CHAPITRE 2 - LA METHODE STATISTIQUE : QUELQUES CONCEPTS IMPORTANTS

2.1 Quelques définitions de base

En statistique comme en tout autre domaine, il est nécessaire de définir des mots et des expressions que l'on pourra utiliser pour désigner des notions spécifiques. Dans la mesure du possible on évitera d'employer trop de "jargon", mais il nous faudra néanmoins utiliser quelques termes techniques. Ces derniers peuvent prêter à confusion car beaucoup d'entre eux sont utilisés dans le langage quotidien : en statistique leur signification diffère légèrement de l'usage courant. Il est donc utile de s'arrêter un instant sur les termes techniques les plus largement utilisés afin que leur contenu soit bien compris.

Lorsque l'on réalise une collecte de données, quel que soit son but, on doit savoir avec précision quel type d'information on recherche, à qui ou à quoi elle se rapporte, de quelle façon on l'obtiendra, et quel groupe d'individus ou d'objets on étudie. Chacun de ces éléments est désigné par un terme technique : on dira que l'on observe un (ou des) caractère(s) sur des unités statistiques appartenant à une population. Les termes observation, caractère, unité statistique et population ont un sens particulier en statistique. Nous allons voir ce qu'ils signifient et de quelle manière on les utilise.

Unité statistique - Ce terme est utilisé pour désigner toute personne, groupe d'individus, objet ou autre élément pour lequel on souhaite obtenir une information numérique. Par exemple : une personne, une famille, un ménage, un village, un bâtiment, une région, une île, un poisson, un bateau, un port, un intervalle de temps comme une semaine ou une année, une église, un établissement commercial, etc. On pourrait donner beaucoup d'autres exemples d'unités statistiques.

Population - Lorsque l'on collecte des données statistiques, on recherche des informations sur un groupe d'unités statistiques - pour désigner ce groupe dans son ensemble, on emploie le mot "population". Dans la langue courante, ce terme est souvent utilisé : on parle par exemple de la population de Tonga ou de Suva. En statistique, on utilisera le mot population pour désigner un groupe d'unités statistiques, quelle que soit leur nature. Ainsi, on pourra parler de la population des ménages d'Apia, de la population de bateaux de pêche à Rarotonga ou encore de la population de tous les poissons pêchés à Tuvalu en 1983 par exemple. Lorsque l'on recueille des données, on prend soin de définir avec précision la population à laquelle on s'intéresse.

Observation - Ce mot désigne l'action de réunir un élément d'information, quelle que soit la méthode utilisée. Habituellement, l'observation est réalisée par une personne, parfois à l'aide d'instruments, mais dans certains cas l'observation et l'enregistrement des données sont réalisés automatiquement par des machines. Il est important de comprendre que, au sens statistique, l'observation désigne n'importe quelle méthode permettant de recueillir les données et non pas le simple fait de regarder puis écrire un résultat. Les méthodes les plus courantes d'observation statistique sont : la mesure, le dénombrement, l'interrogation d'une personne, le relevé de données dans des documents, le questionnaire individuel, etc.

Caractère - On appelle caractère l'une des propriétés ou l'un des aspects de l'unité statistique sur lequel portera l'observation. Par exemple, on pourra observer le poids d'un poisson, la superficie d'une ferme, la valeur totale des marchandises importées dans un port, le revenu annuel d'un ménage, le nombre d'individus habitant dans un village, etc. Le plus souvent une même unité statistique présente plusieurs caractères pouvant faire l'objet de l'observation.

Ainsi, par exemple, la collecte de données sur une campagne de pêche à la canne dans les eaux territoriales d'un pays peut concerner l'observation de quelques-uns des caractères suivants :

port de départ;
pays d'immatriculation;
tonnage brut du navire;
effectif de l'équipage;
nombre de jours de pêche dans les eaux territoriales;
espèces capturées;
quantité de poisson capturée;
poids moyen des poissons.

On distingue deux catégories de caractères. La première regroupe les caractères qui ne peuvent s'exprimer que sous une forme numérique. Ces caractères sont dits quantitatifs ou variables. La seconde catégorie comprend les caractères qui ne prennent pas de valeur numérique : contrairement aux précédents, ils doivent être décrits par des mots. On parle alors de caractères qualitatifs.

Dans la liste de caractères donnée ci-dessus à propos de la pêche à la canne, on peut identifier les caractères quantitatifs suivants :

tonnage brut du navire;
effectif de l'équipage;
nombre de jours de pêche;
quantité de poisson capturée;
poids moyen des poissons.

Les caractères qualitatifs sont les suivants :

port de départ;
pays d'immatriculation;
espèces de poissons.

Il est souvent plus pratique, particulièrement lorsque l'on utilise un ordinateur, de travailler sur des données qui sont toutes exprimées sous forme numérique. Pour cette raison, on attribue parfois des codes numériques aux valeurs prises par les caractères qualitatifs. Ainsi on pourrait par exemple attribuer le code 001 à la bonite, 002 au thon jaune, 003 au thon obèse, 004 au germon, et ainsi de suite. Ces numéros de code sont entrés dans l'ordinateur à la place des noms des espèces. Il faut toutefois bien comprendre que la codification ne change en rien la nature du caractère, qui demeure qualitatif : le code ne fait que remplacer un mot.

2.2 Conventions d'écriture

On utilisera dans ce cours un certain nombre de caractères ou symboles conventionnels qui permettent d'alléger la notation des concepts

statistiques. Nous essayerons de nous en tenir à une symbolique aussi simple que possible.

Les signes conventionnels seront introduits progressivement tout au long de ce cours, mais il est utile d'en définir quelques-uns dès à présent.

n, N : Le nombre d'observations prises en considération est noté "n" lorsqu'il s'agit d'un échantillon et "N" lorsqu'il s'agit de la population totale. Ainsi, si la collecte des données porte sur un échantillon de 17 navires de pêche, on écrira n = 17. Si la flotte totale est composée de 80 navires, on notera N = 80.

x Lorsque l'on observe une variable, c'est-à-dire les valeurs prises par le caractère qui nous intéresse, on utilisera le symbole "x" pour désigner le résultat de l'observation. A ce symbole est souvent associé un indice qui correspond au numéro de l'observation. En d'autres termes, x_1 désignera la valeur prise par le caractère pour la première unité statistique observée, c'est-à-dire la première observation; x_2 désigne le résultat de la seconde observation, et ainsi de suite jusqu'à la dernière observation (la "énième" observation) dont le résultat sera désigné par x_n.

Ainsi, si l'on mesure la longueur des poissons en centimètres, et que le premier poisson de l'échantillon présente une longueur de 62 centimètres, on notera $x_1 = 62$. (L'ensemble des observations pourra être désigné par x_1, x_2, x_3, ... x_n.)

y Lorsqu'on étudie deux variables, c'est-à-dire lorsque l'on s'intéresse à deux caractères différents pour chaque unité statistique observée, on notera par "y" la valeur prise par le second caractère. Par exemple, si l'on veut établir un rapprochement entre le poids et la longueur des poissons et si le premier poisson de l'échantillon pèse 3,8 kg on écrira $x_1 = 62$ et $y_1 = 3,8$.

i Indice désignant le numéro de l'observation. On parle de la ième observation. Pour la première observation, i = 1, et ainsi de suite. Le plus souvent, "i" est associé à une variable. X_i renvoie à la valeur prise par la variable x pour la ième observation.

$\sum$ Lettre grecque sigma majuscule signifiant "somme de". Elle est à distinguer de la lettre sigma minuscule, notée σ, qui sera présentée plus loin dans ce cours.

2.3 Graphiques

Dans les chapitres qui suivent, on présentera de nombreuses notions statistiques sous forme de graphiques. On utilisera surtout trois types de graphiques : les graphiques en nuage ou de dispersion, les graphiques cartésiens, et les graphiques en barres ou à colonnes. Nous dirons également quelques mots sur les graphiques à secteurs, qui, bien qu'intervenant assez peu dans la suite du cours, constituent une méthode intéressante pour présenter l'information sous forme graphique.

L'étude des graphiques est un sujet très important, et la manière d'utiliser les graphiques pour présenter des données statistiques peut affecter profondément la compréhension des résultats. Toutefois, nous n'allons pas entreprendre une étude détaillée de ce domaine. Nous aborderons simplement les principes fondamentaux qui guident la construction de ces graphiques, ce qui nous servira de transition pour les chapitres suivants.

Un graphique est utilisé pour illustrer une relation entre des variables. Il se compose des principaux éléments suivants :

(a) Une légende : principalement le numéro du graphique et un titre décrivant ce que le graphique représente;

(b) Deux axes : l'axe horizontal ou "abcisse" ("axe des x") et l'axe vertical ou "ordonnée" ("axe des y"). L'abcisse et l'ordonnée se rencontrent à l'"origine" ("0"). Chaque axe doit être clairement libellé et gradué selon l'échelle choisie pour la variable correspondante;

(c) Les données : figurées sur le graphique conformément aux principles établis pour le type de graphique choisi.

2.3.1 Graphique en nuage

Le graphique en nuage, ou de dispersion, est utilisé pour étudier la nature de la relation entre deux caractères. En général ,on dispose de deux séries d'observations pour un groupe d'unités statistiques. Occupons-nous uniquements des variables. Supposons que l'on dispose d'un échantillon de n unités : pour chaque unité, on observe deux caractères (ou variables) notés x et y. Les observations relatives à la première unité seront notées x_1y_1, celles relatives à la seconde unité x_2y_2, et ainsi de suite. D'une manière générale, on notera x_iy_i le couple d'observations relatif à la ième unité statistique, et il y aura un total n couples d'observations. Si l'on trace un graphique comportant deux axes, une variable étant associée à chaque axe, un couple d'observations sera matérialisé par un point dont les coordonnées seront (x_i, y_i). Ce type de graphique où les n couples d'observations sont représentés par des points est appelé "graphique en nuage" (dit de dispersion) ou encore "nuage de points". La représentation du nuage est très utile lorsque l'on souhaite obtenir une première indication sur le type de relation existant entre les deux variables x et y.

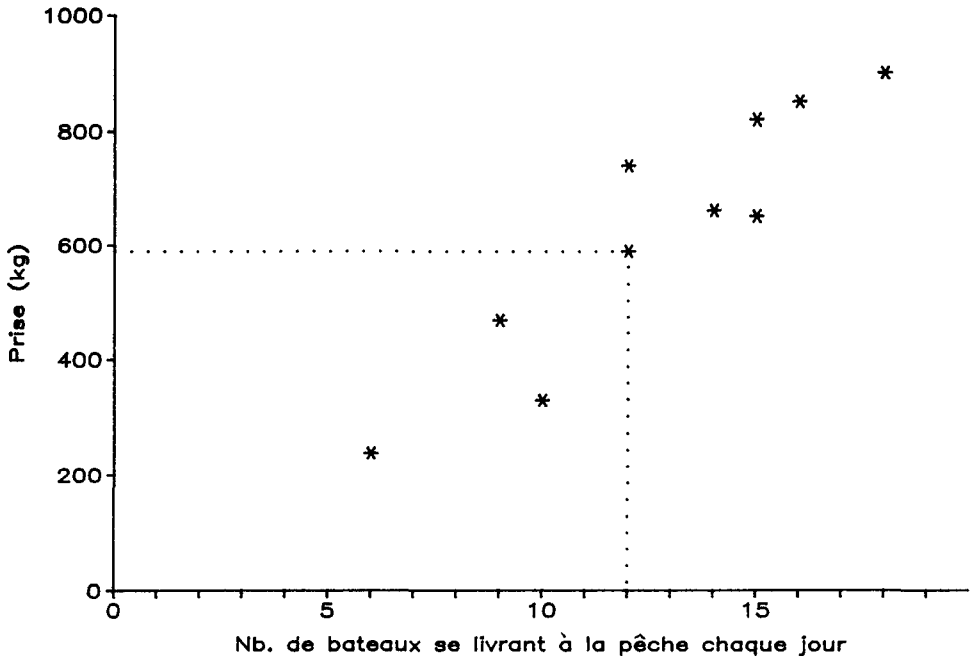
Supposons par exemple que, pour un échantillon de 10 journées, on ait collecté des données sur le nombre de bateaux se livrant à la pêche et la quantité de poisson capturée journellement (tableau 2.1) :

Sur un graphique en nuage, on indiquera le nombre de bateaux se livrant à la pêche en abcisse (x) et la prise en ordonnée (y). Les données correspondant à la première journée sont figurées par un point dont les coordonnées sont : x = 12 (bateaux) et y = 590 (kg). Ce point représente le couple d'observations x_1y_1. Sur le graphique, des lignes en pointillé ont été tracées pour indiquer précisément comment la position du point (x_1y_1) a été déterminée. Dans la pratique on ne représente pas ces lignes : seuls les points apparaissent. La figure 2.1 est le graphique en nuage des 10 couples d'observations.

TABLEAU 2.1 : NOMBRE DE BATEAUX SE LIVRANT A LA PECHE ET PRISE JOURNALIERE D'UNE PECHERIE ARTISANALE

Jour	Nombre de bateaux (x)	Prise totale (kg) (y)
1	12	590
2	15	820
3	10	330
4	12	740
5	18	900
6	14	660
7	6	240
8	15	650
9	16	850
10	9	470

FIGURE 2.1 : NOMBRE DE BATEAUX SE LIVRANT A LA PECHE ET PRISE JOURNALIERE

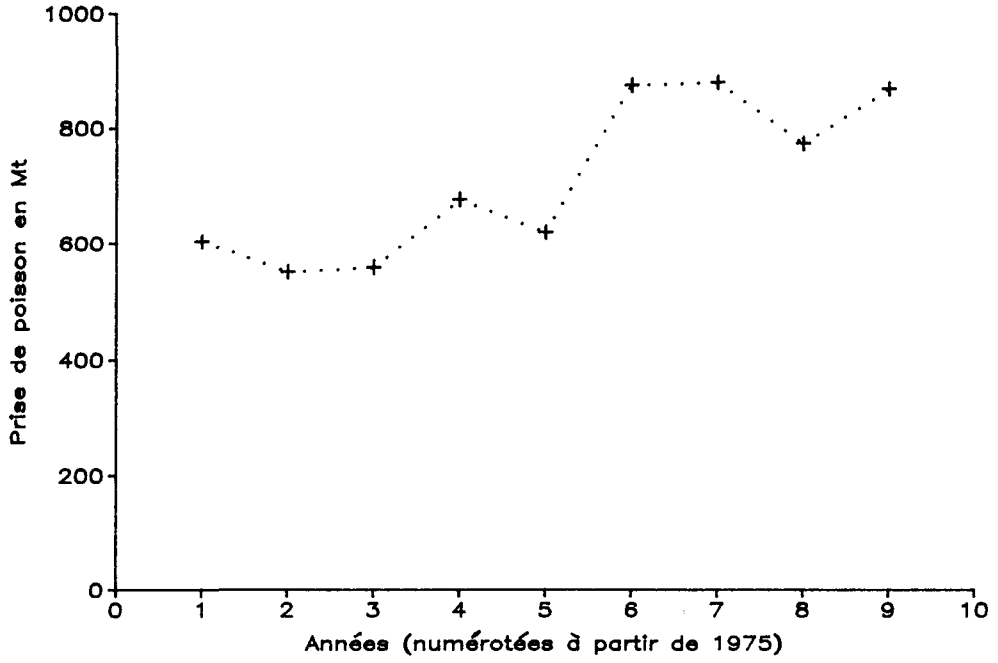


2.3.2 Graphiques cartésiens

Le graphique cartésien diffère peu du graphique en nuage. Tous deux illustrent une relation entre deux variables mais dans le graphique cartésien, les points sont reliés par des segments afin de montrer quelle est la tendance suivie par la relation.

Ce type de graphique est très utile lorsque l'on veut montrer comment des variables évoluent dans le temps. Un exemple est donné à la figure 2.2.

FIGURE 2.2 : PRISE ANNUELLE TOTALE DE POISSON DANS LE PAYS ABC (graphique cartésien)



On notera que les segments qui joignent les points ont une signification réelle : la pente de chaque segment indique si le volume de la prise augmente ou diminue d'une année à l'autre. Sur le graphique précédent, de tels segments n'auraient aucun sens. Ils n'illustreraient pas une tendance.

Les graphiques peuvent être construits selon d'autres méthodes que celle consistant à relier des points par des segments, comme dans le cas présent. Ils se présentent souvent comme des courbes figurant la forme de différentes distributions. Ceci fait l'objet du paragraphe suivant.

2.3.3 Graphiques en barres

Si l'on souhaite présenter graphiquement des données classées selon les valeurs prises par un caractère qualitatif, les graphiques précédents ne conviennent pas. Par exemple, si l'on considère une série statistique donnant la production de poisson selon le district, il n'est pas possible de placer les différents districts en abcisse et de joindre la série de points par une ligne : cette dernière n'aurait aucune signification. Dans un tel cas, la meilleure représentation graphique des données sera le graphique en barres appelé aussi graphique en bâtons ou à tuyaux d'orgue.

L'usage du graphique en barres ne se limite pas aux cas où l'un des caractères observé est qualitatif. On pourra aussi l'utiliser pour décrire des relations entre caractères quantitatifs. La figure 2.3 ci-après reprend sous forme de graphique en barres les données sur les prises annuelles de poisson présentées précédémment dans un graphique cartésien (figure 2.2).

FIGURE 2.3 : PRISE ANNUELLE TOTALE DE POISSON DANS LE PAYS ABC (graphique en barres)

Les bâtons ou colonnes peuvent être rapprochés ou au contraire espacés les uns des autres, et l'on peut les hachurer ou les foncer légèrement pour améliorer la présentation. On examinera plus loin un type particulier de graphique en bâtons, appelé histogramme, dans lequel les données sont présentées sous forme d'une série de bâtons contigus.

2.3.4 Variable dépendante et variable indépendante

Ayant choisi le type de graphique qui convient le mieux pour figurer les données, il nous faut décider lequel des deux caractères sera porté en abcisse (axe des x) et lequel sera en ordonnée (axe des y). Pour déterminer dans quel sens sera construit le graphique, nous devons examiner s'il peut exister une relation quelconque entre les variables ou caractères. Le plus souvent, on s'intéresse aux variations manifestées par l'un des caractères lorsque la valeur de l'autre se modifie. Dans notre exemple sur la prise totale de poisson présenté aux figures 2.2 et 2.3, on cherchait à montrer les variations du volume de la prise en fonction du temps. A la figure 2.1 on s'intéressait aux variations de la prise selon le nombre de bateaux se livrant à la pêche.

Dàns le premier cas (figures 2.2 et 2.3), on peut dire que le volume de la prise dépend du temps (l'année); et du nombre de bateaux engagés, dans le second cas (figure 2.1). De façon plus formelle, on parlera de variable dépendante et de variable indépendante. Dans nos deux exemples, il n'est pas possible d'inverser le sens de la relation; l'écoulement du temps par exemple ne dépend pas du volume de poissons capturés.

Compte tenu de cette relation indépendance/dépendance, on portera toujours la variable indépendante en abcisse (axe des x) et la variable dépendante en ordonnée (axe des y). Il s'agit d'une convention mathématique qui permet une lecture plus facile des graphiques.

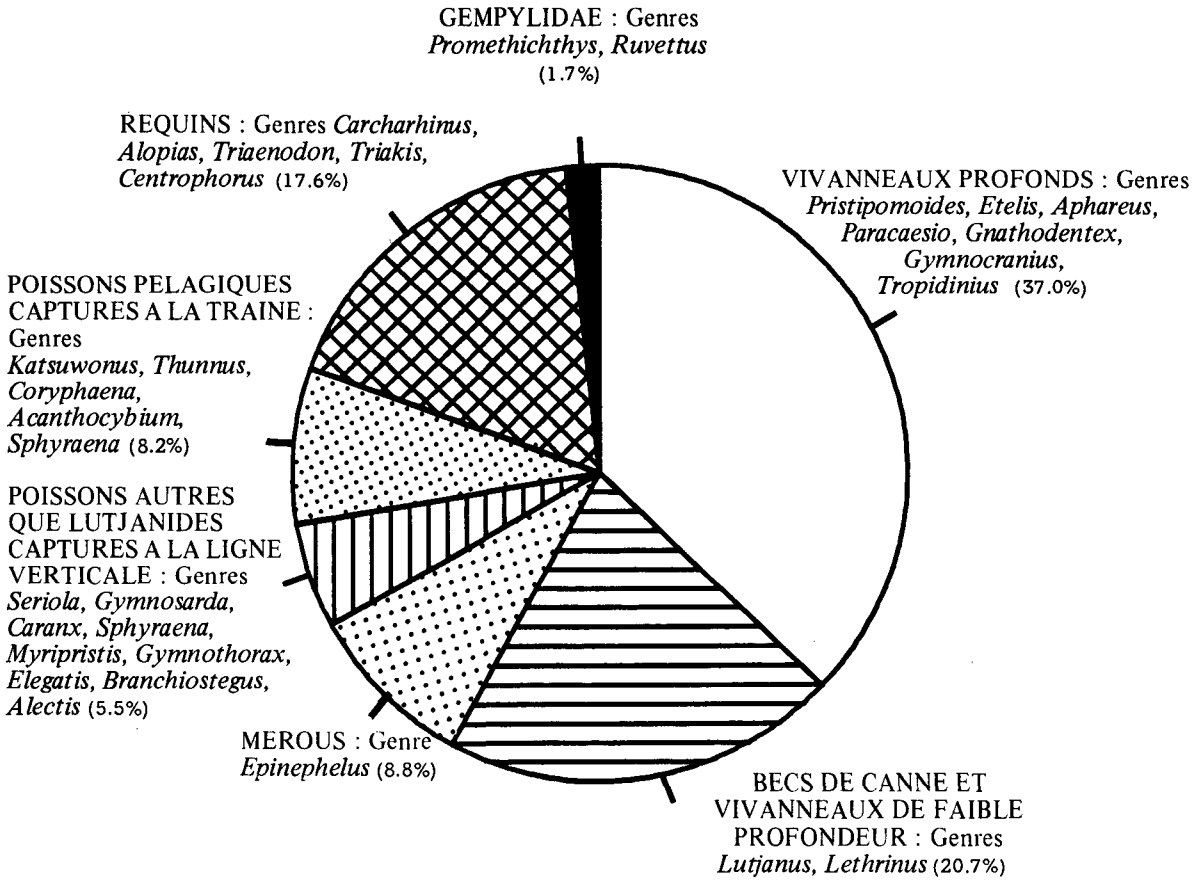
Lorsque le sens de la relation entre les variables est incertain, le choix des axes importe peu. Dans la pratique ce cas est rare, l'une des deux variables pouvant souvent être considérée comme dépendant de l'autre.

Dans un exemple qui sera présenté plus loin, on étudie la relation entre le poids et la taille des poissons. On peut avancer que, dans un tel cas, chacune des deux variables dépend de l'autre et que le sens de la relation ne peut pas être déterminé. Cependant, même dans ce cas, on adoptera une convention selon laquelle la longueur sera toujours placée en abcisse et le poids en ordonnée.

2.3.5 Graphiques à secteurs

Le graphique dit "à secteurs" consiste en un cercle divisé en un certain nombre de secteurs. Chacune des valeurs prises par le caractère étudié se voit attribuer un secteur, et la surface du secteur est proportionnelle à la part de cette valeur dans le total. On peut représenter de cette façon des caractères qualitatifs ou quantitatifs, mais le graphique à secteurs est surtout utile pour la représentation des caractères qualitatifs (figure 2.4).

FIGURE 2.4 : COMPOSITION DE LA PRISE SELON LA PROPORTION DU POIDS TOTAL QUE REPRESENTE CHAQUE ESPECE (DANS CHAQUE GROUPE, LES GENRES SONT CITES PAR ORDRE D'IMPORTANCE DECROISSANTE)



La pratique habituelle, comme sur la figure 2.4, est de commencer le graphique à partir du haut (la position "midi" sur une montre) et, en tournant dans le sens des aiguilles d'une montre, de placer en premier le secteur le plus important.

Un tel graphique à secteurs est très facile à réaliser. La surface d'un secteur est proportionnelle à l'angle intérieur formé par les deux rayons qui délimitent le secteur. Pour obtenir des secteurs dont la surface est proportionnelle à la part que représente chaque valeur du caractère dans le total, il suffit de tracer des secteurs dont les angles sont proportionnels à ces parts. La somme des angles étant égale à 360 degrés, la délimitation de chaque secteur se ramène à un calcul simple.

Le graphique à secteurs étant une représentation de proportions, on pourra le construire à partir des données brutes ou à partir des données exprimées en pourcentages. Afin de préserver la facilité de lecture de ces graphiques, en évite de diviser le cercle en un trop grand nombre de secteurs. On admet généralement un maximum de huit secteurs.

2.4 Arrondi des nombres

Au cours des chapitres suivants il nous faudra parfois arrondir des nombres à un certain nombre de décimales significatives, ou arrondir une décimale donnée à l'unité la plus proche. Lorsque l'on publie des données d'étude, il peut aussi être nécessaire d'arrondir les résultats à une tonne près, ou au millier d'unités près, etc. Afin d'assurer la cohérence de ces opérations, il faut les réaliser selon une méthode standardisée.

Le principe de base consiste à arrondir à l'unité significative la plus proche. Ainsi, le nombre 428 548 arrondi au millier près deviendra 429 000. Si le nombre à arrondir se termine exactement par une demi-unité (428 500, par exemple) la dernière unité significative du nombre arrondi sera, par convention, un chiffre pair (428 000 de préférence à 429 000, par exemple).

Lorsque l'on doit arrondir une série de nombres ainsi que leur somme, il peut apparaître que le total, une fois arrondi, ne soit plus égal à la somme des éléments arrondis. Considérons par exemple la série de nombres suivants à arrondir au millier près :

128 613	arrondi à	129 000
428 548		429 000
37 924		38 000
595 085		?

Selon la règle énoncée plus haut le total arrondi devrait être 595 000, tandis que la somme des nombres arrondis s'élève à 596 000.

Ceci pose un problème de présentation des nombres arrondis qui apparaît souvent dans la pratique. Pour le résoudre, nous adopterons la convention suivante : chaque nombre doit être arrondi selon la règle déjà énoncée (le total est donc bien 595 000, et non 596 000), de telle sorte que la somme des éléments arrondis peut ne pas égaler le total arrondi. L'inconvénient de cette méthode est que le lecteur pourrait alors en déduire qu'une erreur s'est introduite dans les calculs. Pour éviter cette méprise, les nombres sont normalement accompagnés d'une note précisant par exemple que : "La somme des éléments peut différer du total en raison des arrondis".