Statistiques

Marie Delacre

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 recherche mathématique qui permettent de **récolter**, **traiter** et **interpréter** un ensemble de données (généralement vaste).

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

La **population** est l'ensemble des éléments (individus ou objets) auxquels on s'intéresse. Les éléments qui constituent cette population sont appelés **individus** ou **unités statistiques**. 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 intéressent, comme le genre, la situation géographique ou linguistique ou encore l'âge ou 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 mètre et me dire que son tour de taille est de 72 cm. Vous pourriez également vous contenter de dire qu'il est de taille "M" (moyenne). 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, appelé 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 telle qu'illustrée à la Table 1.1. Cette table est une représentation fictive de données socio-démographiques récoltées auprès de 15 employés du secteur automobile.

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

i	genre (W_i)	âge (X_i)	anciennete (Y_i)	langue (Z_i)
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_i (X indicé i) la valeur d'un individu i sur la variable X, par Y_i (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 $Y_3 = 2$, l'âge du 8ème employé de l'échantillon est symbolisé par X_8=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, à savoir la réalisation de graphiques. 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'analyse. 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 connaître 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 observée dans notre échantillon soit bien conforme à la proportion de femmes au sein de la population ? Cela requiert de faire de la **statistique inférentielle**. Cette forme de statistique ne fait pas partie des objectifs de ce cours. Sachez cependant qu’elle existe.

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 est 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 (ces codes numériques ne sont que des étiquettes). 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 grands 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 attribués 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 soit trois fois plus grand que ce dernier.

Les variables quantitatives

Les variables **quantitatives** sont des variables dont les valeurs sont intrinsèquement numériques, c’est-à-dire qu’il fait sens d’effectuer des opérations mathématiques sur leur valeurs. Par exemple, si un particulier a deux poules, que la première pond trois œufs et la deuxième en pond deux, il est correct de dire que le fermier trouvera 5 œufs lorsqu’il se rendra dans son poulailler. La variable “nombre d’œufs” est donc bien une variable quantitative.

Les variables ***discrètes*** ne peuvent prendre que des valeurs isolées et généralement entières, dans un intervalle de valeurs spécifiques dans R[1]. 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 distances, la durée … sont des exemples de variables continues.

Exercices de fin de chapitre

Pour chacune des variables suivantes, déterminez sur quel type d'échelle elles ont été mesurées

	Qualitative		Quantitative	
	Nominale	Ordinale	Discrète	Continue
Statut matrimonial				
Âge				
Genre				
Activité sportive (nulle, légère ou forte)				
Taille (en cm)				
Nombre d'enfants à charge				
Salaire				

[1] R représente l'ensemble des nombres réels.

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'œil 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 attractives (représentations en 3D, avec couleurs inhabituelles, textures bois…). Je déconseille vivement l'usage de ces artifices. Gardez à l'esprit qu'une bonne présentation est la clé d'une bonne communication.

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 dépend 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 (sauf dans certains cas, comme celui du nuage de points qu'on abordera dans le dernier chapitre). La plupart du temps, 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 d'un **tableau de fréquence** et ceux qui reposent sur le calcul de **quantiles**.

Les graphiques construits sur base d'un tableau de fréquences

Variables qualitatives nominales

Simplification des données

Imaginez que 40 participants aient répondu par "oui" ou "non" à la question suivante "Êtes-vous fumeur ?" et que vous vous retrouviez avec l'ensemble de réponses 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 des données brutes. Une manière de résumer cet ensemble d'informations 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*

j	**Fumeur (x_j)**	n_j	f_j
1	non	30	0.75
2	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). Notons que quand on étudie des variables nominales, peu importe l'ordre d'apparition des modalités de la variable dans le tableau de fréquence, cela n'altère en rien le contenu informatif du tableau. Dans la Table 3.1, nous avons représenté d'abord les non-fumeurs, suivi des fumeurs. Il s'agit d'un choix purement arbitraire et la table ci-dessous aurait pu être réalisée également :

j	**Fumeur (x_j)**	n_j	f_j
1	Oui	10	0.25
2	Non	30	0.75

Le nombre de personnes appartenant à chaque catégorie constitue ce que l'on appelle la **fréquence absolue** ou **l'effectif** (ce sont deux synonymes).

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 variable X seront symbolisées 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_j (x indicé j) la jème valeur possible de la variable X. Par exemple, si l'on choisit de nommer X la variable "fumer", $x_2 = oui$. n_j correspond aux fréquences absolues associées à chaque valeur x_j. Par exemple, n_1= 30.

La somme des fréquences absolues correspond au nombre total de personnes constituant l'échantillon (ici, n vaut 40):

$$n = \sum_{j=1}^{k} n_j = n_1 + n_2 = 30 + 10 = 40$$

Avec k étant le nombre de modalités de la variable étudiée.

Dans la formule ci-dessus, on voit apparaître pour la première fois du cours le sigma ($\sum$). Il s'agit du symbole de sommation. $\sum_{j=1}^{n} n_j$ doit être lu comme ceci: "la somme de tous les n_j, avec j pouvant varier de 1 à k". Dans l'exemple de la variable "fumeur", étant donné que cette dernière contient deux modalités, j peut varier de 1 à 2. $\sum_{j=1}^{n} n_j$ est donc égal à $n_1 + n_2$.

La colonne suivante contient la même information mais par rapport à l'ensemble des ménages ; c'est la **fréquence relative**, que l'on note f_j.

$$f_j = \frac{n_j}{n}$$

Par exemple, 30 personnes de l'échantillon sur 40 ne fument pas. La fréquence relative associée à la catégorie de non-fumeurs se notera f_1 et vaudra 30/40= 0.75. On peut également exprimer la fréquence relative en pourcentage :

$$f_j(pourcentage) = \frac{n_j}{n} \times 100$$

Dans ce cas, on dira que 30/40 x 100 = 75 % des personnes de notre échantillon ne fument pas.

Réalisation graphique

Diagramme en bâtons

Le diagramme en bâtons consiste à représenter en abscisse (c'est-à-dire sur l'axe horizontal) les différentes valeurs que peut prendre la variable que l'on étudie (x_j) et en ordonnée (c'est-à-dire sur l'axe vertical) les fréquences associées à chaque valeur x_j (soit n_j, soit f_j, cette dernière pouvant être exprimée soit en termes de proportion comprise entre 0 et 1, soit en termes de pourcentage). Pour chaque valeur x_j, on trace un bâton dont la hauteur varie en fonction de la fréquence associée à cette valeur x_j.

L'ordre d'apparition des différents bâtons importe peu lorsqu'on représente des variables nominales. De plus, étant donné que l'axe des abscisses 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 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 fumeurs et de non fumeurs, dans le second cas, on peut déterminer le pourcentage de personnes qui fument ou ne fument pas dans l'échantillon.

Figure 3.1
Diagramme en bâtons de la distribution de la variable « fumeur »

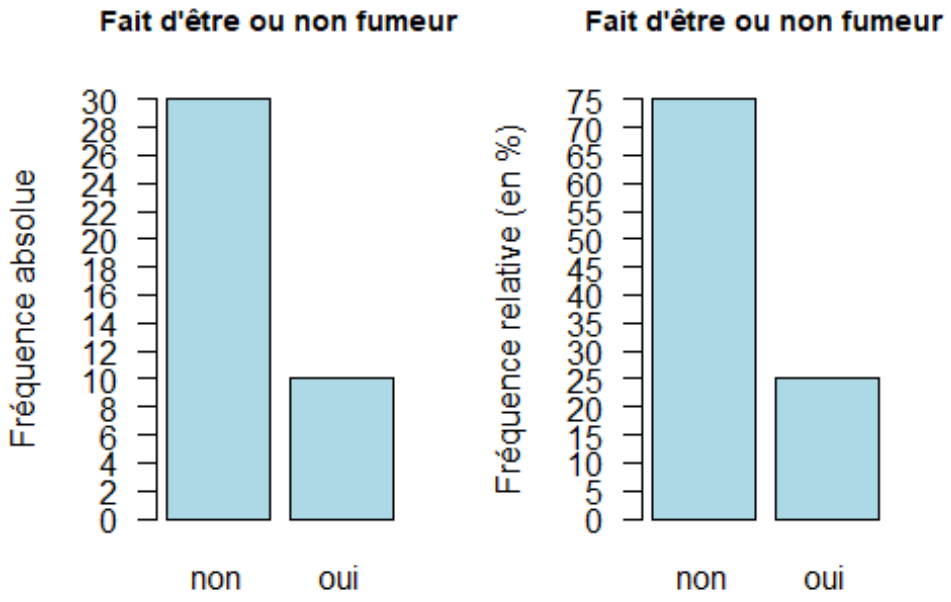
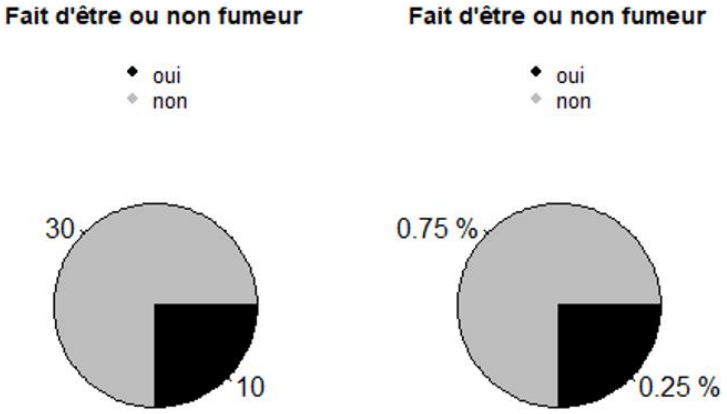


Diagramme circulaire

On trouve occasionnellement dans les rapports des diagrammes circulaires (que l'on appelle aussi graphiques en camembert) pour représenter les variables. Réaliser un diagramme circulaire consiste à subdiviser un cercle en autant de parts 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 : à gauche, le diagramme des fréquences absolues et à droite, celui des fréquences relatives. De même que pour les diagrammes en bâtons, que ce soient les fréquences absolues ou les fréquences relatives importe peu en termes de forme de graphe, mais l'information donnée varie légèrement : dans le premier cas, on peut dénombrer le nombre de fumeurs et de non-fumeurs, dans le second cas, on peut déterminer le pourcentage de personnes qui fument ou ne fument pas dans l'échantillon.

Figure 3.2 *Diagramme circulaire de la variable « fumeur »*



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 individus dans le cas présent. Ces 360° doivent être répartis entre les catégories, de manière proportionnelle à la fréquence de chaque catégorie :

$$40 \rightarrow 360°$$

$$\Leftrightarrow 1 \rightarrow \frac{360°}{40} = 9°$$

$$\Leftrightarrow 30 \rightarrow \frac{360°}{40} \times 30 = 270°$$

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, plutôt que les fréquences absolues, fournit exactement le même résultat, à condition d'utiliser les valeurs exactes (non arrondies). La preuve:

$$100\% \rightarrow 360°$$

$$\Leftrightarrow 1\% \rightarrow \frac{360°}{100} = 3.6°$$

$$\Leftrightarrow 75\% \rightarrow \frac{360°}{100} \times 75 = 270°$$

30 personnes ne fument pas (valeur x_j = non). Cela représente 75% de l'échantillon, l'angle correspondra donc à 75% des 360°, à savoir 270°.

Bien que ce 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 absolues plutôt que sur base des fréquences relatives. Voici les angles qui devraient être associés à chaque catégorie de l'exemple de la Table 3.1:

Table 3.2
Angle de chaque part du diagramme circulaire représentant les données de la Table 3.1

x_j	n_j	angle (en °)
Non	30	270
oui	10	90

Ce type de graphique présente de nombreux inconvénients en termes de communication. Entre autres, il est peu précis : lorsque deux catégoriestab 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 qualitatives ordinales

Table 3.3
Données brutes relatives à la taille de T-shirt d'enfants de 4ème primaires

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

Simplification des données : transformation en tableaux de fréquences

La Table 3.3 répertorie la taille de T-shirt de 40 enfants de 4ème primaire. Notez que j'ai représenté ces données brutes sur plusieurs colonnes distinctes par souci d'économie de place. Cependant, dans un logiciel (Microsoft 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 « Taille ». Rappelez-vous : une ligne par sujet, et une colonne par variable.

La table 3.4 constitue le tableau de fréquence de ces données.

Table 3.4
Transformation de la Table 3.3 en tableau de fréquences

j	Taille de T-shirt (x_j)	n_j	f_j (en %)
1	XS	3	7,5%
2	S	16	40%
3	M	16	40%
4	L	5	12,5%

Comme vous pouvez l'observer, de même que fait précédemment pour les variables nominales, on envisage toutes les tailles de T-shirt possibles et on leur associe une fréquence a solue (n_j) et une fréquence relative (f_j). Par ailleurs, la somme des fréquences absolues correspond toujours au nombre total d'enfants constituant l'échantillon (ici, n vaut 40), et la somme des fréquences relatives correspond toujours à 1 (ou à 100%, si exprimée en pourcentage).

$$n = \sum_{j=1}^{k} n_j = 3 + 16 + 16 + 5 = 40$$

$$\sum_{j=1}^{k} f_j = 7.5\% + 40\% + 40\% + 12.5\% = 100\%$$

Contrairement au cas des variables nominales, l'ordre de présentation de chaque ligne a une importance. On triera les valeurs possibles de la variable représentée par ordre logique (XS apparaitra avant S, S apparaitra avant M, etc.).

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 bâtons 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 d'enfants associés à chaque taille de T-shirt, dans le second cas, on peut déterminer le pourcentage d'enfants qui ont une taille spécifique de T-shirt.

Figure 3.3
Diagramme en bâton des différentes tailles de T-shirt

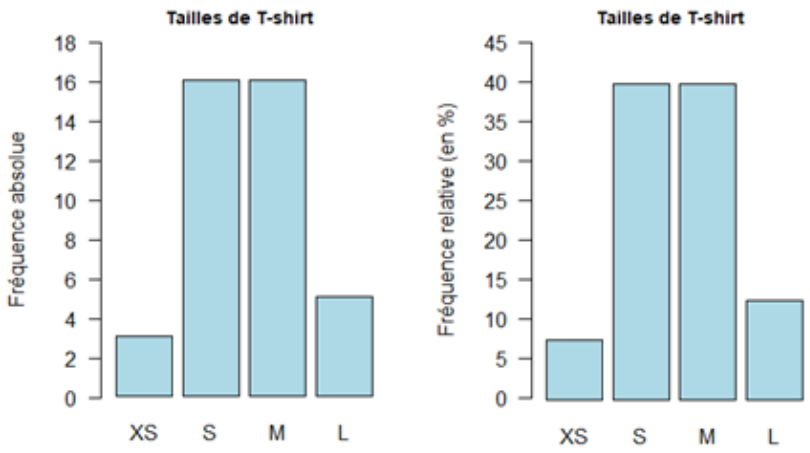
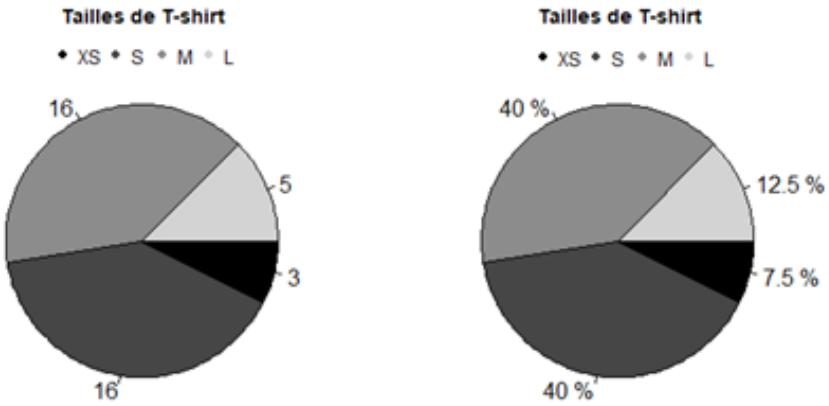


Diagramme circulaire

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 des différentes tailles de T-shirt



La table 3.5 fournit les angles qui devraient être associés à chaque catégorie dans l'exemple de la Table 3.4. Ceux-ci ont été déterminés en procédant à une règle de trois, comme expliqué dans la section aux diagrammes circulaires pour représenter des variables nominales. Essayez de déterminer ces valeurs par vous-même, afin de vous assurer que vous avez bien compris le principe.

Table 3.5
Angle de chaque part du diagramme circulaire représentant chaque modalité de la variable « taille de T-shirt »

x_j	n_j	angle (en °)
XS	3	27
S	16	144
M	6	144
L	5	45

Variables quantitatives discrètes

Table 3.6

Données brutes relatives au nombre d'enfants par ménages, au sein de 40 ménages

i	Enfants (X_i)	i	Enfants (X_i)	I	Enfants (X_i)	i	Enfants (X_i)
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 3.6 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 (Microsoft 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 3.7 constitue le tableau des fréquences de ces données.

Table 3.7
Transformation de la Table 3.6 en tableau de fréquences

j	Nombre d'enfants (x_j)	n_j	f_j (en %)	N_j	F_j (en %)
1	1	14	35.0	14	35.0
2	2	13	32.5	27	67.5
3	3	9	22.5	36	90.0
4	4	3	7.5	39	97.5
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 (ou modalités de la variable) ont été envisagés et on leur associe une fréquence absolue (ou effectif ; n_j), et une fréquence relative (f_j). Par ailleurs, la somme des fréquences absolues correspond toujours au nombre total de ménages constituant l'échantillon (ici, n vaut 40), et la somme des fréquences relatives correspond toujours à 1 (ou à 100%, si exprimée en pourcentage).

$$n = \sum_{j=1}^{k} n_j = 14 + 13 + 9 + 3 + 1 = 40$$

$$\sum_{j=1}^{k} f_j = 35\% + 32.5\% + 22.5\% + 7.5\% + 2.5\% = 100\%$$

De plus, contrairement au cas des variables qualitatives nominales et similairement au cas des variables qualitatives ordinales, l'ordre de présentation de chaque ligne a une importance. On triera les valeurs possibles de la variable représentée par ordre croissant (1 apparaitra avant 2, 2 apparaitra 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: N_j).

$$N_j = \sum_{j=1}^{j} n_j \, , 1 \leq j \leq k$$

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: F_j).

$$F_j = \sum_{j=1}^{j} f_j = \sum_{j=1}^{j} \frac{n_j}{n}, 1 \leq j \leq k$$

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 bâtons importe.

La Figure 3.5 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 un nombre donné d'enfants.

Figure 3.5
Diagramme en bâtons du nombre d'enfants par ménage

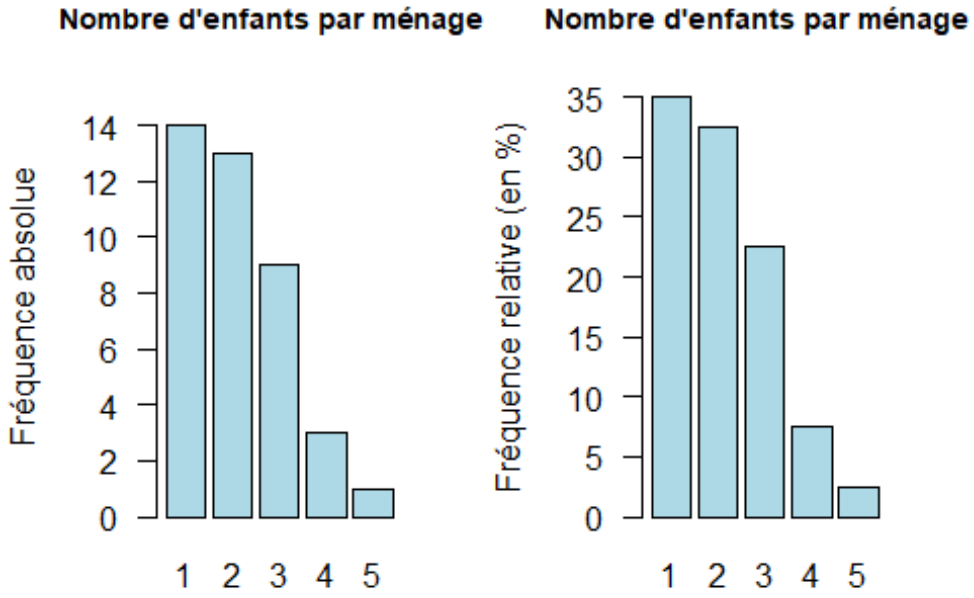
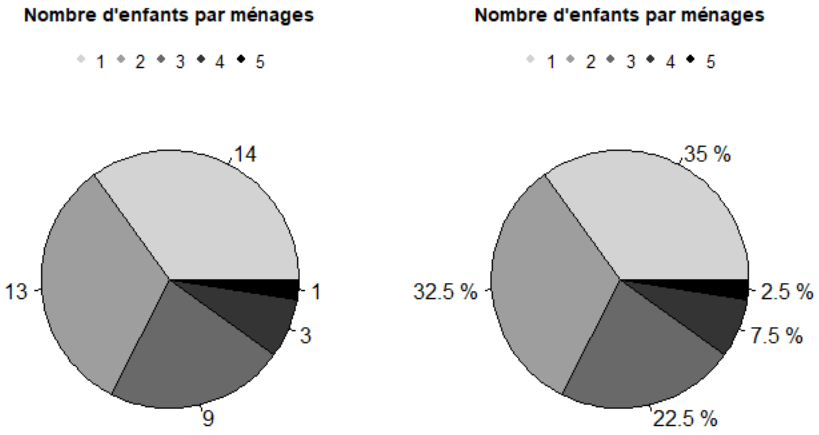


Diagramme circulaire

La Figure 3.6 montre deux diagrammes circulaires de notre distribution : l'un qui représente les fréquences absolues, et l'autre qui représente les fréquences relatives.

Figure 3.6 *Diagramme circulaire du nombre d'enfants par ménage*



La Table 3.8 fournit les angles qui devraient être associés à chaque catégorie dans l'exemple de la Table 3.6. 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. Essayez de déterminer ces valeurs par vous-même, afin de vous assurer que vous avez bien compris le principe.

Table 3.8
Angle de chaque part du diagramme circulaire représentant les données de la Table 3.6

x_j	n_j	**angle (en °)**
1	14	126
2	13	117
3	9	81
4	3	27
5	1	9

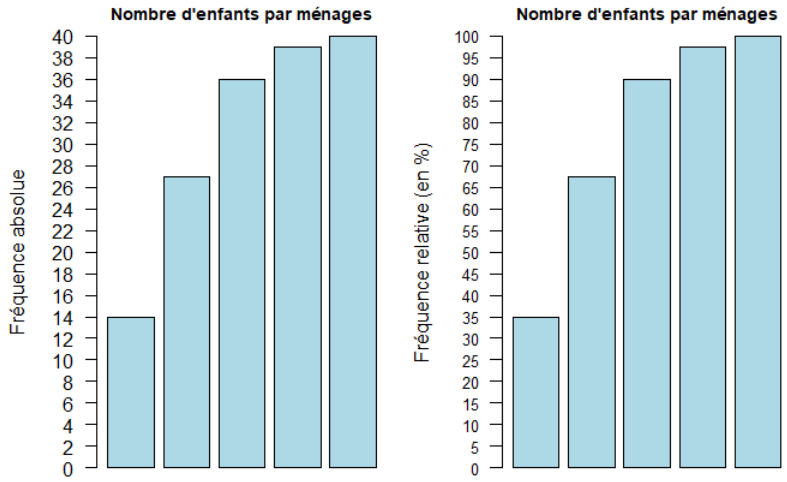
Il a déjà été précisé que ce type de graphique est peu précis et que lorsque deux catégories sont associées à des fréquences relativement similaires, il devient très difficile de repérer leur différence. Dans la Figure 3.6, les deux premières catégories semblent de taille égale, alors qu'elles ne le sont pas. Il était nettement plus aisé de le voir dans la Figure 3.5. C'est dû au fait que notre œil repère plus facilement une différence de 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 modalités de la variable à représenter par des bâtons dont la hauteur varie en fonction non plus de la fréquence (absolue ou relative) de chaque modalité mais en fonction de la fréquence **cumulée** associée à chaque modalité. Le dernier bâton du diagramme des fréquences cumulées aura nécessairement une hauteur égale à n si l'on représente les fréquences absolues, et une hauteur 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 les fréquences 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 bâton et dans le second cas, on peut déterminer le pourcentage de ménages dont le nombre d'enfants atteint au maximum la valeur représentée par le bâton.

Figure 3.7

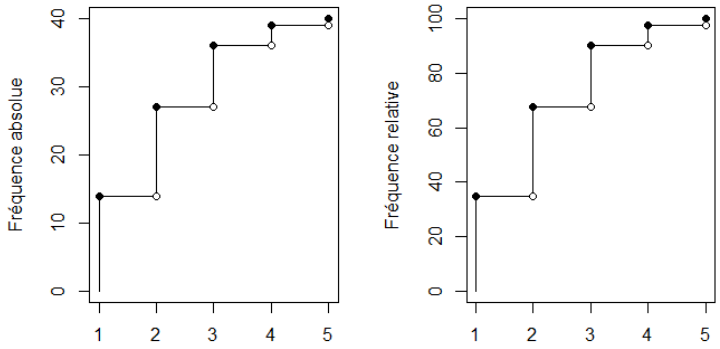
Diagramme des fréquences cumulées du nombre d'enfants par ménage



Il arrive que l'on parle de "diagramme en escalier", et que l'on choisisse une représentation plus proche de celle de la Figure 3.8, dans laquelle un point noir indique la fréquence associée à une catégorie en particulier. L'information fournie par les diagrammes en escalier et les diagrammes en bâtons des fréquences cumulées est identique.

Figure 3.8

Diagramme en escalier représentant les fréquences cumulées du nombre d'enfants par ménage



Variables quantitatives continues

Simplification des données

Dans la mesure où les variables ***continues*** ont une infinité de valeurs possibles, il n'est évidemment pas possible d'associer une fréquence à chaque valeur possible de la variable, similairement à ce que l'on faisait 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 (cela ne fait pas partie de la matière du cours). 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 préexistantes 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 3.9 et la Table 3.10, 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, de l'âge adulte et du troisième âge, on pourra créer 5 classes : une regroupant les individus de 0 à 3 ans, de 3 à 12 ans, de 12 à 18 ans, 18 à 65 ans et de plus de 65 ans (voir la Table 3.11).

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. Par exemple, le centre de la première classe de la Table 3.9 vaut 10000. Connaître cette valeur servira à estimer les **quantiles** de variables continues (ce concept sera très prochainement expliqué) ou la **moyenne**, lorsqu'on ne dispose pas des données brutes.

Table 3.9
Table du PIB par habitants de 45 pays d'Europe

Classe	**Centre de classe**	n_j	f_j **(en%)**	N_j	F_j **(en %)**
[0-20000[	10000	20	44.44	20	44.44
[20000-40000[	30000	9	20	29	64.44
[40000-60000[	50000	11	24.44	40	88.89
[60000-80000[	70000	3	6.67	43	95.56
[80000-100000[	90000	1	2.22	44	97.78
[100000-120000[	110000	1	2.22	45	100.00

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

Table 3.10
Table de fréquence de catégories d'âges d'amplitudes égales dans un échantillon de 20000 personnes

Classe	Centre	n_j	f_j (en%)	N_j	F_j (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 3.11
Table de fréquence de catégories d'âges d'amplitudes inégales dans un échantillon de 20000 personnes

Classe	Centre	n_j	f_j (en%)	N_j	F_j (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 abscisse. 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 les Figure 3.9 et 3.10.

Figure 3.9
Histogramme du PIB (cf. Table 3.9)

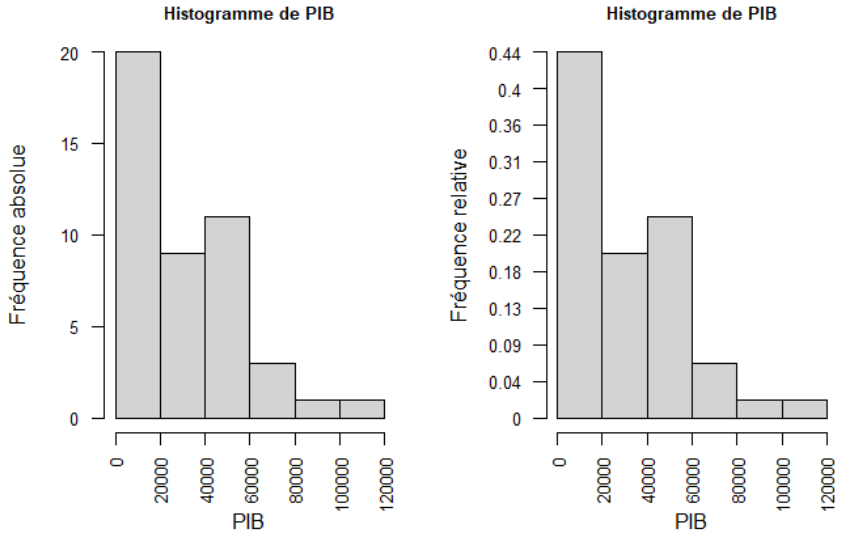
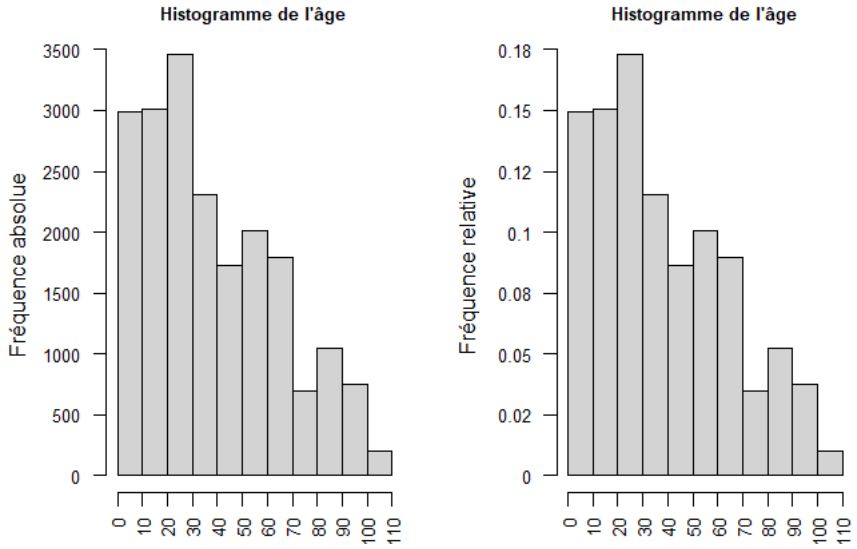


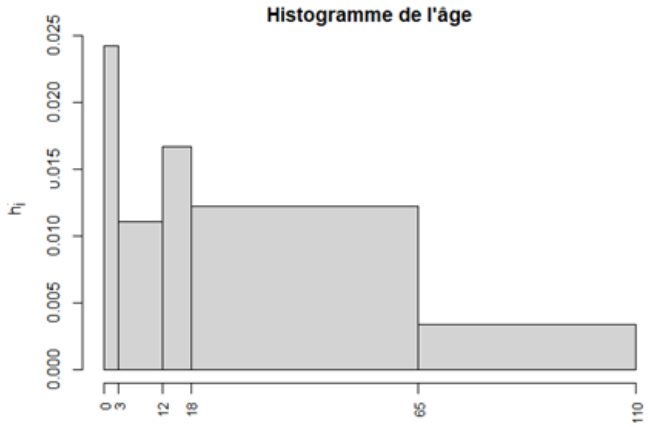
Figure 3.10
Histogramme de l'âge (cf. Table 3.10)



Par contre, lorsque les classes sont de tailles inégales, la hauteur du rectangle 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 entre les classes plus compliquée.

$h_j = \frac{n_j}{a_j}$, avec a_j = amplitude de la classe.

Figure 3.11 *Histogramme de l'âge (cf. Table 3.11)*



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 3.10 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 fréquences**, en reliant par des segments de droite le centre du côté supérieur de chaque rectangle.

Figure 3.12
Histogramme de la distribution du PIB et polygone des fréquences

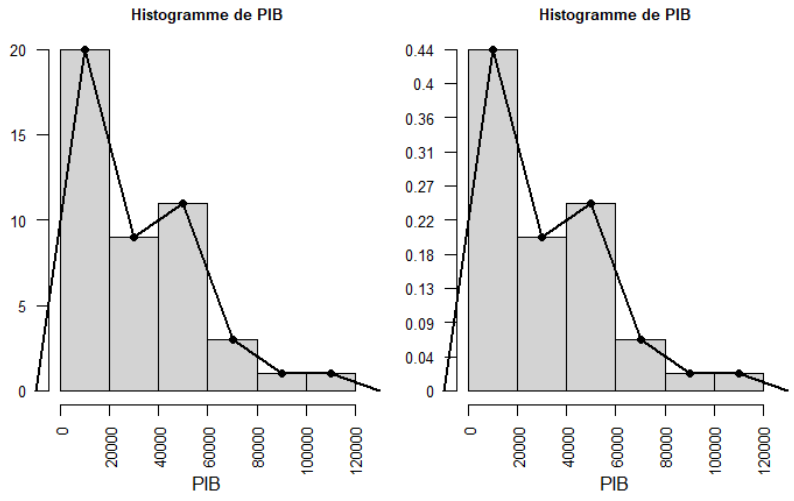
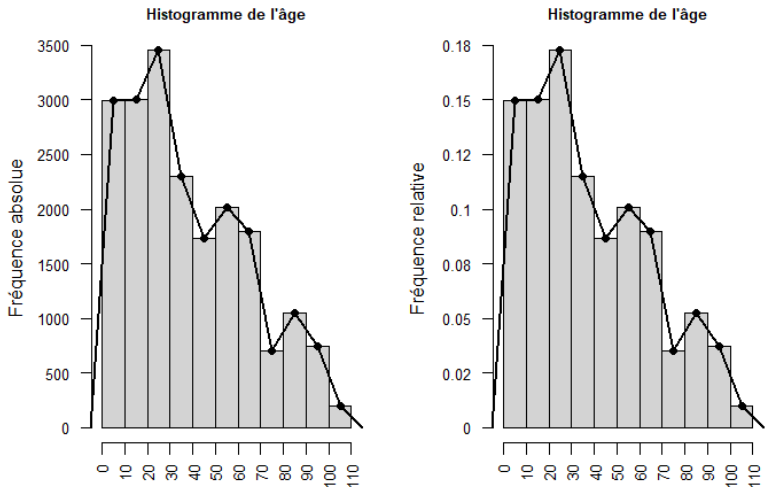


Figure 3.13

Histogramme de la distribution de l'âge (Table 3.10) et polygone des fréquences

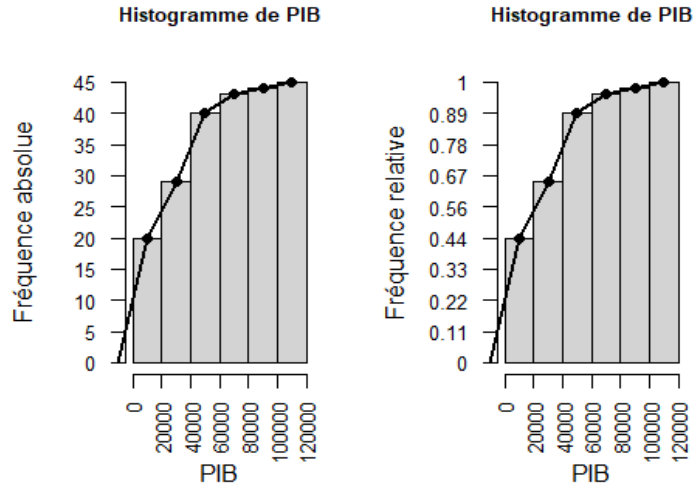


Histogramme des fréquences cumulées

Les histogrammes des fréquences cumulées ressemblent aux diagrammes en bâtons 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 vus dans la section précédente, il est possible de tracer le polygone des fréquences cumulées pour mieux rendre compte du caractère continu de la variable.

Figure 3.14
Histogramme de la distribution du PIB et polygone des fréquences

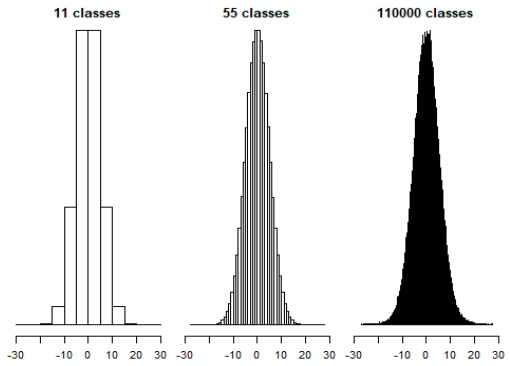


Remarque : les distributions

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 3.15, 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 110000 classes dans l'histogramme de droite). Au fur et à mesure que le nombre de classes augmente, 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 3.15
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 appelées des **quantiles**. Contrairement au calcul des 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 le calcul de 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 d'une série de données brutes

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

$$** Série1 ** 3,7,2,11,9,8,1,13,15$$

La médiane sera la valeur telle qu'il y ait la moitié des observations de la série qui soit de valeur inférieure, et l'autre moitié qui soit de valeur 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 $\frac{n+1}{2}$.

Voici la série ordonnée:

$$1,2,3,7,8,9,11,13,15$$

.

La médiane sera l'observation qui occupera le rang $\frac{9+1}{2}$ (soit la 5^e observation) de la série ordonnée. Or, la 5^e 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):

$$** Série2 ** 3,7,2,6,9,8,11,14,13,15$$

On commence par ordonnée la série :

$$2,3,6,7,8,9,11,13,14,15$$

On calcule ensuite le rang médian : $\frac{10+1}{2} = 5.5$. Cette fois, le rang n'est pas un nombre rond. On en déduit que la médiane sera la valeur entre la 5^e observation (valant 8) et la 6^e 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 :

$$** Série3 ** 2,3,6,7,8,8,11,13,14,15$$

Le rang médian vaut $\frac{10+1}{2} = 5.5$. La médiane sera donc comprise entre les valeurs 8 et 8, et vaudra donc : 8.

Pour chaque série, intéressons-nous uniquement aux valeurs qui se trouvent à gauche de la médiane. Le premier quartile d'une série sera la valeur telle que la moitié des observations sélectionnées lui est inférieure et l'autre moitié lui est supérieure. Similairement, si l'on s'intéresse uniquement aux valeurs à droite de la médiane pour chaque série, le troisième quartile d'une sera sera la valeur telle que 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 :

$$1,2,3,7,8,9,11,13,15$$

- La partie des données à gauche de la médiane est la suivante : 1, 2, 3, 7
  La valeur du premier quartile sera donc égale 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 égale à 12.

Si l'on revient à la série 2, soit la série suivante : $2,3,6,7,8,9,11,13,14,15$

- La partie des données à gauche de la médiane est la suivante : 2, 3, 6, 7, 8
  La valeur du premier quartile sera donc égale 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 égale à 13.

Enfin, si l'on revient à la série 3, soit la série suivante : $2,3,6,7,8,8,11,13,14,15$

- La partie des données à gauche de la médiane est la suivante : 2, 3, 6, 7, 8
  La valeur du premier quartile sera donc égale 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 égale à 13.

Au départ d'un tableau de fréquences

La Table 3.12 est le tableau de fréquence des données de la série 1.

x_j	n_j	f_j	N_j	F_j
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. Vous pourriez vous dire que 56% dépasse 50%. N'oubliez cependant pas que nous étudions ici des séries de données discrètes. Autrement dit, il n'y a pas de valeur intermédiaire possible entre 7 et 8. Vérifiez la valeur de la médiane de la série 1 au départ des séries brutes, et vous verrez qu'elle vaut bien 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~~ **Le premier quartile** 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~~ **Le troisième quartile** vaudra donc 11.

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

x_j	n_j	f_j	N_j	F_j
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~~ **Le premier quartile** 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~~ **Le troisième quartile** vaudra donc 13.

Enfin, la Table 3.14 est le tableau de fréquence des données de la série 2.

x_j	n_j	f_j	N_j	F_j
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 que c'est correct.

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

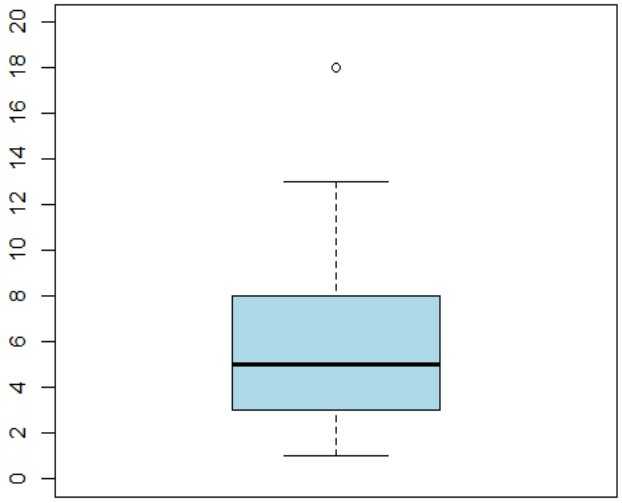
Réalisation graphique : le boxplot (appelée également "boîte à moustaches")

Table 3.15
Tableau de fréquence d'une série de données discrètes

x_j	n_j	f_j	N_j	F_j
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 3.16 est une représentation graphique des données de la Table 3.15.

Figure 3.16
*Boîte à moustaches représentant les données de la Table 3.1*5



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 troisième 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 interquartile** et se calcule comme suit : Q3 - Q1 = 8-3=5. L'écart interquartile 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 3.15). 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 interquartile (soit la longueur de la boîte) de part et d'autre de la boîte. Etant donné que dans l'exemple développé, l'écart interquartile vaut 5, les barrières s'éloigneront de $1{,}5 \times 5 = 7.5$ 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 variable 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'autre 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 cas de 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

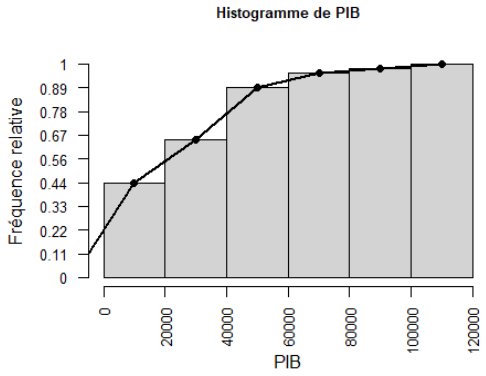
Au départ de données brutes

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 3.9 dont je reproduis ci-dessous le polygone des effectifs cumulés, pour votre commodité. 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 au centre de classe 10000 vaut 44.44 (soit une valeur inférieure à .5 ou 50%). La fréquence cumulée de la valeur 30000 vaut 64.44% (soit une valeur supérieure à .5 ou 50%). La valeur de la médiane se situera donc entre 10000 et 30000 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 $\frac{20000}{20} \times 5.56 = 5560$ sur l'axe des x. Par conséquent, la médiane sera estimée à 10000+5560 = 15560.

En suivant un raisonnement identique, on pourra déterminer la valeur de n'importe quel quantile. Le troisième quartile, par exemple, sera tel qu'il y ait exactement 75% des données à sa gauche et 25% à sa droite. La fréquence cumulée au centre de classe 30000 vaut 64.44% (soit une valeur inférieure à .75 ou 75%). La fréquence cumulée de la valeur 50000 vaut 88.89% (soit une valeur supérieure à .75 ou 75%). La valeur du troisième quartile se situera donc entre 30000 et 50000 et sera telle qu'il y ait exactement 75% des observations à sa gauche, et 25% des observations à sa droite. Une simple règle de trois permet de déterminer sa valeur : il y a une distance de 24.45 unités entre 88.89% et 64.44%. 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 10.56 unités entre 75% et 64.44%. Par règle de trois, on en déduit que cela correspond à une distance de $\frac{20000}{24.45} \times 10.56 = 8636.36$ sur l'axe des x. Par conséquent, le troisième quartile sera estimé à 30000+8638.04 = 38638.04.

Réalisation graphique : le boxplot

À nouveau, le boxplot se déroulera 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.

Synthèse sur les différents types de graphiques possibles

Nature de la variable			
Qualitative	**Quantitative**		
(Nominale ou ordinale)	Discrète	Continue	
X	X		Diagramme en bâtonnets
X	X		Diagramme circulaire
	X		Diagramme des fréquences cumulées/Diagramme en escalier
		X	Histogramme/ Polygone des effectifs
		X	Histogramme des fréquences cumulées/Polygone des effectifs cumulés
	X	X	Boîte à moustaches

Exercices de fin de chapitre

1. Pour 44 domiciliée dans le brabant wallon, on dispose d'une information sur le statut matrimonial.

i	Statut matrimonial (X_i)	i	Statut matrimonial (X_i)	i	Statut matrimonial (X_i)	i	Statut matrimonial (X_i)
1	Veuf	12	marié	23	Divorcé	34	marié
2	Divorcé	13	séparé	24	Célibataire	35	célibataire
3	Divorcé	14	séparé	25	Divorcé	36	séparé
4	Célibataire	15	veuf	26	veuf	37	célibataire
5	Séparé	16	marié	27	séparé	38	veuf
6	Divorcé	17	célibataire	28	marié	39	séparé
7	Marié	18	divorcé	29	divorcé	40	marié
8	Veuf	19	marié	30	célibataire	41	célibataire
9	Célibataire	20	divorcé	31	marié	42	marié
10	Divorcé	21	divorcé	32	divorcé	43	célibataire
11	Divorcé	22	célibataire	33	veuf	44	célibataire

Répondez aux questions suivantes:

1.1. A quel type de variable correspond la variable étudiée ?
1.2. Construisez le tableau de la distribution des fréquences absolues et relatives, associé à ces observations.
1.3. Dessinez le diagramme en barres des fréquences absolues et relatives correspondant. L'ordre d'apparition des différentes catégories est-il important ?
1.4. Dessinez le diagramme circulaire. En quoi ce type de graphique est-il problématique ?

2. Considérons la série statistique ordonnée suivante relative aux âges des membres d'un club sportif (à traiter comme une variable **continue**) :

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

2.1. Groupez cette série en classes d'amplitude 5, dont les centres valent respectivement 19,24,29,34 et 39 (à vous de déterminer les valeurs des bornes inférieures et supérieures de chaque classe, de sorte à pouvoir procéder aux regroupements). Construisez le tableau de la distribution observée des fréquences absolues (effectifs) et relatives correspondant.

2.2. Dessinez l'histogramme et le polygone des effectifs de cette distribution observée.

2.3. Dessinez l'histogramme des fréquences cumulées, et le polygone des effectifs cumulés.

2.4. Sur base des valeurs de la série, en considérant la variable comme étant continue, calculez : la médiane, les premier et troisième quartiles, l'écart interquartile. Quel type de graphique permet de représenter toutes ces valeurs ?

3. Soit le tableau de fréquences suivants, relatif au nombre d'enfants à charge dans une famille.

age (x_j)	n_j
0	80
1	95
2	125
3	75
4	60
5	65

3.1. Sur base des informations de ce tableau, calculez les fréquences absolues cumulées, les fréquences relatives, les fréquences relatives cumulées exprimées en pourcentage.

3.2. Pour combien de familles a-t-on déterminé le nombre d'enfants à charge ? Quelle est la notation pour décrire le nombre total de familles ?

3.3. Dans combien de familles y a-t-il exactement 3 enfants à charge ? Donnez la notation mathématique adéquate pour décrire cette quantité, et déterminez sa valeur.

3.4. Dans combien de familles y a-t-il strictement moins de 2 enfants ? Donnez la notation mathématique adéquate pour décrire cette quantité, et déterminez sa valeur.

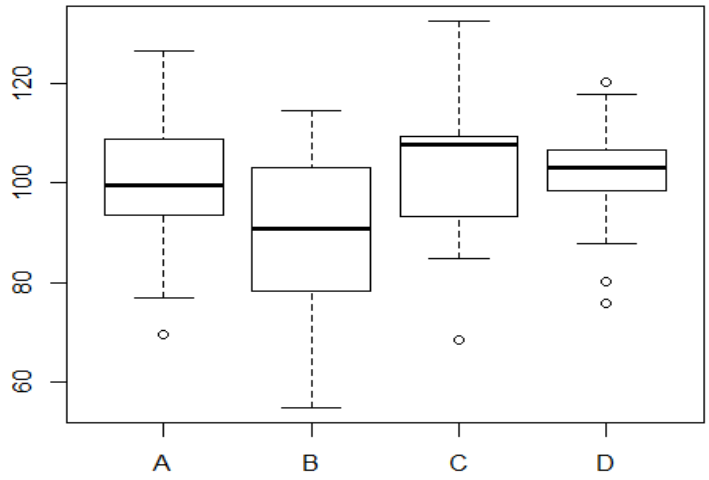
3.5. Dans quelle proportion de familles y a-t-il 2 enfants ou moins ? Donnez la notation mathématique adéquate pour décrire cette quantité, et déterminez sa valeur.

3.6. Représentez les fréquences cumulées sous forme de diagramme des fréquences cumulées.

3.7. Sur base des valeurs de la série, calculez : la médiane, les premier et troisième quartiles, l'écart interquartile. Quel type de graphique permet de représenter toutes ces valeurs ?

4. Dans le graphique suivant, les diverses boîtes à moustaches représentent le QI de 4 groupes de sujets (qu'on appellera A, B, C et D).

Boîte à moustaches représentant le QI de 4 groupes : A, B, C et D



4.1. Combien vaut (approximativement) la médiane du groupe A ?

4.2. Considérant chaque groupe séparément (A, B, C et D), y a-t-il des sujets ayant un QI anormalement bas par rapport au reste du groupe ? Y a-t-il des sujets ayant un QI anormalement haut ?

4.3. Quel est le groupe ayant le plus grand écart interquartile ?

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** : mesures qui permettent de représenter au mieux un ensemble données par une valeur unique.
2) Les mesures de **dispersion** : mesures 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
Y_i	1	5	2	8	8	9	9	1	1	1	8	8	7	10	8

(b) Tableau de fréquence

y_j	1	2	3	4	5	6	7	8	9	10
n_j	4	1	0	0	1	0	1	5	2	1
f_j	0.27	0.07	0.00	0.00	0.07	0.00	0.07	0.33	0.13	0.07

Pour rappel, on utilise une lettre majuscule indicée par *i* pour décrire les données brutes (Y_i = la valeur du i^{ème} 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 (y_j = la jème valeur que peut prendre la variable Y).

Mode

Le mode correspond à la classe la plus représentée. Dans la Table 4.1 (a ou b), vous constaterez 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 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
X_i	19	25	21	27	30	32	30	19	21	21	24	27	26	30	26

(b) Tableau de fréquence

y_j	19	20	21	22	23	24	25	26	27	28	29	30	31	32
n_j	2	0	3	0	0	1	1	2	2	0	0	3	0	1
f_j	0.1	0.0	0.2	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.2	0.0	0.0
	3	0	0	0	0	7	7	3	3	0	0	0	0	7

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, Y_{14}= 100 ou même 1000 au lieu de 10. Le mode resterait identique, ce serait 8. Ça 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 chaque observation de la variable dont on veut calculer la moyenne, et en divisant cette somme par le nombre d'observations additionnées. La moyenne arithmétique 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 $\overline{Y}$.

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

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$\overline{Y} = \frac{1+5+2+8+8+9+9+1+1+1+8+8+7+10+8}{15} = \frac{86}{15} = 5.73$$

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 x_j et les fréquences absolues

$$\overline{X} = \frac{\sum_{j=1}^{k} n_j \times x_j}{n}$$

En utilisant les valeurs x_j et les fréquences relatives

$$\overline{X} = \sum_{j=1}^{k} f_j \times x_j$$

En appliquant ces formes à l'ancienneté fournie dans la Table 4.1, on obtient ceci:

En utilisant les valeurs x_j et les fréquences absolues

$$\overline{Y} = \frac{4 \times 1 + 1 \times 2 + 0 \times 3 + 0 \times 4 + 1 \times 5 + 0 \times 6 + 1 \times 7 + 5 \times 8 + 2 \times 9 + 1 \times 10}{15}$$

$$= \frac{4 + 2 + 0 + 0 + 5 + 0 + 7 + 40 + 18 + 10}{15} = \frac{86}{15} = 5.73$$

En utilisant les valeurs x_j et les fréquences relatives

$$\overline{Y}$$
$$= 0.27 \times 1 + 0.07 \times 2 + 0 \times 3 + 0 \times 4 + 0.07 \times 5 + 0 \times 6 + 0.07 \times 7 + 0.33 \times 8 + 0.13 \times 9 + 0.07 \times 10 = 5.73$$

Bien entendu, les trois formules donnent une valeur identique de la moyenne (bien sûr, à condition d'utiliser des valeurs non arrondies pour les fréquences relatives). C'est parce qu'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 Y_{14}= 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 graphiquement avant de déterminer des valeurs algébriques.

Mesures de dispersion

Une mesure de dispersion sert à quantifier à quel point les données ont tendance à 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 illustrer cela, considérons les deux séries suivantes qui représentent (fictivement) les notes (sur 10) obtenues par 10 étudiants en mathématiques et en français.

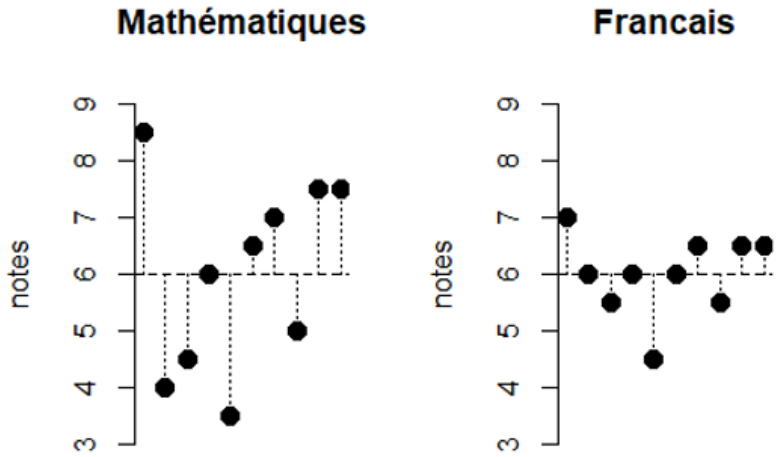
Table 4.3
Notes (/10) obtenues par 10 des étudiants en mathématiques et en français

i	Math (X_i)	Français (Z_i)
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 généralement, 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 1 points de la moyenne alors qu'en mathématiques, les notes peuvent s'éloigner jusqu'à 2.5 points 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

Étendue

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

$$étendue = max(X_i) - min(X_i)$$

Dans l'exemple des notes de cours de la Table 4.3, on trouve que l'étendue des notes en français vaut 7 - 4.5=2.5, et que celle des notes en mathématiques vaut 8.5 - 3.5= 5.

L'avantage de cette mesure est qu'elle est très simple à calculer. En contrepartie, 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'œil 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).

$$EIQ = Q_3 - Q_1$$

Dans l'exemple des notes de cours de la Table 4.3, on trouve que l'écart interquartile des notes en français vaut 6.5-5.5=1, et que celui des notes en mathématiques vaut 7.5-4.5= 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.

$$E_i = X_i - \overline{X_i}$$

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

i	$math(X_i)$	E_i
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 français

i	$francais(Z_i)$	E_i
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 : $\sum_{i=1}^{n} E_i = 2.5 - 2 - 1.5 + 0 - 2.5 + 0.5 + 1 - 1 + 1.5 + 1.5 = 0$

Pour le cours de français : $\sum_{i=1}^{n} E_i = 1 + 0 - 0.5 + 0 - 1.5 + 0 + 0.5 - 0.5 + 0.5 + 0.5 = 0$

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

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

$$EMA = \frac{\sum_{i=1}^{n} | X_i - \overline{X}|}{n}$$

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

$$\frac{2.5 + 2 + 1.5 + 0 + 2.5 + 0.5 + 1 + 1 + 1.5 + 1.5}{10} = 1.4.$$

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

$$\frac{1 + 0 + 0.5 + 0 + 1.5 + 0 + 0.5 + 0.5 + 0.5 + 0.5}{10} = 0.5$$

.

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'écart-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 écart-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:

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

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

$$\frac{6.25 + 4 + 2.25 + 0 + 6.25 + 0.25 + 1 + 1 + 2.25 + 2.25}{10} = 2.55$$

.

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 : $\sqrt{2.55} = 1.6$. 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 :

$$\frac{1 + 0 + 0.25 + 0 + 2.25 + 0 + 0.25 + 0.25 + 0.25 + 0.25}{10} = 0.45$$

.

L'écart-type, soit la racine carrée de la variance vaut $\sqrt{0.45} = 0.67$. 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 aberrantes. Cette sensibilité est accentuée par le fait que l'on élève les écarts au carré. Ces mesures restent malgré tout les mesures de dispersion les plus utilisées en pratique.

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**.

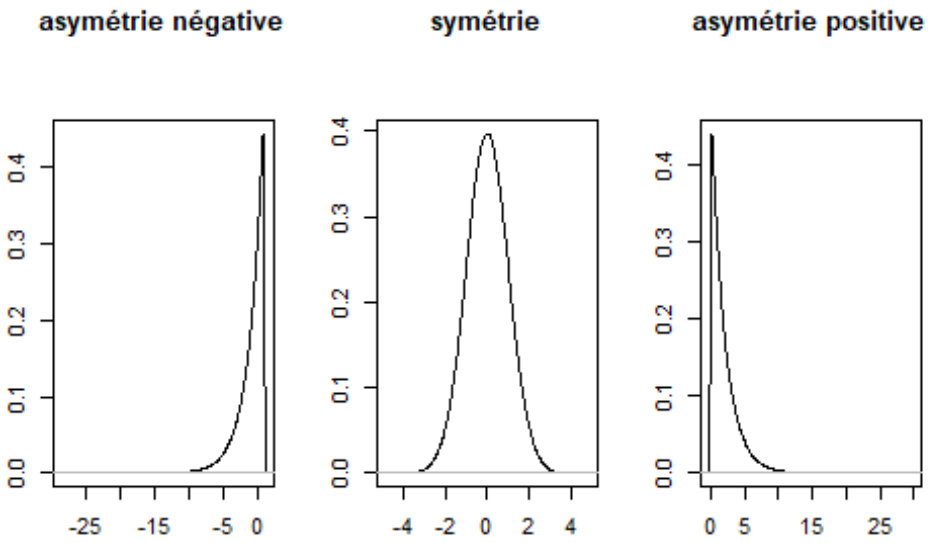
Mesure d'asymétrie

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 :

$$G1 = \frac{M_3}{S^3}$$

$$avec, M_3 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^3}{n}$$

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

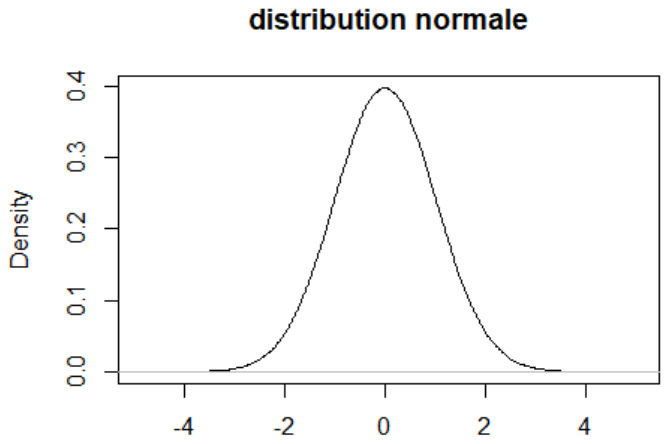
Il n'est pas important que vous puissiez calculer cela manuellement. Par contre, vous devez pouvoir interpréter sa valeur :

- Lorsque la distribution est parfaitement symétrique, le coefficient G1 vaut exactement 0.
- Lorsque la distribution présente une asymétrie positive, la valeur du coefficient G1 est supérieure à 0.
- Lorsque la distribution présente une asymétrie négative, la valeur du coefficient G1 est inférieure à 0.

Mesure d'aplatissement

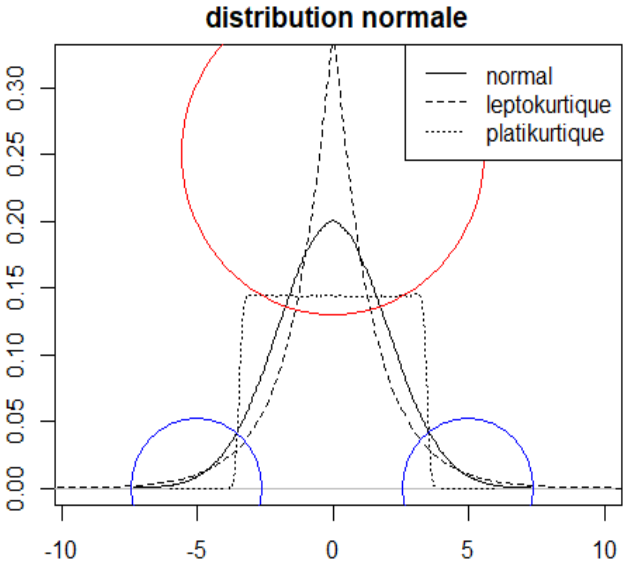
Pour correctement comprendre la mesure d'aplatissement, il faut savoir que la distribution de référence est la distribution normale (ou courbe en cloche, représentée sur la Figure 4.3)

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:

$$G2 = \frac{M_4}{S^4} - 3$$

$$avec, M_4 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^4}{n}$$

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

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.

Exercices de fin de chapitre

1. Voici une série statistique :

6,1,3,15,6,12,3,13,15,15

1.1. Calculez la moyenne de trois manières différentes.
1.2. Calculez la médiane ainsi que le mode de la série.
1.3. Calculez l'étendue et l'écart interquartile de la série.
1.4. Calculez la variance et l'écart-type de la série.

2. Voici une série statistique:

9,12,15,8,75

2.1. Calculez la moyenne de trois manières différentes. La valeur est-elle représentative de la série ? Commentez.
2.2. Calculez la médiane ainsi que le mode de la série.
2.3. Calculez l'étendue et l'écart interquartile de la série.
2.4. Calculez la variance et l'écart-type de la série.

3. Voici un tableau de fréquence :

3.1. Calculez la moyenne de la série.
3.2. Calculez la médiane ainsi que le mode de la série.
3.3. Calculez l'étendue et l'écart interquartile de la série.

4. Voici trois distributions. L'une des trois distributions a une valeur G2 (asymétrie) de 1.26. Une autre a une valeur G2 de -0.82et enfin, la dernière a une valeur G2 de 0. Veuillez determiner quelle valeur se rapporte à chaque distribution.

Chapitre 5 : Relation Entre Deux Variables Continues : Le Coefficient r de Pearson

Jusque là, nous avons passé beaucoup de temps à étudier comment il était possible de représenter graphiquement ou algébriquement les variables prises isolément. Le dernier objectif de ce cours est d'étudier comment deux variables peuvent varier **en même temps**, ou **covarier**. Cela relève de la statistique **bivariée** (par opposition à statistique **univariée**, lorsqu'on étudie les variables séparément les unes des autres).

Revenons à l'exemple de la Table 4.3 dans laquelle nous avions étudié la note de 10 étudiants à deux cours: mathématiques et français.

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

i	Math (X_i)	Français (Z_i)
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

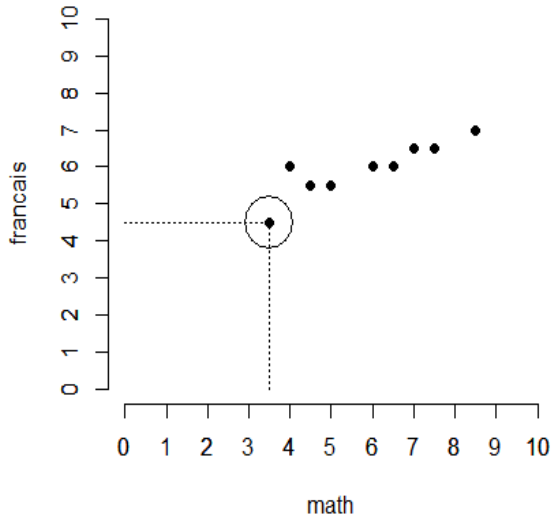
Nous avons déjà déterminé précédemment dans quelle mesure les scores de chacun des cours variaient d'un individu à l'autre (nous avions trouvé, respectivement pour les cours de mathématiques et de français, des variances de 2.55 et 0.45 et donc par conséquent, des écart-types de respectivement 2 et 0.64). On se demande à présent s'il existe un **lien** entre la manière dont ces deux cours varient simultanément d'un sujet à l'autre. Autrement dit, on se demande si ces deux variables **covarient**.

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.

La Figure 5.1 montre le graphe qui reprend les valeurs aux cours de mathématiques et de français les unes par rapport aux autres. On y voit un nuage de points, chaque point 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 français (coordonnée sur l'axe des y).

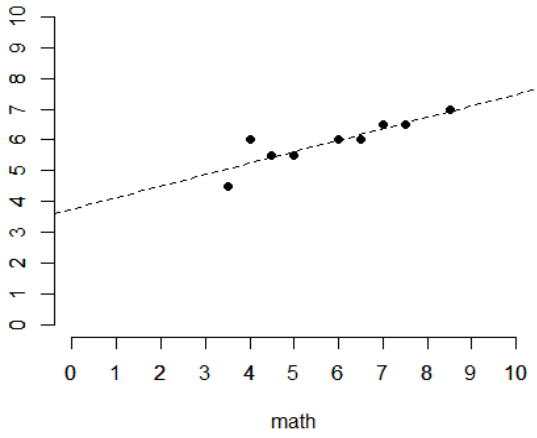
Figure 5.1
Nuage de points représentant la relation entre la note à un cours de math et la note à un cours de français



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 bons points en mathématiques semblent avoir également de moins bons points en français.

Lorsque les points d'un nuage semblent se répartir plus ou moins autour d'une droite, on parle de **relation linéaire**. Le meilleur moyen de représenter l'allure générale de la corrélation est de tracer la droite qui se rapproche le plus possible de l'ensemble des points et que l'on appelle dès lors 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

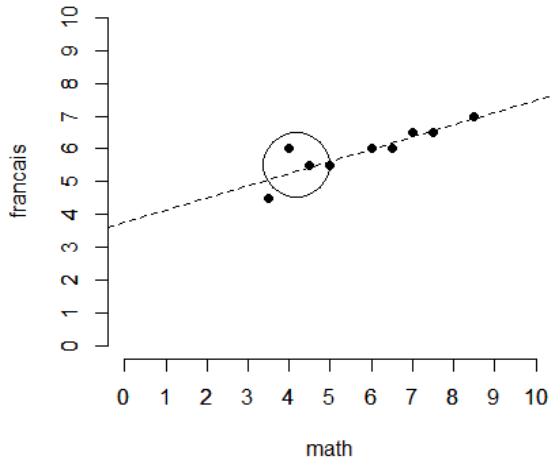


On constate que l'ensemble des points semblent se rapprocher d'une droite croissante. Cela signifie que le score des deux variables tend à varier dans le même sens : lorsque la note au cours de

mathématiques augmente, il en est de même pour la note au cours de français. On parle alors de relation linéaire **positive**.

Attention : il s'agit d'une **tendance** des données. Autrement dit, ce n'est pas nécessairement vrai pour chaque point du nuage. Regardez les deux points entourés dans la reproduction du nuage de points ci-dessous. Le point le plus à gauche dans le cercle représente un étudiant ayant obtenu une note de 4 en mathématiques, et une note de 6 en français. Le point le plus à droite dans le cercle représente un étudiant ayant obtenu une note de 4.5 en mathématiques, soit une note plus élevée que le premier sujet. Pourtant, la note qu'il a obtenu en français est plus faible (5.5). Lorsque les points sont parfaitement alignés, on parole de relation linéaire **parfaite**.

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



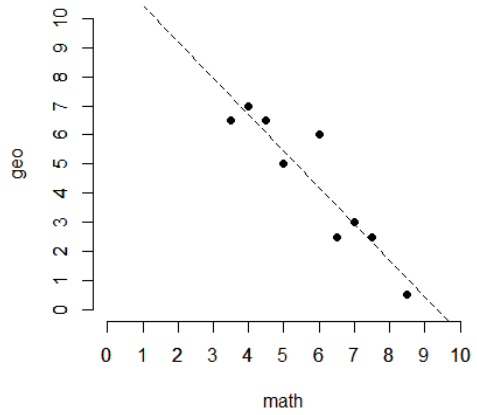
Imaginons à présent que nous souhaitions étudier la relation entre le cours de mathématiques et un cours de géographie suivi par les 10 mêmes étudiants.

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

i	math(X_i)	geographie(W_i)
1	8.5	0.5
2	4.0	7.0
3	4.5	6.5
4	6.0	6.0
5	3.5	6.5
6	6.5	2.5
7	7.0	3.0
8	5.0	5.0
9	7.5	2.5
10	7.5	2.5

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 5.4
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. Cela signifie que le score des deux variables tend à varier dans des sens opposés : lorsque le score de mathématiques augmente, celui de géographique diminue, et réciproquement). On parle alors de relation linéaire négative entre deux variables. De même qu'expliqué précédemment, lorsque les points seront parfaitement alignés, on parlera alors de relation linéaire négative **parfaite**.

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

Table 4.6
Notes (/10) obtenues par 10 des étudiants en éducation physique et en anglais

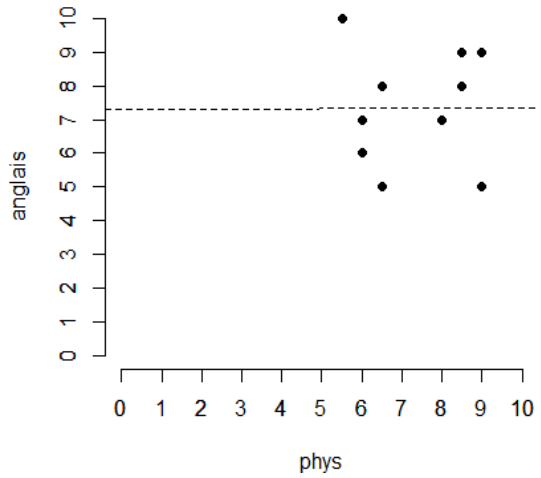
i	**education physique (V_i)**	**Anglais (V_i)**
1	5.5	10
2	9.0	5
3	6.0	6
4	6.0	7
5	6.5	5
6	6.5	8
7	8.0	7
8	8.5	9
9	8.5	8
10	9.0	9

Contrairement aux deux situations précédentes, il ne semble pas possible de résumer correctement le nuage de points par une droite croissante ou décroissante. Il ne semble pas y avoir de relation linéaire entre les notes obtenues au cours de géographiques et celles obtenues au cours d'histoire.

Cela se confirme par la droite qui se rapproche le plus près possible de tous les points, on trouve une droite de pente pratiquement nulle. C'est ce qui arrive en absence de **relation linéaire** entre deux variables.

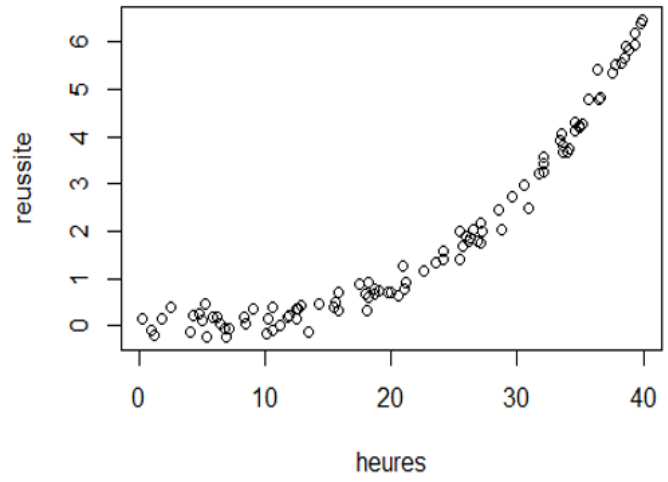
Figure 5.5

Nuage de points représentant la relation entre la note à un cours d'éducation physique et la note à un cours d'anglais



Depuis le début de ce chapitre, le terme "linéaire" revient régulièrement, parce que nous essayons de résumer le nuage de points par une droite. Il est important de bien comprendre qu'il peut exister une relation autre que linéaire entre deux variables. Imaginons la Figure 5.6 qui représente la relation entre le temps d'études du cours de statistiques et le résultat à l'examen.

Figure 5.6 *Nuage de points représentant la relation entre le nombre d'heures d'études et la note obtenue à l'examen de statistiques*



On pourrait imaginer qu'en commençant à étudier, assimiler tous ces nouveaux concepts statistiques demande beaucoup d'énergie et de temps, si bien que parmi les étudiants qui n'ont étudié que quelques heures, une augmentation du temps d'étude ne fait augmenter la note que faiblement. Par contre, chez ceux qui ont dépassé la période frustrante de familiarisation avec toutes ces notations mathématiques compliquées, on observe que chaque nouvelle heure consacrée à étudier ce cours amène à une augmentation de plus en plus importante de la note finale. Le lien entre les deux matières est indéniable, par contre, il ne sera que mal représenté par une droite. Ceci est une illustration de relation **non linéaire**. Ce type de relation ne fait pas partie de la matière que

verrons dans ce cours. Il est cependant important de s'en rendre compte. En prendre conscience aide à comprendre qu'une absence de relation se traduira toujours par une relation linéaire nulle. Par contre, une relation linéaire nulle ne signifie pas forcément qu'il n'y a pas de relation entre deux variables: il est toujours possible qu'il existe une relation autre que linéaire. Or, le seul moyen de s'en rendre compte, c'est de réaliser un nuage de points.

Détermination algébrique de la relation entre deux variables

Quantifier la relation entre deux variables

Covariance

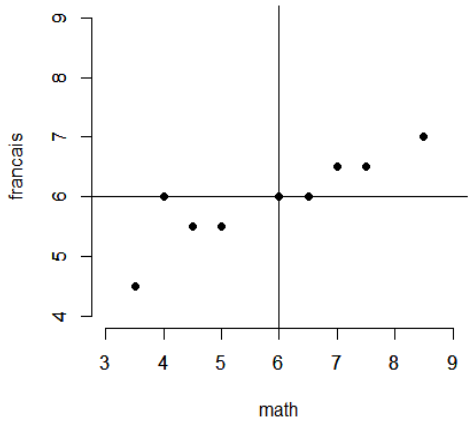
La covariance entre deux variables se calcule comme suit:

$$S_{XY} = \frac{\sum_{i=1}^{n}(X - \overline{X})(Y - \overline{Y})}{n}$$

Voyons à présent en quoi cette expression représente une mesure du lien qui existe entre la variable X et la variable Y.

Nous avions précédemment déterminé graphiquement (voir la Figure 5.1) que la relation linéaire entre les cours de mathématiques et de français était positive pour les dix étudiants considérés à titre illustratif. Ci-dessous, reproduisons cette figure mais en la divisant en 4 quadrants distincts : tous les points qui tombent dans les deux quadrants du haut correspondent aux étudiants dont le score obtenu en français est supérieur à la moyenne, et tous les points qui tombent dans les deux quadrants du bas correspondent aux étudiants dont le score obtenu en français est inférieur à la moyenne. De plus, tous les points qui tombent dans les deux quadrants de droite correspondent aux étudiants dont le score obtenu en mathématiques est supérieur à la moyenne, et tous les points qui tombent dans les deux quadrants de gauche correspondent aux étudiants dont le score obtenu en mathématiques est inférieur à la moyenne.

Figure 5.7 *Reproduction du nuage de points représentant la relation entre la note à un cours de math et la note à un cours de français, découpé en 4 quadrants*



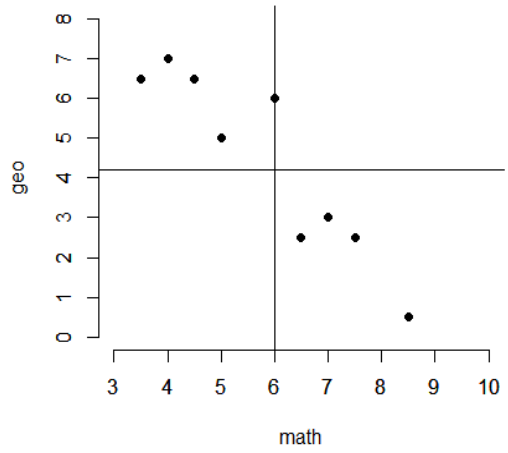
Dans la mesure où la relation linéaire est positive entre les deux cours, lorsque le score en mathématiques est supérieur à la moyenne, le score associé en français sera généralement lui aussi supérieur à la moyenne. Pour tous les points du quadrant supérieur droit, $X_i - \overline{X}$ sera supérieur à

zéro et il en sera de même pour $Y_i - \overline{Y}$. Or, +*+=+. Par ailleurs, lorsque le score en mathématiques est inférieur à la moyenne, le score associé en français sera généralement lui aussi inférieur à la moyenne. Pour tous les points du quadrant inférieur gauche, $X_i - \overline{X}$ sera négatif, et il en sera de meme pour $Y_i - \overline{Y}$. Or, -*-=+. Cela signifie que le terme $\sum_{i=1}^{n} =$ consistera à additionner toute une série de nombres positifs, et par voie de conséquence, que la mesure de covariance sera positive.

Nous avions également déterminé précédemment graphiquement (voir la Figure 5.2) que la relation linéaire entre les cours de mathématiques et de géographie était négative pour les dix étudiants considérés à titre illustratif. Ci-dessous, reproduisons cette figure mais en la divisant en 4 quadrants distincts, en respectant la même logique que précédemment : tous les points qui tombent dans les deux quadrants du haut correspondent aux étudiants dont le score obtenu en géographie est supérieur à la moyenne, et tous les points qui tombent dans les deux quadrants du bas correspondent aux étudiants dont le score obtenu en géographie est inférieur à la moyenne. De plus, tous les points qui tombent dans les deux quadrants de droite correspondent aux étudiants dont le score obtenu en mathématiques est supérieur à la moyenne, et tous les points qui tombent dans les deux quadrants de gauche correspondent aux étudiants dont le score obtenu en mathématiques est inférieur à la moyenne.

Dans la mesure où la relation linéaire est négative entre les deux cours, lorsque le score en mathématiques est supérieur à la moyenne, le score associé en géographie sera généralement inférieur à la moyenne. Pour tous les points du quadrant inférieur droit, $X_i - \overline{X}$ sera positif et $Y_i - \overline{Y}$ sera négatif. Or, +*-=-. Par ailleurs, lorsque le score en mathématiques est inférieur à la moyenne, le score associé en français sera par contre supérieur à la moyenne. Pour tous les points du quadrant supérieur gauche, $X_i - \overline{X}$ sera négatif, par contre $Y_i - \overline{Y}$ sera positif. Or, -*+=+. Cela signifie que le terme $\sum_{i=1}^{n}$ consistera à additionner toute une série de nombres négatifs, et par voie de conséquence, que la mesure de covariance sera négative.

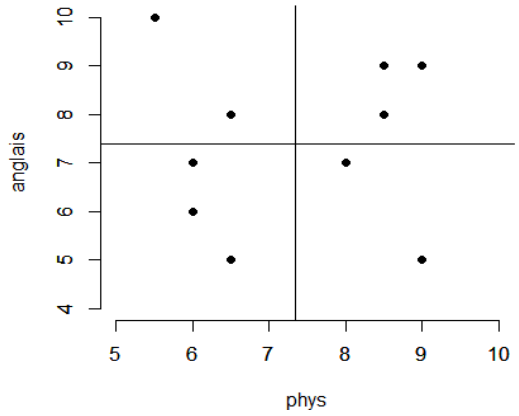
Figure 5.8 *Reproduction du nuage de points représentant la relation entre la note à un cours de math et la note à un cours de géographie, découpé en 4 quadrants*



Comme dernier exemple, nous avions déterminé précédemment graphiquement une absence de relation linéaire entre les cours d'éducation physique et celui d'anglais. Cela signifiait que lorsque la note au cours d'éducation physique augmentait, celle du cours d'anglais pouvait tantôt

augmenter, tantôt diminuer. Graphiquement, cela amène à observer un nuage de points avec des points qui se répartissent aléatoirement dans les 4 quadrants (voir la Figure 5.9).

Figure 5.9 *Reproduction du nuage de points représentant la relation entre la note à un cours de géographie et la note à un cours d'anglais, découpé en 4 quadrants*



Il a déjà été déterminé que pour les quadrants inférieur gauche et supérieur droit, le produit $(X - \overline{X})(Y - \overline{Y})$ était positif, et qu'au contraire, pour les quadrants supérieur gauche et inférieur droit, le produit était négatif. Cela signifie que le terme $\sum_{i=1}^{n}$ consistera à additionner toute une série de nombres dont certains sont positifs et d'autres sont négatifs, et par voie de conséquence, que la mesure de covariance sera proche de zéro (puisque les positifs et les négatifs s'annulent).

EN RESUME :

- Une mesure de covariance négative traduit une relation linéaire négative entre deux variables.
- Une mesure de covariance positive traduit une relation linéaire positive entre deux variables.
- Une mesure de covariance nulle traduit une relation linéaire nulle entre deux variables.

Corrélation

La covariance est une mesure très efficace du lien entre deux variables. Cependant, elle souffre d'un défaut : elle donne une valeur difficilement interprétable en tant que telle. On sait qu'elle augmente, diminue ou s'annule en fonction du lien, mais il n'y a pas de valeur plafond ni de valeur plancher. Admettons que la valeur d'une covariance soit de 1352 (nombre purement choisi au hasard), cela ne permet de dire qu'une seule chose : que le lien est positif. Mais ce lien est-il fort, faible, intéressant, inintéressant ? En fait, tout dépend de l'unité de mesure des variables. Si nous étudions des variables à l'aide de valeurs qui tournent autour de 1 ou 2, alors une covariance de 1352 est probablement un lien extrêmement fort entre les deux variables. En revanche, si les variables ont des valeurs qui tournent autour de 10000000 ou 20000000, alors cette même covariance de 1352 représentera probablement un lien tout à fait dérisoire, que l'on peut considérer comme proche de zéro.

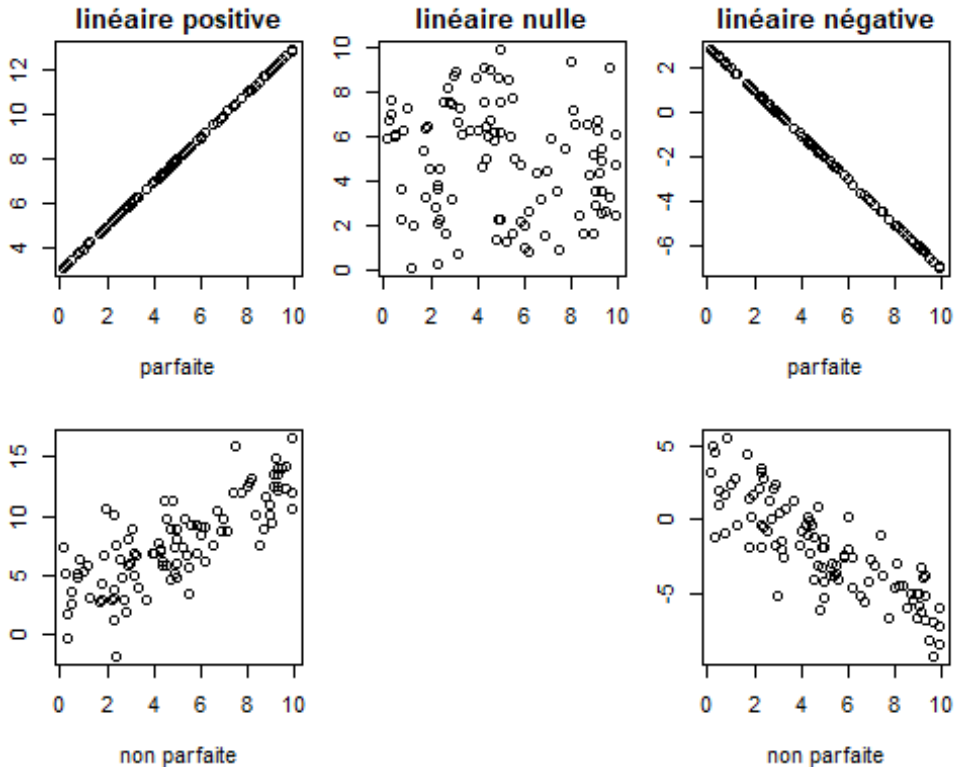
Heureusement, il existe une solution qui permet de contourner ce problème. En effet, si l'on divise la covariance par les écarts-types respectifs de ces deux variables, on obtient une mesure réduite dont l'interprétation est beaucoup plus évidente, que l'on appelle la mesure de **corrélation r de Pearson** :

$$r_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{[(X - \overline{X})(Y - \overline{Y})]/n}{S_X S_Y}$$

Cette mesure est nécessairement comprise entre -1 et 1.

Dans ce cas, -1 correspond à une corrélation linéaire négative parfaite entre deux variables (tous les points du nuage seront parfaitement superposés à une droite de régression décroissante), +1 correspond à une corrélation linéaire positive parfaite (tous les points du nuage seront parfaitement superposés à une droite de régression croissante) et 0 correspond à une absence de corrélation linéaire.

Figure 5.10 *Synthèse*



Toutes les valeurs comprises entre 0 et +1 (non inclus) ou entre 0 et -1 (non inclus) correspondent aux situations où il existe une corrélation linéaire entre les deux variables, mais non parfaite (tous les points tendront à se rapprocher d'une droite croissante ou décroissante, mais sans être parfaitement superposés à cette droite).

Plus les mesures r de Pearson se rapprochent de -1 ou de 1, plus la relation linéaire entre les deux variables sera dite forte. Bien qu'il existe de nombreuses balises, nous nous fierons à celles définies par un dénommé Cohen, d'après qui …

- Une mesure de corrélation comprise entre -.3 et +.3 est faible.
- Une mesure de corrélation comprise entre -.3 et -.5 ou comprise entre .3 et .5 est dite moyenne.
- Une mesure de corrélation inférieure à -.5 ou supérieure à +.5 est dite forte.

A titre d'exercice, vérifiez que $S_{Math,Francais} = 0.95$, $S_{Math,Geographie} = -0.93$ et $S_{Educationphysique,Anglais} = 0$.

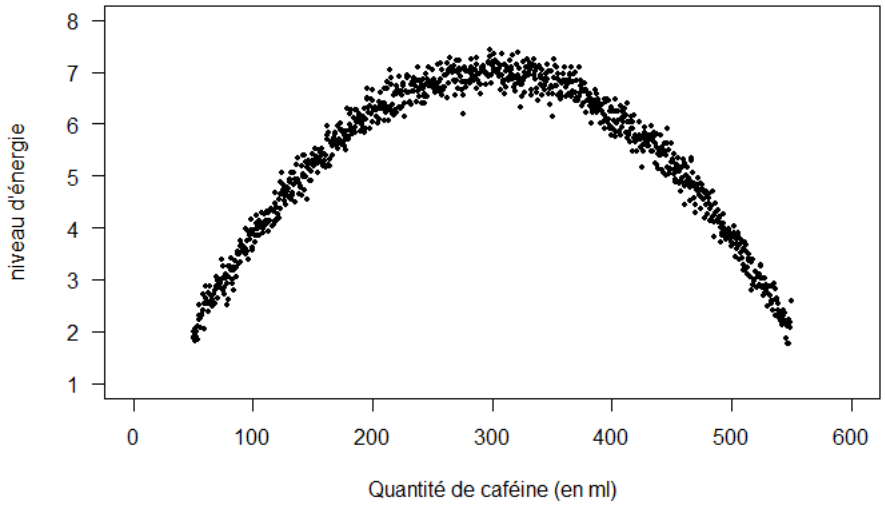
En transformant ces mesures de covariances en mesure de corrélation, on trouve que la corrélation entre les cours de mathématiques et de français vaut $r_{Math,Francais} = \frac{0.95}{1.6 \times 0.67} = 0.89$, soit une corrélation forte

La corrélation entre les cours de mathématiques et de géographie vaut $r_{Math,Geo} = \frac{-3.2}{1.6 \times 2.15} = -0.93$, soit une corrélation forte

Enfin, la corrélation entre les cours d'éducation physique et d'anglais vaut $r_{phys,anglais} = \frac{0.01}{1.3 \times 1.62} = 0.005$, soit une corrélation faible (voir même très faible, pratiquement nulle).

Exercices de fin de chapitre

1. Voici un graphique qui représente la relation entre la quantité de caféine ingurgitée (exprimée en ml) sur le temps de midi et le taux d'énergie lors de la reprise du travail (pour le déterminer, les participants devaient compléter une échelle allant de 0 = très endormi à 10 = en pleine forme), en début d'après-midi.



1.1. Que pouvez-vous dire de cette relation ?

1.2. Pensez-vous que le coefficient r de Pearson représentera cette série de manière adéquate ?

2. On mesure la corrélation de Pearson pour évaluer la relation entre deux variables. On trouve une valeur très proche de 0. Comment interpréter cette valeur ?

3. Voici les notes obtenues par 10 étudiants (en %) en mathématiques et en économie.

3.1. Représenter graphiquement la relation entre les deux variables. 3.2. Calculez la covariance entre les cours de mathématiques et d'économie. 3.3. Calculez la corrélation entre les cours de mathématiques et d'économies.

Références

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

Leys, C. (n.a.). *Syllabus d'analyse de données, première partie*. Université Libre de Bruxelles

Ritter, C. (n.a.). Communicating statistics by graphs and tables.

Support office. Retiré sur: https://support.office.com/.