

# Introduction aux logiciels pour les statistiques

Version du 25/03/2020

**Sofiane MAAZI**

**Sofiane.Maazi@polytech-lille.fr**

**Polytech Lille (GIS2A3)**

**2019 – 2020**







***1<sup>ère</sup> partie***

**Préambule**

# Préambule

- **Volume horaire** : 1 séances de 3 h de cours / 3 séances de 2h00 de TP / 1 séance de 2 heures d'examen (à distance ?)
- **Modalité d'évaluation** : un QCM (« contrôle continu ») et un TP de R noté (« examen »)
- **À connaître impérativement (QCM)** : les passages marqués du symbole 
- **À maîtriser impérativement (examen)** : l'activité, les scripts des TP, les scripts du cours identifiables par le symbole. 
- **Aide à la lecture du cours** : plan du cours / exemples en bleu / remarques en vert
- **Bibliographie et sources** : disponibles en fin de diaporama



# Préambule

Ce cours de statistique descriptive est très **important** !

La statistique descriptive (appelée statistique exploratoire ) consiste à décrire (représenter et résumer) par des indicateurs statistiques les données quand elles sont nombreuses.



Dans la 1<sup>ère</sup> partie du cours , nous décrirons les variables une à une. Il s'agira d'analyser à l'aide d'indicateurs statistiques une variable d'intérêt et de la représenter sous forme de tableaux et/ou graphiques (statistique univariée).

Dans la 2<sup>nd</sup> partie du cours , nous décrirons les variables 2 à 2. Il s'agira de mesurer - si elles existent - les liaisons entre 2 variables d'intérêt et de les représenter éventuellement sous forme de tableaux et/ou graphiques (statistique bivariée).



*Il est possible (et souhaitable !) d'analyser les variables non pas une par une ou deux par deux mais toutes en même temps. C'est l'objet du cours de statistique exploratoire multivariée !*

# Préambule

## Confidentialité de l'information ou la loi Informatique et Liberté

Le statisticien manie des données informatiques potentiellement confidentielles. Il doit donc veiller à respecter une certaine déontologie. De manière plus générale, tout fichier informatique sur des individus quels qu'ils soient, doit obligatoirement faire l'objet d'une déclaration auprès de la Commission nationale de l'informatique et des libertés (CNIL), instituée par la loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés qui la qualifie d'autorité administrative indépendante ([www.cnil.fr](http://www.cnil.fr)).



**Activité du cours**



A decorative graphic on the left side of the slide, consisting of several concentric circles in different shades of blue. The innermost circle is a light blue, and it is surrounded by several darker blue rings of varying thicknesses, creating a sense of depth and movement.

**2<sup>ème</sup> partie**

# **Statistiques univariées**

# 2.1 La statistique

Il existe plusieurs définitions de la statistique ( Petit Robert) :

- **Vieilli** : Étude méthodique des faits sociaux, par des procédés numériques (classements, dénombrements, inventaires chiffrés, recensements), destinée à renseigner et aider les gouvernements.
- **Moderne** : Branche des mathématiques appliquées qui utilise le calcul des probabilités pour établir des hypothèses à partir d'événements réels et faire des prévisions concernant des circonstances analogues.

Ce cours s'inscrit dans l'état d'esprit de la première définition à laquelle il faudrait ajouter... le recours à l'informatique pour traiter des gros volumes de données !

*Le mot au singulier renvoie davantage à la discipline alors qu'au pluriel il désigne les valeurs des différents indicateurs statistiques qui seront vus en détails dans les prochaines diapositives !*



## 2.2 Unités statistiques

Une **unité statistique** est un objet qui présente une valeur pour un caractère étudié. On l'appelle également d'individu statistique (au sens humain) car historiquement la statistique exploratoire a trait à la démographie.

L'ensemble des unités statistique est appelé **population**. En pratique, il est difficile de disposer d'une population et donc de l'exhaustivité des données. On se contente alors d'un **échantillon**, c'est-à-dire d'un sous ensemble de la population.

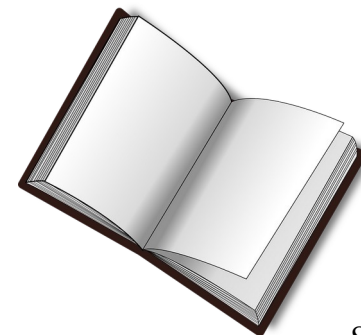
Le cardinal d'une population ou d'un échantillon est appelé **taille**.



Quand vous réalisez une étude statistique, il est important de préciser **le champs de l'étude**, c'est-à-dire l'ensemble des unités statistiques considérées !



Dans un fichier informatique, l'individu statistique est souvent représenté par un **identifiant unique** ( numéro client , numéro de sécurité social...).





## 2.2 Unités statistiques

### Exemples d'unités statistiques :

- un foyer fiscal pour une étude sur le taux d'imposition en France.
- une ampoule pour une étude sur la durée de vie des ampoules d'un certain constructeur.

### Exemples sur l'importance du champ d'étude et/ou du choix de l'échantillon :

1) Insee, Pôle Emploi : qui a les «bons» chiffres du chômage ? , Le Monde [ en ligne ]

2) Dans une ville on interroge des jeunes sur l'intention de vote pour les prochaines élections municipales. 85 % souhaitent voter pour le 1<sup>er</sup>, 10% pour le 2<sup>nd</sup> candidat et 5% pour le 3<sup>ème</sup> candidat. À la lumière de ces résultats, le journaliste annonce que le 1<sup>er</sup> candidat bénéficie d'une forte côte de popularité auprès des plus jeunes concitoyens. Commentez.

*Réponse : Il manque des informations comme le nombre de jeunes interrogés et le choix de l'échantillon !*



## 2.3 Variables d'intérêts

On appelle **série statistique** la suite des valeurs  $(x_1, x_2, \dots, x_n)$  prises par une variable  $X$  sur les unités d'observation. Ces valeurs constituent les **modalités** de la variable  $X$ . La taille de la série est notée  $n$ . L'absence de valeur de  $X$  pour une unité statistique est appelée **valeur manquante**.

Une variable  $X$  est également appelée **caractère  $X$** . Une variable est **qualitative** quand ses modalités ne sont pas numériques (donc non mesurables). En particulier, une variable est **qualitative ordinale** si les modalités peuvent être ordonnées. Sinon, la variable est **qualitative nominale**.



Exemple de variable qualitative ordinales : Dans une table de données, on dispose des résultats au baccalauréat. La variable MENTION contient les modalités « Assez Bien », « Bien » et « Très Bien ». Elle est qualitative ordinale.

Exemple de variable qualitative nominale : Le numéro de sécurité sociale et la nomenclature des PCS.

Une variable qualitative chiffrée n'est pas forcément de type numérique dans une table de données !



## 2.3 Variables d'intérêts

Une variable est **quantitative** quand ses modalités sont numériques (donc mesurables). En particulier, une variable est **quantitative discrète** si les valeurs des modalités sont entières. Sinon la variable est **quantitative continue**.



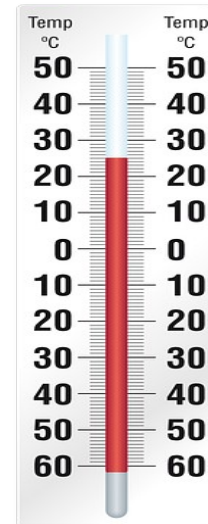
Exemple de variable quantitative discrète : la variable prend comme modalités les valeurs obtenues lors de plusieurs lancers successifs d'un dé à six faces.

Exemple de variable quantitative continue : la température.

Dans la cas des variables quantitatives continues, il arrive souvent qu'on les découpe en classes. La variable créée est alors **qualitative ordinale**.



Exemple : la variable âge est souvent découpée en classes d'âges.



## 2.4 Distribution d'une variable qualitative

**Rappel :** On appelle série statistique (ou vecteur des observations) une suite de valeurs observée d'un caractère X. On la note  $(x_1, x_2, \dots, x_n)$  ou  $\{x_1, x_2, \dots, x_n\}$ .

On considère un caractère qualitatif X à k modalités  $(m_1, m_2, \dots, m_k)$  et  $\{x_1, x_2, \dots, x_n\}$  la série statistique.

On note  $n_j$  l'effectif des individus vérifiant la  $j^{\text{ème}}$  modalité de X. On a bien entendu :  $\sum_{j=1}^k n_j \doteq n$

On appelle distribution de X, la suite des effectifs  $(n_j)_{j \in \llbracket 1, k \rrbracket}$  ou encore la suite des fréquences

$$(f_j)_{j \in \llbracket 1, k \rrbracket} \text{ où } f_j = \frac{n_j}{n}$$



*En pratique, on représente les effectifs ( ou fréquences) relatifs ) la place des effectifs absolus. Il ne faut pas oublier alors de mentionner l'effectif total de la série.*

## 2.4 Distribution d'une variable qualitative

On peut représenter une variable qualitative par un tableau ou un graphique.

On l'illustre par le palmarès général des écoles d'ingénieurs 2019 dressé par l'Étudiant. Les 174 écoles se sont vues attribuer des points selon les critères qu'elles remplissaient. Elles ont ensuite été classées ( *Tableau 1*) par groupe du meilleur ( A+) à au moins bon ( C). Certaines écoles n'ont pas souhaité renseigner toutes les informations requises.

Groupe	Effectif	Fréquence	Effectifs cumulés	Fréquences cumulées
A+	14	8.0	14	8.0
A	19	10.9	33	18.9
B	76	43.7	109	62,6
C	65	37.4	174	100




Tableau 1 -Répartition des écoles d'ingénieurs par groupe de niveau. *Source : l'Étudiant, 2019*

## 2.4 Distribution d'une variable qualitative

On peut remplacer le tableau précédent par le tableau ci-contre (Tableau 2). Il faut alors bien préciser l'effectif total dans la légende et ajouter une ligne total dans le tableau !

Tableau 2 - Répartition des écoles d'ingénieurs par groupe de niveau. Effectif total : 174 écoles. *Source : l'Étudiant, 2019.*

Groupe	Fréquence
A <sup>+</sup>	8.0
A	10.9
B	43.7
C	37.4
<b>Total</b>	<b>100</b>

Deux représentations graphiques sont possibles pour une variable qualitative. Le **diagramme en colonnes** et le **diagramme en secteurs**. 

*Vos graphiques doivent toujours être numérotés, toujours comporter un titre et une source, souvent comporter une note de lecture. Vous devez toujours les citer dans le texte du document mais ils doivent pouvoir être lus indépendamment du texte ! Il faut donc bien les documenter ! Quand vous imprimez vos graphiques en noir et blanc, il est **inutile** de les agrémenter de couleurs. Variez plutôt sur le mode de remplissage (motifs à rayures, à petits pois...) ! Le titre d'un graphique ne doit jamais faire référence au type de graphique !*

## 2.4 Distribution d'une variable qualitative

Un **diagramme en colonnes** (*Figure 1*) ne doit pas être confondu avec un histogramme qui est une des représentations graphiques possibles d'une variable quantitative !



Répartition des écoles d'ingénieurs par groupe de niveau.

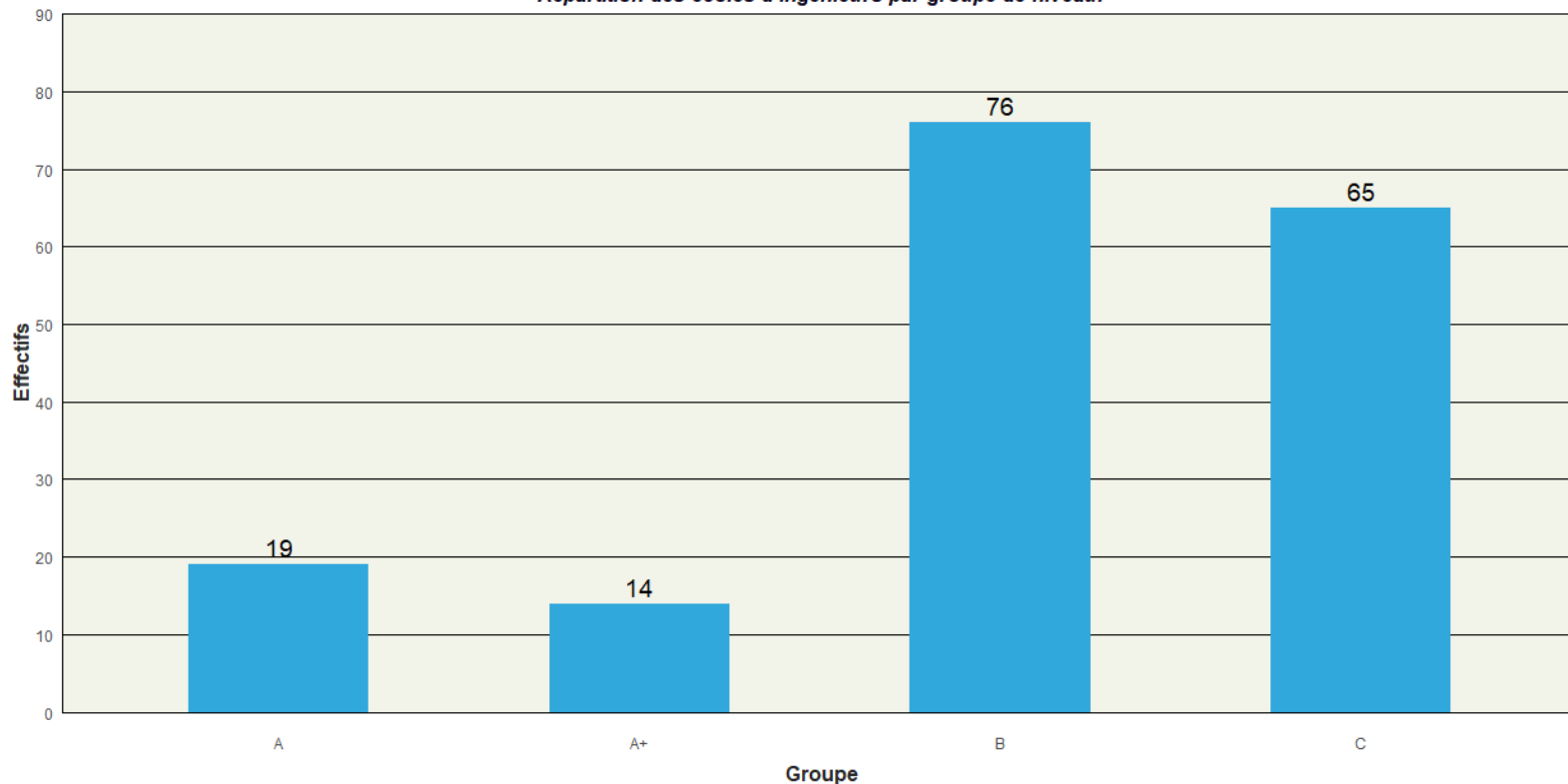


Figure 1 - Répartition des écoles d'ingénieurs par groupe de niveau. Effectif total : 174 écoles.  
*Source : l'Étudiant, 2019.*

## 2.4 Distribution d'une variable qualitative

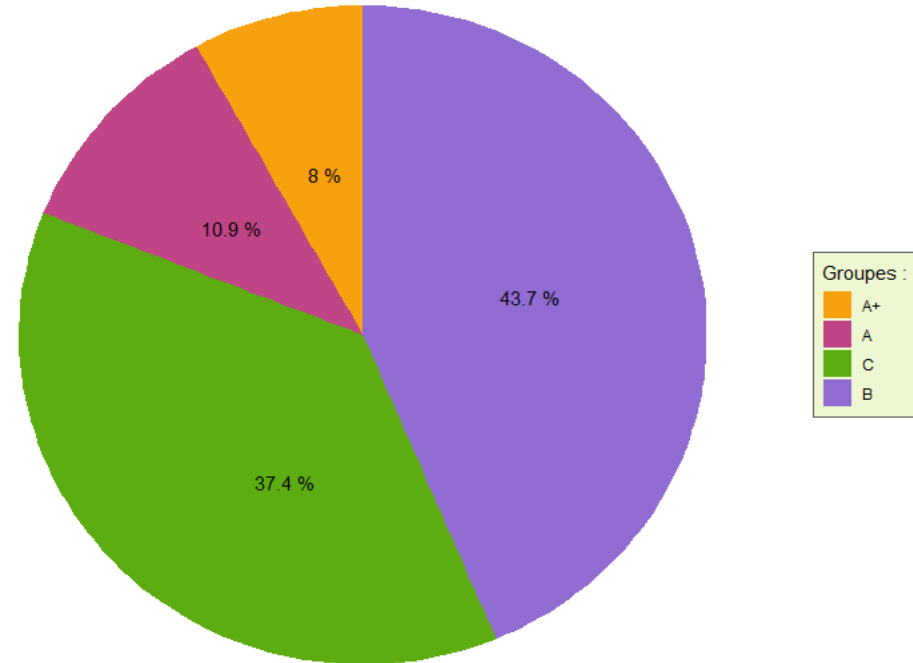
Un **diagramme en secteurs** (*Figure 2*) est aussi appelé « camembert ».

Il est tentant de le représenter en 3 dimensions mais la représentation obtenue est totalement fallacieuse ! En effet, une valeur chiffrée élevée peut apparaître faible selon l'angle d'inclinaison choisi (et vice-versa) !



R

Figure 2 - Répartition des écoles d'ingénieurs par groupe de niveau.  
Effectif total : 174 écoles.  
*Source : l'Étudiant, 2019.*





## 2.5 Distribution d'une variable quantitative

Comme pour les variables qualitatives ordinales, on peut calculer les effectifs, les effectifs cumulés, les fréquences et les fréquences cumulées associées à **une variable quantitative discrète**. Quand la variable est discrète, les effectifs sont alors représentés en R par des **bâtonnets**.



Une **variable quantitative continue** peut prendre une infinité de valeurs possibles. Le domaine de la variable est alors R ou un intervalle de R. Il est alors souvent intéressant de **procéder à des regroupements en classes pour faire des représentations graphiques**.

On considère un caractère quantitatif X et  $\{x_1, x_2, \dots, x_n\}$  la série statistique associée. On note  $C_1, C_2, \dots, C_k$  (avec k appartient à  $\mathbb{N}^*$ ), les classes d'intervalles disjoints dans lesquelles se répartissent les observations  $x_i$ . On note  $n_j$  le nombre d'observations appartenant à la classe  $C_j$ .

On appelle distribution de X, la suite des effectifs  $(n_j)_{j \in \llbracket 1, k \rrbracket}$  ou encore la suite des fréquences

$$(f_j)_{j \in \llbracket 1, k \rrbracket} \quad \text{où} \quad f_j = \frac{n_j}{n} \quad \text{avec} \quad \sum_{j=1}^k n_j = n$$

## 2.5 Distribution d'une variable quantitative

1) La distribution du caractère  $X$  peut être représentée par un tableau et/ou par un graphique. Ce dernier est cependant à privilégier pour son aspect visuel immédiat, mais en réalité, les deux représentations sont complémentaires !

2) Découper une variable quantitative en classes induit implicitement de créer une autre variable qualitative ordinale.

3) **Découper une variable quantitative en classes fait perdre de l'information.**



4) Les classes doivent être ordonnées !

5) Aucune valeur manquante dans les données ici.



## 2.5 Distribution d'une variable quantitative

On peut représenter une variable quantitative par une multitude de graphiques.

On s'intéressera ici à l'histogramme, ses courbes dérivées et la boîte à moustache.

### a) Histogramme

Une variable quantitative contient en général un grand nombre de valeurs. On les regroupe donc en classes qui peuvent être d'amplitudes variables. On note  $a_j$  l'amplitude de la classe  $C_j$ . On appelle alors histogramme de la variable  $X$  un graphique d'occurrences comportant en ordonnées les fréquences relatives  $(f_j/a_j)_{j \in [1, k]}$  à des classes de même amplitude en abscisses.

Autrement dit, un histogramme est une suite de rectangles - dans un repère *ad hoc* – qui représente les effectifs des classes par des rectangles dont la surface (et non la hauteur) représente l'effectif.

*Lorsque les classes n'ont pas la même amplitude, il faut respecter la contrainte de proportionnalité entre l'aire des rectangles et les effectifs. En effet,  $\sum_{j=1}^k \frac{f_j}{a_j} = 1$ . En pratique, les logiciels affichent*

*toujours des histogrammes avec des classes de même amplitudes mais font apparaître souvent les fréquences relatives  $(f_j/a_j)_{j \in [1, k]}$  au lieu des fréquences  $(f_j)_{j \in [1, k]}$ .*

## 2.5 Distribution d'une variable quantitative

Pour déterminer le nombre optimal  $u$  de classes, il n'existe pas de règle absolue ! Néanmoins deux critères expérimentaux sont souvent utilisés dans les logiciels de statistiques :

- la règle de Yule :  $u = 2,5 * n^{\frac{1}{4}}$

- la règle de Sturge :  $u = 1 + \log_2(n)$

avec  $n$  la taille de la série statistique



*L'histogramme permet de se faire rapidement une idée de la distribution d'une variable quantitative : les tendances centrales, la dispersion et la forme .*

L'histogramme à la diapositive suivante a été réalisé à partir du jeu de données *diamonds* de R. Ce jeu de données contient les caractéristiques de 5400 diamants. En particulier, la clarté et le prix du diamant.



## 2.5 Distribution d'une variable quantitative

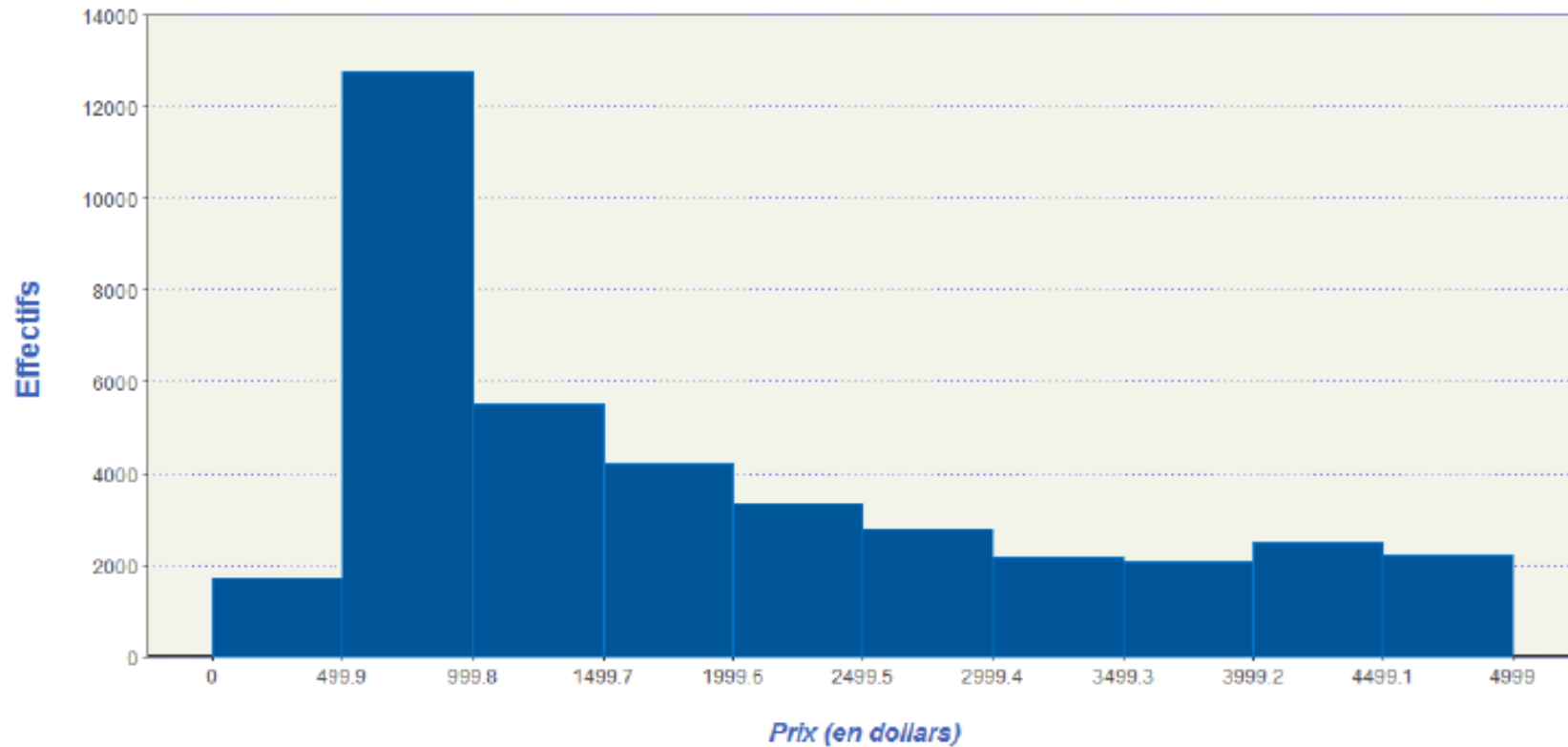


Figure 3 - Répartition du prix des diamants dont le prix est strictement inférieur à 5 000 dollars. Effectifs : 3 913 diamants. *Source : Jeu de donnée diamonds de R avec 5 400 diamants.*

## 2.5 Distribution d'une variable quantitative

### b) Boîte à moustaches



La boîte à moustaches a été inventée en 1977 par John Tukey. C'est un diagramme représentant la distribution d'un caractère quantitatif où figurent les **trois quartiles**, les **valeurs adjacentes** inférieure et supérieure délimitant les moustaches et éventuellement les **valeurs extrêmes**.

**Les boîtes à moustaches sont usitées notamment pour comparer les distributions de plusieurs variables ou d'une même variable entre différents groupes. Elles peuvent aussi être utilisées pour représenter la dispersion d'une unique variable.** La fonction qui produit ces graphiques est la fonction *boxplot*.

On considère une série statistique  $(x_1, \dots, x_n)$  **ordonnée** de  $n$  valeurs.

Le **1<sup>er</sup> quartile  $Q_1$**  de cette série est la valeur qui sépare cette série en deux groupes

- Le premier groupe contient un quart des effectifs (25 %)
- Le deuxième groupe contient trois quarts des effectifs (75 %)



Le **2<sup>ème</sup> quartile  $Q_2$  ou médiane  $Me$**  est la valeur de la série qui sépare cette série en deux groupes de même effectif.

## 2.5 Distribution d'une variable quantitative

Le **troisième quartile**  $Q_3$  de cette série est la valeur qui sépare cette série en deux groupes :

- Le premier groupe contient trois quarts des effectifs (75 %)
- Le deuxième groupe contient un quart des effectifs (25 %)



La **valeur adjacente inférieure** (respectivement supérieure) la plus petite (resp. la plus grande) valeur supérieure (resp. inférieure) à la quantité  $Q_1 - 1.5(Q_3 - Q_1)$  (resp.  $Q_1 + 1.5(Q_3 - Q_1)$ )

$$VAI = \min(x_i \in (x_1, \dots, x_n) : x_i \geq Q_1 - 1.5(Q_3 - Q_1))$$

$$VAS = \max(x_i \in (x_1, \dots, x_n) : x_i \leq Q_3 + 1.5(Q_3 - Q_1))$$

Une **valeur extrême** est une valeur de la série inférieure supérieure) à la valeur adjacente inférieure (resp. supérieure). Attention ! Cette définition n'est pas universelle car la notion de valeur extrême est relativement subjective et dépend du domaine sur lequel porte l'étude statistique.



## 2.5 Distribution d'une variable quantitative



### Exemple 1 – Les notes des élèves

On relève et on ordonne les notes sur 20 d'un DS mathématiques d'un groupe de 10 élèves : 9, 9.5, 10, 10, 12, 13, 15, 18, 18.5, 19.75. La taille de la série est donc  $n=10$ .

Si  $\frac{n}{4}$  est un entier  $p$  alors  $Q_1$  est le terme de rang  $p$  et  $Q_3$  est le terme de rang  $3p$ .

Si  $\frac{n}{4}$  n'est pas un entier,  $Q_1$  est le terme de rang immédiatement supérieur à  $\frac{n}{4}$  et  $Q_3$  est le terme de rang immédiatement supérieur à  $\frac{3n}{4}$ .



Ici  $\frac{n}{4}=2.5$  arrondi à l'entier supérieur vaut 3. Par conséquent  $Q_1$  est la 3<sup>ème</sup> valeur de la série à savoir 10. Cela signifie que 25% des élèves ont obtenu une note inférieure ou égale à 10.

Ici  $\frac{3n}{4}=7.5$  arrondi à l'entier supérieur vaut 8. Par conséquent  $Q_3$  est la 8<sup>ème</sup> valeur de la série à savoir 18. Cela signifie que 75% des élèves ont obtenu une note inférieure ou égale à 18.



## 2.5 Distribution d'une variable quantitative

Si l'effectif total  $n$  est un nombre impair, la médiane est le terme de rang  $\frac{n+1}{2}$ .

Si l'effectif total  $n$  est un nombre pair, la médiane est le centre de l'intervalle formé par les termes de rang  $\frac{n}{2}$  et  $\frac{n}{2} + 1$ .

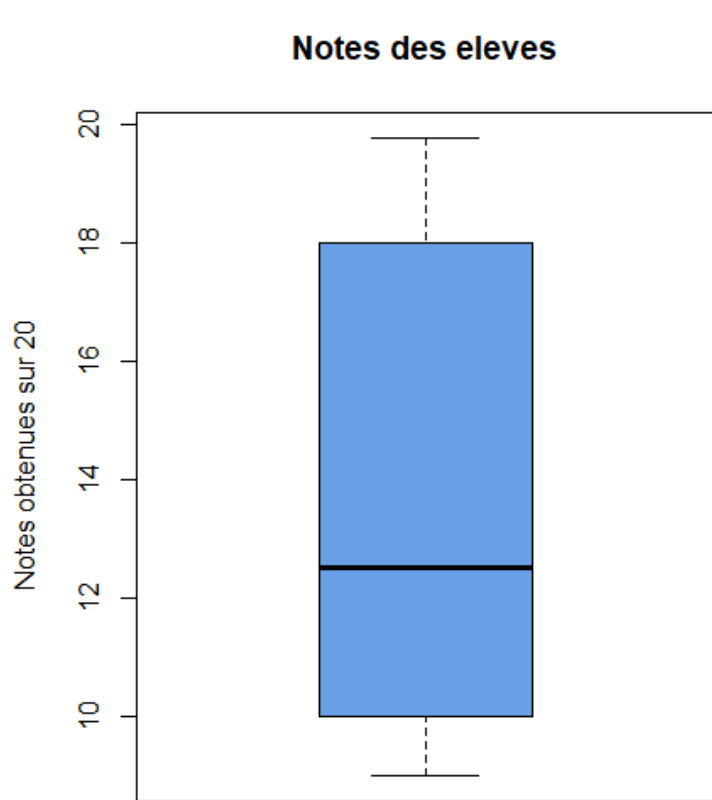


Ici la médiane est le centre de l'intervalle formé par la 5<sup>ème</sup> et la 6<sup>ème</sup> note. Autrement dit, la médiane est égale à la moyenne de la 5<sup>ème</sup> et la 6<sup>ème</sup> note. Ainsi, la médiane vaut 12.5. Cela signifie que 50% des élèves ont obtenu une note inférieure ou égale à 12.5.

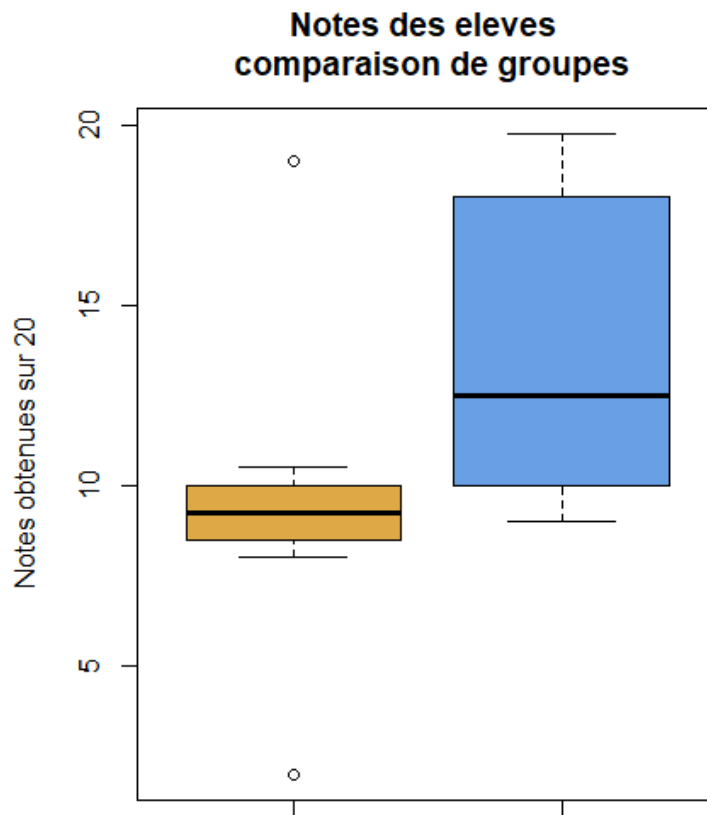
On a  $Q_1 - 1.5(Q_3 - Q_1) = 10 - 1.5(18 - 10) = -2$  Or, aucune valeur de la série est inférieure à cette valeur. Donc il n'y a pas de valeurs extrêmes inférieures.

On a  $Q_3 + 1.5(Q_3 - Q_1) = 18 + 1.5(18 - 10) = 30$  Or, aucune valeur de la série est supérieure à cette valeur. Donc il n'y a pas de valeurs extrêmes supérieures.

## 2.5 Distribution d'une variable quantitative



Sources : Données fictives pour illustrer le cours !



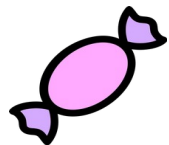
Sources : Données fictives pour illustrer le cours !


Figure 5 - Répartition des notes des élèves.  
Effectifs : 10 notes .  
*Source : Jeu de données fictif.*



## 2.6 Les résumés statistiques

Dans le cas des variables quantitatives, il existe 2 approches complémentaires pour synthétiser l'information. Nous avons abordé jusqu'ici une méthode de représentation basée sur des graphiques et des tableaux. Nous étudierons désormais une méthode analytique fondée sur des **indicateurs statistiques**. Ils sont également appelés **résumés statistiques** ou tout simplement **statistiques**. On distingue les indicateurs de tendance centrales, de position et de dispersion. Les indicateurs sont complémentaires. Par complément, une variable quantitative ne peut pas être résumée à partir d'un seul indicateur dans une étude statistique !



**On appelle statistique une quantité  $T$  fonction de la série statistique  $(x_1, \dots, x_n)$ .** 

Exemple : La moyenne statistique simple

$$T(x_1, \dots, x_n) = \frac{1}{n} \sum_{i=1}^n x_i$$

Dans un sachet de bonbons, on relève le poids (en grammes) des 10 bonbons : 1.3 – 3.4 – 3.2 – 1.8 – 1.2 – 1.7 – 1.5 – 1.3 – 1.1 – 1.5. On note  $s$  cette série statistique.

Un bonbon dans le sachet pèse en moyenne 1.8 grammes :

$$T(s) = \frac{1.3 + 3.4 + 3.2 + \dots + 1.1 + 1.5}{10} = \frac{18}{10} = 1.8 \text{ g}$$

## 2.6 Les résumés statistiques

On dit qu'une statistique est résistante si elle est peu influençable par les valeurs extrêmes.

*La « résistance » ne doit pas être confondu avec la « robustesse » (qui concerne une méthode statistique) même si les 2 termes renvoient à la même notion ! On dit ainsi qu'une méthode statistique est robuste si elle est peu influençable par les valeurs extrêmes. Par exemple, la régression linéaire est une méthode statistique non robuste.*

Cette notion de résistance est très importante ! En effet, les valeurs extrêmes d'une série statistique ont en général une influence non négligeable sur les résultats d'une étude !

### Exemple 1

La médiane est une statistique résistante. En effet, la médiane de la série (1, 5,6) est 5. Si on ajoute deux valeurs,  $10^6$  et  $10^9$ , alors la médiane est 6. En revanche, la moyenne passe de 4 à plus de deux cents millions.



### Exemple 2 : La moyenne statistique simple et la médiane

On relève la masse (en grammes) de 9 chocolats dans une boîte : 9.4 – 9.6 – 9.6 – 9.7 – 9.9 – 10 – 10 – 10 – 10 . On note  $s$  cette série statistique.

$$T(s) = \frac{9.4 + 9.6 + \dots + 10 + 10}{9} = \frac{88.2}{9} = 9.8 \text{ g}$$

$$M(s) = 5^{\text{ème}} \text{ valeur de la série} = 9.9 \text{ g}$$

Un chocolat de la boîte pèse en moyenne 8.69 grammes. Le poids médian d'un chocolat est de 9.9 grammes. Que ce passe-t-il si j'ajoute un chocolat de 1.8g et de 30g ?

$$T(s) = \frac{1.8 + 9.4 + 9.6 + \dots + 10 + 30}{11} = \frac{120}{11} = 10.91 \text{ g}$$

La médiane est inchangée alors que la moyenne a varié !

$$M(s) = 5^{\text{ème}} \text{ valeur de la série} = 9.9 \text{ g}$$

## 2.6 Les résumés statistiques

### 1) Statistiques de tendance centrale

#### a) La moyenne

Il serait plus correct de parler plutôt des moyennes, car il existe plusieurs types de moyenne. Ils découlent de la définition générale de la moyenne appelée aussi moyenne pondérée généralisée.

Soit une série statistique  $x=(x_1,...,x_n)'$  d'une variable d'intérêt  $X$ . Soit  $w=(w_1,...,w_n)'$  des poids relatifs associés au vecteur des valeurs observées.

La moyenne pondérée généralisée d'ordre  $r \in \mathbb{R}_*$  la quantité :

$$\overline{x}_r = \sum_{i=1}^n (w_i x_i)^{\frac{1}{r}} \quad \text{avec} \quad \sum_{i=1}^n w_i = 1$$



Si  $r=1$  alors  $\overline{x}_1 = \sum_{i=1}^n w_i x_i = \langle w | x \rangle$  où  $\langle . | . \rangle$  est

le produit scalaire usuel dans l'espace de

Hilbert  $\mathbb{R}^n$ . C'est la définition de la **moyenne arithmétique pondérée**. Si en plus tous les poids relatifs sont égaux à  $\frac{1}{n}$ , alors c'est la définition de la **moyenne arithmétique simple** !

## 2.6 Les résumés statistiques

La moyenne arithmétique simple vérifie 2 propriétés fondamentales :

1) Elle est le barycentre (ou centre de gravité) des points  $x_i$ . C'est-à-dire :

$$\sum_{i=1}^n (x_i - \bar{x}) = 0$$



2) Elle minimise une quantité appelée risque quadratique.

$$\bar{x} = \underset{a \in \mathbb{R}}{\operatorname{Argmin}} \sum_{i=1}^n (x_i - a)^2 = 0$$



**Exercice : Démontrer les 2 résultats susmentionnés.**

*Exemple prosaïque :* Dans une station de métro, un touriste pressé ignore de quel côté du quai il doit se placer pour être face à la sortie de la station d'arrivée. Sans information a priori, il est clair qu'il devra se situer au milieu du quai. Cette stratégie s'avère la plus efficace s'il devait répéter cette opération plusieurs fois dans la journée.



## 2.6 Les résumés statistiques

La **moyenne géométrique pondérée** est souvent utilisée pour calculer des moyennes de taux de croissance.

Soit le vecteur de  $n$  observations  $x \in \mathbb{R}_*^+$ . On appelle moyenne géométrique pondérée la quantité :

$$\overline{x}_G = \overline{x}_0 = \lim_{r \rightarrow 0} \overline{x}_r = \prod_{i=1}^n x_i^{w_i}$$



*La quantité  $\ln(\overline{x}_0)$  est appelée moyenne arithmétique pondérée des log valeurs de la série.*



Supposons que les taux d'intérêt pour 3 années consécutives soient respectivement de 2, 17, 6 et 22%. Quelle somme obtiendrai-je dans 4 ans si je place aujourd'hui 100 euros ?

Dans 1 ans :  $100 * 1.02 = 102 \text{ euros}$

Dans 2 ans :  $100 * 1.02 * 1.17 = 119.34 \text{ euros}$

Dans 3 ans :  $100 * 1.02 * 1.17 * 1.06 = 126.50 \text{ euros}$

Dans 4 ans :  $100 * 1.02 * 1.17 * 1.06 * 1.22 = 154.33 \text{ euros}$

Moyenne arithmétique des taux :

$$\overline{x}_r = \frac{1.02 + 1.17 + 1.06 + 1.22}{4} = 1.1175$$

Moyenne géométrique des taux :

$$\overline{x}_G = 1.02 * 1.17 * 1.06 * 1.22^{\frac{1}{4}} = 1.114584$$

Appliquons 4 fois les taux trouvés aux 100 euros :


$$100 * 1.1175^4 = 155.9517$$

$$100 * 1.114584^4 = 154.33$$

Le bon taux est bien  $\overline{x}_G$  !

## 2.6 Les résumés statistiques

On appelle **moyenne harmonique pondérée** la quantité :

$$\overline{x_H} = \overline{x_{-1}} = \frac{1}{\left(\sum_{i=1}^n \frac{w_i}{x_i}\right)}$$




Un cycliste parcourt 4 étapes de 100 km. Les vitesses respectives pour ces étapes sont de 10 km/h, 20 km/h, 30 km/h, 40 km/h. Quelle est sa vitesse moyenne ?

Il a parcouru les étapes respectivement en 10h, 5h, 3h20 et 2h30. Il a donc parcouru les 400 km en 10h+5h+3h20+2h30 soit 20.8333h.

Sa vitesse moyenne est donc  $\frac{400}{20.8333} = 19.2 \text{ km/h}$ .

La moyenne arithmétique des vitesses est :

$$\frac{10+20+30+40}{4} = 25 \text{ km/h}.$$

La moyenne harmonique des vitesses est :

$$\frac{4}{\frac{1}{10} + \frac{1}{20} + \frac{1}{30} + \frac{1}{40}} = 19.2 \text{ km/h}.$$

La moyenne harmonique est donc appropriée pour calculer la vitesse moyenne.



## 2.6 Les résumés statistiques

On appelle **moyenne quadratique pondérée** la quantité :

$$\overline{x_Q} = \overline{x_2} = \sqrt{\sum_{i=1}^n w_i x_i^2}$$



*On pourrait assimiler cette quantité à une norme quadratique (avec la métrique diag  $(w_1, \dots, w_n)$ ).*

Cette moyenne est surtout utilisée pour mesurer des écarts de mesure. (voir la partie du cours consacrée à l'écart-type plus loin).

Si on compare les moyennes lorsque tous les  $x_i$  sont strictement positifs, on a :

$$\overline{x_H} \leq \overline{x_G} \leq \overline{x_A} \leq \overline{x_Q}$$



Les quatre moyennes ont des valeurs proches lorsque les valeurs de la série diffèrent peu les unes des autres.

À noter l'existence des **moyennes tronquées** qui sont utilisées lorsque la série statistique comporte quelques valeurs extrêmes qui travestissent la réalité du phénomène étudié. On appelle moyenne tronquée à  $\alpha$  %, la moyenne calculée sur  $(100 - \alpha)\%$  de la taille de la série. La troncature est en générale effectuée en queue de distribution unilatéralement ou bilatéralement (dans ce cas la troncature est symétrique).

## 2.6 Les résumés statistiques

### b) La médiane

La médiane  $Me$  d'une série  $(x_1, \dots, x_n)$  statistique **ordonnée** est la valeur de la série qui divise la population en deux sous – populations de même effectif. **La médiane est une statistique résistante !**

$$Me = \begin{cases} x_{\frac{n+1}{2}} & \text{si } n \text{ est impair} \\ \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} & \text{si } n \text{ est pair} \end{cases}$$

Dans le cas où la série statistique est partitionnée en classes, on parlera non pas de médiane mais de classe médiane (ou intervalle médian). Toutefois, il est possible de déterminer la médiane (appartenant à l'intervalle médian) par **interpolation linéaire**.



### c) Le mode

R

On appelle mode d'une distribution, la valeur du caractère  $X$  dont l'occurrence est la plus élevée. Dans certaines distributions, il peut y avoir plusieurs modes (on parle de distribution pluri-modale).

Si le mode d'une distribution d'un caractère quantitatif discret est aisé à déterminer, celui d'un caractère continu est remplacé par une classe modale ou parfois le centre de cette dernière. Et dans ce cas, le mode dépend complètement du partitionnement en classes.

**Le mode est une statistique résistante !**

## 2.6 Les résumés statistiques

### 2) Statistiques de position

Ils sont également appelés **fractiles**. Ce sont des statistiques de position.



Formellement... soit une variable quantitative  $X$ . On appelle le quantile de  $X$  d'ordre  $\alpha$  ( $\alpha \in [0,1]$ ) la valeur  $q_X(\alpha)$  telle qu'une proportion  $\alpha$  de la population présente une valeur de  $X$  inférieure ou égale à  $q_X(\alpha)$ .

Encore plus formellement... avec  $F_X$  (la fonction de répartition) et  $F_X^{-1}$  (inverse généralisée de  $F_X$ ):

$$q_X(\alpha) = F_X^{-1}(\alpha) = \inf \{x : F_X^{-1}(x) \geq \alpha\}$$



### Exemples

Les **quartiles** sont les 3 valeurs  $Q_{i \in [1,3]}$  qui divisent la population en 4 sous-groupes de taille identique. Ces 3 fractiles sont respectivement d'ordre 0.25, 0.5 et 0.75.  $Q_2$  est également la médiane.

Les **déciles**  $D_{i \in [1,9]}$  sont les quantiles d'ordre 0.1 à 0.9 respectivement qui séparent la population en 10 groupes de même effectif. Le 5<sup>ème</sup> décile est la médiane.

Les **centiles**  $C_{i \in [1,99]}$  sont les fractiles d'ordre 0.01 à 0.99 respectivement qui partitionnent la population en 100 sous-populations de même taille.

## 2.6 Les résumés statistiques

*9 méthodes de calculs existent pour calculer les quantiles. Les logiciels de statistique utilisent la même formule de base pour les calculer mais avec un paramétrage différent !*

*La formule de base pour une série statistique ordonnée est:*

$$q_i(\alpha) = (1 - \mu) x_{(j)} + \mu x_{(j+1)} \quad \text{avec}$$

$$i \in \llbracket 1, 9 \rrbracket \quad (\text{type de quantile})$$

$$\frac{j - m}{n} \leq \alpha \leq \frac{j - m + 1}{n}$$

*j la partie entière de  $n\alpha$   
m est spécifique à chaque type de quantile*



*Par défaut, la fonction quantile de R utilise le type 7 où  $m = 1 - \alpha$  et  $\mu = n\alpha + m - j$*

*Elle utilise l'interpolation linéaire à partir des points  $(p_j, x_{(j)})$  où  $p_j = \frac{j-1}{n-1}$*

## 2.6 Les résumés statistiques

On résume souvent une série statistique par la moyenne. Néanmoins, ce paramètre statistique n'informe en rien sur la dispersion des valeurs autour de la moyenne ! L'utilisation d'indicateurs de dispersion est donc nécessaire.

### 3) Statistiques de dispersion

#### a) L'écart-type et la variance

L'**écart-type** est statistique de dispersion la plus répandue. Elle est définie à partir de la variance pondérée notée  $\sigma_X^2$  ( respectivement  $s_X^2$  ) quand il s'agit d'une population (respectivement d'un échantillon) .

$$\sigma_X^2 = \sum_{i=1}^n w_i (x_i - \bar{x})^2 = \|x - \bar{x} 1_n\|_M^2$$

où  $\bar{x}$  est la moyenne arithmétique pondérée,  
 $1_n = (1, 1, \dots, 1)$  et  $M = \text{diag}(w_i)_i$

avec bien souvent

$$\forall i \in \llbracket 1, n \rrbracket, w_i = \frac{1}{n}$$

Comme la variance n'est pas de même unité que la moyenne, on lui préfère l'écart-type. L'écart-type (« *standard deviation* » en anglais) est la racine carrée de la variance.



## 2.6 Les résumés statistiques

*La variance et donc l'écart-type ne sont pas des statistiques résistantes !*

R

*Dans la plupart des logiciels statistiques, la variance calculée par défaut est la variance*

*corrigée ( $\sigma_X^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ ) au lieu de  $\sigma_X^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$  La raison sera évoquée*

*dans le cours de statistique inférentielle.*

Il n'y a pas de réponse universelle pour interpréter la valeur d'un écart-type. Toutefois, il est clair que la **dispersion sera d'autant plus forte que l'écart-type sera élevé.**



**Exemple 1 :** S'il s'agit de comparer deux distributions (de même moyenne, par exemple 6) d'écarts-types différents (par exemple 11 et 13 respectivement), on peut seulement affirmer que l'une des distributions est plus dispersée que l'autre. Mais, dans cet exemple, la différence n'est pas flagrante.



# Sources et bibliographie

Le cours s'appuie essentiellement sur les références [1] et [2] ci-dessous. Certains passages ont d'ailleurs été conservés ou ré-écrits pour être accessibles aux étudiants dans le cadre du cours. L'ensemble des images du cours proviennent du site internet pixabay.com [\[en ligne\]](#)

[1] **Résumé du Cours de Statistique Descriptive**, Yves Tillé , 18 janvier 2008

[2] **Statistique exploratoire uni et bivariée**, Jocelyn Julienne , Décembre 2011

[3] **Introduction à la programmation en R**, Vincent Goulet [\[en ligne\]](#)

[4] **Introduction au logiciel R** [\[en ligne\]](#)

[5] **Introduction aux graphiques avec R**, Christophe Chesneau [\[en ligne\]](#)

[6] **Graphiques avec ggplot2** [\[en ligne\]](#)

[7] **Manipuler des données avec dplyr** [\[en ligne\]](#)

[8] **Site pédagogique pour les utilisateurs francophones de R** [\[ en ligne\]](#)

[9] **Rapport de corrélation** [\[ en ligne\]](#)

[10] **Analyse bivariée** [\[ en ligne\]](#)

[11] **Cours de Python** [\[en ligne\]](#)

[12] **Apprendre à programmer avec Python 3** [\[en ligne\]](#)

[13] **Programmer en Python**, Valérie Monbet [\[en ligne\]](#)

