STV

Notes méthodologiques

Sang Thrombose Vaisseaux 2010 ;
22, n° 10 : 536-43

Les différents types de variables, leurs représentations graphiques et paramètres descriptifs

Julien Labreuche

INSERM-U 698, 46, rue Henri Huchard 75018 Paris, France
<julien.labreuche@bch.aphp.fr>

Résumé. Quel que soit le phénomène étudié, on est amené à collecter et analyser un ensemble de données communément appelé en statistique « ensemble de variables ». Il existe principalement deux groupes de variables : les variables quantitatives et les variables qualitatives, qui peuvent être subdivisés en plusieurs sous-groupes. La reconnaissance des différents types de variables est la première étape d'une analyse de données car le choix des outils et méthodes statistiques à utiliser en dépend. Cette note aborde les différents types de variables que l'on peut rencontrer en recherche médicale. Elle décrit également les principaux graphiques (histogramme, boîte à moustaches, diagramme en bâtons, diagramme circulaire) et paramètres descriptifs (effectif, fréquence, moyenne, écart type, médiane, étendue interquartile) utilisés pour résumer l'information collectée.

Mots clés : variables, variables quantitatives, variables qualitatives, graphiques, paramètres descriptifs

Abstract

Variables: their graphical representations and descriptive statistics

Whatever the research question, it is necessary to collect and analyze a data set commonly named in statistics "set of variables". There are two main groups of variables: the quantitative variables and the qualitative variables, which can be subdivided into several subgroups. The statistical methods used differ according to the type of variables. The first step of data analysis is to indentify the different types of variables since the choice of statistical methods will depend. This statistical note describe the different types of variables that may be encountered in medical research. The main graphical representations (histogram, box plot, bar chart, pie chart) and descriptive statistics (frequency, percentage, mean, standard deviation, median, interquartile range) used to summarized the data are also presented.

Key words: variables, quantitative variables, qualitative variables, graphical representations, descriptive statistics

Tirés à part :
J. Labreuche

doi: 10.1684/stv.2010.0541

La notion de variable est indispensable en statistique. Elle se définit comme une caractéristique observable ou mesurable pouvant prendre plusieurs valeurs chez des individus[1] issus d'une population d'intérêt [1]. La recherche médicale se pratique le plus souvent sur un sous-ensemble de la population appelé échantillon (*sample* en anglais). On est amené à rencontrer plusieurs types de variables qui sont associés à des méthodes statistiques différentes.[1]

L'objectif de cette note méthodologique est de fournir les notions de base sur la reconnaissance des différents types de variables, préalable nécessaire aux choix des méthodes statistiques. Nous aborderons également les principaux graphiques et paramètres descriptifs qui leur sont associés.

Les différents types de variables

Il existe deux groupes de variables, les variables quantitatives et les variables qualitatives.

Les variables quantitatives

Une variable est dite quantitative lorsque ses valeurs sont des nombres qui peuvent être ordonnés et additionnés (c'est-à-dire qui ont un sens en tant que nombre) [2]. On différencie deux types de variables quantitatives : les variables quantitatives continues et discrètes.

En théorie, on parle d'une variable quantitative continue quand celle-ci correspond à un continuum de valeurs (typiquement issues d'une mesure), et de variable quantitative discrète lorsque la variable ne peut prendre qu'un nombre fini de valeurs isolées, souvent entières, dans l'intervalle où elle varie (typiquement issues d'un comptage). Cependant, en pratique, la distinction entre variable discrète et variable continue est artificielle. En effet, les unités et les précisions des instruments de mesure étant limités, il arrive que des variables dites continues ne puissent prendre que des valeurs isolées (exemple, la pression artérielle mesurée en mmHg). Ainsi, on différencie les variables quantitatives continues de celles discrètes par le nombre de valeurs possibles qu'elles peuvent prendre. Lorsque ce nombre sera faible, c'est-à-dire que l'on peut énumérer les valeurs, la variable sera considérée comme discrète et, dans le cas contraire, comme continue. Par exemple, le nombre de globules blancs qui est par nature une variable discrète sera considéré comme une variable continue puisque le nombre de valeurs possibles est difficilement énumérable.

[1]En statistique, l'individu correspond à une unité d'observation qui peut être des individus en tant que tel, ou des animaux, des cellules…

Les variables qualitatives

Une variable est dite qualitative lorsque ses valeurs sont des qualités appelées modalités (situation professionnelle : actif, étudiant, retraité...). Les modalités peuvent être exprimées sous forme littérale ou numérique. Comme pour les variables quantitatives, on différencie deux types : les variables qualitatives nominales et ordinales.

Les variables qualitatives nominales se caractérisent par des modalités qui ne peuvent pas être ordonnées, comme la couleur des yeux (bleu, marron, vert…). Par opposition, les variables qualitatives ordinales se caractérisent par des modalités qui peuvent être ordonnées comme, par exemple, le niveau d'étude (primaire, secondaire, supérieur). Lorsque les modalités sont remplacées par des nombres (exemple : 1 = primaire, 2 = secondaire, 3 = supérieur), elles se différencient des variables quantitatives discrètes par l'absence d'information sur la distance séparant les nombres. Il ne s'agit que d'une codification sans valeur arithmétique.

Dans la littérature scientifique, il est fréquent de trouver les termes de « variable dichotomique » ou « variable binaire » qui correspondent à une variable qualitative qui ne peut prendre que deux modalités souvent codées 0 et 1 (exemple ; 1 pour homme et 0 pour femme) [3].

Les principales représentations graphiques

Les représentations graphiques, permettent une appréhension globale des données. Ces représentations sont indispensables en statistique car le choix des paramètres descriptifs et des tests statistiques en dépend.

Les variables quantitatives

L'histogramme

L'histogramme est la représentation graphique la plus courante pour une variable quantitative *(figure 1)*. Il se caractérise par une juxtaposition de rectangles contigus de surfaces proportionnelles, dont les bases correspondent aux valeurs et les hauteurs aux fréquences (ou aux effectifs). Pour obtenir un graphique lisible, il est souvent nécessaire de regrouper les valeurs en classes d'amplitudes égales (revenant à transformer une variable continue en variable discrète). Le nombre de classes choisi dépend de la taille de l'échantillon (pour garantir un nombre suffisant d'individus par classe) et du niveau de précision désiré [2]. Dans certaines situations, on peut être amené à construire des histogrammes dont les classes sont d'amplitudes inégales. Dans ce cas, la hauteur du rectangle n'est par proportionnelle à l'effectif de la classe mais à sa densité (*i.e.* l'effectif divisé

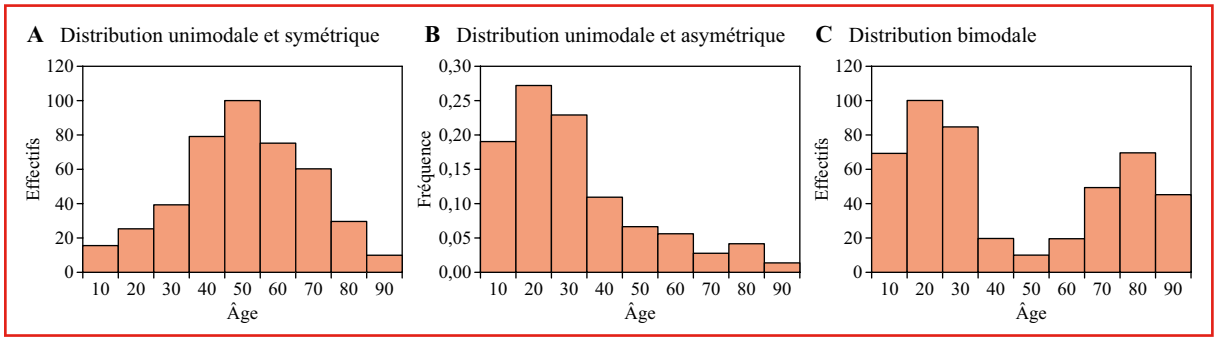

Figure 1. Exemple d'histogrammes. Les exemples présentés dans cette note méthodologique ont été élaborés à partir de données fictives. La figure 1A est caractéristique d'une distribution normale et la figure 1B d'une distribution log-normale.

par l'amplitude de la classe). Il devient plus difficile de comparer les fréquences des classes les unes par rapport aux autres.

Le principal avantage de l'histogramme est qu'il permet de visualiser la forme de la distribution (nombre de pics, symétrie, aplatissement) qui sera ensuite comparée aux principales distributions théoriques, dont la plus connue est la distribution normale (la normalité d'une variable étant un prérequis fréquent dans l'utilisation des tests statistiques) [4]. Nous verrons aussi que l'étude de la forme de la distribution orientera le choix des paramètres descriptifs.

La boîte à moustaches

La boîte à moustaches (communément appelée *Box Plot* en anglais) permet de représenter sur un seul graphique plusieurs paramètres descriptifs. Elle est également parfois appelée diagramme en boîte, diagramme de Tukey (nom de son inventeur) ou boîte à pattes. La boîte à moustaches, dessinée au-dessus ou à côté d'un axe, est constituée comme son nom l'indique, d'une boîte et de deux moustaches.

La boîte est délimitée par le premier et le troisième quartile (notés Q1 et Q3) et décrit 50 % des valeurs. La ligne à l'intérieur de la boîte représente la médiane ; sa position nous renseigne sur la symétrie de la distribution. Les extrémités des deux moustaches délimitent les valeurs dites extrêmes qui sont déterminées à partir de Q1 et de Q3. L'extrémité de la moustache inférieure est la valeur qui est supérieure à Q1-1.5*(Q3-Q1). L'extrémité de la moustache supérieure est la valeur qui est inférieure à Q3+1.5*(Q3-Q1) *(figure 2)*. À titre de remarque, lorsqu'il n'y a pas de valeur extrêmes, les moustaches correspondent aux minimum et maximum des valeurs.

Cependant, il est fréquent de trouver d'autres définitions des moustaches que celles du graphique original ; en effet, les 5^{e} et 95^{e} percentiles sont souvent utilisés. Il convient ainsi de préciser la définition retenue en légende de la figure.

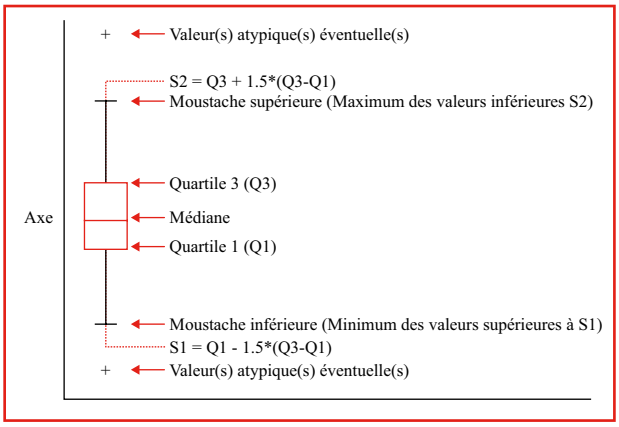

Figure 2. Les éléments constitutifs du graphique original [5].

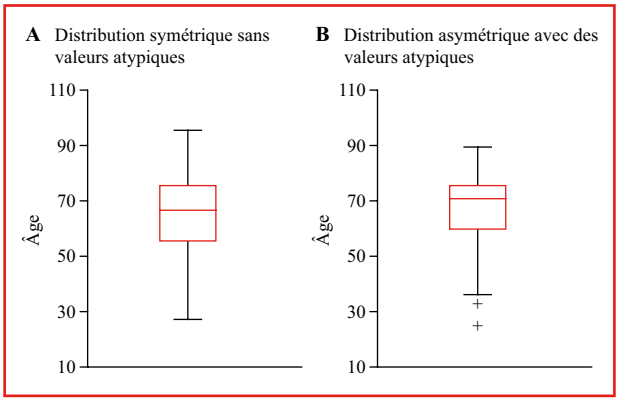

Figure 3. Exemple de boîte à moustaches.

Comparativement à l'histogramme, la boîte à moustaches permet de visualiser les valeurs extrêmes et de représenter côte à côte plusieurs distributions. En revanche, excepté la symétrie, elle ne nous renseigne pas sur la forme de la distribution *(figure 3)*.

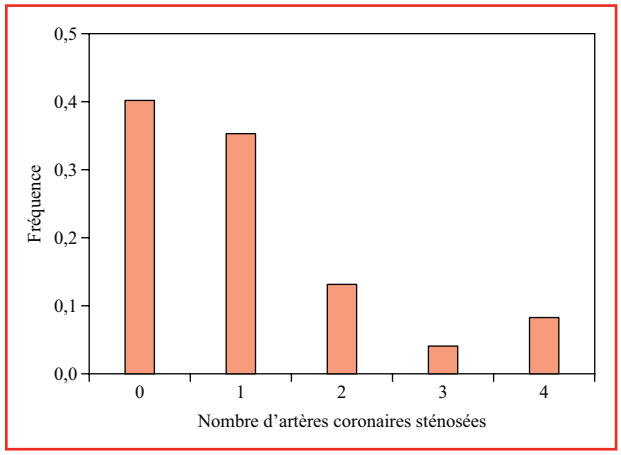

Figure 4. Exemple de diagramme en bâtons.

Le diagramme en bâtons

Le diagramme en bâtons est utilisé dans le cas d'une variable quantitative discrète *(figure 4)*. Il repose sur le même principe que l'histogramme mais les rectangles sont remplacés par des segments (bâtons). Le principal avantage de ce diagramme est qu'il traduit le caractère « isolé » des valeurs.

Les variables qualitatives

Les représentations graphiques des variables qualitatives permettent de visualiser et comparer les fréquences des différentes modalités. Le plus souvent, elles consistent à faire correspondre aux fréquences des surfaces proportionnelles.

Le diagramme circulaire

Il consiste à représenter l'échantillon total par un cercle dont la surface sera divisée en plusieurs secteurs d'angles proportionnels aux fréquences de chacune des modalités *(figure 5)*. On le trouve aussi sous le nom de « diagramme en secteur » ou « diagramme en camembert » en référence aux portions de ce fromage. Il est fréquent que le diagramme circulaire soit dessiné en perspective ayant pour conséquence de ne pas conserver la proportionnalité des surfaces. Il devient alors plus difficile de comparer les fréquences des modalités les unes par rapport aux autres.

Le diagramme circulaire est rarement utilisé pour représenter les variables qualitatives ordinales pour lesquelles l'ordre des modalités a son importance. Ce type de graphique n'est pas adapté lorsque le nombre de modalités devient trop important ; on préférera représenter les fréquences sur un diagramme en barres.

Le diagramme en barres

Classiquement, le diagramme en barres est une juxtaposition de rectangles de surfaces proportionnelles aux fréquences de chacune des modalités. On prend la précaution de séparer les rectangles par des espaces équidistants pour le différencier de l'histogramme et rappeler que les modalités sont des qualités et non des quantités *(figure 6)*. Comme pour le diagramme à bâtons et l'histogramme avec des classes d'amplitudes égales, la hauteur d'un rectangle nous ren-

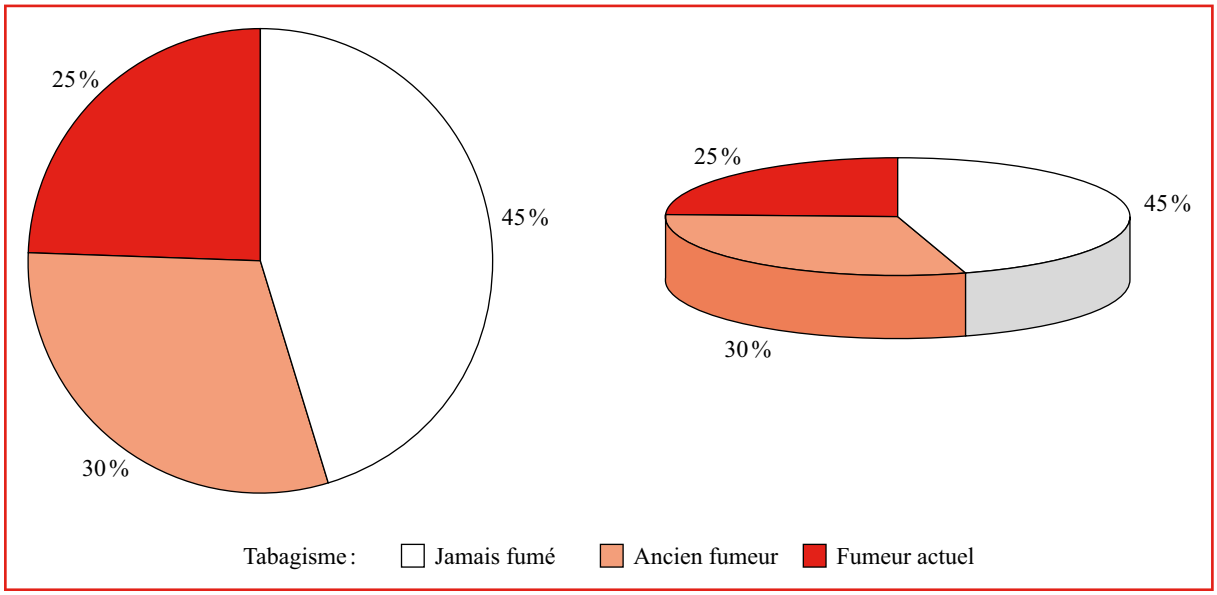

Figure 5. Exemple de diagrammes circulaires.

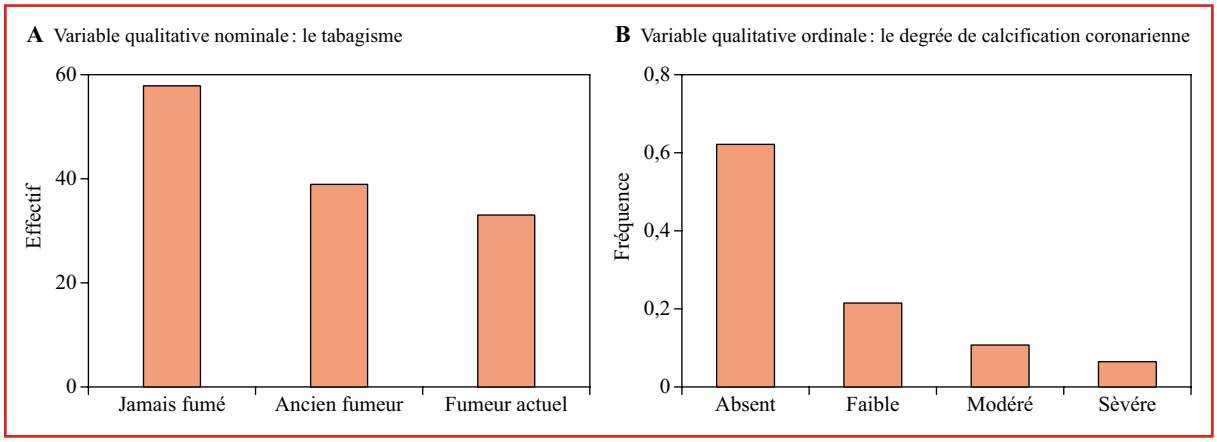

Figure 6. Exemple de diagrammes en barres juxtaposées.

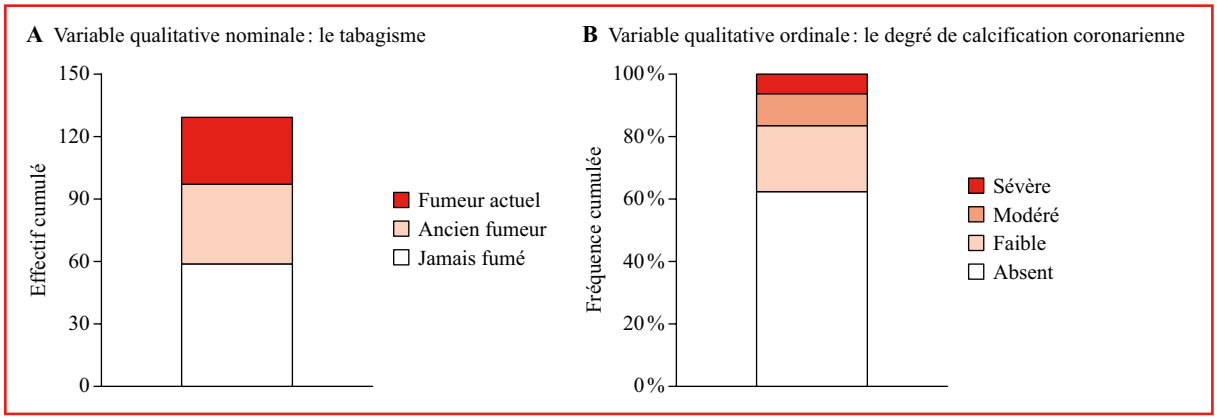

Figure 7. Exemple de diagrammes en barres superposées.

seigne sur la fréquence (ou l'effectif) d'une modalité[2]. Lorsque la variable est qualitative ordinale, il est d'usage de respecter l'ordre des modalités.

Le principal inconvénient de ce type de graphique est qu'il est difficile de représenter côte à côte plusieurs distributions. Cet inconvénient incite parfois à changer la représentation en empilant les barres *(figure 7)*. Comme pour le diagramme circulaire, ce type de graphique ne convient pas lorsque le nombre de modalités est important.

Les paramètres descriptifs

Les paramètres descriptifs permettent de résumer en quelques chiffres les distributions des variables qui constituent l'échantillon. Selon le type de la variable, on utilisera des paramètres différents. Ainsi, on calculera des effectifs et des fréquences selon les modalités d'une variable qualitative[3]. Pour une variable quantitative, il existe un éventail plus large de paramètres descriptifs séparés en 2 groupes : les paramètres de position et les paramètres de dispersion. Ce sont ces paramètres que nous allons détailler.

Les paramètres de position

Paramètres de tendance centrale

La moyenne et la médiane sont les mesures de tendance centrale les plus communes qui définissent une valeur

[2]Qu'il s'agisse du diagramme en barres, en bâtons, ou de l'histogramme, on peut avoir une vision erronée de la différence de fréquences de deux valeurs lorsque l'origine de l'axe qui les représente est différente de la valeur nulle.

[3]À noter qu'en anglais, le mot *frequency* est la traduction du terme français « effectif » et non du terme « fréquence ». Les termes anglophones *relative frequency*, *proportion* ou *percent* peuvent être utilisés pour traduire le mot « fréquence » (voir lexique français-anglais).

autour de laquelle se répartissent les observations. Le mode est également un paramètre de tendance centrale qui est très peu utilisé en pratique.

La moyenne arithmétique est la somme des observations divisée par le nombre d'individus :

$$\overline{X} = \frac{\sum_{i}^{n} X_i}{n}$$

où X_i sont les observations et n la taille de l'échantillon

La médiane est la valeur qui partage en deux parties égales l'échantillon. Le calcul de la médiane nécessite de classer les observations par ordre croissant et de différencier deux situations :

$n/2$ n'est pas un entier la médiane correspond précisément à la valeur centrale soit $X_{(n+1)/2}$

$n/2$ est un entier, la médiane correspond à la moyenne des deux valeurs centrales :

$$\frac{X_{n/2} + X_{n/2+1}}{2}$$

Le mode correspond à la valeur la plus fréquente dans le cas d'une variable discrète, et au sommet (*i.e.* la classe la plus dense) de l'histogramme dans le cas d'une variable continue (on parlera plutôt de classe modale que de mode).

Si la médiane est plus difficile à calculer que la moyenne, elle présente l'avantage d'être moins sensible aux valeurs extrêmes et à la forme de la distribution (on privilégiera la médiane à la moyenne en cas de distribution asymétrique[4]).

La médiane est également un paramètre qui est utilisé dans le cas d'une variable discrète et d'une variable qualitative ordinale avec un nombre important de modalités.

Autres paramètres de position

Les autres paramètres de position les plus courants sont les quartiles et les percentiles. Ils reposent sur le même principe que la médiane, à savoir qu'ils partagent la série des observations en plusieurs groupes d'effectifs égaux[5].

Les quartiles sont les trois valeurs qui partagent la distribution en quatre groupes d'effectifs égaux, Le premier quartile sépare 25 % des valeurs les plus faibles et 75 % des valeurs les plus élevées, le second quartile correspond à la médiane, et le troisième quartile sépare 75 % des valeurs les plus faibles et 25 % des valeurs les plus élevées.

Les percentiles sont les 99 valeurs qui partagent la distribution en 100 groupes d'effectifs égaux. Un percentile est la valeur sous laquelle un certain pourcentage des observations se trouve : le 25^{e} percentile correspond au premier quartile, le 50^{e} percentile correspond à la médiane ou second quartile et le 75^{e} percentile correspond au troisième quartile.

Comme pour la médiane, le procédé de calcul des quartiles et des percentiles est différent selon que le rapport « nombre d'observations/nombres de groupes » soit un entier ou pas ($n/4$ pour les quartiles et $n/100$ pour les percentiles). Concernant les quartiles, si $n/4$ n'est pas un entier, le premier quartile est la valeur de rang immédiatement supérieur à $n/4$; si $n/4$ est un entier, le premier quartile correspond à la moyenne des deux valeurs de rang $n/4$ et de rang immédiatement supérieur à $n/4$ (pour le troisième quartile, il suffit de remplacer n par $3n$). Une approche similaire est utilisée pour les percentiles.

Les paramètres de dispersion

Les paramètres de dispersion évaluent la répartition des observations autour des valeurs centrales. On associe l'écart type à la moyenne et l'intervalle interquartile à la médiane.

L'écart type

L'écart type se déduit du calcul de la variance[6] en prenant sa racine carrée. On le note habituellement *SD* (de l'anglais *standard deviation*). La variance se définit littéralement comme la moyenne des carrés des écarts à la moyenne. Elle se traduit par la formule suivante :

$$\frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$$

où est la moyenne arithmétique[7]. D'où l'écart type :

$$SD = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}}$$

Une interprétation de la valeur d'un écart type peut être donnée pour des distributions normales, à savoir que l'intervalle défini par $[\overline{X} - 2 * SD; \overline{X} + 2 * SD]$ contient environ 95 % des données. Par conséquent, une simple comparaison entre la moyenne et l'écart type d'une distribution

[4]On peut aussi transformer la variable pour obtenir une distribution normale et calculer la moyenne de cette nouvelle variable [6]. Lorsque la forme de la distribution n'est pas unimodale (*i.e.* plusieurs pics sont observés), la médiane comme la moyenne ne sont pas adaptées. Dans cette situation, il est préférable de transformer la variable quantitative en une variable qualitative.

[5]Il est aussi fréquent de trouver les tertiles, quintiles et déciles qui découpent la série d'observations en trois, cinq et dix parties d'effectifs égaux [7].

[6]La variance est très utilisée dans les calculs statistiques mais ne l'est pas comme paramètre descriptif car elle ne s'exprime pas dans la même unité que la variable.

[7]Si on travaille sur l'ensemble de la population, situation exceptionnelle en statistique, le dénominateur sera n au lieu de (n-1). Cette différence s'explique par le fait que, à partir d'un échantillon, on estime les paramètres descriptifs que l'on aurait obtenus sur l'ensemble de la population.

permet de détecter un écart flagrant à la normalité ; si la moyenne est deux fois plus petite que l'écart type, il est peu probable que la distribution soit normale[8] [8].

Il est important de souligner que, dans les articles scientifiques, l'écart type est souvent confondu avec un autre paramètre statistique : l'écart type de la moyenne [9]. Pour éviter une confusion entre ces deux termes, on utilise préférentiellement le terme d'erreur type de la moyenne par analogie avec l'anglais (*standard error of the mean* qui est souvent écrit en abrégé : SEM) [10]. L'écart type quantifie la variabilité des observations d'un échantillon autour de sa moyenne et l'erreur type quantifie la variabilité des moyennes que l'on aurait obtenues avec différents échantillons. L'erreur type est une mesure de précision de l'estimation de la moyenne d'une population [11], qui se calcule simplement en divisant l'écart type par la racine carrée de l'effectif total de l'échantillon ($\frac{SD}{\sqrt{n}}$). Il permet de déduire un intervalle à l'intérieur duquel se situe la vraie valeur de la moyenne (celle que l'on obtiendrait sur l'ensemble de la population), avec un certain degré de confiance (probabilité). Le plus fréquent est de reporter l'intervalle dans lequel la probabilité de trouver la vraie valeur est de 95 % ; cet intervalle se définit par $[\overline{X} - 2 * SEM ; \overline{X} + 2 * SEM]$, autrement appelé « intervalle de confiance à 95 % »[9].

L'intervalle interquartile

L'intervalle interquartile est l'écart entre le premier et le troisième quartile. Il contient 50 % des données (25 % de part et d'autre de la médiane). L'amplitude de l'intervalle est parfois utilisée comme mesure de dispersion ; on parlera d'étendue interquartile.

*L'étendue (***range** *en anglais)*

Il est également d'usage d'utiliser l'étendue qui représente l'écart entre la plus grande et la plus petite valeur. Elle est par définition très sensible aux valeurs extrêmes.

Exemple de calcul des principaux paramètres descriptifs associés à une variable quantitative

Soit un échantillon représentant respectivement les âges de 10 individus, rangé par ordre croissant :

Âge	20	30	30	45	50	55	60	65	70	80
Rang	1	2	3	4	5	6	7	8	9	10

La moyenne : $\overline{X} = \frac{20+30+30+45+50+55+60+65+70+80}{10} = 50,5$.

La médiane : le rapport $n/2 = 5$ étant une valeur entière, la médiane correspond à la moyenne des données de rang 5 et 6. $med = \frac{50+55}{2} = 52,5$.

Le premier quartile (Q1) : on calcule dans un premier temps le rapport (n/4). Le résultat étant une valeur fractionnaire (10/4 = 2,5), Q1 correspond à la donnée de rang 3, Q1 = 30.

Le troisième quartile (Q3) : on calcule dans un premier temps le rapport (3n/4). Le résultat étant une valeur fractionnaire (30/4 = 7,5), Q3 correspond à la donnée de rang 8, Q3 = 65.

L'écart type :

$$\sigma = \sqrt{\frac{(20-50,5)^2+(30-50,5)^2+...+(70-50,5)^2+(80-50,5)^2}{10-1}} = 19,4$$

L'intervalle interquartile (IIQ) : se déduit du calcul de Q1 et de Q3, IIQ = [30 ; 65]. ■

Lexique des équivalents anglais de termes statistiques français

Français	Anglais
Effectif	Frequency (ou absolute frequency)
Fréquence	Relative frequency, percent, proportion
Échantillon	Sample
Moyenne	Mean
Écart type	Standard deviation
Erreur standard	Standard error
Médiane	Median
Intervalle de confiance	Confidence interval
Intervalle interquartile	Interquartile interval
Étendue interquartile	Interquartile range
Étendue	Range

Conflits d'intérêts : aucun

Références

1. Altman DG, Bland JM. Variables and parameters. *Br Med J* 1999 ; 318 : 1667.

2. Falissard B. *Comprendre et utiliser les statistiques dans les sciences de la vie*. Paris : Masson, 1998.

3. Larson MG. Descriptive statistics and graphical displays. *Circulation* 2006 ; 114 : 76-81.

4. Altman DG, Bland JM. The normal distribution. *Br Med J* 1995 ; 310 : 298.

[8]Cette comparaison n'a de sens que pour une variable quantitative qui ne prend que des valeurs positives.

[9]Ce calcul n'est valide que si la taille de l'échantillon est suffisamment grand ($n > 30$ est souvent retenu) ou, dans le cas contraire, si la variable suit une loi normale. Il est également possible de calculer des intervalles de confiance pour les autres paramètres mais cela dépasse le cadre de cette note méthodologique.

5. Le Guen M. La boîte à moustaches pour sensibiliser à la statistique. *Bull Method Sociol* 2002 ; 73 : 43-64.

6. Bland JM, Altman DG. Transformation, means, and confidence intervals. *Br Med J* 1996 ; 312 : 1079.

7. Altman DG, Bland JM. Quartiles, quintiles, centiles, and other quantiles. *Br Med J* 1994 ; 309 : 996.

8. Altman DG, Bland JM. Detecting skewness from summary information. *Br Med J* 1996 ; 313 : 1200.

9. Altman DG, Bland JM. Standard deviations and standard errors. *Br Med J* 2005 ; 331 : 903.

10. Chatellier G, Durieux P. Moyenne, médiane, et leurs indices de dispersion : quand les utiliser et comment les présenter dans un article scientifique ? *Rev Mal Respir* 2003 ; 20 : 421-4.

11. Medina LS, Zurakowski D. Measurement variability and confidence intervals in medicine : why should radiologists care? *Radiology* 2003 ; 226 : 297-301.