Cours de Statistique

Christine Tuleau-Malot[1]

December 13, 2011

[1]Université de Nice Sophia-Antipolis, France

Contents

Chapter 1

Introduction aux probabilités

Introduction
L'objectif de ce chapitre n'est pas de donner un cours de probabilité, mais seulement de définir les notions principales de probabilités qui seront nécessaires au cours de statistique. En effet, si les probabilités sont totalement absentes des statistiques descriptives, elles sont essentielles aux statistiques inférentielles. D'où ce premier chapitre.

1.1 Variable aléatoire

Mis sous forme numérique, le résultat d'une épreuve aléatoire, symbolisé ou non par un nombre, se prêtera ensuite aux calculs, comme celui de la moyenne associée aux différents résultats possibles. C'est la raison pour laquelle on souhaite, dans la majorité des cas, traduire l'événement réalisé par une valeur numérique.
Par exemple, dans le cas du lancer d'une pièce de monnaie, on peut coder par 1 le côté pile et par 0 le côté face. En ce qui concerne le lancer de dé, il existe un codage naturel puisque le résultat a ici un caractère numérique ($\{1, 2, 3, 4, 5, 6\}$). Cependant, on peut envisager d'autres codages comme 0 si le résultat est pair et 1 s'il est impair.
La valeur numérique associée à un résultat est arbitraire et correspond à un codage des événements qui va se faire au moyen d'une application, usuellement notée X, qui va associer un nombre à chacun des événements élémentaires, soit :

$$X \ : \ \Omega \ \longrightarrow \ \mathbb{R}$$

où Ω est l'univers, autrement dit l'ensemble des résultats possibles.

Le résultat ω de l'expérience ayant un caractère aléatoire (puisque l'on ne connaît pas à l'avance le résultat qui va apparaître), il en va de même pour la valeur numérique associée $X(\omega)$. Ainsi, il est intéressant de calculer la probabilité que X prenne une certaine valeur ou appartienne à un certain intervalle.
Pour pouvoir définir cette probabilité sur l'ensemble $X(\Omega)$, il faut pouvoir revenir en arrière sur l'espace Ω puisque une probabilité se définit sur l'espace $(\Omega, \mathcal{A})$, où $\mathcal{A}$ est la tribu associée à l'espace Ω. Cette notion théorique ne sera pas définie plus en avant ici, car dans la pratique, elle n'est pas nécessaire.
Donc, on va imposer une condition à l'application X qui sera alors appelée variable aléatoire.

$\Rightarrow$ ici, variable = fonction

Plus de détails seront donnés dans le cours de probabilité qui sera dispensé au second semestre. A présent, nous allons entrer dans du concret. Pour ce faire, nous allons distinguer tout au long de ce cours deux cas :

1er cas : $X(\Omega)$ est dénombrable, à savoir :

- $X(\Omega)$ est fini
- ou $X(\Omega) \subset \mathbb{N}$
- ou $X(\Omega) \subset \mathbb{Z}$
- ou $X(\Omega)$ est en bijection avec $\mathbb{N}$ (cela signifie qu'il existe une application p de $X(\Omega)$ dans $\mathbb{N}$ telle que pour tout élément y de $\mathbb{N}$, il existe un et un seul élément x de $X(\Omega)$ tel que $p(x) = y$)

Dans chacun de ces cas, on parle de **variable aléatoire discrète**.

2ème cas : $X(\Omega) \subset \mathbb{R}$
Dans ce cas, on parle alors de **variable aléatoire continue**.

1.2 Variable aléatoire discrète

1.2.1 Définition

Définition 1.
On appelle ***variable aléatoire discrète*** *sur* $(\Omega, \mathcal{A})$*, une application* $X : \Omega \to \mathbb{R}$ *telle que :*

- $X(\Omega)$ *est dénombrable*
- $\forall\, x \in \mathbb{R},\ X^{-1}(x)$ *est un événement, autrement dit* $X^{-1}(x) \in \mathcal{P}(\Omega)$.

Exemple 1.
Soit l'expérience ainsi définie : on lance un dé et on code l'expérience de la manière suivante

- *si on obtient un numéro impair, alors* X *prend la valeur 1*
- *si on obtient un numéro pair, alors* X *prend la valeur 0*

Ainsi $X(\Omega) = \{0; 1\}$, $X^{-1}(1) = \{1; 3; 5\}$ *et* $X^{-1}(0) = \{2; 4; 6\}$.

1.2.2 Loi de probabilité

Pour une définition plus généraliste, le lecteur se référera à un cours de probabilités à proprement parlé. Cependant, voici ce qu'il est bon de savoir pour la suite de ce cours.
L'ensemble $X(\Omega)$ étant dénombrable, il est possible de représenter ses éléments par l'ensemble des $x_i,\ i \in \mathcal{N}$.
On définit alors la **loi de probabilité** P_X de X par les probabilités individuelles :

$$p_i = P_X(X = x_i) = P(X^{-1}(x_i)), \quad \text{pour } i \in \mathbb{N}$$

Remarque 1.
Lorsque $X(\Omega)$ ne comprend qu'un petit nombre de valeurs, cette loi de probabilité encore appelée distribution, est en général représentée sous forme de tableau.

Exemple 2.
Soit l'expérience ainsi définie : on lance un dé équilibré et on définit X comme étant le numéro de la face visible sur le dessus.
Alors, $X(\Omega) = \{1; 2; 3; 4; 5; 6\}$ et $p_1 = P(X = 1) = P(\text{la face 1 est la face visible}) = \frac{1}{6}$.

Remarque 2.
La précision que le dé est équilibré est très importante car cela permet d'en déduire que chacune des faces à la même chance d'apparaître. Or, il est à noter que le calcul classique d'une probabilité est le suivant : soit A un événement, $P(A) = \frac{\text{cardinal de } A}{\text{cardinal de l'univers}}$ avec
cardinal de A = le nombre d'éléments élémentaires constituant A.

Exemple 3.
Soit l'expérience ainsi définie : on lance un dé équilibré et on définit X par :

- *$X = 1$ si le dé fait apparaître 5 ou 6*
- *$X = 0$ sinon*

Alors, $X(\Omega) = \{0; 1\}$, $p_0 = P(X = 0) = P(\text{le dé fait apparaître 1,2,3 ou 4}) = \frac{4}{6} = \frac{2}{3}$ et $p_1 = P(X = 1) = \frac{1}{3}$.

Proposition 1.
Une loi de probabilité vérifie :

- $\forall\ i \in \mathbb{N},\ p_i \in [0, 1]$
- $\sum_{i \in \mathbb{N}} p_i = 1$

1.2.3 Fonction de répartition

Définition 2.
*On appelle **fonction de répartition** de la variable aléatoire X, la fonction F_X définie par :*

$$\forall\ x \in \mathbb{R},\ F(x) = P(X \leq x) = \sum_{i\ \setminus\ x_i \leq x} p_i$$

Proposition 2.
Une fonction de répartition F_X satisfait les points suivants :

- *la fonction est croissante et en escalier*
- *la fonction est continue à droite (mathématiquement cela signifie qu'en un point de continuité à droite x, $\lim\limits_{y \to x,\ y>x} F_X(y) = F_X(x)$ et concrètement, cela signifie que lorsque l'on trace la fonction F_X, dans un voisinage à droite du point x, on ne lève pas le crayon pour arriver au point de coordonnées $(x, F_X(x))$.)*

- $\forall\ x \in \mathbb{R},\ 0 \leq F(x) \leq 1$

- $\lim\limits_{x \to -\infty} F_X(x) = 0$ *et* $\lim\limits_{x \to +\infty} F_X(x) = 1$

Remarque 3.
Toute fonction f qui vérifie tous les points précédents est une fonction de répartition, et de ce fait, il existe une variable aléatoire X telle que $F_X = f$.
A partir d'une fonction de répartition, on peut retrouver la loi de probabilité de la variable. En effet, si $x_1 < x_2 < \ldots < x_n < \ldots$, on a :

$$\forall\ i \geq 2,\ P(X = x_i) = P(X \leq x_i) - P(X \leq x_{i-1}) = F_X(x_i) - F_X(x_{i-1})$$

et $P(X = x_1) = P(X \leq x_1) = F_X(x_1)$.

Remarque 4.
Attention, dans la définition que je vous ai présentée, on a $F_X(t) = P(X \leq t)$ et donc l'inégalité est large. Parfois, on rencontre non pas cette définition, mais la définition suivante : $F_X(t) = P(X < t)$. Il faut donc faire particulièrement attention car cela change une des propriétés, à savoir que la fonction devient continue à gauche et non plus à droite et par ailleurs, pour retrouver la loi de probabilité associée, on a alors

$$\forall\ i \geq 1,\ P(X = x_i) = P(X \leq x_{i+1}) - P(X \leq x_i) = F_X(x_{i+1}) - F_X(x_i)$$

Exemple 4.
Si l'on reprend l'exemple du lancer du dé équilibré avec comme variable X la valeur de la face visible sur le dessus, on a alors :

$\forall x < 1,\ F(x) = 0$

$\forall 1 \leq x < 2,\ F(x) = \frac{1}{6}$

$\forall 2 \leq x < 3,\ F(x) = \frac{2}{6}$

$\forall 3 \leq x < 4,\ F(x) = \frac{3}{6}$

$\forall 4 \leq x < 5,\ F(x) = \frac{4}{6}$

$\forall 5 \leq x < 6,\ F(x) = \frac{5}{6}$

$\forall 6 \leq x,\ F(x) = 1$

Bien faire attention que cela est en accord avec la définition donnée et non avec l'autre convention!

1.2.4 Espérance et Variance

Définition 3.
On appelle ***espérance mathématique*** *de la variable (aléatoire) X, la quantité,* ***si elle existe*** *:*

$$\mathbb{E}(X) = \sum_{i \in \mathbb{N}} x_i * p_i$$

où $\{x_i,\ i \in \mathbb{N}\}$ est l'ensemble des valeurs possibles de la variable X et p_i les probabilités associées.

Remarque 5.
La notion d'existence est très importante. En effet, lorsque la variable prend une infinité de valeurs, la somme qui intervient dans la définition de l'espérance est donc une somme qui porte sur une infinité de termes. Or, on sait qu'une telle somme peut alors valoir l'infini, et dans ce cas, on dit que la somme n'existe pas.
Donc, lorsque la variable ne prend qu'un nombre fini de valeurs, l'existence sera toujours vérifiée. Par contre, lorsqu'il y a une infinité de valeurs possibles, il faudra faire bien attention.

Interprétation :

- la valeur de l'espérance d'une variable X est une valeur numérique unique. Il s'agit de la **moyenne en probabilité** de la variable X. Nous verrons un peu plus tard dans le cours le lien qu'il existe avec la moyenne arithmétique classique.
- $\mathbb{E}(X)$ peut se voir comme le centre de gravité, ou barycentre, des points ω_i d'abscisse x_i affectés des poids p_i.

Remarque 6.
Si $\forall i \in \mathbb{N},\ x_1 \leq X \leq x_n$, autrement dit si x_1 est la plus petite valeur possible et x_n la plus grande, alors $\mathbb{E}(X) \in [x_1, x_n]$.

Exemple 5.
Reprenons l'exemple du lancer du dé équilibré précédent.

$$\mathbb{E}(X) = \sum_{i=1}^{6} i * \frac{1}{6} = \frac{1}{6} * \frac{6*7}{2} = 3,5$$

En effet, $\forall i \in \{1, \ldots, 6\}, x_i = i$ et $p_i = \frac{1}{6}$.

Remarque 7.
A l'origine, l'espérance mathématique a été introduite pour traduire la notion de gain moyen, ou l'espérance de gain, la variable X représentant alors la valeur du gain lors d'une expérience. Par exemple, considérons deux joueurs notés A et B qui jouent à un jeu de dé. On décide que le joueur B gagne les mises si le résultat du dé est supérieur ou égal à 3. Dans l'autre cas, c'est le joueur A qui remporte les mises.
Quelles doivent être les mises a et b respectivement des joueurs A et B pour que le jeu soit équitable?

Soit X_A la variable représentant le gain du joueur A et X_B celle du joueur B. On a :

$$X_A = \begin{cases} a+b & si\ le\ dé \leq 2 \\ 0 & si\ le\ dé \geq 3 \end{cases}$$

et

$$X_B = \begin{cases} a+b & si\ le\ dé \geq 3 \\ 0 & si\ le\ dé \leq 2 \end{cases}$$

D'où :

$$P(X_A = 0) = \frac{4}{6},\ P(X_A = a+b) = \frac{2}{6},\ P(X_B = 0) = \frac{2}{6},\ P(X_B = a+b) = \frac{4}{6}$$

Ainsi :

$$\mathbb{E}(X_A) = 0 * \frac{2}{3} + (a+b)\frac{1}{3} = \frac{a+b}{3} et\ \mathbb{E}(X_B) = 0 * \frac{1}{3} + (a+b)\frac{2}{3} = 2\frac{a+b}{3}.$$

Le jeu est équitable si la mise de départ est égale à l'espérance de gain, soit

$$\begin{cases} \frac{a+b}{3} = a \\ 2\frac{a+b}{3} = b \end{cases}$$

Soit $b = 2a$.
Ce résultat est intuitif car le joueur B a deux fois plus de chance de gagner que le joueur A!

Proposition 3.
Soit X, Y deux variables aléatoires discrètes.

- $\forall a \in \mathbb{R},\ \mathbb{E}(a) = a$
- $\forall a \in \mathbb{R},\ \mathbb{E}(X + a) = \mathbb{E}(X) + a$
- $\forall a \in \mathbb{R},\ \mathbb{E}(aX) = a\mathbb{E}(X)$
- $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$
- $\forall \lambda \in \mathbb{R},\ \forall \mu \in \mathbb{R}, \mathbb{E}(\lambda X + \mu Y) = \lambda\mathbb{E}(X) + \mu\mathbb{E}(Y)$
- *soit g une fonction possédant de bonnes propriétés (que nous ne détaillerons pas ici) telle que l'espérance de $g(X)$ existe, on a $\mathbb{E}(g(X)) = \sum_{i\in\mathbb{N}} p_i * g(x_i)$*

Définition 4.
*On appelle **variance mathématique** de la variable (aléatoire) X, la quantité, **si elle existe** :*

$$\mathbb{V}(X) = \sum_{i\in\mathbb{N}} p_i * (x_i - \mathbb{E}(X))^2$$

où $\{x_i,\ i \in \mathbb{N}\}$ est l'ensemble des valeurs possibles de la variable X et p_i les probabilités associées.

Remarque 8.
Soit X une variable discrète, on a $\mathbb{V}(X) = \mathbb{E}((X - \mathbb{E}(X))^2)$ et donc $\mathbb{V}(X) \geq 0$.

Interprétation :
La variance est un indicateur qui mesure la dispersion des valeurs prises par la variable X autour de sa valeur moyenne (moyenne en probabilité).

Exemple 6.
Si l'on se réfère encore une fois à l'exemple du lancer du dé équilibré, on a :

$$\mathbb{V}(X) = \sum_{i=1}^{6} \frac{1}{6}(i - 3,5)^2 = \frac{35}{12}$$

Proposition 4.
Soit X une variable aléatoire discrète, on a :

- $\forall a \in \mathbb{R},\ \mathbb{V}(a) = 0$
- $\forall a \in \mathbb{R},\ \mathbb{V}(X + a) = \mathbb{V}(X)$
- $\forall a \in \mathbb{R},\ \mathbb{V}(aX) = a^2\mathbb{V}(X)$
- $\mathbb{V}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$ *(attention à la place des parenthèses et des puissances 2)*

1.2.5 Quelques lois usuelles

Loi uniforme discrète

Définition 5.
Soit X une variable aléatoire. On dit que X suit une ***loi uniforme discrète*** *sur l'ensemble $\mathcal{A}$ de cardinal k, ce que l'on note $X \sim \mathcal{U}(\mathcal{A})$ si :*

- *soit x une valeur possible de X, alors $x \in \mathcal{A}$*
- *$\forall x \in \mathcal{A}$, on a $P(X = x) = \frac{1}{k}$*

Interprétation :
Cela signifie que toutes les valeurs contenues dans l'ensemble $\mathcal{A}$ sont équi-probables, c'est à dire qu'elles ont la même probabilité d'apparaître au cours de l'expérience.

Exemple 7.
Un exemple classique d'une telle loi est le tirage de la première boule lors du tirage du loto. Au départ il y a 49 boules dans l'urne. Chaque numéro allant de 1 à 49 à la probabilité $\frac{1}{49}$ d'être tiré.

Remarque 9.
La loi uniforme discrète la plus classique est celle sur l'ensemble $\mathcal{A} = \{1, 2, \ldots, k\}$.

Proposition 5.
Soit X une variable aléatoire de loi uniforme sur l'ensemble $\mathcal{A} = \{x_1, \ldots, x_k\}$.

- *l'espérance de X existe et vaut $\mathbb{E}(X) = \frac{\sum_{i=1}^{k} x_i}{k}$*
- *dans le cas particulier où $\forall i \in \{1, \ldots, k\},\ x_i = i$, on a $\mathbb{E}(X) = \frac{k+1}{2}$*
- *la variance de X existe et vaut $\mathbb{V}(X) = \frac{x_1^2 + x_2^2 + \ldots + x_k^2}{k} - (\mathbb{E}(X))^2$*
- *dans le cas particulier où $\forall i \in \{1, \ldots, k\},\ x_i = i$, on a $\mathbb{V}(X) = \frac{k^2-1}{12}$*
- *la fonction de répartition F a pour expression :*

$$F(x) = \begin{cases} 0 & si\ x < x_1 \\ \frac{1}{k} & si\ x_1 \leq x < x_2 \\ \frac{2}{k} & si\ x_2 \leq x < x_3 \\ \frac{i}{k} & si\ x_i \leq x < x_{i+1}\ \forall i \in \{3, \ldots (k-1)\} \\ 1 & si\ x_k \leq x \end{cases}$$

Preuve.
Comme l'ensemble des valeurs possibles pour X est en nombre fini, l'espérance et la variance existent.
Par définition, on :

$$\mathbb{E}(X) = \sum_{i \in \mathbb{N}} x_i * p_i$$

or, $\forall i \in \{1, \ldots, k\}$, $p_i = \frac{1}{k}$ d'où le résultat dans le cadre général.
En ce qui concerne le cas particulier où $\forall i \in \{1, \ldots, k\}$, $x_i = i$, on utilise en plus le fait que $\sum_{i=1}^{k} i = \frac{k*(k+1)}{2}$. Pour ce qui est de la variance, on utilise le fait que $\mathbb{V}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$ ainsi que le fait que $\mathbb{E}(X^2) = \sum_{i \in \mathbb{N}} x_i^2$. Ces éléments permettent d'obtenir le résultat du cadre général.
Pour le cas particulier, on utilise le résultat suivant à savoir $\sum_{i=1}^{k} i^2 = \frac{k*(k+1)*(2k+1)}{6}$. Cette formule se démontre très facilement par récurrence. Ensuite :

$$\begin{aligned} \mathbb{V}(X) &= \frac{k(k+1)(2k+1)}{6k} - \frac{(k+1)^2}{4} \\ &= \frac{(k+1)(2k+1)}{6} - \frac{(k+1)^2}{4} \\ &= \frac{k+1}{12} * (4k + 2 - 3k - 3) \\ &= \frac{k+1}{12}(k-1) \\ &= \frac{k^2-1}{12} \end{aligned}$$

Pour ce qui est de la fonction de répartition, il suffit de revenir à la définition et de se rendre compte que la fonction saute d'un pas de $\frac{1}{k}$ à chaque fois que l'on rencontre une des valeurs prise par la variable. ■

Loi de Bernoulli

Définition 6.
On dit qu'une variable X suit une ***loi de Bernoulli*** *(attention à l'orthographe) de* ***paramètre*** *p, avec $p \in [0,1]$, ce que l'on note $X \sim \mathcal{B}(p)$ si :*

- *$X(\Omega) = \{0;1\}$ (à savoir les valeurs possibles pour la variable sont 0 ou 1)*
- *$P(X=0) = 1-p$ et $P(X=1) = p$*

Exemple 8.
De façon générale, cette variable permet de modéliser toutes les expériences ne comprenant que deux issues possibles qui seront alors codées par 0 en cas d'échec et 1 en cas de succès.

- *Un premier exemple concret est : on lance un dé équilibré et on se déclare vainqueur si le résultat sur la face supérieure du dé est pair. Dans ce cas, $p = \frac{1}{2}$.*
- *Un second exemple concret est : on lance un dé équilibré et on se déclare vainqueur si le résultat sur la face supérieure est 1, 2, 3 ou 4. Dans ce cas, $p = \frac{4}{6}$*

Pourquoi ces deux exemples, simplement pour attirer votre attention sur le fait que la valeur de p n'est pas obligatoirement $\frac{1}{2}$ et qu'en général, il faut identifier ce paramètre!

Proposition 6.
Soit X une variable de loi de Bernoulli de paramètre p, on a :

- *l'espérance de X existe et vaut p*
- *la variance de X existe et vaut $p(1-p)$ (remarque on note communément $q = 1-p$)*
- *la fonction de répartition F a pour expression :*

$$F(x) = \begin{cases} 0 & si\ x < 0 \\ 1-p & si\ 0 \leq x < 1 \\ 1 & si\ 1 \leq x \end{cases}$$

Preuve.
L'existence de l'espérance et de la variance provient du fait que le nombre de valeurs possibles pour X est 2.

$$\begin{aligned} \mathbb{E}(X) &= \sum_{i \in \mathbb{N}} x_i * p_i \\ &= 0 * (1-p) + 1 * p \\ &= p \end{aligned}$$

$$\begin{aligned}\mathbb{V}(X) &= \mathbb{E}(X^2) - (\mathbb{E}(X))^2 \\ &= 0^2 * (1-p) + 1^2 * p - p^2 \\ &= p - p^2 \\ &= p(1-p)\end{aligned}$$

■

Loi binomiale

Définition 7.
On dit qu'une variable X suit une ***loi de Binomiale*** *de* ***paramètres*** *n* ***et*** *p, avec $p \in [0,1]$ et $n \in \mathbb{N}^*$, ce que l'on note $X \sim \mathcal{B}(n,p)$ si :*

- *$X(\Omega) = \{0; 1; \ldots; n\}$ (à savoir les valeurs possibles pour la variable sont tous les entiers entre 0 et n, bornes comprises)*
- *$\forall k \in \{0; 1; \ldots; n\},\ P(X = k) = C_n^k p^k (1-p)^{n-k}$*

Remarque 10.
$C_n^k = \frac{n!}{k!*(n-k)!}$ avec $k! = 1 * 2 * \ldots * k$. Il s'agit du nombre de combinaisons de k parmi n, à savoir le nombre de façons de tirer k boules parmi n sans notion d'ordre.

Modélisation :
Cette loi de probabilité modélise des situations où l'expérience globale que l'on considère se décompose en la succession de n épreuves de Bernoulli de paramètre p (à savoir épreuves à deux issues avec une probabilité de succès égale à p) **identiques** et **"indépendantes"** et où l'on s'intéresserait au nombre de succès.

Remarque 11.

- *Dire que des épreuves sont indépendantes signifie que l'issue d'une épreuve n'a aucune influence sur la survenue et l'issue d'une autre épreuve.*
- *Ainsi définie, une variable X de loi binomiale de paramètres n et p peut se décomposer de la manière suivante :*
$$X = X_1 + X_2 + \ldots + X_n$$
où $\forall i \in \{1, 2, \ldots, n\}$, X_i est une variable de Bernoulli de paramètre p et l'ensemble de ces variables sont "indépendantes".
- *C'est à partir de la modélisation qu'il est facile de retrouver la formule de probabilité.*

Preuve.

Si l'on s'intéresse à l'événement $X = k$ pour $k \in \{0; 1; \ldots; n\}$, cela signifie que l'on a, au cours de n épreuves, k succès et $n - k$ échecs. Or, la probabilité d'un succès est p et celle d'un échec est $1 - p$. Ensuite, le nombre de façons d'avoir k succès et $n - k$ échecs parmi n épreuves est C_n^k. Ensuite, toutes les combinaisons de k succès ont la même probabilité de survenir, et cette probabilité est de $p^k(1-p)^{n-k}$.
Ces considérations nous permettent d'obtenir la formule de probabilité annoncée. ■

Proposition 7.
Soit X une variable aléatoire de loi binomiale de paramètres n et p.

- *l'espérance de X existe et vaut $\mathbb{E}(X) = np$*
- *la variance de X existe et vaut $\mathbb{V}(X) = np(1-p)$*

Preuve.
Puisque X suit une loi binomiale de paramètres n et p, il existe n variables aléatoires indépendantes et identiquement distribuées $X_1, \ldots, X_n$ telles que $X = X_1 + \ldots + X_n$. D'où :

$$\begin{aligned}
\mathbb{E}(X) &= \mathbb{E}(X_1 + \ldots + X_n) \\
&= \sum_{i=1}^{n} \mathbb{E}(X_i) \quad \text{(d'après les propriétés de l'espérance)} \\
&= \sum_{i=1}^{n} p \quad \text{(car les } X_i \text{ sont des variables de Bernoulli de paramètre } p) \\
&= np
\end{aligned}$$

$$\begin{aligned}
\mathbb{V}(X) &= \mathbb{V}(X_1 + \ldots + X_n) \\
&= \sum_{i=1}^{n} \mathbb{V}(X_i) \quad \text{(d'après les propriétés de la variance} \\
&\qquad\qquad \text{et parce que les variables } X_i \text{ sont indépendantes)} \\
&= \sum_{i=1}^{n} p(1-p) \quad \text{(car les } X_i \text{ sont des variables de Bernoulli de paramètre } p) \\
&= np(1-p)
\end{aligned}$$

■

Loi de Poisson

Définition 8.
*On dit qu'une variable X suit une **loi de Poisson** de **paramètre** λ, avec $\lambda > 0$, ce que l'on note $X \sim \mathcal{P}(\lambda)$ si :*

- *$X(\Omega) = \mathbb{N}$ (à savoir les valeurs possibles pour la variable sont tous les entiers positifs ou nul)*

- $\forall k \in \mathbb{N},\ P(X = k) = \frac{\lambda^k * e^{-\lambda}}{k!}$

Remarque 12.
Comment on arrive t'on à la loi de Poisson?
Supposons que l'on veuille déterminer le nombre de fois où un événement se produit, mais que, en dépit de certaines ressemblances, la situation diffère de celle évoquée dans le cadre de la loi binomiale selon les différences ci-dessous :

- *déclarer que l'événement s'est produit un certain nombre de fois ne résulte pas de l'examen d'une quantité dénombrable "d'individus", mais de l'examen d'un ensemble continu ou encore d'un ensemble représentable sous forme d'un intervalle du type* $]0, t[$*, tel que la longueur, l'aire, ...*
- *l'accomplissement de l'événement "succès" dans un certain sous-intervalle de* $]0, t[$ *n'influe pas la réalisation de l'événement dans un autre sous-intervalle (notion d'indépendance)*
- *la probabilité que l'événement se produise dans un intervalle de temps très petit est presque nulle (notion d'événement assez rare)*
- *le nombre de réalisations de l'événement "succès" dans l'intervalle* $]0, t[$ *est connu!*

Par exemple, tout ceci se produira si l'on souhaite déterminer le nombre de tremblements de terre se produisant dans les caraïbes en un an, ou encore le nombre d'appels enregistrés en une semaine au 911 d'une région donnée.

Une modélisation est :

1. *On divise l'intervalle* $]0, t[$ *en n sous-intervalles. Afin qu'il n'y ait que 2 issues possibles (réalisation ou non-réalisation), et afin qu'on ne risque pas de rencontrer plus d'une seule réalisation de l'événement par sous-intervalle, nous ferons tendre n vers* $+\infty$ *puisque nous savons que, par postulat, la probabilité de réalisation sur un intervalle très petit est quasiment nulle!*
2. *Soit X le nombre de réalisation sur ces n sous-intervalles de* $]0, t[$ *et soit* λ *le nombre moyen de réalisations sur* $]0, t[$*. Comme probabilité qu'un événement se produise sur un sous-intervalle, on pose* $p = \frac{\lambda}{n}$*.*
3. *Soit* $x \in \{0, \ldots, n\}$*, on va chercher à déterminer l'expression de* $P(X = x)$ *à l'aide de la variable X sachant que* $X \sim \mathcal{B}(n, p)$ *avec* $n \to +\infty$*. Ainsi, pour* $x \in \{0, \ldots, n\}$*,* $P(X = x) = C_n^x p^x (1 - p)^{(n-x)}$*.*
4. *Comme on fait tendre n vers* $+\infty$*, on peut utiliser la formule de Stirling qui dit que* $n! \sim \sqrt{2\pi n} \left(\frac{n}{e}\right)^n$*.*
5. *En utilisant le fait que* $a^b = \exp b.ln(a)$*, on obtient la formule annoncée.*

Proposition 8.
Soit X une variable aléatoire de loi de Poisson de paramètre λ*.*

- *l'espérance de X existe et vaut* $\mathbb{E}(X) = \lambda$
- *la variance de X existe et vaut* $\mathbb{V}(X) = \lambda$

1.3 Variable aléatoire continue (ou à densité)

1.3.1 Définition

Définition 9.
On appelle ***variable aléatoire continue*** *toute application* $X : \Omega \to \mathbb{R}$ *telle que pour tout intervalle* $I \subset \mathbb{R}$*, on ait* $X^{-1}(I)$ *soit un événement.*

Exemple 9.
La durée de vie d'une ampoule ou encore le salaire d'un individu tiré au sort dans une population sont représentés par des variables aléatoires continues. En effet, si l'on se donne deux valeurs réelles quelconques, on peut toujours trouver une valeur intermédiaire qui soit une valeur possible pour la variable considérée.

1.3.2 Loi de probabilité

Pour une variable aléatoire continue, sa loi de probabilité est déterminée par deux éléments que sont :

- $X(\Omega)$: l'ensemble des valeurs admissibles pour la variable X. Cet ensemble est donné sous forme d'intervalle fini ou non.
- une **fonction de densité** encore appelée fonction de probabilité.

Définition 10.
On dit qu'une fonction f *est une* ***fonction de densité*** *si :*

- $\forall x \in \mathbb{R},\ f(x) \geq 0$
- $\int_{-\infty}^{+\infty} f(x).dx = 1$

Soit X *une variable aléatoire de densité* f*, on définit alors la probabilité de* X *d'appartenir à un intervalle* $[a, b]$ *(avec* $a \leq b$*) de la manière suivante :*

$$P(a \leq X \leq b) = \int_a^b f(x).dx$$

Remarque 13.

- *les deux propriétés d'une fonction de densité peuvent encore s'énoncer de la manière suivante : l'aire comprise sous la courbe représentative de* f *et l'axe des abscisses vaut 1*
- *Soit* $a < b$*, on a :*
  - $P(X \in [a, b[) = P(X \in]a, b[) = P(X \in]a, b]) = P(X \in [a, b])$*, autrement dit pour une variable continue le fait de considérer des intervalles fermés ou ouverts ne changent rien dans le calcul d'une probabilité.*
  - $P(X = a) = 0$*, autrement dit une variable aléatoire continue ne charge aucun point.*

Exemple 10.
Soit f la fonction définie par :

$$f(x) = \begin{cases} \frac{a}{x} & si \ -e \leq x < -1 \\ x + 1 - a & si \ -1 \leq x < 0 \\ 0 & sinon \end{cases}$$

Déterminer la valeur du paramètre a afin que la fonction f soit une fonction de densité.

Résolution :
On sait que f ne sera une fonction de densité que si :

- $\forall x \in \mathbb{R}, \ f(x) \geq 0$
- $\int_{-\infty}^{+\infty} f(x).dx = 1$

Calculons $\int_{-\infty}^{+\infty} f(x).dx$ en fonction du paramètre a.

$$\begin{aligned} \int_{-\infty}^{+\infty} f(x).dx &= \int_{-\infty}^{-e} f(x).dx + \int_{-e}^{-1} f(x).dx + \int_{-1}^{0} f(x).dx + \int_{0}^{+\infty} f(x).dx \\ &= \int_{-e}^{-1} \frac{a}{x}.dx + \int_{-1}^{0} (x + 1 - a).dx \\ &= a\left[ln(|x|)\right]_{-e}^{-1} + \left[\frac{x^2}{2} + (1-a)x\right]_{-1}^{0} \\ &= -a - \frac{1}{2} + (1 - a) \\ &= -2a + \frac{1}{2} \end{aligned}$$

Ainsi, f ne pourra être une fonction de densité que si $-2a + \frac{1}{2} = 1$ soit $a = \frac{-1}{4}$.

Attention : *Ceci ne suffit pas à prouver qu'avec $a = \frac{-1}{4}$, f est une fonction de densité. En effet, il faut encore s'assurer de la positivité globale de la fonction!*
Soit $x \in \mathbb{R}$:

- *si $x \in [-e, -1[$, $f(x) = \frac{-1}{4x}$. Puisque x est négatif, on a $f(x) \geq 0$.*
- *si $x \in [-1, 0[$, $f(x) = x + \frac{5}{4}$, donc $f(x) \geq 0$.*
- *si $x \geq 0$ ou $x < -e$, $f(x) = 0$, donc $f(x) \geq 0$.*

Exemple 11.
Soit X une variable aléatoire continue dont la fonction de densité est donnée par la fonction f définie ci-avant. Calculer $P(X \in [-2; \frac{-1}{2}])$.

Résolution :

$$\begin{aligned}
P\left(X \in [-2; \frac{-1}{2}]\right) &= \int_{-2}^{\frac{-1}{2}} f(x).dx \\
&= \int_{-2}^{-1} \frac{-1}{4x} dx + \int_{-1}^{\frac{-1}{2}} \left(x + \frac{5}{4}\right) dx \\
&= \frac{-1}{4} \left[ln(|x|)\right]_{-2}^{-1} + \left[\frac{x^2}{2} + \frac{5}{4}x\right]_{-1}^{\frac{-1}{2}} \\
&= \frac{1}{4}ln(2) + \left(\frac{1}{8} - \frac{5}{8}\right) - \left(\frac{1}{2} - \frac{5}{4}\right) \\
&= \frac{1}{4}ln(2) - \frac{1}{2} + \frac{3}{4} \\
&= \frac{1}{4}ln(2) + \frac{1}{4} \\
&\simeq 0.42
\end{aligned}$$

1.3.3 Fonction de répartition

Définition 11.
*La **fonction de répartition** F d'une variable aléatoire continue X, de fonction de densité f, est définie par :*

$$\forall\ x \in \mathbb{R},\ F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t).dt$$

Remarque 14.
$\forall x \in \mathbb{R}$, on a aussi $F(x) = P(X < x)$.

Proposition 9.
Une fonction de répartition satisfait les points suivants :

- *la fonction est croissante*
- *la fonction est continue*
- $\forall\ x \in \mathbb{R},\ 0 \leq F(x) \leq 1$
- $\lim\limits_{x \to -\infty} F(x) = 0$ *et* $\lim\limits_{x \to +\infty} F(x) = 1$
- $\forall a \leq b,\ P(X \in [a, b]) = F(b) - F(a)$
- $f = F'$ *en tout point x où la fonction F est dérivable*

Exemple 12.
On reprend la fonction de densité précédente. Déterminer la fonction de répartition associée.

Résolution :
Soit $x \in \mathbb{R}$, par définition on a : $F(x) = \int_{-\infty}^{x} f(t).dt$.

1er cas : si $x < -e$, on a $F(x) = \int_{-\infty}^{x} 0.dt = 0$

2ème cas : si $x \in [-e; -1[$, on a :

$$\begin{aligned} F(x) &= \int_{-\infty}^{x} f(t).dt \\ &= \int_{-\infty}^{-e} 0.dt + \int_{-e}^{x} \frac{-1}{4t} dt \\ &= \frac{-1}{4} \left[ln(|t|)\right]_{-e}^{x} \\ &= \frac{-1}{4}(ln(|x|) - 1) \end{aligned}$$

3ème cas : si $x \in [-1; 0[$, on a :

$$\begin{aligned} F(x) &= \int_{-\infty}^{x} f(t).dt \\ &= \int_{-\infty}^{-e} 0.dt + \int_{-e}^{-1} \frac{-1}{4t} dt + \int_{-1}^{x} \left(t + \frac{5}{4}\right) dt \\ &= \frac{-1}{4} \left[ln(|t|)\right]_{-e}^{-1} + \left[\frac{t^2}{2} + \frac{5t}{4}\right]_{-1}^{x} \\ &= \frac{1}{4} + \frac{x^2}{2} + \frac{5x}{4} - \frac{1}{2} + \frac{5}{4} \\ &= \frac{x^2}{2} + \frac{5x}{4} + 1 \end{aligned}$$

4ème cas : si $x > 0$, $F(x) = 1$ par définition d'une fonction de densité (ou par calcul)

Remarque 15.
Utilisation de la fonction de répartition :
Soit X une variable aléatoire continue admettant pour fonction de densité la fonction f suivante :

$$f(x) = \begin{cases} x & si\ 0 \leq x < 1 \\ \frac{x}{15} & si\ 1 \leq x < 4 \\ 0 & sinon \end{cases}$$

Déterminer la loi de la variable aléatoire $Y = 2X + 1$.

Résolution :
Pour déterminer la loi de la variable Y, on va chercher à calculer la fonction de répartition de

Y, notée G, en fonction de celle de X, notée F.
Soit $y \in \mathbb{R}$,

$$\begin{aligned} G(y) &= P(Y \leq y) \\ &= P(2X+1 \leq y) \\ &= P\left(X \leq \frac{y-1}{2}\right) \\ &= F\left(\frac{y-1}{2}\right) \end{aligned}$$

Calcul de F :

1er cas : si $x < 0$, on a $F(x) = \int_{-\infty}^{x} 0.dt = 0$

2ème cas : si $x \in [0;1[$, on a $F(x) = \frac{x^2}{2}$

3ème cas : si $x \in [1;4[$, on a :

$$\begin{aligned} F(x) &= \int -\infty^{x} f(t).dt \\ &= \int_{-\infty}^{0} 0.dt + \int_{0}^{1} tdt + \int_{1}^{x} \frac{t}{15}dt \\ &= \left[\frac{t^2}{2}\right]_0^1 + \left[\frac{t^2}{30}\right]_1^x \\ &= \frac{1}{2} + \frac{x^2}{30} - \frac{1}{30} \\ &= \frac{x^2}{2} + \frac{14}{30} \end{aligned}$$

4ème cas : si $x \geq 4$, $F(x) = 1$ par définition d'une fonction de densité (ou par calcul)

On en déduit donc :

$$G(y) = \begin{cases} 0 & si\ \frac{y-1}{2} < 0 \Leftrightarrow y < 1 \\ \frac{(y-1)^2}{8} & si\ \frac{y-1}{2} \in [0,1[\Leftrightarrow y \in [1;3[\\ \frac{(y-1)^2}{120} + \frac{14}{30} & si\ \frac{y-1}{2} \in [1,4[\Leftrightarrow y \in [3;9[\\ 1 & si\ \frac{y-1}{2} \geq 4 \Leftrightarrow y \geq 9 \end{cases}$$

En procédant par dérivation de la fonction précédente aux points de continuité, on obtient :

$$G(y) = \begin{cases} \frac{(y-1)}{4} & si\ y \in]1;3[\\ \frac{(y-1)}{60} & si\ y \in]3;9[\\ 0 & sinon \end{cases}$$

Attention : Aux points de recollement, la différentiabilité n'est pas assurée. De ce fait, on ouvre les intervalles lors de la dérivation, puisque cela ne modifie en rien les calculs ensuite.

1.3.4 Espérance et Variance

Définition 12.
On appelle ***espérance mathématique*** *de la variable aléatoire X de fonction de densité f, la quantité,* ***si elle existe*** *:*

$$\mathbb{E}(X) = \int_{-\infty}^{+\infty} x.f(x).dx$$

Proposition 10.
Soit X, Y deux variables aléatoires continues.

- $\forall a \in \mathbb{R},\ \mathbb{E}(X + a) = \mathbb{E}(X) + a$
- $\forall a \in \mathbb{R},\ \mathbb{E}(aX) = a\mathbb{E}(X)$
- $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$
- $\forall \lambda \in \mathbb{R},\ \forall \mu \in \mathbb{R}, \mathbb{E}(\lambda X + \mu Y) = \lambda\mathbb{E}(X) + \mu\mathbb{E}(Y)$
- *soit g une fonction possédant de bonnes propriétés (que nous ne détaillerons pas ici) telle que l'espérance de $g(X)$ existe, on a $\mathbb{E}(g(X)) = \int_{-\infty}^{+\infty} g(x).f(x).dx$*

Définition 13.
On appelle ***variance mathématique*** *de la variable aléatoire X de fonction de densité f, la quantité,* ***si elle existe*** *:*

$$\mathbb{V}(X) = \int_{-\infty}^{+\infty} (x - \mathbb{E}(X))^2.f(x).dx$$

Remarque 16.
Soit X une variable aléatoire continue, on a, comme dans le cas discret, $\mathbb{V}(X) = \mathbb{E}((X-\mathbb{E}(X))^2)$ et donc $\mathbb{V}(X) \geq 0$.

Proposition 11.
Soit X une variable aléatoire continue, on a :

- $\forall a \in \mathbb{R},\ \mathbb{V}(X + a) = \mathbb{V}(X)$
- $\forall a \in \mathbb{R},\ \mathbb{V}(aX) = a^2\mathbb{V}(X)$
- $\mathbb{V}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$ *(attention à la place des parenthèses et des puissances 2)*

1.3.5 Quelques lois usuelles

Loi uniforme continue

Définition 14.
Soit X une variable aléatoire continue. On dit que X suit une ***loi uniforme continue*** *sur l'intervalle $[a, b]$, ce que l'on note $X \sim \mathcal{U}([a, b])$ si :*

- *soit x une valeur possible de X, alors $x \in [a, b]$*

- *la fonction de densité de la variable X est définie par :*

$$f(x) = \begin{cases} \frac{1}{b-a} & si\ x \in [a,b] \\ 0 & sinon \end{cases}$$

Remarque 17.
Il est nécessaire que les bornes a et b de l'intervalle considéré soient finies;

Proposition 12.
Soit X une variable aléatoire de loi uniforme sur l'intervalle $[a,b]$.

- *l'espérance de X existe et vaut $\mathbb{E}(X) = \frac{b+a}{2}$*
- *la variance de X existe et vaut $\mathbb{V}(X) = \frac{(b-a)^2}{12}$*
- *la fonction de répartition F a pour expression :*

$$F(x) = \begin{cases} 0 & si\ x < a \\ \frac{x-a}{b-a} & si\ a \leq x < b \\ 1 & si\ x \geq b \end{cases}$$

Preuve.
Les calculs ne sont pas à faire en cours, mais ils sont mis dans ces notes de cours.
Puisque la fonction est continue sur un support compact, l'espérance et la variance existent.

$$\begin{aligned} \mathbb{E}(X) &= \int_{-\infty}^{+\infty} t * f(t).dt \\ &= \int_a^b \frac{t}{b-a}.dt \\ &= \frac{1}{b-a}\left[\frac{t^2}{2}\right]_a^b \\ &= \frac{1}{b-a}\frac{b^2-a^2}{2} \\ &= \frac{b+a}{2} \end{aligned}$$

$$\begin{aligned}
\mathbb{E}(X^2) &= \int_{-\infty}^{+\infty} t^2 * f(t).dt \\
&= \int_a^b \frac{t^2}{b-a}.dt \\
&= \frac{1}{b-a}\left[\frac{t^3}{3}\right]_a^b \\
&= \frac{1}{b-a}\frac{b^3-a^3}{3} \\
&= \frac{b^2+ab+a^2}{3}
\end{aligned}$$

D'où :

$$\begin{aligned}
\mathbb{V}(X) &= \mathbb{E}(X^2) - (\mathbb{E}(X))^2 \\
&= \frac{b^2+ab+a^2}{3} - \left(\frac{b+a}{2}\right)^2 \\
&= \frac{b^2+ab+a^2}{3} - \frac{b^2+2ab+a^2}{4} \\
&= \frac{1}{12}\left(4b^2+4ab+4a^2-3b^2-6ab-3a^2\right) \\
&= \frac{b^2-2ab+a^2}{12} \\
&= \frac{(b-a)^2}{12}
\end{aligned}$$

$$F(x) = \begin{cases} \int_{-\infty}^{x} f(t).dt = \int_{-\infty}^{x} 0.dt = 0 & \text{si } x < a \\ \int_{-\infty}^{x} f(t).dt = \int_{a}^{x} \frac{1}{b-a}.dt = \frac{x-a}{b-a} & \text{si } a \leq x < b \\ 1 & \text{si } x \geq b \end{cases}$$

■

Loi Exponentielle

Définition 15.
*On dit qu'une variable continue X suit une **loi exponentielle** de **paramètre** θ, avec $\theta > 0$, ce que l'on note $X \sim \mathcal{E}(\theta)$ si :*

- $X(\Omega) = [0, +\infty[$ *(à savoir les valeurs possibles pour la variable sont les réels positifs ou nuls)*

- *la fonction de densité associée est définie par :*

$$f(x) = \begin{cases} \theta \exp(-\theta.x) & si\ x \geq 0 \\ 0 & sinon \end{cases}$$

Exemple 13.
Ce type de variable aléatoire est souvent utilisée pour représenter une durée de vie, comme la durée de vie d'un matériel donné, la durée de chômage, la durée d'hospitalisation, ...

Proposition 13.
Soit X une variable de loi exponentielle de paramètre θ, on a :

- *l'espérance de X existe et vaut $\frac{1}{\theta}$*
- *la variance de X existe et vaut $\frac{1}{\theta^2}$*
- *la fonction de répartition F a pour expression :*

$$F(x) = \begin{cases} 1 - \exp(-\theta.x) & si\ x \geq 0 \\ 0 & sinon \end{cases}$$

Preuve.
La preuve n'est pas à faire en cours, mais elle est donnée dans ces notes de cours.
L'existence de l'espérance et de la variance provient de la prédominance de l'exponentielle négative sur tout polynôme au voisinage de l'infini.
Soit T un réel positif.

$$\begin{aligned} \int_{-\infty}^{T} f(t).t.dt &= \int_{0}^{T} t.\theta.\exp(-\theta.t).dt \\ &= [-t\exp(-\theta.t)]_0^T + \int_0^T \exp(-\theta.t).dt \\ &= -T\exp(-\theta.T) - \left[\frac{\exp(-\theta.t)}{\theta}\right]_0^T \\ &= -T\exp(-\theta.T) - \frac{\exp(-\theta.T)}{\theta} + \frac{1}{\theta} \end{aligned}$$

On fait alors tendre T vers $+\infty$. On obtient alors que

$$\lim_{T\to+\infty} \int_{-\infty}^{T} f(t).t.dt = \frac{1}{\theta}$$

Puisque la limite existe, elle est la valeur de l'espérance.

Soit T un réel positif.

$$\begin{aligned}\int_{-\infty}^{T} f(t).t^2.dt &= \int_0^T t^2.\theta.\exp(-\theta.t).dt \\ &= \left[-t^2\exp(-\theta.t)\right]_0^T + \int_0^T 2t.\exp(-\theta.t).dt \\ &= -T^2\exp(-\theta.T) + \frac{2}{\theta}\int_{-\infty}^{T} f(t).t.dt\end{aligned}$$

On fait alors tendre T vers $+\infty$. On obtient alors que

$$\lim_{T\to+\infty}\int_{-\infty}^{T} f(t).t^2.dt = \frac{2}{\theta}\mathbb{E}(X)$$

D'où :

$$\mathbb{V}(X) = \frac{2}{\theta^2} - \left(\frac{1}{\theta}\right)^2 = \frac{1}{\theta^2}$$

$$F(x) = \begin{cases} \int_{-\infty}^{x} f(t).dt = \int_0^x 0.dt = 0 & \text{si } x < 0 \\ \int_{-\infty}^{x} f(t).dt = \int_0^x \theta\exp(-\theta.t).dt = [-\exp(-\theta.t)]_0^x = 1-\exp(-\theta.x) & \text{si } 0 \le x \end{cases}$$

■

loi normale (ou loi gaussienne)

Définition 16.
*On dit qu'une variable X suit une **loi Normale** de **paramètres** μ **et** σ^2, avec $\mu \in \mathbb{R}$ et $\sigma^2 > 0$, ce que l'on note $X \sim \mathcal{N}(\mu, \sigma^2)$ si :*

- $X(\Omega) = \mathbb{R}$
- *la fonction de densité associée est définie par* $\forall x \in \mathbb{R},\ f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$.

Remarque 18.

- *la fonction de répartition n'est pas calculable en raison de la présence, dans l'exponentielle, d'un polynôme du second degré*
- *par voie de conséquence, pour tout réel a et b, si $X \sim \mathcal{N}(\mu, \sigma^2)$, $P(X \in [a, b])$, n'est pas calculable. Par contre, il est possible d'en obtenir des approximations en utilisant par exemple la méthode des trapèzes afin d'évaluer les intégrales.*
- *la fonction de densité est symétrique par rapport à l'axe $x = \mu$, c'est à dire que $\forall x \in \mathbb{R},\ f(\mu + x) = f(\mu - x)$.*

Proposition 14.
Soit X une variable aléatoire de loi normale de paramètres μ et σ^2.

- *l'espérance de X existe et vaut $\mathbb{E}(X) = \mu$*
- *la variance de X existe et vaut $\mathbb{V}(X) = \sigma^2$*

Remarque 19.

- *il existe une loi normale plus particulière. Il s'agit de la loi normale centrée réduite qui correspond aux paramètres $\mu = 0$ et $\sigma^2 = 1$.*
- *soit X une loi normale de paramètres μ et σ^2. Si l'on pose $Y = \frac{X-\mu}{\sigma}$, alors Y est une variable aléatoire de loi normale centrée réduite.*
- *Soit X une loi normale centrée réduite. Soit μ un réel et σ^2 un réel strictement positif. Si l'on pose $Y = \sigma X + \mu$, alors Y est une variable aléatoire de loi normale de paramètres μ et σ^2.*
- *On constate ainsi qu'il est possible, via une transformation affine, de se ramener d'une loi normale quelconque à une loi normale centrée réduite. Ceci est très important car comme dit ci-avant, il est impossible de calculer des probabilités exactes. Cependant, à l'aide de méthodes d'approximation, à l'image de la méthode des trapèzes, on peut obtenir des approximations. Pour autant, mettre en application ces méthodes à l'aide d'une papier et d'un crayon se révèle difficile. C'est la raison pour laquelle on dispose d'une table associée à la loi normale centrée réduite qui donne des approximations de probabilités. A l'aide de cette table, dont vous pouvez trouver un exemplaire en annexe, il est alors possible d'évaluer les probabilités pour toute loi normale.*
- *Soit $X \sim \mathcal{N}(0,1)$, $\forall a \in \mathbb{R},\ P(X < -a) = P(X > a)$.*

1.3.6 Quelques résultats importants de probabilités

Inégalité de Markov

Proposition 15.
Soit X une variable aléatoire positive qui possède une espérance, alors :

$$\forall \lambda > 0,\ P(X \geq \lambda \mathbb{E}(X)) \leq \frac{1}{\lambda}$$

ou encore :

$$P(X \geq \lambda) \leq \frac{\mathbb{E}(X)}{\lambda}$$

Remarque 20.
Cette inégalité est utile d'un point de vue théorique mais non pratique car :

- *si $\lambda \leq 1$, l'inégalité sous sa première forme ne dit rien car cela revient à comparer une probabilité à une grandeur supérieure à 1!*

- *en général, la valeur numérique donnée par la borne supérieure est bien supérieure à la valeur de la probabilité!*

Remarque 21.
Extension :
Soit X une variable aléatoire de signe quelconque telle que $\mathbb{E}(|X|^k)$ existe (ce qui peut encore se dire que $|X|$ possède un moment d'ordre k). Alors :

$$\forall \lambda > 0,\ P(|X|^k \geq \lambda) \leq \frac{\mathbb{E}(|X|^k)}{\lambda}$$

ou :

$$\forall \epsilon > 0,\ P(|X| \geq \epsilon) \leq \frac{\mathbb{E}(|X|^k)}{\epsilon^k}$$

Inégalité de Bienaymé-Tchebychev

Proposition 16.
Soit X une variable aléatoire dont la variance existe, alors :

$$\forall \epsilon > 0,\ P(|X - \mathbb{E}(X) \geq \epsilon) \leq \frac{\mathbb{V}(X)}{\epsilon^2}$$

Remarque 22.

- *Cette inégalité résulte uniquement de l'application de l'extension de l'inégalité de Markov à la variable $X - \mathbb{E}(X)$ avec $k = 2$.*
- *Cette nouvelle inégalité relie la probabilité de X de s'écarter de sa moyenne en probabilité $\mathbb{E}(X)$, à sa variance qui est justement un indicateur de dispersion de $\mathbb{E}(X)$.*
- *Si on pose $\epsilon = \sigma(X) = \sqrt{\mathbb{V}(X)}$, on obtient :*

  $$P(|X - \mathbb{E}(X)| \geq \sigma(X)) \leq 1$$

  Ainsi, cette inégalité n'a d'intérêt que lorsque $\epsilon \geq \sigma(X)$.

Remarque 23.
Ci-dessous, nous allons voir la limitation pratique de cette inégalité.
Soit X une variable aléatoire de loi normale de paramètres m et σ^2. Soit a un paramètre positif.

$$P(|X - m| \geq a\sigma) = 1 - P(|U| < a)$$

où U est une variable aléatoire de loi normale centrée réduite.
Considérons $a = 1,5$.

$$
\begin{aligned}
P(|X-m| \geq a\sigma) &= 1 - P(-a < U < a) \\
&= 1 - (P(U < a) - P(U \leq -a)) \\
&= 1 - (P(U < a) - P(U \geq a)) \text{ } \textit{par symétrie de la loi normale centrée réduite} \\
&= 1 - P(U < a) + P(U \geq a) \\
&= 1 - P(U < a) + 1 - P(U < a) \\
&= 2 - 2 * P(U < a) \\
&= 2 - 2 * 0.9332 \text{ } \textit{par lecture de la table} \\
&= 0.1336
\end{aligned}
$$

Or, le majorant donné par l'inégalité de Bienaymé-Tchebychev est :

$$\frac{\mathbb{V}(X)}{a^2\sigma^2} = \frac{1}{a^2} = 0.444444...$$

Ceci prouve que la portée pratique de cette inégalité est faible!

Exemple 14.
Une utilisation cependant de cette inégalité.
On lance un dé équilibré à 6 faces. On s'intéresse à l'événement suivant : la fréquence d'apparition du 6 est comprise entre 1/6-0.01 et 1/6+0.01. Combien de lancers suffit-il de réaliser pour pouvoir affirmer que la probabilité de l'événement précédent soit supérieure ou égale à 95%?

Soit n le nombre de lancers à réaliser.
Soit X la variable aléatoire égale au nombre de fois où le 6 est apparu au cours des n lancers. La fréquence d'apparition du 6 est donc $\frac{X}{n}$.
On cherche n de sorte que :

$$P\left(\left|\frac{X}{n} - \frac{1}{6}\right| \leq 0.01\right) \geq 0.95$$

On sait que X est une variable aléatoire de loi binomiale de paramètres n et $1/6$. Donc, $\mathbb{E}(X) = \frac{n}{6}$.
Or

$$P\left(\left|\frac{X}{n} - \frac{1}{6}\right| \leq 0.01\right) = P\left(\left|X - \frac{n}{6}\right| \leq n * 0.01\right) = P(|X - \mathbb{E}(X)| \leq n * 0.01)$$

Or, d'après l'inégalité de Bienaymé-Tchebychev, nous savons que $P(|X - \mathbb{E}(X)| \geq \epsilon) \leq \frac{\mathbb{V}(X)}{\epsilon^2}$. D'où :

$$
\begin{aligned}
P(|X - \mathbb{E}(X)| \leq n * 0.01) &\geq 1 - P(|X - \mathbb{E}(X)| > n * 0.01) \\
&\geq 1 - \frac{\mathbb{V}(X)}{(0.01n)^2}
\end{aligned}
$$

Il suffit par conséquent de prendre n de sorte que $1 - \frac{\mathbb{V}(X)}{(0.01n)^2} \geq 0.95$. Or $\mathbb{V}(X) = \frac{5n}{36}$. Donc, n est tel que :

$$\begin{aligned} 1 - \frac{5n}{36 * n^2 * 0.01^2} &\geq 0.95 \\ \frac{5n}{36 * n^2 * 0.01^2} &\leq 0.05 \\ \frac{50000}{36 * 0.05} &\leq n \\ \frac{1000000}{36} &\leq n \\ 27777,7 &\leq n \end{aligned}$$

Puisque n est un entier, cela signifie que l'on pourra porter une telle affirmation moyennant d'avoir réaliser 27778 lancers de dé!

Convergence en probabilité

Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variable aléatoire qui converge vers X. Cela signifie concrètement que X_n se "rapproche" de X lorsque n augmente, à savoir que $|X_n - X|$ devient d'autant plus petit. Cependant, X_n et X sont des variables aléatoires. Donc, il nous faut considérer l'événement $|X_n - X| \leq \epsilon$ et la convergence se traduira par le fait que cet événement sera réalisé avec une probabilité qui tend vers 1 lorsque n tend vers $+\infty$.

Définition 17.
On dit que (X_n) ***converge en probabilité*** *vers X, ce que l'on note $X_n \underset{p}{\longrightarrow} X$ si :*

$$\forall \epsilon > 0,\ P(|X_n - X| < \epsilon) \underset{n\to+\infty}{\longrightarrow} 1$$

ou de manière équivalente :

$$\forall \epsilon > 0,\ P(|X_n - X| > \epsilon) \underset{n\to+\infty}{\longrightarrow} 0$$

Exemple 15.
Soit (X_n) la suite de variables aléatoires définie par $\forall n \in \mathbb{N}^*$, $P(X_n = 0) = 1 - \frac{1}{n}$ et $P(X_n = n) = \frac{1}{n}$.
Montrons que cette suite converge en probabilité.

Soit $\epsilon > 0$ et soit $n > \epsilon$, on a :

$$P(|X_n| \geq \epsilon) = P(X_n = n) = \frac{1}{n} \underset{n\to+\infty}{\longrightarrow} 0$$

Donc $X_n \underset{p}{\longrightarrow} 0$!

Loi des Grands Nombres

Définition 18.
Soit (X_n) une suite de variables aléatoires mutuellement indépendantes telle que $\forall n \in \mathbb{N}$, $\mathbb{E}(X_n) = \mu$ et $\mathbb{V}(X) = \sigma^2$. Alors :

$$\bar{X_n} \underset{p}{\longrightarrow} \mu$$

où $\bar{X_n} = \frac{X_1 + \ldots + X_n}{n}$.

Remarque 24.
Ce résultat sera largement utilisé dans la partie de ce cours relative à l'estimation.

Convergence en loi

Définition 19.
Soit (X_n) une suite de variables aléatoires de fonction de répartition F_n. Soit X une variable aléatoire de fonction de répartition F.
*On dit que (X_n) **converge en loi** vers X si en tout point de continuité de F, noté x, on a $(F_n(x))$ converge vers $F(x)$.*
Ceci se note $X_n \underset{\mathcal{L}}{\longrightarrow} X$.

Remarque 25.
Si $X_n \underset{p}{\longrightarrow} X$, alors $X_n \underset{\mathcal{L}}{\longrightarrow} X$

Théorème de la Limite Centrale

La Loi des Grands Nombres énonce la convergence de $\bar{X_n}$ vers la moyenne théorique.

Théorème 1.
Soit (X_n) une suite de variables aléatoires indépendantes et identiquement distribuées telle que $\forall n \in \mathbb{N}$, $\mathbb{E}(X_n) = \mu$ et $\mathbb{V}(X) = \sigma^2$. Alors :

$$\sqrt{n}\frac{\bar{X_n} - \mu}{\sigma} \underset{\mathcal{L}}{\rightsquigarrow} \mathcal{N}(0; 1)$$

Remarque 26.
C'est du fait de ce théorème que la loi normale centrée réduite tient une telle place.

Application :
Soit X une variable aléatoire de loi binomiale de paramètres n et p.
On sait que $X = Y_1 + \ldots + Y_n$ avec Y_i des variables aléatoires de Bernoulli indépendantes de paramètre p.

Par application de Théorème de la Limite Centrale, on a :

$$\begin{aligned}
\sqrt{n}\frac{\frac{\sum_{i=1}^{n} Y_i}{n} - p}{\sqrt{pq}} &\overset{\mathcal{L}}{\rightsquigarrow} \mathcal{N}(0;1) \\
\sqrt{n}\frac{\frac{X}{n} - p}{\sqrt{pq}} &\overset{\mathcal{L}}{\rightsquigarrow} \mathcal{N}(0;1) \\
\frac{X - np}{\sqrt{npq}} &\overset{\mathcal{L}}{\rightsquigarrow} \mathcal{N}(0;1)
\end{aligned}$$

Autrement dit, on peut approcher une loi binomiale de paramètres n et p par une loi normale de paramètres np et npq.
Cette approximation sera valide du moment que $n \geq 30$, $np \geq 5$ et $nq \geq 5$.

Exemple 16.
On lance 400 fois une pièce de monnaie équilibrée et on veut calculer la probabilité que la fréquence d'apparition de pile soit comprise entre 0.45 et 0.55.

On pose :

$$X_i = \begin{cases} 1 & \text{si pile au lancer numéro } i \\ 0 & \text{si face au lancer numéro } i \end{cases}$$

On a $X_i \sim \mathcal{B}(1/2)$.
On pose $X = \sum_{i=1}^{400} X_i$, *ce qui n'est autre que la variable égale au nombre de fois où pile apparaît au cours des 400 lancers.*
On sait que $X \sim \mathcal{B}(400; 1/2)$, *que* $\mathbb{E}(X) = 200$ *et que* $\mathbb{V}(X) = 100$.
On veut calculer $P(0.45 \leq \frac{X}{400} \leq 0.55)$.

$$\begin{aligned}
P\left(0.45 \leq \frac{X}{400} \leq 0.55\right) &= P(0.45 * 400 \leq X \leq 0.55 * 400) \\
&= P(180 \leq X \leq 220) \\
&= P\left(\frac{180 - 200}{10} \leq \frac{X - 200}{10} \leq \frac{220 - 200}{10}\right) \\
&= P\left(\frac{-20}{10} \leq \frac{X - 200}{10} \leq \frac{20}{10}\right) \\
&= P\left(-2 \leq \frac{X - 200}{10} \leq 2\right)
\end{aligned}$$

Or, la variable $\frac{X-200}{10}$ *peut être approchée par une loi normale centrée réduite si l'on se réfère au théorème de la limite centrale car* $400 > 30$, $200 > 5$ *et* $200 > 5$. *Donc, considérons la variable* Z *de loi normale centrée réduite.*

$$
\begin{aligned}
P(-2 \leq Z \leq 2) &= P(Z \leq 2) - P(Z \leq -2) \\
&= P(Z \leq 2) - P(Z > 2) \\
&= P(Z \leq 2) - (1 - P(Z \leq 2) \\
&= 2 * P(Z \leq 2) - 1 \\
&= 2 * 0.9772 - 1 \\
&= 0.9544
\end{aligned}
$$

Chapter 2

Statistique descriptive

L'objectif de ce chapitre est comme son nom l'indique de pouvoir décrire "correctement" des jeux de données. Pour cela, nous allons explorer les différentes représentations graphiques possibles ainsi que les indicateurs statistiques caractéristiques. Bien entendu, nous ajouterons aux descriptions, non seulement une partie d'interprétation mais encore un volet plus théorique sur les inconvénients et avantages des divers indicateurs.

2.1 Vocabulaire

Avant d'aller plus en avant dans la suite de ce chapitre, nous allons commencer par définir un certain nombre de notions que nous allons rencontrer tout au long des deux chapitres à venir.

- population : il s'agit de l'ensemble sur lequel porte l'étude statistique qui est menée.
- échantillon : il s'agit de la partie de la population qui participe effectivement à l'étude menée.

  Exemple 17.
  On s'intéresse à la couleur des yeux des habitants de la région Provence-Alpes Côté d'Azur. On interroge 400 habitants de cette région.
  La population est ici l'ensemble des habitants de la région PACA, tandis que l'échantillon n'est que les 400 personnes interrogées.

- variable ou caractère : il s'agit tout simplement de l'objet de l'étude.

  Exemple 18.
  Sur l'exemple précédent, la variable ou caractère est la couleur des yeux.

- Individu ou élément statistique : Il s'agit d'un élément de l'échantillon.

  Remarque 27.
  Il faut faire attention que ce que l'on appelle individu en statistique n'est pas forcément une personne physique.

Exemple 19.
Sur l'exemple précédent, un individu est une personne interrogée.
Par contre, si l'on s'intéresse au chiffre d'affaire d'une entreprise française, cette fois-ci, un individu n'est autre qu'une entreprise française.

- nature d'une variable: Il s'agit de déterminer si la variable d'étude est de nature quantitative continue, quantitative discrète, qualitative catégorielle ou qualitative ordinale. Comment distinguer les 4 natures possibles :
  - Une variable est quantitative si l'on s'intéresse à quelque chose que l'on peut mesurer physiquement. Au contraire, une variable est qualitative si l'étude porte sur une variable que l'on ne peut pas mesurer ou si celle-ci est relative à un jugement.
  - Lorsqu'une variable est quantitative, on dira qu'elle est :
    * discrète si les valeurs possibles pour la variable sont en petit nombre, ou si les valeurs sont espacées (ce que l'on appelle encore discrète)
    * continue si les valeurs possibles sont des réels appartenant à un intervalle
  - Lorsqu'une variable est qualitative, on dira qu'elle est :
    * catégorielle si les modalités (nom donné aux valeurs dans le cadre des variables qualitatives) prises par la variable ne peuvent pas être hiérarchisées
    * ordinale si les modalités peuvent être ordonnées

Voici quelques exemples de variables pour chacune des natures possibles.

qualitative catégorielle		qualitative ordinale
catégorie socio-professionnelle des parents		mention au bac
lieu de résidence des parents (banlieu, campagne, ville)		jugement sur un vin (mauvais, passable, bon, excellent)

quantitative discrète		quantitative continue
nombre d'enfants dans une famille		taille d'une personne
nombre de succès à un jeu de pile ou face sur 10 lancers		poids d'une personne

2.2 Représentations graphiques

Lorsque l'on dispose de jeux de données, la façon la plus simple de synthétiser ces données est de les représenter visuellement à l'aide de graphiques. Cependant, en fonction de la nature de la variable considérée, il n'est pas possible de faire n'importe quelle représentation graphique. Tout l'objet de ce paragraphe est de définir quelles représentations sont adaptées aux différentes

natures de variable, et surtout de procéder correctement à la réalisation des graphiques.
Ce qu'il est important de retenir, c'est qu'avant de procéder à l'aspect graphique, il faut commencer par établir ce que l'on appelle un tableau de représentation qui contient les éléments permettant la réalisation graphique.

2.2.1 Variable qualitative catégorielle

Tableau de représentation

Lorsque l'on est en présence d'une variable qualitative catégorielle, le tableau de représentation comprend 3 colonnes comme le montre le tableau ci-dessous :

modalité	effectif	angle
a_1	n_1	α_1
a_2	n_2	α_2
$\vdots$	$\vdots$	$\vdots$
a_n	n_n	α_n

où :

- n en indice représente le nombre de modalités différentes
- a_1, a_2, ..., a_n sont les différentes modalités
- n_i est l'effectif associé à la modalité a_i, autrement dit le nombre de fois où la modalité a_i est apparue dans le jeu de données
- α_i est l'angle associé à la modalité a_i selon la formule suivante :

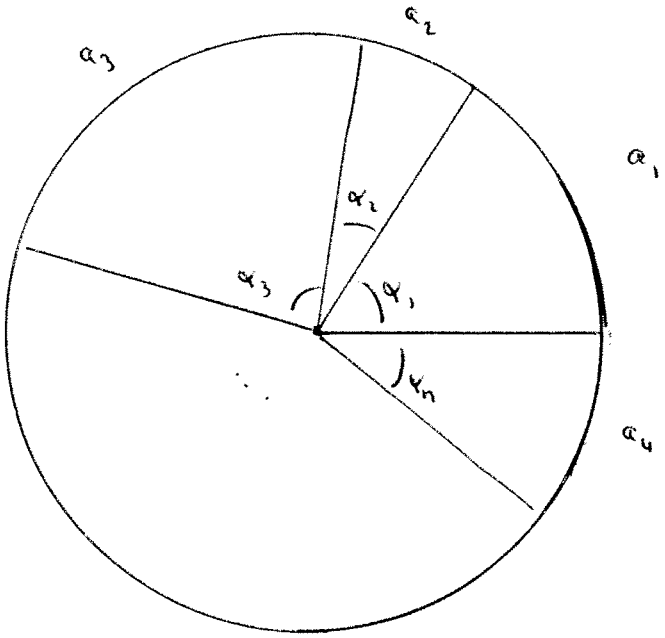
$$\alpha_i = 360\frac{n_i}{N}$$

avec N l'effectif total, soit encore $N = \sum_{i=1}^{n} n_i$.

Représentation graphique

Pour une telle variable, la représentation graphique adaptée est le diagramme circulaire, encore appelé camembert.
En voici une illustration.

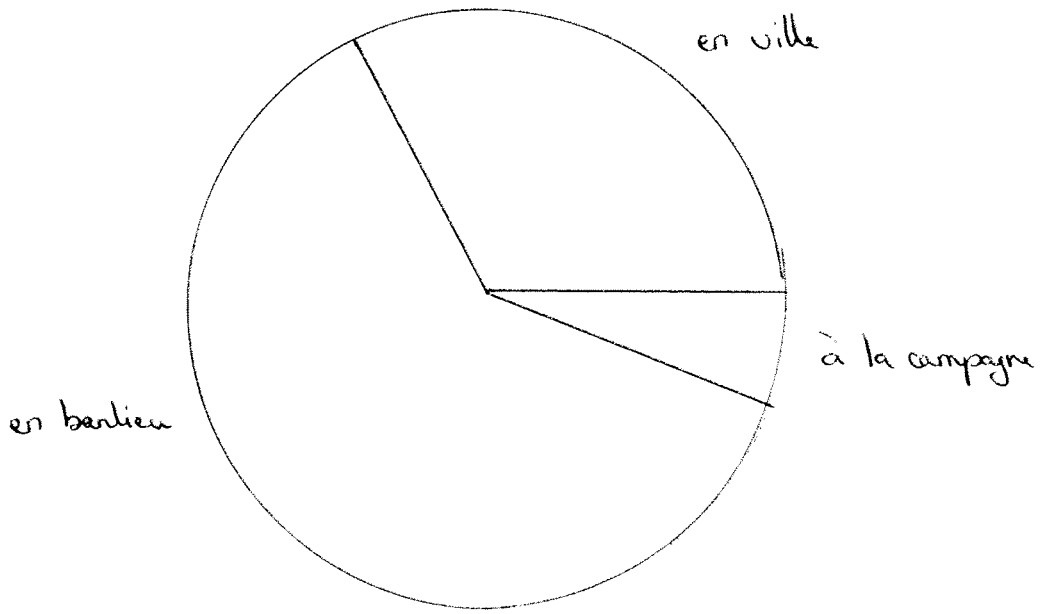
Représentation par camembert



Exemple 20.
Supposons qu'après dépouillement des données, on obtienne le tableau de représentation suivant :

modalité	*effectif*	*angle*
en ville	*80*	α_1
en banlieu	*150*	α_2
à la campagne	*20*	α_3

On a alors $N = 250$ *et donc* $\alpha_1 = 360\frac{80}{250} = 115.2$, $\alpha_2 = 360\frac{150}{250} = 216$ *et* $\alpha_3 = 360\frac{20}{250} = 28.8$
On obtient alors la représentation graphique suivante :



2.2.2 Variable qualitative ordinale

Tableau de représentation

Lorsque l'on est en présence d'une variable qualitative ordinale, le tableau de représentation comprend encore une fois 3 colonnes comme le montre le tableau ci-dessous. Cependant, la troisième colonne ne porte plus le même nom et les valeurs afférentes ne se calculent pas selon la même formule.

modalités	effectif	fréquence
a_1	n_1	f_1
a_2	n_2	f_2
$\vdots$	$\vdots$	$\vdots$
a_n	n_n	f_n

où :

- n en indice représente le nombre de modalités différentes
- a_1, a_2, …, a_n sont les différentes modalités classées selon la hiérarchie implicite de la plus petite à la plus grande

- n_i est l'effectif associé à la modalité a_i, autrement dit le nombre de fois où la modalité a_i est apparue dans le jeu de données
- f_i est la fréquence associé à la modalité a_i selon la formule suivante :

$$f_i = \frac{n_i}{N}$$

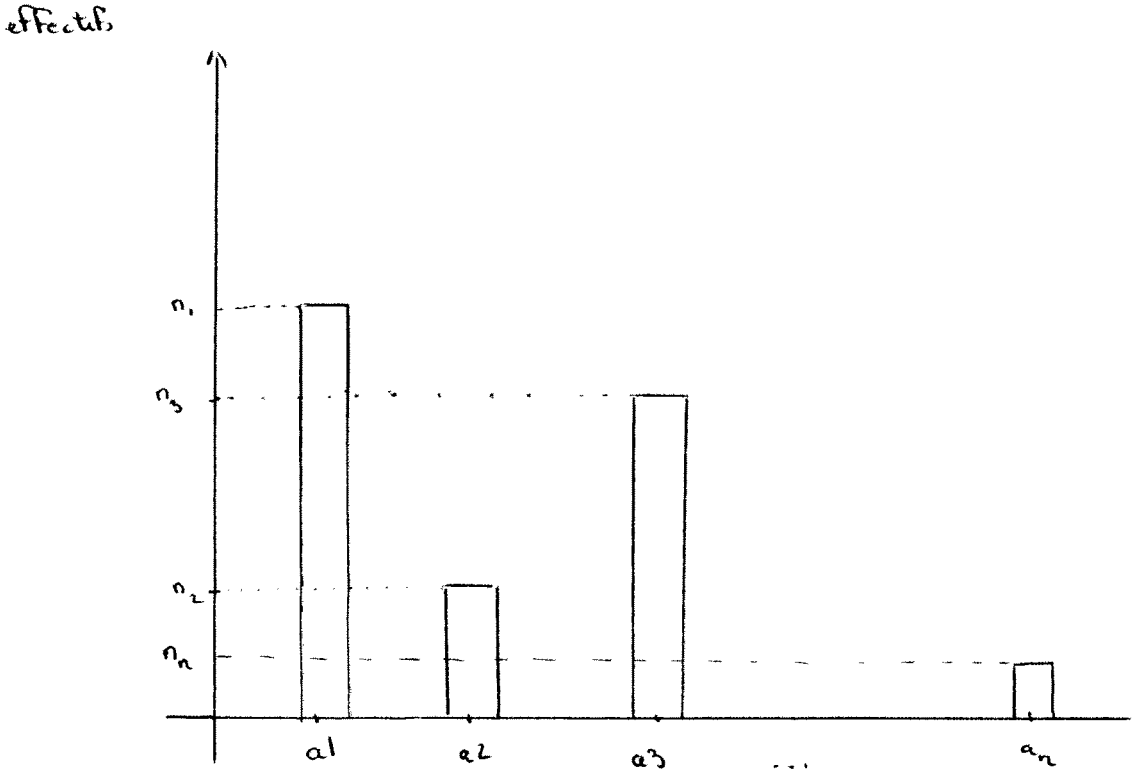
avec N l'effectif total, soit encore $N = \sum_{i=1}^{n} n_i$.

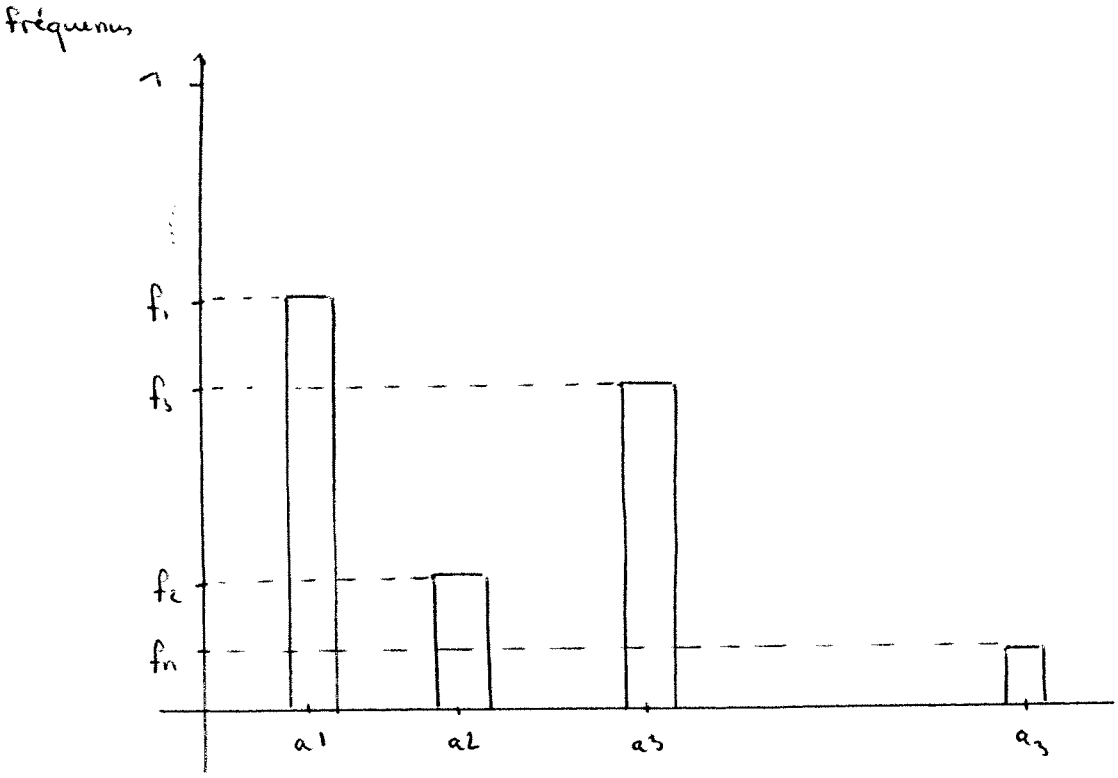
Remarque 28.

- *le fait de répertorier les modalités selon leur hiérarchie n'est pas quelque chose d'obligatoire, cependant cela est conventionnel.*
- $\sum_{i=1}^{n} f_i = 1$

Représentation graphique

La représentation graphique qui est adaptée à ce type de variable est le diagramme en bâtons. Deux représentations graphiques distinctes sont possibles, à savoir le diagramme en bâtons des effectifs ou le diagramme en bâtons des fréquences. Ci-dessous apparaît une représentation de chacune des deux possibilités.





La différence notable entre les deux illustrations est que pour ce qui est de la représentation des effectifs, la hauteur des bâtons n'est autre que les valeurs des effectifs, tandis que pour ce qui est de la représentation des fréquences, la hauteur des bâtons est égale aux différentes valeurs des fréquences.

Remarque 29.

- *Il est conventionnel que sur l'axe des abscisses du graphique, les modalités apparaissent selon l'ordre implicite. Cependant, bien faire attention à ne pas graduer l'axe des abscisses et à ne pas y mettre de flèche en extrémité, car l'ordre n'est qu'implicite et l'écart entre deux modalités n'est en rien significatif.*
- *Si les deux représentations graphiques (effectif et fréquence) sont possibles dans l'absolu, il y a une représentation graphique qui est privilégiée. Il s'agit de celle des fréquences, car elle permet notamment la comparaison aisée d'échantillons de taille différente. En effet, si la taille du second jeu de données est le double de celui du premier jeu de données, si les deux jeux de données sont relatifs à la même expérience et que les individus de l'échantillon ont été tirés au hasard, les effectifs associés au second jeu de données seront approximativement le double des effectifs du premier jeu. Autrement dit, sur la représentation en effectif, les bâtons associés au second jeu de données seront approximativement*

deux fois plus hauts que ceux du premier jeu. Par contre, en ce qui concerne la représentation en fréquence, les bâtons auront sensiblement la même hauteur pour les deux jeux de données. Donc, il est plus facile de conclure à la similarité sur la représentation en fréquence que sur celle en effectif.

Donc, la représentation graphique qui sera qualifiée d'adaptée dans le cas d'une variable qualitative ordinale sera le diagramme en bâtons des fréquences!

Remarque 30.
Il est possible dans la théorie d'inverser les deux représentations graphiques adaptées précédentes cependant, dans la pratique on ne le fera pas car on perd dans le diagramme circulaire la hiérarchie implicite et le diagramme en bâtons en crée une.

2.2.3 Variable quantitative discrète

Tableau de représentation

En ce qui concerne le tableau de représentation associé à une variable quantitative discrète, il comporte 4 colonnes. Par rapport au tableau de représentation d'une variable qualitative ordinale, il suffit d'ajouter une colonne dénommée fréquence cumulée croissante, très souvent réduite à fréquence cumulée ainsi que le montre le tableau suivant :

valeurs	effectif	fréquence	fréquence cumulée
a_1	n_1	f_1	F_1
a_2	n_2	f_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_n	n_n	f_n	F_n

où :

- n en indice représente le nombre de valeurs différentes
- a_1, a_2, ..., a_n sont les différentes valeurs classées par ordre croissant
- n_i est l'effectif associé à la valeur a_i, autrement dit le nombre de fois où la modalité a_i est apparue dans le jeu de données
- f_i est la fréquence associé à la modalité a_i selon la formule suivante :

  $$f_i = \frac{n_i}{N}$$

  avec N l'effectif total, soit encore $N = \sum_{i=1}^{n} n_i$
- F_i est la fréquence cumulée associée à la valeur a_i, à savoir la somme des fréquences associées à des valeurs inférieures ou égales à a_i, ce qui s'écrit mathématiquement parlant

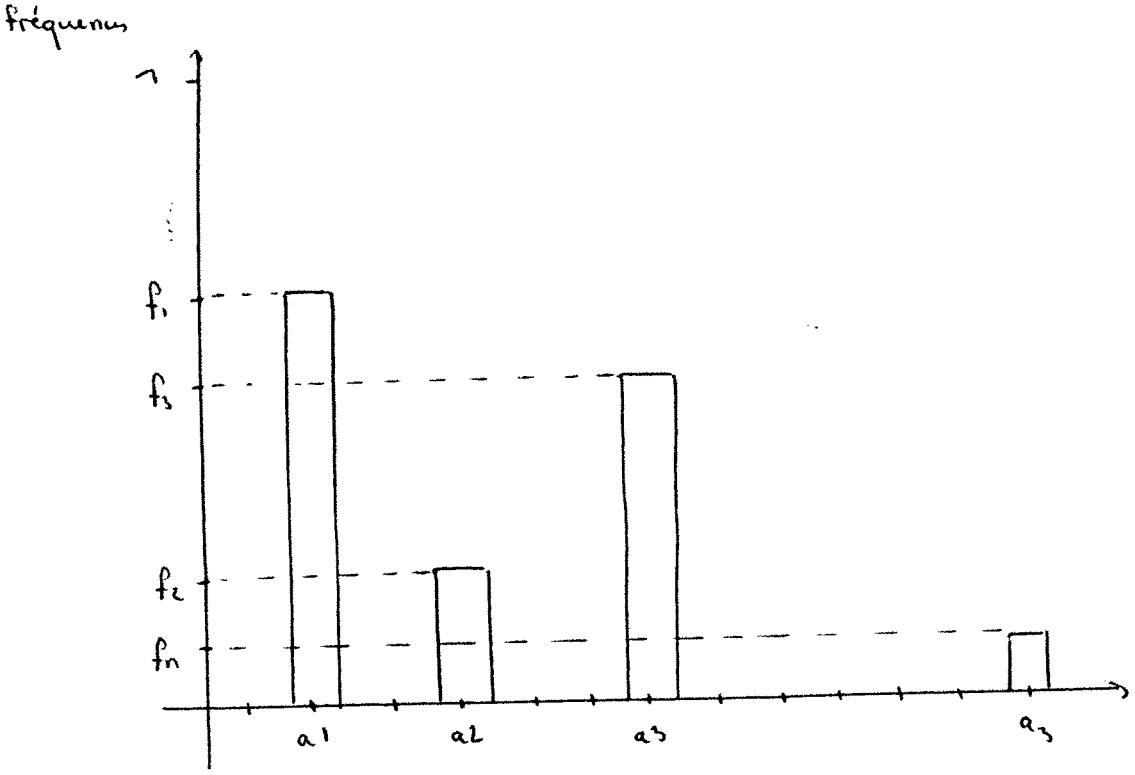
  $$F_i = \sum_{j \leq i} f_j$$

Remarque 31.
Pour une variable qualitative, la notion de fréquence cumulée n'a aucun sens puisque la notion d'inférieur ou égal n'a aucune raison d'être.

Représentation graphique

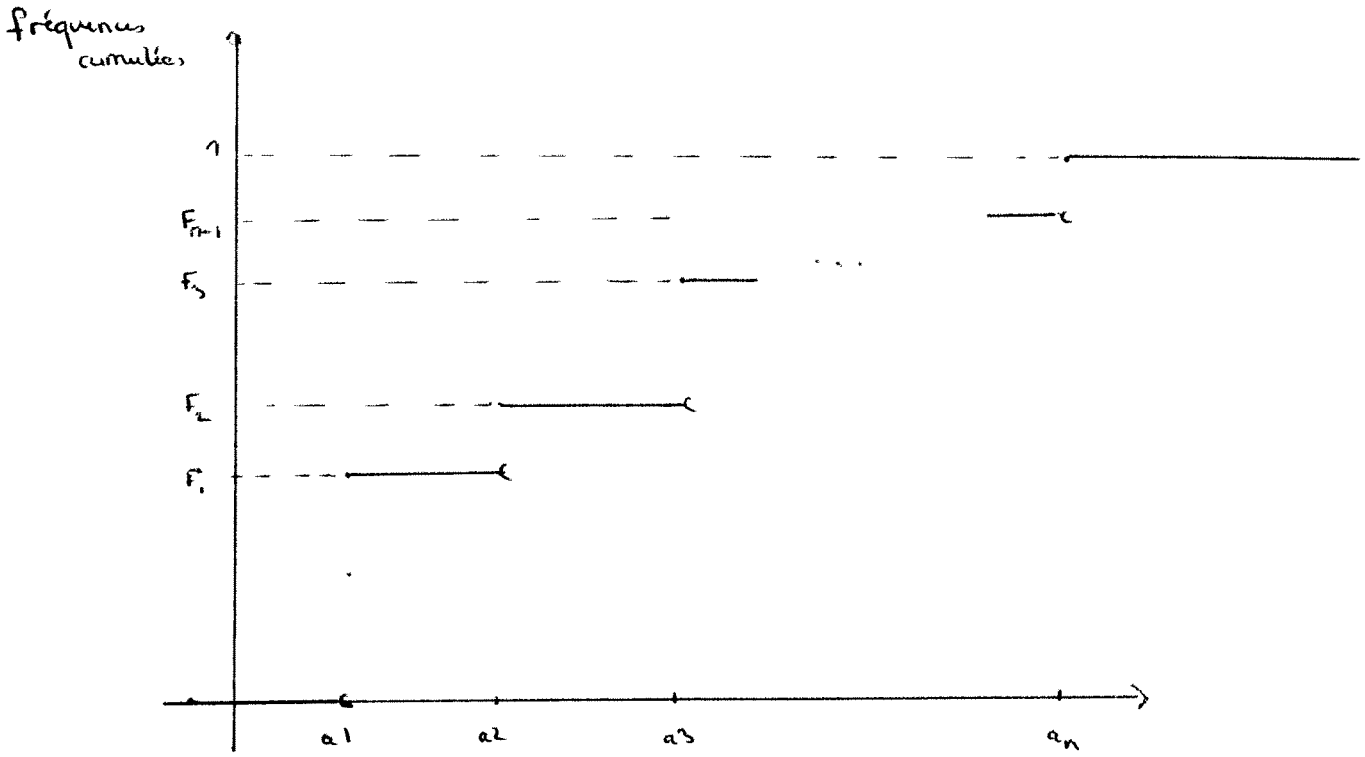
Dans le cadre d'une variable quantitative discrète, il existe deux représentations graphiques adaptées. La première est le diagramme en bâtons des fréquences et la seconde est la courbe des fréquences cumulées.
Voici une représentation de chacune d'elles.

diagramme en bâtons des fréquences



Remarque 32.
A la différence du cadre des variables qualitatives ordinales, ici l'axe des abscisses du graphique est gradué et porte une flèche en son extrémité. En effet, les valeurs étant numériques, cette fois-ci la place des valeurs a un sens et surtout l'écart entre deux valeurs est significatif.

Courbe des fréquences cumulées



La courbe des fréquences cumulées est une fonction en escalier qui est continue à droite, positive et qui aura pour valeur maximale 1.
Si l'on se rappelle du chapitre précédent relatif aux probabilités, cette fonction est à rapprocher d'une fonction de répartition. Le lien est le suivant : si l'on considère que les observations sont différentes réalisations d'une variable X, la courbe de fréquences cumulées est une approximation de la fonction de répartition. L'approximation sera d'autant meilleure que le nombre de données sera important. La courbe des fréquences cumulées est encore appelée fonction de répartition empirique. Le terme empirique signifie que cela a été obtenu à partir de données.

Si l'on veut aller un peu plus loin dans le lien entre les deux chapitres, on peut dire que les fréquences sont des approximations des probabilités associées à une variable aléatoire X discrète.

2.2.4 Variable quantitative continue

Tableau de représentation

Ce paragraphe est de loin celui qui est le plus compliqué, car très souvent, dans les années antérieures, certains rudiments vous en ont été donnés, mais ces derniers ne sont pas valables dans le cadre général des variables quantitatives continues. Donc, le mieux pour éviter les erreurs est d'oublier ce qui a pu vous être enseigné sur ce sujet et de repartir de zéro.

En ce qui concerne une variable quantitative continue, le tableau de représentation se compose de 6 colonnes.
Mais, avant de pouvoir vous montrer ce tableau, il faut commencer par s'attarder sur une des difficultés du cadre des variables de nature continue. En effet, il est impossible d'énumérer l'ensemble des valeurs possibles, car dans la théorie, il s'agit de tous les réels à l'intérieur d'un intervalle. Donc, il va être impossible de lister les différentes valeurs dans une première colonne, ainsi que cela a pu être fait dans toutes les situations précédentes. Donc, il va falloir constituer des classes, et ce sont ces classes qui vont servir de référence pour le tableau de représentation. Dans un premier temps, nous allons supposer que les différentes classes vous sont données. Nous verrons un petit peu plus loin une méthode pour les constituer.

Voici un tableau de représentation adapté à une variable de nature continue.

classe	center	effectif	fréquence	fréquence cumulée	hauteur
$[a_1; a_2[$	c_1	n_1	f_1	F_1	h_1
$[a_2; a_3[$	c_2	n_2	f_2	F_2	h_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[a_p; a_{p+1}[$	c_p	n_p	f_p	F_p	h_p

où :

- p est le nombre de classes ($p \leq n$)
- a_1, a_2, …, a_{p+1} sont différents réels croissants correspondant aux limites des classes
- c_i est le centre de la i-ème classe, à savoir $c_i = \frac{a_i + a_{i+1}}{2}$
- n_i est l'effectif associé à la i-ème classe, autrement dit le nombre de valeurs du jeu de données comprises entre a_i compris et a_{i+1} exclu
- f_i est la fréquence associé à la i-ème classe selon la formule suivante :

  $$f_i = \frac{n_i}{N}$$

  avec N l'effectif total, soit encore $N = \sum_{i=1}^{p} n_i$
- F_i est la fréquence cumulée associée à la i-ème classe, à savoir mathématiquement parlant :

  $$F_i = \sum_{j \leq i} f_j$$

- h_i est la hauteur associée à la i-ème classe et se calcule selon la formule :

$$h_i = \frac{f_i}{a_{i+1} - a_i}$$

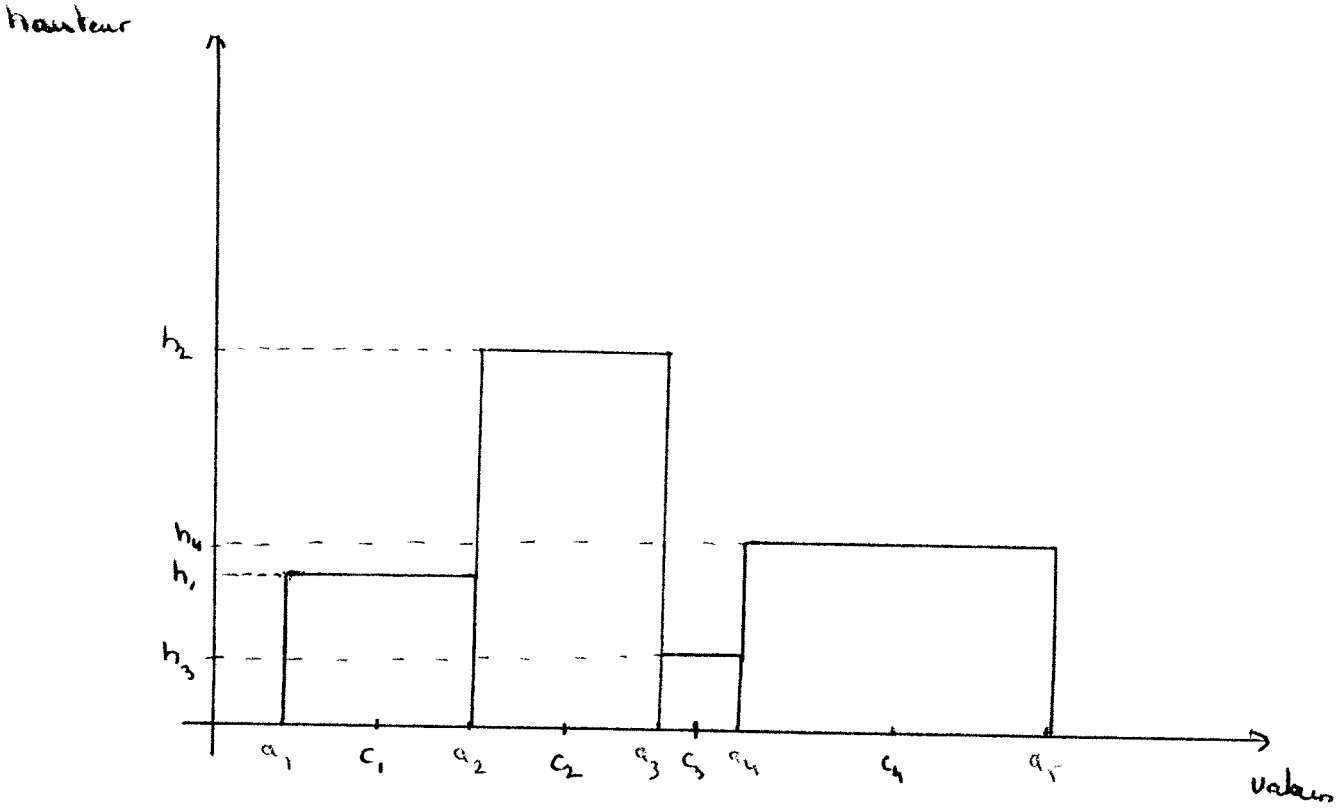
Remarque 33.

- *La fréquence cumulée associée à la i-ème classe, à savoir $[a_i; a_{i+1}[$ s'interprète de la façon suivante : il s'agit de la fréquence de l'événement "être inférieur strict à a_{i+1}". En effet, cette fréquence cumulée est obtenue une fois que l'on a balayé toutes les valeurs jusqu'à la valeur a_{i+1} exclue.*
- *la hauteur d'une classe n'est autre que la fréquence de la classe divisée par l'amplitude de cette dernière. Cette division sert de renormalisation et elle sera indispensable notamment dans le cas où les classes ne sont pas de même amplitude.*

représentation graphique

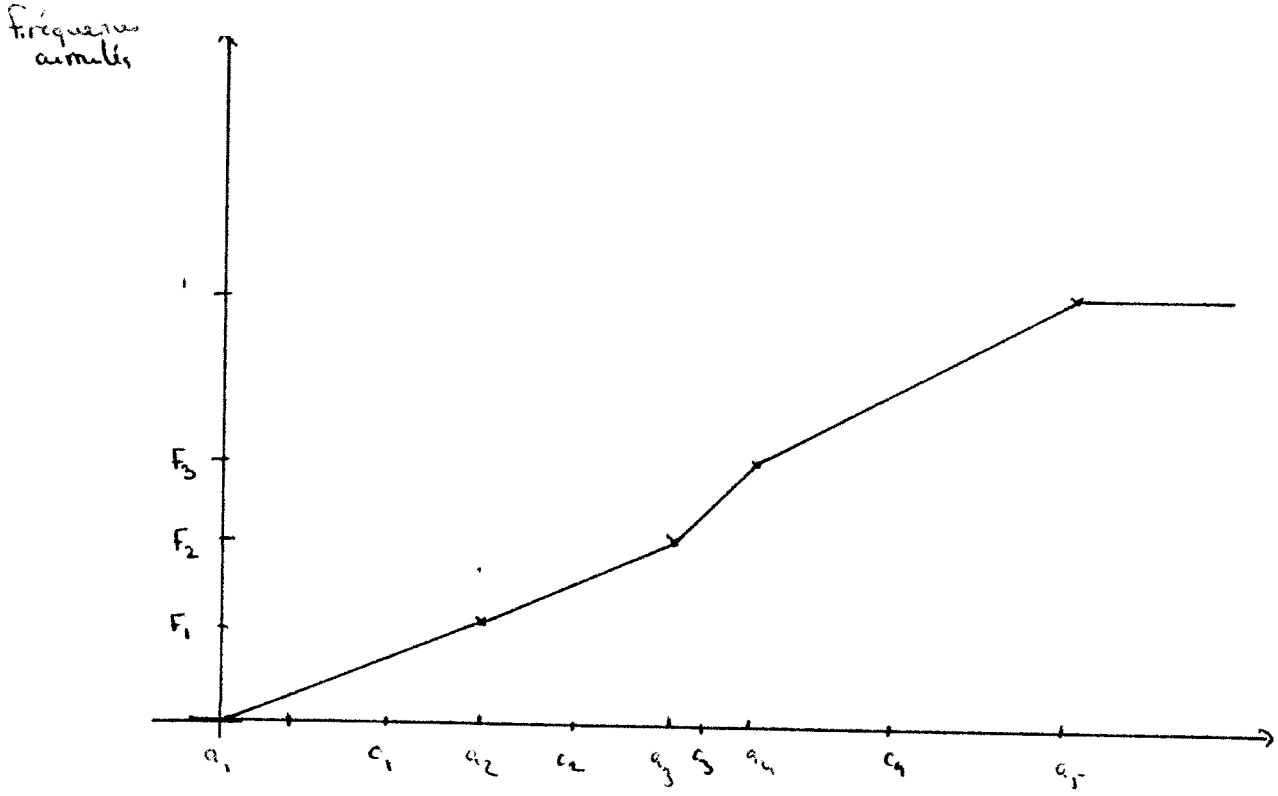
Tout comme dans le cas des variables aléatoires discrètes, il y a deux représentations graphiques adaptées. Comme précédemment, on retrouve la courbe des fréquences cumulées, mais par contre, en lieu et place du diagramme en bâtons, on trouve l'histogramme. Attention, c'est principalement sur la construction de l'histogramme que les erreurs apparaissent, car très souvent, de mauvaises représentations vous ont été enseignées dans le passé, mais pas tout le temps.

Histogramme



Remarque 34.
Ici, les bâtons sont bien évidemment collés puisqu'il existe une continuité entre les valeurs en abscisse, chose que l'on ne rencontrait pas dans le cadre précédent.
Par ailleurs, avec la normalisation en ordonnée, à savoir la division par l'amplitude de chacune des classes, ce que l'on a, c'est un histogramme qui a une aire égale à 1. Cette remarque est très importante, car lorsque l'on s'intéressera à l'aspect modélisation, on cherchera par exemple si l'on connaît une loi dont la fonction de densité se rapproche de l'allure de notre histogramme. Or, par définition, une fonction de densité est d'aire 1. D'où la caractère primordial de cette renormalisation. Par ailleurs, pour encore mieux vous en convaincre, une petit exemple viendra un peu plus loin.

Courbe des fréquences cumulées



Remarque 35.
Attention, la courbe des fréquences cumulées est ici une ligne brisée qu'il ne faut surtout pas chercher à lisser. Le fait que deux points soient joints par un segment de droite repose sur une hypothèse très forte quant à la répartition des données. En effet, on suppose qu'à l'intérieur d'une classe, les données sont réparties de façon homogène. On sait que dans la réalité, ce n'est pas le cas, mais cela demeure l'approximation usuelle.

En ce qui concerne le cadre des variables continues, une question fondamentale se pose. En effet, comment construire des classes qui soient adaptées à notre jeu de données?

Il n'existe pas une seule réponse possible, mais disons qu'il existe une technique largement usitée que je vais vous exposer à présent. Cette technique se décompose en deux grandes étapes :

1. Détermination du nombre de classes :
   Soit n le nombre de données dont on dispose.
   Soit k le nombre de classes à considérer.
   Il existe un lien entre la valeur de k et celle de n. Ce lien est donné par la formule de Sturge qui dit : $k \approx 1 + 3,22 * log_{10}(n)$.
   Le signe ne peut pas être celui = pour la simple et unique raison qu'en général, $1 + 3,22 * log_{10}(n)$ n'est pas un entier, alors que le nombre de classes doit en être un. Par ailleurs, bien faire attention que dans la formule, ce n'est pas le logarithme népérien qui apparaît, à savoir la fonction ln, mais bien le logarithme en base 10.

2. Constitution des classes :
   Maintenant que le nombre de classes a été déterminé, nous allons pouvoir déterminer les limites de chacune des classes. Voici la façon de procéder :

   - Soit m la plus petite valeur observée et M la plus grande.
   - Soit ε un "très petit" paramètre (afin de le définir, on peut par exemple prendre $\varepsilon = (M - m) * 10^{-3}$).
   - Soit $a = \frac{M-m+2\varepsilon}{k}$. Cette quantité a correspond à l'amplitude de chacune des classes, car par cette méthode on constitue des classes de même amplitude.
   - On pose alors $a_1 = m - \varepsilon$, puis pour $i \in \{2, \ldots, k+1\}$, on pose $a_i = a_1 + (i-1) * a$. Bien évidemment, on montre par le calcul que $a_{k+1} = M + \varepsilon$.

2.3 Indicateurs statistiques

A présent que nous savons synthétiser un jeu de données à l'aide de représentations graphiques, nous allons regarder comment procéder avec des quantificateurs numériques. Nous donnerons dans la suite de cette section non seulement les définitions, mais également les avantages et inconvénients de chacun ainsi que si possible l'interprétation qu'il est possible d'en faire.

Avant d'aller plus en avant dans cette partie du cours, il est à notre que dans le cadre des variables qualitatives, il n'existe qu'un seul indicateur statistique; il s'agit du **mode** qui, dans ce cas, a pour définition :

$$\text{mode} = \text{modalité associée au plus grand effectif.}$$

Remarque 36.
Attention, le mode n'est pas forcément unique!

Dans toute la suite de cette partie du cours, nous allons considérer une variable quantitative.

Les indicateurs statistiques sont classés en trois grandes familles. Il y a :

- les paramètres de **tendance centrale**
- les paramètres de **position**
- les paramètres de **dispersion**

2.3.1 les mesures de tendance centrale

Dans cette famille, on compte 3 mesures bien distinctes que sont la **moyenne**, le **mode** et la **médiane**.
L'idée est ici d'essayer de résumer l'ensemble de données à l'aide d'une seule grandeur.

Formules :

- le **mode** : on le note généralement M_o.
  - variable discrète : il s'agit de la valeur de la variable la plus souvent rencontrée, autrement la valeur associée à l'effectif le plus élevé.
  - variable continue : on ne parlera plus de mode mais de **classe modale** et par définition, la classe modale est la classe pour laquelle le bâton est le plus haut dans l'histogramme.

  Remarque 37.
  Attention, la classe modale n'est pas nécessairement la classe associée à l'effectif le plus élevé! En effet, cette seconde définition n'est vraie que pour des classes de même amplitude, mais devient totalement erronée dans le cadre de classes d'amplitude distincte!

- **la moyenne** : on la note en générale $\bar{x}$.
  - variable discrète : $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ ou $\bar{x} = \frac{\sum_{i=1}^{n} n_i a_i}{n}$, avec x_i toutes les réponses recueillies (ce sont donc les données brutes), n l'effectif total, a_i les différentes valeurs prises et n_i leur effectif associé.
  - variable continue : $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ sur données brutes et $\bar{x} = \frac{\sum_{i=1}^{k} c_i n_i}{n}$ sur données regroupées en classes, avec c_i le centre de la i-ème classe, n_i son effectif et k le nombre de classes.

- la **médiane** : on la note M_d.
  Par définition, c'est la valeur qui divise l'échantillon en deux sous-échantillons contenant chacun 50% des données. Cette grandeur satisfait $F(M_d) = 0.5$ avec F la fonction des fréquences cumulées.
  - variable discrète : La fonction des fréquences cumulées étant une fonction en escalier, elle ne prend pas forcément la valeur 0.5. Une façon de calculer la médiane est la suivante :
    1. On cherche dans le tableau de représentation la valeur de la fréquence cumulée qui est immédiatement ≥ 0.5.

2. la médiane est alors la valeurs associée à cette fréquence cumulée.

– variable continue : Comme ici la fonction des fréquences cumulées est strictement croissante entre la plus petite et la plus grande donnée et qu'elle vaut 0 en la plus petite donnée et 1 en la plus grande, on est assuré qu'il existe un antécédent par F à 0.5. Pour le trouver, voici la manière de procéder :

1. <u>on détermine la classe modale :</u>
   on repère dans la tableau de représentation la fréquence cumulée immédiatement ≥ 0.5. La classe modale est alors la classe associée à cette fréquence cumulée. Notons $[a, b[$ cette classe.
2. <u>on détermine la valeur de la médiane :</u>
   sur l'intervalle $[a, b[$, la fonction des fréquences cumulées est linéaire. Donc, pour trouver le point x de $[a, b[$ tel que $F(x) = 0.5$, on va utiliser l'interpolation linéaire.
   * on écrit l'équation de la droite passant par les points $(a, F(a))$ et $(b, F(b))$, à savoir $y = cx + d$.
   * on cherche l'expression de c et d en fonction de a, $F(a)$, b et $F(b)$. On a :

$$\begin{cases} F(a) = ca + d \\ F(b) = cb + d \end{cases}$$

   Soit :

$$\begin{cases} d = F(a) - ca \\ F(b) = cb + F(a) - ca \end{cases}$$

   Soit :

$$\begin{cases} d = F(a) - ca \\ c = \frac{F(b)-F(a)}{b-a} \end{cases}$$

   D'où :

$$\begin{cases} d = \frac{F(a)b-F(b)a}{b-a} \\ c = \frac{F(b)-F(a)}{b-a} \end{cases}$$

   * on cherche alors x tel que $0.5 = cx + d$.

$$\begin{aligned} x &= \frac{0.5 - d}{c} \\ &= \frac{0.5 - F(a) + ca}{c} \\ &= a + \frac{0.5 - F(a)}{c} \\ &= a + \frac{b - a}{F(b) - F(a)}(0.5 - F(a)) \end{aligned}$$

   * Ainsi, $M_d = a + \frac{b-a}{F(b)-F(a)}(0.5 - F(a))$.

<u>Avantages et Inconvénients :</u>

- à propos du mode :
  - valeur simple à calculer
  - ne dépend pas des valeurs extrêmes
  - valeur non nécessairement unique mais plus ils sont nombreux et moins ils ont d'importance
  - ne donne pas de renseignement sur l'ensemble des données
- à propos de la médiane :
  - ne dépend pas des valeurs extrêmes
  - valeur plus **représentative** que la moyenne si le jeu de données présente des valeurs extrêmes **et** que la distribution est asymétrique.
  - un grand écart entre la valeur de la médiane et celle de la moyenne peut indiquer la possibilité de présence de données extrêmes. Attention, la réciproque est fausse.
  - ne donne pas de renseignement sur l'ensemble des données
- à propos de la moyenne :
  - valeur qui tient compte de l'ensemble des données
  - valeur très dépendante des valeurs extrêmes
  - valeur la plus populaire des trois même si parfois elle n'est pas représentative

Remarque 38.
Dans le cas de données obtenues par simulation d'une loi normale (et s'il y a suffisamment de données), les trois valeurs sont identiques.

2.3.2 les mesures de position

Les mesures de position permettent de situer les données les unes par rapport aux autres.

Formules :

- les **quartiles** : ils sont au nombre de 3 et permettent de diviser les données en quatre parties de même effectif. On les note $Q_1,\ Q_2,\ Q_3$.
- les **déciles** : ils sont au nombre de 9 et permettent de diviser les données en dix parties de même effectif. On les note $D_1,\ D_2, \ldots, D_9$.
- les **centiles** : ils sont au nombre de 99 et permettent de diviser les données en cent parties de même effectif. On les note $C_1,\ C_2, \ldots, C_{99}$.

Plus généralement, on parle de quantiles. Le **quantile d'ordre** α (exprimé en pourcentage) est la valeur qui permet d'avoir au moins une proportion α des données sous la valeur du quantile. Pour la détermination du quantile d'ordre α, il convient de procéder à l'identique de la médiane en remplaçant 0.5 par α.

Ainsi, le premier quartile est le quantile d'ordre 25%, le troisième quartile est le quantile d'ordre 75%, le 7-ième décile est le quantile d'ordre 70% et le 47-ème centile est le quantile d'ordre 47%.

Remarque 39.
Vous pouvez constater que la médiane n'est autre que le second quartile!

Avantages et Inconvénients :

- valeurs non dépendantes des valeurs extrêmes
- l'examen des écarts entre les quartiles renseigne sur la concentration des données et sur leur symétrie
- seul un examen attentif de l'analyste permet de détecter les quantiles intéressants

2.3.3 les mesures de dispersion

Ces mesures visent à quantifier la tendance qu'on les données à s'étaler, à se disperser de part et d'autre d'une valeur centrale.

Formules :

- l'**étendue** : on la note en générale E.
  - variable discrète : il s'agit de l'écart entre la plus grande valeur et la plus petite valeur du jeu de données
  - variable continue : il s'agit de l'écart entre la plus grande limite de classe et la plus petite limite de classe
- la **variance** : on la note s^2.
  - variable discrète : $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$ ou $s^2 = \frac{\sum_{i=1}^{p} n_i(a_i - \bar{x})^2}{n-1}$
  - variable continue : $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$ ou $s^2 = \frac{\sum_{i=1}^{p} n_i(c_i - \bar{x})^2}{n-1}$ (cas des données regroupées en classe)
- l'**écart-type** : on le note s et ce n'est rien d'autre que la racine carrée de la variance.
- le **coefficient de variation** : il se note $C.V.$ et a pour définition $C.V. = \frac{s}{\bar{x}} * 100$
- l'**écart inter-quartile** : il se note Q et a pour définition $Q = Q_3 - Q_1$.

Avantages et Inconvénients :

- à propos du l'étendue :
  - valeur simple à calculer
  - dépend des valeurs extrêmes mais ne donne aucune information sur la dispersion des valeurs intermédiaires

- à propos de la variance et de l'écart-type :
  - mesure qui prend en compte toutes les données
  - il s'agit d'une mesure absolue car elle possède une unité qui est l'unité des données au carré pour la variance et la même unité que les données pour l'écart-type
  - il s'agit d'une mesure de dispersion vis à vis de la moyenne
  - mesure qui est difficilement interprétable si elle est prise seule
  - permet de comparer la dispersion des valeurs d'une même variable pour deux échantillons d'une même population

  Remarque 40.
  Une utilité de l'écart-type :
  Quelle que soit l'allure de la distribution, au moins $\left(1 - \frac{1}{k^2}\right) * 100\%$ *des données se trouvent dans l'intervalle* $[\bar{x} - ks, \bar{x} + ks]$.
  Ainsi, si $k = 2$, *au moins* 75% *des données se trouvent dans l'intervalle* $[\bar{x} - 2s, \bar{x} + 2s]$.
  Pour $k = 3$, *au moins* 88.8% *des données se trouvent dans l'intervalle* $[\bar{x} - 3s, \bar{x} + 3s]$.

- à propos du coefficient de variation :
  - il s'agit d'une mesure relative car elle ne possède pas d'unité
  - ne se calcule que dans certains cas (variable avec une échelle de rapport et que la valeur 0 signifie absence de)
  - qualifie l'homogénéité d'une distribution
  - permet de quantifier la fiabilité de la moyenne
- à propos de l'écart-interquartile :
  - joue un peu le rôle de l'écart-type lorsque la médiane est plus représentative que la moyenne
  - ne se calcule que dans certains cas (variable avec une échelle de rapport et que la valeur 0 signifie absence de)
  - qualifie l'homogénéité d'une distribution
  - permet de quantifier la fiabilité de la moyenne

2.3.4 quelques autres mesures

Il existe deux autres mesures qui sont couramment utilisées. Il s'agit du coefficient d'asymétrie d'une part et du coefficient d'applatissement ou kurtosis d'autre part. Ces deux mesures sont des paramètres de forme ainsi que nous allons le voir.

Visuellement, on peut qualifier une courbe de symétrique ou d'asymétrique. Plus l'asymétrie sera forte, et plus la médiane sera plus représentative que la moyenne.
Pour quantifier cette asymétrie, on utilise le coefficient d'asymétrie qui a pour définition :

$$C.D. = \left(\frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - \bar{x}}{s} \right)^3 \right) \quad \text{sur données brutes}$$

On dit alors que la courbe est :

- asymétrique à gauche si $C.D. < 0$
- symétrique si $C.D. = 0$
- asymétrique à droite si $C.D. > 0$

Remarque 41.
Cette expression du coefficient d'asymétrie est la version d'un estimateur non biaisé. Cette version est celle qui est notamment utilisée par Excel.

Une autre mesure de forme consiste à se comparer à la loi qui est très certainement la plus utilisée en statistique, à savoir la loi normale. Cette comparaison se fait via le coefficient d'applatissement encore dénommé kurtosis, dont la définition est :

$$C.A. = \left(\frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum \left(\frac{x_i - \bar{x}}{s} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)} \right) \quad \text{sur données brutes}$$

On dit alors que la courbe est :

- plus aplatie que la loi normale si $C.A. < 0$
- avec un étalement comparable à la loi normale si $C.A. = 0$
- plus pointue que la loi normale si $C.A. > 0$

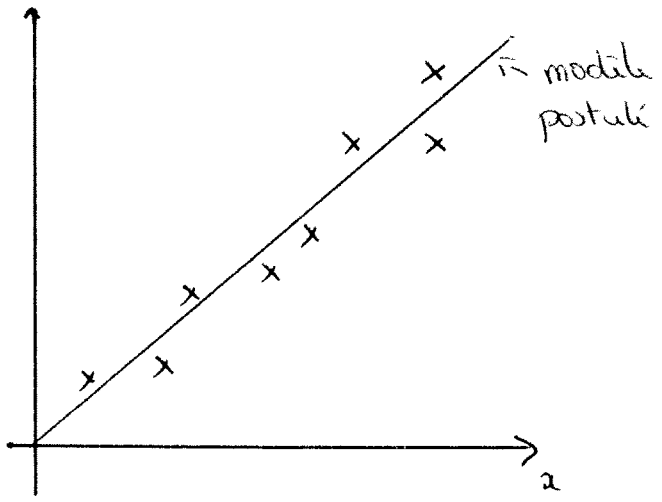
Remarque 42.
Cette expression du coefficient d'applatissement est la version d'un estimateur non biaisé. Cette version est celle qui est notamment utilisée par Excel.
Par ailleurs, ce coefficient ne sera à calculer que si auparavant, la courbe avait été déclarée symétrique car la densité d'une loi normale est quant à elle symétrique.

2.4 Régression linéaire simple

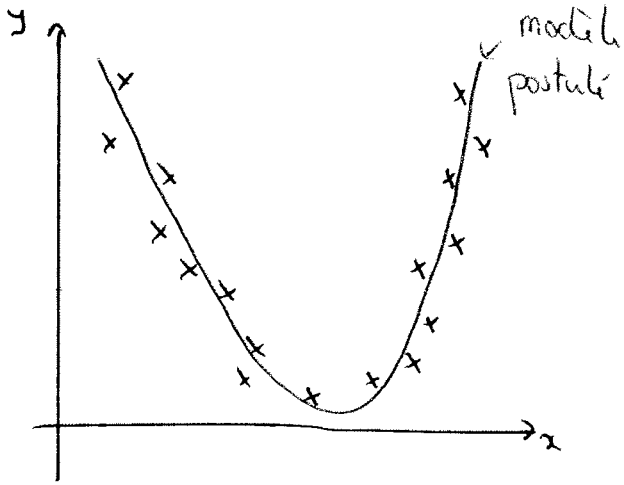
Dans ce qui précède, nous avons considéré une seule variable. Dans cette section, nous allons étudier le lien qu'il peut exister entre deux variables. Cependant, nous n'allons pas considérer toutes les dépendances possibles, mais uniquement celle qui a trait à la régression linéaire.

Considérons deux variables X et Y pour lesquelles nous cherchons à savoir s'il existe un lien "fonctionnel", à savoir s'il existe une fonction f telle que $Y = f(X)$.

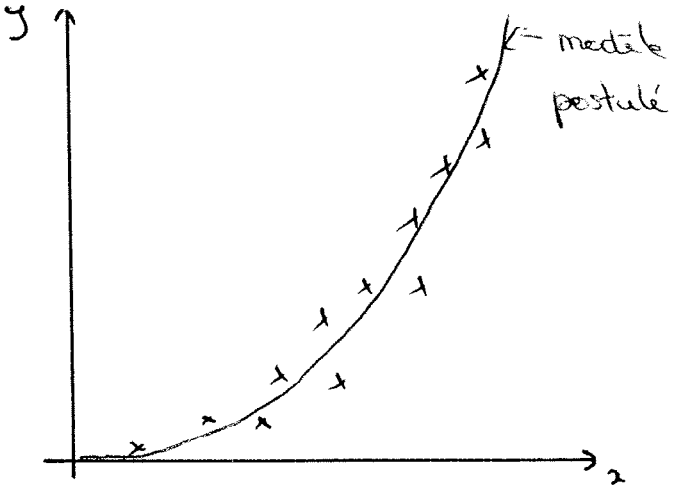
La première idée pour répondre à cette question consiste à tracer le nuage de points associé. Si l'on note n le nombre d'observations de chacune des variables, $\{x_i,\ i \in \{1, \ldots, n\}\}$ l'ensemble des observations de la variable X et $\{y_i,\ i \in \{1, \ldots, n\}\}$ l'ensemble des observations de la variable Y, le nuage de points consiste en la représentation graphique des différents couples (x_i, y_i).
Vous pouvez voir ci-dessous des illustrations de nuages de points.



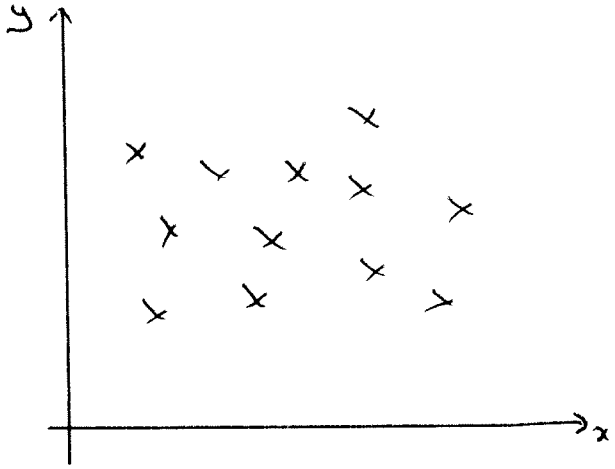
lien fonctionnel linéaire



lien fonctionnel quadratique



lien fonctionnel exponentiel



aucun lien fonctionnel

Ce premier examen purement graphique peut ne pas fournir pas de résultats concrets tout comme il peut déboucher sur plusieurs hypothèses. Cependant, une démarche purement mathématique devra toujours étayer les constatations visuelles. C'est ce que nous allons regarder dans la suite de cette section.

2.4.1 Recherche algébrique : cadre général

Le principe général de la recherche d'un lien fonctionnel est le le **principe des moindres carrés**. L'idée est de trouver, à partir des observations, une estimation fidèle de la fonction f, estimation que l'on notera $\hat{f}$.
Pour trouver cette estimation, on pose $\hat{Y} = \hat{f}(X)$ et on cherche $\hat{f}$ telle que la distance entre les vrais points observés y_i et ceux estimés, à savoir $\hat{y}_i = \hat{f}(x_i)$, soit la plus faible possible.
Mathématiquement, cela s'écrit :

$$\text{On cherche } \hat{f} \text{ telle que } \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \text{ soit minimale.}$$

La courbe de représentation de la fonction $\hat{f}$ est alors appelée **courbe de régression**.

2.4.2 Coefficient de détermination

Supposons à présent avoir déterminer la fonction $\hat{f}$. Nous aimerions pouvoir quantifier la qualité de cette ajustement de f par $\hat{f}$. Cette quantification va se faire par l'intermédiaire de ce que l'on appelle le **coefficient de détermination**.

L'idée est la suivante : Nous savons que les valeurs de la variable Y se disperse autour de leur moyenne $\bar{y}$ et nous savons quantifier cette dispersion grâce à la variance, à savoir

$$s_Y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}.$$

D'autre part, si le lien entre X et Y que l'on note $\hat{Y} = \hat{f}(X)$ existe vraiment, les valeurs de Y chercherons également à se disperser autour de la courbe. La question est alors de savoir dans quelle mesure la dispersion des valeurs prises par Y est-elle influencée par les valeurs de la variable X.
Pour répondre à cette question, il faut se reporter à ce que l'on appelle l'**équation de la variance** qui dit :

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\text{variance totale}} = \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\text{variance inexpliquée}} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{\text{variance expliquée par le lien entre X et Y}}$$

Preuve.

$$\begin{aligned}
\sum_{i=1}^{n}(y_i-\bar{y})^2 &= \sum_{i=1}^{n}(y_i-\hat{y}_i+\hat{y}_i\bar{y})^2 \\
&= \sum_{i=1}^{n}((y_i-\hat{y}_i)+(\hat{y}_i-\bar{y}))^2 \\
&= \sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+2*\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y}) \\
&= \sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2
\end{aligned}$$

En effet, on montre que $\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})=0$. ■

On définit le **coefficient de détermination**, noté r^2 par :

$$r^2=\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

Proposition 17.

- *r^2 est un réel compris entre 0 et 1*
- *r^2 indique la proportion de la variation totale de Y expliquée par son lien avec la variable X, via la courbe de régression*
  - *si r^2 est proche de 100%, les variations de Y sont très fortement expliquées par celles de X*
  - *si r^2 est proche de 75%, les variations de Y sont fortement expliquées par celles de X*
  - *si r^2 est proche de 50%, les variations de Y sont moyennement expliquées par celles de X*
  - *si r^2 est proche de 25%, les variations de Y sont très faiblement expliquées par celles de X*
  - *si r^2 est proche de 0%, les variations de Y sont nullement expliquées par celles de X*

2.4.3 Cas particulier de la régression linéaire simple

Equation de la droite de régression

Ce que l'on appelle la **régression linéaire simple** est le cas où la fonction $\hat{f}$ a pour équation un polynôme de degré 1, à savoir $\hat{f}(x)=a+bx$. Cela revient à dire que l'on cherche des estimations de Y sous la forme $\hat{Y}=a+bX$.

Proposition 18.
*L'**équation de la droite de régression** entre Y et X trouvée par la méthode des moindres carrés est donnée par :*

$$\hat{Y} = a + bX$$

*avec comme **pente** :*

$$b = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} (x_i)^2 - n(\bar{x})^2}$$

*et comme **ordonnée à l'origine** :*

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

Preuve.
Soit $\hat{Y} = a + bX$, l'équation qui passe le plus près de chacun des points.
Nous cherchons les valeurs de a et b pour lesquelles la quantité $S(a,b) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ sera minimale.
Calculons pour commencer les dérivées partielles de S.

$$\begin{aligned}
\frac{dS}{da}(a,b) &= \frac{d}{da}\left(\sum_{i=1}^{n}(y_i - (a+bx_i))^2\right) \\
&= \sum_{i=1}^{n}\frac{d}{da}\left((y_i - (a+bx_i))^2\right) \\
&= \sum_{i=1}^{n} 2(y_i - (a+bx_i).\frac{d}{da}(y_i - (a+bx_i)) \\
&= \sum_{i=1}^{n} 2(y_i - (a+bx_i).(-1) \\
&= -2\sum_{i=1}^{n}(y_i - (a+bx_i) \\
&= -2\sum_{i=1}^{n} y_i + 2\sum_{i=1}^{n} a + 2b\sum_{i=1}^{n} x_i \\
&= -2\sum_{i=1}^{n} y_i + 2na + 2b\sum_{i=1}^{n} x_i \\
&= -2n\bar{y} + 2na + 2bn\bar{x}
\end{aligned}$$

De même :

$$\begin{aligned}
\frac{dS}{db}(a,b) &= \frac{d}{db}\left(\sum_{i=1}^{n}(y_i-(a+bx_i))^2\right)\\
&= \sum_{i=1}^{n}\frac{d}{db}\left((y_i-(a+bx_i))^2\right)\\
&= \sum_{i=1}^{n}2(y_i-(a+bx_i).\frac{d}{db}\left(y_i-(a+bx_i)\right)\\
&= \sum_{i=1}^{n}2(y_i-(a+bx_i).(-x_i)\\
&= -2\sum_{i=1}^{n}y_ix_i+2a\sum_{i=1}^{n}x_i+2b\sum_{i=1}^{n}(x_i)^2\\
&= -2\sum_{i=1}^{n}y_ix_i+2na\bar{x}+2b\sum_{i=1}^{n}(x_i)^2
\end{aligned}$$

A présent, nous allons annuler simultanément ces deux dérivées partielles, à savoir résoudre :

$$\begin{cases}\frac{dS}{da}(a,b)=0\\ \frac{dS}{db}(a,b)=0\end{cases}$$

Soit :

$$\begin{cases}-2n\bar{y}+2na+2bn\bar{x}=0\\ -2\sum_{i=1}^{n}y_ix_i+2na\bar{x}+2b\sum_{i=1}^{n}(x_i)^2=0\end{cases}$$

$$\begin{cases}\bar{y}=a+b\bar{x}\\ \sum_{i=1}^{n}y_ix_i=na\bar{x}+b\sum_{i=1}^{n}(x_i)^2\end{cases}$$

$$\begin{cases}a=\bar{y}-b\bar{x}\\ \sum_{i=1}^{n}y_ix_i=n(\bar{y}-b\bar{x})\bar{x}+b\sum_{i=1}^{n}(x_i)^2\end{cases}$$

$$\begin{cases}a=\bar{y}-b\bar{x}\\ \sum_{i=1}^{n}y_ix_i=n\bar{y}\bar{x}+b\left(\sum_{i=1}^{n}(x_i)^2-n(\bar{x})^2\right)\end{cases}$$

$$\begin{cases}a=\bar{y}-b\bar{x}\\ b=\frac{\sum_{i=1}^{n}y_ix_i-n\bar{y}\bar{x}}{\sum_{i=1}^{n}(x_i)^2-n(\bar{x}))^2}\end{cases}$$

■

Remarque 43.
On peut montrer par le calcul que :

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

et que :

$$\sum_{i=1}^{n}((x_i)^2 - n(\bar{x}))^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Ceci conduit à plusieurs expressions possibles pour le coefficient de pente, en l'occurrence ici b.

Proposition 19.
La droite de régression passe toujours par le point $(\bar{x}, \bar{y})$.

Coefficient de corrélation

Dans le cadre général, nous avons vu que le coefficient de détermination permettait de quantifier l'intensité du lien entre les variables X et Y à travers la modélisation par $\hat{f}$. Mais, un des inconvénients de ce coefficient de détermination est qu'il faut avoir entièrement déterminer la fonction $\hat{f}$ pour pouvoir évaluer ce dernier.
Dans le cadre du modèle de régression linéaire simple, il existe une seconde façon d'aborder la problématique de l'intensité du lien entre X et Y. Cette seconde méthode implique le **coefficient de corrélation**.

Définition 20.
Le ***coefficient de corrélation*** *a pour expression :*

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Proposition 20.

- $-1 \leq r \leq 1$
- $(\textit{coefficient de corrélation})^2 = \textit{coefficient de détermination}$

Preuve.
La preuve de ces résultats sera faite lors des TD. ■

Remarque 44.
Le calcul du coefficient de corrélation n'a de sens que si le modèle envisagé est un modèle linéaire.

Interprétation :

- Si le coefficient de corrélation est proche de 1 en **valeur absolue**, alors le modèle linéaire sera une bonne modélisation. Et si en valeur absolue, le coefficient de corrélation vaut exactement 1, cela signifie que la modélisation est parfaite.

- si par contre, le coefficient de corrélation vaut 0, ou est proche de 0, cela signifie qu'il n'existe pas de lien linéaire entre les deux variables considérées. Cela ne veut pas dire pour autant qu'il n'y a pas de lien. Ce lien peut être, par exemple, quadratique et cela ne sera pas détecté par le coefficient de corrélation.

Remarque 45.
Le coefficient de détermination nécessité la détermination totale de $\hat{f}$ pour pouvoir être calculé. Le calcul du coefficient de corrélation ne nécessite quant à lui que les données initiales. C'est la raison pour laquelle, la première modélisation tentée est celle linéaire.
Par ailleurs, on parle de corrélation positive lorsque $r > 0$, ce qui signifie que les deux variables évoluent dans le même sens. On parle de corrélation négative lorsque $r < 0$, ce qui signifie que les deux variables évoluent en sens inverse (quand l'un croît l'autre décroît et réciproquement).

Chapter 3

Statistique inférentielle

3.1 Estimation poncutelle

3.1.1 Introduction

Une manière d'introduire l'estimation est la suivante.
Jusqu'à présent, nous avons travaillé sur des données, sans nous soucier du modèle dont elles étaient issues.
Or, un des soucis du statisticien est de mettre en correspondance des observations et un modèle probabiliste.
Donc, le problème qui nous intéresse pourrait s'énoncer de la manière suivante : à partir d'observations $x_1,\ x_2, \ldots, x_n$ d'une certaine variable aléatoire X (associée au phénomène étudié), obtenus grâce à des expériences aléatoires identiques et indépendantes, quelle loi théorique P, inconnue, peut-on adopter comme loi parente? Une autre façon de dire les choses est : quelle est la loi de probabilité associée à ce phénomène de production de données? Ce problème peut alors se schématiser ainsi :

$$(x_1,\ x_2, \ldots,\ x_n) \longrightarrow P?$$

Si le choix de P devait s'opérer parmi l'ensemble des lois de probabilité existantes (ensemble notée $\mathcal{P}$), le problème serait alors difficile à résoudre et surtout, il nécessiterait un très grand nombre d'observations n. On s'attaquerait à un problème qualifié de "non-paramétrique". Mais, compte tenu d'informations a priori dont dispose le statisticien, le choix de la loi P va se restreindre à une famille donnée $(P_\theta,\ \theta \in \Theta)$ indexée par un indice θ parcourant un ensemble Θ bien déterminé.

Exemple 21.

- $\{P_\theta,\ \theta \in \Theta\} = \{\mathcal{E}(\lambda),\ \lambda > 0\}, \Rightarrow \Theta = \mathbb{R}_*^+$
- $\{P_\theta,\ \theta \in \Theta\} = \{\mathcal{N}(\mu, \sigma^2),\ \mu \in \mathbb{R},\ \sigma^2 > 0\}, \Rightarrow \theta = (\mu, \sigma^2)\ et\ \Theta = \mathbb{R} * \mathbb{R}_*^+$

Par conséquent, la loi P est entièrement déterminée par la donnée de θ qui sera appelé paramètre de la distribution. Nous rentrons alors dans le domaine de la statistique paramétrique.

Afin que le problème soit parfaitement identifiable, nous supposons que pour $\theta \neq \theta'$, on a $P_\theta \neq P_{\theta'}$.
Mais alors, le problème considéré peut se reformuler ainsi :

$$(x_1, x_2, \ldots, x_n) \quad \longrightarrow \quad \theta \ ?$$

Ceci est ce que l'on appelle un problème d'estimation!
Principalement deux situations sont à distinguer :

- on choisit une seule valeur pour le paramètre $\theta \Longrightarrow$ **estimation ponctuelle**
- on choisit un sous-ensemble de Θ comme valeurs acceptables de $\theta \Longrightarrow$ **estimation par intervalle de confiance**

Ces deux problèmes sont résolus par la donnée d'une application $T : E^n \longrightarrow F$ qui associera, à un n-échantillon $(X_1, X_2, \ldots, X_n)$, une ou plusieurs variable(s) aléatoire(s) à valeur(s) numérique(s).

Remarque 46.
$E = X(\Omega)$ et $F \subset \mathbb{R}$ ou $F \subset \mathbb{R}^k$

3.1.2 Définition d'un estimateur

Afin de définir la notion d'estimateur, nous allons regarder ce qui se passe sur un exemple.
Cet exemple est issu des sondages, qui régulièrement, cherchent à connaître l'opinion publique vis à vis du président de la République.
Comme ensemble Ω, nous allons prendre l'ensemble des électeurs français. Nous notons A, le sous-ensemble des électeurs favorables au président. Nous considérons la variable aléatoire X définie par :

$$\begin{array}{rccl} X: & \Omega & \longrightarrow & \mathbb{R} \\ & \omega & \longrightarrow & 1 \text{ si } \omega \in A \\ & & & 0 \text{ si } \omega \notin A \end{array}$$

Remarque 47.
Ici ω est un électeur français.

La loi de probabilité de X est connue puisque'il s'agit d'une loi de Bernoulli $\mathcal{B}(p)$. Le paramètre θ qui définit ici entièrement cette loi est $\theta = p = P(X = 1) = P(A)$. Ainsi, dans cet exemple, la famille de lois retenue est $\{\mathcal{B}(p),\ p \in [0; 1]\}$. $(\Theta = [0; 1])$

Pour avoir une idée de la vraie valeur de ce paramètre, on interroge n électeurs tirés au hasard dans Ω et on associe à chacun de ces électeurs une variable de Bernoulli X_i, de même loi que X.
Si on constate sur le sondage que 48% des personnes interrogées sont favorables au président, on en déduira qu'environ un français sur deux est favorable au président.
En langage statistique, on dit que l'on estime p par la valeur 48%. Cette valeur, calculée sur la base de n observations, est donc une **estimation** de p, obtenue à partir de la fréquence empirique $f_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, fréquence qui sera appelée **estimateur** de p.

Remarque 48.

$\frac{1}{n}\sum_{i=1}^{n} X_i$*: estimateur (il s'agit d'une variable aléatoire)*

$\frac{1}{n}\sum_{i=1}^{n} x_i$*: estimation (il s'agit d'une réalisation d'une variable aléatoire, à savoir une valeur numérique)*

Considérons à présent ce second exemple.
Supposons qu'avant d'avoir à choisir un véhicule automobile, on se fixe un critère de choix basé sur N, le nombre moyen de pannes par an que l'on est susceptible de rencontrer avec un modèle donné.
On mène une étude chez un concessionnaire, à savoir, on prélève n dossiers de véhicules au hasard et on note pour chacun d'eux le nombre N_i de pannes subies la dernière année de mise en circulation. La loi de Poisson étant adaptée pour modéliser le nombre de pannes, la famille de lois retenue est $\{\mathcal{P}(\theta),\ \theta \in \mathbb{R}_+\}$. Ici $\theta = \mathbb{E}[N]$.
On estime θ par la moyenne des valeurs observées sur l'échantillon :

$$\bar{N_n} = \frac{1}{n}\sum_{i=1}^{n} N_i$$

Dans ces deux exemples, on a donc construit un modèle statistique où la variable X suit une loi P_θ. Pour se faire une idée de la valeur inconnue de θ, qui permet la détermination totale de la loi, on utilise un échantillon de cette loi. A partir des observations $x_1, x_2, \ldots, x_n$, on calcule alors une certaine valeur numérique qui sera considérée comme valeur approchée de θ, et qui sera appelée estimation de θ. La règle qui permet d'effectuer ce calcul est un estimateur.

Définition 21.
Un ***estimateur*** *de θ est une application T_n de E^n dans F qui, à un échantillon $(X_1, X_2, \ldots, X_n)$ de la loi P_θ, associe une variable aléatoire dont on peut déterminer la loi.*

Remarque 49.
La loi de la variable $T_n(X_1, X_2, \ldots, X_n)$ dépend de celle de X et donc de θ.
Chaque réalisation $T_n(x_1, x_2, \ldots, x_n)$ est une estimation de θ.
Cette définition est générale, et elle ne donne pas de méthode de construction. Cependant, comme les deux exemples précédents l'ont montré, l'expression de l'estimateur découle très naturellement de l'interprétation que l'on peut donner du paramètre θ.

Dans les deux exemples, nous avons eu $\theta = \mathbb{E}[X]$, c'est à dire la moyenne théorique de la loi. Comme estimateur de ce paramètre, nous retenons très logiquement la moyenne empirique, à savoir celle de l'échantillon :

$$T_n(X_1, X_2, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Cependant, dans certains cas, il est impossible de donner une interprétation du paramètre. On est alors amené à utiliser une méthode de construction. Nous allons en détailler 2 :

- ***méthode des moments** : généralisation de celle intuitive*
- ***méthode du maximum de vraisemblance***

3.1.3 Méthode de construction

Méthode des moments

Dans le cas où le paramètre à estimer est $\theta = \mathbb{E}[X]$, nous avons vu que l'estimateur naturel est la moyenne empirique $\bar{X}_n$.
De même, pour estimer un paramètre $\theta = \mathbb{V}[X]$, un estimateur naturel est la variance empirique, à savoir

$$(S'_n)^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

avec $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Soit $m_k = \mathbb{E}[X^k]$ le moment d'ordre k et soit $\mu_k = \mathbb{E}[(X - \mathbb{E}[X])^k]$ le moment centré d'ordre k. Plus généralement, on cherche une valeur de k pour laquelle m_k ou μ_k soit réellement une fonction du paramètre à estimer θ. Un estimateur de θ, noté $\hat{\theta}_n$, est donné par la résolution de l'équation en θ :

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} X_i^k &= m_k \quad \text{si} \quad m_k \text{ est une fonction de } \theta \\ \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^k &= \mu_k \quad \text{si} \quad \mu_k \text{ est une fonction de } \theta \end{aligned}$$

Exemple 22.
Soit $X \sim \mathcal{E}(\theta)$.
On soit que $\mathbb{E}[X] = \frac{1}{\theta}$. Ainsi, m_1 est une fonction de θ. Un estimateur $\hat{\theta}_n$ de θ est donné par la résolution de :

$$\frac{1}{\hat{\theta}_n} = \bar{X}_n$$

Soit $\hat{\theta}_n = \frac{1}{\bar{X}_n}$.

Remarque 50.
On aurait aussi pu dire $V[X] = \frac{1}{\theta^2}$.
On aurait alors résolu :

$$\frac{1}{(\hat{\theta}_n)^2} = (S'_n)^2$$

Soit $\hat{\theta}_n = \frac{1}{\sqrt{(S'_n)^2}} = \frac{1}{S'_n}$.

Remarque 51.
Cette méthode se justifie par les propriétés de convergence des moments empiriques vers les moments théoriques associés.

Méthode du Maximum de Vraisemblance

La **vraisemblance** $L(x_1, x_2, \ldots, x_n; \theta)$ représente la probabilité d'observer le n-uplet $(x_1, x_2, \ldots, x_n)$ pour une valeur fixée de θ.
Dans la situation inverse où ici on a observé $(x_1, x_2, \ldots, x_n)$ sans connaître la valeur de θ, on va attribuer à θ la valeur qui semble la plus vraisemblable compte tenu de nos observations. Autrement dit, nous allons attribuer à θ la valeur pour laquelle la vraisemblance est la plus forte.
La règle est la suivante :

- à $(x_1, x_2, \ldots, x_n)$ fixé, on considère la vraisemblance L comme une fonction de θ,
- on approche θ par $\hat{\theta}_n$ la valeur qui maximise L.

On en déduit la définition suivante :

Définition 22.
On appelle ***estimateur du maximum de vraisemblance****, toute fonction* $\hat{\theta}_n$ *de* $(X_1, X_2, \ldots, X_n)$ *qui vérifie :*

$$L(x_1, x_2, \ldots, x_n; \hat{\theta}_n) = \max_{\theta \in \Theta} L(x_1, x_2, \ldots, x_n; \theta)$$

Remarque 52.

- *Cette définition ne renseigne en aucune façon ni sur l'existence, ni sur l'unicité de cet estimateur!*
- *la recherche de l'estimateur du maximum de vraisemblance peut se faire directement par la recherche du maximum de* L *dans le cas où* L *est 2 fois dérivable par rapport à* θ*, en résolvant* $\frac{\partial L}{\partial \theta}(x_1, x_2, \ldots, x_n; \hat{\theta}_n) = 0$ *avec* $\frac{\partial^2 L}{\partial \theta^2}(x_1, x_2, \ldots, x_n; \hat{\theta}_n) < 0$.
  Cependant, la vraisemblance comme nous allons pouvoir le voir, étant un produit, il est préférable de résoudre le problème suivant :
  $\frac{\partial ln(L)}{\partial \theta}(x_1, x_2, \ldots, x_n; \hat{\theta}_n) = 0$ *avec* $\frac{\partial^2 ln(L)}{\partial \theta^2}(x_1, x_2, \ldots, x_n; \hat{\theta}_n) < 0$.

Définition 23.
Soit $(X_1, X_2, \ldots, X_n)$ un n-échantillon dont la loi dépend d'un paramètre θ. La vraisemblance de $(X_1, X_2, \ldots, X_n)$, notée $L(X_1, X_2, \ldots, X_n; \theta)$, est la loi de probabilité de ce n-uplet. Cette dernière est définie par, $\forall (x_1, x_2, \ldots, x_n)$:

$$L(x_1, x_2, \ldots, x_n; \theta) = \begin{cases} \prod_{i=1}^{n} P(X_i = x_i|\theta) & \text{si } X \text{ est une variable discrète} \\ \prod_{i=1}^{n} f_\theta(x_i) & \text{si } X \text{ est une variable continue de fonction de densité } f_\theta \end{cases}$$

Exemple 23.
Soit $(X_1, X_2, \ldots, X_n)$ un n-échantillon de loi $\mathcal{E}(\theta)$ avec $\theta > 0$.

1. *Déterminer un estimateur $\hat{\theta}_1$ de θ par la méthode des moments.*
2. *Déterminer un estimateur $\hat{\theta}_2$ de θ par la méthode du maximum de vraisemblance.*

Par la méthode des moments :
Soit X une variable aléatoire de loi $\mathcal{E}(\theta)$. Alors, on sait que $\mathbb{E}(X) = \frac{1}{\theta}$ et $\mathbb{V}(X) = \frac{1}{\theta^2}$.
Puisque $\mathbb{E}(X)$ et $\mathbb{V}(X)$ sont des fonctions du paramètre inconnu θ, on peut appliquer la méthode des moments à partir de $\mathbb{E}(X)$ ou de $\mathbb{V}(X)$.
Ainsi, puisque $\mathbb{E}(X) = \frac{1}{\theta}$, on a $\hat{\theta}_1$ qui est solution de :

$$\frac{1}{\hat{\theta}_1} = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

soit

$$\hat{\theta}_1 = \frac{1}{\frac{1}{n} \sum_{i=1}^{n} X_i} = \frac{1}{\bar{X}_n}.$$

Par ailleurs, puisque $\mathbb{V}(X) = \frac{1}{\theta^2}$, on peut aussi avoir $\hat{\theta}_1$ qui est solution de :

$$\frac{1}{\hat{\theta}_1^2} = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \bar{X}_n\right)^2,$$

soit

$$\hat{\theta}_1 = \frac{1}{\sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(X_i - \bar{X}_n\right)^2}} = \frac{1}{(S'_n)}.$$

Ainsi, sur cet exemple, on voit qu'il n'y a pas unicité forcément de l'estimateur obtenu par la méthode des moments. Tout dépend du moment considéré. Mais, la règle est de toujours privilégier le moment d'ordre le plus petit possible, car en général, les calculs associés sont plus simples.

Par la méthode du maximum de vraisemblance :
Soit X une variable aléatoire de loi $\mathcal{E}(\theta)$. Alors, la fonction de densité a pour expression :

$$f_\theta(x) = \begin{cases} \theta \exp(-\theta x) & \text{si } x > 0 \\ 0 & \text{sinon.} \end{cases}$$

Par conséquent, la vraisemblance a pour expression :

$$L(x_1, x_2, \ldots, x_n; \theta) = \prod_{i=1}^{n} f_\theta(x_i),$$

Soit :

$$L(x_1, x_2, \ldots, x_n; \theta) = \theta^n \exp -\theta \sum_{i=1}^{n} x_i \prod_{i=1}^{n} \mathbb{1}_{]0,+\infty[}(x_i).$$

Pour simplifier les calculs, nous n'allons pas travailler sur la vraisemblance, mais sur la log-vraisemblance, à savoir $g(\theta) = \ln L(x_1, x_2, \ldots, x_n; \theta)$. En raison de la croissance de la fonction logarithme, maximiser la fonction g équivaut à maximiser la vraisemblance.

Or $g(\theta) = n \ln \theta - \theta \sum_{i=1}^{n} x_i + \ln \prod_{i=1}^{n} \mathbb{1}_{]0,+\infty[}(x_i)$. Ainsi, la fonction est dérivable et :

$$\frac{\partial g}{\partial \theta}(\theta) = \frac{n}{\theta} - \sum_{i=1}^{n} x_i.$$

Ainsi, résoudre $\frac{\partial g}{\partial \theta}(\theta) = 0$ équivaut à $\theta = \frac{n}{\sum_{i=1}^{n} x_i}$.
Attention, cela signifie qu'en $\theta = \frac{n}{\sum_{i=1}^{n} x_i}$, la fonction de vraisemblance est extrémale, mais nous ne savons pas si elle y est bien maximale. Il faut donc regarder, en ce point, si la dérivée seconde de g est bien négative.
Puisque $\frac{\partial^2 g}{\partial \theta^2}(\theta) = \frac{-n}{\theta^2}$, c'est bon.
Ainsi, l'estimateur obtenu par la méthode du maximum de vraisemblance est $\hat{\theta}_2 = \frac{1}{X_n}$.

Avec cette exemple, on voit que parfois, l'estimateur donné par la méthode du maximum de vraisemblance est identique à celui obtenu par la méthode des moments, mais que ce n'est pas systématique.
De même, il n'y a pas toujours existence et/ou unicité de l'estimateur du maximum de vraisemblance.
Par ailleurs, la méthode utilisée ici pour déterminer le point rendant la fonction de vraisemblance maximale ne peut pas toujours être appliquée car il est impératif que la fonction de vraisemblance soit dérivable pour tout θ. Or, dès que la fonction de densité considérée a un support qui dépend du paramètre θ, nous ne serons plus dans ce cadre; il faudra alors utiliser des considérations de monotonie pour pouvoir répondre.

3.1.4 Propriétés des estimateurs

Nous avons vu comment construire un ou plusieurs estimateurs, et ceci pour un paramètre inconnu θ. Mais, a présent comment mesurer la qualité d'un estimateur et surtout, lorsque l'on en a plusieurs, comment choisir le meilleur ou le cas échéant, les améliorer?

Biais

Pour pouvoir considérer que $T_n(x_1, x_2, \ldots, x_n)$ est une valeur approchée de θ, il faut que les valeurs prises par T_n ne s'écartent pas trop de θ.
Cependant, puisque T_n est une variable aléatoire, on ne peut imposer qu'une condition sur sa valeur moyenne, à savoir son espérance.
Ceci nous conduit à définir le **"biais"** d'un estimateur comme l'écart entre son espérance et la vraie valeur du paramètre, à savoir $b(T_n) = \mathbb{E}(T_n) - \theta$.

Définition 24.
*Un estimateur T_n de θ est dit **sans biais** si, pour tout $\theta \in \Theta$ et tout entier n, on a :*

$$\mathbb{E}(T_n) = \theta \quad \text{soit} \quad b(T_n) = 0.$$

Remarque 53.
Soit $(X_1, \ldots, X_n)$ un n-échantillon de X. Si le paramètre à estimer θ n'est autre que l'espérance de X, alors $\bar{X}_n$ est toujours un estimateur sans biais de θ.
En effet :

$$\mathbb{E}(\bar{X}_n) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \frac{1}{n}\sum_{i=1}^{n}\theta = \theta.$$

Remarque 54.
Soit $(X_1, \ldots, X_n)$ un n-échantillon de X. Si le paramètre à estimer θ n'est autre que la variance de X, alors l'estimateur $T_n = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2$ est un estimateur biaisé ($b(T_n) \neq 0$).

En effet :

$$\begin{aligned}
\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2 &= \sum_{i=1}^{n}\left(X_i - \mathbb{E}(X) + \mathbb{E}(X) - \bar{X}_n\right)^2 \\
&= \sum_{i=1}^{n}\left(X_i - \mathbb{E}(X)\right)^2 + \sum_{i=1}^{n}\left(\mathbb{E}(X) - \bar{X}_n\right)^2 + 2\sum_{i=1}^{n}\left(X_i - \mathbb{E}(X)\right)\left(\mathbb{E}(X) - \bar{X}_n\right) \\
&= \sum_{i=1}^{n}\left(X_i - \mathbb{E}(X)\right)^2 + n\left(\mathbb{E}(X) - \bar{X}_n\right)^2 + 2\left(\mathbb{E}(X) - \bar{X}_n\right)\sum_{i=1}^{n}\left(X_i - \mathbb{E}(X)\right) \\
&= \sum_{i=1}^{n}\left(X_i - \mathbb{E}(X)\right)^2 + n\left(\mathbb{E}(X) - \bar{X}_n\right)^2 + 2\left(\mathbb{E}(X) - \bar{X}_n\right)\left(n\bar{X}_n - n\mathbb{E}(X)\right) \\
&= \sum_{i=1}^{n}\left(X_i - \mathbb{E}(X)\right)^2 + n\left(\mathbb{E}(X) - \bar{X}_n\right)^2 - 2n\left(\mathbb{E}(X) - \bar{X}_n\right)^2 \\
&= \sum_{i=1}^{n}\left(X_i - \mathbb{E}(X)\right)^2 - n\left(\mathbb{E}(X) - \bar{X}_n\right)^2
\end{aligned}$$

D'où :

$$\begin{aligned}
\mathbb{E}(T_n) &= \frac{1}{n}\mathbb{E}\left(\sum_{i=1}^{n}(X_i - \mathbb{E}(X))^2 - n\left(\mathbb{E}(X) - \bar{X}_n\right)^2\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(X_i - \mathbb{E}(X)\right)^2 - \mathbb{E}\left(\left(\mathbb{E}(X) - \bar{X}_n\right)^2\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{V}(X_i) - \mathbb{V}(\bar{X}_n) \\
&= \theta - \frac{\theta}{n} \\
&= \frac{n-1}{n}\theta
\end{aligned}$$

Ainsi, puisque $\mathbb{E}(T_n) \neq \theta$, cet estimateur est biaisé.

Définition 25.
Un estimateur T_n de θ est dit ***asymptotiquement sans biais*** *si, pour tout $\theta \in \Theta$ on a :*

$$\lim_{n\to+\infty} \mathbb{E}(T_n) = \theta.$$

Remarque 55.
Soit $(X_1, \ldots, X_n)$ un n-échantillon de X. On suppose que le paramètre à estimer θ n'est autre que la variance de X.
On a vu que $T_n = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2$ est un estimateur biaisé de θ et que $\mathbb{E}(T_n) = \theta - \frac{\theta}{n}$.
Si l'on pose $\tilde{T}_n = T_n + \frac{\theta}{n}$, alors $\mathbb{E}(\tilde{T}_n) = \theta$ mais pour autant, $\tilde{T}_n$ n'est pas un estimateur sans biais de θ car ce n'est tout simplement pas un estimateur. En effet, $\tilde{T}_n$ dépend du paramètre inconnu θ et donc cela ne convient pas pour un estimateur.
Par contre, si l'on pose $\tilde{T}_n = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2$ soit $\tilde{T}_n = \frac{n}{n-1}T_n$, alors $\tilde{T}_n$ devient un estimateur sans biais de θ.

Convergence

Intuitivement, on pense que si la taille n de l'échantillon augmente, alors l'information sur θ va augmenter et donc l'estimateur devrait "d'une certaine manière" se "rapprocher" de θ. Cependant, puisqu'un estimateur est une variable aléatoire, la notion de se rapprocher est celle de la convergence en probabilité.

Définition 26.
Un estimateur T_n de θ est ***convergent*** *si :*

$$T_n \underset{p}{\to} \theta \quad \textit{soit} \quad \forall \varepsilon > 0,\ P(|T_n - \theta| > \varepsilon) \underset{n\to+\infty}{\to} 0.$$

Théorème 2.
Tout estimateur sans biais de θ dont la variance tend vers 0 est convergent

Preuve.
Ceci découle immédiatement de l'inégalité de Bienaymé-Tchebychev. ■

Remarque 56.
Il est possible, dans le théorème de ne considérer que des estimateurs asymptotiquement sans biais en lieu et place des estimateurs sans biais.

Optimalité

Qualité d'un estimateur :
La qualité d'un estimateur va se mesurer au moyen d'une distance au paramètre θ, par exemple $|T_n - \theta|$ ou $(T_n - \theta)^2$.
Afin d'obtenir une valeur numérique, on va considérer la moyenne en probabilité de cette distance.

Définition 27.
On appelle ***erreur quadratique moyenne*** *de l'estimateur T_n du paramètre θ, la quantité :*

$$EQ(T_n) = \mathbb{E}\left((T_n - \theta)^2\right)$$

Proposition 21.
Soit T_n un estimateur de θ, alors :

$$EQ(T_n) = \mathbb{V}(T_n) + (b(T_n))^2$$

Preuve.

$$\begin{aligned} EQ(T_n) &= \mathbb{E}\left((T_n - \theta)^2\right) \\ &= \mathbb{E}\left((T_n - \mathbb{E}(T_n) + \mathbb{E}(T_n) - \theta)^2\right) \\ &= \mathbb{E}\left((T_n - \mathbb{E}(T_n))^2\right) + \mathbb{E}\left((\mathbb{E}(T_n) - \theta)^2\right) + 2\mathbb{E}\left((T_n - \mathbb{E}(T_n))(\mathbb{E}(T_n) - \theta)\right) \\ &= \mathbb{V}(T_n) + (\mathbb{E}(T_n) - \theta)^2 + 2(\mathbb{E}(T_n) - \theta)\mathbb{E}\left(T_n - \mathbb{E}(T_n)\right) \\ &= \mathbb{V}(T_n) + (b(T_n))^2 \end{aligned}$$

■

Remarque 57.
Si T_n est un estimateur sans biais du paramètre θ, alors $EQ(T_n) = \mathbb{V}(T_n)$.
Si on privilégie les estimateurs sans biais, un tel estimateur sera optimal si sa variance est la plus faible.

Définition 28.
Soit T_n et T'_n deux estimateurs sans biais de θ.
*On dit que T_n est **plus efficace** que T'_n si, à partir d'un certain rang pour n et pour tout $\theta \in \Theta$, on a $\mathbb{V}(T_n) \leq \mathbb{V}(T'_n)$.*

3.2 Estimation par intervalle de confiance

3.2.1 Introduction

Avant de vous présenter la théorie sur cette partie du cours, nous allons commencer par un exemple introductif.

Un industriel commande un lot de tiges métalliques qu'il ne peut utiliser que si leur longueur est comprise entre 23.6mm et 23.7mm.
Ces tiges ont été fabriquées par une machine qui, lorsqu'elle est réglée à la valeur m produit des tiges dont la longueur peut être considérée comme une variable aléatoire de loi $\mathcal{N}(m, \sigma^2)$ où l'écart-type σ est une caractéristique de la machine, de valeur connue $\sigma = 0.02mm$.
Compte tenu de la symétrie de la distribution gaussienne, la proportion de tiges utilisables par l'industriel sera maximale si le réglage est effectué à $m_0 = 23.65$.
Ne connaissant pas cette valeur, à la réception d'un lot de tiges métalliques, l'industriel prélève au hasard n tiges dont il mesure les longueurs x_1, x_2,..., x_n. Ce sont des réalisations des variables aléatoires X_1, X_2,..., X_n. L'idée est de se faire une idée du paramètre de réglage m à partir des données.
D'après la modélisation de la longueur d'une tige, m n'est autre que l'espérance de la loi sous-jacente, et donc pour obtenir un estimateur, l'industriel calcule la moyenne des longueurs observées, à savoir $\bar{x}_n = 32.63$. Il en conclut que s'il n'est pas réaliste de croire que la vraie valeur du paramètre m est exactement 23.63, elle doit malgré tout être très proche.
Il lui paraît raisonnable de conclure qu'il y a 95 chances sur 100 que la valeur de m soit comprise entre $23.63 - a$ et $23.63 + b$.
Le problème consiste alors à fixer des valeurs précises de a et b et on comprend bien que ces valeurs vont dépendre du nombre de chances que l'on fixe d'avoir la vrai valeur dans l'intervalle.
L'intervalle ainsi obtenu est appelé **intervalle de confiance** et sa probabilité de contenir la vraie valeur du paramètre est le **niveau de confiance**.
La longueur de cet intervalle est proportionnelle à ce niveau de confiance.
On peut toujours donner un intervalle de confiance qui contient avec certitude la vraie valeur du paramètre m en le rendant suffisamment large. Mais alors, cet intervalle n'est pas informatif!
Donc, il va falloir trouver un compromis entre le longueur de l'intervalle et le niveau de confiance.

Définition 29.
Soit une famille paramétrique quelconque de loi $\{P_\theta,\ \theta \in \Theta\}$.
*Un **intervalle de confiance** pour le paramètre θ, au **niveau de confiance** $1-\alpha$ ($1-\alpha \in]0,1[$) est un intervalle qui a la probabilité $1-\alpha$ de contenir la vraie valeur de θ.*

3.2.2 Principe de construction

Dans l'exemple introductif, l'intervalle de confiance était $[\bar{x}_n - a; \bar{x}_n + b]$ et il devait contenir la vraie valeur de m avec une probabilité $1 - \alpha$.
La détermination de a et b va se faire à partir de la valeur du niveau de confiance $1 - \alpha$, valeur fixée par la statisticien.
La condition est :

$$\begin{aligned} 1 - \alpha &= P(\bar{X}_n - a < m < \bar{X}_n + b) \\ 1 - \alpha &= P(-b < \bar{X}_n - m < a) \end{aligned}$$

Une condition pour déterminer 2 valeurs $\Rightarrow$ une infinité de solution.

La loi de $\bar{X}_n - m$ étant symétrique, on choisit $b = a$.
La loi de $\bar{X}_n$ est $\mathcal{N}(m, \sigma^2/n)$, on a $\sqrt{n}\frac{\bar{X}_n - m}{\sigma} \sim \mathcal{N}(0, 1)$.
D'où :

$$\begin{aligned} 1 - \alpha &= P\left(-\frac{\sqrt{n}a}{\sigma} < \sqrt{n}\frac{\bar{X}_n - m}{\sigma} < \frac{\sqrt{n}a}{\sigma}\right) \\ &= P\left(Z < \frac{\sqrt{n}a}{\sigma}\right) - P\left(Z < -\frac{\sqrt{n}a}{\sigma}\right) \text{ avec } Z = \sqrt{n}\frac{\bar{X}_n - m}{\sigma} \\ &= P\left(Z < \frac{\sqrt{n}a}{\sigma}\right) - P\left(Z > \frac{\sqrt{n}a}{\sigma}\right) \\ &= P\left(Z < \frac{\sqrt{n}a}{\sigma}\right) - \left(1 - P\left(Z < \frac{\sqrt{n}a}{\sigma}\right)\right) \\ &= 2P\left(Z < \frac{\sqrt{n}a}{\sigma}\right) - 1 \end{aligned}$$

Si on note F la fonction de répartition associée à Z, on a :

$$F\left(\frac{\sqrt{n}a}{\sigma}\right) = 1 - \frac{\alpha}{2}.$$

Donc

$$\frac{\sqrt{n}a}{\sigma} = F^{-1}\left(1 - \frac{\alpha}{2}\right).$$

Ainsi, pour $1 - \alpha = 95\%$, soit $\alpha = 0.05$ et $n = 100$, on a $F^{-1}(1 - \frac{\alpha}{2}) = F^{-1}(0.975) = 1.96$ et on en déduit que $a = 0.004$.
Ainsi l'intervalle est $23.626 < m < 23.634$.

Principe :
Le point de départ est fourni par un estimateur T_n de θ, construit à partir de $(X_1, X_2, \ldots, X_n)$ et dont on connait la loi en fonction de θ.
Ainsi on détermine les valeurs $t_1 = t_1(\theta)$ et $t_2 = t_2(\theta)$ tels que :

$$P(t_1 \leq T_n \leq t_2) = 1 - \alpha$$

Ensuite, pour obtenir un intervalle de θ, on inverse les bornes :

$$P(a(T_n) \leq \theta \leq b(T_n)) = 1 - \alpha$$

$\Rightarrow [a(T_n), b(T_n)]$ est l'intervalle de confiance pour θ au niveau de confiance $1 - \alpha$.
Attention, il s'agit d'un intervalle aléatoire!

Remarque 58.
Il y a un choix arbitraire à faire car au final, on cherche t_1 et t_2 tels que $P(T_n \leq t_1) + P(T_n \geq t_2) = \alpha$ soit $P(T_n \leq t_1) = \alpha_1$ et $P(T_n \geq t_2) = \alpha_2$ avec $\alpha_1 + \alpha_2 = \alpha$. Cette décomposition possède une infinité de choix!

- *Intervalle bilatérale : ($\alpha_1 > 0$ et $\alpha_2 > 0$)*
  - *cas symétrique : $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$ (en général lorsque la distribution de la loi de T_n est symétrique ou si pas de précision particulière donnée*
  - *cas non symétrique : $\alpha_1 \neq \alpha_2$ et tels que $\alpha_1 + \alpha_2 = \alpha$ (aucune raison particulière)*
- *Intervalle unilatéral : ($\alpha_1 \alpha_2 = 0$)*
  - *à droite : ($\alpha_1 = 0$ et $\alpha_2 = \alpha$) (raison : c'est l'interprétation donnée à θ comme par exemple la résistance d'un matériau qui doit être supérieure à un seuil minimal ($\theta \geq a(T_n)$)*
  - *à gauche : ($\alpha_2 = 0$ et $\alpha_1 = \alpha$) (raison : c'est l'interprétation donnée à θ comme par exemple la proportion de défectueux dans un lot que l'on souhaite inférieure à un seuil maximal ($\theta \leq b(T_n)$)*

3.2.3 Intervalle pour une proportion

Nous allons voir que l'inversion des bornes de $t_1(\theta)$ et $t_2(\theta)$ en $a(T_n)$ et $b(T_n)$ n'est pas toujours si facile que cela?

Supposons avoir effectué un sondage pour connaître les intentions de vote pour un candidat A. A chaque individu i interrogé, on associe une variable aléatoire X_i de loi de Bernoulli de paramètre p :

$$X_i = \begin{cases} 1 & \text{si vote pour } A \\ 0 & \text{sinon.} \end{cases}$$

Donc, un estimateur ponctuel de p est $\hat{p}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$ car p est l'espérance de X_i. Il s'agit en plus d'un estimateur sans biais et convergent de p.
Construisons un intervalle de confiance au niveau $1 - \alpha = 95\%$.
Ainsi, on cherche $t_1 = t_1(\theta)$ et $t_2 = t_2(\theta)$ tels que $P(t_1 \leq \hat{p}_n \leq t_2) = 0.95$.
Nous allons utiliser la loi de $n\hat{p}_n$ puisque $n\hat{p}_n \sim \mathcal{B}(n, p)$. Donc, posons $S_n = n\hat{p}_n$.
On cherche donc t_1 et t_2 tels que $P(nt_1 \leq S_n \leq nt_2) = 0.95$.

Plaçons nous dans le cadre des intervalles symétriques.
Ainsi $P(S_n \leq nt_1) \leq 0.025$ et $P(S_n \geq nt_2) \leq 0.025$.
On doit alors chercher le plus grand entier n_1 tel que $\sum_{i=0}^{n_1} C_n^i p^i (1-p)^{n-1} \leq 0.025$ et le plus petit entier n_2 tel que $\sum_{i=0}^{n_2} C_n^i p^i (1-p)^{n-1} \geq 0.975$.
Ceci est très compliqué au niveau du calcul et on utilise alors ce que l'on appelle des abaques qui sont des courbes qui donnent les valeurs de $t_2(p)$ et $t_1(p)$ en fonction de p.
Par exemple, ici si l'on prend $n = 100$ et $\bar{x}_n = 0.44$, on trouve comme intervalle pour p : $[0.341; 0.538]$.

Pour éviter l'utilisation des abaques, une autre manière de procéder revient à utiliser le théorème limit central.
On dit que $\sqrt{n}\frac{\hat{p}_n - p}{\sqrt{p(1-p)}} \sim \mathcal{N}(0,1)$.
On cherche alors a tel que :

$$P\left(-a < \sqrt{n}\frac{\hat{p}_n - p}{\sqrt{p(1-p)}} < a\right) = 1 - \alpha.$$

On lit alors dans la table associée à une loi $\mathcal{N}(0,1)$, la valeur du quantile d'ordre $1 - \alpha/2$ qui correspond à la valeur de a car :

$$\begin{aligned}
P\left(-a < \sqrt{n}\frac{\hat{p}_n - p}{\sqrt{p(1-p)}} < a\right) &= P\left(\sqrt{n}\frac{\hat{p}_n - p}{\sqrt{p(1-p)}} < a\right) - P\left(\sqrt{n}\frac{\hat{p}_n - p}{\sqrt{p(1-p)}} < -a\right) \\
&= P\left(\sqrt{n}\frac{\hat{p}_n - p}{\sqrt{p(1-p)}} < a\right) - P\left(\sqrt{n}\frac{\hat{p}_n - p}{\sqrt{p(1-p)}} > a\right) \\
&= P\left(\sqrt{n}\frac{\hat{p}_n - p}{\sqrt{p(1-p)}} < a\right) - \left(1 - P\left(\sqrt{n}\frac{\hat{p}_n - p}{\sqrt{p(1-p)}} < a\right)\right) \\
&= 2P\left(\sqrt{n}\frac{\hat{p}_n - p}{\sqrt{p(1-p)}} < a\right) - 1
\end{aligned}$$

Soit $P(\sqrt{n}\frac{\hat{p}_n - p}{\sqrt{p(1-p)}} < a) = 1 - \alpha/2$.

Notons $q_{1-\alpha/2}$ la valeur de ce quantile. Donc, on a $P(\hat{p}_n - \frac{\sqrt{p(1-p)}}{\sqrt{n}} q_{1-\alpha/2} < p < \hat{p}_n + \frac{\sqrt{p(1-p)}}{\sqrt{n}} q_{1-\alpha/2}) = 1 - \alpha$.
Le problème est que les bornes de l'intervalle dépendent de p ce qui n'est pas possible pour un intervalle de confiance.
Il y a trois solutions possibles :

- on résout séparément les deux inéquations du second degré
- on approche p par $\hat{p}_n$ (approximation valable si $np(1-p) \geq 3$) et ainsi l'intervalle de

confiance devient :

$$\left[\hat{p}_n - \frac{\sqrt{\hat{p}_n(1-\hat{p}_n)}}{\sqrt{n}} q_{1-\alpha/2}; \hat{p}_n + \frac{\sqrt{\hat{p}_n(1-\hat{p}_n)}}{\sqrt{n}} q_{1-\alpha/2}\right]$$

L'application numérique dans les mêmes conditions que pour les abaques donne $[0.343; 0.537]$.

- on remplace $p(1-p)$ par sa borne supérieure à savoir $1/4$ et ainsi l'intervalle de confiance devient :

$$\left[\hat{p}_n - \frac{1}{2\sqrt{n}} q_{1-\alpha/2}; \hat{p}_n + \frac{1}{2\sqrt{n}} q_{1-\alpha/2}\right]$$

L'application numérique dans les mêmes conditions que pour les abaques donne $[0.342; 0.538]$.

3.2.4 Intervalles associés aux paramètres d'une loi normale

Dans la suite, nous allons considérer $(X_1, X_2, \ldots, X_n)$ un n-échantillon de loi $\mathcal{N}(m, \sigma^2)$ et nous allons chercher à estimer les paramètres m et σ^2 séparément.

Cependant, avant de pouvoir mener à bien ce projet, nous avons besoin de quelques nouvelles lois de probabilité.

Quelques lois particulières

Définition 30.
Soit $(X_1, X_2, \ldots, X_n)$ un n-échantillon de loi $\mathcal{N}(0,1)$.
Posons $Y = X_1^2 + X_2^2 + \ldots + X_n^2$.
*Alors la loi de Y est une loi du **Chi-deux** à n degrés de liberté, ce que l'on note $\chi^2(n)$.*

Remarque 59.
Si $(X_1, X_2, \ldots, X_n)$ est un n-échantillon de loi $\mathcal{N}(m, \sigma^2)$, alors $Y = \left(\frac{X_1-m}{\sigma}\right)^2 + \left(\frac{X_2-m}{\sigma}\right)^2 + \ldots + \left(\frac{X_n-m}{\sigma}\right)^2$ suit une loi du Chi-deux à n degrés de liberté.

Proposition 22.
Soit $Y \sim \chi^2(n)$, alors :

- $\mathbb{E}(Y) = n$,
- $\mathbb{V}(Y) = 2n$.
- *Soit $(X_1, X_2, \ldots, X_n)$ un n-échantillon de loi $\mathcal{N}(m, \sigma^2)$.*
  Si l'on pose $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$, alors $\frac{n-1}{\sigma^2} S_n^2 \sim \chi^2(n-1)$.

Définition 31.
Soit U une variable aléatoire de loi $\mathcal{N}(0,1)$ et soit V une variable aléatoire de loi $\chi^2(n)$.
On suppose de plus que U et V sont deux variables aléatoires indépendantes.

Posons $Y = \frac{U}{\sqrt{V/n}}$.
Alors la loi de Y *est une loi de* ***Student*** *à* n *degrés de liberté, ce que l'on note* $\mathcal{T}(n)$.

Proposition 23.
Soit $Y \sim \mathcal{T}(n)$, *alors :*

- $\mathbb{E}(Y) = 0$ *si* $n > 1$,
- $\mathbb{V}(Y) = \frac{n}{n-2}$ *si* $n > 2$.
- *Soit* $(X_1, X_2, \ldots, X_n)$ *un* n*-échantillon de loi* $\mathcal{N}(m, \sigma^2)$. *On sait que* $\bar{X}_n \sim \mathcal{N}(m, \sigma^2/n)$ *et que* $\frac{n-1}{\sigma^2} S_n^2 \sim \chi^2(n-1)$. *Par ailleurs, on peut montrer que* $\bar{X}_n$ *et* S_n^2 *sont des variables aléatoires indépendantes. De ce fait :*

$$\frac{\sqrt{n}\frac{\bar{X}_n - m}{\sigma}}{\sqrt{\frac{(n-1)S_n^2/\sigma^2}{n-1}}} = \sqrt{n}\frac{\bar{X}_n - m}{S_n} \sim \mathcal{T}(n-1).$$

Définition 32.
Soit U *une variable aléatoire de loi un* $\chi^2(n_1)$ *et* V *une variable aléatoire de loi un* $\chi^2(n_2)$.
On suppose de plus que U *et* V *sont deux variables aléatoires indépendantes.*
Soit $Y = \frac{U/n_1}{V/n_2}$. *Alors* Y *suit une loi de* ***Fisher*** *à* n_1 *et* n_2 *degrés de liberté, ce que l'on note* $\mathcal{F}(n_1, n_2)$.

Remarque 60.
Si $X \sim \mathcal{F}(n_1, n_2)$ *alors* $1/X \sim \mathcal{F}(n_2, n_1)$.

Proposition 24.
Soit $Y \sim \mathcal{F}(n_1, n_2)$, *alors :*

- $\mathbb{E}(Y) = \frac{n_2}{n_2-2}$ *si* $n_2 > 2$,
- $\mathbb{V}(Y) = \frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)(n_2-4)}$ *si* $n_2 > 4$.

Intervalle pour m à σ^2 connu

Ce cadre correspond à ce lui de l'exemple introductif et on a donc déjà le résultat, à savoir :

Proposition 25.
Un intervalle de confiance de niveau $1-\alpha$, *centré en* $\bar{X}_n$ *de* m *est :*

$$\left[\bar{X}_n - \frac{\sigma}{n} q_{1-\alpha/2}; \bar{X}_n + \frac{\sigma}{n} q_{1-\alpha/2}\right]$$

où $q_{1-\alpha/2}$ *est le quantile d'ordre* $1-\alpha/2$ *d'une loi* $\mathcal{N}(0,1)$ *à savoir que* $F(q_{1-\alpha/2}) = 1-\alpha/2$ *avec* F *la fonction de répartition associée à une* $\mathcal{N}(0,1)$.

Intervalle pour m à σ^2 inconnu

L'intervalle de confiance ne peut plus convenir car il dépend du paramètre inconnu σ!
Mais, si l'on regarde les calculs menés pour obtenir cet intervalle, ils reposent sur le fait que $\sqrt{n}\frac{\bar{X}_n - m}{\sigma} \sim \mathcal{N}(0,1)$. C'est certes toujours vrai, mais la variable dépend du paramètre inconnu σ. Donc, il faudrait pouvoir faire disparaître σ en le remplaçant par quelque chose de connu mais tout en continuant à connaître la loi de la nouvelle variable induite.
Or justement, on sait que $\sqrt{n}\frac{\bar{X}_n - m}{S_n} \sim \mathcal{T}(n-1)$. C'est donc sur cette nouvelle variable que l'on va construire l'intervalle de confiance.
On cherche donc t tel que $P(-t < \sqrt{n}\frac{\bar{X}_n - m}{S_n} < t) = 1 - \alpha$.
Or :

$$\begin{aligned} P\left(-t < \sqrt{n}\frac{\bar{X}_n - m}{S_n} < t\right) &= P\left(\sqrt{n}\frac{\bar{X}_n - m}{S_n} < t\right) - P\left(\sqrt{n}\frac{\bar{X}_n - m}{S_n} < -t\right) \\ &= P\left(\sqrt{n}\frac{\bar{X}_n - m}{S_n} < t\right) - P\left(\sqrt{n}\frac{\bar{X}_n - m}{S_n} > t\right) \\ &= P\left(\sqrt{n}\frac{\bar{X}_n - m}{S_n} < t\right) - \left(1 - P\left(\sqrt{n}\frac{\bar{X}_n - m}{S_n} < t\right)\right) \\ &= 2P\left(\sqrt{n}\frac{\bar{X}_n - m}{S_n} < t\right) - 1 \end{aligned}$$

Donc, on a :

$$P\left(\sqrt{n}\frac{\bar{X}_n - m}{S_n} < t\right) = 1 - \alpha/2.$$

Ainsi, t est le quantile d'ordre $1 - \alpha/2$ d'une loi de Student à $n-1$ degrés de liberté.
D'où :

Proposition 26.
Un intervalle de confiance de niveau $1-\alpha$, centré en $\bar{X}_n$ de m est :

$$\left[\bar{X}_n - \frac{S_n}{n}t_{1-\alpha/2}; \bar{X}_n + \frac{S_n}{n}t_{1-\alpha/2}\right]$$

où $t_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ d'une loi $\mathcal{T}(n-1)$ à savoir que $F_t(q_{1-\alpha/2}) = 1 - \alpha/2$ avec F_t la fonction de répartition associée à une $\mathcal{T}(n-1)$.

Intervalle pour σ^2 à m connue

Pour le moment, nous nous sommes intéressés au problème de l'estimation de l'espérance, à variance connue ou inconnue. Mais, on peut aussi le problème inverse à savoir le problème de l'estimation de la variance.
C'est ce qui nous occupe dans les deux sections à venir.

Commençons donc par le cas le plus "simple" de l'estimation de la variance à espérance supposée connue.

D'après les chapitres précédents, on sait que $\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - m)^2$ est un estimateur de la variance.
On peut d'ailleurs montrer que cet estimateur est sans biais, convergent et efficace.
De plus, on a $\frac{n}{\sigma^2}\hat{\sigma}_n^2 \sim \chi^2(n)$.

On cherche alors a et b tels que $P(a < \frac{n}{\sigma^2}\hat{\sigma}_n^2 < b) = 1 - \alpha$.
Posons $\alpha_1 = P(\frac{n}{\sigma^2}\hat{\sigma}_n^2 < a)$ et $\alpha_2 = P(\frac{n}{\sigma^2}\hat{\sigma}_n^2 > b)$. Ainsi $\alpha_1 + \alpha_2 = \alpha$.
Puisque la distribution d'une loi du Chi-deux n'est pas symétrique, on procède à un choix arbitraire pour α_1 et α_2. Le cas classique est de prendre $\alpha_1 = \alpha_2 = \alpha/2$.
On utilise alors la table associée à une loi du Chi-deux pour la détermination de aa et b, et il en résulte que l'intervalle de confiance pour σ^2 au niveau de confiance $1 - \alpha$ est de la forme :

$$\left[\frac{n}{b}\hat{\sigma}_n^2; \frac{n}{a}\hat{\sigma}_n^2\right].$$

Intervalle pour σ^2 à m inconnue

L'intervalle précédent ne peut pas convenir car $\hat{\sigma}_n^2$ dépend de m maintenant inconnue!
Pour remédier à la difficulté, en lieu et place de $\hat{\sigma}_n^2$, on utilise $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$. On sait de plus que $\frac{n-1}{S_n^2}\hat{\sigma}_n^2 \sim \chi^2(n-1)$.

On cherche alors a et b tels que $P(a < \frac{n-1}{\sigma^2}S_n^2 < b) = 1 - \alpha$.
Posons $\alpha_1 = P(\frac{n}{\sigma^2}S_n^2 < a)$ et $\alpha_2 = P(\frac{n}{\sigma^2}S_n^2 > b)$. Ainsi $\alpha_1 + \alpha_2 = \alpha$.
Puisque la distribution d'une loi du Chi-deux n'est pas symétrique, on procède à un choix arbitraire pour α_1 et α_2. Le cas classique est de prendre $\alpha_1 = \alpha_2 = \alpha/2$.
On utilise alors la table associée à une loi du Chi-deux pour la détermination de aa et b, et il en résulte que l'intervalle de confiance pour σ^2 au niveau de confiance $1 - \alpha$ est de la forme :

$$\left[\frac{n-1}{b}S_n^2; \frac{n-1}{a}S_n^2\right].$$