Chapitre 7
Probabilité

Serge COHEN

IPANEMA, UAR3461,
Institut photonique d'analyse non-destructive européen des matériaux anciens

1 Statistique et probabilité

L'objet principal de la statistique descriptive est de fournir une représentation synthétique, directement compréhensible, d'un ensemble d'individus* formant une population*, au sens statistique (*cf.* 4). Par extension, ces outils peuvent être utilisés pour comparer deux populations, mettre en évidence leur similitude et leurs différences. Toutefois, le plus souvent, après avoir proposé une description de la population étudiée, on souhaite pouvoir étendre les résultats. Il s'agit là, par exemple, de généraliser les résultats obtenus à une population plus large non intégralement étudiée, et/ou d'extrapoler ce que pourraient être les résultats sur une population avec d'autres caractéristiques, ou encore d'exploiter la description de la population afin d'identifier le ou les processus qui lient ces différentes caractéristiques. Pour l'ensemble de ces tâches, il est nécessaire de modéliser, explicitement ou implicitement, les données statistiques en vue d'étendre les résultats au-delà de la description.

La théorie des probabilités est née au 16e siècle de l'étude des jeux. Elle visait alors à proposer une description utile d'un phénomène aléatoire, comportant donc une part significative d'incertitude. Ainsi la construction d'un modèle probabiliste d'un jeu de cartes ou d'un jeu de dés permettra de justifier de la façon de répartir les gains du jeu si la partie ne peut pas être achevée. Les outils de la théorie des probabilités se sont progressivement étendus dans d'autres domaines d'application tels que l'étude des phénomènes climatiques ou météorologiques. Au début du 20e siècle, Kolmogorov refonde la théorie des probabilités et lui donne la forme qui est encore actuellement utilisée. Le formalisme de Kolmogorov permet d'unifier le traitement des variables discrètes* et des variables continues*. Dès lors, et comme pour la statistique descriptive,

l'application de la théorie des probabilités s'ouvre à de nombreux domaines comme l'économie, la sociologie, la géographie, mais aussi la biologie, la physique…

Là où la statistique fournit l'ensemble des outils qui permettent de réunir et d'analyser des ensembles cohérents de données relatives à des populations, la théorie des probabilités s'intéresse à l'étude des phénomènes caractérisés par le hasard et l'incertitude. Ces deux domaines des mathématiques se complètent quand il s'agit de modéliser l'incertitude qui existe dans un ensemble cohérent de données réelles. Ainsi, la théorie des probabilités fournit les outils permettant de préciser l'incertitude qui découle de l'estimation des caractéristiques d'une population à partir d'un échantillon de celle-ci. Ces mêmes outils forment la base des méthodes de régression et de classification de la statistique qui permettent d'extrapoler des résultats déjà établis à de nouvelles données, et qui font l'objet des chapitres 9 et 10. Et ce sont encore les outils de la théorie des probabilités qui permettent de décrire le risque pris en inférant un processus à partir de données observées (*cf.* 8).

2 Lois de probabilités

2.1 Probabilité et variable aléatoire

Dans la théorie des probabilités, le hasard ou l'incertitude est introduit à travers la réalisation d'une expérience aléatoire qui aboutit à un événement ou issue. Une telle expérience est appelée aléatoire, car on ne connaît pas *a priori*, c'est-à-dire avant de réaliser l'expérience, l'issue qu'elle va produire. En revanche, on connaît avant même de réaliser l'expérience l'ensemble de tous les événements que l'expérience pourrait produire. On appelle cet ensemble, l'ensemble fondamental ou l'univers des possibles, le plus souvent noté Ω ; une issue –élément de Ω– étant notée ω. La théorie des probabilités s'attache alors à modéliser mathématiquement ce qu'on connaît de l'expérience à travers la fonction de probabilité, c'est-à-dire une fonction (noté $\mathbb{P}$) qui pour un sous-ensemble $A \subset \Omega$[1] associe une valeur réelle entre 0 et 1, avec les propriétés suivantes :

- Pour tout $A \subset \Omega$, $0 \leq \mathbb{P}(A) \leq 1$[2] (*positivité*) ;
- La probabilité de l'univers est de 1 : $\mathbb{P}(\Omega) = 1$[3] (*masse unitaire*) ;
- La probabilité de l'union de deux sous-ensembles disjoints est la somme des probabilités de chacun des sous-ensembles : pour tout $A_1 \subset \Omega$ et $A_2 \subset \Omega$, si $A_1 \cap A_2 = \emptyset$ alors $\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2)$[4] (*additivité*).

Un exemple concret d'expérience aléatoire est le jet d'une pièce de monnaie. Cette expérience a deux issues possibles, *pile* ou *face* : $\Omega = \{\textbf{pile}, \textbf{face}\}$. Si la pièce est

[1] On lit : (A) est inclu dans Ω

[2] Lire : Pour tout ensemble A inclus dans Ω, la probabilité de A est supérieur ou égale à 0 et inférieure ou égale à 1

[3] Lire : La probabilité de l'ensemble Ω est de 1

[4] Lire : Pour tout ensembles A_1 et A_2 inclus dans Ω, si l'intersection de A_1 et A_2 est vide, alors la probabilité de l'union de A_1 et A_2 est la somme des probabilités de A_1 et de A_2

équilibrée, c'est-à-dire s'il est aussi probable d'avoir une issue que l'autre, l'expérience peut être modélisée par la fonction de probabilité suivante :

- $\mathbb{P}(\{\textbf{pile}\}) = \mathbb{P}(\{\textbf{face}\})$
- $\mathbb{P}(\{\textbf{pile}\}) + \mathbb{P}(\{\textbf{face}\}) = \mathbb{P}(\{\textbf{pile}, \textbf{face}\}) = \mathbb{P}(\Omega) = 1$
- on en déduit : $\mathbb{P}(\{\textbf{pile}\}) = \mathbb{P}(\{\textbf{face}\}) = \frac{1}{2}$, soit une probabilité de $0,5$ chance de tomber sur pile ou face.

On a un exemple similaire avec le jet d'un dé à six faces, dont l'issue est l'obtention d'un nombre entier entre 1 et 6. Si le dé est équilibré, le même type de raisonnement permet de modéliser l'expérience par la probabilité : $\mathbb{P}(\{1\}) = \mathbb{P}(\{2\}) = \ = \mathbb{P}(\{6\}) = \frac{1}{6}$, soit une probabilité de $0,16666$ de tomber sur chacune des faces du dé.

A contrario l'expérience visant à mesurer la vitesse du son dans un milieu continu (par exemple de l'air à une pression et une température bien définies) en mesurant la distance parcouru par un son pendant une durée de $5s$ dans le milieu étudié n'est pas une expérience aléatoire, à moins de se concentrer sur l'étude des petites variations entre des mesures successives. En effet, le résultat de l'expérience est déterminée par la vitesse du son qui est précisément définie et les petites variations de mesure observées ne correspondent pas à une variation de la vitesse du son, mais à des processus liés à la mesure elle-même. Nous reviendrons sur cet exemple en fin de ce chapitre.

2.1.1 Variable aléatoire

Souvent l'issue d'une expérience aléatoire peut naturellement être exprimée par un nombre. Pour notre expérience du jet d'un dé, on représente naturellement l'issue ω de l'expérience par un nombre entier entre 1 et 6. Il arrive aussi souvent qu'on ne soit pas intéressé par l'issue elle-même de l'expérience, mais par une caractérisation numérique de celle-ci. C'est par exemple le cas si on lance au hasard un ballon, l'issue ω étant le lieu où retombe le ballon ; on peut ne pas être réellement intéressé par le lieu, mais en revanche vouloir connaître la distance parcourue qu'on notera $X(\omega)$.

Dans un cas comme dans l'autre, la fonction X associe le nombre $X(\omega)$ à l'issue ω de l'expérience aléatoire. Une telle fonction est appelée **variable aléatoire**. Il y a quelques aspects utiles à noter :

- Ce qui est *aléatoire* dans la variable aléatoire, ce n'est pas tant la fonction X que le fait qu'elle soit appliquée à une expérience aléatoire. Autrement dit, si la répétition de l'expérience amène plusieurs fois la même issue ($\omega_1 = \omega_2$), alors, pour chacune de ces expériences, la variable aléatoire aura toujours la même valeur ($X(\omega_1) = X(\omega_2)$) ;
- Pour autant, la fonction X n'a pas besoin d'être explicitée, ni même connue précisément. En d'autres termes, on peut s'attacher à décrire les probabilités d'obtenir un ensemble de valeurs pour X sans avoir à caractériser la fonction de probabilité de ω ni même savoir calculer $X(\omega)$ pour un ω donné.

La variable aléatoire permet donc de modéliser la caractérisation physique ou concrète de l'issue d'une expérience aléatoire.

En statistique, la **variable aléatoire** est un outil qui permet de généraliser des résultats obtenus par exploitation d'une variable statistique* (*cf.* 4), c'est-à-dire d'explorer et de caractériser quels auraient été les résultats si la collecte de donnée avait été différente. La variable aléatoire propose une représentation phénoménologique de l'incertitude liée à la variable statistique (observée), quelle que soit la source de cette incertitude. La question, importante, des sources d'incertitudes (imprécision de la mesure, échantillonnage …) ne relève pas de la théorie des probabilités, mais de la combinaison du domaine d'application et de la statistique.

Souvent, on note E l'ensemble des valeurs que peut prendre la fonction X, mathématiquement on définit alors la variable aléatoire X comme la fonction $\Omega \to E$, $\omega \mapsto x = X(\omega)$[5]. Par extension on note alors $\mathbb{P}(x) = \mathbb{P}(\{\omega \in \Omega / X(\omega) = x\})$[6], autrement écrit, la probabilité de x est la probabilité que l'issue de l'expérience aléatoire conduise à une valeur de ω vérifiant $X(\omega) = x$. Enfin, pour un sous-ensemble $\mathcal{X}$ de E, on note $\mathbb{P}(\mathcal{X}) = \mathbb{P}(\{\omega \in \Omega / X(\omega) \in \mathcal{X}\})$[7] ; en toutes lettres, la probabilité du sous-ensemble $\mathcal{X}$ de valeur de la variable aléatoire X est la probabilité que l'issue (ω) de l'expérience aléatoire conduise à une valeur de la variable $(X(\omega))$ appartenant au sous-ensemble considéré $(\mathcal{X})$.

3 Caractérisation d'une variable aléatoire

Comme la variable statistique, la variable aléatoire est d'abord caractérisée par l'ensemble, E, des valeurs qu'elle peut prendre.

- S'il y a un nombre fini de valeurs possibles, on parle alors de **variable aléatoire finie** ;
- Si le nombre de valeurs est infini (∞), mais peut être codé par un entier, on parle de **variable aléatoire dénombrable** ;

  On parle de **variable aléatoire discrète** pour réunir les deux types précédents ;
- Enfin, s'il est impossible d'utiliser un entier (qu'il soit fini ou non) pour représenter la valeur de la variable, on parle de **variable aléatoire continue**.

Là où la variable statistique sera caractérisée par les valeurs qu'elle prend dans le jeu de données observées, la variable aléatoire sera, elle, caractérisée par sa **loi de probabilité**. Celle-ci prend une forme différente suivant que la variable aléatoire est discrète ou continue.

Dans le premier cas, celui d'une variable aléatoire discrète, les valeurs prises par la variable X sont les x_i avec un indice entier. On définit la **distribution de probabilité**,

[5] Lire : X est la fonction de Ω dans E qui, à ω, associe la valeur $x = X(\omega)$

[6] Lire : La probabilité de x est la probabilité de l'ensemble des éléments ω de Ω pour lesquels la fonction X prend la valeur x

[7] Lire : La probabilité du sous-ensemble $\mathcal{X}$ est la probabilité de l'ensemble formé par les éléments ω de Ω pour lesquels la fonction X prend une valeur $(X(\omega))$ de $\mathcal{X}$

comme la fonction de $E \rightarrow [0,1]$[8] qui donne la probabilité d'observer une valeur précise de la variable aléatoire : $p(x_i) = \mathbb{P}(x_i)$.

Dans le second cas, celui d'une variable aléatoire continue, cette fonction n'a presque jamais de sens ; en effet la probabilité que la variable aléatoire prenne une valeur donnée, et infiniment précise, est nulle, à de rares exceptions près. La variable aléatoire est alors caractérisée par sa **densité de probabilité** et/ou par sa **fonction de répartition**.

La **densité de probabilité** de la variable X est la fonction p de $E \rightarrow \mathbb{R}^+$ qui vérifie pour toute valeur de ϵ suffisamment petite $\mathbb{P}(]x-\epsilon; x+\epsilon[) \approx 2\,\epsilon\,p(x)$; plus rigoureusement cette densité est définie par la limite $p(x) = \lim_{\epsilon \to 0} \frac{\mathbb{P}(]x-\epsilon;x+\epsilon[)}{2\,\epsilon}$.

La **fonction de répartition** (aussi appelée **distribution cumulative**), P de $E \rightarrow [0,1]$, de la variable X est définie par $P(x) = \mathbb{P}(]-\infty, x[)$.

Si la fonction de répartition est d'utilisation un peu moins intuitive que la densité de probabilité, elle a toutefois quelques propriétés intéressantes. Il s'agit d'une fonction croissante dont la valeur va de 0 à 1 : $\lim_{x\to-\infty} P(x) = 0$ et $\lim_{x\to+\infty} P(x) = 1$. Un autre aspect utile de la fonction de répartition, c'est qu'elle est également définie pour une variable discrète à valeur entière. Dans ce cas elle prend la forme d'une fonction constante par morceau, les sauts se faisant pour les valeurs auxquelles la distribution de probabilité est définie. Cette dernière propriété fait de la fonction de répartition un outil intuitif pour relier une variable aléatoire continue à son pendant statistique qui, même s'il s'agit d'une variable continue, ne prend qu'un nombre fini de valeurs au sein d'un jeu de donnée.

Ainsi l'expression loi d'une variable aléatoire recouvre indifféremment ces trois caractérisations possibles d'une variable aléatoire : distribution de probabilité, densité de probabilité et loi de répartition.

3.1 Exemples et représentation graphique d'une loi de probabilité

Nous allons, ici, donner quelques exemples de loi de probabilité, pour des variables aléatoires correspondant à des sources d'incertitudes biens connues.

3.1.1 Loi uniforme : tirage d'une pièce ou d'un dé

Il s'agit de la loi de probabilité d'une variable finie, ayant la même probabilité de prendre toutes ses valeurs possibles. C'est par exemple la loi de probabilité modélisant l'incertitude du tirage à pile ou face d'une pièce équilibrée ; ou encore la loi de probabilité représentant le tirage d'un dé non truqué. Pour un dé à N faces, l'incertitude d'un tirage du dé est alors représentée par la distribution de probabilité $, p(1) = ... = p(N) = \frac{1}{N}$.

3.1.2 Loi de Poisson

La loi de Poisson, nommée d'après le mathématicien Siméon Denis Poisson (1781–1840), permet de modéliser l'incertitude liée à un processus générant des événements avec une fréquence moyenne connue, et pour lequel le temps entre deux événements succes-

[8] Fonction de l'ensemble E dans l'interval des nombres réels compris entre 0 et 1

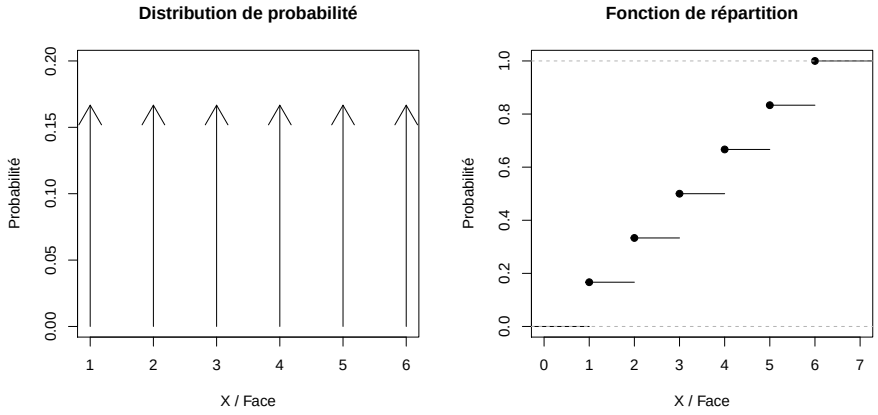


FIG. 1 : Représentation de la loi de probabilité décrivant le tirage d'un dé à 6 faces équilibré.

sifs ne dépend pas des événements eux-mêmes. Ainsi on utilise la loi de Poisson pour décrire une variable aléatoire qui décrit le comportement du nombre d'événements se produisant dans un intervalle de temps fixé, si ces événements se produisent avec une fréquence moyenne ou espérance (voir la définition de l'espérance *infra*) connue, et indépendamment du temps écoulé depuis l'événement précédent [9].

Cette loi de probabilité dépend de la fréquence moyenne de l'événement observé, ce paramètre étant le plus souvent représenté par la lettre λ.

La probabilité d'observer k événements si le paramètre de la loi de Poisson est λ est donnée par :

$$p(k) = \frac{\exp(-\lambda)\lambda^k}{k!}$$

3.1.3 Loi normale, ou loi Gaussienne

Comme on le verra plus tard, une loi de probabilité est particulièrement importante du fait de résultats théoriques en probabilité. Il s'agit de la loi Gaussienne qui est aussi appelée loi Normale du fait de son importance dans la théorie des probabilités. Il s'agit d'une loi qui décrit une variable aléatoire continue, non bornée, et dont la densité de probabilité est définie suivant deux paramètres, μ et σ. Sans rentrer dans les précisions que nous verrons un peu plus tard, le paramètre μ correspond à la position du *centre* de la loi, quand à σ, il permet de choisir la *largeur* de celle-ci.

La densité de probabilité de la loi normale est donnée par la fonction :

[9] Note des éditeurs : la loi de Poisson est également utilisée pour analyser le caractère aléatoire ou non de la distribution spatiale d'un semis de points, *cf.* 11.

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

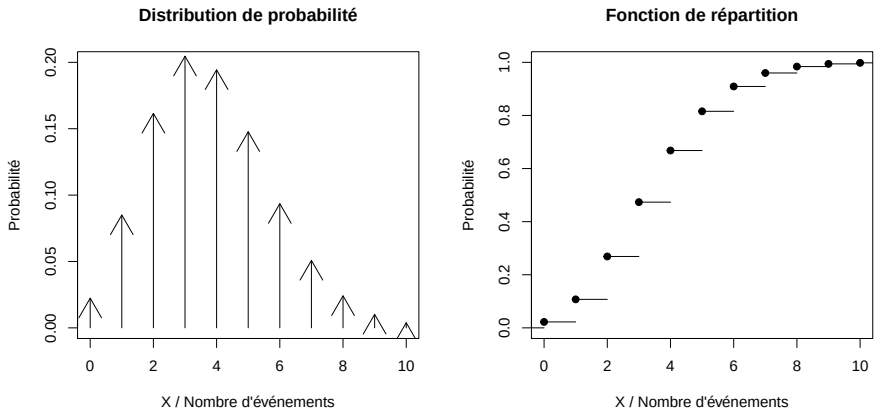


FIG. 2 : Représentation de la loi de probabilité de Poisson de paramêtre $\lambda = 3,8$, c'est-à-dire la probabilité d'observé un nombre précis d'évenement sur une unité de temps quand la fréquence moyenne des évenements est $3,8$ et que les événements sont indépendants les uns des autres.

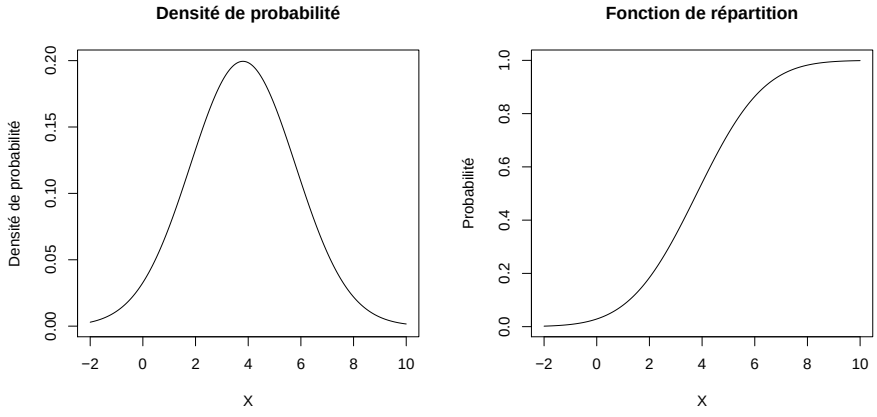


FIG. 3 : Représentation de la loi normale de paramêtre $\mu = 3,8$ et $\sigma = 2$.

3.2 Espérance et moments d'une variable aléatoire

Une variable aléatoire, X, peut éventuellement être partiellement caractérisée par son espérance qu'on note $\mathbb{E}(X)$ et dont l'expression est intrinsèquement liée à la loi de la variable aléatoire. Intuitivement, l'espérance donne une information sur la *position* de la variable aléatoire.

Représentation de ce qu'est l'espérance

Soit, pour une variable aléatoire discrète, c'est-à-dire une variable dont les valeurs éventuelles peuvent être indicées par un nombre entier $i \in \mathbb{N}$:

$$\mathbb{E}(X) = \sum_{i \in \mathbb{N}} x_i \, p(x_i)$$

Dans le cas d'une variable aléatoire continue, les valeurs prises par la variable ne peuvent pas être indicées par un entier et la somme discrète doit être remplacée par une somme intégrale :

$$\mathbb{E}(X) = \int_{\mathbb{R}} x \, p(x) \, \mathrm{d}x$$

Par extension on définit l'espérance d'une fonction d'une variable aléatoire. Si g est une fonction de $E \to \mathbb{R}$, $x \mapsto g(x)$, alors l'espérance de la fonction g, suivant la loi de probabilité de la variable X est donnée, respectivement pour une variable discrète ou continue, par :

$$\mathbb{E}_X(g) = \sum_{i \in \mathbb{N}} g(x_i) \, p(x_i) \ \textit{(variable discrète)}$$

$$\mathbb{E}_X(g) = \int_{x \in E} g(x) \, p(x) \, \mathrm{d}x \ \textit{(variable continue)}$$

Une famille particulière de fonction nous intéresse, c'est la famille des fonctions de puissance, $x \mapsto x^n$. Pour ces fonctions, on définit les moments, ainsi que les moments centrés. Le moment d'ordre n de la variable X est l'espérance de la fonction g définie par $x \mapsto |x|^n$, ainsi l'espérance d'une variable est, par définition, son moment d'ordre 1. Le moment centré d'ordre n est l'espérance de la fonction définie par $x \mapsto |x - \mathbb{E}(X)|^n$. D'une façon générale, le terme *centré* veut dire qu'on modifie les valeurs prises par la variable pour qu'elles soient centrées sur 0.

NB Dans cette section, on a défini l'espérance d'une variable aléatoire, ou celle d'une fonction d'une variable aléatoire sans détailler dans quels cas cette espérance existe, c'est-à-dire peut être calculée, ou pas. Dans sa généralité, la question de l'existence de l'espérance est assez complexe et sort totalement du cadre de cet ouvrage. Certaines variables aléatoires conceptuellement simples n'ont pas pour autant d'espérance ; un exemple est celui d'une expérience consistant à tirer au hasard un entier avec une probabilité uniforme. Toutefois il y a deux réserves à noter :

- Si l'expérience de notre exemple est conceptuellement simple, elle est en revanche pratiquement irréalisable. Une expérience physiquement finie ne pourra pas aboutir à une issue infiniment grande, ce qui revient à borner la valeur de la variable aléatoire, auquel cas l'espérance pourra être calculée ;

- Concernant les lois de probabilité exploitées dans le cadre de l'analyse statistique, il est rare que l'espérance ne puisse pas être calculée.

3.2.1 Cas particulier du moment centré d'ordre 2 : la variance

Le moment centré d'ordre 2 d'une variable aléatoire, comme son espérance, tient un rôle particulièrement important : c'est lui qui est le plus souvent utilisé pour caractériser l'échelle de variabilité de la variable aléatoire autour de son espérance. L'importance du moment centré d'ordre 2 justifie qu'on lui donne un nom spécifique, il s'agit de la variance*, et pour une variable aléatoire X on note sa variance $var(X)$. Suivant que la loi de probabilité de la variable X est discrète ou continue, sa variance est donnée, respectivement par :

$$\mathrm{Var}(X) = \sum_{i \in \mathbb{N}} (x - \mathbb{E}(X))^2 \, p(x_i) \text{ (\textit{variable discrète})}$$

$$\mathrm{Var}(X) = \int_{x \in \mathbb{R}} (x - \mathbb{E}(X))^2 \, p(x) \, \mathrm{d}x \text{ (\textit{variable continue})}$$

On est, là aussi, dans la même situation que celle de l'espérance : la variance n'est pas toujours définie ; même si la grande majorité des variables aléatoires utilisées en statistique peuvent tout de même être caractérisées par leur espérance et leur variance. Là où l'espérance informe sur la *position* de la variable aléatoire, sa variance et son écart-type renseigne sur son *échelle*.

Pour des raisons qui apparaitront plus clairement dans la section suivante, il est souvent plus expressif d'utiliser la racine carrée de la variance, plutôt que la variance elle-même. Ainsi on définit l'écart-type d'une variable aléatoire X comme la racine carrée de sa variance et la lettre grecque σ est utilisée pour sa notation mathématique :

$$\sigma(X) = \sqrt{\mathrm{Var}(X)}$$

3.2.2 Quelques propriétés de l'espérance et de la variance

Soit une variable aléatoire X $(\Omega \to \mathbb{R}, \omega \mapsto X(\omega))$ dont l'espérance $\mathbb{E}(X)$ existe, et deux nombres réels quelconques α et β, alors la fonction $\Omega \to \mathbb{R}, \omega \mapsto \alpha\, X(\omega) + \beta$ est aussi une variable aléatoire (qu'on notera $\alpha\, X + \beta$) et, de plus, son espérance existe et vaut $\mathbb{E}(\alpha\, X + \beta) = \alpha\ \mathbb{E}(X) + \beta$.

Cette propriété permet de mieux comprendre l'utilité de la version centrée d'une variable. En prenant $\alpha = 1$ et $\beta = -\mathbb{E}(X)$ on a $\mathbb{E}(X - \mathbb{E}(X)) = 0$, c'est-à-dire que l'espérance de la variable centrée est nulle.

Si X admet une variance, on a alors : $\mathrm{Var}(\alpha\, X + \beta) = \alpha^2\ \mathrm{Var}(X)$, et son écart type vérifie donc $\sigma(\alpha\, X + \beta) = \alpha\ \sigma(X)$. On voit dans cette dernière expression l'intérêt de l'écart-type, l'expression l'utilisant est plus simple et elle montre que si la variable X porte une unité (par exemple il s'agit d'une distance en mètre) alors $\sigma(X)$ porte exactement la même unité, ce qui en rend l'exploitation plus intuitive.

Cette propriété permet de définir la version *centrée réduite* d'une variable aléatoire. Si X est une variable aléatoire, la variable $Y = \frac{X-\mathbb{E}(X)}{\sigma(X)}$ est la représentation centrée réduite de X avec les propriétés suivantes : $\mathbb{E}(Y) = 0$ et $\mathrm{Var}(Y) = 1$.

Soit deux variables aléatoires, $X_1 : \Omega_1 \to \mathbb{R}$, $\omega_1 \mapsto X_1(\omega_1)$ et $X_2 : \Omega_2 \to \mathbb{R}$, $\omega_2 \mapsto X_2(\omega_2)$, alors la fonction $X : (\Omega_1 \times \Omega_2) \to \mathbb{R}$, $(\omega_1, \omega_2) \mapsto X_1(\omega_1) + X_2(\omega_2)$ est une variable aléatoire (dont la réalisation nécessite la réalisation conjointe de X_1 et X_2) qu'on note $X = X_1 + X_2$. Si, de plus, les espérances de X_1 et de X_2 sont définies, alors on peut calculer l'espérance de X et on a la relation $\mathbb{E}(X_1 + X_2) = \mathbb{E}(X_1) + \mathbb{E}(X_2)$.

Si, de plus, les variables X_1 et X_2 sont indépendantes, notion qui sera explicitée *infra*, on a alors la relation suivante sur la variance : $\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2)$, que l'on peut aussi exprimer en termes d'écart-type par : $\sigma(X_1 + X_2) = \sqrt{\mathrm{Var}(X_1) + \mathrm{Var}(X_2)} = \sqrt{\sigma(X_1)^2 + \sigma(X_2)^2}$.

4 Variables indépendantes et identiquement distribuées

Il convient maintenant de préciser les liens existant entre une variable statistique et sa modélisation par une variable aléatoire. La loi d'une variable aléatoire décrit l'incertitude sur la mesure, ou l'observation, avant que cette mesure ou cette observation ne soit faite. La statistique, elle, ne s'intéresse pas à une seule des mesures (ou observations), mais à un ensemble cohérent de celles-ci ; de plus, une fois l'observation effectuée, il n'y a plus d'incertitude sur la valeur prise par la mesure.

Chaque observation, chaque mesure correspond à une expérience, dont la valeur est une réalisation. Ainsi chacune est modélisée par une variable aléatoire. Si l'incertitude liée à chacune des mesures est la même (avant de réaliser la mesure et de connaître la valeur de celle-ci), les variables aléatoires liées aux mesures sont dites identiquement distribuées, c'est-à-dire qu'elles partagent toute une seule et même loi de probabilité. Cette loi de probabilité représente alors ce qu'on connaît *a priori* des incertitudes, c'est donc une généralisation des valeurs déjà mesurées.

Si, de plus, il n'y pas de dépendance entre deux mesures de la variable statistique, alors les variables aléatoires liées à chacune des mesures sont indépendantes, c'est-à-dire que la réalisation d'une variable aléatoire n'affecte pas la réalisation d'une autre variable aléatoire.

Quand ces deux conditions sont réalisées conjointement, les variables aléatoires correspondantes à chacune des mesures de la série statistique sont *indépendantes et identiquement distribuées*, ce qui est abrégé par *iid*.

4.1 Loi des grands nombres

La *loi des grands nombres* est un résultat de la théorie des probabilités qui permet, à nouveau, de relier les notions de variable aléatoire et de variable statistique. Cette loi relie la moyenne arithmétique d'une série de variables aléatoires *iid* à l'espérance de ces variables.

Soit une série infinie de variables aléatoires *iid* (cette série peut, en particulier, représenter les valeurs observées pour une même variable statistique) $(X_1, X_2, ...)$ alors

la moyenne arithmétique des n premières variables est elle-même une variable aléatoire qu'on notera $\bar{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} \frac{X_i}{n}$. La loi des grands nombres s'intéresse alors à la relation qui existe entre la moyenne, $\bar{X_n}$, et l'espérance d'une seule variable, $\mathbb{E}(X_1)$, (comme les variables sont *iid*, elles ont toutes la même espérance, $\mathbb{E}(X_1) = \mathbb{E}(X_2) = ... = \mathbb{E}(X_n)$).

Les propriétés de l'espérance et de la variance vue *supra* permettent déjà d'établir que pour toute valeur de n, aussi grande soit-elle, on a les deux propriétés suivantes :

$$\mathbb{E}(\bar{X_n}) = \frac{1}{n}\, n\, \mathbb{E}(X_1) = \mathbb{E}(X_1)$$

$$\mathrm{Var}(\bar{X_n}) = \frac{1}{n^2}\, n\, \mathrm{Var}(X_1) = \frac{\mathrm{Var}(X_1)}{n}$$

$$\sigma(\bar{X_n}) = \frac{\sigma(X_1)}{\sqrt{n}}$$

La loi des grands nombres va plus loin, en comparant directement la moyenne (et non son espérance) à l'espérance d'une variable. Cette loi existe sous plusieurs formes dont les variations dépassent le cadre du présent ouvrage, et nous allons nous limiter à la *loi forte des grands nombres*. La loi forte des grands nombres stipule que la moyenne de variables aléatoires *iid* converge presque sûrement vers l'espérance de chacune des variables quand le nombre de variables croît vers l'infini. Mathématiquement, cela s'écrit :

$$\mathbb{P}\left(\lim_{n\to\infty} \bar{X_n} = \mathbb{E}(X_1)\right) = 1$$

NB On parle de convergence *presque sûre* et non pas *simplement* de *sûr*, car quand l'ensemble des possibles est **infini**, certaines issues peuvent ne pas arriver même si leur probabilité est égale à 1. Il est plus facile de comprendre ce résultat par sa réciproque : une expérience aléatoire ayant une infinité d'issues possibles peut avoir une issue de probabilité nulle. Nous vous proposons l'exemple suivant. Supposons que la question la plus importante de l'univers ait un entier comme réponse, et que c'est là la seule information qu'on ait par rapport à cette réponse, on modélise l'incertitude sur la réponse par une variable aléatoire entière avec une loi de probabilité uniforme[10]. La probabilité que la réponse ne soit pas 42 est égale à 1 et, pourtant, c'est bien 42 la bonne réponse [11].

[10] Cette loi est sans doute un peu compliqué à apréhender. Comme pour une variable continue, la probabilité d'une seule valeur, ou même d'un nombre fini de valeurs, est nulle. En revanche des ensembles infinis d'entiers ont une probabilité non nulle. Par exemple la $\mathbb{P}(\mathrm{N} \text{ est paire}) = 0,5$

[11] *cf. Le Guide du voyageur galactique* de Douglas Adam, 1979.

4.2 Théorème central limite (loi Normale)

Si on veut caractériser encore plus finement la variable aléatoire $\bar{X_n}$, on peut s'interesser à sa loi de probabilité. Les résultats exposés dans la précédente section nous indiquent déjà que cette loi va être de plus en plus serrée autour de l'espérance de X_1. Pour affiner le résultat, on s'interesse alors à la version *centrée réduite* de $\bar{X_n}$, d'autant qu'on a déjà des résultats concernant l'espérance et la variance de $\bar{X_n}$. On s'interesse donc à la variable

$$\bar{Y_n} = \frac{\bar{X_n} - \mathbb{E}(\bar{X_n})}{\sigma(\bar{X_n})}$$

On peut également définir $\bar{Y_n}$ directement à partir de l'espérance et de l'écart-type de X_1 :

$$\bar{Y_n} = \frac{\sqrt{n}}{\sigma(X_1)} \left(\bar{X_n} - \mathbb{E}(X_1)\right)$$

Le théorème central limite, sous des hypothèses le plus souvent vérifiées, nous donne alors le résultat suivant. Pour tout interval $[a, b] \subset \mathbb{R}$, on a le résultat suivant

$$\mathbb{P}(\bar{Y_n} \in [a, b]) \xrightarrow[n\to\infty]{} \int_{x=a}^{b} \phi(x)dx$$

avec $\phi(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})$, c'est-à-dire la densité de probabilité d'une loi Normale centrée en 0 et d'écart-type égale à 1.

On peut également lire ce résultat ainsi : pour tout interval $[a, b]$ inclu dans $\mathbb{R}$, la probabilité que $\bar{Y_n}$ soit dans $[a, b]$ tend, quand n tend vers l'infini, vers la probabilité qu'une variable suivant la loi Normale centrée en 0 et de variance 1 soit dans ce même interval.

NB Le théorème central limite est *central* de par le rôle central qu'il tient dans la théorie des probabilité.

Ce résultat explique pourquoi la loi Normale permet de bien approximer un grand nombre de loi de probabilité.

5 Deux variables aléatoires (et plus...)

Prenons deux variables aléatoires, X et Y. Nous avons exposé *supra*, en définissant la notion de variables *iid*, que ces deux variables aléatoires sont indépendantes si la réalisation de X n'est aucunement affectée par la réalisation de Y et réciproquement. Quand, au contraire, la réalisation de X affecte la réalisation de Y (ou que la réalisation de Y affecte celle de X), on parle de variables aléatoires dépendantes ou encore de variables aléatoires liées. Autrement dit, si la connaisance de la réalisation de X (ou

Y) modifie l'incertitude sur les valeurs que peut prendre Y (ou X) alors les variables X et Y ne sont pas indépendantes.

Prenons un exemple : nous disposons de 5 boules blanches et 5 boules noires disposées dans deux urnes, l'une blanche et l'autre noire. L'urne blanche contient 4 boules blanches et 1 noire, l'urne noire contient 1 boule blanche et 4 boules noires ; expérience représentée en figure 4. Demandons à quelqu'un de prendre au hasard une boule : nous ne connaissons ni la couleur de la boule ni celle de l'urne sélectionnée par l'opérateur. Si X est la variable aléatoire représentant la couleur de l'urne et Y la couleur de la boule, alors les variables X et Y ne sont pas indépendantes. En effet, si on ne connait pas la réalisation de X, c'est-à-dire si on ne sait pas de quelle urne la boule a été tirée, la variable Y - la couleur de la boule tirée - a autant de chance d'être noire que blanche ($\mathbb{P}(Y = \text{blanc}) = 0,5$). En revanche si on connait la réalisation de X -on connait l'urne d'où provient la boule-, l'incertitude sur Y change : si, par exemple, c'est l'urne blanche, il y a une probabilité de $\frac{4}{5}$ que la boule soit blanche ($\mathbb{P}(Y = \text{blanc}) = 0,8$). Réciproquement, tant qu'on ne connait pas la couleur de la boule, il y a 1 chance sur 2 que l'urne blanche ait été choisie ($\mathbb{P}(X = \text{blanc}) = 0,5$ et $0,5$ que ce soit la noire), mais si on sait, par exemple, que la boule tirée est blanche ($Y = \text{blanc}$) alors ils est beaucoup plus probable que ce soit l'urne blanche ($X = \text{blanc}$) qui ait été sélectionnée par l'opérateur (nous verrons plus bas le calcul précis de ces probabilités).

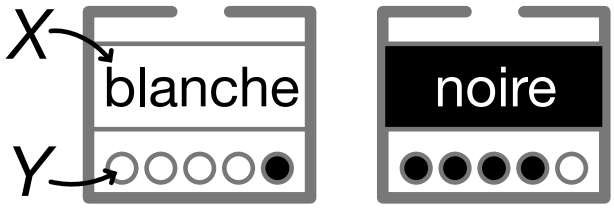


FIG. 4 : Schéma de l'expérience aléatoire proposée comme exemple de deux variables liées. X est la variable aléatoire représentant la couleur de l'urne, Y celle correspondant à la couleur de la boule.

5.1 Loi de probabilité jointe

Comme les variables ne sont pas indépendantes, il est naturel de caractériser l'incertitude en ce qui concerne la réalisation *simultanée* des deux variables, c'est à dire d'attribuer une probabilité aux valeurs à la fois des deux variables aléatoires. C'est ce qu'on appelle la loi de probabilité jointe, aussi appelée loi de probabilité conjointe, des deux variables. La caractérisation d'une paire de variables aléatoires qui ne sont pas indépendantes suit le même processus que celui que nous avons exposé en section 3. Si les deux variables, X et Y, sont discrètes, on peut caractériser l'incertitude des valeurs des variables à travers la distribution de probabilité jointe de X et Y qui donne la probabilité d'observer à la fois une valeur précise de X et une valeur précise de Y :

$$p(x_i, y_j) = \mathbb{P}(X = x_i, Y = y_j)$$

où x_i et y_j correspondent aux valeurs précises prises respectivement par X et Y.

Dans le cas où X et Y sont des variable aléatoires continues on définit la densité de probabilité conjointe à travers une limite, comme est définie la densité de probabilité d'une seule variable continue, également en section 3 :

$$p(x,y) = \lim_{\substack{\epsilon \to 0 \\ \epsilon' \to 0}} \frac{\mathbb{P}(X \in]x-\epsilon; x+\epsilon[, Y \in]y-\epsilon'; y+\epsilon'[)}{2\epsilon\epsilon'}$$

5.2 Probabilité marginale et probabilité conditionnelle

L'utilisation de deux variables aléatoires liées pour caractériser l'incertitude n'interdit pas de ne s'intéresser qu'à une seule des deux variables. Dans l'exemple ci-dessus, c'est ce qu'on se propose de faire quand on cherche la probabilité de tirer une boule blanche (ou une noire), sans rien connaitre de l'urne selectionnée. Autrement dit, on s'intéresse à ne caractériser que l'incertitude sur la variable Y. La loi de probabilité caractérisant cette incertitude s'appelle la **loi de probabilité marginale** de la variable (ici Y), et s'écrit comme la loi d'une seule variable, si besoin indexée par la lettre de la variable à laquelle on s'interesse ; dans cet exemple on écrirait p_Y définie par $p_Y(y_j) = \mathbb{P}(Y = y_j)$.

On peut également s'intéresser à l'incertitude liée à l'une des deux variables une fois que l'autre a été déterminée. C'est ce qui est proposé à la fin de l'exemple de la figure 4, quand on cherche à caractériser l'incertitude de la couleur de l'urne (représentée par la variable X) une fois que la couleur de la boule (la valeur prise par la variable Y) est connue. Cette incertitude est caractérisée par une loi de probabilité qu'on appelle **loi de probabilité conditionnelle** de la variable X. On utilise la barre verticale, |, pour séparer ce qui est encore incertain, à gauche, de ce qui est déjà connu, à droite. Ainsi on note $p(\text{urne blanche} \,|\, \text{boule blanche})$ la probabilité que l'urne sélectionnée soit blanche une fois qu'on sait que la boule qui a finalement été tirée est elle-même blanche. La signification de cette barre verticale est étendue à l'ensemble des notations :

$$p(x_i \,|\, y_i) = \mathbb{P}(X = x_i \,|\, Y = y_i)$$

En notant E_X et E_Y les valeurs prises respectivement par X et Y, ces lois de probabilité ont les caractérisques suivantes :

- Pour les lois marginales : $\sum_{x_i \in E_X} p_X(x_i) = 1$ et $\sum_{y_i \in E_Y} p_Y(y_i) = 1$
- Pour les lois conditionnelles : $\forall y_j \in E_Y, \sum_{x_i \in E_X} p(x_i \,|\, y_j) = 1$ [12] et réciproquement, $\forall x_i \in E_X, \sum_{y_j \in E_Y} p(y_j \,|\, X = x_i) = 1$.

De plus, ces lois de probabilités conditionnelles et marginales sont liées à la loi jointe par les relations :

[12] On lit : quelque soit la valeur y_j, la somme des probabilités conditionnelles sachant que $Y = y_j$ vaut 1.

$$p_X(x_i) = \sum_{y_j \in E_Y} p(x_i, y_j)^{13} \ (\textit{marginalisation}) \, ; \tag{1.1}$$

$$p(x_i, y_j) = p_Y(y_j) \ \times \ p(x_i \,|\, Y = y_j)^{14} \tag{1.2}$$

Les propriétés exprimées ci-dessus pour des variables discrètes sont également vérifiées pour des variables continue, une fois les formules adaptées en utilisant une sommation continue (intégrale). Il est également possible d'étendre ces résultats à plus de deux variables aléatoires avec éventuellement un mélange de variables discrètes et continues. Ainsi n variables aléatoires, $X_1, ..., X_n$, peuvent être étudiées conjointement, auquel cas il devient utile de noter $\mathbf{X} = (X_1, ..., X_n)$ le vecteur aléatoire de dimension n aggrégeant l'ensemble des variables en une seule notation.

Revenons à l'exemple de la figure 4, nous pouvons maintenant la décrire à l'aide des fonctions de probabilité marginales, conditionnelles et jointes. D'après le processus de tirage aléatoire décrit, un premier choix est fait à égale probabilité entre les deux urnes. Autrement dit, la probabilité marginale de X est uniforme : $p_X(\text{urne blanche}) = p_X(\text{urne noire}) = \frac{1}{2}$. Une fois l'urne choisie, le choix de la boule est elle aussi uniforme. Il s'agit là de l'expression de la probabilité du tirage de la boule (variable Y) conditionnellement au choix de l'urne : si l'urne blanche a été choisie $p(\text{boule blanche} \,|\, \text{urne blanche}) = \frac{4}{5}$ et $p(\text{boule noire} \,|\, \text{urne blanche}) = \frac{1}{5}$ [15]. Cette probabilité conditionnelle permet de répondre à la deuxième question posée en début de section (probabilité de la couleur de la boule une fois connue la couleur de l'urne). On peut maintenant calculer la probabilité jointe à partir de l'équation (1.2) : $p(x, y) = p_X(x) \ \times \ p(y \,|\, x)$, ce qui permet d'obtenir les probabilités jointes présentées dans le tableau 1.

Tab. 1 : Calcul des probabilités jointes correspondant à l'expérience décrite par la figure 4. Une fois les probabilité jointes calculées, on peut compléter le tableau par la dernière ligne donnant les probabilités marginales de la seconde variable en utilisant l'équation (1.1).

urne (X) \ boule (Y)	blanche	noire	p_X
blanche	$0,5 \times 0,8 = 0,4$	$0,5 \times 0,2 = 0,1$	$0,5$
noire	$0,5 \times 0,2 = 0,1$	$0,5 \times 0,8 = 0,4$	$0,5$
p_Y	$0,4 + 0,1 = 0,5$	$0,1 + 0,4 = 0,5$	

6 Loi de Bayes

La loi de Bayes permet de formaliser la troisième proposition de notre exemple : « si on sait, par exemple, que la boule tirée est blanche, alors ils est beaucoup plus probable que ce soit l'urne blanche (X = blanc) qui ait été sélectionnée par l'opérateur ». Cette proposition exprime la densité conditionnelle $p(\text{urne blanche} \,|\, \text{boule blanche})$.

[13] La probabilité marginale de x_i est égale à la somme pour tous les y_j de la probabilité jointe de x_i et y_j. On dit également que la probabilité marginale de x_i est obtenue par marginalisation de la variable Y au sein de la probabilité jointe.

[14] Et réciproquement on a $p(x_i, y_j) = p_X(x_i) \ \times \ p(y_j \,|\, X = x_i)$

[15] Si c'est l'urne noire qui est choisie $p(\text{boule blanche} \,|\, \text{urne noire}) = \frac{1}{5}$ et $p(\text{boule noire} \,|\, \text{urne noire}) = \frac{4}{5}$.

En utilisant l'equation (1.2) on peut exprimer la probabilité jointe de deux façons distinctes qui mettent en relation les deux probabilités conditionnelles :

$$p(x,y) = p_X(x) \times p(y \,|\, x) = p_Y(y) \times p(x \,|\, y)$$
$$p(x \,|\, y) = \frac{p_X(x) \times p(y \,|\, x)}{p_Y(y)} \tag{1.3}$$

À l'aide de cette formule et des valeurs de probabilités marginales calculées au tableau 1, on peut calculer la probabilité pour que l'urne choisie soit blanche si la seule information connue est que la boule est blanche $p(\text{urne blanche} \,|\, \text{boule blanche}) = \frac{0{,}5 \times 0{,}8}{0{,}5} = 0,8$.

Dans la formule de Bayes (1.3), le dénominateur ($p_Y(y)$) est constant car il ne dépend que de y dont la valeur est exactement connue. Ainsi on peut comparer des probabilités conditionnelles pour une même condition sans même avoir à calculer la probabilité marginale $p_Y(y)$, si bien qu'on peut considérer que ce dénominateur est essentiellement un terme d'échelle qui permet de satisfaire la propriété $\sum_{x_i \in E_X} p(x_i \,|\, y) = 1$ vue *supra*.

7 Statistique Bayésienne et vraisemblance

Jusqu'à maintenant nous avons proposé d'utiliser les outils de la théorie des probabilité pour caractériser l'incertitude liée à la réalisation d'expériences aléatoires, c'est-à-dire d'expériences dont le résultat n'est, au moins en partie, pas strictement défini. La **statistique Bayésienne** propose d'étendre l'utilisation des outils de la probabilité pour décrire l'incertitude existante sur la valeur des paramètres d'un modèle.

Revenons à l'exemple de la section 2.1 où nous avons réalisé une série de mesure de la vitesse du son en mesurant la distance parcourue par le son pendant $5s$. La distance parcourue est toujours la même, en revanche la mesure elle-même introduit des effets aléatoires (délais lors de la production ou de la détection du son... ce qu'on appelle communément *erreur de mesure* ou *bruit de mesure*) si bien que la distance mesurée a une petite composante aléatoire introduisant une incertitude sur la valeur de la vitesse qu'on extrait de chaque mesure. La proposition de la théorie Bayesienne est d'utiliser les outils de la probabilité pour modéliser l'incertitude de notre *connaissance* de la vitesse du son, même si celle-ci n'est intrinséquement pas aléatoire.

Pour mettre cela en équation : en notant v la vitesse du son (en m/s), la distance (en m), notée d, parcourue en $5s$ est $d = 5 \times v$, et par conséquent $v = d/5$. Cette distance n'a rien d'aléatoire[16], en revanche la distance mesurée est entachée d'une erreur qui est aléatoire et c'est donc une variable aléatoire : $D = d + E = 5 \times v + E$ où E représente l'erreur aléatoire de mesure. Dans l'équation $D = 5 \times v + E$, D et E sont des variables aléatoires et v est un paramètre : il a une valeur unique et toujours la

[16] Il en va de même pour la vitesse qui n'a rien d'aléatoire non plus. C'est pour différencier ces valeurs *vraies* que nous utilisons la graphie d et v ; là où d et v représenterons une réalisation des variables aléatoires D et V.

même si on réalise plusieurs fois l'expérience et même si cette valeur est inconnue. Une série de mesures de D peut alors permettre de proposer une valeur approchée de v, on parle alors d'inférence* (ou encore d'estimation) et la statistique Bayésienne propose de modéliser l'incertitude restante sur la *vraie valeur* de v (c'est-à-dire, la différence entre la vraie valeur de v et sa valeur approchée) à l'aide d'une distribution de probabilité, *comme si* v était une variable aléatoire, et on utilise donc la variable aléatoire V pour représenter l'état de notre connaissance (exprimer l'incertitude le plus précisément possible) sur la valeur de v. Dans ce cadre, la probabilité $p(d, v)$ est la densité de probabilité jointe que la distance mesurée soit d et que la vitesse estimée du son soit v et l'ensemble des probabilités conditionnelles et marginales correspondent aux définitions *supra*. Parmi l'ensemble de ces distributions de probabilité, la densité $p(d \,|\, v)$ est le plus souvent la mieux caractérisée, en effet :

$$p(d \,|\, v) = \mathbb{P}(E = d - 5 \times v)$$

Il s'agit de la densité de probabilité de l'erreur de mesure qui est caractéristique du dispositif instrumental et qui est souvent connue. En revanche, c'est la probabilité conditionnelle $p(v \,|\, d)$ qui représente l'incertitude restante sur la valeur de v une fois exploitées les mesures de D. C'est la loi de Bayes (1.3) qui va permettre de la calculer.

$$p(v \,|\, d) = \frac{p_V(v) \times p(d \,|\, v)}{p_D(d)} \tag{1.4}$$

Cette equation est centrale pour la statistique Bayésienne[17], si bien que chaque terme de cette équation a un nom, à l'exception du dénominateur $p_D(d)$ qui, comme nous l'avons vu précédemment, est constant et correspond essentiellement à une normalisation.

$p(d \,|\, v)$: fonction de vraisemblance ; elle exprime la probabilité d'obtenir la mesure effectuée dans l'hypothèse où la valeur prise pour le paramètre est bien celle considéré (ici v) ;

$p_V(v)$: probabilité *a priori* ; elle exprime l'incertitude existante sur la valeur du paramètre (ici v) avant même que la mesure ne soit faite ;

$p(v \,|\, d)$: probabilité *a posteriori* ; elle caractérise ce qui reste d'incertitude sur la valeur du paramètre (ici v) une fois prise en compte la mesure effectuée.

Ce formalisme permet donc de combiner les informations provenant de multipes expériences (et sources) (*cf.* 1.2, 1.4, 1.5 et 12). De plus, si les informations sont indépendantes, la probabilité *a posteriori* obtenue sur la base d'une information peut être utilisée comme probabilité *a priori* pour l'exploitation des informations pas encore exploitées . Par ailleurs, la probabilité *a posteriori* est une représentation plus

[17] Et c'est parce que la loi de Bayes est si centrale dans cette demarche qu'on parle de statistique Bayésienne

complète de l'état de notre connaissance sur la valeur du paramètre, par rapport à une valeur approchée même si celle-ci est assortie d'une barre d'erreur ou même d'un intervalle de confiance.