Chapitre 4
Statistique descriptive uni-variée

Philippe LANOS
Archéosciences-Bordeaux, Université Bordeaux-Montaigne & Géosciences-Rennes, Université de Rennes 1

Outre le résumé d'information qui permet de caractériser une série de données par quelques nombres en perdant le moins d'information possible, la statistique descriptive univariée permet de définir la place d'un individu dans la population, de comparer les distributions d'une variable sur différentes populations, de caractériser la place d'un échantillon dans une population. La statistique descriptive ouvre aussi la voie à l'identification de la loi de probabilité (les liens entre statistique et probabilité font l'objet du chapitre 7) et de ses paramètres (position, dispersion) à partir des données observées qui en sont une réalisation (c'est le domaine de l'inférence statistique qui est abordé au chapitre 8). Il s'agit aussi ici de mettre en évidence le processus qui a engendré les données et de faire ressortir les différences, les écarts, les anomalies à partir d'une description statistique rigoureuse de la population étudiée, tout en n'oubliant pas que les ensembles archéologiques étudiés posent souvent de vraies questions de représentativité (Selz, 2013) qui doivent être contrôlées par l'archéologue. Les ouvrages qui présentent l'analyse statistique descriptive uni-variée sont très nombreux. C'est pourquoi nous recommandons ici une sélection d'ouvrages qui ont un lien plus étroit avec les questions statistiques soulevées dans le domaine de l'archéologie et de l'archéométrie et qui permettront au lecteur d'approfondir les notions présentées dans ce chapitre : Baxter (2003), Bellanger & Tomassone (2014), Chenorkian (1996), Demoule *et al.* (2002), Djindjian (1991), Guerreau (2004), Massonie (1990), Rivoirard & Stoltz (2012), Saporta (2011), Vessereau (2002), Vidal (2004).

1 Notions fondamentales : population, individu, variable statistique, modalité

Ce chapitre décrit les notions fondamentales intervenant dans un problème de statistique. La statistique descriptive permet de dégager, de façon concise et simplifiée,

les informations essentielles issues d'une collection ou série de données. Dans le texte, nous indiquons entre crochets les termes équivalents ou synonymes.

1.1 Données - observations

Les **données** désignent l'ensemble des observations qui sont constituées soit de mesures physico-chimiques ou de comptage (issues d'une expérience, de relevés, d'enregistrements), soit d'observations qualitatives (issues d'une enquête, d'un dépouillement de documents...). Ces **observations** portent sur une ou plusieurs variables [caractères, attributs, descripteurs] attachées à un individu [unité statistique]. Par exemple, on mesure la masse (c'est la variable) d'une monnaie (c'est l'individu). L'ensemble des individus sur lesquels on a mesuré ou observé une ou plusieurs variables constitue la population [population statistique, ensemble statistique]. Dans l'exemple précédent, ce sera l'ensemble des monnaies pesées. La question du nettoyage d'une base de données se pose souvent. Il s'agit d'identifier, quand c'est possible, les données aberrantes, les individus atypiques et de traiter les données manquantes. Les données aberrantes, *outliers* en anglais, correspondent à des valeurs hors tendance qui peuvent par exemple résulter d'erreurs de mesure. Les données atypiques sont par définition des observations peu fréquentes, c'est-à-dire des points qui ne suivent pas la distribution caractéristique du reste des données, mais il faut noter qu'il est généralement difficile de les distinguer des données aberrantes. Les données manquantes désignent une absence d'observation sur un ou des individus particuliers. La gestion des données manquantes est une question difficile qui nécessite différents types de solutions (Little & Rubin, 2019). On peut retirer soit les variables étudiées, soit les individus concernés. Ou bien on peut imputer des valeurs aux données manquantes (par la moyenne, par tirage conditionnel, par analyse factorielle) en modifiant a minima la structure originelle globale de variance-covariance (Clavel et al., 2014). Enfin, on peut encore réaliser des analyses en présence de données manquantes.

1.2 Population - échantillon

Il est important de préciser que le terme **population** est pris au sens statistique dans l'ensemble de la seconde partie de ce manuel : il désigne un ensemble, une collection d'individus sur lesquels vont porter les observations (figure 1). Cette collection peut donc être globale (on considère tous les individus observables) ou bien il peut s'agir d'un échantillon [sous-ensemble] extrait d'une population plus large selon des critères qui doivent être définis (individus disponibles à l'observation, choix raisonné de certains individus, tirage aléatoire...). La statistique descriptive s'applique à la population tout comme à l'échantillon, c'est-à-dire qu'il s'agit dans les deux cas de travailler sur un ensemble d'individus sélectionnés.

La notion d'**échantillon** au sens d'un tirage aléatoire ou arbitraire fait appel à l'inférence statistique qui est présentée au chapitre 8 : dans ce cas, l'objectif est d'inférer une connaissance sur la population d'origine (ou de référence) à partir des caractéristiques statistiques de l'échantillon. Cela suppose que l'échantillon soit représentatif de la population, autrement dit qu'il soit possible de passer d'une partie (l'échantillon) au tout (la population de référence) au moyen d'une méthode d'échantillonnage.

Il est important de noter que la **taille de la population** [cardinal de l'ensemble statistique, effectif de la population], qui est égale au nombre des individus qui la composent, doit rester fixe tout au long de l'étude (sinon les pourcentages calculés seront faux). Par conséquent, dans toute étude statistique, 1) il faut définir exactement ce qu'est l'individu, 2) définir avec précision les critères d'appartenance des individus à la population ou à son échantillon : l'individu est donc défini en fonction de la problématique posée. Par exemple, la population sera constituée de toutes les monnaies retrouvées (échantillon) ou existantes (population) dans une unité stratigraphique particulière d'un site donné. Dans le cas d'un échantillon extrait d'une population, la taille de l'échantillon sera notée n et la taille de la population (d'origine) sera notée N quand celle-ci est connue. On gardera la notation n dans la suite de ce chapitre.

1.3 Variable statistique

Une étude statistique qui porte sur une seule **variable** est appelée statistique **univariée** [unidimensionnelle], sur deux variables, statistique **bivariée** [bidimensionnelle], et sur plus de deux variables, statistique **multivariée** [analyse des données multidimensionnelles].

Dans une population composée de n individus, chaque individu sera indexé (numéroté) par un indice k prenant des valeurs entre 1 et n. Les variables étudiées seront notées par des lettres majuscules $X, Y, Z \ldots$ Sur l'individu k, on mesure ou bien on observe la valeur prise par la variable X : cette valeur sera notée x_k (lire : x indice k). Les valeurs $x_1, x_2, \ldots, x_n$ prises par la variable X constitue la **série de données**, ou **série statistique**, présentée généralement sous la forme d'un tableau de données. Si on observe plusieurs variables X, Y, Z, les valeurs pour l'individu 1 seront notées : x_1, y_1, z_1, et ainsi de suite. L'usage veut que chaque individu soit présenté sur une ligne différente du tableau de données, et que chaque variable occupe une colonne.

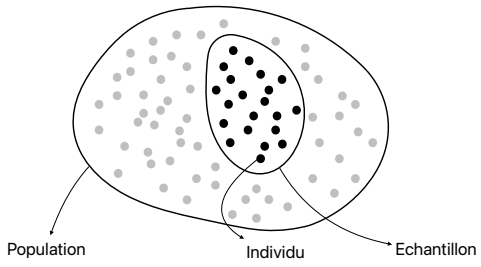


FIG. 1 : Dans une étude statistique, il faut définir la population (ensemble statistique de taille connue N ou bien de taille inconnue), l'échantillon (sous-ensemble statistique de taille n), les individus (unité statistique). En fonction de la problématique archéologique, l'échantillon sera la population elle-même ou bien sera un extrait de la population (collection d'individus disponibles à l'observation, ou choix raisonné ou encore tirage aléatoire). Sur chaque individu, on pourra observer ou mesurer un ou plusieurs caractères (variable statistique) dont il faut définir la nature qualitative, quantitative discrète ou bien quantitative continue. Les données sont les valeurs observées ou mesurées sur les individus, pour chaque caractère.

La valeur x_k prise par une variable X pour un individu k dépend de la nature ou du type de la variable étudiée :

Une variable qui ne résulte pas d'une mesure mais qui est décrite par une observation prenant des **modalités** [catégories], exclusives les unes des autres, est dite **variable qualitative** [variable catégorielle, facteur]. Par exemple, on décrira la couleur d'un objet par plusieurs modalités : rouge, orange, jaune, vert, bleu… On parle dans ce cas de **variable qualitative nominale** car on ne peut pas définir d'ordre entre les modalités. Ces modalités peuvent aussi s'exprimer par un codage comme 1, 2, 3. Si les modalités correspondent à des critères d'appréciation qui peuvent être ordonnés, on a affaire à une **variable qualitative ordinale**. Mais attention, un calcul de moyenne sur ces valeurs de code n'a aucun sens. Les modalités ou le codage nécessitent la définition d'une **nomenclature** ou d'un catalogue de référence par rapport auxquels les individus seront comparés. Par exemple, un tesson de sigillée sera rattaché à tel type de production grâce à un catalogue. Une couleur sera attribuée en fonction d'un nuancier. Si la couleur d'un individu observé ne correspond à aucune des modalités définies au préalable, il faudra ajouter une modalité « indéterminée » de telle sorte que tous les individus de la population étudiée pourront entrer dans une et une seule modalité : on parle de **disjonction** et d'**exhaustivité** des modalités, ou encore de tableau de données complet (Massonie, 1990). Enfin, si les modalités prennent les valeurs 0 ou 1 (vrai ou faux ; présence ou absence), on a affaire à une **variable disjonctive** [dichotomique, binaire, booléenne].

Une variable qui ne prend que des valeurs numériques entières ou isolées est dite **variable discrète**. Ces valeurs résultent souvent d'un processus de comptage. C'est par exemple le cas du nombre d'enfant par famille : ici, l'unité statistique (l'individu) est la famille, et la variable est le nombre d'enfants qui prend les valeurs $0, 1, 2$ ou plus. Dans ce cas, la taille de la population (au sens statistique) correspondra au nombre de familles.

Une variable qui prend des valeurs numériques sur l'ensemble des nombres réels, ou sur un sous-ensemble des réels, est dite **variable continue** [numérique, métrique, scalée]. Les valeurs sont fournies par des appareils de mesure (par exemple une balance pour mesurer une masse). Elles sont exprimées avec un certain nombre de chiffres après la virgule, qui est déterminé par la précision des appareils de mesure : ce sont les chiffres significatifs après la virgule. À noter que des zéros terminaux peuvent être significatifs. Si une mesure de masse est faite au millième de gramme près, alors le zéro de $x = 13,530$ g est significatif. Notons enfin qu'une variable discrète peut être (et est souvent) traitée comme une variable continue lorsque le nombre de valeurs mesurées devient très grand.

Le traitement statistique peut intervenir après une transformation mathématique des variables continues, appelées alors données « brutes » (avant transformation), observées sur les individus. Une **transformation de variable** peut être utile dans plusieurs cas :

- pour mettre en évidence de façon plus claire une relation entre deux variables statistiques X et Y, en considérant par exemple le logarithme des valeurs des deux variables, soit $u_k = \ln(x_k)$ et $v_k = \ln(y_k)$.
- pour centrer, normaliser ou dé-corréler des variables. Par exemple, dans le domaine de l'analyse des compositions élémentaires (*cf.* 2.1.5), on peut effectuer une transformation log-ratio centrée (clr : centered log-ratio) des valeurs de composition x_k : $\mathrm{clr}(x_k) = \ln\left(\frac{x_k}{g(x)}\right)$ avec $g(x) = \sqrt[n]{x_1 \cdot x_2 \ldots x_n}$. La fonction ln correspond au logarithme (népérien) et $g(x)$ est une moyenne géométrique (*cf.* infra).
- pour effectuer un changement de repère en analyse factorielle : les axes factoriels définissent de nouvelles variables qui sont des combinaisons linéaires des variables observées (*cf.* 6).

1.4 Le traitement statistique

Pour résumer : la nature des variables étudiées est fondamentale en statistique car elle détermine en grande partie le type des traitements mathématiques qui pourront être effectués sur les données. Par exemple, on ne pourra pas calculer une moyenne sur une variable qualitative nominale ou ordinale. Ainsi, dans tout problème statistique, il faut d'abord identifier la population et les individus qui la composent, ensuite les variables et leur nature, puis définir les unités et la précision des mesures dans le cas des variables quantitatives, ou bien le codage et la nomenclature de référence dans le cas des variables qualitatives. Enfin, il faut définir les classes de valeurs dans le cas des variables quantitatives, les modalités dans le cas des variables qualitatives. Les mesures et les observations constituent le tableau de données, la série statistique. Le dépouillement de ce tableau, par tri à plat dans le cas d'une seule variable, tri croisé dans le cas de deux variables, permet de construire la distribution statistique des effectifs et des fréquences. Cette distribution peut être résumée sous la forme de graphiques. Enfin, pour synthétiser l'information, en particulier dans le cas des variables quantitatives, on pourra calculer des paramètres résumant la série, par exemple des paramètres de position [tendance centrale] et de dispersion.

2 Tableau de données, distribution statistique et graphiques

2.1 Variables qualitatives

Le **tableau de données** dépend du codage utilisé. En tableau 1, il comprend n lignes correspondant aux n individus et deux colonnes : la première indique le numéro de l'individu, la seconde la valeur de la modalité observée sur l'individu. En tableau 2, nous avons la même information décomposée selon un **tableau disjonctif** avec trois colonnes R, B, V composées des valeurs binaires 0 et 1 . La somme en colonne correspond au nombre d'individus ayant la modalité concernée. La somme en ligne est bien sûr toujours égale à 1 (un individu).

Tab. 1 : Tableau de données

individu	Couleur
1	R
2	V
3	V
4	B
5	R
6	V
7	V
8	V
9	B
10	R
11	V

Tab. 2 : Tableau disjonctif des données

individu	R	V	B
1	1	0	0
2	0	1	0
3	0	1	0
4	0	0	1
5	1	0	0
6	0	1	0
7	0	1	0
8	0	1	0
9	0	0	1
10	1	0	0
11	0	1	0

Le dépouillement du tableau de données par tri à plat (à la main ou par informatique) permet d'obtenir le **tableau de fréquences** [distribution statistique, distribution empirique, distribution de fréquences] qui associe à chaque modalité i les effectifs n_i et les fréquences f_i (tableau 3). L'**effectif** est le nombre d'individus par modalité. On parle aussi de fréquence absolue. La **fréquence** (ou fréquence relative), est la proportion des individus par modalité, nombre compris entre 0 et 1. Mais on l'exprime plus souvent en **pourcentage**, en gardant par exemple un ou deux chiffres après la virgule, par la formule :

$$f_i(\%) = \frac{n_i}{n} \times 100$$

Ainsi, la fréquence f_i peut s'exprimer en pourcent entre 0 et 100% ou bien en **proportion** entre 0 et 1 (soit le pourcentage divisé par 100). Mais attention, dans tous les calculs, il faut bien considérer f_i au sens de la proportion.

Dans un tableau de fréquences, l'indice i représente l'index de la modalité. Ainsi, n_1 représente l'effectif de la modalité R, c'est-à-dire le nombre d'individus ayant la couleur rouge.

Tab. 3 : Tableau de fréquences d'une variable qualitative

Variable couleur	Effectif n_i	Fréquence $f_i(\%)$
R	3	27,27
V	6	54,55
B	2	18,18
n =	**11**	**100,00**

Ces résultats sont plus faciles à analyser et à comparer avec ceux d'autres populations grâce à différents graphiques : diagramme en barres, diagramme circulaire, diagramme en radar. Dans le **diagramme en barres** [bande, tuyaux d'orgue] des fréquences, chaque modalité est représentée par une barre séparée des autres (pour éviter toute confusion avec l'histogramme), de largeur arbitraire (choix esthétique) et dont la hauteur est proportionnelle à la fréquence. A noter qu'il ne s'agit absolument pas d'un histogramme (erreur courante à ne pas commettre) et en conséquence les barres ne doivent pas être reliées par des droites ou des courbes. Dans le **diagramme circulaire** [autrement appelé « camembert »], chaque modalité est représentée par un secteur dont l'angle est proportionnel à la fréquence. Dans les deux cas, on veut jouer sur le fait que l'œil compare facilement des surfaces. La lecture de ces graphiques est améliorée par l'utilisation de couleurs et/ou de trames. Si on veut comparer la « structure » d'une variable qualitative, autrement dit la répartition en fréquence de ses modalités, sur deux populations A et B de tailles différentes n_A et n_B, il est possible de dessiner des diagrammes circulaires dont le rapport des surfaces est égal au rapport des tailles (figure 2). Ceci implique que le rapport des rayons des cercles est égal à la racine carrée du rapport des tailles : $\frac{r_A}{r_B} = \sqrt{\frac{n_A}{n_B}}$. Ainsi l'œil pourra comparer la répartition des fréquences (surfaces des secteurs) tout en pondérant en fonction des tailles des populations.

Dans le **diagramme en radar** [diagramme en araignée, en étoile] (figure 3), les fréquences associées aux modalités sont portées sur des axes radiaux régulièrement répartis sur le cercle. Une distribution statistique pour une population donnée sera représentée par un polygone rejoignant les fréquences (dont la somme est égale à 100%). Ce graphique permet de comparer efficacement la structure de plusieurs distributions en même temps.

2.2 Variables quantitatives discrètes

Nous prenons ici l'exemple d'une variable quantitative discrète X correspondant au nombre d'anses observées sur un ensemble de céramiques (des vases). La population est définie par un ensemble de 10 vases provenant d'un niveau stratigraphique dans un site. L'individu étudié est le vase. La variable étudiée est le nombre d'anses observées

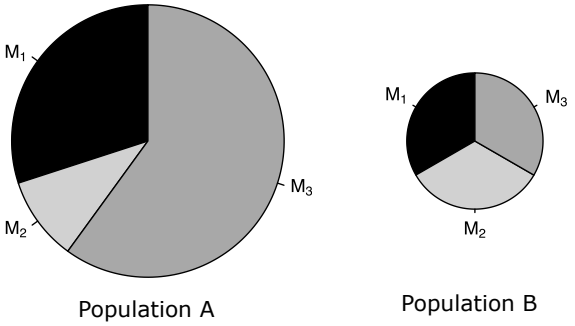


FIG. 2 : Diagrammes circulaires établis pour deux populations A et B. Le rapport des surfaces est égal au rapport des tailles n_A et n_B, ici choisi égal à 9 (si la population A contient 900 individus, la population B en contient alors 100). Pour respecter ce rapport, le rayon de A doit être égal à 3 fois le rayon de B. La variable qualitative étudiée prend trois modalités dont les fréquences sont représentées, pour chaque diagramme, par la surface des secteurs associés. Par exemple, dans la population A, la fréquence f_1 (en %) de la modalité M1 se calcule graphiquement comme l'angle d'ouverture du secteur concerné divisé par 360°, soit $f_1 = (108°/360°) \times 100 = 30\%$. Avec ces deux diagrammes, nous pouvons comparer la structure (distribution en fréquences) de la variable étudiée dans les deux populations tout en ayant la vision de la taille (de la masse) de chacune des populations.

sur un vase. Il s'agit d'une mesure de comptage exprimée sous la forme d'un nombre entier. Pour le vase k, la valeur observée est notée $\mathrm{O_k}$. Autrement dit, les 10 individus "vase" sont numérotés par l'indice k, k allant de 1 à n (ici, $n = 10$).

Le dépouillement du tableau de données (tableau 4), qui peut se faire comme précédemment via un tableau disjonctif, permet de construire le tableau des fréquences (tableau 5) qui s'enrichit ici de deux colonnes supplémentaires contenant la fréquence cumulée F_i et le produit $x_i f_i$ destiné au calcul de la moyenne. La première colonne donne les différentes valeurs possibles de la variable X : $\mathrm{x_1} = 0, \mathrm{x_2} = 1, \mathrm{x_3} = 2, \mathrm{x_4} = 3$. Ici, l'indice i parcourt les différentes valeurs discrètes (isolées) prises par la variable X. La deuxième colonne fournit l'effectif n_i qui correspond au nombre d'individus ayant un nombre d'anses donné x_i.

La fréquence f_i, exprimée en pourcentage (%), est calculée de la même manière que pour la variable qualitative. La fréquence cumulée est alors définie par :

$F_1 = f_1 = 10\% \quad F_3 = f_1 + f_2 + f_3 = 60\%$

$F_2 = f_1 + f_2 = 40\% \quad F_4 = f_1 + f_2 + f_3 + f_4 = 100\%$

Symboliquement, on écrira : $F_i = \sum_{q=1}^{i} f_q$, le signe Σ voulant dire « somme ». On dira « somme de $q = 1$ à i des fréquences simples f indice q ». La fréquence cumulée F_i indique la proportion de la population de vases ayant un nombre d'anses inférieur ou égale à x_i. Par exemple pour calcule F_2, il y a 60% des vases ayant 2 anses ou moins, c'est-à-dire ayant 0,1 ou 2 anses. Ici la somme de ces fréquences est égale à $F_2 = 60\%$.

La dernière colonne contient le produit (multiplication) de la valeur x_i par la proportion (fréquence en pourcent divisée par 100) $f_i = \frac{n_i}{n}$. La somme de ces produits donne

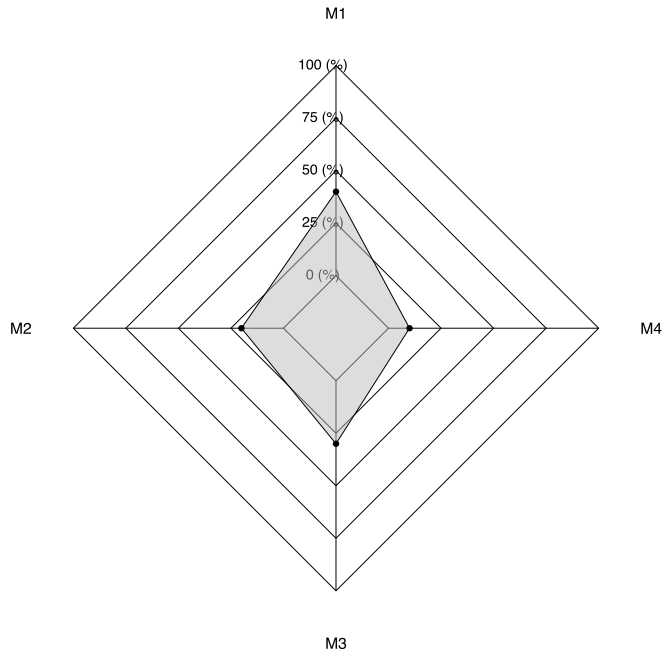


FIG. 3 : Diagramme en radar (ou araignée) : les sommets du polygone représentent les modalités de la variable qualitative étudiée : ici on a quatre modalités notées M1, M2, M3, M4. Chaque axe est gradué en % de 0 à 100. La distribution en fréquence de la variable sera représentée par un polygone dont la somme des coordonnées sur les axes est égale à 100% (zone grisée). Il est ainsi possible de superposer des polygones de fréquences afin de comparer la structure de la variable étudiée sur plusieurs populations.

TAB. 4 : Tableau de données d'une variable discrète

Individu k	Nombre anses O_k
1	3
2	3
3	0
4	1
5	1
6	2
7	3
8	3
9	1
10	2

directement la moyenne empirique pondérée (c'est-à-dire observée) $\bar{x}$ définie par les expressions équivalentes suivantes :

Tab. 5 : Tableau de fréquences construit à partir du tableau de données 4

Variable nbre anses x_i	Effectif n_i	Fréquence $f_i(\%)$	Fréquence Cumulée F_i	$f_i x_i$ /100
0	1	10,0	10,0	0,00
1	3	30,0	40,0	0,30
2	2	20,0	60,0	0,40
3	4	40,0	100,0	1,20
n =	**10**	**100, 00**		**1, 90**

$$\bar{x} = \frac{n_1x_1 + n_2x_2 + n_3x_3 + n_4x_4}{n} = f_1x_1 + f_2x_2 + f_3x_3 + f_4x_4 = \sum_{i=1}^{4} f_ix_i = 1,9$$

Le nombre moyen d'anses par vase est de 1,9. Ce paramètre statistique est un nombre abstrait car le nombre réel d'anses par pot est forcément entier ! Mais c'est un indicateur d'une information de tendance centrale pour la variable étudiée. Nous sommes en fait habitués à manipuler ce type de nombre abstrait sans que cela ne pose véritablement problème, comme quoi notre cerveau est capable d'abstraction !

Deux types de graphiques sont utilisés pour représenter le tableau des fréquences d'une variable quantitative discrète : le diagramme en bâton et la courbe cumulative. Dans ces graphiques, l'axe horizontal, appelé axe des abscisses, représente les valeurs x_i de la variable en des points discrets (isolés). L'axe vertical, appelé axe des ordonnées, représente l'effectif, la fréquence ou encore la fréquence cumulée.

Dans le **diagramme en bâtons**, à chaque valeur x_i on fait correspondre un segment vertical de longueur proportionnelle à l'effectif ou bien à la fréquence de cette valeur. Pour passer de l'effectif à la fréquence, il faut effectuer un changement d'échelle sur l'axe des ordonnées. Pour tracer le diagramme en bâtons des fréquences simples du tableau 5 , on détermine la fréquence maximale observée qui est $f_4 = 40\%$. On lui assigne un segment de longueur $l_4 = 5$ cm par exemple. Par conséquent, la longueur l_2 du segment correspondant à la valeur x_2 sera donnée par la règle de trois :

$$l_2 = f_2 \times \frac{l_4}{f_4} = f_2 \times \frac{5 \text{ cm}}{40\%} = 0,125 f_2 = 3,75 \text{ cm}$$

D'où le diagramme en bâtons en figure 4.

La **courbe cumulative** [courbe cumulée, fonction de répartition] est un graphique en escalier dont les paliers horizontaux ont pour ordonnées (position sur l'axe vertical) les fréquences cumulées F_i. La différence entre deux paliers F_{i-1} et F_i correspond donc à un saut d'amplitude égale à la fréquence simple f_i (figure 5).

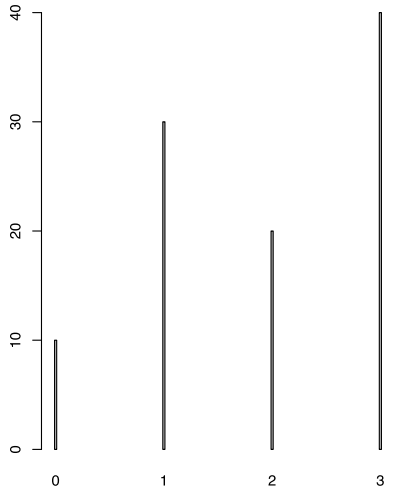


FIG. 4 : Diagramme en bâtons des fréquences : à chaque valeur de la variable quantitative discrète xi , on fait correspondre un segment vertical dont la longueur est proportionnelle à la fréquence fi (ici en %) (colonne 3 du tableau 5). La hauteur des segments permet ainsi de visualiser rapidement la structure (distribution en fréquences) de la variable étudiée.

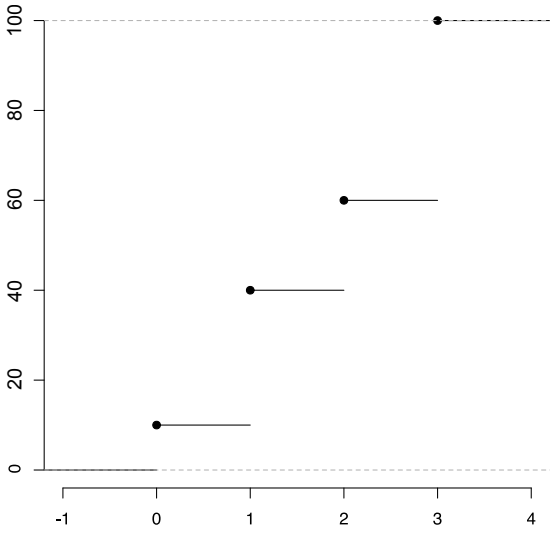


FIG. 5 : Courbe cumulative pour une variable quantitative discrète : fonction en escalier dont les paliers horizontaux ont pour ordonnées les fréquences cumulées Fi (colonne 4 du tableau 5). L'axe des ordonnées (axe vertical) est ici gradué en proportion entre 0 et 1, ce qui correspond à des fréquences entre 0 et 100 %. Le saut entre chaque palier est égal à la fréquence simple f_i.

2.3 Variables quantitatives continues

On prendra comme exemple de variable continue le volume (mesuré en litres) de conteneurs en céramique (par exemple des amphores). Les valeurs sont prises sur l'ensemble $\mathbb{R}$ des réels et sont arrondies à une décimale après la virgule, soit une mesure ayant une précision au dixième de litre près. Le tableau de données (tableau 6) indique la valeur du volume mesuré sur chacun des 60 conteneurs étudiés.

Tab. 6 : Tableau de données d'une variable quantitative continue

Individu k	Volume V_k (litres)
1	77,1
2	30,9
...	...
...	...
59	46,0
60	55,4

Ce tableau peut aussi se présenter sous la forme d'une simple liste (série statistique) de 60 valeurs V_k présentées sur une ou plusieurs lignes (tableau 7).

Tab. 7 : série statistique correspondant au tableau 6.

77.1	30.9	85.0	36.3	30.9	31.5	34.3	34.2	51.8	48.9	61.7	49.4	27.6	35.9	33.6	50.9	40.0	38.7	45.4	38.2
64.0	50.9	62.6	17.4	22.2	18.6	22.5	31.7	39.9	56.0	23.1	19.6	25.0	33.6	24.7	45.3	53.7	20.6	40.9	42.1
21.2	16.3	38.1	73.3	39.0	25.2	50.0	28.7	33.9	31.3	58.7	17.3	58.5	28.5	44.6	46.1	47.7	38.0	46.0	55.4

Deux types d'analyse peuvent être menés pour décrire et synthétiser ces données : soit par un regroupement des valeurs observées V_k selon des classes (ce qui amène à construire un histogramme), soit par des calculs directs opérés sur les valeurs individuelles V_k elles-mêmes, sans regroupement (ce qui amène à construire des box-plots).

Pour construire le tableau de fréquences, on regroupe les volumes observés par classes de valeurs. Par convention, une **classe** est un intervalle fermé à gauche et ouvert à droite du type : $[b_i, b_{i+1}\ [$. L'effectif n_i de la classe i est le nombre des individus dont le caractère, la variable, prend une valeur supérieure ou égale à la borne b_i et inférieure strictement à la borne supérieure b_{i+1}. Le **centre d'une classe** bornée est $c_i = \frac{(b_i+b_{i+1})}{2}$ (moyenne des deux bornes). L'**amplitude d'une classe** est $a_i = b_{i+1} - b_i$ (*cf.* colonne 5 en tableau 8). Elle intervient dans la construction de l'histogramme. On utilise en général des classes d'amplitudes identiques. Le choix du nombre n_C de classes est déterminé en fonction du nombre n de valeurs observées. Une formule souvent utilisée propose de prendre : $n_C = \sqrt{n}$ ou bien $n_C = 2n^{\frac{1}{3}}$ (règle de Rice). Mais il est possible de faire varier ce nombre pour trouver le meilleur compromis entre lissage (peu de classes) et ajustement (beaucoup de classes) de l'histogramme aux données.

En dernière colonne du tableau de fréquences, nous ajoutons la densité de fréquence d_i qui est égale au rapport entre la fréquence et l'amplitude de classe : $d_i = \frac{f_i}{a_i}$. Cette densité est d'autant plus grande que l'effectif dans la classe est grand et que la largeur de la classe est petite. Elle sert à construire l'histogramme.

TAB. 8 : Tableau de fréquences d'une variable quantitative établi à partir des valeurs du tableau 7.

Classe i de centre c_i	Effectif n_i	Fréquence $f_i(\%)$	Fréquence Cumulée F_i	Amplitude a_i	Densité $d_i = f_i/a_i$
$[0, 10[$	0	0,00	0,00	10	0,000
$[10, 20[$	5	8,33	8,33	10	0,833
$[20, 30[$	11	18,33	26,67	10	1,833
$[30, 40[$	18	30,00	56,67	10	3,000
$[40, 50[$	11	18,33	75,00	10	1,833
$[50, 60[$	9	15,00	90,00	10	1,500
$[60, 70[$	3	5,00	95,00	10	0,500
$[70, 80[$	2	3,33	98,33	10	0,333
$[80, 90[$	1	1,67	100,00	10	0,167
$[90, 100[$	0	0,00	100,00	10	0,000
	60	**100, 00**			

Il est important de noter que le regroupement des valeurs x_i en classe revient à transformer la variable quantitative continue X en variable qualitative, la classe définissant alors une modalité. Cette opération est couramment effectuée en analyse des données (*cf.* 5 & 6) pour analyser en même temps variables qualitatives et variables quantitatives observées sur les individus d'une population.

Deux types de graphiques sont utilisés pour représenter le tableau de fréquences d'une variable quantitative continue : l'histogramme et la courbe des fréquences cumulées.

L'**histogramme** permet de visualiser la répartition, la distribution des fréquences d'une variable continue - on parlera ici de l'histogramme des fréquences, qui est identique à celui des effectifs à un facteur d'échelle près sur l'axe des ordonnées. A chaque classe, on fait correspondre un rectangle basé sur cette classe et dont la hauteur est proportionnelle à la densité d_i, autrement dit dont la surface est proportionnelle à la fréquence de la classe (figure 6). En anticipant un peu, si on estime la probabilité que la variable X prenne une valeur x_i dans la classe i par la proportion $f_i = \frac{n_i}{n}$, cette probabilité sera représentée par la surface du rectangle (et non pas par la hauteur qui, elle, correspond à la densité de fréquence). Ce point est essentiel pour bien interpréter les distributions de probabilité, par exemple celles fournies sur le temps par les méthodes de datation (*cf.* 12). Si les amplitudes de classe sont inégales (par exemple par regroupement de certaines classes), les surfaces des rectangles devront toujours rester proportionnelles aux fréquences de telle sorte que la surface totale de l'histogramme reste toujours égale à 1, c'est-à-dire 100%. L'histogramme apparaît comme une succession de rectangles contiguës dont la hauteur varie en fonction de la densité.

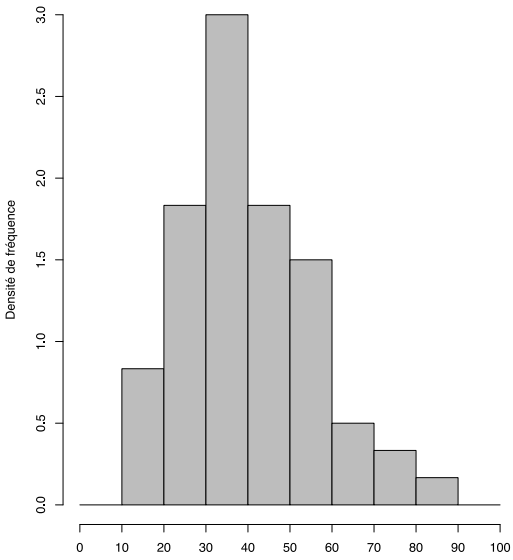


FIG. 6 : Histogramme : graphe obtenu à partir des colonnes 1 et 6 du tableau 8. La variable quantitative continue (volume en litres) est représentée sur l'axe horizontal entre 0 et 100, par classe de 10 litres. Chaque rectangle grisé est de surface proportionnelle à la fréquence simple f_i (colonne 3 du tableau 8). La hauteur du rectangle, appelée densité de fréquence, est donc proportionnelle à la fréquence simple f_i divisée par la largeur de classe a_i. Il faut bien distinguer la densité de fréquence (représentée par la hauteur d'un rectangle) de la fréquence elle-même (représentée par la surface de ce rectangle).

La **courbe des fréquences cumulées** [courbe cumulative, fonction de répartition] $F(x)$ est égale à la proportion, à la fréquence des mesures x_i inférieures à la valeur x (figure 7). Pour sa construction, l'ordonnée $F(x)$ de la courbe cumulative correspond à la surface de l'histogramme située « à gauche » de x, c'est-à-dire pour des valeurs inférieures à x. C'est une courbe monotone non décroissante (donc avec possibilité de plateaux) composée de segments de droite dans chaque classe. Ceci résulte du fait qu'on ne localise plus les valeurs individuelles x_i au sein d'une classe i (on voit ici qu'on perd un peu d'information dans le passage du tableau de données au tableau de fréquences - mais c'est pour y voir plus clair !). On suppose alors que ces valeurs x_i sont uniformément distribuées dans la classe, avec pour moyenne le centre de classe. Cette distribution uniforme est précisément représentée par la fonction rectangle, c'est-à-dire par une densité constante sur l'intervalle de classe. Sous cette hypothèse, la surface du rectangle entre la borne b_i et x croît linéairement avec x pour atteindre la valeur f_i à la borne b_{i+1}. Par conséquent, à partir de la courbe cumulative, il est facile de déduire la valeur de x qui partage l'effectif de la population en deux parties égales : il faut prendre la valeur de x telle que $F(x) = 50\%$. Cette valeur de la variable s'appelle la médiane (voir infra).

Notons qu'il y a une correspondance entre l'histogramme (en statistique descriptive), la densité de probabilité (en théorie des probabilités) et la notion de dérivée (en

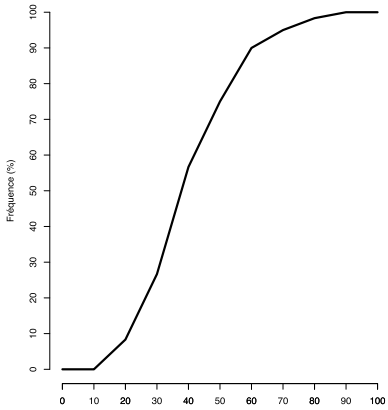


FIG. 7 : Courbe des fréquences cumulées : graphe obtenu à partir des colonnes 1 et 4 du tableau 8. C'est une courbe monotone non décroissante constituée de segments de droite dans chaque classe. Pour une valeur particulière de volume x, cette courbe donne la fréquence (cumulée) des amphores dont le volume est inférieur ou égal à x. Inversement, si la fréquence cumulée est fixée à 50 %, alors la valeur x correspondante est appelée la médiane (voir section 3.1) : c'est la valeur de la variable qui partage la population étudiée en deux effectifs égaux (50 %, 50 %).

analyse des fonctions). De même qu'il y a une correspondance entre la courbe des fréquences cumulées, la fonction de répartition et la notion d'intégrale. Ainsi, la dérivée de la courbe cumulée est égale à l'histogramme (qui est une fonction continue par morceaux - c'est-à-dire par classe). Inversement, l'intégrale de l'histogramme (calcul de l'aire) entre $-\infty$ et x redonne la courbe cumulée $F(x)$. La même information est contenue dans l'histogramme et la courbe cumulative, le passage de l'un à l'autre se faisant par dérivation-intégration. Toutefois, l'habitude veut qu'il semble plus facile d'interpréter un histogramme qu'une courbe cumulative. De fait, il est plus facile de percevoir visuellement des changements de surface sur un histogramme que des variations de pente sur la courbe cumulée correspondante.

Un exemple typique de courbe cumulative est la courbe granulométrique qui est le plus souvent un diagramme dit « semi-log » où les dimensions des grains sont représentées sur l'axe des abscisses en échelle logarithmique de base 10 (en général en μm) et où les fréquences cumulées de 0 à 100% sont placées sur l'axe des ordonnées, en échelle arithmétique. Il n'est pas lisible de superposer des histogrammes pour comparer la structure de populations différentes. En revanche, la superposition de courbes cumulatives reste souvent assez claire s'il n'y a pas trop de courbes et si l'on utilise des traits et des couleurs pour les différencier. On peut ainsi facilement visualiser et comparer la composition en taille de grains de différents échantillons (populations) de sédiment.

3 Paramètres statistiques

Dans le cas de variables quantitatives, parallèlement au traitement graphique, nous pouvons résumer l'information par des paramètres statistiques [indicateurs statistiques] qui caractérisent soit une tendance centrale, c'est-à-dire une position sur l'échelle des valeurs prises par X (mode, médiane, moyenne), soit une dispersion des valeurs (étendue, quantiles, intervalle, variance), soit encore une forme de distribution de ces valeurs (dissymétrie, aplatissement). La connaissance de ces paramètres synthétiques permet de comparer rapidement différentes populations. Ils peuvent aussi servir d'estimateur de paramètres de lois de probabilité pour lesquelles les valeurs observées seraient des réalisations. Ce dernier point renvoie à ce qu'on appelle la modélisation et l'inférence statistique.

3.1 Les paramètres de tendance centrale [de position]

Le premier de ces paramètres est le **mode** [valeur dominante], noté m : c'est la valeur de la variable qui correspond à la « densité de fréquence » maximale. C'est le bâton le plus élevé (le pic) ou bien le rectangle le plus élevé dans l'histogramme. Dans ce dernier cas, on parle de classe modale. Attention, ce n'est pas la fréquence la plus élevée qui, elle, mesure une surface, autrement dit une probabilité. Une distribution de fréquence présentant un seul mode est dite **monomodale** [unimodale]. Quand la distribution présente deux maximums, on dit qu'elle est **bimodale**. S'il y a plus de deux modes, la distribution est dite **multimodale**. Un mode correspond à un point de pente maximale (point d'inflexion) sur la courbe cumulative. Il faut noter que la valeur du mode est très sensible au regroupement des données en classes.

La **médiane**, notée M, est la valeur de la variable qui partage l'effectif en deux parties égales $(50\%, 50\%)$, les valeurs observées x_i du tableau de données étant rangées, triées dans un ordre croissant (ou décroissant). Si le nombre n de valeurs est impair, la médiane existe, sinon il faut faire une interpolation entre les deux valeurs encadrantes. Si la variable est quantitative, on peut aussi déduire la médiane de la courbe cumulative en cherchant x tel que $F(x) = 50\%$, soit graphiquement, soit par interpolation linéaire (règle de trois).

Enfin, la **moyenne** est un nombre dont le calcul utilise l'ensemble des valeurs observées de la variable, que l'on pondère par les effectifs et qu'on qualifie souvent de **moyenne empirique pondérée**. C'est souvent une valeur abstraite qui n'existe pas toujours dans les données. Il existe différentes formules de moyenne pondérée qui sont utilisées en fonction du problème traité.

Rappelons les notations : n représente le nombre total des individus étudiés. L'indice i désigne la modalité x_i. Le nombre de modalités est noté p. La modalité peut être une valeur isolée x_i ou bien un centre de classe c_i (dans ce cas, on remplace x_i par c_i dans les formules suivantes).

La **moyenne arithmétique** $\bar{x}$: $\quad \bar{x} = \frac{1}{n}\sum_{i=1}^{p} n_i x_i = \sum_{i=1}^{p} f_i x_i$

Il faut lire : la moyenne arithmétique pondérée x barre est égale à la somme de i égal 1 à p des produits de la fréquence f indice i par la valeur x indice i. On peut simplifier en écrivant $\sum_i \ldots$ en omettant p comme dans les trois formules qui suivent.

La **moyenne quadratique** Q : $\quad Q^2 = \frac{1}{n}\sum_i n_i x_i^2 = \sum_i f_i x_i^2$

La **moyenne harmonique** H telle que : $\quad \frac{1}{H} = \frac{1}{n}\sum_i n_i \frac{1}{x_i} = \sum_i f_i \frac{1}{x_i}$

La **moyenne géométrique** G telle que : $\log(G) = \frac{1}{n}\sum_i n_i \log(x_i) = \sum_i f_i \log(x_i)$

où $n = n_1 + n_2 + \cdots + n_p = \sum_{i=1}^{p} n_i$

La moyenne arithmétique pondérée $\bar{x}$ est le plus souvent utilisée car elle est liée à la notion de centre de gravité, de barycentre. La fréquence $f_i = \frac{n_i}{n}$ fonctionne alors comme un **coefficient de pondération**, comme une masse placée à la coordonnée x_i. Les masses n_i sont réduites (normalisées) par la masse totale n de telle sorte que la somme des pondérations f_i soit égale à 1 . Si la valeur x_i augmente par rapport aux autres valeurs, la masse (réduite) se déplace et donc la moyenne $\bar{x}$ se déplace aussi (on peut prendre comme image une balance à fléau où le point de suspension de la balance donne la position de la moyenne). La moyenne arithmétique pondérée est donc très sensible aux valeurs extrêmes, hors tendance, appelées aussi *outliers* en anglais, à la différence du mode ou de la médiane qui sont calculés à partir des effectifs. Toutefois, cette sensibilité diminue lorsque le nombre n de valeurs augmente.

La somme algébrique des écarts d'une série de valeurs x_i à la moyenne $\bar{x}$ est nulle. De plus, la somme des carrés des écarts d'une série de valeurs x_i à une valeur arbitraire x_0 est minimale si cette valeur x_0 est égale à la moyenne $\bar{x}$. Autrement dit, la moyenne est la valeur de la variable qui minimise la variance (*cf.* infra). Ainsi, la moyenne arithmétique est l'abscisse du centre de gravité de l'aire de l'histogramme des fréquences.

Si les valeurs x_i sont considérées individuellement, la fréquence, ou pondération, de chaque valeur sera égale à $\frac{1}{n}$. Si les valeurs sont regroupées en p classes, la moyenne arithmétique sera calculée à partir des centres de classe par la formule : $\bar{x} = \frac{1}{n}\sum_i^p n_i c_i$. Ce calcul de moyenne sera moins exact que celui effectué directement avec les valeurs individuelles, ce d'autant plus que l'amplitude des classes sera grande (effet du regroupement en classes). Dans l'exemple de la série du tableau 7, la moyenne individuelle est de 39,91 alors que la moyenne par les centres de classe donne 39,83. La différence reste donc assez faible après regroupement en classe de largeur 10 litres.

La moyenne quadratique est utilisée par exemple dans le calcul de la variance (*cf.* infra).

La moyenne harmonique est utilisée par exemple quand on combine des erreurs sur les âges ^{14}C. Si on effectue deux datations radiocarbone sur un même objet, on obtiendra des âges $A_1 \pm \sigma_{A1}$ et $A_2 \pm \sigma_{A2}$ différents du fait des erreurs expérimentales de mesure. S'agissant d'un même objet ayant un seul âge, on montre que l'âge moyen A sera donné par la moyenne arithmétique pondérée :

$$A = \frac{\frac{A_1}{\sigma_{A1}^2} + \frac{A_2}{\sigma_{A2}^2}}{\left(\frac{1}{\sigma_{A1}^2} + \frac{1}{\sigma_{A2}^2}\right)}$$

et que l'erreur σ_A sur cet âge moyen sera la racine carrée de la moyenne harmonique des variances (*cf.* supra) :

$$\sigma_A = \sqrt{\frac{1}{\left(\frac{1}{\sigma_{A1}^2} + \frac{1}{\sigma_{A2}^2}\right)}}$$

Ici, la pondération est égale à l'inverse de la variance sur l'âge. Plus un âge est précis (variance petite), plus sa « masse » (inverse de la variance) est grande et donc plus il contribue au calcul de la moyenne A et de σ_A. Ces résultats servent ensuite à la calibration en date calendaire (*cf.* 1.4 & 12).

La moyenne géométrique est utilisée pour calculer le rapport moyen de variations dans une série chronologique (par exemple taux bancaire moyen lorsque ce taux varie sur une période donnée). Autre utilisation : lorsqu'une série statistique présente une grande dissymétrie (valeurs x_i concentrées puis progressivement étalées dans une queue de distribution, à gauche ou à droite), il est possible que le logarithme des x_i forme une série à peu près symétrique : dans ce cas, la moyenne géométrique sera plus représentative que la moyenne arithmétique.

Il est important de noter qu'il y a une relation d'ordre entre les quatre moyennes, l'égalité étant obtenue que si toutes les valeurs x_i sont identiques :

$$H \leq G \leq \underline{x} \leq Q$$

3.2 Les paramètres de dispersion

Pour caractériser la **dispersion** des données, il est nécessaire d'introduire d'autres paramètres qui sont l'étendue, les quantiles, les écarts moyens, la variance, l'écart-type et le coefficient de variation, qui font tous appel à la notion d'écart.

L'**étendue** e d'une série est la différence entre la plus grande valeur et la plus petite valeur de la variable : $e = \max(x_i) - \min(x_i)$. L'intervalle $[\min(x_i), \max(x_i)]$ permet de déterminer la plage contenant 100% des valeurs. Il est très sensible aux valeurs extrêmes.

Les **quantiles** désignent les quartiles, déciles et centiles. Les **quartiles** Q_1, Q_2, Q_3 sont les valeurs de la variable qui partagent l'effectif rangé par ordre croissant en quatre sous-ensembles de 25% chacun. Le quartile Q_2 correspond à la médiane. Si la distribution (l'histogramme) est dissymétrique, Q_1 et Q_3 ne seront pas symétriques par rapport à la médiane. Les **déciles** partagent la population en dix sous-ensembles égaux, les **centiles** partagent la population en cent sous-ensembles égaux.

L'**intervalle interquartile** permet une mesure de dispersion qui élimine l'influence des valeurs extrêmes. Cet intervalle qui contient 50% des valeurs est égal à $Q_3 - Q_1$. Plus cet intervalle est petit, plus la concentration (la densité de fréquence) des valeurs est grande.

Il existe une représentation graphique de cet intervalle appelée **boite à moustaches**, ou encore **Boxplot** en anglais (figure 8). On résume la série statistique à partir des quartiles Q_1, Q_2, Q_3 en traçant un rectangle allant de Q_1 à Q_3, coupé par un trait vertical au niveau de la médiane Q_2. On ajoute deux segments (moustaches) aux extrémités menant aux valeurs extrêmes ou bien au premier et dernier décile, ou encore au 1$^{\text{er}}$ et 99 ème centile qui sont approchés respectivement par $Q_1 - 1,5\,(Q_3 - Q_1)$ et $Q_3 + 1,5\,(Q_3 - Q_1)$ dans l'hypothèse d'une distribution gaussienne. La boite à moustaches permet de comparer facilement des distributions de populations de tailles différentes (toutes sont normalisées à 100%) et est largement utilisée en parallèle de l'histogramme.

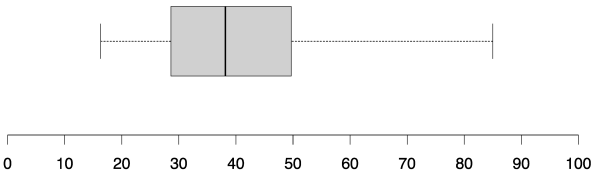


FIG. 8 : Box-Plot ou boite à moustaches : graphe qui permet de résumer une série statistique à partir des quartiles Q_1, Q_2 et Q_3. Il est constitué d'un rectangle construit sur Q_1 et Q_3 et divisé par un trait vertical correspondant à la médiane Q_2. Ici, nous avons effectué les calculs à partir du tableau 8 : soit $Q_1 = 28,7$, $Q_2 = 38,2$, $Q_3 = 49,7$. Les deux segments (moustaches) aux extrémités mènent aux valeurs minimale et maximale, soit 16,3 et 85,0. Ce graphe a été dessiné avec le logiciel R en utilisant la fonction *Box-Plot* dans laquelle on choisit l'option *« range = 0 »*.

L'**écart absolu moyen** $E_{\underline{x}}$ est la moyenne arithmétique des écarts, en valeur absolue, à la moyenne arithmétique de la série : $E_{\underline{x}} = \frac{1}{n}\sum_i^p n_i\,|x_i - \underline{x}|$. Mais cette formule, pour naturelle qu'elle soit, est rarement utilisée car elle se prête mal au calcul algébrique à cause de la valeur absolue. C'est pourquoi on lui préfère la variance.

La **variance** V_x est la moyenne arithmétique des écarts quadratique (écarts élevés au carré) à la moyenne arithmétique de la série, soit :

$$V_x = \frac{1}{n}\sum_i^p n_i\,(x_i - \bar{x})^2$$

Le calcul de la variance est accéléré par la formule de König qui fait intervenir la moyenne quadratique des x_i :

$$V_x = \left(\frac{1}{n}\sum_i^p n_i x_i^2\right) - \bar{x}^2$$

Ainsi, la variance est égale à la moyenne pondérée des carrés moins le carré de la moyenne. Mais si l'unité de mesure de la variable est le litre, la variance elle s'exprimera en litre ^{2} (litre au carré), ce qui n'est pas homogène avec la mesure en litre. C'est pourquoi on repasse en litre en prenant la racine carrée de la variance : L'écart-type s_x, ou *standard-deviation* en anglais, est alors défini par :

$$s_x = \sqrt{V_x}$$

L'**écart-type** indique comment, en moyenne, les valeurs de la variable sont groupées autour de la moyenne arithmétique. C'est donc une mesure de dispersion. L'écart-type permet d'apprécier le risque qu'une valeur de la variable, prise au hasard, soit proche ou éloignée de la moyenne.

La notion de variance (et d'écart-type) est plus souvent utilisée que celle d'écart absolu moyen pour la raison principale qu'elle permet des manipulations mathématiques plus simples (par exemple lorsque l'on combine des variances). D'autre part, la variance est reliée à la notion d'inertie en mécanique, à l'instar de la moyenne arithmétique pondérée qui est reliée à la notion de centre de gravité. Pour suivre l'analogie mécanique, plus les valeurs x_i seront dispersées, plus le moment d'inertie de ces valeurs autour de la moyenne $\bar{x}$ sera grand, autrement dit, plus l'énergie nécessaire pour faire tourner cet ensemble de valeurs autour de la moyenne sera grande.

Le **coefficient de variation** correspond à un écart-type relatif, c'est-à-dire à une mesure de dispersion relative. Il est défini comme le rapport entre l'écart-type s_x et la moyenne $\bar{x}$:

$$CV = \frac{s_x}{\bar{x}}$$

C'est un nombre positif sans dimension, sans unité, qui permet de comparer des séries statistiques différentes. En particulier, il permet de comparer la dispersion de séries utilisant des unités différentes (par exemple des mètres et des kilogrammes). A noter qu'il n'y a pas d'intervalle de confiance construit à partir du coefficient de variation.

L'**inégalité de Bienaymé-Tchebychev** montre que, quelle que soit la série statistique étudiée, un intervalle construit en retranchant et en ajoutant deux fois l'écart-type à la moyenne arithmétique contient plus de 75% de l'effectif total. Et dans le cas particulier où les valeurs x_i sont générées par une loi de Gauss (loi dite normale), l'intervalle $[\bar{x} - s_x, \bar{x} + s_x]$ contient 68% des données et l'intervalle $[\bar{x} - 1,96s_x, \bar{x} + 1,96s_x]$ contient 95% des données.

On peut aussi déterminer l'intervalle le plus court possible qui contient par exemple 95% des valeurs x_i, triées en ordre croissant (comme pour la détermination des quantiles). Cet intervalle est appelé **Intervalle de crédibilité** (IC) dans le cadre de la statistique bayésienne (*cf.* 12).

3.3 Les paramètres de forme

La forme de la distribution de fréquence (histogramme) entraîne une relation entre le mode m, la médiane M et la moyenne arithmétique $\bar{x}$ en fonction de la forme de l'histogramme (figure 9) :

- Si la distribution est symétrique, les trois paramètres sont égaux. Si la distribution n'est pas trop dissymétrique, on a l'approximation de Pearson : $m =$

$3M - 2\bar{x}$. Cette formule permet de vérifier la vraisemblance des résultats, ou d'émettre une hypothèse sur la forme de la distribution.

- Si la distribution est dissymétrique [oblique] à gauche, soit asymétrique vers la droite [queue de distribution étalée à droite], on a : $m \leq M \leq \bar{x}$.
- Si la distribution est dissymétrique [oblique] à droite, soit asymétrique vers la gauche [queue de distribution étalée à gauche], on a : $\bar{x} \leq M \leq m$.

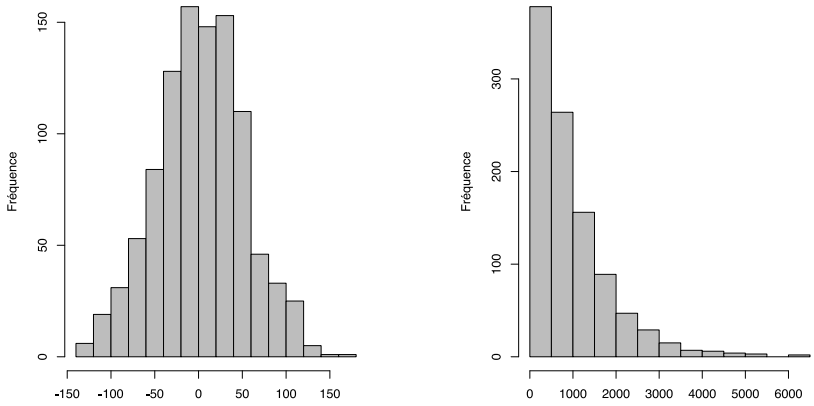


FIG. 9 : Histogrammes symétrique et dissymétrique. Le premier histogramme à tendance symétrique a été obtenu par un tirage aléatoire de 1000 données dans une distribution normale (Gauss), qui est symétrique, de moyenne nulle et d'écart-type égal à 50. Les paramètres mode, médiane et moyenne observés sont très proches. Le second histogramme montre une allure nettement dissymétrique à gauche, c'est-à-dire avec queue de distribution étalée à droite. Il a été obtenu par tirage aléatoire de 1000 données dans une distribution binomiale négative de paramètres ($\mu = 1000, r = 1,2$). Dans ce cas, le mode (= 424,0) est nettement à gauche, la moyenne (= 1016,6) est tirée vers la queue de distribution tandis que la médiane (= 764,0) se situe entre les deux.

Le **coefficient d'asymétrie**, ou **skewness** en anglais, mesure l'asymétrie d'une série statistique. C'est une moyenne pondérée définie par :

$$S = \frac{1}{n} \sum_{i}^{p} n_i \left(\frac{x_i - \bar{x}}{s_x} \right)^3$$

C'est un nombre sans dimension (sans unité car il s'agit d'un écart réduit par l'écart-type) dont les valeurs sont interprétées de la manière suivante :

Si $S = 0$, on a une distribution symétrique qui est comparable à une distribution gaussienne (*cf.* 7).

Si $S > 0$, la distribution est dissymétrique à gauche.

Si $S < 0$, la distribution est dissymétrique à droite.

Le **Kurtosis** est un coefficient qui mesure le **degré d'aplatissement** d'une distribution, défini par la moyenne pondérée suivante :

$$K = \frac{1}{n} \sum_{i}^{p} n_i \left(\frac{x_i - \bar{x}}{s_x} \right)^4$$

Si $K = 3$, la distribution est proche d'une distribution gaussienne.

Si $K > 3$, la distribution est plus pointue que la gaussienne.

Si $K < 3$, la distribution est plus aplatie que la gaussienne.

3.4 Choix des paramètres descriptifs - vers l'analyse de la variance

Le choix entre mode, médiane et moyenne arithmétique dépend de la question posée. En reprenant l'exemple du volume des amphores, la moyenne donne une idée du volume, mais ce volume peut ne pas exister dans les observations et il peut aussi être influencé par des volumes hors tendance, trop grands ou trop petits du fait d'erreurs de mesure par exemple ou bien encore de « contamination » par des individus exogènes. La médiane est donc plus représentative de la répartition réelle des volumes : 50% des volumes sont plus petits et 50% sont plus grands que la médiane.

Si la distribution est monomodale (un seul maximum dans l'histogramme) et fortement dissymétrique, il est préférable d'utiliser le mode et la médiane au lieu de la moyenne car celle-ci sera fortement influencée par la queue de distribution (à gauche ou à droite). Les paramètres étendue, quantiles et variance/écart-type peuvent être, quant à eux, utilisés en même temps de façon complémentaire pour caractériser la dispersion des données.

Si la distribution est multimodale, il se peut que ni la moyenne ni la médiane ne soient dans les données ! Ainsi, dans une distribution de date calendaire multimodale issue d'une calibration d'âge ^{14}C, la date moyenne peut tomber dans un creux. L'utilisation de la moyenne n'a donc pas de sens dans ce cas. On préfèrera identifier les différents modes présents dans la distribution pour répondre à la question : quelles sont les dates les plus probables au vu de l'échantillon étudié ? Dans le cas d'une distribution multimodale du volume des amphores, on peut être amené à décomposer la population en sous-populations qui peuvent être monomodales et plus faciles à décrire. Dans ce cas, la distribution multimodale globale peut être le résultat d'un mélange de deux ou plusieurs sous-populations dont la mise en évidence dans l'histogramme dépendra de l'amplitude choisie pour les classes. Il faut faire des essais pour choisir la meilleure partition en classes. Si une population de taille N est décomposée en r sous-populations de tailles différentes $N_1, \ldots, N_r$ (ces sous-populations deviennent alors des catégories, au sens de variables qualitatives), on montre que la moyenne arithmétique de la population totale est égale à la moyenne pondérée des moyennes des sous-populations, la pondération appliquée à chacune de ces dernières étant égale à $\frac{N_r}{N}$. De plus, la variance totale de la population sera égale à la somme de deux variances : une variance inter-population qui est donnée par la variance pondérée des moyennes des sous-populations ; une variance intra-population qui est donnée par la

moyenne pondérée des variances des sous-populations. Cette décomposition est un pas vers ce qu'on appelle l'**analyse de la variance** (Saporta, 2011).

En résumé :

Avant de commencer l'étude

- Il est essentiel de bien définir la population, l'échantillon, les individus, les variables étudiées et leur nature (qualitative, discrète ou continue), les modalités ou les classes de ces variables. La compréhension des résultats statistiques (graphiques) passe par l'identification de chacun de ces termes.
- Ne pas modifier l'effectif de la population en cours d'étude !
- Il ne faut pas confondre l'effectif (nombre d'individus) pour une modalité, pour une valeur discrète ou pour une classe d'une part, avec l'effectif qui correspondrait à un comptage (observation) dans le cas d'une variable discrète, d'autre part.

Pour éviter les erreurs courantes

- Ne pas confondre diagramme en barres et histogramme.
- Dans un diagramme en barres, séparer les barres et ne pas les relier par une courbe ou des segments de droite.
- Ne pas confondre la densité de fréquence (hauteur de l'histogramme) avec la fréquence (surface de l'histogramme).
- Moyenne et variance n'ont de sens que sur des variables quantitatives (numériques) discrètes ou continues.
- Attention à la signification de la moyenne dans le cas d'une distribution de fréquences multimodale : dans ce cas, la moyenne peut ne pas exister dans les données.