Première partie
Concepts généraux

1 Objectifs

L'objectif des statistiques est de synthétiser un grand nombre de données en un ensemble d'indicateurs permettant une analyse objective.

On peut distinguer :

- des indicateurs fournissant les données les plus fréquentes ou attendues telles que la moyenne, la médiane ou le mode ;
- des indicateurs fournissant une indication sur la répartition des données telle que le maximum/minimum, la variance, l'écart-type, des quantiles, les fréquences ou les fréquences cumulées ;
- des représentations graphiques pertinentes qui permettent d'évaluer rapidement la valeur et la répartition des données.

2 Type de données

2.1 Données continues ou quantitatives

Les données continues sont des données numériques. Elles se retrouvent généralement dans un intervalle donné, mais l'on peut donc trouver une infinité de valeurs dans cet intervalle. Un exemple est donné au point 4.1.

2.2 Données discrètes ou qualitatives

Les données discrètes sont des données discontinues et en quantité limitées.

Il peut s'agir de classes de données (une subdivision, un intervalle arbitraire de données continues, un exemple est donnée en 4.2)

Il peut s'agir également de catégories (comme la profession, le sexe, tout regroupement arbitraire de données, ...). Un exemple est donné au point 4.3.

2.3 Populations et échantillons

En statistique, une **population** est un ensemble de données. Ce terme ne désigne donc pas nécessairement des individus. Il s'agit d'une population de données et non de personnes.

Comme, parfois, la quantité est extrêmement importante, afin de diminuer le calcul ou les mesures, on utilisera un sous-ensemble de données moins important. Ce sous-ensemble est appelé **échantillon**. Si l'échantillon est suffisamment grand, il pourra représenter la population.

On utilisera le symbole n comme identifiant du nombre de données (tant pour une population qu'un échantillon).

Deuxième partie

Statistiques à une variable

1 Minimum, maximum et médiane

Le **minimum** (*min*) est la plus petite valeur d'un échantillon (ou d'une population) de données.

Le **maximum** (*max*) est la plus grande valeur d'un échantillon (ou d'une population) de données.

La **valeur médiane statistique** est la valeur pour laquelle la moitié de la population aura une valeur plus faible et l'autre moitié aura une valeur plus élevée. Il s'agit donc du percentile 50 (cfr 2.1).

- Si la population a un nombre impair de données, on utilisera la valeur de la donnée « au milieu »(la valeur de l'élément $\frac{n+1}{2}$ dans une suite de données ordonnée).
- Si la population a un nombre pair de données, on utilisera la moyenne des deux valeurs médianes (correspondant à la moyenne $\frac{n+1}{2}$ et $\frac{n}{2}$ dans une suite de données ordonnée).

On peut aussi définir une **valeur médiane géométrique** (*med*) qui correspond à :

$$med = min + \frac{(max - min)}{2}$$

2 Catégories, Classes, Fréquence et Mode

Les catégories sont des informations qui ne sont pas liées à des nombres continus, mais à des données discrètes.

Afin de synthétiser des valeurs numériques, il est courant de créer des **classes de valeurs**. On peut, par exemple, utiliser des intervalles d'âge (par tranche de 10 ans, par exemple) plutôt que les valeurs numériques de l'âge.

La **fréquence** correspond à l'occurrence d'un échantillon ou d'une population pour une valeur numérique, une classe ou une catégorie. Elle est souvent exprimée sous forme de pourcentage par rapport à la population ou à l'échantillon total.

$$freq(A) = \frac{n_A}{n}$$

La **fréquence cumulée** ne peut s'appliquer qu'à des classes ou des valeurs numériques (et non à des catégories). En effet, un ordre est nécessaire pour effectuer le cumul. Le cumul peut se faire de façon croissante ou décroissante (l'ordre croissant est souvent préféré). Cette fréquence cumulée correspond donc simplement par l'accumulation des fréquences suivant l'ordre qui a été choisi.

Sur base d'un ordre donné, par exemple A, B, C, D , E, F :

$$freq_{cumul}(D) = freq_{cumul}(C) + freq(D)$$

Le **mode** est la valeur la plus représentée au sein d'une population ou d'un échantillon. Ce mode peut être source de confusion quand plusieurs valeurs sont représentées de façon équivalente. Ce mode peut être utilisé pour des valeurs numériques, des classes ou des catégories.

Exemple L'exemple ci-dessous est choisi avec un nombre faible de valeur afin de faciliter la compréhension. Soit un ensemble d'élèves d'une classe dont l'âge est repris dans le tableau ci-dessous. Il s'agit bien de classes de données et non de valeurs numériques puisqu'il y eu un regroupement par an [1].

18	20	17	18	17	18	17	18	19	18	16	19	20	18	19	18	17	19	16	18

TABLE 1 – L'âge des élèves.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
16	16	17	17	17	17	18	18	18	18	18	18	18	18	19	19	19	19	20	20

TABLE 2 – L'âge ordonné et numéroté des élèves.

Le tableau de fréquence et de fréquence cumulée est donc :

Âge	Fréquence	Fréquence cumulée
16 ans	2	2
17 ans	4	6
18 ans	8	14
19 ans	4	18
20 ans	2	20

TABLE 3 – Fréquence et fréquence cumulée de l'âge des élèves.

Le mode de cette statistique est « 18 ans ».

1. L'âge pourrait être exprimé en valeur décimal (par exemple, 16,5 ans pour 16 ans et 6 mois).

2.1 Quantiles : percentile, décile et quartile

Comme indicateurs de dispersions des données, il est courant d'utiliser les **quantiles**. Il s'agit de fractionner la population (ou l'échantillon) en centième (**percentile**), en dixième (**décile**) ou en quart (**quartile**).

Dans l'exemple précèdent,

- le premier décile correspond à la classe « 16 ans » ;
- les deuxième et troisième déciles correspondent à la classe « 17 ans » ;
- les quatrième, cinquième, sixième et septième déciles correspondent à la classe « 18 ans » ;
- les huitième et neuvième déciles correspondent à la classe « 19 ans » ;
- le dixième et dernier décile correspond à la classe « 20 ans » ;

Dans le cas de quartiles,

- le premier quartile correspond au 25 premiers pourcents ;
- le second quartile va de 25 à 50 % ;
- le troisième quartile va de 50 à 75 % ;
- le quatrième quartile va de 75 à 100 % ;

On peut définir les différents points de valeurs suivants sur une liste de n données :

- $Q0$ correspond au minimum du premier quartile, soit le minimum de la population ou de l'échantillon ;
- $Q1$ correspond au maximum du premier quartile ou au minimum du second quartile, soit la valeur de l'élément $\frac{n+3}{4}$;
- $Q2$ correspond au maximum du second quartile ou au minimum du troisième quartile, soit la valeur de l'élément $\frac{n+1}{2}$, soit la valeur de la médiane statistique ;
- $Q3$ correspond au maximum du troisième quartile ou au minimum du quatrième quartile, soit la valeur de l'élément $\frac{3n+1}{4}$;
- $Q4$ correspond au maximum du quatrième quartile, soit le maximum de la population ou de l'échantillon.

En cas de valeur paire au sein du quartile, et donc en absence d'une valeur précise dans l'échantillon/population, on prendra la valeur moyenne entre les deux valeurs dans l'intervalle. Par exemple,

- si les deux valeurs sont 17 et 19, la valeur sera 18 ;
- si les deux valeurs sont 16,5 et 18, on prendra 17,25 ;
- si les deux valeurs sont 17,4 et 18,4 on prendra 17,9 ;

Dans l'exemple précèdent,

- le premier quartile correspond à l'intervalle $[16, 17]$;

- le second quartile correspond à l'intervalle $[17, 18]$;
- le troisième quartile correspond à l'intervalle $[18, 19]$;
- le quatrième quartile correspond à l'intervalle $[19, 20]$;

La fréquence ou la fréquence cumulées de ces quantiles peut être ensuite représentées dans des tableaux ou des diagrammes.

2.2 Représentations graphiques

2.2.1 Diagrammes en bâtonnets

Une représentation en bâtonnet utilise un diagramme où :

- l'axe horizontal représente les classes ou les catégories. Il est important d'appréhender que cet axe n'est pas « numérique ». la distance entre l'ensemble des classes est fixe et ne dépend pas des valeurs des classes.
- l'axe vertical le nombre d'occurrence (ou le pourcentage) de la population. Cet axe est donc bien numérique.

Une représentation en bâtonnets ne peut s'utiliser que sur des données par classes ou par catégories. Les termes de diagrammes en colonnes ou en barres sont aussi utilisés.

Fréquence simple Dans le cas d'une fréquence simple, on reprendra la fréquence de chaque classe ou catégorie.

Si on reprend l'exemple du point 2, on obtiendra le diagramme en bâtonnets suivant.

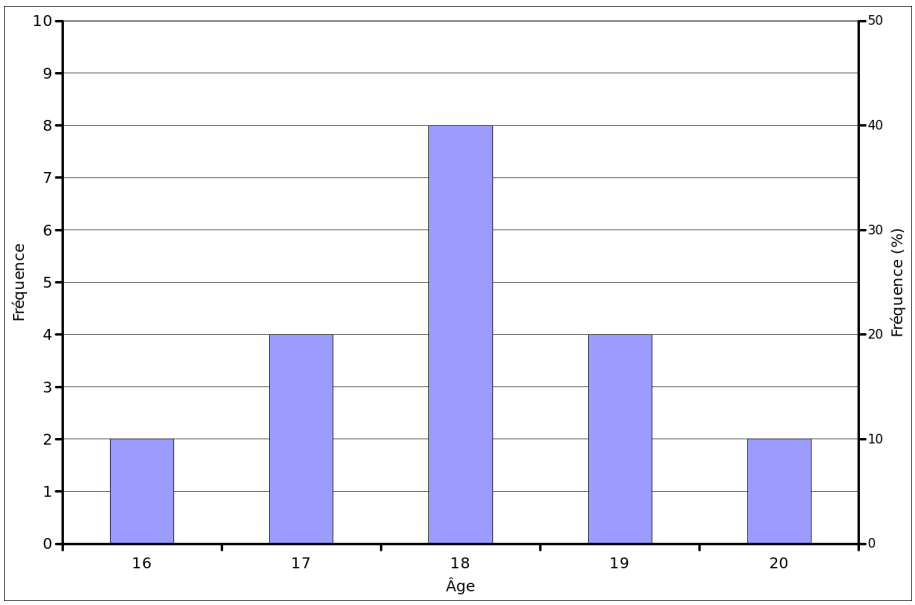

FIGURE 1 – Représentation en bâtonnet de la fréquence d'âge des élèves.

Fréquence cumulée Dans le cas d'une fréquence cumulée, on reprendra la fréquence cumulée de chaque classe. On ne peut faire de diagramme cumulé sur des catégories étant donné l'absence d'ordre des catégories.

Si on reprend l'exemple du point 2, on obtiendra le diagramme en bâtonnets suivant.

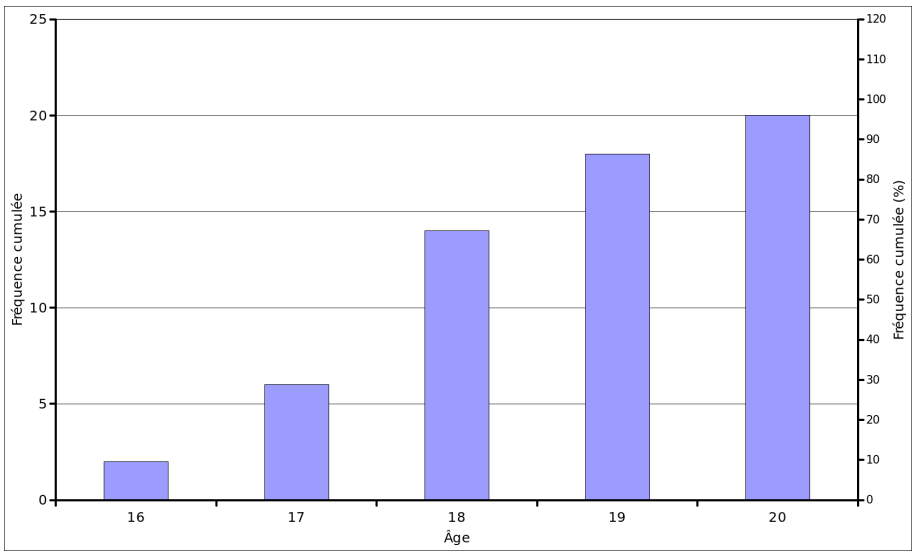

FIGURE 2 – Représentation en bâtonnet de la fréquence cumulées d'âge des élèves.

2.2.2 Diagrammes en quartier

Une représentation en quartiers utilise un diagramme où un cercle est divisé en un certain nombre de parts. L'amplitude de l'angle de chaque part est proportionnel à la fréquence de la catégorie ou de la classe. Chaque part peut contenir la fréquence nominale ou en pourcents.

Une représentation en quartiers ne peut s'utiliser que pour représenter la fréquence des données par classes ou par catégories. Cette représentation ne peut pas s'utiliser pour une fréquence cumulée.

Si on reprend l'exemple du point 2, on obtiendra le diagramme en quartiers suivant. Les diagrammes en quartiers sont parfois appelés « en camembert ».

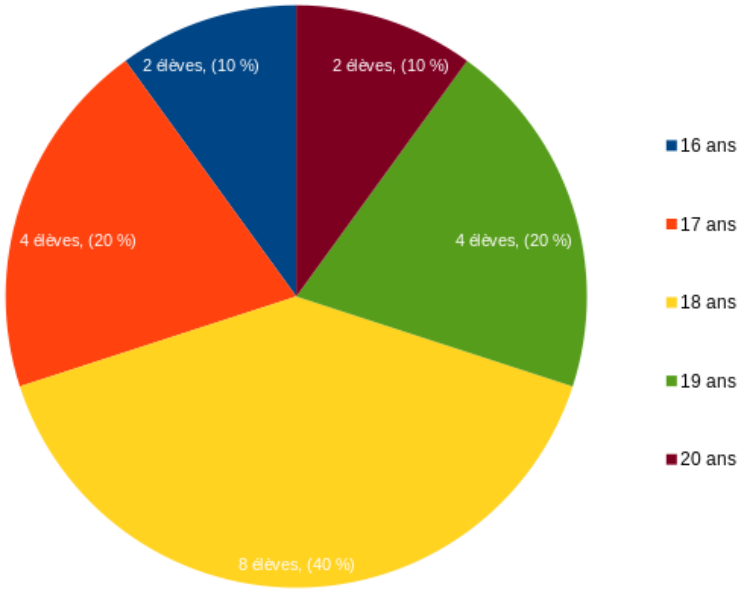

FIGURE 3 – Représentation en quartier de la fréquence d'âge des élèves.

2.2.3 « Boîtes à moustaches »

La représentation en « boîte à moustaches » [2] permettent de représenter à la fois la valeur attendue et la dispersion des données. Ce type de diagramme permet de comparer rapidement différentes populations ou échantillons, mais ne peut s'appliquer qu'à des classes ou des catégories.

Le schéma suivant montre les différentes informations contenues dans ces diagrammes en boîtes.

De bas en haut, on retrouvera donc

- le minimum ;
- le deuxième quartile ;
- la médiane statistique ;
- le troisième quartile ;
- le maximum.

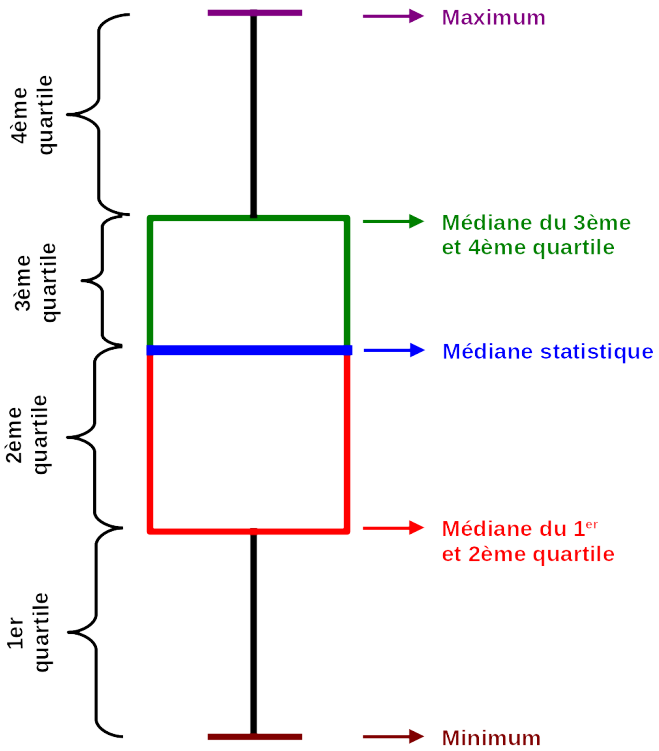

FIGURE 4 – Signification de la « boîte à moustache ».

2. Les termes de diagrammes en boîtes ou boxplot sont aussi utilisés.

3 Moyenne, variance et écart-type

Le calcul de la moyenne et de l'écart-type ne peut se faire que sur des données continues.

3.1 Calcul de la moyenne

Le calcul de la moyenne ($\bar{x}$) d'une population ou d'un échantillon est simplement la somme des valeurs divisé par le nombre de valeur.

$$\bar{x} = \frac{\sum x}{n}$$

Si n_x est la fréquence d'un valeur,

$$\bar{x} = \frac{\sum n_x.x}{n}$$

3.2 Calcul de la variance et de l'écart-type

La variance (var) est la somme des différences à la moyenne au carré divisé par le nombre de valeur.

$$var = \frac{\sum(\bar{x} - x)^2}{n}$$

Si n_x est la fréquence d'un valeur,

$$var = \frac{\sum n_x.(\bar{x} - x)^2}{n}$$

L'écart-type (σ) est simplement la racine de la variance.

$$\sigma = \sqrt{var}$$

3.3 Interprétation de la moyenne et de l'écart-type

Lorsque les données se répartissent uniformément autour d'une seule valeur (distribution normale, cfr. cours de Probabilités en 6ème année), la moyenne ($\bar{x}$) correspond à la valeur attendue, c-à-d à la valeur la plus probable. Dans ce cadre, les valeurs de moyenne, de médiane statistique (P50) et de médiane géométrique tendent à s'égaliser.

Dans le cas d'une distribution normale [3], l'écart-type (σ) peut être être assimilé à des percentiles.

En effet, les intervalles suivants correspondent à des percentiles précis :

- ➞ $[\bar{x} - \sigma; \bar{x} + \sigma]$, reprend 68,27% de la population (ou de l'échantillon)
- ➞ $[\bar{x} - 2\sigma; \bar{x} + 2\sigma]$, reprend 95,45% de la population (ou de l'échantillon)
- ➞ $[\bar{x} - 3\sigma; \bar{x} + 3\sigma]$, reprend 99,73% de la population (ou de l'échantillon)

3. On peut définir une distribution normale comme un ensemble de données symétrique sur la valeur centrale et ou cette valeur centrale converge vers la moyenne, la médiane statistique, la médiane géométrique et le mode

3.4 Représentation graphique

Les moyennes et écart-types peuvent être représenté par des **barres d'erreur** qui sont des formes simplifiées de « boîtes à moustaches ». La barre centrale correspond à la moyenne($\bar{x}$) et les deux extrémités à $\bar{x}-\sigma$ et $\bar{x}+\sigma$. Parfois, $2\,\sigma$ (95% des données) ou $3\,\sigma$ peuvent être utilisé, mais ils doivent alors être spécifiés.

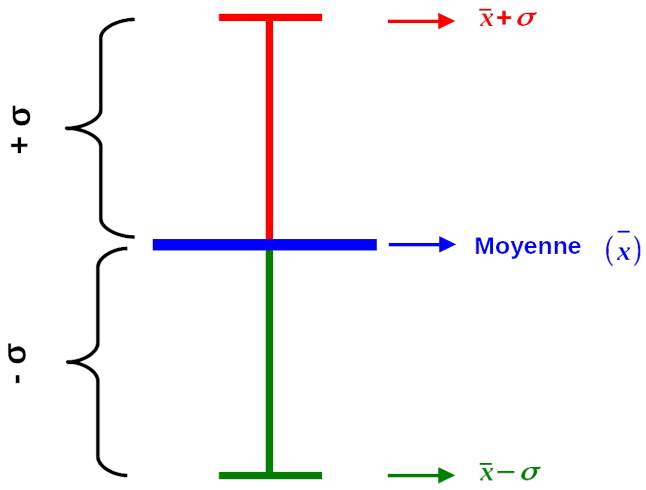

FIGURE 5 – Signification de la barre d'erreur.

4 Exemples concrets

4.1 Statistiques de valeurs numériques

Soit 3 classes dont les cotations en pourcent sont reprises dans le tableau ci-dessous (par facilité, ces cotations ont déjà été ordonnées).

Numéro	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Classe A	15	31	43	44	46	47	52	55	56	57	61	64	67	68	73	75	79	84	87
Classe B	14	29	32	42	43	47	54	56	68	72	78	79	81	88	90	94			
Classe C	38	39	41	43	45	45	46	49	50	51	52	53	63	65	71	75	83		

TABLE 4 – Les cotations ordonnées et numérotées par classe.

Sur base de ces données, on peut calculer la moyenne et l'écart-type des 3 classes d'élèves.

	Classe A		Classe B		Classe C	
	x	$(\bar{x}-x)^2$	x	$(\bar{x}-x)^2$	x	$(\bar{x}-x)^2$
	15	1858,06	14	2156,44	38	239,34
	31	734,70	29	988,32	39	209,40
	43	228,17	32	808,69	41	155,52
	44	198,96	42	339,94	43	109,63
	46	146,54	43	304,07	45	71,75
	47	123,33	47	180,57	45	71,75
	52	37,27	54	41,44	46	55,81
	55	9,64	56	19,69	49	19,99
	56	4,43	68	57,19	50	12,04
	57	1,22	72	133,69	51	6,10
	61	8,38	78	308,44	52	2,16
	64	34,75	79	344,57	53	0,22
	67	79,12	81	422,82	63	90,81
	68	97,91	88	759,69	65	132,93
	73	221,85	90	873,94	71	307,28
	75	285,43	94	1126,44	75	463,52
	79	436,59			83	871,99
	84	670,54				
	87	834,91				
Somme	1104	6011,79	967	8865,94	909	2820,24
n	19		16		17	
/n	$\bar{x} = 58{,}11$	$var = 316{,}41$	$\bar{x} = 60{,}44$	$var = 554{,}12$	$\bar{x} = 53{,}47$	$var = 165{,}90$
		$\sigma = 17{,}79$		$\sigma = 23{,}54$		$\sigma = 12{,}88$

TABLE 5 – Calcul de la moyenne et de l'écart-type des 3 classes.

Ensuite, on peut représenter les barres d'erreurs de chaque classe.

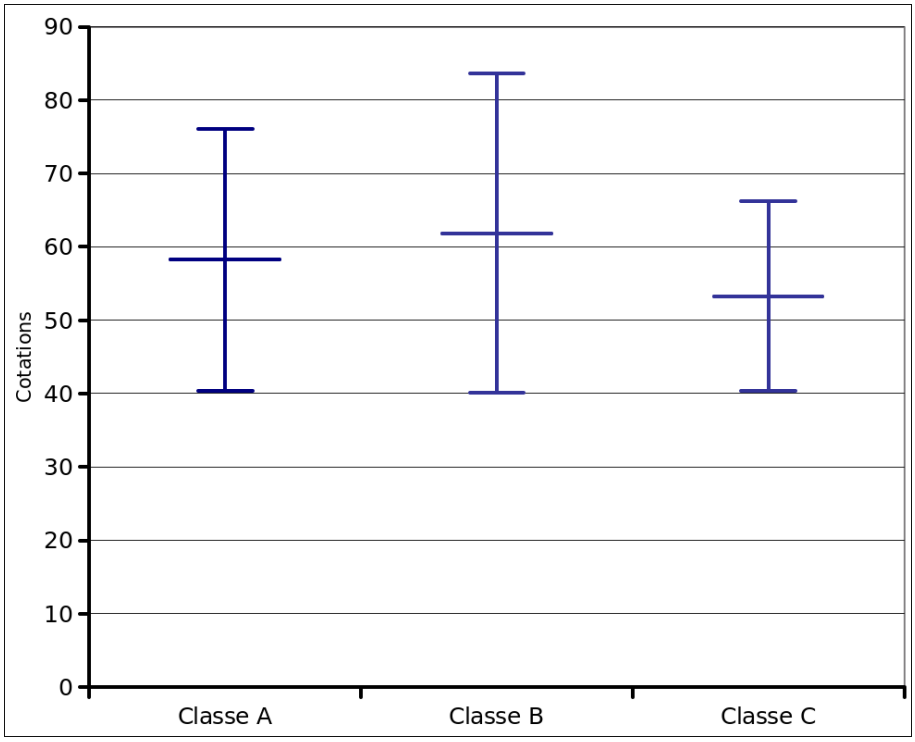

FIGURE 6 – Barres d'erreur des 3 classes.

4.2 Statistiques de classes de valeurs numériques

Sur base des données reprises en 4.1, on peut regrouper ces données en classes de points par pas de 10 points. Le tableau ci-dessous reprend la fréquence de chacune de ces classes de points pour chaque classe d'élèves.

	Classe A	Classe B	Classe C
0-9	0	0	0
10-19	1	1	0
20-29	0	1	0
30-39	1	1	2
40-49	4	3	6
50-59	4	2	4
60-69	4	1	2
70-79	3	3	2
80-89	2	2	1
90-100	0	2	0

TABLE 6 – Fréquences des classes de points par classe d'élèves.

Ces données peuvent être représentées par un diagramme en bâtonnets des classes de points.

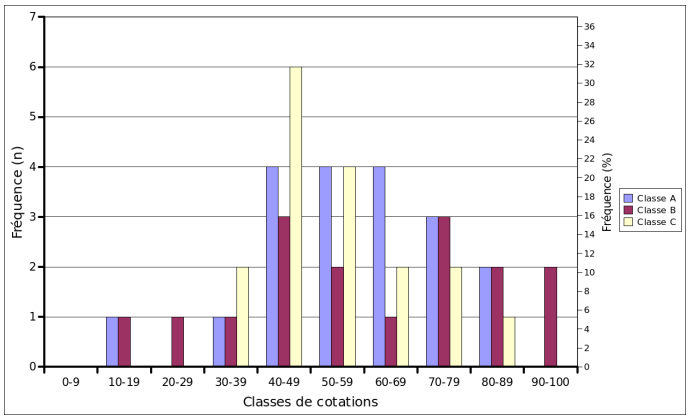

FIGURE 7 – Représentation des fréquences des classes de points par classe d'élèves.

Le tableau ci-après reprend les fréquences cumulées des classes de points pour chaque classe d'élèves.

	Classe A	Classe B	Classe C
0-9	0	0	0
10-19	1	1	0
20-29	1	2	0
30-39	2	3	2
40-49	6	6	8
50-59	10	8	12
60-69	14	9	14
70-79	17	12	16
80-89	19	14	17
90-100	19	16	17

TABLE 7 – Fréquences cumulées des classes de points par classe d'élèves.

Ces données de fréquences cumulées peuvent être représentées par un diagramme en bâtonnets des classes de points.

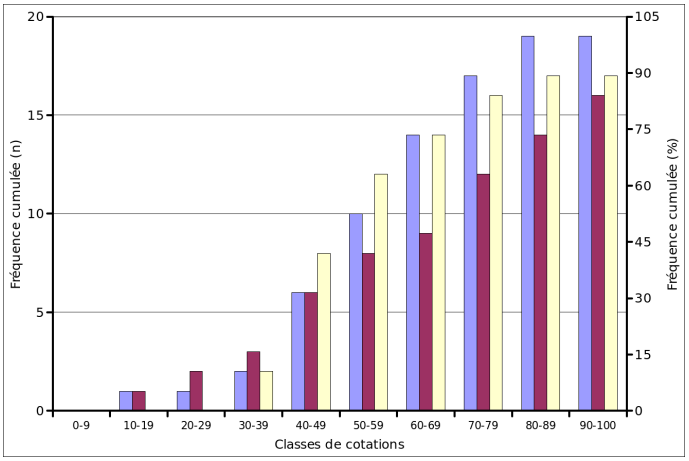

FIGURE 8 – Représentation cumulée des fréquences des classes de points par classe d'élèves.

En calculant les quartiles de ces classes de points, on peut représenter ces données sous forme de diagramme en « boîtes à moustaches ».

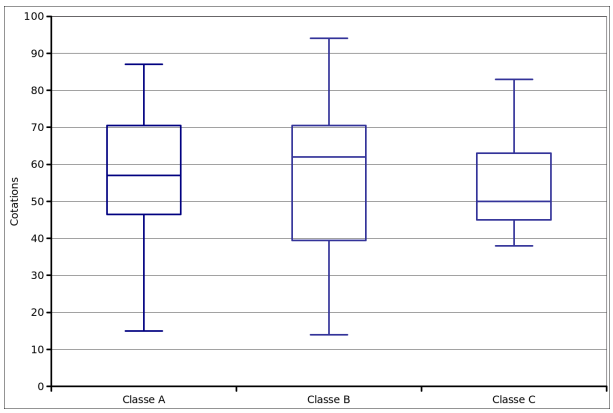

FIGURE 9 – Représentation en « boîtes à moustaches »des classes de points par classe d'élèves.

4.3 Statistiques de catégories

Le tableau ci-après reprend la fréquence des communes des élèves pour chaque classe.

Commune	Classe A	Classe B	Classe C
Saint-Josse	7	4	5
Schaerbeeck	6	5	6
Bruxelles-ville	3	3	3
Autres	3	4	3

TABLE 8 – Fréquences des communes des élèves par classe.

Les diagrammes suivants reprennent la fréquence des communes des élèves pour chaque classe.

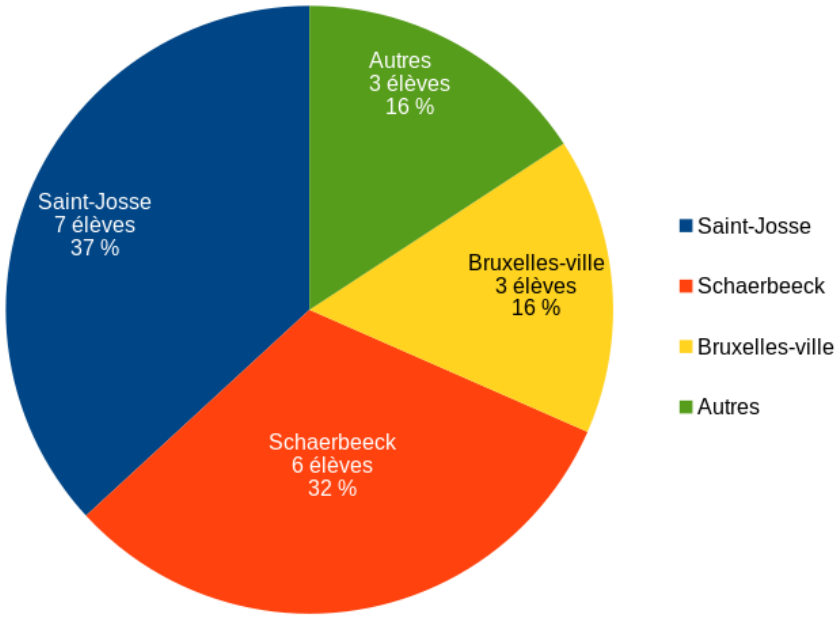

FIGURE 10 – Répartition en quartier des communes des élèves de la classe A.

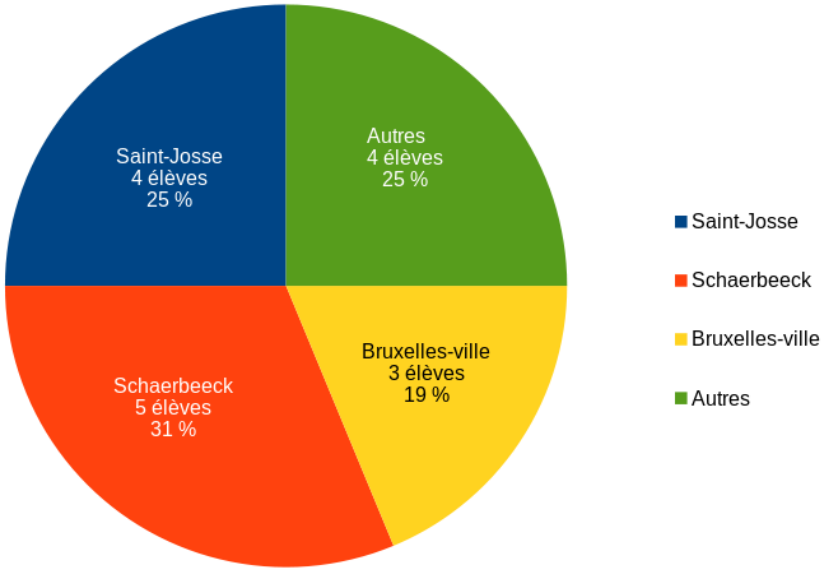

FIGURE 11 – Répartition en quartier des communes des élèves de la classe B.

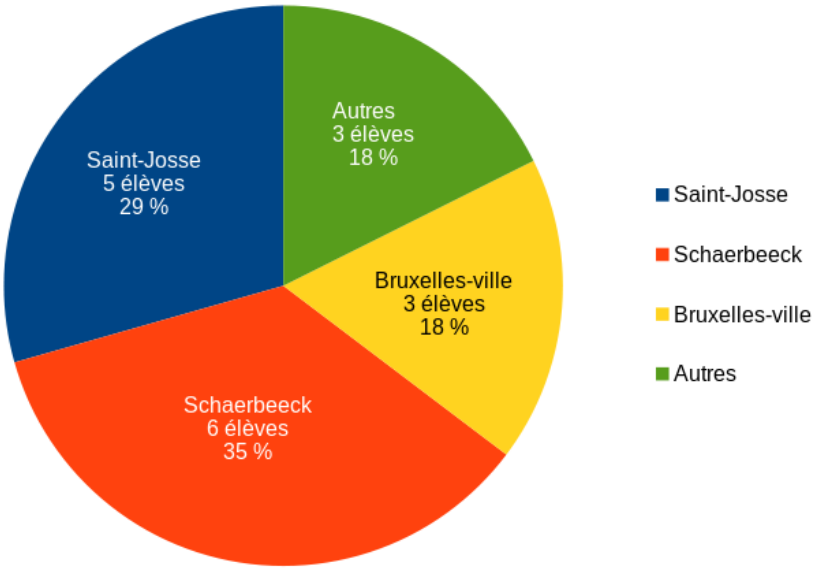

FIGURE 12 – Répartition en quartier des communes des élèves de la classe C.

Troisième partie

Statistiques à deux variables

1 Objectif des statistiques à 2 variables

L'objectif principal des statistiques à deux variables est de vérifier s'il existe un lien de corrélation entre deux variables. Si ce lien existe, on peut tenter de quantifier la relation entre les deux variables, par exemple, avec une droite de régression linéaire.

2 Représentation graphique

Il est d'usage de représenter les données de deux variables numériques par des diagrammes « en nuage de points ».

Une des deux données est représentée sur l'abscisse (axe des X), l'autre sur l'ordonnée (axe des Y).

3 Régression linéaire

Une régression ou ajustement est une méthode permettant de synthétiser un ensemble de données par une fonction. Dans le cas d'une régression linéaire, il s'agit d'une fonction linéaire, c'est à dire une fonction de la forme :

$$y = a.x + b$$

3.1 Méthode de Mayer

La méthode de Mayer consiste :

- à créer deux sous-échantillons, l'un sur les valeurs inférieures et l'autre sur les valeurs supérieures de la population/échantillon initial ;
- à calculer la moyenne des deux sous-échantillons pour chacune des deux variables ;
- les deux moyennes ($\bar{x}_1$ et $\bar{y}_1$) du premier sous-échantillon détermine le point p_1 ;
- les deux moyennes ($\bar{x}_2$ et $\bar{y}_2$) du second sous-échantillon détermine le point p_2 ;
- en reliant les deux points p_1 et p_2 ont détermine une droite de régression entre les deux variables.

Cette méthode ne garantit pas que la droite de régression obtenue passe par les moyennes de la population/échantillon initial (même si elle s'en approche souvent fort).

Exemple Nous allons prendre comme exemple les données de température moyenne mondiale de l'air fournies par la NASA (https ://data.giss.nasa.gov/gistemp/graphs/) entre 1993 et 2022.

Sur l'axe des abscisse (x), nous avons l'année.

Sur l'axe des ordonnées (y), nous avons l'écart de température par rapport à une période de référence [4].

Ci-dessous le tableau de données.

	Année	***Écart inf.(°C)***	***Année***	***Écart sup.(°C)***
	1993	0,23	2008	0,54
	1994	0,32	2009	0,66
	1995	0,45	2010	0,72
	1996	0,33	2011	0,61
	1997	0,46	2012	0,65
	1998	0,61	2013	0,68
	1999	0,38	2014	0,75
	2000	0,39	2015	0,9
	2001	0,54	2016	1,02
	2002	0,63	2017	0,92
	2003	0,62	2018	0,85
	2004	0,53	2019	0,98
	2005	0,68	2020	1,02
	2006	0,64	2021	0,85
	2007	0,66	2022	0,89
Moyenne	**2000**	**0,498**	**2015**	**0,803**

TABLE 9 – Écart de température mondiale de l'air par rapport à l'année 1951-1980.

Les deux points reliés par une droite sont donc (2000 ; 0,498) et (2015 ; 0,803). La moyenne globale des années est **2007,5**. La moyenne globale des écarts de température est **0,65 °C**.

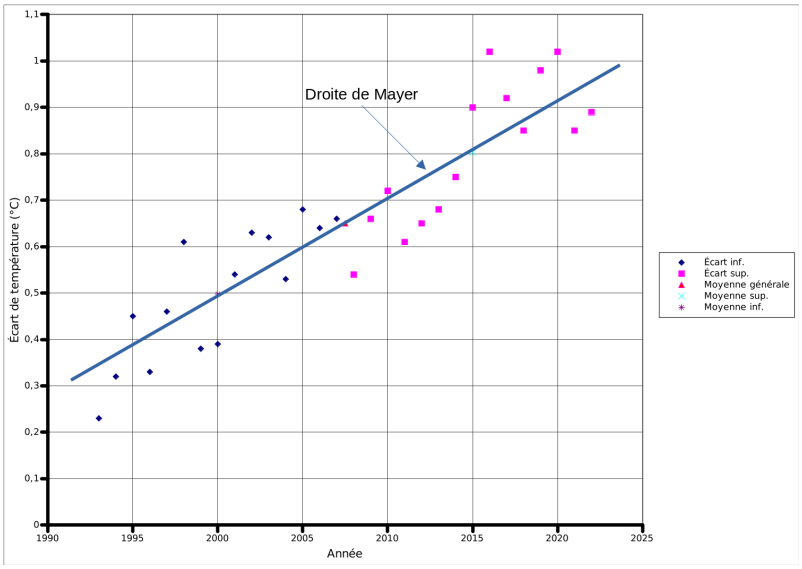

FIGURE 13 – Droite de régression par la méthode de Mayer.

4. 1951-1980

3.2 Méthode des moindres carrés

La méthode des moindres carrés consiste à déterminer une droite de régression :

- qui passe par le couple de moyenne $(\bar{x},\, \bar{y})$;
- qui minimise le carré de la différence entre les valeurs de la droite et les valeurs de l'échantillon/population.

C'est la méthode la plus employée et reconnue comme la plus valide si la relation est linéaire.

3.2.1 Covariance

La covariance de deux variables est un nombre permettant de quantifier les écarts conjoints par rapport à leur moyenne. Elle correspond à la somme de la différence des produits des valeurs et des produits des moyennes divisé par la taille de l'échantillon/population.

$$Cov(x, y) = \frac{\sum(x.y - \bar{x}.\bar{y})}{n}$$

3.2.2 Méthode de calcul de la droite de régression

On peut montrer que la pente de la droite de régression aura alors comme valeur la division de la covariance des deux valeurs par la variance de la variable x.

$$a = \frac{Cov(x, y)}{V(x)}$$

Sachant que par définition,

$$\bar{y} = a.\bar{x} + b$$

on peut facilement démontrer que

$$b = \bar{y} - a.\bar{x}$$

3.2.3 Coefficient de corrélation

Le coefficient de corrélation (r) est une valeur numérique entre -1 et 1. Il détermine la corrélation entre les variables x et y. Il correspond au rapport de la covariance sur le produit des écart-types.

$$r = \frac{Cov(x, y)}{\sigma_x.\sigma_y}$$

On peut distinguer plusieurs situations :

- si r est entre 1 et 0.5 : la corrélation est forte et positive entre les deux variables ;
- si r est entre $0,5$ et 0 : la corrélation est faible et positive entre les deux variables ;
- si r est entre 0 et $-0,5$: la corrélation est faible et négative entre les deux variables ;
- si r est entre $0,5$ et -1 : la corrélation est forte et négative entre les deux variables.

La représentation graphique en « nuages de points » permet de voir rapidement (mais sans quantifier) le type de relation entre les deux variables.

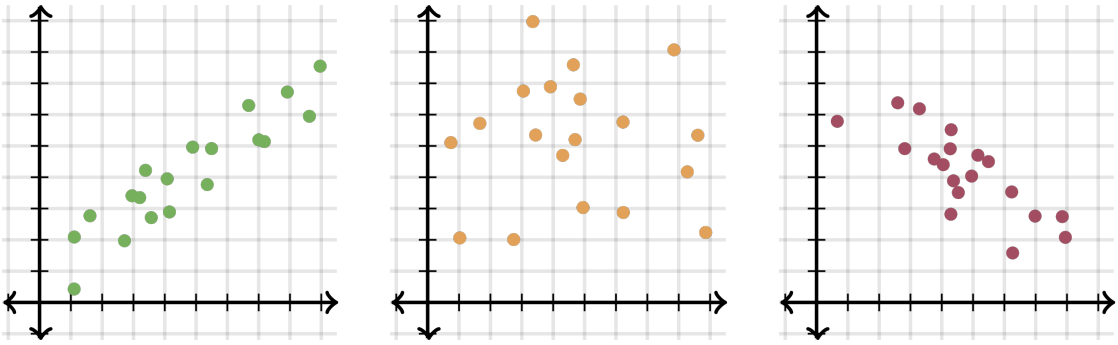

FIGURE 14 – Représentation graphique typique d'une corrélation respectivement forte et positive, faible, forte et négative.

Corrélation et causalité Il est important de bien comprendre que la corrélation n'est pas la causalité. La corrélation est purement mathématique.

La causalité est due à un processus qui est à l'œuvre. Un des facteurs est identifié comme cause (quantifié par une variable) aura comme conséquence la modification (augmentation ou diminution) de l'autre variable.

En cas de corrélation forte, il existe quatre situations :

- x est la cause de la modification de y (conséquence),
- ou inversement y est la cause de la modification de x.
- Il existe un autre facteur qui lie les deux variables. Par exemple, la vente de lunettes solaires est corrélé de façon forte et directe à la vente de pommade de protection solaire. La cause de ces 2 variables est liée à la météo (l'insolation solaire).
- Le lien est juste fortuit, est juste une coïncidence. Par exemple, le cycle menstruel moyen chez la femme est de 28 jours, la période orbitale de la Lune autour de la Terre est 27,3 jours. Vu la variabilité de la durée des cycles menstruels, cela produira un coefficient de corrélation proche de 1. Mais, il n'existe pas, a priori, de lien causal entre les deux [5].

Pour résumer, si la corrélation est du domaine des statistiques, la causalité est du domaine de la science du domaine considéré (biologie, physique, sociologie, économie, ...).

3.2.4 Coefficient de détermination

Le coefficient de détermination est simplement le coefficient de corrélation au carré (r^2). Il varie donc entre 0 et 1.

Lorsqu'il est proche de 0, le pouvoir prédictif de la droite de régression est faible et lorsqu'il est proche de 1, le pouvoir prédictif de la droite de régression est fort.

5. Les cycles menstruels des mammifères sont très variables. Par exemple, chez le chimpanzé, il est de 36 jours.

Exemple Nous reprenons les données de température moyenne mondiale de l'air fournies par la NASA. L'axe des x est toujours l'année. L'axe des y est l'écart de température par rapport aux années de différences 1951-1980.

Nous calculons les moyennes, les variances, les écart-type et la covariance directement dans le tableau.

	Année	***Écart (°C)***	$(\mathbf{x}-\bar{\mathbf{x}})^2$	$(\mathbf{y}-\bar{\mathbf{y}})^2$	$(\mathbf{x}.\mathbf{y})-(\bar{\mathbf{x}}.\bar{\mathbf{y}})$
	1993	0,23	210,25	0,177	-847,154
	1994	0,32	182,25	0,109	-667,464
	1995	0,45	156,25	0,040	-407,794
	1996	0,33	132,25	0,103	-646,864
	1997	0,46	110,25	0,036	-386,924
	1998	0,61	90,25	0,002	-86,764
	1999	0,38	72,25	0,073	-545,924
	2000	0,39	56,25	0,068	-525,544
	2001	0,54	42,25	0,012	-225,004
	2002	0,63	30,25	0,000	-44,284
	2003	0,62	20,25	0,001	-63,684
	2004	0,53	12,25	0,014	-243,424
	2005	0,68	6,25	0,001	57,856
	2006	0,64	2,25	0,000	-21,704
	2007	0,66	0,25	0,000	19,076
	2008	0,54	0,25	0,012	-221,224
	2009	0,66	2,25	0,000	20,396
	2010	0,72	6,25	0,005	141,656
	2011	0,61	12,25	0,002	-78,834
	2012	0,65	20,25	0,000	2,256
	2013	0,68	30,25	0,001	63,296
	2014	0,75	42,25	0,010	204,956
	2015	0,9	56,25	0,062	507,956
	2016	1,02	72,25	0,137	750,776
	2017	0,92	90,25	0,073	550,096
	2018	0,85	110,25	0,040	409,756
	2019	0,98	132,25	0,109	673,076
	2020	1,02	156,25	0,137	754,856
	2021	0,85	182,25	0,040	412,306
n=30	2022	0,89	210,25	0,057	494,036
Somme	**60225**	$\mathbf{19,51}$	$\mathbf{2247,5}$	$\mathbf{1,320}$	$\mathbf{49,755}$
/n	$\bar{\mathbf{x}}=\mathbf{2007,5}$	$\bar{\mathbf{y}}=\mathbf{0,650}$	$\mathbf{V(x)=74,917}$	$\mathbf{V(y)=0,044}$	$\mathbf{Cov(x,y)=1,658}$
$\sqrt{}$	-	-	$\sigma_{\mathbf{x}}=\mathbf{8,655}$	$\sigma_{\mathbf{y}}=\mathbf{0,210}$	-

TABLE 10 – Écart de température mondiale de l'air par rapport à l'année 1951-1980.

La pente a peut être calculée.

$$a=\frac{Cov(x,y)}{V(x)}$$
$$a=\frac{1,658}{74,917}$$
$$a=0,022$$

L'ordonnée à l'origine b se déduit alors.

$$b = \bar{y} - a.\bar{x}$$

$$b = 0,650 - 0,022.2007,5$$

$$b = -43,792$$

Le coefficient de corrélation peut aussi être calculé.

$$r = \frac{Cov(x,y)}{\sigma_x.\sigma_y}$$

$$r = \frac{1,658}{8,655.0,210}$$

$$r = 0,913$$

Le coefficient de détermination vaut donc :

$$r^2 = 0,834$$

La figure suivante représente la droite de régression des écarts de température en fonction des années.

On peut constater que tant le coefficient de corrélation que le coefficient de détermination sont proche de 1 et donc que la corrélation est élevée.

Mais, il est bien clair que ce sont pas les années qui sont la cause des écarts de température, ni les écarts de température qui sont la cause des années. Dans ce cas-ci, c'est le taux de dioxyde (CO_2) qui augmente d'année en année qui est la cause de l'augmentation de température.

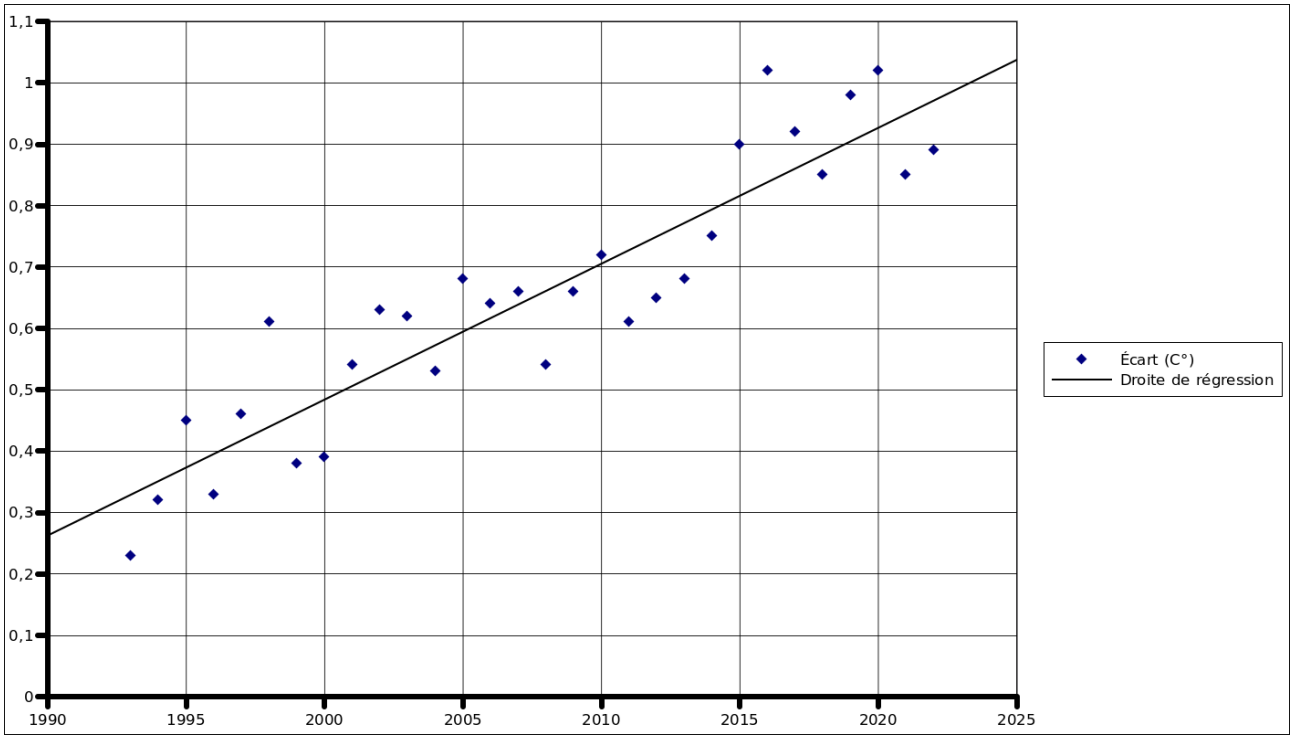

FIGURE 15 – Droite de régression par la méthode des moindres carrés.

Table des matières

I Concepts généraux 1

1 Objectifs **1**

2 Type de données **1**
2.1 Données continues ou quantitatives 1
2.2 Données discrètes ou qualitatives 1
2.3 Populations et échantillons 1

II Statistiques à une variable 2

1 Minimum, maximum et médiane **2**

2 Catégories, Classes, Fréquence et Mode **2**
2.1 Quantiles : percentile, décile et quartile 4
2.2 Représentations graphiques 6
2.2.1 Diagrammes en bâtonnets 6
2.2.2 Diagrammes en quartier 7
2.2.3 « Boîtes à moustaches » 8

3 Moyenne, variance et écart-type **9**
3.1 Calcul de la moyenne 9
3.2 Calcul de la variance et de l'écart-type 9
3.3 Interprétation de la moyenne et de l'écart-type 9
3.4 Représentation graphique 10

4 Exemples concrets **11**
4.1 Statistiques de valeurs numériques 11
4.2 Statistiques de classes de valeurs numériques 13
4.3 Statistiques de catégories 15

III Statistiques à deux variables 17

1 Objectif des statistiques à 2 variables **17**

2 Représentation graphique **17**

3 Régression linéaire **17**
3.1 Méthode de Mayer 17
3.2 Méthode des moindres carrés 19
3.2.1 Covariance 19
3.2.2 Méthode de calcul de la droite de régression 19
3.2.3 Coefficient de corrélation 19
3.2.4 Coefficient de détermination 20