Statistiques

Marie Delacre

###### Pagebreak

# Chapitre 1: qu’est-ce que la statistique, et à quoi sert-elle?

## Qu’est-ce que la statistique?

La statistique est l’ensemble des instruments de recherches mathématiques qui permettent de **récolter**, **traiter** et **interpréter** un ensemble de données (généralement vaste).

### La récolte des données: les concepts de population, individus et échantillons

La **population** est l’ensemble des éléments (individus ou objets) auxquels on s’intéresse. Les éléments qui constitue cette population sont appelés **individus** ou **unité statistique**. Admettons par exemple que l’on souhaite étudier les caractéristiques de l’ensemble des employés belges du secteur automobile. Chaque employé de nationalité belge travaillant dans ce secteur constituera une unité statistique, et leur ensemble constituera la population d’intérêt.

Il est bien souvent compliqué, voire impossible de recueillir des informations pour l’entièreté de la population tant celle-ci peut être vaste. Pour cette raison, on prélèvera généralement, aléatoirement, une partie de taille gérable de cette population que l’on appelle l’**échantillon**, et l’on mesurera, pour chaque individu de l’échantillon, un ensemble d’informations (des **variables**) qui nous intéresse, comme le genre, la situation géographique ou linguistique ou encore l’âge et le salaire moyen.

Les variables que l’on récolte au sein de l’échantillon peuvent être de diverses natures. Autrement dit, il existe plusieurs manières de les mesurer. Imaginez par exemple que je vous demande d’estimer la silhouette d’un individu. Vous pourriez prendre un mêtre et me dire que son tour de taille est de 72 cm. Vous pourriez également vous contenter de dire qu’il est de taille “Medium”. Décider de la manière dont on mesure les variables aura des implications importantes en termes de possibilités d’analyses et ne doit pas être laissé au hasard. Un chapitre ultérieur sera dédié aux variables et à leurs natures possibles (chapitre 2).

### Le traitement des données, appelée également la statistique descriptive

Après avoir récolté des données, on se retrouve avec une série de variables pour chaque individu. Ces données pourront être encodées dans une base de données tel qu’illustré à la Table 1.1. Cette table est une représentation fictive de données socio-démographiques récoltées auprès de 15 employés du secteur automobile.

###### Pagebreak

Table 1.1  
*Données socio-démographiques récoltées auprès de 15 employés du secteur automobile*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| i | genre () | âge () | anciennete () | langue () |
| 1 | masculin | 19 | 1 | allemand |
| 2 | masculin | 25 | 5 | français |
| 3 | féminin | 21 | 2 | français |
| 4 | masculin | 27 | 8 | néerlandais |
| 5 | masculin | 30 | 8 | français |
| 6 | masculin | 32 | 9 | néerlandais |
| 7 | féminin | 30 | 9 | allemand |
| 8 | féminin | 19 | 1 | néerlandais |
| 9 | féminin | 21 | 1 | néerlandais |
| 10 | masculin | 21 | 1 | français |
| 11 | masculin | 24 | 8 | français |
| 12 | féminin | 27 | 8 | néerlandais |
| 13 | féminin | 26 | 7 | néerlandais |
| 14 | masculin | 30 | 10 | néerlandais |
| 15 | féminin | 26 | 8 | néerlandais |

Notez que chaque ligne de la Table 1.1 représente un individu de l’échantillon, et que chaque colonne représente une variable. C’est de la sorte que l’on représente généralement les données brutes (soit les données qui ont été récoltées directement auprès des participants) dans les logiciels de traitement de données. Par ailleurs, nous prendrons également l’habitude d’utiliser certaines notations. On symbolise généralement les variables par des lettres majuscules choisies à la fin de l’alphabet (par exemple, X, Y, Z…). On utilise l’indice *i* pour se référer à un individu quelconque de l’échantillon. On symbolise par (X indicé i) la valeur d’un individu i sur la variable X, par (Y indicé i) la valeur d’un individu i sur la variable Y, etc. Par exemple, l’ancienneté du 3ème employé de l’échantillon est symbolisée par = 2, l’âge du 8ème employé de l’échantillon est symbolisée par =19, etc. Gardez tout cela à l’esprit pour la suite de ce cours! Les données brutes sont généralement peu parlantes en elles-mêmes. Afin de pouvoir faire passer un message à travers ces données, il sera nécessaire de les simplifier, de les résumer. C’est l’objectif de la **statistique descriptive**.

L’une des manières de présenter avantageusement les données est de se centrer sur une approche visuelle, soit la réalisation de graphique. Cela fera l’objet du chapitre 3. Au delà de la représentation graphique des données, plusieurs caractéristiques sont importantes à déterminer algébriquement. On peut par exemple se demander quel est l’âge moyen des participants de l’échantillon, ou encore quelle est la proportion de femmes au sein de celui-ci. Cela fera l’objet du chapitre 4.

Nous avons déjà évoqué le fait que la *nature* des variables en jeu aura des implications sur les possibilités d’analyses. Le choix des représentations graphiques et des résumés algébriques que l’on pourra réaliser à l’aide des données dépendront de cette nature!

### L’interprétation des données, appelée également la statistique inférentielle

Il est très important de bien comprendre que la simple description d’un échantillon ne permet pas en tant que telle de tirer des conclusions générales sur l’ensemble de la population. Or, s’il est très intéressant de connaitre les caractéristiques de l’échantillon que l’on étudie, il sera souvent plus intéressant encore de pouvoir inférer les caractéristiques de notre échantillon à la population toute entière. Comment s’assurer que la proportion de femmes observées dans notre échantillon soit bien conforme à la proportion de femmes au sein de la population? Cela recquiert de faire de la **statistique inférentielle**. Cette forme de statistique ne fait pas partie des objectifs de ce cours. Sachez cependant qu’elle existe.

###### Pagebreak

# Chapitre 2: mesure des variables

Mesurer une variable correspond à attribuer un code, le plus souvent chiffré, à chaque individu. Le nom “variable” vient du fait que ce code peut varier d’un individu à l’autre, chaque code possible étant appelé une **modalité** de la variable.

Nous distinguerons les variables **qualitatives** et les variables **quantitatives**.

### Les variables qualitatives

Les variables **qualitatives** sont les variables qui permettent de distinguer les individus sur base de la catégorie à laquelle ils appartiennent.

On parle de variables ***nominales*** lorsqu’il n’est pas possible d’établir un ordre logique croissant ou décroissant entre les modalités de la variable qualitative. Par exemple, la variable “genre”, qui permet de distinguer les hommes des femmes, est de type nominal.

Bien qu’on leur attribue parfois des codes numériques, effectuer des opérations algébriques sur les valeurs de ces variables n’aurait pas de sens. Si je conviens arbitrairement d’attribuer le code “1” à tous les individus du genre masculin, et le code “2” à tous les individus du genre féminin, je ne fais qu’établir une convention sans aucun sens mathématique. Je pourrais tout aussi bien décider que le genre féminin soit codé “1” et que le genre masculin soit codé “4”. Il n’y a aucun lien arithmétique entre ces classes. La valeur “2” n’a par exemple pas le statut de “double de la valeur 1”, elle n’est même pas “plus grande que la valeur 1”, elle est juste différente qualitativement.

On parle de variables ***ordinales*** lorsqu’il est possible d’établir un ordre logique croissant ou décroissant entre les modalités de la variable qualitative. Par exemple, une variable “taille” qui permet de distinguer trois groupes d’enfants suivant qu’ils soient petits, moyens ou grand est de type ordinal, dans la mesure où l’on sait que les grands sont plus grands que les moyens, étant eux-mêmes plus grands que les petits.

Cette fois, si l’on décide d’attribuer des codes numériques à ces variables, il importera que l’ordination des chiffres reflète la relation entre les catégories: les codes attribués aux catégories basses devront être plus petits que ceux attribuées aux catégories plus élevées. Par exemple, on attribuera le code “1” aux petits, “2” aux moyens et “3” aux grands. Pourtant, il n’est toujours pas possible d’effectuer des opérations algébriques sur les valeurs de ces variables. En effet, si l’on peut déduire rapidement sur base des codes qu’un enfant ayant la valeur “3” sur la variable taille est plus grand qu’un enfant ayant la valeur “1”, cela ne signifie pas pour autant qu’il est trois fois plus grand que ce dernier.

## Les variables quantitatives

Les variables **quantitatives** sont les variables dont les valeurs sont intrinsèquement numériques, c’est-à-dire qu’il fait sens d’effectuer des opérations mathématiques sur ses valeurs. Par exemple, si un fermier dispose de deux poules, que la première pond trois oeufs et la deuxième en pond deux, il est correct de dire que le fermier trouvera 5 oeufs lorsqu’il se rendra dans son poulailler.

Les variables ***discrètes*** ne peuvent prendre que des valeurs isolées et généralement entières, dans un intervalle de valeurs spécifiques dans R. Ce sont typiquement les variables de comptage (le nombre d’enfants dans une école, le nombre de filles au sein d’une famille…)

Les variables ***continues***, au contraire, peuvent prendre (au moins théoriquement) n’importe quelle valeur numérique possible (soit une infinité de valeurs), dans un intervalle de valeurs spécifiques dans R. L’âge, les distance, le temps… sont des exemples de variables continues.

###### Pagebreak

# Chapitre 3: exploration graphique des données à une dimension

Les représentations graphiques sont généralement très appréciées, parce que lorsqu’elles sont réalisées judicieusement, elles permettent en un coup d’oeil de se faire une idée de la tendance principale des données.

Il existe une multitude de représentations graphiques si bien qu’il n’est pas toujours évident de savoir quelle représentation choisir. Par ailleurs, certains logiciels proposent des solutions plus ou moins fantaisistes qui peuvent sembler tententes (représentations en 3D, avec couleurs inhabituelles, textures bois…). Je déconseille vivement l’usage de ces artifices. Gardez qu’une bonne présentation est la clé d’une bonne communication. Afin de vous guider vers des choix judicieux, je concluerai ce chapitre sur l’exploration des graphique par quelques conseils qui m’avaient été donnés lors d’un atelier, il y a quelques années, par un brillant scientifique de l’UCL dénommé Christian Ritter.

Pour l’heure, rappelons deux éléments importants à garder à l’esprit lorsqu’on réalise un graphique. Premièrement, toutes les formes de graphiques ne conviennent pas à toutes les formes de données. Le choix du graphique devra dépendre de la nature des variables (on ne représentera pas de la même façon une variable discrète et une variable continue, par exemple). Il faudra donc bien comprendre la nature des variables que l’on souhaite représenter. Deuxièmement, un graphique constitue une représentation **simplifiée** des données. Cela signifie qu’on ne va pas y représenter directement les données brutes. Après avoir identifié la nature de la variable, la réalisation du graphique se déroulera donc en deux étapes: 1) simplification des données et 2) réalisation du graphique en tant que tel.

Nous développerons deux catégories de graphiques: ceux qui sont construits sur base de **tableaux de fréquence** et ceux qui reposent sur le calcul de **quantiles**.

## Les graphiques construits sur base des tableaux de fréquences

### Variables qualitatives

#### Simplification des données

Imaginez que 40 participants aient répondu par “oui” ou “non” à la question suivante “Etes vous fumeur?” et que vous vous retrouvez avec l’ensemble de réponse suivant:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| non | non | non | oui | non | non | oui | non | non | non |
| oui | non | oui | oui | oui | non | non | oui | non | non |
| non | non | oui | oui | oui | non | non | non | non | non |
| non | non | non | non | non | non | non | non | non | non |

C’est ce qu’on appelle les données brutes. Une manière de résumer cet ensemble d’information est de calculer un **tableau de fréquences**, tel que représenté dans la Table 3.1.

Table 3.1 *Transformation des données brutes relatives aux fumeurs en tableau de fréquence*

|  |  |  |
| --- | --- | --- |
| fumeur ( |  |  |
| non | 30 | 0.75 |
| oui | 10 | 0.25 |

Il s’agit d’un tableau dans lequel chaque ligne représente une valeur possible de la variable. Comme vous pouvez l’observer, deux lignes ont été envisagées (une pour les fumeurs, une pour les non fumeurs) et le nombre de personnes appartenant à chaque catégorie constitue ce que l’on appelle la **fréquence absolue**. La somme des fréquences absolues correspond au nombre total de personnes constituant l’échantillon (ici, n vaut 40).

La colonne suivante contient la même information mais par rapport à l’ensemble des ménages, c’est la **fréquence relative**.

Par exemple, 30 personnes de l’échantillon sur 40 ne fument pas, or 30/40 x 100 = 75 % (75% des personnes ne notre échantillon ne fument pas).

De même que pour les données brutes, nous prendrons l’habitude d’utiliser certaines notations. Dans un tableau de fréquence, on symbolise généralement l’ensemble des valeurs que peut prendre la variable par une lettre minuscule (Par exemple, les différentes valeurs que peut prendre la vaiable X seront symboliées par x). On utilise l’indice *j* pour se référer à chaque valeur possible, j pouvant varier de 1 à k (dans le cas présent, k = 2). On symbolise par (x indicé j) la jème valeur possible de la variable X. Par exemple, si l’on choisit de nommer X la variable “fumer”, . correspond aux fréquences absolues associées à chaque valeur . Par exemple, = 30. Enfin, correspond aux fréquences relatives. Par exemple = 0.25.

Notons que quand on étudie des variables nominales, peu importe l’ordre d’apparition des modalités de la variable dans le tableau de fréquence. Dans l’exemple, nous avons représenté d’abord les non fumeurs, suivi des fumeurs. L’inverse aurait pu être fait également sans pour autant altérer le contenu informatif du tableau.

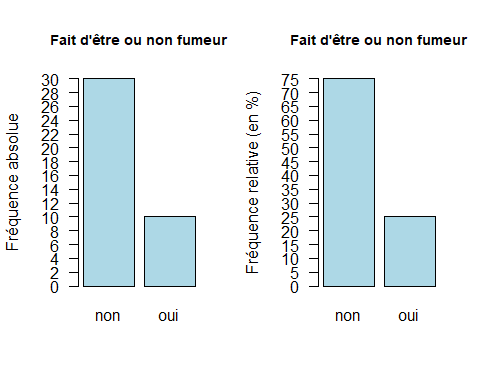
#### Réalisation graphique

##### Diagramme en bâtons

Le diagramme en bâtons consiste à représenter en abcisse les différentes valeurs que peut prendre la variable que l’on étudie() et en ordonnée les fréquences associées à chaque valeur (soit , soit ). Pour chaque valeur , on trace un bâtonnet dont la hauteur varie en fonction de la fréquence associée à cette valeur .

L’ordre dans lequel les différents bâtonnets importe peu lorsqu’on représente des variables nominales. De plus, étant donné que l’axe des abcisses n’est pas normé, la largeur des bâtons importe peu également (même si pour le confort des yeux, on les gardera généralement uniformes).

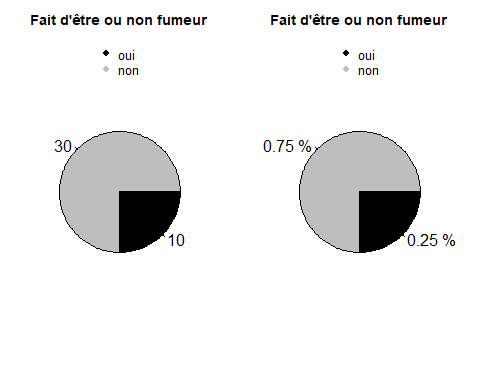
La figure Figure 3.1 montre deux diagrammes en bâtons de notre distribution. Que l’on représente les fréquences absolues ou les fréquences relatives sur l’axe des ordonnées importe peu en termes de forme de graphe. En revanche, l’information donnée est un petit peu différente: dans le premier cas, on peut dénombrer le nombre de fumeur et de non fumeur, dans le second cas, on peut déterminer le pourcentage de personnse qui fument ou ne fument pas dans l’échantillon.

Figure 3.1  
*Diagramme en barre de la distribution de la Table 3.1* 

##### Diagramme circulaire

On trouve occasionnellement dans les rapports des diagrammes circulaires (qu’on appelle aussi graphiques en camembert) pour représenter les variables. Réaliser un diagramme circulaire consiste à subdiviser un cercle en autant de part qu’il n’y a de modalités possibles, chaque part ayant une taille proportionnelle à la fréquence associée à la modalité qu’elle représente.

La Figure 3.2 montre deux diagrammes circulaires de notre distribution. De même que pour les diagrammes en bâtons, que ce soient les fréquences absolues ou les fréquences relatives importe peu de termes de forme de graphe, mais l’informatin donnée varie légèrement: dans le premier cas, on peut dénombrer le nombre de fumeur et de non fumeur, dans le second cas, on peut déterminer le pourcentage de personnse qui fument ou ne fument pas dans l’échantillon.

Figure 3.2 *Diagramme circulaire de la distribution de la Table 3.3* 

Déterminer l’angle pour chaque catégorie peut se faire en procédant à une règle de trois. Un cercle fait exactement 360° et doit représenter l’ensemble des individus constituant l’échantillon, soit 40. Ces 360° doivent être répartis entre les catégories, de manière proportionnelle à la fréquence de chaque catégorie. Par exemple, 30 personnes ne fument pas (valeur = non). Cela représente 75% de l’échantillon, l’angle correspondra donc à 75% des 360°.

Naturellement, dans la mesure où les parts sont de taille proportionnelle à la fréquence associée à chaque catégorie, déterminer l’angle pour chaque catégorie en utilisant les fréquences relatives, plutét que les fréquences absolues, fournit exactement le même résultat, à condition d’utiliser les valeurs exactes (non arrondies). La preuve:

Bien que ça ne soit pas le cas dans cet exemple, les fréquences relatives contiendront parfois un nombre important (voire infini) de décimales. Il est dès lors généralement plus facile de déterminer les angles de chaque part sur base des fréquences abolues plutôt que sur base des fréquences relatives. La Table 3.4 fournit les angles qui devraient être associés à chaque catégorie dans l’exemple de la Table 3.1:

Table 3.4  
*Angle de chaque part du diagramme circulaire représententant les données de la Table 3.1*

|  |  |  |
| --- | --- | --- |
|  |  | angle (en °) |
| non | 30 | 270 |
| oui | 10 | 90 |

Ce type de graphique présente de nombreux inconvénients en termes de communication. Entre autres, ils sont peu précis: lorsque deux catégories sont associées à des fréquences relativement similaires, il devient très difficile de repérer leur différence. Cela sera illustré dans la section dédiée aux diagrammes circulaires pour représenter des variables quantitatives discrètes.

### Variables quantitatives discrètes

Table 3.2  
*Données brutes relatives au nombre d’enfants par ménages, au sein de 40 ménages*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| i | Enfants () | i | Enfants () | i | Enfants () | i | Enfants () |
| 1 | 5 | 11 | 1 | 21 | 1 | 31 | 1 |
| 2 | 2 | 12 | 3 | 22 | 1 | 32 | 1 |
| 3 | 2 | 13 | 3 | 23 | 2 | 33 | 2 |
| 4 | 3 | 14 | 1 | 24 | 2 | 34 | 2 |
| 5 | 1 | 15 | 4 | 25 | 1 | 35 | 1 |
| 6 | 2 | 16 | 3 | 26 | 2 | 36 | 2 |
| 7 | 1 | 17 | 3 | 27 | 3 | 37 | 3 |
| 8 | 2 | 18 | 1 | 28 | 1 | 38 | 1 |
| 9 | 2 | 19 | 2 | 29 | 3 | 39 | 3 |
| 10 | 2 | 20 | 1 | 30 | 4 | 40 | 4 |

#### Simplification des données: transformation en tableaux de fréquences

La Table 3.2 répertorie le nombre d’enfants vivant au sein de 40 ménages fictifs. Notez que j’ai représenté ces données brutes sur plusieurs colonnes distinctes par soucis d’économie de place. Cependant, dans un logiciel (excel, ,par exemple), il conviendrait de n’encoder que deux colonnes: une première relative au numéro (ou id) attribué à chaque sujet, et une deuxième relative à la variable “nombre d’enfants”. Rappelez-vous: une ligne par sujet, et une colonne par variable!

La Table 3.3 constitue le tableau de fréquence de ces données.

###### Pagebreak

Table 3.3  
*Transformation de la Table 3.2 en tableau de fréquence*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| nombre d’enfants () |  | (en %) |  | (en %) |
| 1 | 14 | 35.0 | 14 | 35.0 |
| 2 | 13 | 32.5 | 27 | 67.5 |
| 3 | 9 | 22.5 | 36 | 90.0 |
| 4 | 3 | 7.5 | 39 | 97.5 |
| 5 | 1 | 2.5 | 40 | 100.0 |

Comme vous pouvez l’observer, de même que fait précédemment pour les variables nominales, les nombres possibles d’enfants par ménage ont été envisagés et on leur associe une fréquence absolue (), et une fréquence relative (). Par ailleurs, la somme des fréquences absolues correspond toujours au nombre total de ménages constituant l’échantillon (ici, n vaut 40).

Contrairement au cas des variables nominales, l’ordre de présentation de chaque ligne a une importance. On triera généralement les valeurs possibles de la variable représentée par ordre croissant (1 apparait avant 2, 2 apparait avant 3, etc.).

De plus, on pourra calculer ce qu’on appelle les **fréquences cumulées**, soit le cumul de plusieurs fréquences.

Les fréquences absolues cumulées sont la somme de plusieurs fréquences absolues (notation: ).

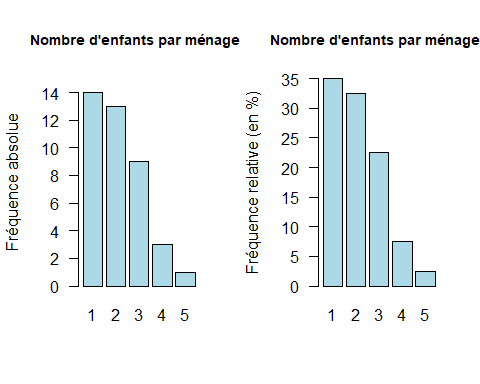
Cela permet de déterminer rapidement, par exemple, que 39 ménages sont constitués de maximum 4 enfants. Enfin, cette information peut être exprimée par rapport à l’ensemble des ménages. On dira alors que 97.5 % des ménages sont constitués de maximum 4 enfants. Cette dernière expression correspond aux **fréquences relatives cumulées**, soit la dernière colonne du tableau (notation: ).

#### Représentation graphique

##### Diagramme en bâtons

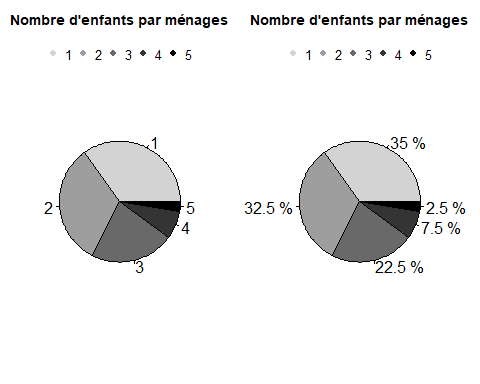
Dans la mesure où il existe un ordre logique entre les différentes valeurs de la variable représentée, l’ordre des bâtonnets importe (contrairement au cas des variables nominales).

La Figure 3.3 montre deux diagrammes en bâtons de notre distribution. On constate à nouveau que l’allure générale du graphique est la même, que l’on représente les fréquences absolues ou relatives sur le diagramme. En revanche, l’information donnée est un petit peu différente: dans le premier cas, on peut dénombrer le nombre de ménages qui ont un nombre donné d’enfants, dans le second cas, on peut déterminer le pourcentage de ménages qui ont ce nombre d’enfants.

Figure 3.3  
*Diagramme en barre de la distribution de la Table 3.3* 

##### Diagramme circulaire

La Figure 3.4 montre deux diagrammes circulaires de notre distribution: l’un qui représente les fréquences absolues, et l’autre qui représente les fréquences relatives.

Figure 3.4 *Diagramme circulaire de la distribution de la Table 3.3* 

La Table 3.5 fournit les angles qui devraient être associés à chaque catégorie dans l’exemple de la Table 3.3. Ceux-ci ont été déterminés en procédant à une règle de trois, comme expliqué dans la section dédiée aux diagrammes circulaires pour représenter des variables nominales.

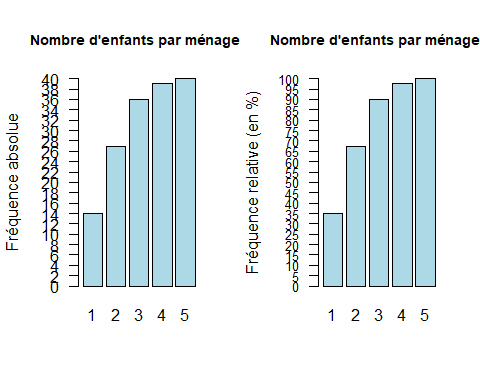
Table 3.5  
*Angle de chaque part du diagramme circulaire représententant les données de la Table 3.3*

|  |  |  |
| --- | --- | --- |
|  |  | angle (en °) |
| 1 | 14 | 126 |
| 2 | 13 | 117 |
| 3 | 9 | 81 |
| 4 | 3 | 27 |
| 5 | 1 | 9 |

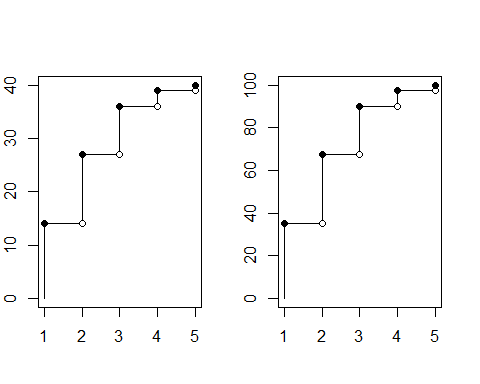
Il a déjà été précisé ce type de graphique est peu précis et que lorsque deux catégories sont associées à des fréquences relativement similaires, il devient très difficile de repérer leur différence. Dans la Figure 3.4, les deux premières catégories semblent de taille égale, alors qu’elles ne le sont pas. Il est nettement plus aisé de le voir dans la Figure 3.3. C’est dû au fait que notre oeil repère plus facilement une différence d’hauteur entre deux objets alignés (deux rectangles du diagramme en bâtons) qu’une différence d’aire dans un cercle. Pour cette raison, je vous recommanderai de privilégier autant que possible le diagramme en bâtons. La simplicité est la clé d’un bon graphique!

##### Diagramme des fréquences cumulées

Le diagramme des fréquences cumulées consiste à représenter les différentes valeurs que peut prendre la variable à représenter par des bâtonnets dont la hauteur varie en fonction non plus de la fréquence (absolue ou relative) mais en fonction de la fréquence **cumulée** associée à cette valeur. Le dernier bâton du diagramme des fréquences cumulées aura nécessairement une hauteur égale à n si l’on représente les fréquences absolues, et une auteur égale à 1 si l’on représente les fréquences relatives. Une fois encore, l’allure générale du graphique est la même, que l’on représente les fréquences absolues ou relatives sur le diagramme. En revanche, l’information donnée est un petit peu différente: dans le premier cas, on peut dénombrer le nombre de ménages dont le nombre d’enfants atteint au maximum la valeur représentée par le bâtonnet, dans le second cas, on peut déterminer le pourcentage de ménages dont le nombre d’enfants atteint au maximum la valeur représentée par le bâtonnet.

Figure 3.5  
*Diagramme des fréquences cumulées de la distribution de la Table 3.3* 

Il arrive que l’on parle de “diagramme en escalier”, et que l’on choisisse plutôt ce mode de représentation, dans lequel un point noir indique la fréquence associée à une catégorie en particulier. L’information fournie par les diagrammes en escalier et les diagrammes des fréquences cumulées est identique.

Figure 3.6  
*Diagramme en escalier représentant les fréquences cumulées de la distribution de la Table 3.3* 

### Variables quantitatives continues

#### Simplification des données

Dans la mesure où les variables ***continues***, il n’est pas possible d’associer une fréquence à chaque valeur possible de la variable, similairement à ce que l’on fait pour des variables discrètes. On procédera alors à des regroupements des valeurs en classes (d’amplitude à déterminer), et on déterminera les fréquences associées à ces classes.

Il existe de nombreux critères pour déterminer le nombre de classes. De manière générale, plus on aura de classes, plus grande sera la précision. Lorsque l’on doit effectuer les calculs manuellement, on se contentera généralement d’une dizaine de classes. Le choix peut également être guidé par des raisons théoriques(catégories pré-existantes dans la nature). Lorsqu’on utilise des logiciels informatiques prévus à cet effet, on peut par contre travailler avec beaucoup plus de classes (soit des classes ayant une amplitude nettement plus faible), ce qui permettra de se faire une idée bien plus précise de l’allure des données, comme nous y reviendrons à la fin de cette sous-section (voir la partie sur les distributions statistiques).

Lorsque c’est possible, il est préférable d’utiliser des classes ayant toutes la même amplitude, tel que dans la Table 3.6 et la Table 3.7, notamment pour des raisons de lisibilité graphique. Cependant, il peut arriver que pour des raisons théoriques, nous soyons amenés à envisager des classes de tailles différentes. Par exemple, si l’on veut distinguer des caractéristiques de la petite enfance, enfance, adolescence, âge adulte et troisième âge, on pourra créer 5 classes: une regroupant les individus de 0 à 3 ans, de ceux de 3 à 12 ans , de 12 à 18 ans, 18 à 65 ans et de plus de 65 ans (voir la Table 3.8).

Enfin, le **centre de classe** est la valeur qui se trouve à égale distance de la borne inférieure et de la borne supérieure d’une classe. Connaître cette valeur servira à estimer les **quantiles** de variables continues (ce concept sera très prochainement expliqué).

Table 3.6  
*Table des PIB par habitants de 45 pays d’Europe*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Classe | Centre de classe |  | (en%) |  | (en %) |
| [0-20000[ | 10000 | 20 | 100 | 20 | 44.44 |
| [20000-40000[ | 30000 | 9 | 100 | 29 | 64.44 |
| [40000-60000[ | 50000 | 11 | 100 | 40 | 88.89 |
| [60000-80000[ | 70000 | 3 | 100 | 43 | 95.56 |
| [80000-1e+05[ | 90000 | 1 | 100 | 44 | 97.78 |
| [1e+05-120000[ | 110000 | 1 | 100 | 45 | 100.00 |

Source: <https://fr.tradingeconomics.com/country-list/gdp-per-capitaécontinent=europe>

Table 3.7  
*Table de fréquence de catégories d’âges dans un échantillon de 20000 personnes*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Classe | Centre |  | (en%) |  | (en %) |
| [0-10[ | 5 | 2994 | 14.97 | 2994 | 14.97 |
| [10-20[ | 15 | 3006 | 15.03 | 6000 | 30.00 |
| [20-30[ | 25 | 3455 | 17.27 | 9455 | 47.27 |
| [30-40[ | 35 | 2304 | 11.52 | 11759 | 58.80 |
| [40-50[ | 45 | 1732 | 8.66 | 13491 | 67.45 |
| [50-60[ | 55 | 2016 | 10.08 | 15507 | 77.53 |
| [60-70[ | 65 | 1794 | 8.97 | 17301 | 86.50 |
| [70-80[ | 75 | 699 | 3.50 | 18000 | 90.00 |
| [80-90[ | 85 | 1051 | 5.25 | 19051 | 95.25 |
| [90-100[ | 95 | 749 | 3.74 | 19800 | 99.00 |
| [100-110[ | 105 | 200 | 1.00 | 20000 | 100.00 |

Table 3.8  
*Table de fréquence de catégories d’âges dans un échantillon de 20000 personnes*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Classe | Centre |  | (en%) |  | (en %) |
| [0-3[ | 1.5 | 1450 | 7.25 | 1450 | 7.25 |
| [3-12[ | 7.5 | 2001 | 10.01 | 3451 | 17.25 |
| [12-18[ | 15.0 | 1999 | 9.99 | 5450 | 27.25 |
| [18-65[ | 41.5 | 11501 | 57.50 | 16951 | 84.75 |
| [65-110[ | 87.5 | 3049 | 15.25 | 20000 | 100.00 |

#### Réalisation graphique

##### Histogramme

Le graphique le plus couramment utilisé pour représenter des variables continues est l’histogramme. Un histogramme ressemble à un diagramme en bâtons, mais au lieu d’avoir des bâtons isolés qui représentent une valeur unique, on a des rectangles, collés les uns aux autres (pour rendre compte du caractère continu de la variable) qui représentent une classe de valeurs.

Les classes de valeurs sont représentées en abcisse. Lorsque toutes les classes sont de taille égale, la hauteur des rectangles est proportionnelle aux fréquences (absolues ou relatives; les fréquences correspondent à l’ordonnée), tel qu’on peut le voir dans la Figure 3.3 ainsi que dans la Figure 3.8 .

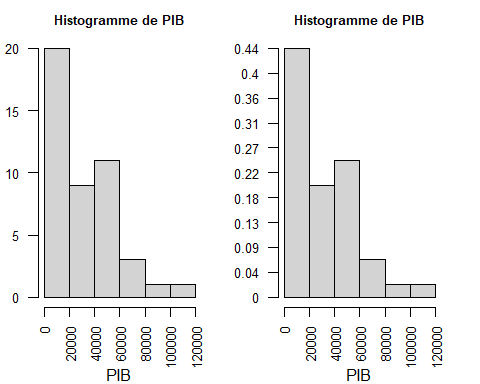
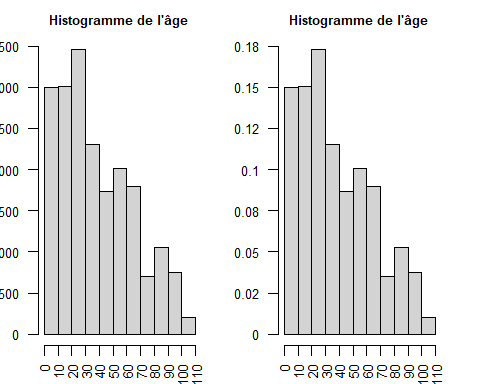
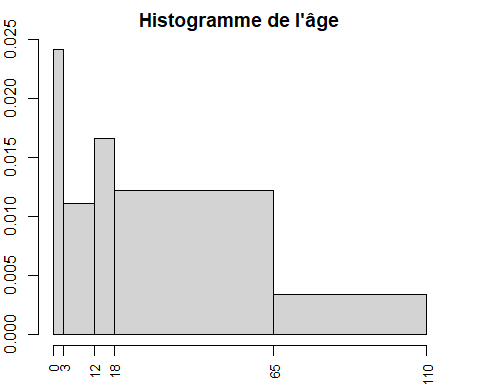
Figure 3.7  
*Histogramme de la distribution de la Table 3.6* 

Figure 3.8  
*Histogramme de la distribution de la Table 3.7* 

Par contre, lorsque les classes sont de tailles inégales, la hauteur du rectancle est proportionnelle à la densité de la classe (soit l’effectif divisé par l’amplitude de la classe), ce qui rend la comparaison des fréquences des classes plus compliquée.

Figure 3.9 *Histogramme de la distribution de la Table 3.8* 

Dans la mesure où les données sont continues, cela signifie que toutes les valeurs de la variable représentée sont considérées comme possibles. Par exemple, le premier rectangle de la Figure 3.8 englobe toutes les personnes dont l’âge est compris entre 0 et 10 ans (incluant les personnes ayant 2,38574 ans), et lorsque je me déplace du premier centre de classe vers le deuxième centre de classe, j’envisage n’importe quelle valeur comprise entre 5 et 15. Pour mieux rendre compte de la continuité de la variable, on peut tracer ce qu’on appelle le **polygone des effectifs**, en reliant par des segments de droite le centre du côté supérieur de chaque rectangle.

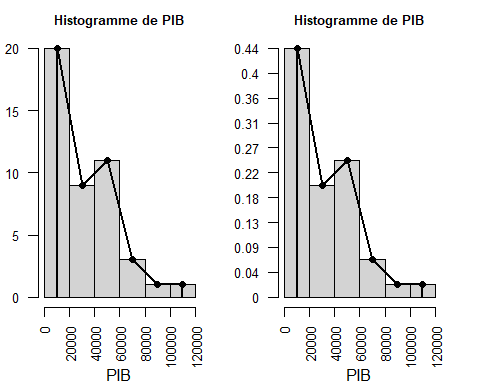
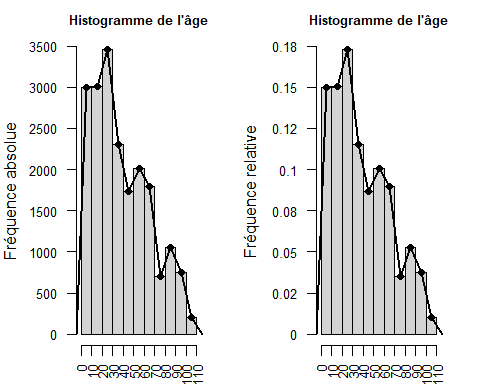
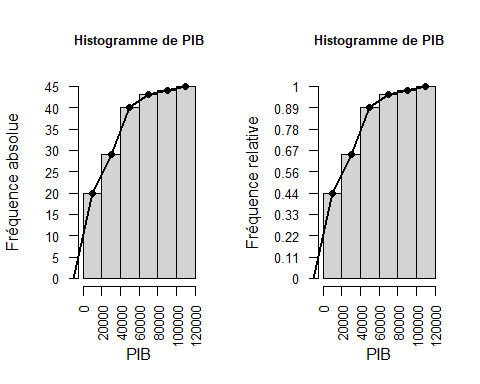
Figure 3.10  
*Histogramme de la distribution de la Table 3.6 et polygone des effectifs* 

Figure 3.12  
*Histogramme de la distribution de la Table 3.7 et polygone des effectifs* 

##### Histogramme des effectifs (ou fréquences) cumulé(e)s

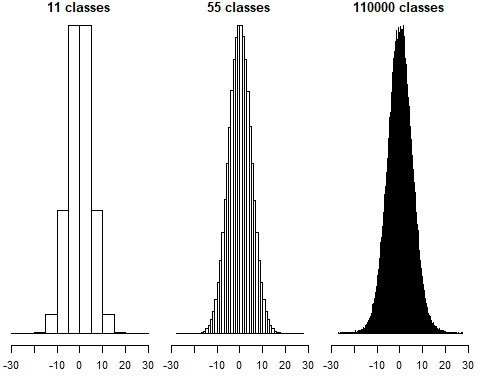
Les histogrammes des effectifs (ou fréquences) cumulées ressemblent aux diagrammes des fréquences cumulées, si ce n’est que cette fois, les différents bâtons sont collés les uns aux autres pour rendre compte du caractère continu de la variable.

De même que pour les histogrammes vu dans la section précédente, il est possible de tracer le polygone des effectifs (ou fréquences) cumulé(e)s pour mieux rendre compte du caractère continu de la variable.

Figure 3.10  
*Histogramme de la distribution de la Table 3.6 et polygone des effectifs* 

#### Remarque: les distributions

Il a été brièvement mentionné que lorsque le nombre de classe augmente (ce qui revient à dire que l’amplitude de chaque classe diminue), l’allure générale des données se dessine avec plus de précision. Cela est illustré dans la Figure 3.11, où les mêmes données sont représentées à l’aide de trois histogrammes, dont le nombre de classes varie (11 classes dans l’histogramme de gauche,55 classes dans l’histogramme du milieu et 1.110^{5} classes dans l’histogramme de droite). Au fur et à mesure que le nombre de classes augmentent, l’allure de l’histogramme se rapproche d’une courbe, jusqu’à se lisser parfaitement. Cette courbe lisse représente ce qu’on appelle la distribution des données.

Figure 3.11  
*De l’histogramme vers la distribution de données* 

## Les graphiques construits sur base des quantiles

Outre les tableaux de fréquence, une autre manière utile de décrire une série de données ou une distribution est de la diviser en un certain nombre d’intervalles contenant tous le même nombre d’observations. Les bornes de ces intervalles sont appelés des **quantiles**. Contrairement aux tableaux de fréquence, le calcul des quantiles n’a de sens que pour décrire des variables **quantitatives** (discrètes ou continues).

Certains quantiles sont particulièrement connus, on retrouve parmi eux:

* La **Médiane** qui découpe la distribution en deux parties contenant chacune 50% des observations. Ce quantile, très important, donne une mesure de ce qu’on appelle la **tendance centrale** (cette notion sera développée au chapitre suivant).
* Les **Quartiles** qui découpent la distribution en 4 parties contenant chacune 25% des observations.
* Les **Déciles** qui découpent la distribution en 10 parties contenant chacune 10% des observations
* Les **Percentiles** qui découpent la distribution en 100 parties contenant chacune 1% des observations.

Remarquez que le 25ème percentile est le premier quartile, le cinquantième percentile est le cinquième décile ou le deuxième quartile ou encore la médiane, etc.

Bien que le principe soit identique quel que soit le quantile envisagé, nous ne développerons que la **médiane** et les **quartiles**, parce qu’ils sont nécessaires à la réalisation de boîtes à moustaches que nous élaborerons ensuite. Par ailleurs, on ne calculera pas ces quantiles exactement de la même manière suivant que les données soit discrètes ou continues.

### Variables quantitatives discrètes

#### Simplification des données: calcul de la médiane et des quartiles

##### Au départ de données brutes

Soit une série statistique contenant un nombre impair de sujets (n=9):

La médiane va consister à déterminer une valeur telle qu’il y ait la moitié des observations de la série dont la valeur lui est inférieure, et l’autre moitié dont la valeur lui est supérieure. Dans un premier temps, il faudra ordonner la série par ordre croissant. Ensuite, on pourra déterminer la position à laquelle se trouve la médiane. Il s’agira de la valeur qui occupe le rang .

Voici la série ordonnée:

.

La médiane sera l’observation qui occupera le rang (soit la observation) de la série ordonnée. Or, la observation de la série ordonnée a pour valeur 8. La médiane vaut donc 8.

Attention: il ne faut pas confondre le **rang médian** (soit la position occupée par la médiane dans la série) et la **médiane** (la valeur de l’observation située au rang médian).

Soit une autre série statistique, contenant cette fois un nombre pair de sujets (n=10):

On commence par ordonnée la série:

On calcule ensuite le rang médian: . Cette fois, le rang n’est pas un nombre rond. On en déduit que la médiane sera la valeur entre la observation (valant 8) et la observation de la série (valant 9). Par convention, on prend la moyenne de ces deux valeurs, ce qui donne 8.5.

Comme ces deux exemples permettent de l’illustrer, avec un nombre impair d’observations, la valeur de la médiane correspond nécessairement à une valeur observée de la variable. Par contre, avec un nombre pair d’observations, la valeur de la médiane ne correspond “pas nécessairement” à une valeur observée de la médiane. Attention, pas nécessairement ne veut pas dire jamais: imaginons que les deux valeurs de part et d’autre du rang moyen soient identiques, comme dans la série ordonnée suivante:

.  
Dans ce cas, étant donné que le rang médian vaut 5.5, la médiane sera la moyenne entre 8 et 8, soit 8.

La valeur du premier quartile d’une série sera celle qui est telle que si l’on s’intéresse exclusivemenet aux valeurs à gauche de la médiane, la moitié des observations sélectionnées lui est inférieure et l’autre moitié lui est supérieure. Similairement, la valeur du troisième quartile sera celle qui est telle que si l’on s’intéresse exclusivement aux valeurs à droite de la médiane, la moitié des observations sélectionnées lui est inférieure et l’autre moitié lui est supérieure.

Par exemple, si l’on revient à la série 1, soit la série suivante:

* La partie des données à gauche de la médiane est la suivante: 1, 2, 3, 7

La valeur du premier quartile sera donc égal à 2.5.

* La partie à droite de la médiane est la partie suivante: 9, 11, 13, 15

La valeur du troisième quartile sera donc égal à 12.

Si l’on revient à la série 2, soit la série suivante:

* La partie des données à gauche de la médiane est la suivante: 2, 3, 6, 7, 8

La valeur du premier quartile sera donc égal à 6.

* La partie à droite de la médiane est la partie suivante: 9, 11, 13, 14, 15

La valeur du troisième quartile sera donc égal à 13.

Enfin, si l’on revient à la série 3, soit la série suivante:

* La partie des données à gauche de la médiane est la suivante: 2, 3, 6, 7, 8

La valeur du premier quartile sera donc égal à 6.

* La partie à droite de la médiane est la partie suivante: 8, 11, 13, 14, 15

La valeur du troisième quartile sera donc égal à 13.

##### Au départ de tableaux de fréquences

La Table 3.11 est le tableau de fréquence des données de la série 1.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
| 1 | 1 | 0.11 | 1 | 0.11 |
| 2 | 1 | 0.11 | 2 | 0.22 |
| 3 | 1 | 0.11 | 3 | 0.33 |
| 7 | 1 | 0.11 | 4 | 0.44 |
| 8 | 1 | 0.11 | 5 | 0.56 |
| 9 | 1 | 0.11 | 6 | 0.67 |
| 11 | 1 | 0.11 | 7 | 0.78 |
| 13 | 1 | 0.11 | 8 | 0.89 |
| 15 | 1 | 0.11 | 9 | 1.00 |

La fréquence cumulée associée à la valeur 7 vaut 0.44. Dans la mesure où la médiane est la valeur telle qu’il y ait 50% des observations à gauche de celle-ci, elle sera nécessairement supérieure à 7. La fréquence cumulée associée à la valeur 8 vaut 0.56. La médiane vaudra donc 8.

En suivant un raisonnement identique:

* la fréquence cumulée associée à la valeur 2 vaut 0.22. La fréquence cumulée associée à la valeur 3 vaut 0.33. La médiane vaudra donc 3.
* la fréquence cumulée associée à la valeur 9 vaut 0.67. La fréquence cumulée associée à la valeur 11 vaut 0.78. La médiane vaudra donc 11.

La Table 3.13 est le tableau de fréquence des données de la série 3.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
| 2 | 1 | 0.1 | 1 | 0.1 |
| 3 | 1 | 0.1 | 2 | 0.2 |
| 6 | 1 | 0.1 | 3 | 0.3 |
| 7 | 1 | 0.1 | 4 | 0.4 |
| 8 | 2 | 0.2 | 6 | 0.6 |
| 11 | 1 | 0.1 | 7 | 0.7 |
| 13 | 1 | 0.1 | 8 | 0.8 |
| 14 | 1 | 0.1 | 9 | 0.9 |
| 15 | 1 | 0.1 | 10 | 1.0 |

La fréquence cumulée associée à la valeur 7 vaut 0.4. Dans la mesure où la médiane est la valeur telle qu’il y ait 50% des observations à gauche de celle-ci, elle sera nécessairement supérieure à 7. La fréquence cumulée associée à la valeur 8 vaut 0.6. La médiane vaudra donc 8.

En suivant un raisonnement identique:

* la fréquence cumulée associée à la valeur 3 vaut 0.2. La fréquence cumulée associée à la valeur 6 vaut 0.3. La médiane vaudra donc 6.
* la fréquence cumulée associée à la valeur 11 vaut 0.7. La fréquence cumulée associée à la valeur 13 vaut 0.8. La médiane vaudra donc 13.

Enfin, la Table 3.12 est le tableau de fréquence des données de la série 2.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
| 2 | 1 | 0.1 | 1 | 0.1 |
| 3 | 1 | 0.1 | 2 | 0.2 |
| 6 | 1 | 0.1 | 3 | 0.3 |
| 7 | 1 | 0.1 | 4 | 0.4 |
| 8 | 1 | 0.1 | 5 | 0.5 |
| 9 | 1 | 0.1 | 6 | 0.6 |
| 11 | 1 | 0.1 | 7 | 0.7 |
| 13 | 1 | 0.1 | 8 | 0.8 |
| 14 | 1 | 0.1 | 9 | 0.9 |
| 15 | 1 | 0.1 | 10 | 1.0 |

Ici, on constate que la fréquence cumulée associée à la valeur 8 vaut exactement 0.5. Dans ce cas particulier, par convention, on estimera que la médiane est la moyenne entre cette valeur associée à une fréquence cumulée d’exactement .5 et la valeur suivante de la série. Dans le cas présent, la médiane sera donc la moyenne entre 8 et 9, soit 8.5. Vérifiez la valeur de la médiane que nous avions trouvé en analysant les données brutes de la série 3, vous verrez bien qu’il s’agit bien de la valeur que nous avions trouvé.

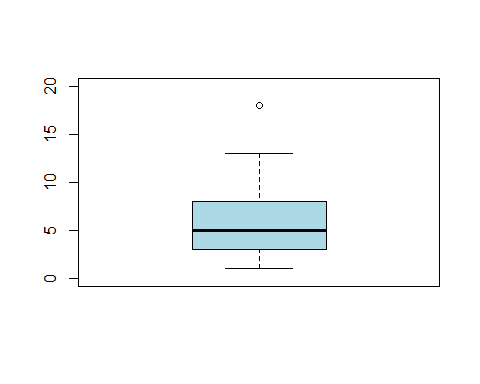
La valeur des premier et troisième quartiles correspondent aux premières lignes de fréquences cumulées qui dépassent respectivement .25 et .75, soit 6 et 13.

#### Réalisation graphique: le boxplot (appelée également “boîte à moustaches”)

Table 3.14  
*Tableau de fréquence d’une série de données discrètes*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
| 1 | 4 | 0.07 | 4 | 0.07 |
| 2 | 7 | 0.13 | 11 | 0.20 |
| 3 | 9 | 0.16 | 20 | 0.36 |
| 4 | 5 | 0.09 | 25 | 0.45 |
| 5 | 3 | 0.05 | 28 | 0.51 |
| 6 | 4 | 0.07 | 32 | 0.58 |
| 7 | 8 | 0.15 | 40 | 0.73 |
| 8 | 2 | 0.04 | 42 | 0.76 |
| 9 | 7 | 0.13 | 49 | 0.89 |
| 10 | 4 | 0.07 | 53 | 0.96 |
| 13 | 1 | 0.02 | 54 | 0.98 |
| 18 | 1 | 0.02 | 55 | 1.00 |

La Figure 3.14 est une représentation graphique des données de la Table 3.14.

Figure 3.14  
*Boîte à moustache représentant les données de la Table 3.14* 

L’axe vertical représente les différentes valeurs que peut prendre la variable étudiée. Etant donné que la valeur la plus basse de la série est 1, et que sa valeur la plus haute est 18, cela signifie que l’ensemble des observations de la série devra être représenté entre ces deux valeurs.

La **boîte centrale** (en bleu) s’étend du premier quartile (la limite inférieure de la boîte centrale = 3) au troisème quartile (= la limite supérieure de la boîte centrale = 8). Elle correspond donc aux 50% des données centrales de la distribution. La distance qui sépare Q1 et Q3 s’appelle **l’écart inter-quartile** et se calcule comme suit: Q3 - Q1 = 8-3=5. L’écart inter-quartile est une mesure de dispersion. Ce type de mesure sera décrit ultérieurement.

La barre à l’intérieur de la boîte représente la **médiane** (valant 5 dans l’exemple de la Table 3.14). La position de la médiane à l’intérieur de la boîte indique le degré de symétrie ou d’asymétrie de la portion centrale de la distribution.

Les **moustaches** (les lignes qui sortent de part et d’autre de la boîte centrale) dépendent des **barrières** qui sont situées à une distance de 1,5 l’écart inter-quartile (soit la longueur de la boîte) de part et d’autre de la boîte. Etant donné que dans l’exemple développé, l’écart inter-quartile vaut 5, les barrières s’éloigneront de points des extrémités de la boîte centrale. En conséquence:

* la barrière supérieure vaut 8+7.5= 15.5
* la barrière inférieure vaut 3-7.5= -4.5

Cependant, une fois les barrières déterminées, il se peut qu’elles ne correspondent à aucune valeur existante de la distribution. Nous allons donc observer les valeurs adjacentes à l’intérieur des barrières. Les scores de la variables étudiée qui se situent à l’**intérieur** des barrières varient de 1 à 13. Donc, la valeur adjacente inférieure est de 1 et la valeur adjacente supérieure est de 13. C’est à ces valeurs adjacentes que correspondent les extrémités visibles des moustaches.

Les **valeurs extrêmes** sont les points de part et d’autres des moustaches. Ce sont des valeurs qui sont supérieures à la barrière supérieure, ou inférieures à la barrière inférieure. Elles correspondent à des valeurs dont le score semble anormalement élevé ou bas, compte tenu de l’ensemble des observations étudiées. Ce serait par exemple le casde l’âge d’un individu ayant 51 ans, dans une étude consacrée à des adolescents.

Le boxplot a pour avantage de permettre de détecter très rapidement ces valeurs extrêmes. Par ailleurs, elle permet de se représenter mentalement l’allure de la distribution: est-elle symétrique ou asymétrique?

### Variables quantitatives continues

#### Simplification des données: calcul de la médiane et des quartiles

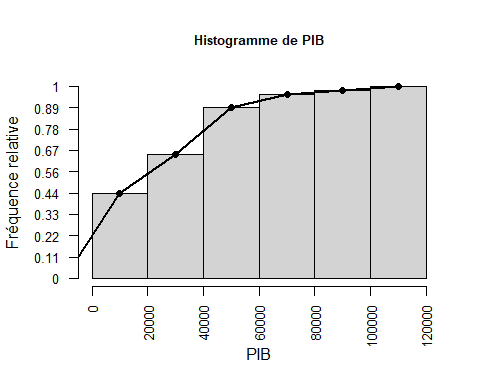
##### Au départ de données brutes

Lorsque les données brutes sont fournies, le calcul de la médiane se réalise exactement de la même manière que pour les données quantitatives discrètes. Je vous renvoie donc à la section antérieure sur les variables quantitatives discrètes.

##### Au départ de tableaux de fréquences

Au début du chapitre, il a été vu que réaliser un tableau de fréquences de données continues implique préalablement de regrouper les valeurs par classe. Bien que ces regroupements constituent une perte d’information, il est malgré tout possible d’estimer la valeur de la médiane et des quantiles, en utilisant le **polygone des effectifs**.

Revenons une fois encore à l’exemple du PIB de la Table 3.6 dont je reproduis ci-dessous le polygone des effectifs cumulés, pour votre convénience. Comme vous l’aurez compris, travailler sur le graphique dans lequel l’axe des y représente les fréquences absolues cumulées, ou dans celui sur lequel l’axe des y représente les fréquences relatives cumulées revient au même. Comme je trouve personnellement plus facile d’utiliser celui qui représente les fréquences relatives cumulées, j’opte pour ce choix.



La fréquence cumulée de la valeur 10000 vaut 44.44 (soit une valeur inférieure à .5 ou 50%). La fréquence cumulée de la valeur 30000 vaut 20 (soit une valeur supérieure à .5 ou 50%). Je sais donc que la valeur de la médiane se situera entre 10^{4} et 310^{4} et sera telle qu’il y ait exactement 50% des observations à sa gauche, et 50% des observations à sa droite. Une simple règle de trois permet de déterminer sa valeur: il y a une distance de 20 unités entre 64.44% et 44.44%. Sur l’axe des x, cette distance est projetée sur une distance de 20000 (soit l’écart entre les deux centres de classes).

Il y a une distance de 5.56 unités entre 50% et 44.44%. Par règle de trois, on en déduit que cela correspond à une distance de sur l’axe des x. Par conséquent, la médiane sera estimée à 10000+5560 = 15560.

En suivant un raisonnement identique, l’on estime les valeurs du premier et troisième quartile… A FAIRE!

#### Réalisation graphique: le boxplot

A nouveau, le boxplot se déroulement de manière identique à ce qui a été vu dans le chapitre sur les variables quantitatives discrètes. Seule la manière de déterminer la médiane ainsi que les premier et troisième quartile diffère.

###### Pagebreak

# Chapitre 4: exploration algébrique des données à une dimension

Jusqu’à présent, nous n’avons envisagé la présentation d’une distribution statistique que sous forme graphique. Cependant, plusieurs caractéristiques sont importantes à déterminer algébriquement. On peut distinguer trois grandes catégories d’indicateurs algébriques essentiels:

1. Les mesures de **tendance centrale**:mesure qui permet de représenter au mieux un ensemble données par une valeur unique
2. Les mesures de **dispersion**: mesure du degré auquel les données s’éloignent de la tendance centrale
3. Les mesures d’**asymétrie** et d’**aplatissement** : mesures liées à la forme de la distribution.

## Mesures de tendance centrale

Parmi elles, se trouvent la moyenne, le mode et la médiane. En tant que quantile particulier, la médiane a déjà été décrite dans le chapitre 3. Pour illustrer le mode et la moyenne, reprenons l’exemple de l’ancienneté de la Table 1.1. Pour votre facilité, je reproduis ci-dessous ces données, sous forme de données brutes (a) et de tableau de fréquences (b):

Table 4.1  
*Extrait de la Table 1.1 : ancienneté de 15 employés*

***(a) Données brutes***

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| i | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|  | 1 | 5 | 2 | 8 | 8 | 9 | 9 | 1 | 1 | 1 | 8 | 8 | 7 | 10 | 8 |

***(b) Tableau de fréquence***

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|  | 4 | 1 | 0 | 0 | 1 | 0 | 1 | 5 | 2 | 1 |

Pour rappel, on utilise une lettre majuscule indicée par *i* pour décrire les données brutes ( = la valeur du ième sujet sur la variable Y), et une lettre minuscule indicée par *j* lorsqu’on parle des valeurs de la variable Y dans un tableau de fréquence ( = la jème valeur que peut prenre la variable Y).

### Mode

Le mode correspond à la classe la plus représentée. Dans la Table 4.1 (a ou b), vous consaterez aisément que 8 est le nombre d’années d’ancienneté le plus représenté, puisque 5 employés ont cette ancienneté. C’est donc le mode de la distribution.

Remarquez qu’une distribution peut être multimodale. Par exemple, dans la série de la table Table 4.2, il y a deux valeurs plus représentées que les autres: ce sont les valeurs 21 et 30. Quand il y a deux modes, on parle plus spécifiquement de distribution bimodale.

Table 4.2  
*Extrait de la Table 1.1 : âge 15 employés*

***(a) Données brutes***

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| i | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|  | 19 | 25 | 21 | 27 | 30 | 32 | 30 | 19 | 21 | 21 | 24 | 27 | 26 | 30 | 26 |

***(b) Tableau de fréquence***

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|  | 4 | 1 | 0 | 0 | 1 | 0 | 1 | 5 | 2 | 1 |

Le grand avantage du mode est d’être insensible aux valeurs extrêmes. Imaginons que par exemple, j’ai encodé par erreur que l’ancienneté du sujet 14, = 100 ou même 1000 au lieu de 10. Le mode resterait identique, ce serait 8. Ca ne sera pas le cas pour la moyenne, comme nous le verrons plus tard. En revanche, le grand désavantage du mode est qu’il ne dépend que de la ou des quelques valeurs les plus représentées et est totalement insensible au reste de la distribution (même des valeurs qui n’ont rien d’aberrantes).

### Moyenne

Lorsqu’on parle de moyenne statistique, on entend toujours (à notre niveau) la moyenne arithmétique. Si l’on part des données brutes, la moyenne arithmétique peut se calculer en prenant la somme des valeurs de la variable dont on veut calculer la moyenne, et en divisant cette somme par le nombre de valeurs qui constituent cette somme. La moyenne arithématique se symbolise par une barre horizontale placée au dessus de la lettre majuscule qui représente la variable étudiée. Dans les formules, nous utilisons généralement la lettre “X”, mais cette lettre peut être remplacée par n’importe quelle autre lettre représentant une variable). Pour reprendre l’exemple de la Table 4.1, dans la mesure où l’ancienneté est représentée par la lettre Y, la moyenne de cette variable se notera .

**Formule de la moyenne arithmétique calculée à partir des données brutes de l’ancienneté**

Lorsqu’il y a un grand nombre de données, appliquer cette formule manuellement devient relativement ardu (imaginez que vous deviez additionner le score de 300 participants, et diviser ensuite cette somme par 300, cela devient long!). Dans ce cas, on peut utiliser les données présentées sous forme de distribution de fréquences, et calculer la moyenne, soit sur base des fréquences absolues, soit sur base des fréquences relatives.

**Formule de la moyenne arithmétique calculée à partir d’un tableau de fréquence**

***En utilisant les valeurs et les fréquences absolues***

***En utilisant les valeurs et les fréquences relatives***

Appliquées à l’ancienneté fournie dans la table Table 4.1, on obtient ceci:

***En utilisant les valeurs et les fréquences absolues***

***En utilisant les valeurs et les fréquences relatives***

Bien entendu, les trois formules donnent une valeur identique de la moyenne, et heureusement!Il s’agit de trois manières différentes de calculer la même chose.

La moyenne présente plusieurs inconvénients.

Premièrement, elle est très sensible aux valeurs aberrantes. Rappelez-vous que lorsque nous avons envisagé le mode, ce dernier était parfaitement insensible à un = 100 ou 1000 qui remplacerait le 10 dans les données brutes. Dans le cas de la moyenne, ce n’est plus le cas du tout. En effet, le remplacement du 10 en 100 fait passer la moyenne de 5.73 à 11.73 et le remplacement du 10 en 1000 fait passer la moyenne à 71.73. Constatez que ces valeurs ne permettent pas du tout de représenter correctement l’ancienneté habituelle des employés (et heureusement!). C’est donc un désavantage sérieux qui nous oblige à être attentifs aux valeurs aberrantes.

Deuxièmement, si l’on travaille avec des distributions asymétriques ou multimodales, la moyenne ne les représentera pas correctement.Imaginons une classe constituée de 30 étudiants dans laquelle la moitié a obtenu 0/10 à une interrogation, et où l’autre moitié a obtenu 10/10 (dans cette classe, la distribution est bimodale). La moyenne vaudra exactement 5/10. Pourtant, aucun étudiant n’a obtenu une note proche de 5/10.

Nous voyons bien à quel point il est important d’étudier la forme d’une distribution avant de déterminer des valeurs algébriques.

## Mesures de dispersion

Une mesure de dispersion sert à quantifier à quel points les données ont tendance à s’étaler, se disperser autour de la tendance centrale. Ces mesures sont importantes car la mesure de tendance centrale résumera d’autant mieux les données que la dispersion des données autour de celle-ci sera faible.

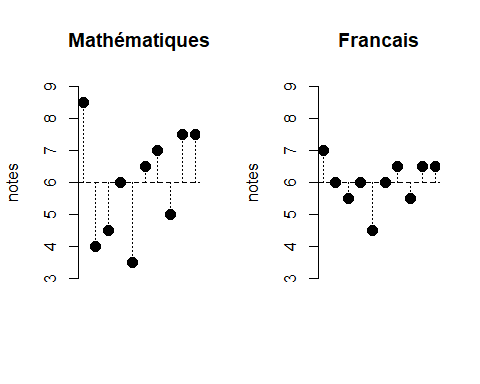
Pour l’illustrer, considérons les deux séries suivantes qui représentent (fictivement) les notes (sur 10) obtenues par 10 étudiants en mathématiques et en français.

Table 4.3  
*Notes (/10) obtenues par 10 des étudiants en mathématiques et en français*

|  |  |  |
| --- | --- | --- |
|  | math( | francais( |
| 1 | 8.5 | 7.0 |
| 2 | 4.0 | 6.0 |
| 3 | 4.5 | 5.5 |
| 4 | 6.0 | 6.0 |
| 5 | 3.5 | 4.5 |
| 6 | 6.5 | 6.0 |
| 7 | 7.0 | 6.5 |
| 8 | 5.0 | 5.5 |
| 9 | 7.5 | 6.5 |
| 10 | 7.5 | 6.5 |

Je vous laisse le soin de calculer la moyenne de ces deux séries et de constater que dans les deux cours, elle vaut exactement 6. Par contre, la dispersion des données autour de cette tendance centrale n’est pas la même pour les deux cours.

Dans la Figure 4.1, les notes de chaque cours sont représentées. La ligne horizontale représente la moyenne. Les points au dessus de la ligne horizontale sont les notes supérieures à la moyenne et les points en dessous de cette ligne sont les notes inférieures à la moyenne. On observe graphiquement que les notes s’éloignent plus de la moyenne pour le cours de mathématiques que pour le cours de français. Cela se confirme algébriquement: en français, les notes s’éloignent au maximum de 1points de la moyenne alors qu’en mathématiques, les notes peuvent s’éloigner jusqu’à 2.5points de la moyenne. Autrement dit, la dispersion des notes est plus forte pour le cours de mathématiques que pour le cours de français.

Figure 4.1  
*Etalement des notes de la Table 4.3 autour de la moyenne (6)* 

Quantifier la dispersion des données peut se faire de plusieurs manières plus ou moins sophistiquées. Nous en ciblerons quatre:  
*L’étendue* L’écart interquartile *L’écart moyen absolu* La variance et l’écart-type

### Etendue

L’étendue est une mesure très élémentaire qui consiste à mesurer la différence entre la plus grande et la plus petite des valeurs observées.

Dans l’exemple des notes de cours de la Table 4.3, on trouve que l’étendue des notes en français vaut 2.5, et que celle des notes en mathématiques vaut 5.

L’avantage de cette mesure est qu’elle est très simple à calculer. En contre-partie, elle présente des inconvénients importants. Premièrement, elle ne dépend que deux valeurs de la série (la plus grande et la plus petite) et ne prend pas du tout en compte toutes les autres valeurs. Deuxièmement, elle est très sensible aux valeurs extrêmes: il suffit qu’un seul sujet ait un score atypique pour que l’étendue augmente considérablement. Vous pouvez donc regarder cet indicateur lorsque vous abordez en un coup d’oeil la distribution de vos résultats, mais n’accordez aucun crédit à cet indicateur dès lors que vous entrez dans une analyse plus fine.

### Ecart interquartile

L’écart interquartile consiste à mesurer l’écart entre le troisième et le premier quartile (c’est la boite centrale de la boîte à moustaches décrite précédemment, soit les 50% des données centrales de la distribution).

Dans l’exemple des notes de cours de la Table 4.3, on trouve que l’écart interquartile des notes en français vaut 1, et que celui des notes en mathématiques vaut 3.

Contrairement à l’étendue des données, l’écart interquartile présente l’avantage de ne pas être sensible aux valeurs extrêmes.

### Ecart moyen absolu (EMA)

Un autre moyen de calculer la dispersion est de mesurer les écarts par rapport à la moyenne (E), tel que fait à la deuxième colonne de la Table 4.4 et de la Table 4.5.

Table 4.4  
*Notes en % obtenues par 10 des étudiants en mathématiques*

|  |  |  |
| --- | --- | --- |
|  |  |  |
| 1 | 8.5 | 2.5 |
| 2 | 4.0 | -2.0 |
| 3 | 4.5 | -1.5 |
| 4 | 6.0 | 0.0 |
| 5 | 3.5 | -2.5 |
| 6 | 6.5 | 0.5 |
| 7 | 7.0 | 1.0 |
| 8 | 5.0 | -1.0 |
| 9 | 7.5 | 1.5 |
| 10 | 7.5 | 1.5 |

Table 4.5  
*Notes en % obtenues par 10 des étudiants en francais*

|  |  |  |
| --- | --- | --- |
|  |  |  |
| 1 | 7.0 | 1.0 |
| 2 | 6.0 | 0.0 |
| 3 | 5.5 | -0.5 |
| 4 | 6.0 | 0.0 |
| 5 | 4.5 | -1.5 |
| 6 | 6.0 | 0.0 |
| 7 | 6.5 | 0.5 |
| 8 | 5.5 | -0.5 |
| 9 | 6.5 | 0.5 |
| 10 | 6.5 | 0.5 |

Une information intéressante, en vue de déterminer la dispersion, serait de pouvoir estimer l’écart moyen des observations par rapport à la moyenne. Cependant, nous nous retrouvons confrontés à un problème: calculer la moyenne des écarts consisterait à additionner tous ces écarts, et à diviser la somme des écarts par le nombre de termes additionnés. Or, étant donné que certains écarts sont supérieurs à 0 et que d’autres sont inférieurs à 0, lorsque l’on tente d’additionner tous les écarts par rapport à la moyenne, on se retrouve invariablement avec un résultat nul:

Pour le cours de math:

Pour le cours de français:

Il existe deux moyens de contourner ce problème. Soit prendre la valeur absolue de chaque écart, soit élever toutes les erreurs au carré (et donc ne plus se soucier du signe, puisqu’un chiffre négatif élevé au carré devient positif).

La première stratégie est celle choisie lorqu’on calcule l’EMA. l’EMA est une estimation de l’écart à la moyenne, en moyenne, par sujet. Il se calcule comme suit:

Appliquée aux notes de mathématiques de la Table 4.3, cela donne

.

Cela revient à dire que les notes s’écartent en moyenne de 1.4 points par rapport à la moyenne de 6.

Appliquée aux notes de français de la Table 4.3, cela donne

.

Cela revient à dire que les notes s’écartent en moyenne de 0.5 points par rapport à la moyenne de 6.

Bien que cette mesure est une excellente façon de se représenter la dispersion, elle est totalement supplantée par l’écar-type. Cela provient du fait que l’écart-type est dérivé de la variance qui jouit de propriétés mathématiques qui lui font jouer un rôle central en statistique théorique.

### Variance et ecart-type

Nous avons envisagé de résoudre le problème de la somme nulle des erreurs (des écarts à la moyenne) en élevant chaque écart au carré. C’est la stratégie utilisée lorsque l’on calcule la variance des données. La variance consiste donc à calculer la valeur moyenne du carré des écarts entre les données observées et leur moyenne. Elle se calcule comme suit:

Appliquée aux notes de mathématiques de la Table 4.3, cela donne:

.

Il y a cependant un petit problème: la moyenne de la somme des carrés des écarts (SCE) est un chiffre dont l’unité est le carré de l’unité des données étudiées. Trouver une variance de 2.55 revient à dire que les notes s’écartent en moyenne de 2.55 “points au carré” de la moyenne de 6. C’est donc très difficilement interprétable. Pour résoudre ce problème, on va ramener cet indicateur à la même unité que l’unité de la moyenne, c’est-à-dire les “points”. Pour ce faire, il suffit d’extraire la racine carrée de la variance: . Cette mesure constitue **l’écart-type**. Nous pouvons maintenant dire que les sujets s’écartent en moyenne de 1.6 points autour de la moyenne de 6.

Appliquée aux notes de français de la Table 4.3, cela donne:

.

L’écart-type, soit la racine carrée de la variance vaut . Nous pouvons maintenant dire que les sujets s’écartent en moyenne de 0.67 points autour de la moyenne de 6.

Remarquez que la mesure de l’écart-type est un petit peu différente de celle que nous obtenions en utilisant l’EMA (Pour le cours de math, écart-type = 1.6 vs. EMA = 1.4; pour le cours de français, écart-type = 0.67 vs. EMA = 0.5). L’écart-type est un petit peu plus conservateur, c’est-à-dire qu’il surestime un petit peu l’erreur par rapport à l’EMA.

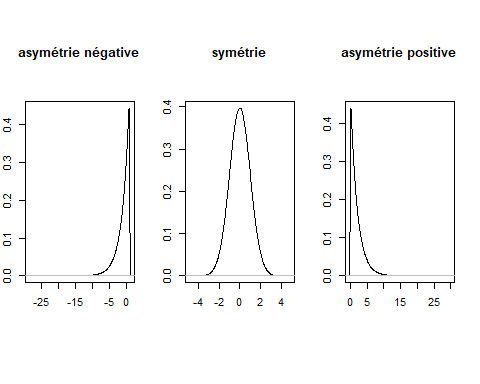
La variance et l’écart-type souffrent de la même limite que la moyenne: ils sont tous les deux sensibles aux valeurs abberrantes. Cette sensibilité est accentuée par le fait que l’on élève les écarts au carré!

## Mesures d’asymétrie et d’aplatissement

Les distributions statistiques peuvent avoir plusieurs formes. Nous avons déjà envisagé la différence entre les distributions unimodales et bimodales. Il est également possible de distinguer les formes des distributions sur base d’autres paramètres, tels que l’**asymétrie** et l’**aplatissement**.

### Mesure d’asymétrie

Une distribution symétrique est telle que la moyenne et la médiane sont confondues. Sur la partie centrale de la Figure 4.2, on observe graphiquement que la première moitié de la distribution partie à gauche de la médiane) est la parfaite symétrie orthogonale de la deuxième moitié de la distribution (la partie à droite de la médiane).

Figure 4.2  
*Exemple de distributions à asymétrie négative (gauche) symétrique (milieu) et à asymétrie positive (droite)* 

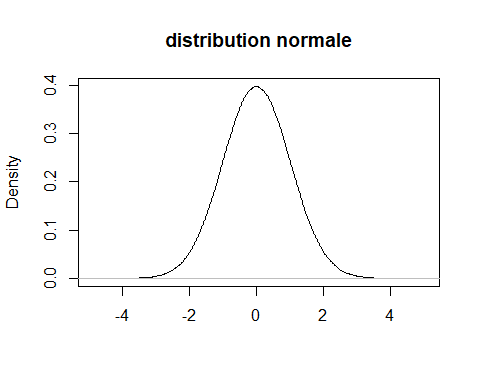
Une distribution peut également présenter une asymétrie négative (si la queue de la distribution tend vers les valeurs négatives de l’axe x; extrêmité gauche de la Figure 4.2) ou une asymétrie positive (si la queue de la distribution tend vers les valeurs positives de l’axe x; extrêmité droite de la Figure 4.2).

Bien qu’il existe plusieurs mesures d’asymétrie, la plus courante est celle proposée par Pearson, dont voici la formule:

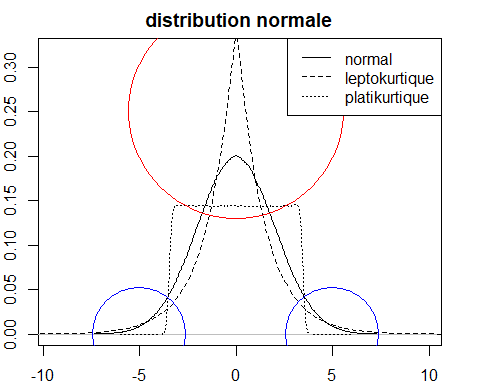
Il n’est pas important que vous puissiez calculer cela manuellement. Par contre, vous devez pouvoir interpréter sa valeur:  
*Lorsque la distribution est parfaitement symétrique, le coefficient G1 vaut exactement 0.*  
Lorsque la distribution présente une asymétrie positive, le coefficient G1 a une valeur supérieure à 0.  
\*Lorsque la distribution présente une asymétrie négative, le coefficient G1 a une valeur inférieure à 0.

### Mesure d’aplatissement

Pour correctement comprendre la mesure d’aplatissement, il faut savoir que la distribution de référence est la distribution normale (ou courbe en cloche, représentée sur la Figure 4.3)

Figure 4.3  
*Distribution normale (appelée aussi courbe en cloche)* 

Une distribution dont le kurtosis vaut 0 est une distribution qui n’est ni plus plate ni plus pointue que la fameuse courbe en cloche que l’on appelle **distribution normale**. Une distribution dite **leptokurtique** est une distribution qui est plus pointue et avec des extrémités plus épaisses que la distribution normale. Enfin, une distribution dite **platikurtique** est une distribution aplatie, dont le centre et les extrémités sont moins fournies que ceux d’une distribution normale.

Figure 4.4  
*Exemple de distributions d’aplatissement normale, leptokurtique et platikurtiques* 

De nouveau, il existe plusieurs mesures de kurtosis. Voici celle proposée par Pearson:

De même que pour la mesure d’asymétrie, il n’est pas important que vous puissiez calculer cela manuellement. Par contre, vous devez pouvoir interpréter sa valeur:  
*Lorsque la distribution a l’aplatissement d’une distribution normale, le coefficient G2 vaut exactement 0.*  
Lorsque la distribution est leptokurtique, le coefficient G2 a une valeur supérieure à 0.  
\*Lorsque la distribution est platikurtique, le coefficient G2 a une valeur inférieure à 0.

###### Pagebreak

# Chapitre 5: relation entre deux variables continues: le coefficient r de Pearson

Jusque là, nous avons passé beaucoup de temps à étudier les variables. Nous avons vu comment il était possible de les représenter graphiquement et aussi comment les étudier algébriquement. Tout cela relève de la statistique descriptive **univariée**.

Le dernier objectif de ce cours est d’apprendre à observer comment deux variables peuvent varier **en même temps**, ou covarier. Cela relève de la statistique **bivariée**.

Revenons à l’exemple de la Table 4.4 dans laquelle nous avions étudié la note de 10 étudiants à deux cours: le cours de mathématiques et le cours de français.

Nous avons déjà précédemment étudié comment les notes de chacun de ces cours pouvaient varier individuellement d’un étudiant à l’autre. La question que nous nous posons maintenant est de savoir comment les notes de ces deux cours **covarient**. Autrement dit, existe-t-il un lien entre les notes aux deux cours ou encore, lorsque les notes à un cours augmentent, est-ce que celles de l’autre cours tendent à augmenter également?

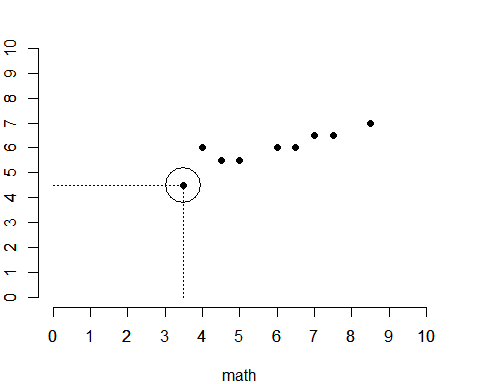
## Représentation graphique de la relation entre deux variables

Une première manière d’appréhender la question est de se représenter graphiquement la situation. Lorsque l’on souhaite pouvoir étudier le lien entre deux variables, la meilleure représentation graphique consiste à tracer deux axes perpendiculaires, l’un représentant l’une des deux variables, l’autre représentant l’autre variable.

Table 4.3  
*Notes (/10) obtenues par 10 des étudiants en mathématiques et en français*

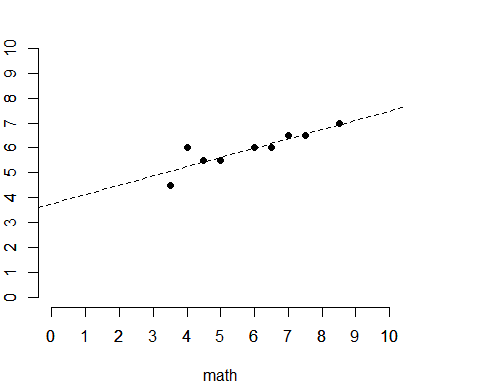
|  |  |  |
| --- | --- | --- |
|  | math( | francais( |
| 1 | 8.5 | 7.0 |
| 2 | 4.0 | 6.0 |
| 3 | 4.5 | 5.5 |
| 4 | 6.0 | 6.0 |
| 5 | 3.5 | 4.5 |
| 6 | 6.5 | 6.0 |
| 7 | 7.0 | 6.5 |
| 8 | 5.0 | 5.5 |
| 9 | 7.5 | 6.5 |
| 10 | 7.5 | 6.5 |

La Figure 5.1 montre le graphe qui reprend les deux valeurs les unes par rapport aux autres. On y voit un nuage de points, chacun de ceux-ci ayant une double coordonnée: sur l’axe des x, et sur l’axe des y. Par exemple, le point entouré correspond à un individu ayant obtenu la note 3.5 au cours de mathématiques (coordonnée sur l’axe des x) et 4.5 au cours de francais (coordonnée sur l’axe des y).

Figure 5.1  
*Nuage de points représentant la relation entre la note à un cours de math et la note à un cours de français*  


En observant le nuage de points dans son ensemble, il apparait que celui-ci n’est pas du tout aléatoire. Globalement, les individus ayant les meilleurs points en mathématiques semblent également avoir des points élevés en français. A contrario, ceux qui ont les moins bon points en mathématiques semblent avoir également de moins bons points en français.

Pour mieux encore appréhender la relation qui existe entre deux variables, il est possible de tracer, au départ du nuage de points, une droite qui se rapproche le plus possible de l’ensemble des points (selon des règles très précises), que l’on appelera la **droite de régression**.

Figure 5.2  
*Nuage de points représentant la relation entre la note à un cours de math et la note à un cours de français*  


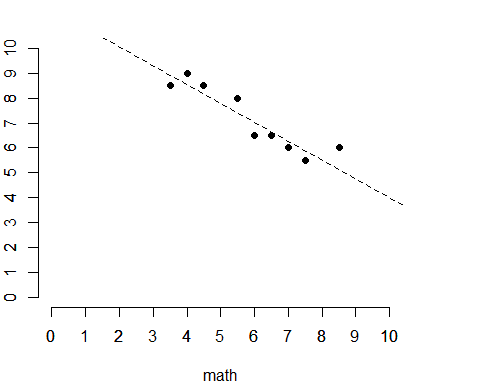
On constate que l’ensemble des points semblent se rapprocher d’une droite de régression dont la pente est croissante. Ca sera le cas chaque fois que la relation linéaire entre deux variables est positive (càd quand le score des deux variables tend à varier dans le même sens).

Imaginons à présent que nous souhaitions étudier la relation entre le cours de mathématiques et un cours de géographie suivi par les 10 mêmes étudiants.

Table 4.6  
*Notes (/10) obtenues par 10 des étudiants en mathématiques et en français*

|  |  |  |
| --- | --- | --- |
|  | math( | geographie( |
| 1 | 8.5 | 6.0 |
| 2 | 4.0 | 9.0 |
| 3 | 4.5 | 8.5 |
| 4 | 6.0 | 6.5 |
| 5 | 3.5 | 8.5 |
| 6 | 6.5 | 6.5 |
| 7 | 7.0 | 6.0 |
| 8 | 5.5 | 8.0 |
| 9 | 7.5 | 5.5 |
| 10 | 7.5 | 5.5 |

Comme précédemment, l’étude du nuage de points et de la droite de régression associée aidera à mieux se représenter la relation entre les deux variables (voir r).

Figure 5.3  
*Nuage de points représentant la relation entre la note à un cours de math et la note à un cours de géographie*  


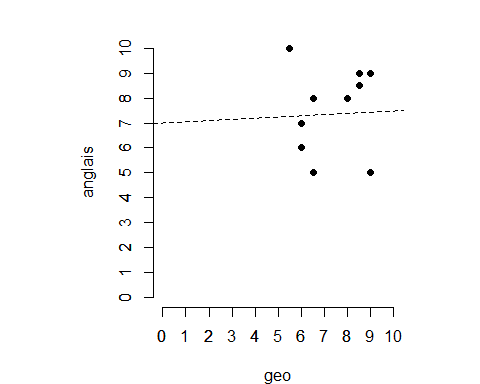
Cette fois, les points semblent se rapprocher d’une droite de régression décroissante.Ca sera le cas chaque fois que la relation linéaire entre deux variables est négative (càd quand le score des deux variables tend à varier dans des sens opposés: lorsque score de mathématiques augmente, celui de géographique diminue, et réciproquement).

Comme dernier cas de figure, imaginons que nous souhaitions étudier la relation entre le cours de géographie et d’anglais, toujours suivi par les 10 mêmes étudiants.

Table 4.6  
*Notes (/10) obtenues par 10 des étudiants en mathématiques et en français*

|  |  |  |
| --- | --- | --- |
|  | geographie( | anglais( |
| 1 | 5.5 | 10.0 |
| 2 | 9.0 | 5.0 |
| 3 | 6.0 | 6.0 |
| 4 | 6.0 | 7.0 |
| 5 | 6.5 | 5.0 |
| 6 | 6.5 | 8.0 |
| 7 | 8.0 | 8.0 |
| 8 | 8.5 | 9.0 |
| 9 | 8.5 | 8.5 |
| 10 | 9.0 | 9.0 |

Encore une fois, l’étude du nuage de points et de la droite de régression associée aidera à mieux se représenter la relation entre les deux variables (voir r).

Figure 5.4  
*Nuage de points représentant la relation entre la note à un cours de géographie et la note à un cours d’anglais*  


Contrairement au deux situations précédentes, les points ne se rapprochent ni d’une droite croissante, ni d’une droite décroissante. Il ne semble pas y avoir de relation entre les notes obtenues au cours de géographiques et celles obtenues au cours d’histoire. Par ailleurs, la droite de régression a une pente quasiment nulle.

## Détermination algébrique de la relation entre deux variables

###### Pagebreak

# Références

Labreuche, J. (2010). Les différents types de variables, leurs représentations graphiques et paramêtres descriptifs. Sang Thrombose Vaisseaux, 22(10), 536-543.

Syllabus de Christophe

Séminaire de Christian Ritter sur les graphiques

Support office