régions, ou des pays différents), *physiques* (comparer des performances pour des imprimantes d'ordinateurs) ou autres. On peut aussi, dans le même diagramme, comparer les modalités à la moyenne de ces modalités sur une série d'informations plus large.

Les segments de droite qui relient les valeurs des modalités forment les *pourtours* d'**un polygone**, bien apparent à l'œil, comme on le voit dans la figure ci-après. C'est la **superposition des figures ainsi formées** qui permet les comparaisons. L'effet visuel est renforcé par le **choix de couleurs** dans ces segments de droite, ou même dans les surfaces des polygones[1]. Dans l'exemple ci-après, nous prenons *quatre modalités d'activités professionnelles*. Nous aboutissons donc à quatre axes et nous comparons les valeurs des modalités pour deux agglomérations différentes (comparaison spatiale).

Diagramme en « radar »

Figure 3.5 – Répartition du personnel salarié de l'entreprise X par type d'activité au 1er janvier 2007 et par lieu d'établissement



	En Arles	À Paris
Ouvriers	20	25
Employés	25	22
Cadres	20	38
Cadres sup	5	15
TOTAL	70	100

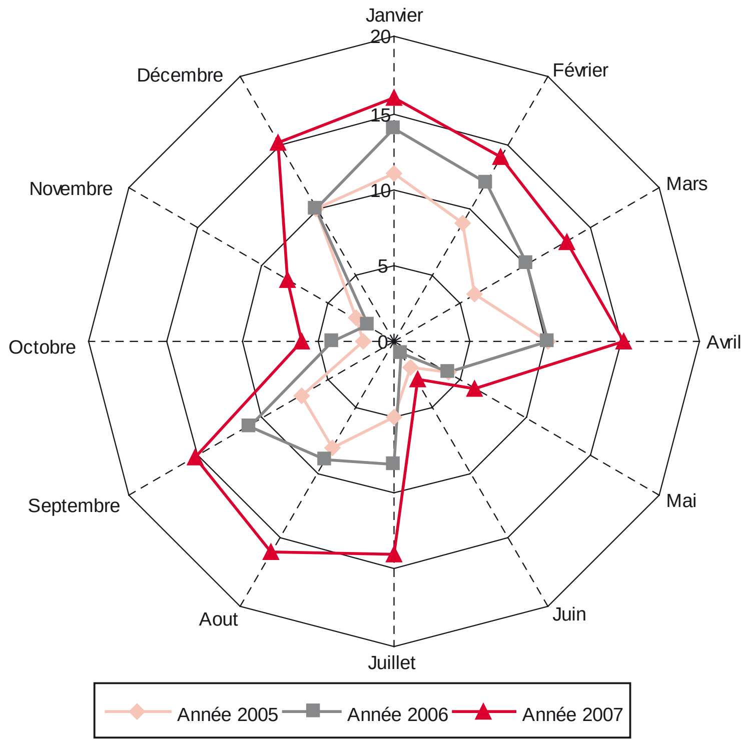
Dans le même ordre d'idée, le graphique suivant (figure 3.6) montre la **décomposition mensuelle sur trois ans**, du chiffre d'affaires d'un commerce, en milliers d'euros.

1. Voir les exercices 5 et 6, page 117 et suivantes, pour des résolutions manuelles et avec Excel.

Diagramme en toile d'araignée

Figure 3.6 – Chiffre d'affaires du commerce Y (en milliers d'euros)

Mois \ Année	2005	2006	2007
Janvier	11	14	16
Février	9	12	14
Mars	6	10	13
Avril	10	10	15
Mai	4	4	6
Juin	2	1	3
Juillet	5	8	14
Août	8	9	16
Septembre	7	11	15
Octobre	2	4	6
Novembre	3	2	8
Décembre	10	10	15



Douze axes et *trois polygones* rendent compte de l'évolution : il est possible de comparer les différentes performances, mois par mois, sur les trois années[1].

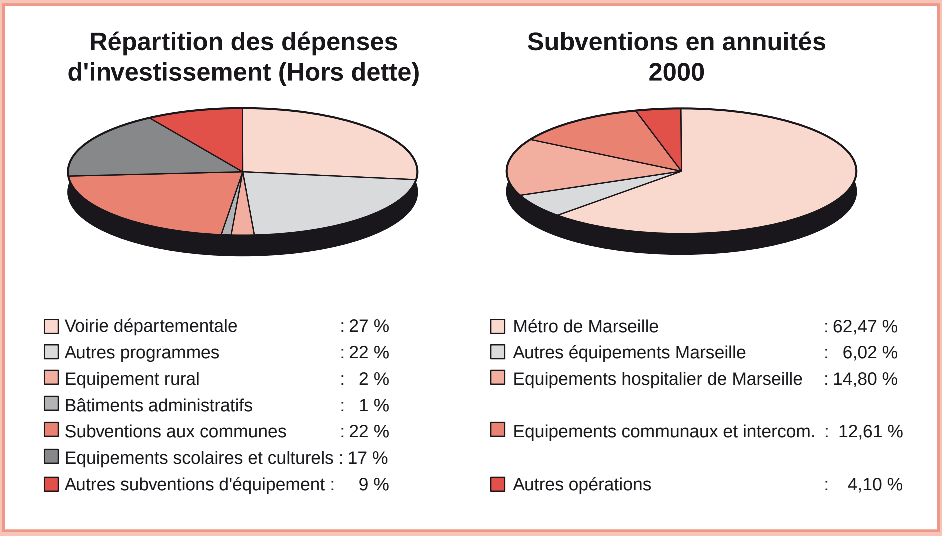
1. C'est-à-dire de comparer 3 « mois de janvier » entre eux, 3 « mois de février » entre eux, etc. On retrouvera cette notion de *correction de variations saisonnières* au chapitre 6, avec la notion de CVS, page 252.

C – Les diagrammes angulaires

Voilà deux exemples de diagrammes angulaires à une dimension : le premier nous servira d'*observation*, le second de *construction*.

Diagrammes à secteurs circulaires en 3D

Figure 3.7 – Source : *Rapport financier 2000*, Conseil général des Bouches-du-Rhône.



Si vous demandez à un Français non averti : « Comment s'appelle ce genre de figure ? », il y a une probabilité non nulle qu'il vous réponde : « Camembert ! ». Évidemment, vu la quantité de variétés de fromages au lait de vache que nous connaissons et l'image traditionnelle de cette appellation, cette réponse spontanée n'a rien d'étonnant ; mais posez la même question à un Mexicain ou un Danois, vous n'obtiendrez pas forcément la même réponse. Beaucoup d'Anglo-Saxons, adoptant également *une référence gustative*, l'appellent *Pie Chart*, ou encore *Cake Chart*, c'est-à-dire *graphique en tarte* ou *en gâteau*.

Oublions les goûts, mais gardons les couleurs : quel est, finalement, le nom de ce diagramme ? C'est un « **diagramme à secteurs circulaires** ».

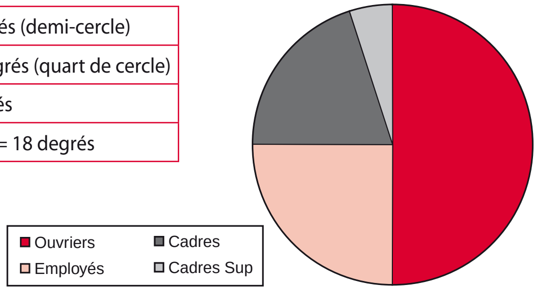
Dans l'exemple ci-dessus, les données chiffrées sont repérées par leurs couleurs dans la légende. On peut aussi, quand il n'y a pas d'effectifs trop faibles, les inscrire dans les secteurs. On peut aussi **agréger les tout petits secteurs** pour équilibrer le dessin.

La construction de ces diagrammes est simple : on multiplie chaque fréquence par 360 (un cercle, c'est 360 degrés). Ainsi, sur l'exemple de l'entreprise X précédent, il vient :

Construction d'un diagramme à secteurs circulaires

Figure 3.8 – Répartition du personnel salarié de l'entreprise X par type d'activité au 1er janvier 2007

50 % d'ouvriers, soit 0,5 × 360 = 180 degrés (demi-cercle)
25 % d'employés, soit 0,25 × 360 = 90 degrés (quart de cercle)
20 % de cadres, soit 0,20 × 360 = 72 degrés
5 % de cadres supérieurs, soit 0,05 × 360 = 18 degrés



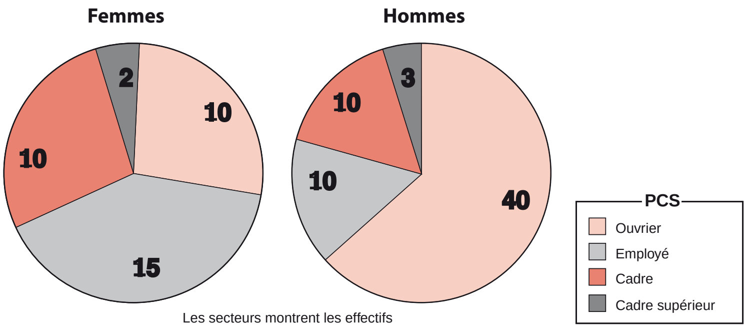
Les trois diagrammes à secteurs circulaires que nous venons de voir restent à une dimension. Comment passer à deux dimensions ?

1) La première méthode consiste à aligner plusieurs cercles d'égales surfaces, les uns à côté des autres, comme le montre le diagramme ci-dessous :

Tableau à deux dimensions et double diagramme à secteurs circulaires

Figure 3.9 – Répartition du personnel salarié de l'entreprise X par type de PCS et par sexe au 1er janvier 2007

Effectif		Sexe		Total
		femme	homme	
PCS	ouvrier	10	40	50
	employé	15	10	25
	cadre	10	10	20
	cadre supérieur	2	3	5
Total		37	63	100

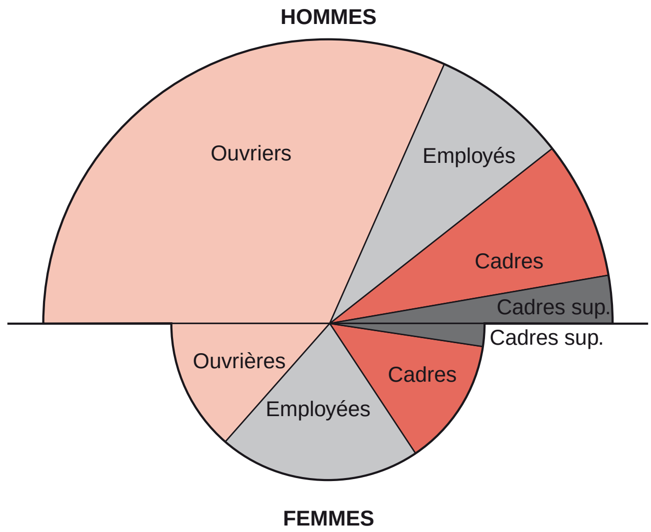


2) La deuxième méthode est de fabriquer des **diagrammes à secteurs semi-circulaires** : toujours dans le cas de l'entreprise X, on peut ajouter la dimension « sexe » en réservant 180 degrés pour les hommes et 180 degrés pour les femmes. Au lieu de multiplier les fréquences<par 360, on multipliera par 180.

De plus, la surface de chaque demi-cercle étant proportionnelle à l'effectif, ce type de diagramme permet de **bien visualiser les différences d'effectifs** : on voit bien ici qu'il y a plus d'hommes que de femmes. Il vient :

Diagramme à secteurs semi-circulaires

Figure 3.10 – Répartition du personnel salarié de l'entreprise X par type d'activité et par sexe au 1er janvier 2007 (sur mêmes données que figure 3.9)



Le *diagramme semi-circulaire* est une technique très intéressante pour représenter *deux dimensions* sur un diagramme circulaire.

Notons que ce type de diagramme offre directement la possibilité de *rajouter une information visuelle supplémentaire* (celle de différencier les effectifs) et qu'il est **regrettable** que la plupart des logiciels informatiques ne le proposent pas.

L'application pratique de l'exercice 4, page 111, constitue un intéressant exemple de choix et de construction par étapes de tels diagrammes, dans le domaine qualitatif.

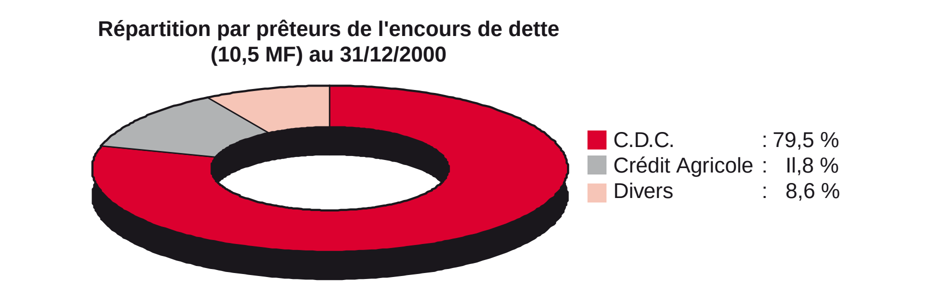
Il existe un autre type de diagramme angulaire, le « **diagramme en anneaux**[1] », qui peut être à une ou à plusieurs dimensions, mais qui est surtout adapté à un petit nombre de modalités : l'anneau, qui représente 100 % des observations, est divisé en secteurs. Chaque secteur possède la proportion de chaque modalité.

1. *Ring chart* ou *Doughnut chart* en anglais.

Voici un diagramme à un seul anneau (trois modalités du caractère « *prêteurs de l'encours de la dette* » dont l'anneau du prêteur majoritaire apparaît nettement) :

Diagramme en anneaux à une dimension

Figure 3.11 – Source : *Rapport financier 2000*, Conseil général des Bouches-du-Rhône.



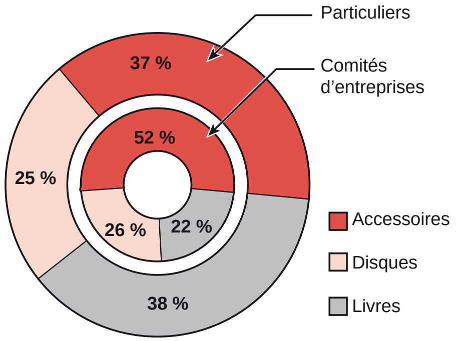
Ce type de diagramme s'adapte au cas de **deux dimensions**, mais n'a d'intérêt que s'il fait ressortir d'assez fortes variations entre les situations de chaque anneau.

Observons un **diagramme à deux anneaux,** qui fait apparaître deux dimensions : il s'agit de chiffres fictifs concernant les ventes *de trois types* d'articles de bureau, en pourcentage du total des ventes, pour un magasin dont la clientèle est essentiellement composée de *deux groupes* d'acheteurs, les *particuliers* et les *comités d'entreprise*.

Diagramme en anneaux à deux dimensions

Figure 3.12 – Ventes d'articles de bureau

Vente / Articles	Pour comités d'entreprises	Pour particuliers
Accessoires	52 %	37 %
Livres	22 %	38 %
Disques	26 %	25 %



Les anneaux sont ici *séparés*. Ils peuvent aussi être « accolés » mais, dans ce dernier cas, les couleurs qui se retrouvent juxtaposées doivent être bien contrastées pour que l'on puisse apprécier les différences ; cela n'est pas toujours évident, selon l'endroit où « tombent » les juxtapositions.

Par ailleurs, à plus de trois anneaux, la figure devient pratiquement illisible.

D – Les diagrammes de surface

On arrive au cas *continu*.

Rappelons qu'un caractère **quantitatif** est dit « **continu** » et prend dès lors le nom de « **variable** » (voir chapitre 2) s'il existe une infinité de valeurs intermédiaires, et donc que l'on peut pousser **sa mesure à un certain nombre de décimales**. C'est le cas de la variable « âge » par exemple, lorsqu'on la décline par **classes juxtaposées** comme dans les tableaux des pages 49 et suivantes du chapitre 2.

C'est aussi le cas des variables de mesure physique (poids, taille, âge), des variables de durée (temps, périodes), des variables d'espace (longueurs, hauteurs, aires), de monnaie (chiffres d'affaires, salaires), par exemple.

Les variables sont regroupées en « classes semi-ouvertes à droite[1] ». Les classes sont *contiguës* : par exemple, pour les âges, la classe [0, 10 ans[est attenante à la classe [10, 15 ans[, qui est elle-même attenante à la classe [15, 40 ans[, si l'on a affaire à des classes *d'amplitudes inégales*.

Le diagramme adapté à ces situations s'appelle « **histogramme** ». Il *ressemble* aux « tuyaux d'orgues » (rectangles), mais le message visuel est complètement différent et il ne faut **surtout pas confondre** les deux supports : l'histogramme donne une image pour laquelle **l'œil humain repère la surface des rectangles**, et non leurs *hauteurs*, comme pour les bâtons ou les tuyaux d'orgues.

Les **classes se jouxtant** obligatoirement, la dernière observation d'une classe est la même (à 2 ou 3 chiffres après la virgule), que la première de la classe suivante. La figure géométrique finale, formée par toutes les classes, se lit comme **LA SURFACE TOTALE** de l'ensemble des rectangles. Elle représente donc l'ensemble des cas possibles que peut prendre le phénomène, c'est-à-dire, en fréquences, **100 %**.

Construisons un histogramme, à partir des données du tableau suivant :

Tableau 3.1 : Âge des 110 élèves inscrits dans les filières professionnelles des « Métiers de la Communication » de l'école « Com.C.Chic » durant l'année 2006 (Données fictives)

Âges (x_i)	Nombre (n_i)
[15, 20[	10
[20, 25[	30
[25, 30[	40
[30, 35[	20
[35, 40[	7
[40, 45[	3
	$\Sigma =$ 110

- Tout d'abord, nous remarquons que **les classes sont** d'amplitudes **égales**.

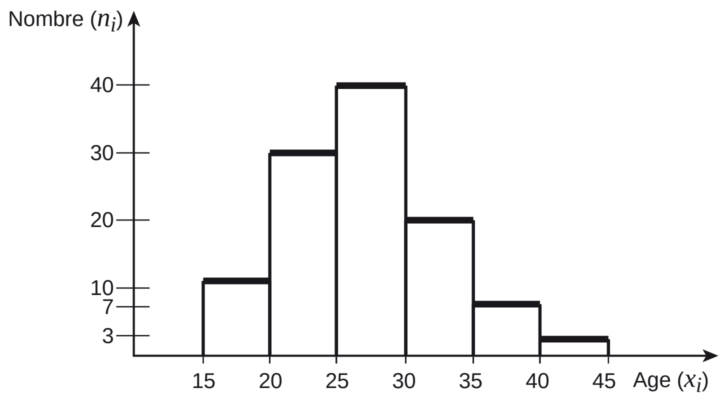
1. Comme on l'a vu dans le chapitre 2, la borne de droite est exclue, de telle manière que l'individu âgé par exemple de 20 ans, ne fait pas partie de la classe [10, 20[, mais de la classe [20, 30[.

- Nous choisissons les échelles des deux axes de coordonnées et portons les valeurs correspondantes. On peut voir que l'étendue de la série est assez large (de [15, 20[à [40, 45[), car il s'agit d'une école « professionnelle », comprenant des filières de formation continue.
- Nous construisons les rectangles de manière à ce que leurs hauteurs soient celles de la colonne « effectifs ».

Il vient :

Histogramme

Figure 3.13 – Âge des 110 élèves inscrits dans les filières professionnelles des « Métiers de la Communication » de l'école « Com.C.Chic » durant l'année 2006 (Données fictives)



L'histogramme **est la surface totale**, visuellement repérée, formée à partir des rectangles de même largeur puisque les classes sont d'amplitudes égales.

On en déduit que la surface des trois premiers rectangles correspond à l'idée de *cumuler* les effectifs des trois premières classes (10 + 30 + 40), et l'on pourra dire : « 80 élèves sur 110, soit 73 %, ont **moins de** 30 ans ». On appelle ce résultat une « **fréquence cumulée ascendante**[1] ».

On comprend que la notion « **plus de** » correspondra à une « **fréquence cumulée descendante** » (il n'y a que 10 élèves qui ont plus de 35 ans, par exemple).

Le grand problème en matière d'histogramme réside dans la « **correction des classes inégales** ».

On va saisir toute la puissance de lecture en termes de surfaces en se prêtant à **l'exercice suivant** : imaginons que nous voulions **regrouper les deux dernières classes** ; c'est assez légitime puisqu'elles ont des effectifs très faibles.

On aura dans le tableau *un effectif de 10 élèves (7 + 3)* pour la dernière classe ainsi regroupée.

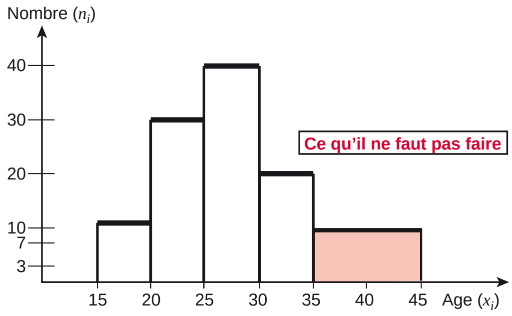
1. On retrouve cette notion de *fréquence cumulée* dans le chapitre 4 pour le calcul de la médiane.

1 – Voici d'abord ce qu'il ne faut pas faire, tout simplement parce que ce serait *faux* !

Histogramme mal corrigé

Figure 3.14 – Âge des 110 élèves inscrits dans les filières professionnelles des « Métiers de la Communication » de l'école « Com.C.Chic » durant l'année 2006 (Données fictives)

Âges (x_i)	Nombre (n_i)
[15, 20[	10
[20, 25[	30
[25, 30[	40
[30, 35[	20
[35, 45[	**10**
	Σ = 110



Pourquoi ce diagramme est-il *faux* ? Parce que l'on n'a pas respecté les surfaces : on a **reporté les hauteurs** (7 + 3 = 10), et, évidemment, à bases égales, cela a **augmenté la surface des deux dernières classes** (surfaces colorées). L'histogramme n'a plus le même équilibre (plus la même surface) ; l'œil ne voit plus la même chose : le message visuel est faussé et, par là même, la communication aussi.

C'est un piège dans lequel tombent bien des utilisateurs, d'autant plus facilement que certains logiciels informatiques font également la même erreur (bien vérifier, dans la pratique…).

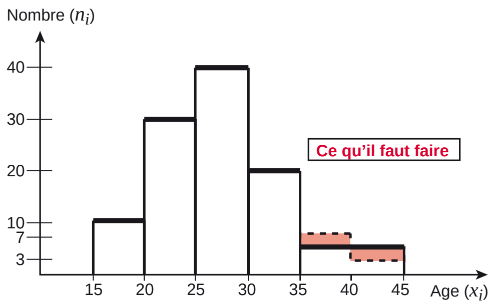
2 – Voici maintenant ce qu'il faut faire :

Puisqu'on a doublé les classes (plus exactement multiplié par 2 « l'amplitude » unitaire des classes), il faut diviser par 2 la hauteur correspondante (10/2 = 5). **La hauteur du dernier rectangle doit donc être de 5**, comme le montre la figure 3.15.

Histogramme bien corrigé

Figure 3.15 – Âge des 110 élèves inscrits dans les filières professionnelles des « Métiers de la Communication » de l'école « Com.C.Chic » durant l'année 2006 (Données fictives)

Âges (x_i)	Nombre (n_i)	Hauteur
[15,20[	10	10
[20,25[	30	30
[25,30[	40	40
[30,35[	20	20
[35,45[	10	**5**
	Σ = 110	



Le trait en pointillés reprend la forme de l'histogramme originel et montre que l'équilibre est maintenant conservé (on dit qu'il y a *compensation des aires*).

Par ailleurs, au niveau purement *esthétique,* les logiciels informatiques permettent de donner libre cours à notre imagination pour présenter des diagrammes très « **design** » : on trouve une grande variété de dessins très originaux, depuis les formes à plat, ou *obliques*, ou à secteurs éclatés, jusqu'aux *effets 3D.*

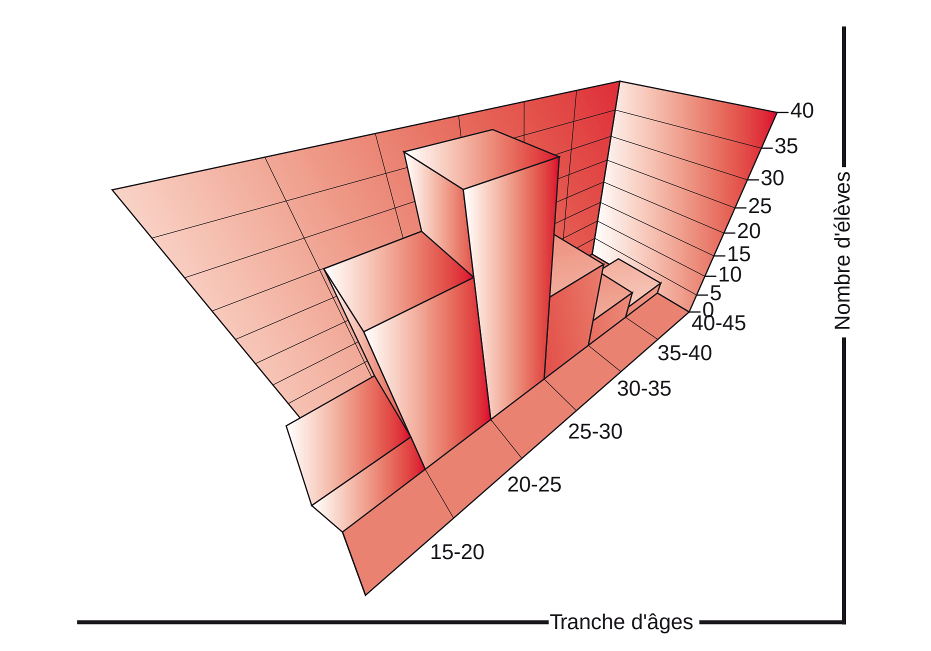
Les résultats sont plus ou moins réussis au *niveau de la communication*. N'oublions jamais qu'un diagramme doit faire passer un message simple, *compris en dix secondes une fois le titre lu.*

L'histogramme suivant, sur les données précédentes (avec classes égales), donne **un aperçu** de ce que l'on peut obtenir en 3D, avec un logiciel de base. Au-delà de l'aspect (relativement) *esthétique,* apprécié ou non, de ce genre de représentation, il faut être **extrêmement prudent**, surtout quand les classes sont **d'amplitudes inégales**, car les logiciels ne sont pas tous suffisamment performants pour effectuer les nécessaires corrections que l'on a vues précédemment.

La plupart des logiciels connus ont en effet la fâcheuse tendance à oublier que les histogrammes représentent, pour l'œil humain qui les regarde, une **surface** et non une **succession de hauteurs.** Cette surface doit toujours rester équilibrée quand on agrandit ou raccourcit les classes (c'est-à-dire les bases des rectangles).

Histogramme en 3D

Figure 3.16 – Âge des 110 élèves inscrits dans les filières professionnelles des « Métiers de la Communication » de l'école « Com.C.Chic » durant l'année 2006 (Données fictives)

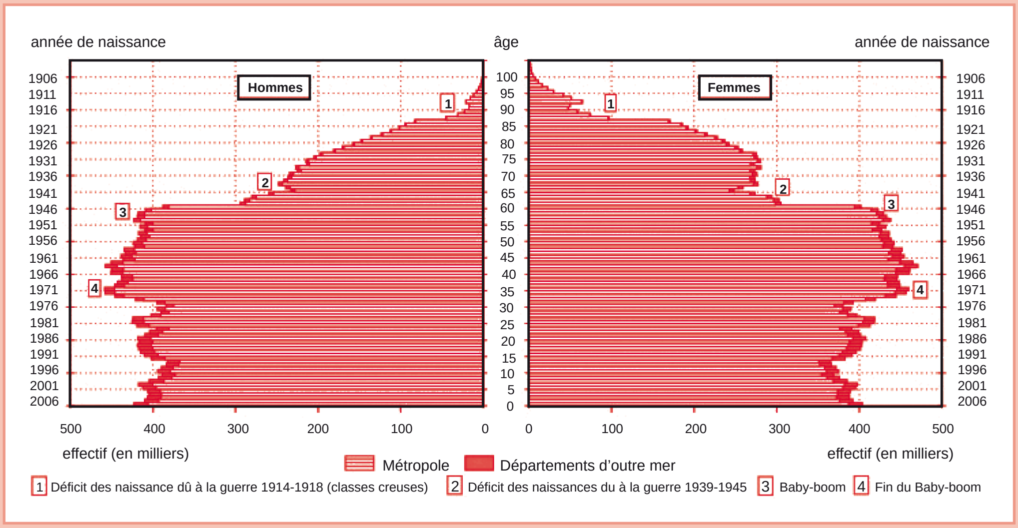


Notons enfin (voir page suivante) que les « **pyramides des âges** », que l'on voit à peu près partout disposées selon une figure traditionnelle, ne sont en fait que des *doubles histogrammes renversés* : un histogramme pour les modalités « hommes » (renversé sur la gauche) et un autre, en face, pour les modalités « femmes » (renversé sur la droite). Rappelons qu'on appelle souvent ces rectangles renversés des « **barres** », tout comme dans le cas des tuyaux d'orgues, probablement pour s'accorder aux *Bar Charts* des Anglo-Saxons.

Dans l'exemple proposé ci-après, les « **barres** » sont très fines, vu l'étendue assez large (heureusement pour notre espérance de vie) de la série considérée.

Double histogramme renversé

Figure 3.17 – Bilan démographique 2006 : un excédent naturel record



Source : *INSEE Première*, n° 1 118, janvier 2007.

E – Les diagrammes des séries chronologiques[1]

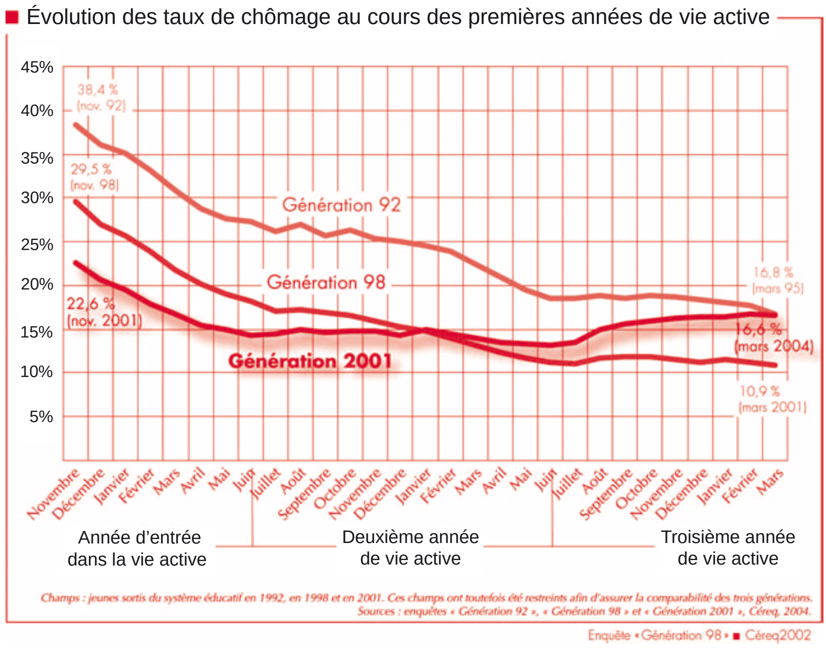
Les représentations de ces séries sont des sortes de diagrammes de hauteur un peu spéciales car, comme leur nom l'indique, les séries chronologiques (*chronos, le temps, en grec*) ne représentent que des **évolutions de phénomènes repérés dans le temps**. Ce temps, ici, est généralement divisé en années, trimestres, mois, ou jours.

Voici (figure 3.18) un exemple de série chronologique où le phénomène est repéré en mois :

1. Voir également chapitre 6, page 260 et suivantes.

Diagramme de séries chronologiques

Figure 3.18 – Source : CEREQ, « Enquête génération », 2004.



Profitons de l'observation de ce diagramme pour *construire* une série chronologique :

L'axe des abscisses (horizontal, en bas) repère les échelles de temps. L'axe des ordonnées (vertical) est réservé au report des valeurs du phénomène étudié.

C'est donc **une hauteur que l'on reporte** mais, contrairement aux tuyaux d'orgues, ou aux bâtons, on ne trace pas de lignes verticales ; on indique **uniquement le point correspondant**, et l'on relie par **des segments de droite** (*jamais de courbes*) les points obtenus[1].

Dans la plupart des cas, les valeurs du phénomène (chiffre d'affaires, production industrielle, nombre d'étudiants, ou autres) ne commencent pas à zéro, mais bien plutôt à un chiffre assez important. On débute donc les graduations de l'axe vertical par un chiffre « rond » proche de la valeur observée la plus faible. Cette opération s'appelle la **coupure de l'axe**. Cette coupure doit en principe **être bien précisée** sur le graphique pour éviter les exagérations d'interprétation parfois dues aux effets d'optique. Les unités de mesure (milliers, millions, etc.) doivent également être précisées de manière claire.

1. Cela est la manière la plus traditionnelle de représenter des séries chronologiques. Rappelons qu'en figure 3.6, nous avions utilisé un graphique de type « radar » pour permettre des comparaisons temporelles. Cependant, ces radars ne sont « lisibles » que quand les valeurs mensuelles ou annuelles sont assez bien séparées visuellement, alors que les graphiques traditionnels ne souffrent pas de cet inconvénient.

Soit les données trimestrielles suivantes concernant la *fréquentation touristique*, sur quatre années, pour une région de montagne. Les unités sont des *nuitées d'hôtels*.

Tableau 3.2 : Fréquentation touristique en nuitées

Trimestres	Année 2003	Année 2004	Année 2005	Année 2006
Premier	1 090	1 200	1 250	1 320
Deuxième	1 300	1 450	1 390	1 550
Troisième	920	980	1 100	1 110
Quatrième	1 000	1 070	1 160	1 220

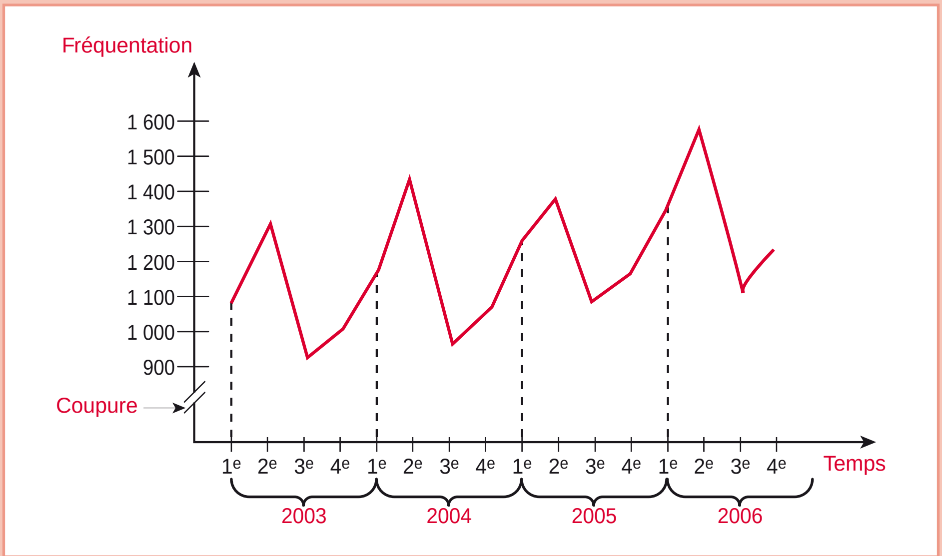
Construisons le diagramme de cette série chronologique. Le temps est partagé en trimestres sur quatre années : nous aurons donc 16 points à reporter et à joindre par des segments de droite (figure 3.19).

Les valeurs s'échelonnent de 920 à 1 550 selon les périodes. Pour ne pas donner une impression visuelle « **d'aplatissement** », il faut étaler le plus possible l'intervalle de variation des valeurs du phénomène. C'est le rôle de la coupure de l'axe et du choix de l'échelle des ordonnées : ici, nous repérerons la coupure aux alentours de la valeur 850, pour équilibrer le graphique. Cependant, notons que ce repérage est facile pour un graphique dont on connaît les valeurs comme ici, mais peut réserver des surprises pour un graphique où l'on porte les valeurs au jour le jour, ou mois par mois, en même temps que l'on découvre les données.

Le diagramme est le suivant :

Construction d'un diagramme de série chronologique

Figure 3.19 – Fréquentation touristique en nuitées



Le graphique fait ressortir de façon visuelle, mieux que le *radar*, ici, les **variations répétitives**[1] d'année en année dues au rythme de fréquentation hivernale.

La réponse à la question 1 de l'exercice 1, page 261, (chapitre 6), permet de compléter les considérations précédentes sur un cas pratique, chiffré.

§ 2 – LES REPRÉSENTATIONS GRAPHIQUES À TRADUCTION INDIRECTE

On utilise, pour les constructions de ces diagrammes, des **outils de transcription** autres que les instruments que permet la géométrie de base.

- Pour les graphiques *semi-logarithmiques*, on utilise une échelle spéciale ou un *papier* « semi-log ».
- Pour les *diagrammes figuratifs,* on utilise des figurines (pictogrammes).
- Pour les *cartogrammes*, le support est une carte géographique.
- Pour les diagrammes factoriels, on utilise des méthodes de *maximisation de distances* sur des repères factoriels dont on ne connaît pas à l'avance le nom des axes ; ceux-ci étant révélés par l'analyse multidimensionnelle *a posteriori*.

A – Les graphiques semi-logarithmiques

Rassurons-nous, malgré le nom qu'ils portent, l'utilisation de ces graphiques n'impose en aucune manière de savoir calculer des logarithmes, ou d'être imbattable sur l'équation de la fonction « Log ».

Ici, il s'agit **d'un problème d'échelle** : au lieu d'utiliser les échelles arithmétiques, que l'on connaît depuis l'école maternelle, et où la distance entre la graduation 1 et la graduation 2 est la même que celle qui se trouve entre 2 et 3 et ainsi de suite, on utilise un *autre système de graduation*.

À la place de la proportionnalité précédente (**proportionnalité arithmétique**), on repère des graduations proportionnelles aux logarithmes des valeurs (**proportionnalité logarithmique**). Dans le premier cas, la distance entre les graduations 4 et 3, par exemple, donnera : **4 – 3 = 1 cm** sur une feuille de papier où l'échelle est arithmétique. Dans le deuxième cas (échelle logarithmique), cette distance sera : **log 4 – log 3 = 1 cm**. Comprenons bien que les *logarithmes* n'apparaîtront pas sur le graphique ; seules les **valeurs 3 et 4** apparaîtront.

Pourquoi se torturer ainsi ? Quel est l'avantage de remplacer des échelles arithmétiques que l'on connaît si bien ? La réponse tient en l'analyse de *deux inconvénients fondamentaux* des échelles habituelles :

1. Variations dites « saisonnières » ; voir chapitre 6.

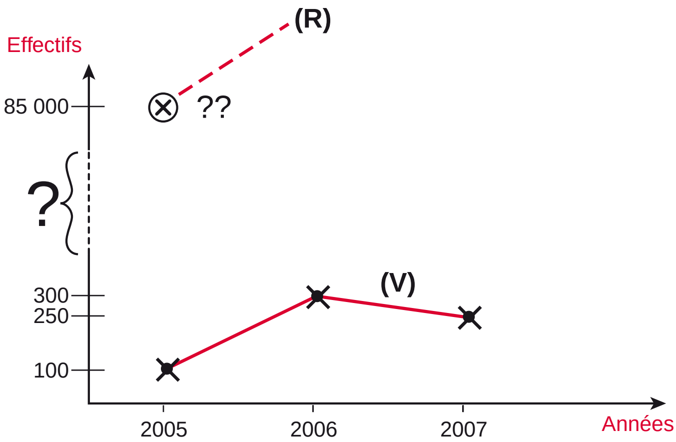
1 – Le cas de la « saturation des axes »

Prenons un exemple : voici les **effectifs des agents de l'État dans la ville** « **V** » **et dans la région** « **R** » sur trois années. Il est intéressant de se demander si ces effectifs augmentent ou diminuent de la même manière (proportionnellement ou pas) en même temps ; et quoi de plus « visuel » qu'un petit diagramme pour en juger ! Essayons de le construire sur une échelle habituelle (arithmétique) avec les données suivantes :

L'impossible représentation graphique des données

Figure 3.20 – Effectifs des agents de l'État dans la ville V et dans la région R (Graphique impossible)

Dates	Nombre d'agents en V	Nombre d'agents en R
2005	100	90 000
2006	300	920 000
2007	250	85 000



On aura beau tenter toutes les combinaisons possibles de graduation, **les deux représentations visuelles ne « rentreront » jamais dans le même graphique** : si l'on essaye de faire une échelle verticale (axe des ordonnées) longue de plusieurs mètres… le résultat ne sera pas du tout pratique et n'aura pas un message visuellement efficace ! Si l'on change d'échelle au milieu de l'axe, ce qui est interdit par les règles mathématiques, les proportions ne seront pas respectées.

Quelle est la solution ? La solution, c'est le diagramme semi-logarithmique comme on le verra plus loin.

2 – Le cas des « parallèles non proportionnelles »

Un autre problème peut surgir quand on **visualise des comparaisons graphiques sur support arithmétique** : que peut-on, par exemple, conclure sur l'évolution dans le temps des productions de deux entreprises A et B ? On voit sur le graphique des **lignes parallèles.**

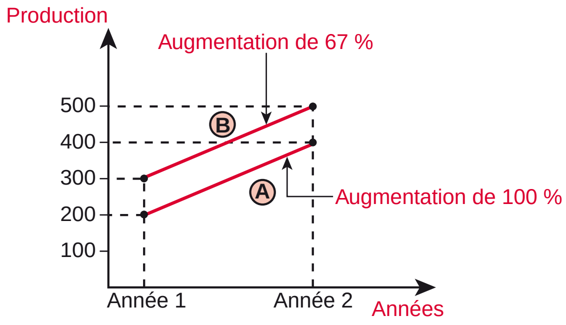
Le problème des variations relatives

Figure 3.21 – Production de deux entreprises A et B

L'entreprise A produit 200, puis 400
L'entreprise B produit 300, puis 500

Visuellement les deux évolutions apparaissent parallèles ; or, les productions ont augmenté dans des proportions bien différentes : 66,7 % et 100 %(doublement).

(500 : 300) – 1 = 0,6666, soit 66,7 %
(400 : 200) – 1 = 1, soit 100 %



Le message visuel est ici *trompeur* : on dirait (à première vue) que les deux productions progressent à la même vitesse. Or, il n'en est rien[1].

Les échelles arithmétiques traduisent mal les variations relatives. Quelle est la solution ? La solution, c'est encore le diagramme semi-logarithmique.

1. Un autre exemple de ce type, sur les taux de départ en vacances par catégories sociales, apparaît en correction de l'exercice 9, page 127 et suivantes. On y voit d'abord une évolution parallèle dans le schéma 3.14, qui devient non parallèle, *et donc incorrecte*, dans le schéma final, en échelle arithmétique.

3 – L'utilisation des supports et du papier semi-log

Pour bien comprendre

ÉCHELLES ARITHMÉTIQUES : *C'est le domaine de l'addition* (+)	**ÉCHELLES LOGARITHMIQUES :** *C'est le domaine de la multiplication* (×)
Elles décrivent des partitions : *Exemple :* l'enfant a mangé 20 % de la tarte aux pommes. S'il en reprend 5 %, il aura pris **deux parts** de tarte aux pommes 20 % + 5 % = 25 % *Il aura mangé un quart de la tarte.*	**Elles décrivent des évolutions :** *Exemple :* l'enfant a grandi de 10 % par an sur trois ans. Son **taux de croissance** annuel est de 0,10 et l'évolution aboutit à : $(1 + 0{,}10) \times (1 + 0{,}10) \times (1 + 0{,}10) = (1 + 0{,}10)^3$ **Sa taille de départ a été multipliée par 1,33**
ON ADDITIONNE OU ON SOUSTRAIT	**ON MULTIPLIE OU ON DIVISE**
Les échelles arithmétiques traduisent bien les écarts absolus : Si 1 cm vaut 10 €, 2 cm valent 20 €	**Les échelles logarithmiques traduisent bien les variations relatives :** 1/10 = 10/100 = 100/1000 = 10 %

On a vu que les échelles semi-log étaient construites selon le principe suivant : au lieu de reporter une même distance sur l'axe vertical comme c'est le cas pour les graphiques arithmétiques, on reporte une distance proportionnelle aux logarithmes :

10-------------------100---------------------1 000---------------------10 000

Comme ce sont des logarithmes décimaux, les distances physiques (sur le papier) entre les multiples de 10 sont égales.

Nota : il faut se représenter cet axe *en vertical* puisque c'est celui des ordonnées, c'est-à-dire celui de la quantification du phénomène ; l'axe des abscisses reste en échelle arithmétique.

Chaque intervalle entre deux graduations de 10 s'appelle un « **module** ». On voit donc déjà que l'on pourra placer sur la même feuille, par exemple « **à trois modules** » comme sur la figure 3.22, page 90, des quantités de l'ordre de 10, jusqu'à des quantités de l'ordre de 9 999 ; ce qui résout notre problème de **saturation des axes**, et l'exercice de la page 119 devient tout à fait faisable.

Par ailleurs, les propriétés mathématiques des logarithmes font que deux droites, ou deux courbes, qui se présentent de façon parallèle sur le papier **correspondent bien à des évolutions proportionnelles**. Ainsi le deuxième inconvénient décrit page 87 disparaît lui aussi.

Un dernier détail sur les distances : sur un papier arithmétique (papier millimétrique par exemple), si 3 – 4 donnent un centimètre, 5 – 4 aussi, et 145 – 144 aussi. Par contre, les différences entre les *log* ne donnent pas une proportionnalité constante :

- En effet : log 2 – log 1 = 0,30103, la distance pourra être de 3,01 millimètres sur l'échelle.
- De même : log 3 – log 2 = 0,176, la distance pourra être de 1,76 millimètre sur l'échelle.
- Et : log 4 – log 3 = 0,1249, la distance pourra être de 1,125 millimètre sur l'échelle.

On prend donc conscience que les graduations de chaque module sont **de plus en plus resserrées** de bas en haut sur l'axe vertical, ce qui ne pose pas de problème de lecture comme on le verra.

4 – Exercice : construction d'un diagramme semi-logarithmique[1]

Soit les données ci-après, représentant le nombre de visiteurs ayant fréquenté un complexe touristique, dont l'activité débute en 1998 et qui connaît un succès de fréquentation particulièrement remarquable.

Tableau 3.3 : Visiteurs ayant fréquenté le complexe touristique « Montmer »

Années	Nombre de visiteurs	Années	Nombre de visiteurs
1998	350	2003	18 500
1999	850	2004	22 000
2000	1 200	2005	95 500
2001	2 000	2006	74 000
2002	20 000	2007	60 000

Tout d'abord essayons de présenter l'évolution sur un graphique habituel (arithmétique)… **On voit vite que c'est impossible**, vu l'écart de valeur entre 95 500 et 350 : l'échelle habituelle ne le permet donc pas.

Passons donc au graphique semi-logarithmique.

Utilisons un **papier** « **semi-log** » du type de ceux qui sont vendus dans le commerce (voir en page 90) pour faire cet exercice. Ensuite on pourra, *si l'on veut*, utiliser des logiciels spécialisés pour construire de tels diagrammes ; mais le but ici est surtout de **comprendre leur logique d'élaboration,** de pouvoir les reconnaître et les lire dans les revues ou ouvrages scientifiques.

Le papier semi-log de la page 90 comprend trois modules dont les ordres de grandeurs ne sont pas repérés. On a donc *le choix de la première graduation*. Dans la mesure où « log 0 » n'existe pas, chaque module commence par un 1.

Si le 1 du premier module représente la valeur 1, le 1 du deuxième module représentera la valeur 10, le 1 du troisième la valeur 100.

Si l'on commence à 100, le 1 du deuxième module sera 1 000 et le 1 du troisième 10 000.

Pour l'exercice, on va considérer que :

- Le 1 du premier module représente la valeur 100.
- Le 1 du deuxième module représentera donc la valeur 1 000.
- Le 1 du troisième, la valeur 10 000, et la dernière valeur de ce module est 99 999.

Ainsi, on peut être sûr que toutes les observations seront présentes dans le graphe. L'axe des abscisses sera construit selon une logique arithmétique, par exemple en séparant **chaque année de référence par six graduations.**

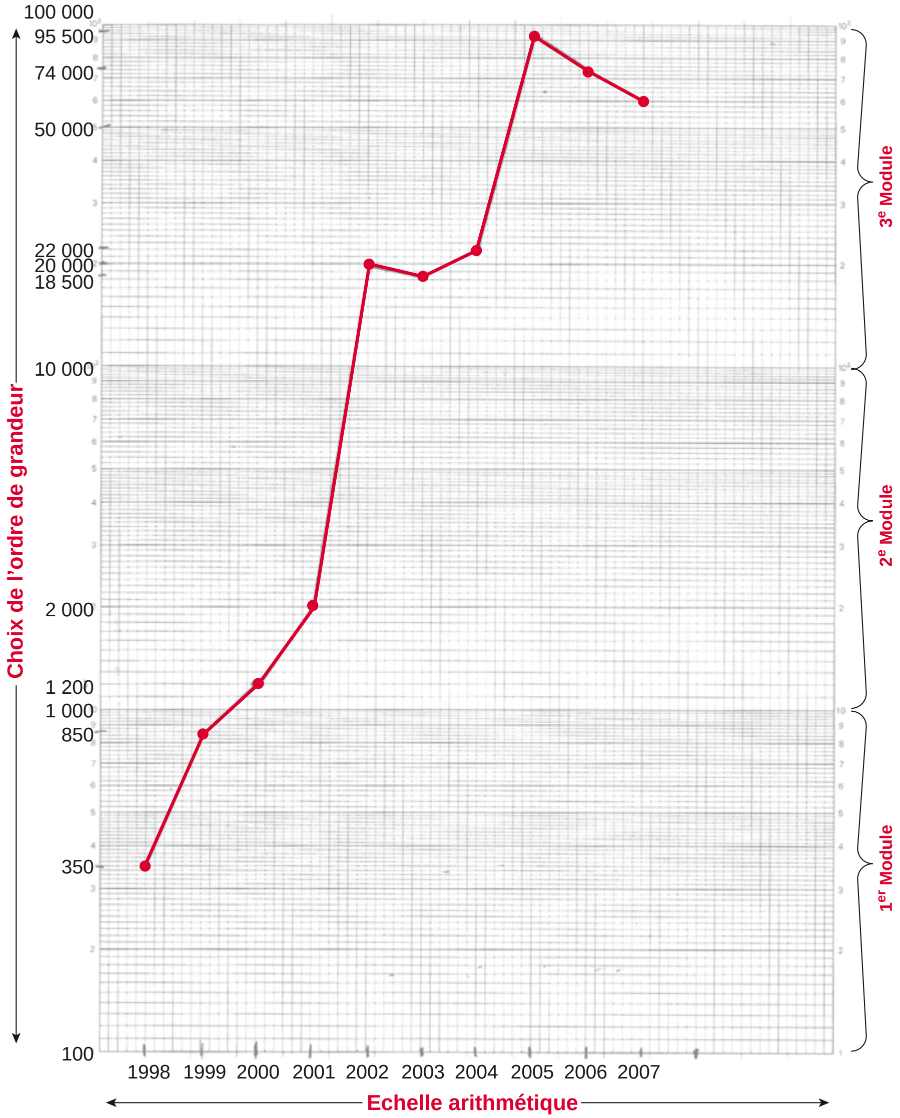
1. La résolution par Excel, sur un autre exemple chiffré, apparaît en fin de correction de l'exercice 7, page 123, « **Construction automatisée** ».

On reportera ensuite, **point par point**, les valeurs du phénomène étudié ; ici le nombre de visiteurs. Il faut prendre garde au deuxième module de ne pas confondre la valeur 1 200 avec la valeur 2 000.

On écrira les valeurs en face de chaque graduation choisie et l'on reliera par des segments de droite, et non des courbes, les points ainsi repérés, pour **obtenir l'allure finale du phénomène**, comme le montre le graphique ci-dessous.

Exercice de construction sur papier semi-logarithmique

Figure 3.22–Visiteurs ayant fréquenté le complexe touristique « Montmer »



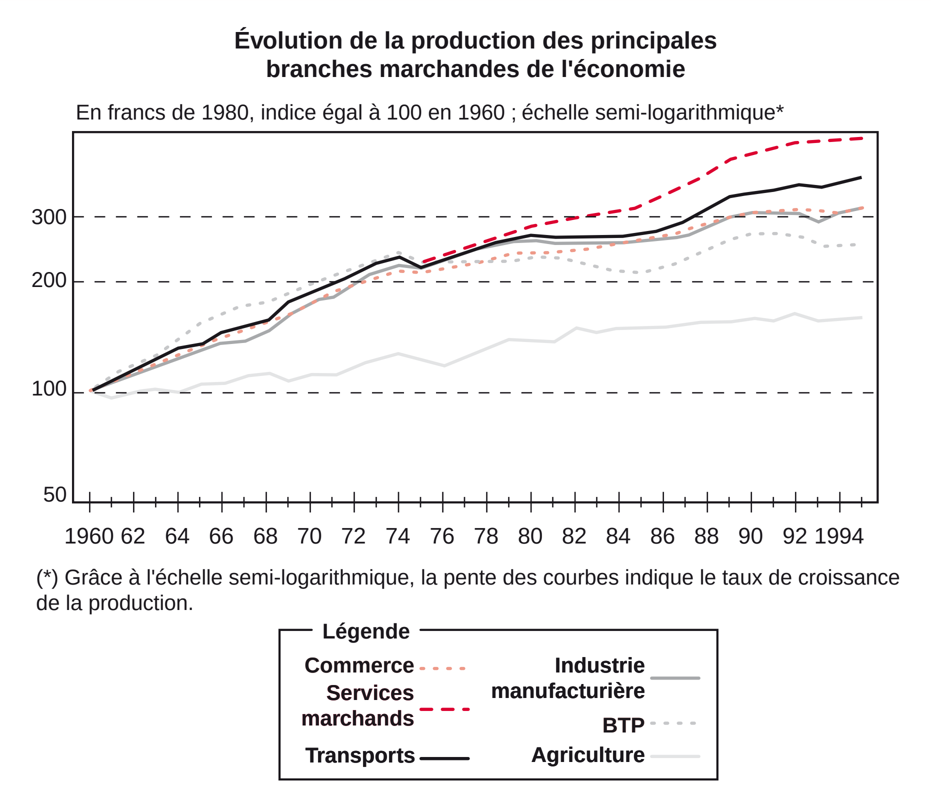
On comprend bien, d'une part, que l'utilisation de papiers semi-logarithmiques évite tout calcul de logarithme et, d'autre part, que **les valeurs se lisent directement sur le graphe comme sur une échelle habituelle**, sans que l'on soit particulièrement gêné par l'effet de resserrement des graduations. Les valeurs sont donc « à leur place », et c'est ça l'important.

Il faut savoir que le *passage* par les logarithmes entraîne un effet **géométrique** sur la courbe. Des longueurs égales sont représentées par des rapports égaux dans le cas du semi-logarithmique : **ce sont des variations relatives égales**. Dans le cas arithmétique, **ce sont des variations absolues égales**. La conséquence est que, par exemple, un phénomène dont **le taux de croissance annuel est constant** sera représenté par une **droite** en graphique semi-logarithmique, alors qu'il est représenté par une **courbe** en graphique arithmétique[1].

Ci-après, à titre *d'exemple de lecture*, un diagramme « semi-log » de l'INSEE.

Reconnaître un diagramme semi-logarithmique

Figure 3.23 – Référence : « L'essor des services publics depuis les années 1960 », in *INSEE Première*, n° 498, décembre 1996.



Source : INSEE, Comptes de la nation.

1. Pour de plus amples détails, et pour une « mise en situation » pratique de construction de ces diagrammes, se reporter à l'exercice 8, page 124, « Construire un diagramme semi-logarithmique ».

B – Les diagrammes figuratifs

Ici, nous sommes loin des applications géométriques simples, dans la mesure où ce sont **des illustrations** qui sont à la base de la transmission du message visuel. Ces types de diagrammes se décomposent en deux groupes, les **pictogrammes** et les **cartogrammes**.

1 – Les pictogrammes

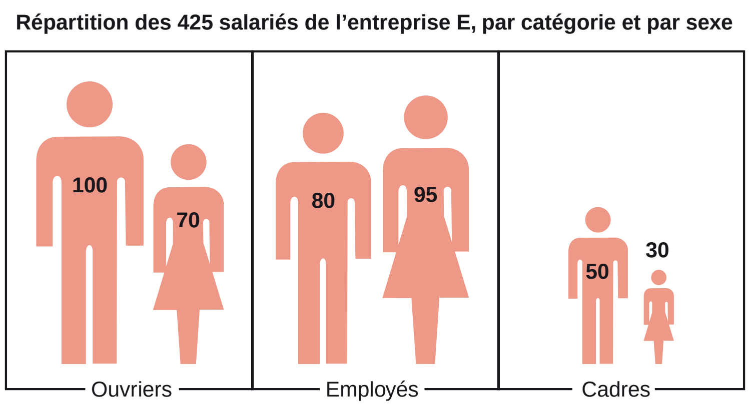
Ces diagrammes sont des **dessins colorés**, plus ou moins travaillés, à base d'images dont les plus ou moins grandes *corpulences*[1] rendent compte des quantités étudiées. Une méthode assez répandue, en ce domaine, est d'employer des **figurines** symbolisant des hommes et des femmes pour décrire des caractères tels que l'emploi, l'âge, le salaire, ou autre.

On obtient, ainsi, au moins deux dimensions : celle du caractère sexe, s'ajoutant à celle du caractère étudié.

La figure suivante simule une situation où les femmes sont moins employées que les hommes dans deux catégories socioprofessionnelles sur trois.

Diagramme de surface et corpulence

Figure 3.24



La rigueur statistique impose que, comme ici, **la surface, et non la hauteur des figurines** soit proportionnelle à l'effectif… Il est désolant de voir que, dans bien des études, cette norme n'est pas respectée.

Les symboles utilisés sont variés : depuis l'image reproduisant l'objet lui-même (automobiles, maisons, produits de consommation, etc.) jusqu'à l'image suggérant une identification en rapport avec le sujet traité dans l'étude, facile à repérer (des fûts de différentes tailles, pour représenter l'augmentation du prix de pétrole, des cheminées d'usine pour l'emploi industriel, etc.). Généralement, le chiffre représentant la quantité est inclus dans la figurine, ou immédiatement proche, s'il n'y a pas assez de place.

1. Nous employons le terme de « corpulence » pour spécifier qu'il s'agit de représentations de *surface* et non de *hauteur*.

Parfois, des dessins quelque peu « torturés » rendent l'identification involontairement difficile : la signification est loin d'être immédiate et la communication passe moins bien. Ici encore, la règle de la lecture en *dix secondes une fois le titre lu* ne doit pas être oubliée.

2 – Les cartogrammes

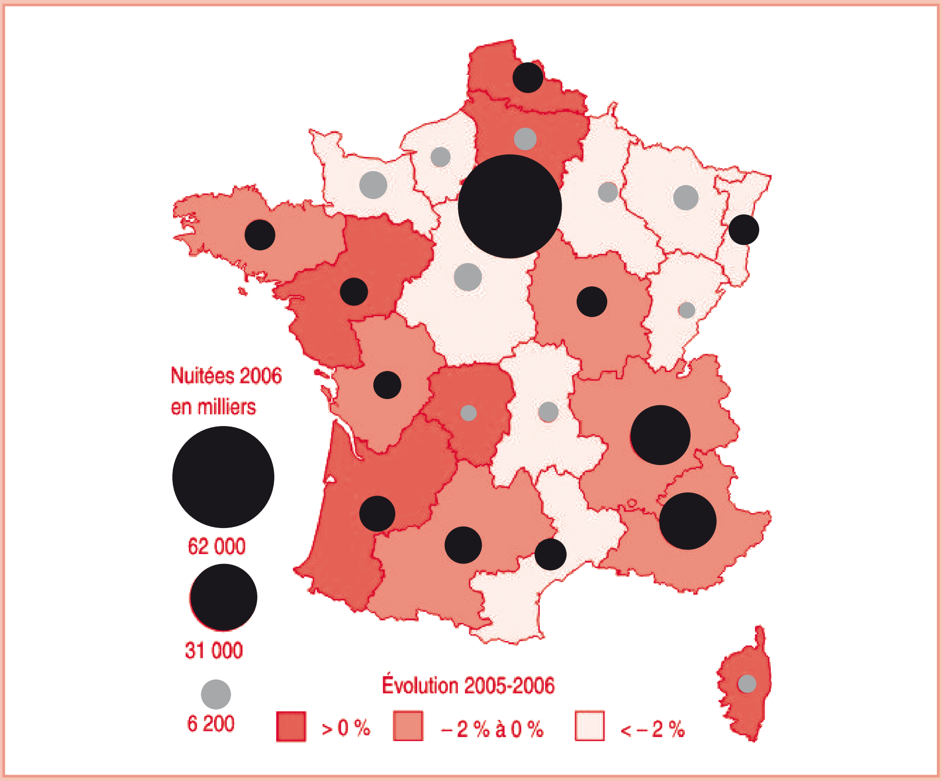
Les informations sont **repérées sur une carte**, soit par des **couleurs**, correspondant à des modalités bien indiquées dans la légende du diagramme, soit par des **cercles** de différents rayons. Dans ce dernier cas, la représentation est encore une **représentation de surface**, et non de hauteur.

La deuxième dimension est évidemment la **dimension spatiale** (villes, régions, pays, autres territoires).

Le cartogramme suivant, tiré de l'INSEE, *combine deux informations* (s'ajoutant à l'aspect régional) : **l'évolution annuelle** repérée par des couleurs, et **les volumes** du phénomène, repérés par des cercles :

Cartogramme de surface

Figure 3.25 – Nuitées dans les hôtels en 2006

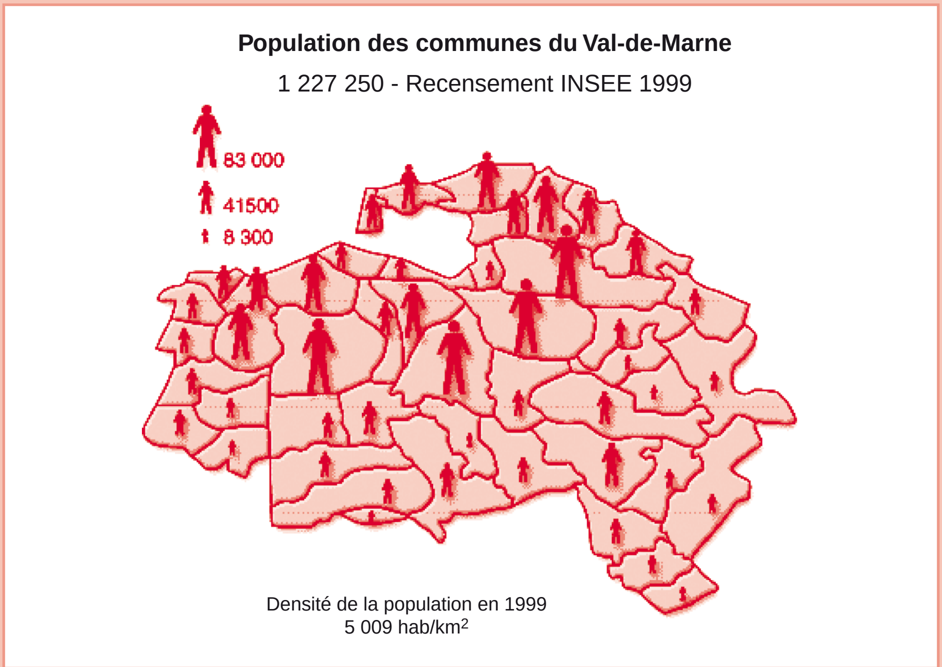


Source : INSEE, Direction du tourisme, partenaires régionaux, enquête de fréquentation hôtelière.
Référence : « L'hôtellerie et les campings en 2006 », in *INSEE Première*, n° 1 125, mars 2007.

Une représentation originale, « mixte » **cartogramme-pictogramme**, est donnée dans le diagramme suivant :

Cartogramme-pictogramme

Figure 3.26



Source : INSEE, Recensement RGP, 1999.

C – Les graphiques des analyses factorielles

Nous entrons ici dans le vaste domaine de « **l'analyse de données** » et de l'application de la *statistique descriptive mathématisée* aux méthodes dites « **multicritères** », ou « **multidimensionnelles** ». Ces méthodes aboutissent à des représentations graphiques modernes[1] et originales, qu'il faut savoir reconnaître quand on les rencontre dans des revues spécialisées ou des rapports. Dès lors, bien plus que l'aspect mathématique de construction, c'est **l'aboutissement graphique** qui nous intéresse dans ce paragraphe.

Les diagrammes que l'on a vus jusqu'à présent décrivent un phénomène par un ou deux caractères, très rarement plus. On parle de *dimensions,* comme on l'a vu dans le chapitre 2. Par exemple, on peut statistiquement « décrire » le niveau d'un étudiant, par une dimension : ses *notes* à l'examen. Or, la réalité est bien plus complexe : un étudiant ne peut pas être résumé par une note, ou par un ensemble de notes ; il est bien autre

1. Bien que la base de ces méthodes fût connue depuis le début du XX[e] siècle (K. Pearson, C. Spearman, LL. Thurstone, H. Hotelling), il faut attendre les années 1965, et la puissance de calcul des ordinateurs, pour qu'elles s'enrichissent et deviennent applicables dans la pratique, notamment avec J. L. Benzécri et son équipe.

chose (heureusement). On dit qu'il ne peut pas être « **réduit** » à **une** dimension. Son **profil** est plus large : il est « **pluri** » ou « **multidimensionnel** ». Ce qui est vrai pour l'étudiant est vrai pour bien des choses que l'on observe dans la réalité : un produit ne se réduit pas à un prix, un pays à sa production intérieure brute (PIB), un logement au nombre de ses pièces, un ouvrage littéraire à son nombre de pages, etc.

Le problème réside dans *l'appréhension* de ces multiples facettes qui constituent le « **profil** » de ce que l'on veut étudier.

Il est, dans la vie, un grand nombre de cas où l'observation statistique de deux individus, **repérés par un unique caractère**, apporte une vision d'éléments en opposition, alors que les profils sont proches : en matière d'élection présidentielle, par exemple, on remarque que certains électeurs d'extrême gauche ont un profil qui n'est pas très éloigné de ceux de certains électeurs d'extrême droite, alors que le caractère « position sur l'axe politique » les oppose au maximum ; de même, deux candidats à un emploi peuvent apparaître très proches d'un point de vue professionnel, et l'on aura du mal à les départager, alors que leurs profils, culturel, artistique ou de mode de vie, sont complètement différents. La vision statistique *unidimensionnelle* demande donc souvent à être dépassée.

Il faut donc synthétiser, et définir des « **critères** », les plus explicatifs et les plus objectifs possibles. Le choix des critères et des caractères qui les *mesurent* (dans le cas quantitatif), ou qui les *traduisent* (dans le cas qualitatif) se fait en fonction de la connaissance du terrain ou par enquête préalable.

Mais comment ne pas se perdre dans tous ces caractères ? C'est tout l'intérêt de ces *analyses de données* : on « **ne sélectionne pas** » tel caractère ou tel autre, mais on « **combine** » ces caractères entre eux, par des procédés algébriques[1], de manière à ce qu'ils rendent compte de leur dépendance ou indépendance. Ils font alors apparaître des « **facteurs communs** » ou au contraire des « facteurs spécifiques ». Le terme de *facteur* donne son nom à celui de l'analyse « **factorielle** » (titre du présent paragraphe).

Quelle **logique** dicte ces combinaisons ? C'est une logique purement « **visuelle** » à la base, on pourrait même dire « photographique » : en effet, si l'on veut prendre un cliché photographique d'un imposant édifice architectural, par exemple, on va se déplacer, se baisser, reculer, se hisser, afin d'obtenir l'image la moins *plate* possible. En d'autres termes, que fait-on ? On cherche à **maximiser les distances** entre les différents points de l'édifice, pour que *la photo présente le meilleur relief*.

C'est cette notion de *distances à maximiser* que l'on reprend dans les analyses factorielles. C'est moins direct et moins intuitif que dans le cas de la photo, car ce sont *des moyennes de carrés de distances entre des points de projection*, mais l'idée est la même : **étaler au maximum l'image[2] sur le diagramme**.

On parvient ainsi à *construire* de nouveaux caractères. Ces nouveaux caractères présentent quatre avantages :

1. *Combinaisons linéaires simples* de type le plus souvent matriciel, dont le détail des calculs algébriques ne nous intéresse pas ici.
2. Étaler l'image revient à maximiser sa dispersion ; on dit que l'on maximise son « **inertie** ».

D'abord, ils sont en nombre moins important que ceux que l'on avait choisis au début ; on a « **réduit** » les données.

Ensuite, ils « **synthétisent** » (ils résument) la réalité de départ en concentrant les informations, et par là même les caractères obtenus deviennent plus riches de sens, plus « **significatifs** ».

Ils sont parfois, nouveaux, non prévus, inattendus ; ce sont alors des caractères **cachés**, que l'analyse aura révélés.

Ils sont, enfin, **indépendants** les uns des autres, par construction, c'est-à-dire que chacun apporte une information *spécifique* et non influencée par les autres caractères.

Il existe plusieurs types d'analyses factorielles[1] ; la plus connue est « l'**analyse en composantes principales** » (**ACP**). La combinaison qui fait apparaître le maximum de facteurs communs s'appelle la « **première composante principale** », ou « **premier axe** » de la représentation graphique. On trouve ensuite la deuxième composante ou deuxième axe, et ainsi de suite.

Il faut bien comprendre que ces axes, révélés par la méthode, n'ont *pas encore de nom* ; tout au plus auraient-ils un nom commun à tous les caractères qui ont participé, à des degrés divers, à leur fabrication, ce qui risquerait de donner une appellation complètement incompréhensible. Il faut donc « **interpréter les axes** » : cette étape est à la fois la plus délicate et la plus pertinente de l'analyse.

L'*interprétation des axes factoriels* doit aboutir à une **signification concrète**, comprise par les utilisateurs sans aucune équivoque. On peut s'aider de calculs spécifiques[2] que fournit un grand nombre de programmes informatiques, mais cela ne permet, en aucune manière, de se passer de la connaissance du terrain : l'utilisateur y est le plus souvent le meilleur guide.

La méthode graphique consiste, ensuite, à **observer la disposition des points** sur le diagramme final, en opérant **des regroupements**, pour en tirer le profil le plus adapté à chaque axe, et c'est le point le plus intéressant de la lecture.

Prenons deux exemples : le premier **fictif** pour évoquer un *essai d'interprétation* des axes factoriels en ACP, le second sur **données réelles** pour revenir dans le concret, et savoir **reconnaître et lire** ce genre de diagramme.

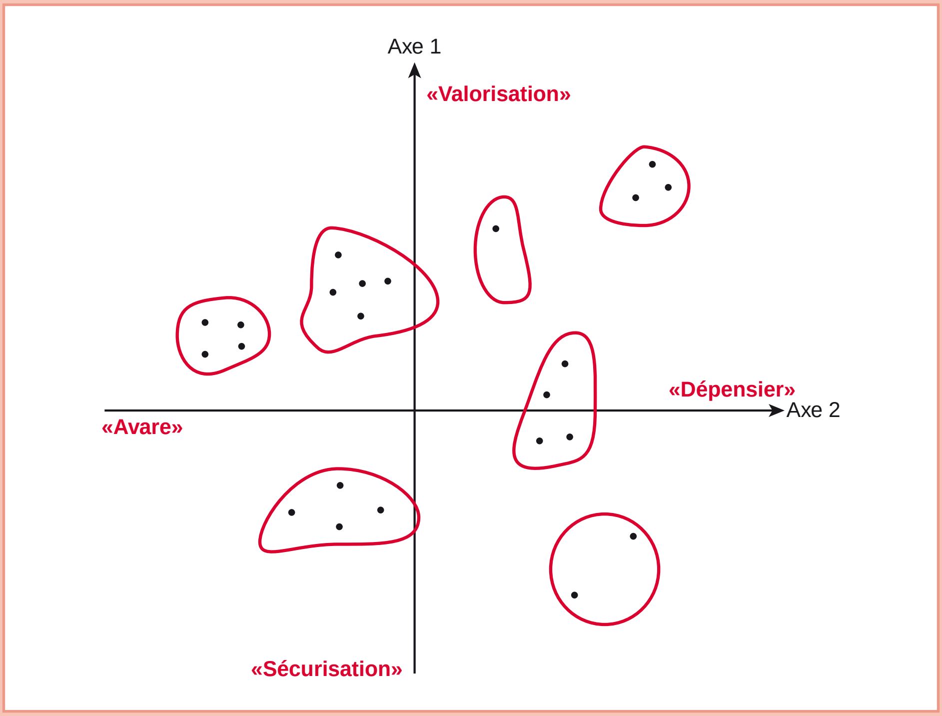
- Pour le premier, imaginons une chaîne de boutiques de vêtements de sport qui s'intéresse aux motifs subjectifs d'achat par types de clientèle. L'enquête portera sur des **critères** qui seront repérés par des **caractères à une dimension** : la catégorie socioprofessionnelle (classée en *supérieure, haute, médiane, basse, très basse*), l'âge, le nombre d'enfants, l'habitat et la distinction *urbain-rural*, le secteur d'activité (*agriculture, commerce, industrie, bâtiment, mode, théâtre, etc.*), le sport pratiqué, le goût pour les voyages (*organisés ou non, de type aventure ou non, courts séjours ou non, etc.*), les types de véhicules du ménage, les types de commerces fréquentés (*alimentaires, restauration, sorties*), et bien d'autres questions que l'on pourrait imaginer, sans que, bien sûr, le questionnaire ne tourne à l'interrogatoire policier.

1. Analyses : *canonique, discriminante, hiérarchique, des correspondances.*
2. Calculs de corrélation (voir ce mot au chapitre 7) entre chaque axe et ses composantes, les caractères initiaux.

On pourrait obtenir le diagramme suivant, en quatre quadrants, où **les regroupements** en « pavés » représentent les *proximités* de profil des groupes recensés, et les axes représentent les « socio-styles » d'achat. Les axes sont échelonnés **par opposition** (un comportement en haut, son contraire en bas).

Type de regroupements graphiques en analyse factorielle

Figure 3.27 – Motifs subjectifs d'achat par types de clientèle



Selon les résultats de l'analyse, le premier axe (ici, l'axe vertical) **pourrait être interprété** comme repérant le caractère **subjectif** de la volonté d'achat : on oppose *la sécurisation* (« achetez ce pull, il est parfait, bien fabriqué, avec des laines de qualité irréprochable ») à la *valorisation* (« achetez ce pull, cela marquera votre appartenance à une élite »).

Le deuxième axe (ici l'axe horizontal) pourrait concerner, non pas le pouvoir d'achat réel du client, mais son **comportement** en la matière, en opposant le *style économe* (« avare » dans le graphique) au *style dépensier*.

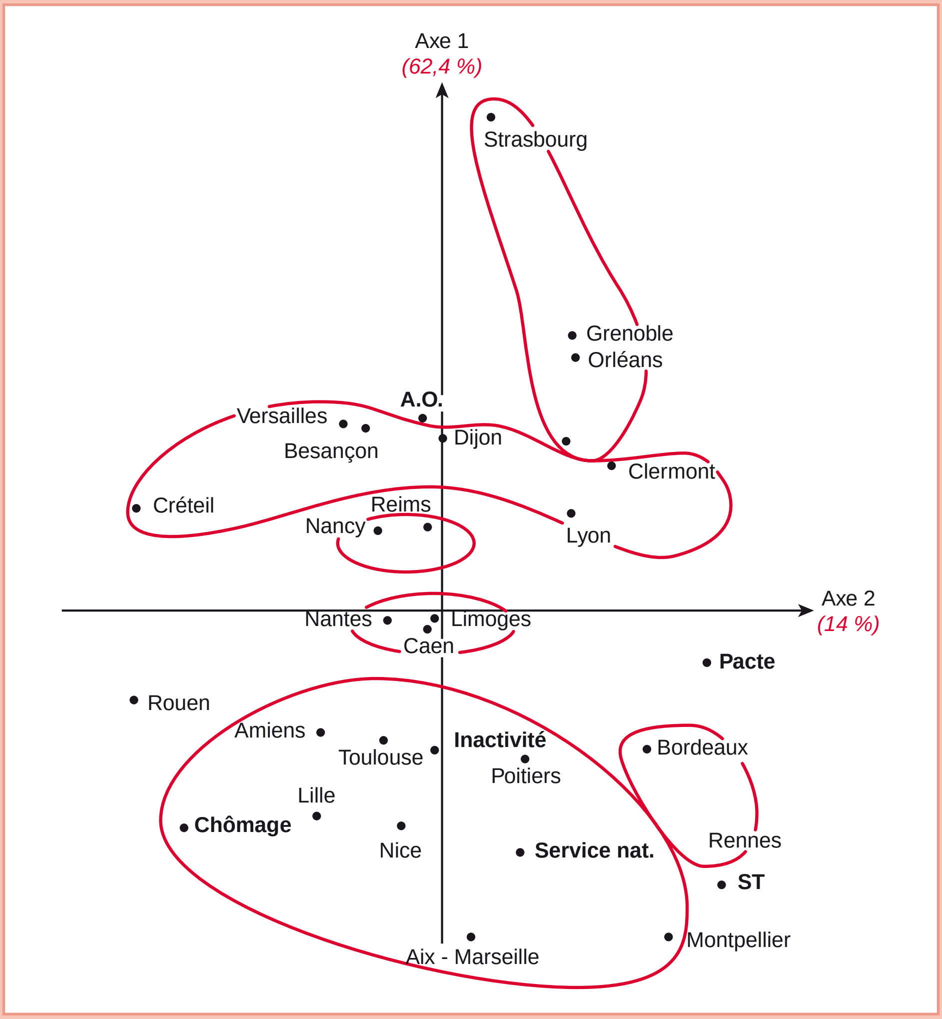
- Le deuxième exemple (voir page suivante) est tiré d'une étude du CEREQ[1], dans laquelle l'auteur, sur la base d'une enquête passée en 1980, analyse les critères de **sortie des jeunes de l'appareil éducatif et leur insertion dans les différentes régions** de France. Les régions sont ici repérées en 25 académies. Les groupes concernés (*garçons*, dans cet exemple) sont déclinés selon les niveaux de formation, le statut

1. CEREQ, « Quand les jeunes formés au niveau CAP-BEP abordent la vie active dans les régions », in *Formation Emploi* n° 8, par Jean Biret, octobre-décembre 1984.

d'actif (ayant effectivement un emploi, ou chômeur), le statut d'inactif (en stage, ou sans activité déclarée), le statut d'occupé (en emploi formation dans le cadre des contrats du 3e pacte pour l'emploi, ou en dehors de la réglementation de ce pacte), le statut de militaire du service national. Le diagramme de type ACP[1] est présenté ci-dessous (figure 3.28) :

Un exemple réel d'analyse en composantes principales

Figure 3.28 – Insertion des jeunes sortis de l'appareil éducatif dans différentes régions de France



Source : CEREQ, *Formation Emploi* n° 8, octobre-décembre 1984

1. C'est, plus exactement, une « **analyse des correspondances** », de type Benzécri, qui peut être considérée comme une ACP sur des modalités de caractères qualitatifs, codées en valeurs entières numérisées.

Les regroupements par proximités de points concernent les régions (plus exactement les 25 académies). L'interprétation du premier axe factoriel (ici vertical) oppose, en haut l'activité occupée en dehors du pacte (**AO**) et, en bas, le service national (**Service nat.**).

Le deuxième axe oppose les formes **pacte** et **chômage** de façon assez nette. L'auteur ajoute que « l'**inactivité** et les formes de type stage (**ST**) ne contribuent quasiment pas à la formation des axes 1 et 2 ». On voit bien, par exemple, dans le pavé du bas, **Lille proche de Nice**, mettant en relief que ce n'est pas le degré de latitude géographique qui explique la relative faiblesse des taux d'occupation masculins, hors pacte.

Loin des conclusions et résultats chiffrés, nous ne présentons ici cet exemple pris dans la réalité que dans un but uniquement **pédagogique** : celui de **nous familiariser** avec la lecture des représentations graphiques d'analyses factorielles.

Hors l'exemple du CEREQ, ajoutons un mot sur les limites de ces analyses factorielles : si la projection des points et le calcul des axes obéissent à une *indiscutable* rigueur mathématique, il n'en va pas toujours de même de *l'utilisation* qui en est faite ensuite : les axes demandent à être « interprétés » et, dans bien des cas, leur « signification » n'est pas directement évidente, ce qui risque de conduire à des approximations discutables. Il vaut mieux alors choisir d'autres modes d'investigation. De plus, quand on regroupe les points les plus proches, il est bon de *vérifier* que les deux axes rendent compte suffisamment bien de la *réalité du terrain*.

À défaut, on risque de considérer comme voisins des points risquant d'être très éloignés sur le troisième axe, voire le quatrième. Un peu comme quand les anciens ont regroupé dans chaque constellation les étoiles qui leur paraissaient proches, vues de la terre, mais que l'on a découvert en fait très éloignées les unes des autres dans l'espace sidéral.

Exercices sur les graphiques, diagrammes et autres messages visuels

***Nota* important :** Les résolutions de calculs par Excel 2007, selon une démarche « pas à pas », sont proposées à la fin des réponses concernées.

Le lecteur non débutant, ayant une connaissance suffisante des logiciels, pourra aisément se passer d'un certain nombre de détails que nous développons ici, et trouver facilement des raccourcis.

Exercice 1 : Cas d'une variable quantitative discrète : choisir le message visuel adapté

Énoncé :

Reprenons les résultats de l'exercice n° 1 page 57 précédent, concernant le nombre d'enfants par familles salariées de l'entreprise Machin. Le tableau final que nous avions construit était le suivant :

Nombre d'enfants par famille des salariés de l'établissement Machin

Nombre d'enfants par famille (x_i)	Nombre de familles (n_i)
0	9
1	7
2	10
3	0
4	4
	TOTAL 30

Source : Recensement réalisé par le comité d'entreprise de l'établissement Machin durant le mois de décembre 2009, auprès du personnel salarié.

Question 1 Construisez le graphique le plus adapté à cette situation.

Question 2 Si l'on bénéficiait de résultats quantitatifs supplémentaires donnant la distance habitat – travail (en quatre classes kilométriques) des familles enquêtées, pourrait-on les faire apparaître sur ce graphique ? Choisissez deux méthodes de représentation et tracez les graphes en inventant des valeurs cohérentes. Justifiez votre réponse.

Question 3 Au vu de ces résultats cohérents, présentez le tableau à deux dimensions dont les valeurs vous ont servi à fabriquer les deux diagrammes précédents.

Corrigé de l'exercice 1

Réponse à la question 1

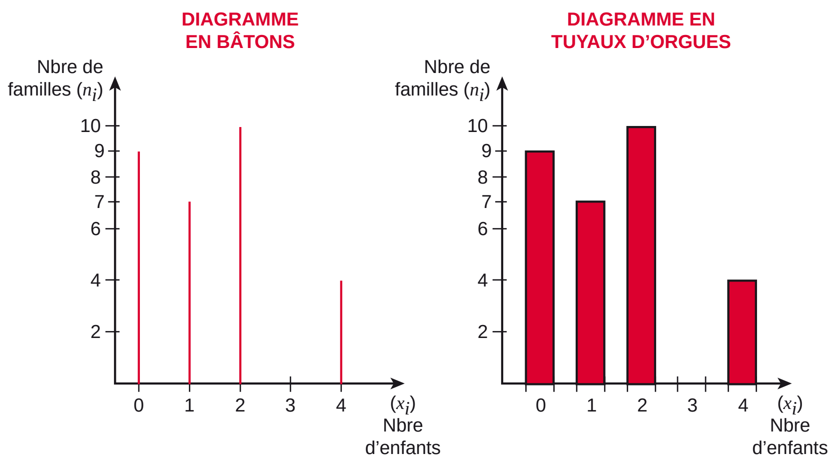
Le caractère est de type **quantitatif discret** (les modalités sont des nombres entiers, et l'on passe de 1 à 2, puis à 3, et ainsi de suite, de façon discontinue, sans chiffres après la virgule) ; rappelons que dans les cas quantitatifs, on assimile les deux termes *caractère* et *variable*.

Le diagramme de référence peut être un « diagramme en bâtons » ou « en tuyaux d'orgues »[1], permettant de reporter sur l'axe vertical **des hauteurs** ; ces hauteurs sont les valeurs des effectifs étudiés, c'est-à-dire le nombre de familles correspondant à chaque modalité.

Ces deux possibilités de graphique sont présentées dans le schéma 3.1 ci-dessous.

Schéma 3.1

Nombre d'enfants par famille des salariés de l'établissement Machin



Source : recensement réalisé par le comité d'entreprise de l'établissement Machin durant le mois de décembre 2009, auprès du personnel salarié.

Notons que le diagramme en tuyaux d'orgues est généralement préféré à celui en simples bâtons, car les bâtons, assez fins, ne se perçoivent pas toujours très bien visuellement.

1. Ou également en « barres » horizontales ; voir page 71.

Résolution par Excel

Rappelons que le lecteur ayant une connaissance suffisante des logiciels pourra aisément se passer d'un certain nombre de détails développés ici. Nous restons volontairement à un niveau « débutant ».

A) **Saisir** dans une colonne les nombres correspondant aux effectifs (9 ;7 ;10 ;0 ;4), puis sélectionner **Insertion** dans la barre des menus, puis **Graphique colonnes**, ensuite ce qu'Excel appelle « histogramme » 2D groupé[1]. Le diagramme apparaît.

B) Cliquer sur **Disposition** dans la barre des menus, puis, dans la barre d'outils (au-dessous), choisir **Titre du graphique**, puis **Au-dessus de graphique**. Remplacer « titre » par *Nombre d'enfants par famille* dans la zone du graphe, et cliquer hors du graphe.

Revenir sur **Disposition**, et choisir dans la barre d'outils **Titre des axes**. Cliquer sur **Titre de l'axe principal**, puis sur **Titre en dessous de l'axe**.

Dans la zone du graphe, déplacer (en faisant apparaître le curseur) « titre de l'axe » vers la droite, et remplacer ces mots par **Nombre d'enfants**.

Cliquer à nouveau sur **Disposition**, **Titre des axes**, et répéter la procédure pour l'axe vertical, dont le titre est **Nombre de familles**, en élargissant la case si nécessaire.

Cliquer hors du graphe et continuer si besoin par des choix sur les **couleurs**, les **styles** et autres détails plus ou moins personnels.

---

1. Qui n'est pas un histogramme, mais un diagramme en tuyaux d'orgues.

Réponse à la question 2

Si l'on bénéficiait de ces résultats supplémentaires, on doublerait l'information, et on passerait donc à un diagramme à deux dimensions.

La première méthode consisterait à *partitionner* les « tuyaux » en autant de segments verticaux que de classes d'âge fournies, et à utiliser des couleurs ou des trames diverses pour les différencier visuellement (voir le schéma 3.2 page suivante).

Si l'on a par exemple les quatre classes de distances suivantes **[de 0 à 5 km [, [de 5 à 10 km [, [de 10 à 20 km [, [20 km et plus [** pour chaque famille, on peut déterminer la part de chaque classe de distances afin de partitionner chaque tuyau d'orgue **proportionnellement**, comme on le voit dans le diagramme page suivante.

Résolution par Excel

A) **Saisir** dans une première colonne le titre [0 ;5[km et les nombres d'effectifs suivants, que l'énoncé ne donne pas mais que nous proposons[1] **(4 ;3 ;4 ;0 ;2)**, puis dans une deuxième colonne, le titre [5 ;10[km et les nombres **(3 ;1 ;3 ;0 ;1)**, puis dans une troisième colonne, le titre [10 ;20[et les nombres **(1 ;2 ;2 ;0 ;0)**, et enfin dans une quatrième, le titre [20 et plus[et les nombres **(1 ;1 ;1 ;0 ;1)**. **Sélectionner** la plage des 4 colonnes et cliquer sur **Insertion** dans la barre des menus, puis sur …

---

1. Ce sont ceux de la réponse à la dernière question, du tableau pages 103 et 104, proposés pour respecter les totaux des données.

…

Graphique colonnes, ensuite sur ce qu'Excel appelle « histogramme » empilé[2]. Le diagramme apparaît.

Choisir le **style**, la **couleur**, et les différentes options que l'on désire.

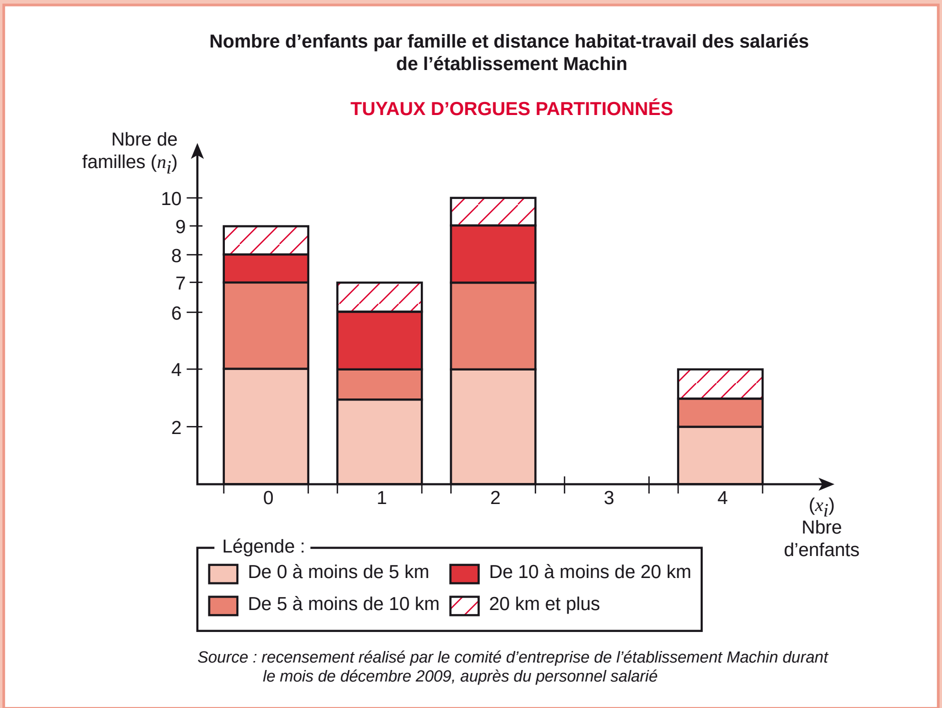
B) Recommencer la procédure B) du paragraphe précédent pour les noms des axes.

---

1. Ce sont ceux de la réponse à la dernière question, du tableau pages 103 et 104, proposés pour respecter les totaux des données.
2. Qui n'est pas un histogramme, mais un diagramme en tuyaux d'orgues.

Résolution par Excel

Schéma 3.2



La deuxième méthode consiste à « accoler » les classes de hauteurs *proportionnelles aux effectifs partiels*, comme le montre le schéma 3.3 (page suivante). Cela est possible ici car nous n'avons que quatre types de classes de distances ; à plus de quatre, le diagramme serait trop chargé.

Attention, il faut que les *hauteurs* dans ce graphe soient en cohérence avec les *partitions* choisies précédemment

Ainsi, par exemple, pour les familles sans enfants, le schéma précédent indiquait que 4 familles habitaient entre 0 et 5 km du lieu de travail : il faut donc que la hauteur du premier tuyau dans le diagramme ci-dessous soit égale à 4 unités, sinon le diagramme sera faux.

Résolution par Excel

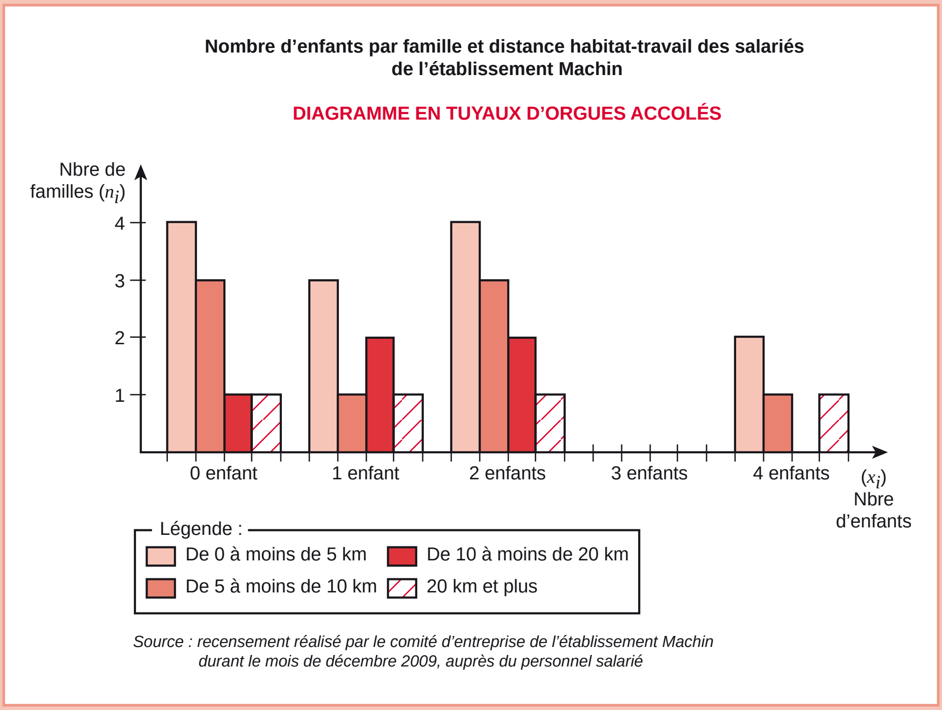
A) **Saisir** dans une première colonne le titre [0 ;5[km et les nombres d'effectifs suivants, que l'énoncé ne donne pas mais que nous proposons[1] **(4 ;3 ;4 ;0 ;2)**, puis dans une deuxième colonne, le titre [5 ;10[km et les nombres **(3 ;1 ;3 ;0 ;1)**, puis dans une troisième, le titre [10 ;20[et les nombres **(1 ;2 ;2 ;0 ;0)**, et enfin dans une quatrième, le titre [20 et plus[et les nombres **(1 ;1 ;1 ;0 ;1)**. **Sélectionner** la plage des 4 colonnes et cliquer sur **Insertion** dans la barre des menus, puis sur **Graphique colonnes**, enfin sur ce qu'Excel appelle « histogramme » groupé[2]. Le diagramme apparaît.

Choisir le **style**, la **couleur**, et les différentes options que l'on désire.

B) Recommencer la procédure du paragraphe B) précédent pour les noms des axes.

1. Ce sont ceux de la réponse à la dernière question, du tableau page 104, proposés pour respecter les totaux des données.
2. Qui n'est pas un histogramme, mais un diagramme en tuyaux d'orgues.

Schéma 3.3



Réponse à la question 3

Ici nous avons choisi, à titre d'exemple, les valeurs suivantes qui **garantissent la cohérence** entre les deux diagrammes :