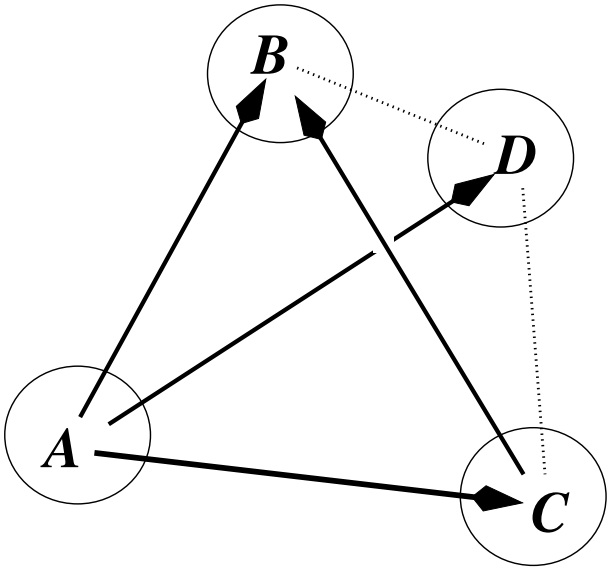
Introduction Descriptive aux Réseaux Bayésiens

Jean-Baptiste DENIS[1]

Pierre PARDON



Rapport technique 2005-5, 52 pp.
Unité Mathématiques et Informatique Appliquées
INRA
Domaine de Vilvert
F-78352 Jouy-en-Josas Cedex
[1] Jean-Baptiste.Denis@Jouy.Inra.Fr

Table des matières

Table des figures

Liste des tableaux

TAB. 1 – Six différentes manières de pratiquer une interprétation de données.
En lignes sont les deux types d'approche que nous avons distingués (cf. §10.6). En colonnes, les types de modèle utilisé (cf. §10.5). Les six possibilités sont envisageables ; notre préférence va à celle du bas à droite.

	modèles univariables	modèles multivariables	réseaux bayésiens
statistique classique			
statistique bayésienne			*

1 Introduction

Dans cette section, nous positionnons et limitons le sujet de ce rapport.

A l'occasion du projet **MAM**[1], nous nous proposons de modéliser le fonctionnement[2] d'un abattoir d'un point de vue salubrité et sécurité des aliments à l'aide d'un réseau bayésien. Avant de nous lancer dans une modélisation complexe, il nous a paru bon de préciser entre nous les concepts basiques de cette approche. Cette note est le résultat de nos divers échanges et réflexions sur le sujet. Elle ne présente pas de vraie originalité, mais elle rassemble des idées qu'on trouve généralement éparses dans la littérature, tentant un exposé simple et détaillé de l'essentiel. Cette présentation se fait au travers d'exemples pour appuyer la réflexion sur des situations concrètes et pour mieux dégager les idées des aspects techniques.

Il est important de souligner dès le départ que ce document ne concerne que la description des réseaux bayésiens. Le problème majeur relatif à l'estimation de leurs paramètres n'est pas abordé.

Certains termes spécialisés mériteraient d'être proprement définis[3]. Nous renvoyons les lecteurs aux nombreux ouvrages sur le sujet, ou encore aux présentations didactiques que l'on peut trouver sur internet : une contribution se trouve à partir de :

`<http ://www.inra.fr/miaj/public/matrisq/jbdenis/notes/welcome.html>`

Mais peut-être faut-il commencer par positionner les **réseaux bayésiens** dans le paysage de l'usage des statistiques[4]. Ils sont en effet souvent assimilés aux approches bayésiennes de la statistique, ce qui nous semble une erreur car il s'agit avant tout d'une modélisation.

Les modélisations statistiques peuvent être classées de diverses façons ; nous en retiendrons deux pour notre propos :

- Le type de modélisation probabiliste : nous distinguerons les modèles univariables, les modèles multivariables, les réseaux bayésiens (cf. §10.5).
- Le type d'approche statistique : nous distinguerons l'approche fréquentiste de la l'approche bayésienne (cf. §10.6).

Le tableau 1 met en lumière le croisement des deux classifications envisagées.

Dans notre contexte, un modèle est une représentation simplifiée, relativement abstraite, parfaitement formalisée d'un processus ou d'un système, en vue de le décrire, de l'expliquer, de le prédire et de faciliter une analyse décisionnelle. Il repose sur un ensemble d'hypothèses qui permettent une représentation simplifiée ; il est souvent bâti autour de paramètres qui sont des variables mathématiques qui lui donnent une certaine

[1] Microbiologie Abattoir Modélisation
[2] dans une optique de maîtrise de process
[3] Dans la mesure du possible, nous essayons de définir les termes techniques lors de leur première utilisation.
[4] Et sans doute aussi de l'intelligence artificielle.

souplesse pour s'adapter à une grand nombre de situation[5] ; enfin le modèle peut incorporer les incertitudes que l'observateur a du système qu'il étudie au travers d'observations. Les modèles probabilistes (ou stochastiques) sont des modèles basés sur des variables aléatoires (précisées par des distributions de probabilité). Quelques rappels sont donnés en §9. Nous appelons une approche statistique, une procédure d'extractation d'information à partir de données à l'aide d'un modèle. Le plus souvent, il s'agit de préciser la valeur de paramètres du modèle grâce aux données disponibles.

Dès à présent, il faut qu'il soit bien clair dans l'esprit du lecteur que nous nous intéressons aux réseaux bayésiens qui ne sont rien d'autre qu'un moyen de construire des modèles probabilistes. Nous ne faisons pas de statistiques, nous ne nous intéressons pas à interpréter des données, ce que permet une approche bayésienne. Ceci dit, il est tout à fait possible de traiter les données adjointes à un réseau bayésien par une approche bayésienne, c'est d'ailleurs ce que nous choisirions.

En §2, nous donnons quelques éléments de vocabulaire et des pistes la démarche générale, le lecteur qui ne connaît pas encore les réseaux bayésiens ne doit pas s'attarder sur les développements qui y sont faits, ils s'éclaireront aux cours des exemples qui suivent. La §3, le réseau bayésien le plus élémentaire est détaillé, cela permet de bien comprendre un fonctionnement basé sur des distributions de probabilité qui se retrouve au coeur de tout réseau bayésien. Les §4, §5 et §6 présentent tout autant en détail les trois réseaux de trois noeuds envisageables. Ceci est fait en variant le type de variables pour généraliser les cas. La §7 est en devenir. La §8 brosse un petit panorama des logiciels disponibles que nous avons repérés. Les deux dernières sections sont des annexes que nous avons considéré utiles pour le sujet, soit pour préciser des notions autour des probabilités (§9) qui sont au coeur de la définition des réseaux bayésiens, soit pour diverses considérations annexes (§10) que le lecteur peut ignorer sans regret.

N.B.1 :Le niveau actuel de ce document, toujours en cours d'élaboration, reste très hétérogène et nous avons le projet d'améliorer la situation. Nous recommandons donc à nos éventuels lecteurs de ne pas hésiter à sauter les paragraphes qui leur sembleraient ésotériques. Nous serions bien entendu très heureux de recueillir toutes leurs réactions, en particulier contestataires.

N.B.2 : Pour faciliter la lecture, nous avons pris soin de distinguer deux niveaux de difficulté indiqués par des tailles de caractères différents. En caractères plus gros est reproduit le niveau élémentaire du document, principalement intuitif, sa lecture est nécessaire. En caractères plus petits se trouvent des développements plus précis, généralement introduits avec un formalisme mathématique, le lecteur peut s'en dispenser sans crainte de perdre les concepts principaux. Egalement, ont été ainsi typographiées les illustrations plus techniques qui demandent un certain temps pour être bien appréhendées. Le lecteur peut les aborder sans les comprendre en détail s'il accepte de nous faire confiance.

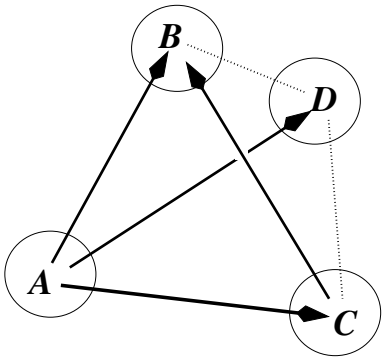
2 Qu'est-ce qu'un réseau bayésien ?

Les réseaux bayésiens doivent leur nom à l'utilisation d'un réseau où les variables d'intérêt sont disposées[6] et aux réflexions de Thomas Bayes autour des probabilités conditionnelles. Si on cherche dans la littérature des informations sur les réseaux bayésiens, on peut trouver : *Les réseaux bayésiens ont été utilisés dans les années 80 pour gérer les incertitudes dans les systèmes expert à base de règles. Ils ont l'avantage de permettre une représentation compacte de certaines probabilités, une visualisation aisée des interactions entre variables, et enfin l'obtention d'une loi de probabilité conjointe des variables à partir de sources hétérogènes (experts et/ou bases de données).* La propriété essentielle des réseaux bayésiens est de lier les informations portées par les différentes variables d'un système. Leur utilisation consiste à répercuter au travers

[5]Dans notre cas, les paramètres seront principalement employés pour spécifier des distributions de probabilité.

[6]La terminologie mathématique est celle de graphe, cf. plus loin.

FIG. 1 – Réseau bayésien simple



du fonctionnement du réseau les informations nouvelles de certaines variables mieux connues sur l'ensemble des autres variables. Notre but dans ce document est de décrire les réseaux bayésiens de façon simple mais cependant précise, à partir d'exemples. Nous sommes en effet persuadés qu'ils représentent une clef majeure dans la mise en place efficace de modélisation de phénomènes complexes.

2.1 quelques définitions

Avant de présenter des exemples, qui éclaireront la notion, il convient d'introduire de manière plus formelle ce que nous entendons par réseaux bayésiens. **Les réseaux bayésiens sont des modèles graphiques associés à des relations probabilistes liant les variables caractéristiques du système étudié.** Il s'agit d'un ensemble de variables aléatoires dénommées **noeuds du réseau**, reliées entre elles par des relations précisées dénommées **arcs du réseau** qui constituent le graphe du réseau. Les arêtes sont orientées : les noeuds au départ des flèches sont les parents des noeuds à la pointe de la flèche (qui sont dénommés logiquement leurs enfants). Le graphe ne doit pas comporter de boucles, c'est-à-dire qu'on ne doit pas pouvoir partir d'un noeud et y revenir en suivant le sens des flèches du graphe. Ce type de graphe est dénommé graphe acyclique orienté (*directed acyclic graph = dag*, en anglais). Toutes les paires de noeuds ne sont pas reliées, ce qui donne des dag aux caractéristiques différentes. La figure 1 propose un réseau bayésien comportant 4 noeuds (A, B, C, D). Les flèches indiquent que A influence directement B, C et D, également que C influence directement B. On ne trouve aucun cycle en suivant les flèches.

On trouve dans le liste suivante trois notions capitales qui correspondent à trois étapes successives d'élaboration du réseau bayésien :

1. les **variables** : ils s'agit de caractéristiques le plus souvent quantitatives (discrètes ou continues) du système qu'on veut modéliser pour un objectif donné. La plupart seront considérées aléatoires. Nous les noterons en majuscules, les valeurs qu'elles prennent[7] seront généralement notées par la minuscule correspondante.
2. le **graphe** : l'ensemble des flèches qui relient les variables précédemment définies. L'existence et l'orientation des flèches traduisent les relations directes exercées entre les variables et le sens dans lequel on va les considérer : la variable aval (enfant) va être définie en fonction des variables amont (parents).

[7] que l'on nomme réalisations.

3. le **modèle** : chaque variable orpheline (sans parent) est définie par une valeur ou une distribution de valeurs ; de même chaque variable enfant est définie (aléatoire ou fixe) en fonction de ses parents. De cette manière, une distribution conjointe est finalement définie sur l'ensemble des variables du système : c'est la modélisation probabiliste.

Le réseau le plus simple est bien entendu celui qui ne comporte qu'un noeud ! Cependant, il n'est pas intéressant car aucune propriété ne peut être associée aux relations entre les noeuds puisqu'elles n'existent pas ! Nous décrirons en détail des graphes comportant deux ou trois noeuds. Notons que si système à modéliser est complexe et nécessite de nombreux noeuds, il est souvent adroit de le découper en sous-systèmes de sorte que les relations intra et inter sous-systèmes soient des réseaux bayésiens indépendants. Le résultat global reste cependant un réseau bayésien[8].

Remarquons que l'interdiction des cycles empêche une relation à double sens entre deux noeuds, même si on serait tenté de la considérer comme interprétable. Par exemple, la taille de l'estomac influe sur la sensation de satiété, mais inversement, à plus long terme, la satiété influe sur la taille de l'estomac. Dans une modélisation par réseau bayésien, il nous faut choisir quel est le sens que nous privilégions. Par commodité, mais abus d'interprétation, nous qualifierons parfois le noeud parent de cause, et le noeud enfant de conséquence. Il est important d'ajouter que l'absence de cycle n'est en rien une restriction de la modélisation probabiliste ; au contraire leur présence empêcherait l'établissement de la distribution conjointe de l'ensemble des variables du réseau.

Détaillons un peu en l'utilisant l'exemple de la taille de l'estomac (notée E) et de l'appétit (noté A). On peut choisir de définir le modèle selon la figure 2 (i) ou (ii), c'est-à-dire en échangeant les rôles de parent et enfant entre les deux noeuds du réseau. Si A est parent, il faudra spécifier sa distribution marginale et la distribution de E conditionnellement à A (notée $E \mid A$). Si E est parent, il faudra spécifier sa distribution et celle de $A \mid E$. Cette réversibilité est liée au fait qu'en modélisation probabiliste et approches statistiques, on ne décrit pas des causalités mais des covariations[9] (cf. §9.6). L'interprétation en terme d'influence est ajoutée subjectivement. Ces notions s'éclairciront au fil des exemples traités.

2.2 démarche générale de construction d'un réseau bayésien

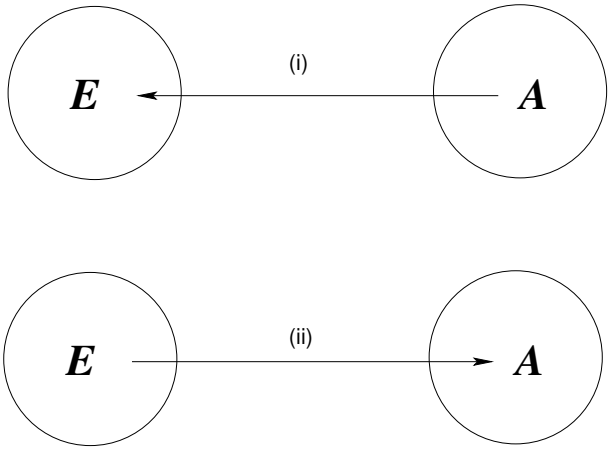
Il est sans doute important de préciser dès maintenant la démarche suivie, sans l'écran du calcul des différentes probabilités. Cette démarche peut s'appliquer à tous les types de graphes, même si certaines étapes sont plus ou moins court-circuitées pour aller plus directement au résultat.

1. **Choix et définition des variables** qui constituent les noeuds du réseau. Certaines d'entre elles peuvent être *cachées*, c'est-à-dire ne pas être observables ; d'autres seront qualifiées de *paramètres* car elles spécifient les distributions des autres variables.
2. **Choix et définition du graphe** des relations, c'est-à-dire des relations directes entre les noeuds proposés à l'étape précédente. Les flèches entre les noeuds précisent la manière dont nous savons spécifier le passage d'information d'un noeud à l'autre de manière cohérente.

[8] L'intérêt de cette modularisation dépend bien entendu de la pertinence (du point de vue du réseau ce qui devrait correspondre à celui du système étudié) du découpage opéré et de la simplification possible des liaisons entre sous-systèmes (qui devraient se définir sur un nombre très réduit de flèches).

[9] Cette notion devrait être formalisée, mais intuitivement nous entendons par covariation, la part commune de variation qui existe entre deux variables : plus deux variables varient simultanément plus la covariation est importante.

FIG. 2 – Deux modèles équivalents

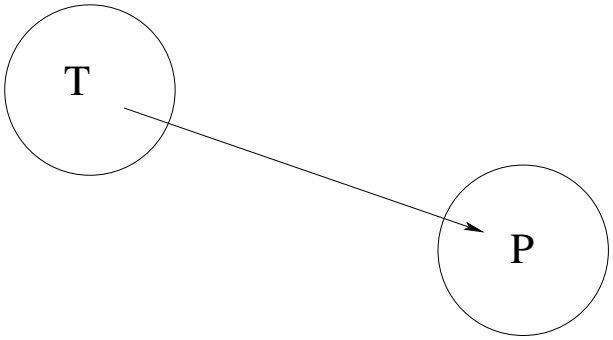


3. **Définition de chaque relation** : c'est indiquer mathématiquement la nature des relations directes. Il peut s'agir d'une relation fonctionnelle, ou d'une relation probabiliste. Si aucune des relations n'est probabiliste, le réseau se réduit à quelques distributions indépendantes et ne justifie pas vraiment le qualificatif de réseau bayésien, car aucun conditionnement ne peut intervenir. **Dans une optique de statistique bayésienne, on distingue les *a priori* qui sont les distributions portées sur les variables cachées et la *vraisemblance* qui correspondent aux distributions des variables observées.**
4. **Calcul d'une probabilité conjointe**. Les définitions précédentes doivent permettre d'achever une distribution de probabilité conjointe de l'ensemble des noeuds du réseau. Utilisant de manière répétée le théorème de Bayes, elle s'obtient simplement par le produit des probabilités de l'ensemble des noeuds. Il n'est pas toujours utile de l'expliciter, mais il est important d'en saisir l'existence.
5. **Calcul d'une probabilité conjointe conditionnelle**. C'est tenir compte du fait qu'on a observé les valeurs prises par certains noeuds du réseau (on dénomme classiquement ces variables des données). Pour tenir compte de cette information nouvelle, on restreind la distribution conjointe des variables non observées à celles qui sont cohérentes avec les observations : c'est l'opération de conditionnement par application du théorème de Bayes. Dans une optique de statistique bayésienne, il s'agit du calcul général de l'*a posteriori*.
6. **Calcul de probabilités marginales**. Il s'agit, pour tirer des conclusions à partir de certaines variables non observées, de calculer les distributions d'un (sous-ensemble de) noeud(s), en oubliant les autres. Attention, si on s'intéresse à deux paramètres simultanément, considérer leurs lois marginales n'est pas rigoureux ; il faut regarder leur loi conjointe, car ils peuvent être très liés.

3 Maillon élémentaire : une histoire de poids

Adoptons les étapes de la démarche générale qui vient d'être présentée et appliquons la à un maillon élémentaire constitué de deux noeuds et d'une flèche entre ces noeuds (figure 3).

FIG. 3 – Réseau bayésien réduit à un maillon élémentaire



3.1 définition des variables

Suppons que nous nous intéressions au poids à la naissance de bébés (notés P) d'une région et d'une époque précisées ; admettons aussi que nous connaissions la taille de ces mêmes bébés (notée T) et que nous nous posions la question d'inférer[10] sur le poids connaissant la taille. Si cette situation n'est peut-être pas très intéressante en soi car peser un bébé n'est pas une opération difficile ou dangereuse, elle représente cependant les nombreux cas de figure où la variable d'intérêt est malaisée à observer et qu'on cherche à lui substituer un indicateur le plus efficace possible.

Le fond du raisonnement est que les deux variables, poids et taille du bébé, sont corrélées ; si donc on connaît seulement la taille d'un bébé, on dispose d'une information utile. Par exemple, les très petits bébés ne pèseront pas lourds (en général), au contraire des très grands bébés (en général).

3.2 définition du réseau (graphe et relations)

Sachant que nous voulons utiliser la variable T pour préciser la variable P, nous allons placer entre les deux noeuds une flèche qui part de T pour arriver sur P, pour traduire le fait que la variable T apporte de l'information sur P. Cf. Fig. 3. Dit autrement, la loi conditionnelle de P sachant T, notée $[P \mid T]$ est différente[11] de la loi de P en elle-même (sans connaissance autre), dite loi marginale.

3.3 cas de variables aléatoires binaires

Par variable aléatoire binaire, on entend une variable aléatoire dont les valeurs possibles sont deux états différents : par exemple $\{0, 1\}$, $\{1, 2\}$, {en bonne santé, malade}, {rouge, vert}, {fille, garçon},... La distribution d'une variable binaire est complètement spécifiée par sa probabilité de se trouver dans un des deux états (l'autre étant bien entendu le complémentaire à 1). On indicera ces variables par le b de binaire : P_b et T_b.

[10] Pour répondre à des questions précisées, l'inférence statistique passe par la modélisation probabiliste des observations disponibles.

[11] Et normalement plus précise.

TAB. 2 – Probabilité conditionnelle de P_b sachant T_b

$P(P_b \mid T_b)$	$P_b = 1$	$P_b = 2$
$T_b = 1$	0.95	0.05
$T_b = 2$	0.56	0.44

3.3.1 définition des variables

Supposons, pour commencer simplement, que les deux variables de notre modèle soient binaires, c'est-à-dire qu'elles ne prennent que deux valeurs symboliques 1 et 2 :

- $T_b = 1$, le bébé est dans la catégorie des petits bébés ; $T_b = 2$, il ne l'est pas.
- $P_b = 1$, le bébé est dans la catégorie des bébés légers ; $P_b = 2$, il ne l'est pas.

3.3.2 influence de T_b sur P_b ($P_b \mid T_b$)

Dans une optique probabiliste, on peut traduire l'influence de T_b sur P_b par une table de probabilités conditionnelles (cf. §9.3) : une proposition est faite dans le tableau 2 . Il s'agit d'une table de deux lignes (associées aux états possibles de la variable T_b) et deux colonnes (associées aux états de la variables P_b). Elle contient 4 probabilités qui s'interprètent de la manière suivante. 0.56 est la probabilité[12] d'avoir un bébé léger alors que l'on sait que le bébé est grand. Très logiquement, l'événement complémentaire $P_b = 2$ sachant $T_b = 2$ est $1 - 0.56 = 0.44$. On note ces quatre probabilités de la manière suivante :

$$P(P_b = 1|T_b = 1) = 0.95$$
$$P(P_b = 2|T_b = 1) = 0.05$$
$$P(P_b = 1|T_b = 2) = 0.56$$
$$P(P_b = 2|T_b = 2) = 0.44$$

Ces valeurs semblent raisonnables dans le sens où elles traduisent bien la relation positive entre les deux variables : la proportion de poids élevés est beaucoup plus importante pour les grandes tailles que pour les petites.

On notera que la somme des cases d'une ligne de la table vaut 1, conséquence du fait qu'on y envisage les probabilités d'événements exclusifs et décrivant l'ensemble des possibles (cf. §9.2). Ceci n'est pas vrai par colonne puisqu'il s'agit des probabilités associées à une même valeur de P dans des circonstances différentes : les probabilités d'avoir un bébé petit selon qu'il est petit ou pas n'ont pas d'autre relation qu'une inégalité à respecter : $P(P_b = 2 \mid T_b = 1) < P(P_b = 1 \mid T_b = 1)$... encore n'est-elle dictée que par le bon sens et n'est pas une contrainte mathématique, et elle pourrait ne pas être vraie pour certaines définitions de variables binaires. Autrement dit, chaque ligne du tableau 2 porte sur deux (sous-) populations disjointes, chacune pouvant donc se définir indépendamment de l'autre. C'est la raison pour laquelle la somme des probabilités par ligne égale 1 ; ce serait 100, si on avait choisi de les exprimer en pourcentages. On aurait pu préférer représenter le tableau 2 comme la juxtaposition des deux probabilités présentées en tableau 3

[12] Un point capital est de savoir comment la valeur 0.56 a été déterminée. Dans cette note, nous ne nous posons pas la question et supposons "tout" connaître !

TAB. 3 – Probabilité conditionnelle de P_b sachant T_b décomposée

$P(P_b\|T_b=1)$	$P_b=1$	$P_b=2$		$P(P_b\|T_b=2)$	$P_b=1$	$P_b=2$
	0.95	0.05			0.56	0.44

TAB. 4 – Importance des probabilités marginales
Deux répartitions d'effectifs très différents peuvent engendrer la même probabilité conditionnelle de $P(P_b \mid T_b)$

Effectifs	$P=1$	$P=2$		Effectifs	$P=1$	$P=2$
$T=1$	95	5		$T=1$	950	50
$T=2$	560	440		$T=2$	56	44

3.3.3 inversion de la relation : influence de P_b sur T_b ($T_b \mid P_b$)

De manière complètement similaire, on pourrait chercher à traduire l'influence du poids sur la taille. Naïvement, on pourrait estimer qu'il suffit de "retourner" la relation précédente. C'est vrai, mais on ne peut pas y parvenir avec les seules probabilités de la table 2 ! Essayons de voir pourquoi.

Insistons de nouveau, cette table $P(P_b|T_b)$ donne comme seule information les deux proportions, chacune correspondant à une ligne du tableau 2 : les proportions de bébés légers parmi les petits (95%), et parmi les grands (56%). Les proportions de bébés lourds sont bien entendu les complémentaires à 100%, soit 5% et 44%. Les proportions réciproques (celles des bébés petits parmi les légers et parmi les lourds) demandent, pour être établies, la connaissance complète de la répartion des bébés suivant les 4 catégories. En effet, comme le démontre les deux répartitions d'effectifs compatibles présentées en table 4, la même probabilité conditionnelle peut-être s'appliquer à des situations très différentes.

Autrement dit, il s'agit d'effectifs de deux populations très dissemblabes dans les répartitions mais construites de telle manière qu'elles induisent la même distribution conditionnelle $P(P_b \mid T_b)$... par contre, on en obtient des probabilités conditionnelles inverses, $P(T_b|P_b)$, très différentes. Dans le premier cas, pour les bébés légers, il y a moins de bébés petits que de bébés grands, mais c'est l'inverse dans le second cas, alors que les proportions de la table 3 sont exactement respectées dans un cas comme dans l'autre. La clef de cet apparent paradoxe[13] se trouve dans les proportions quasiment inverses des petits et grands bébés qui affectent les proportions des bébés légers et des bébés lourds...

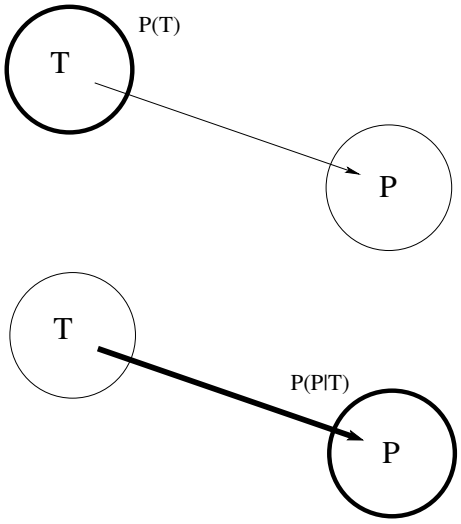
Ceci peut se voir par application directe (répétée) du théorème de Bayes[14] (cf. §9.4). En

[13] Il s'agit du paradoxe de Simpson.

[14] Par exemple :

$$\begin{aligned} P(T_b=1 \mid P_b=1) &= \frac{P(T_b=1, P_b=1)}{P(P_b=1)} \\ &= \frac{P(T_b=1, P_b=1)}{P(T_b=1, P_b=1)+P(T_b=2, P_b=1)} \\ &= \frac{P(P_b=1 \mid T_b=1)\,P(T_b=1)}{P(P_b=1 \mid T_b=1)\,P(T_b=1)+P(P_b=1 \mid T_b=2)\,P(T_b=2)} \end{aligned}$$

FIG. 4 – Spécification du modèle.
Deux distributions de probabilité sont associées à des deux sous-ensembles du réseau (mis en gras).



TAB. 5 – Distribution marginale des tailles (variable binaire)

$P(T_b)$	
$T_b = 1$	0.480
$T_b = 2$	0.420

supplément aux probabilités conditionnelles $P(P_b \mid T_b)$, il faut donc disposer de la distribution marginale de la variable aléatoire T_b.

3.3.4 spécification complète du modèle

En fait, la distribution conditionnelle $P(P_b \mid T_b)$ et la distribution marginale de la variable conditionnante $P(T_b)$ nous fournit la distribution conjointe des deux variables, notée $P(P_b, T_b)$ ou $P(T_b, P_b)$. De cette distribution conjointe se déduisent toutes les autres, en particulier l'autre distribution conditionnelle $P(T_b \mid P_b)$, et bien entendu l'autre distribution marginale $P(P_b)$. La spécification complète du modèle peut se faire de trois manières différentes :

- $P(P_b, T_b)$,
- $P(P_b \mid T_b)$ et $P(T_b)$,
- $P(T_b \mid P_b)$ et $P(P_b)$.

Cette règle est générale et s'applique à tous types de variables binaires, discrètes et continues ou mixtes. Les sous-parties du réseau correspondant au choix de la spécification du modèle sont illustrées dans la figure 4.

Il nous faut donc ajouter la distribution marginale de T_b qui est donnée dans la table 5. Il est alors possible d'en déduire la probabilité conjointe, c'est la table . Il est simple ensuite de trouver à partir de la densité conjointe la probabilité conditionnelle de T_b sachant P_b. Il suffit de normaliser les colonnes (chaque colonne correspond à une valeur fixée de T_b) pour en faire une probabilité. Le résultat en est donné dans la table 7 après échange des lignes et des colonnes

TAB. 6 – Probabilité conjointe de T_b et P_b

$P(T_b, P_b)$	$P_b = 1$	$P_b = 2$
$T_b = 1$	0.456	0.024
$T_b = 2$	0.291	0.229

TAB. 7 – Probabilité conditionnelle de T_b sachant P_b.

$P(T_b \mid P_b)$	$T_b = 1$	$T_b = 2$
$P_b = 1$	0.610	0.390
$P_b = 2$	0.095	0.0905

pour que la variable conditionnante, P_b, se trouve en lignes.

3.4 cas de variables aléatoires discrètes

Les variables aléatoires discrètes généralisent les variables aléatoires binaires en autorisant plus de deux états possibles. Autrement dit, une variable binaire n'est qu'une variable discrète particulière. On indicera ces variables par le d de discrète : P_d et T_d.

3.4.1 définition des variables et du réseau

On pourrait par exemple distinguer 3 niveaux de tailles des bébés : petits (1), moyens (2) et grands (3) ; et aussi distinguer 4 niveaux de poids des bébés : très légers (a), légers (b), lourds (c) et très lourds (d). Pour éviter les confusions, on notera les deux variables dans leur version discrète T_d et P_d. Nous conservons, bien sûr, le graphe utilisé avec les variables binaires (figure 3).

3.4.2 influence de T_d sur P_d ($P_d \mid T_d$)

Comme nous l'avons fait en §3.3.2, dans une optique probabiliste, on peut traduire l'influence de T_d sur P_d par une table de probabilités conditionnelles (cf. §9.3) : une proposition est faite dans le tableau 8 .

Il s'agit d'une généralisation de la table 2 obtenue en éclatant la variable T sur sa première modalité ($1 \rightarrow \{0, 1\}$) et les deux modalités de la variable P ($1 \rightarrow \{a, b, c\}, 2 \rightarrow \{d\}$).

TAB. 8 – Probabilité conditionnelle de P_d sachant T_d
(cas de variables discrètes)

$P(P_d \mid T_d)$	$P_d = a$	$P_d = b$	$P_d = c$	$P_d = d$
$T_d = 0$	0.736	0.214	0.044	0.005
$T_d = 1$	0.272	0.399	0.252	0.077
$T_d = 2$	0.027	0.169	0.363	0.440

TAB. 9 – Distribution marginale de la variable T_d.

$P(T_d)$	
$T_d = 0$	0.182
$T_d = 1$	0.298
$T_d = 2$	0.520

TAB. 10 – Probabilité conjointe de P_d et T_d
La somme de toutes ces probabilités est l'unité.

$P(P_d, T_d)$	$P_d = a$	$P_d = b$	$P_d = c$	$P_d = d$
$T_d = 0$	0.134	0.039	0.008	0.001
$T_d = 1$	0.081	0.119	0.075	0.023
$T_d = 2$	0.014	0.088	0.189	0.229

Par exemple, on vérifie, à partir des tableaux 2 et 8, que :

$$\begin{aligned} P(P_d = d \mid T_d = 2) &= 0.440 \\ &= P(P_b = 2 \mid T_b = 2) \end{aligned}$$

3.4.3 calcul de la distribution conjointe de P_d et T_d

La problématique est identique à celle identifiée dans le cas des variables binaires (cf. §3.3.3), il faut aussi disposer de la distribution marginale de la variable T_d. Supposons que celle-ci soit donnée par le tableau 9 . On peut en déduire par application similaire du théorème de Bayes la table 11 des probabilités conditionnelles de T_d sachant P_d.

Pour cela, on commence par calculer la probabilité conjointe par le théorème de Bayes avec des formules du genre : $P(P_d = b, T_d = 1) = P(P_d = b, \mid T_d = 1)\, P(T_d = 1) = 0.399 \times 0.298 = 0.1189...$ Finalement, on obtient la table 10.

3.4.4 inversion de la relation : influence de P_d sur T_d $(T_d \mid P_d)$

A partir de la distribution conjointe, on déduit assez directement[15] la distribution conditionnelle de T_d par rapport à P_d, notée $[T_d \mid P_d]$. Celle-ci est donnée dans la table 11 des probabilités conditionnelles de T_d sachant P_d.

3.4.5 remarque : calculs sur ordinateur

Tous les calculs précédents sont en général réalisés grâce à des logiciels spécialisés (cf. §8) d'autant plus nécessaires que suivant les modèles utilisés, les calculs peuvent devenir complètement impensables à la main.

[15] Il suffit de diviser chaque probabilité par la somme des probabilités de sa ligne. Par exemple $0.604 = \frac{0.134}{0.134+0.039+0.008+0.001}$.

TAB. 11 – Probabilité conditionnelle de T sachant P
cas de variables discrètes

$P(T_d \mid P_d)$	$T_d = 0$	$T_d = 1$	$T_d = 2$
$P_d = a$	0.585	0.354	0.061
$P_d = b$	0.159	0.484	0.358
$P_d = c$	0.029	0.276	0.695
$P_d = d$	0.004	0.091	0.905

3.5 cas de variables aléatoires continues

Même si la réalité observée est par nature toujours discrète, la précision avec laquelle une taille et un poids sont mesurés rend naturel de considérer ces variables comme des réels mathématiques (sans doute positives !). Nous les qualifierons de continues et les indicerons par c : P_c et T_c.

3.5.1 définition des variables et du réseau

On notera respectivement par T_c et P_c, les variables de taille (exprimée en millimètres) et de poids (exprimé en grammes). Le graphe de la figure 3 est conservé.

3.5.2 influence de T_c sur P_c ($P_c \mid T_c$)

Les distributions de probabilité associées à des variables de type continu ne s'expriment pas par des fréquences mais par des densités[16]. L'équivalent de nos tables conditionnelles sera donc une densité pour P_c qui dépendra de la valeur prise par T_c, de fait considérée comme covariable. Par exemple, qu'elle suive une distribution normale dont l'espérance est une fonction linéaire de T_c, $20T_c - 7000$, et l'écart-type une constante, 300. La distribution normale étant notée par N et pas ses deux paramètres d'espérance et de variance[17], on peut donc écrire :

$$P_c \mid T_c \quad \sim \quad N\left(20T_c - 7000,\ 300^2\right)$$

Ceci reflète bien une liaison croissante entre le poids et la taille : en moyenne le poids en g augmente 20 fois comme la taille en mm. On peut représenter ceci graphiquement (figure 5). Il y a une densité différente pour chaque valeur de T_c, la figure n'en propose que 3 pour les valeurs $T_c = \{490, 500, 550\}$. On remarque que la densité se déplace sur la droite pour des valeurs croissantes de T_c mais garde une forme identique.

3.5.3 inversion de la relation : influence de P_c sur T_c ($T_c \mid P_c$)

Là encore, il faut donner la marginale de T_c pour pouvoir inverser la relation. Si on suppose que $T \sim N(500, 20^2)$, alors on peut appliquer le théorème de Bayes.

$$[T_c \mid P_c] \quad = \quad \frac{[T_c, P_c]}{[P_c]} = \frac{[P_c \mid T_c]\,[T_c]}{[P_c]}$$

[16] La densité $f(x)$ d'une variable aléatoire X continue est la fonction positive telle que pour tout $a \leq b$, $P(X \in [a, b[) = \int_a^b f(x)dx$. Une façon de percevoir la densité d'une variable aléatoire est de la considérer comme l'histogramme infiniment précis des fréquences relatives de la variable aléatoire.

[17] La variance est le carré de l'écart-type.

FIG. 5 – Distributions conditionnelles du Poids
(pour 3 valeurs de Taille différentes dans le cas continu).

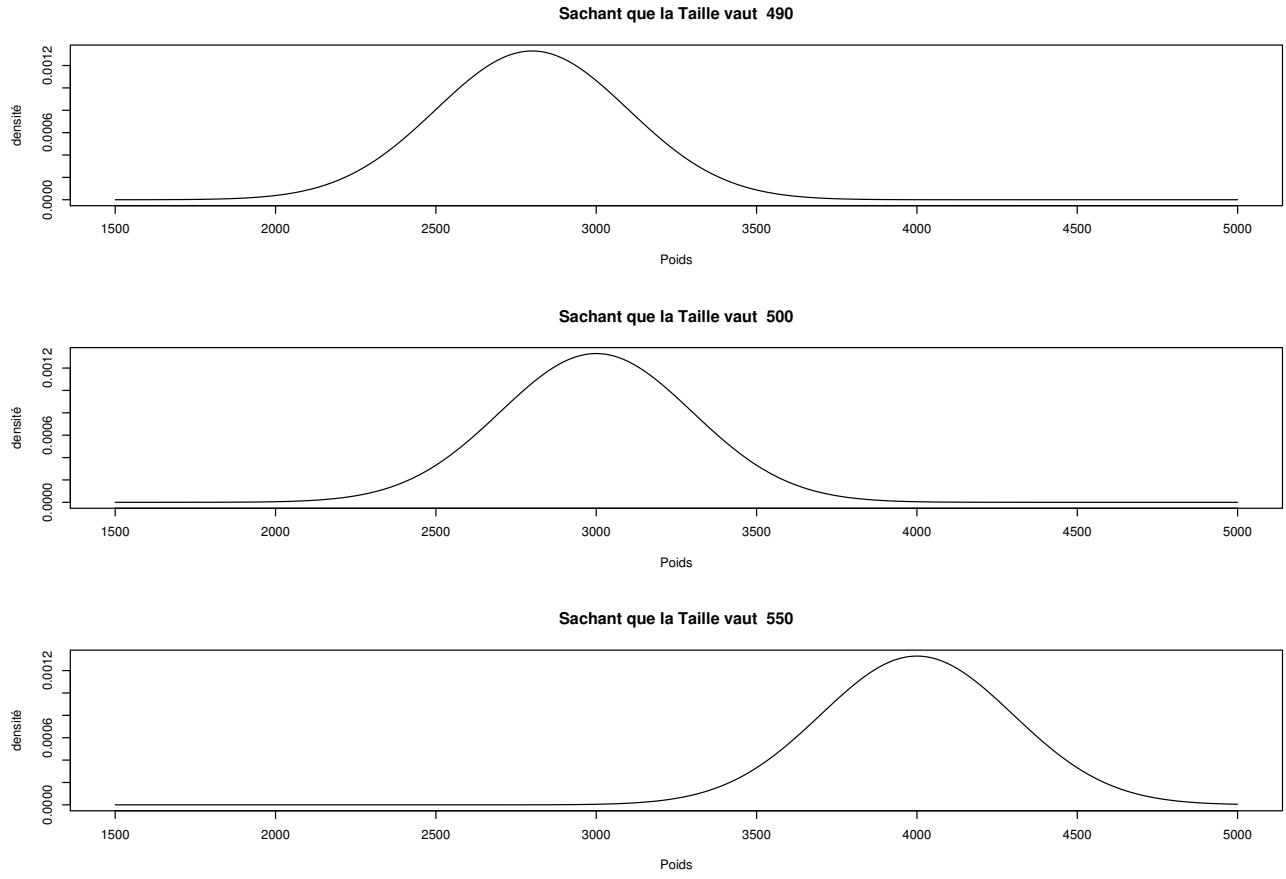
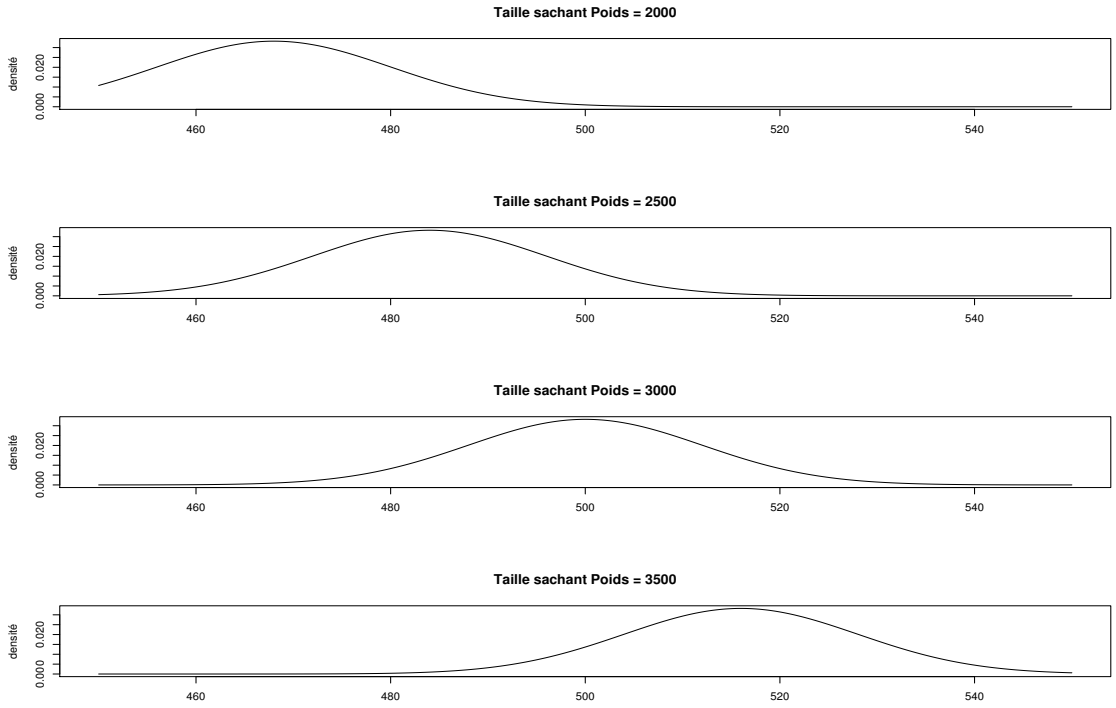


FIG. 6 – Distributions conditionnelles de la Taille
(pour 4 valeurs de Poids différents dans le cas continu)



L'explicitation des densités[18] aboutit à une distribution normale, dont l'espérance est une combinaison linéaire de P_c et l'écart-type est constant :

$$T_c \mid P_c \quad \sim \quad N\left(0.032P_c + 404, 12^2\right)$$

Ce qui est illustré par la figure 6 .

Il est intéressant de noter que la loi conjointe de (P_c, T_c), qui n'a pas été exprimée, est une binormale.

[18] Il faut commencer par calculer la densité conjointe dont on n'a besoin qu'à une constante multiplicative près :

$$\begin{aligned}
[T_c, P_c] &= [P_c \mid T_c]\,[T_c] \\
&= \phi_{P_c|T_c}\left(P, 20T - 7000, 300^2\right)\phi_{T_c}\left(T, 500, 20^2\right) \\
&= \frac{1}{300\sqrt{2\pi}}\exp\left(-\frac{\left(P - (20T - 7000)\right)^2}{2 \times 300^2}\right)\frac{1}{20\sqrt{2\pi}}\exp\left(-\frac{(T-500)^2}{2 \times 20^2}\right) \\
&= K \times \exp\left(-\frac{\left(T - (404 + 0.032P)\right)^2}{2 \times 12^2}\right)
\end{aligned}$$

Ensuite, on applique le résultat de proportionnalité de la loi conditionnelle en fixant P, la variable conditionnante. On reconnaît alors une loi normale ce qui démontre le résultat énoncé.

3.6 remarques

3.6.1 orientation du réseau

Dans notre présentation, les deux variables (T, P) ont joué un rôle différent car nous avions pris le parti de prédire P en fonction de T (cf. §3.1). Mais on aurait tout aussi bien pu prendre le parti inverse. De fait, les deux modélisations sont équivalentes. Intuitivement, on modélise la covariation conjointe de (T, P), il est tout à fait subjectif de prétendre qu'une variable influence l'autre. C'est un objectif que de vouloir "prédire" l'une en fonction de l'autre. La modélisation de la covariation est donnée par la distribution conjointe ; elle pourrait fort bien être construite à partir de $[T \mid P]$ et $[P]$ pour servir à prédire P à partir de T, c'est-à-dire à rechercher $[P \mid T]$. C'est ce qui a déjà été indiqué en §3.3.4. D'autres développements sur ce sujet sont menés en §10.2.1.

3.6.2 continuité n'implique pas normalité

Dans notre présentation lorsque les variables sont continues, nous utilisons des distributions normales mais ce n'est pas obligatoire. C'est plus facile car les résultats sont alors simples et favorisent la présentation. Il existe beaucoup d'autres densités de probabilité pour les variables continues, par exemple la distribution Bêta qui a la particularité de ne charger qu'un segment de la droite réelle[19] (par exemple $[0, 1]$).

4 Réseau à 3 noeuds divergent : fumer nuit à la santé

La situation envisagée dans cette section est celle d'un réseau bayésien de trois noeuds de type divergent (cf. la figure 7). Elle correspond à la modélisation d'une cause commune sur deux variables ; l'une des variables est facilement observable, mais c'est l'autre qui ne l'est pas (par exemple parce que trop coûteuse à obtenir) et qui est l'objectif. Le but est de tenter d'utiliser la variable accessible comme indicateur de celle qui ne l'est pas. Il s'agit donc d'une situation très commune aussi bien en recherche que dans la vie courante.

Supposons qu'on s'intéresse à la capacité respiratoire d'un individu ; cette capacité est notée R, pour **r**espiratoire. Admettons que cette capacité respiratoire est principalement déterminée par sa consommation quotidienne en cigarettes ; nous la noterons S, pour ***s**moke*[20]. Supposons enfin qu'un autre effet d'une consommation excessive de cigarettes soit l'occurrence d'une toux matinale que nous noterons T, pour **t**oux, qui pourrait jouer le rôle d'indicateur.

Pour juger de la capacité respiratoire d'un individu donné, le mieux serait de nous livrer à une expérience physiologique simple, mais admettons que ce ne soit pas possible. Un autre recours serait de la déduire de sa consommation de cigarettes journalières. Admettons aussi que ce ne soit pas possible non plus et que la seule information dont nous disposions soit relative à l'éventuelle toux qui le secoue au réveil.

L'information reste qualitative mais on a envie de penser que
- s'il ne tousse pas le matin, c'est qu'il ne doit pas beaucoup fumer,
- et que s'il ne fume pas beaucoup, il n'est pas impossible de penser qu'il ait une bonne capacité respiratoire.

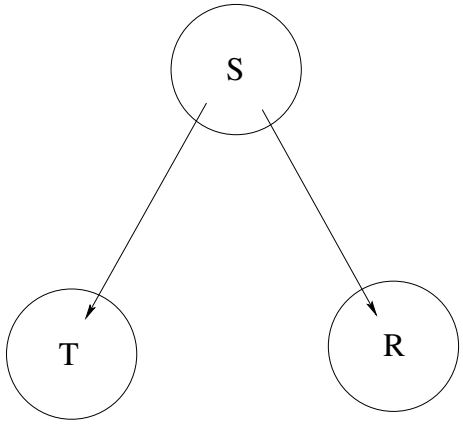
Ou encore que

[19] La droite réelle est l'ensemble des nombres réels de $-\infty$ à $+\infty$.

[20] Nous n'avons pas utilisé le F de **f**umer, craignant une confusion avec l'utilisation du symbole $False$ ou $Faux$ des variables binaires ou faible ou fort comme utilisé plus loin.

FIG. 7 – Réseau bayésien divergent.

Le fait de fumer (S) a une influence directe sur la toux matinale (T) et la capacité respiratoire (R).



- s'il tousse violemment le matin, il est envisageable que ce soit la conséquence d'une consommation excessive de l'herbe à Nicot,
- et que s'il fume beaucoup, sa capacité respiratoire soit amoindrie.

C'est ce genre de raisonnements que nous allons mettre sous forme quantitative en construisant un réseau bayésien de trois noeuds.

4.1 définition des variables

Supposons que les trois variables de notre modèle soient discrètes. Plus précisément :

- $S \in \{0, 1-5, 6-20, +20\}$[21], suivant la consommation quotidienne moyenne de cigarettes de l'individu par jour.
- $T = 0$, si l'individu ne tousse pas le matin et $T = 1$ dans le cas contraire ; c'est une variable binaire.
- $R \in \{f, m, F\}$, suivant que l'individu a une capacité respiratoire faible, moyenne ou Forte.

4.2 définition du réseau

Tentant de formaliser ce qui vient d'être admis, on peut considérer les liaisons entre les trois variables R, S, T traduites par le graphe présenté en figure 7.

Les liaisons entre les variables sont symbolisées par des flèches pour indiquer le sens de l'influence[22] : l'effet va de S vers R, et non l'inverse.

Ce graphe ne définit pas complètement un modèle car il n'indique que les relations directes existant entre les trois variables sans préciser les distributions de probabilité. Il permet cependant de raisonner un peu. Par exemple, pour prédire R, connaître (T, S) équivaut à connaître S tout seul. En effet, il n'y a pas de flèche directe entre R et T : toute l'information que T apporte sur R, c'est au travers de la connaissance qu'il apporte sur S. Dit autrement, si on connaît S, R n'apporte rien de plus ; on parle d'indépendance conditionnelle. Autre exemple issu du même genre de raisonnement : la prédiction de R par T est *moins bonne* que celle de

[21] C'est-à-dire que S appartient à l'une des quatre classes définies comme $\{0\}$, $\{1-5\}$, $\{6-20\}$ ou $\{+20\}$.

[22] en fait, cette présentation n'a rien de statistique (cf. §9.6)

TAB. 12 – Probabilité conditionnelle de R sachant S
Les valeurs en gras sont les modes de chaque distribution.

$P(R\|S)$	$R=f$	$R=m$	$R=F$
$S=0$	0.1	0.4	**0.5**
$S=1-5$	0.2	**0.5**	0.3
$S=6-20$	0.3	**0.5**	0.2
$S=+20$	**0.5**	0.4	0.1

R par S. Le terme *moins bonne* est à préciser mais il devrait l'être de manière cohérente avec cette proposition.

4.3 définition des relations

4.3.1 influence de S sur R ($R \mid S$)

On peut reprendre pour chaque relation, les raisonnements tenus dans le cas d'un maillon simple en §3.4. Dans une optique probabiliste, l'influence de S sur R se traduit par une table de probabilités conditionnelles (cf. §9.3) : une proposition est faite dans le tableau 12

Il s'agit d'une table de quatre lignes (associées aux états possibles de la variable S) et trois colonnes (associées aux états de la variables R). Chaque ligne propose la distribution de la variable R conditionnellement à une valeur de S. Par exemple, la probabilité d'avoir une capacité respiratoire faible sachant qu'on fume plus de 20 cigarettes par jour vaut 0.5 ($P(R=f \mid S=+20)$). On observera que les probabilités les plus élevées de chaque distribution sont placées grossièrement selon une diagonale du tableau 12 ; elles dessinent bien une relation inverse entre les deux variables, ce qui est raisonnable dans la mesure où si on fume, la capacité respiratoire risque d'être diminuée.

On notera que la somme des cases d'une ligne de la table vaut 1, conséquence du fait qu'on y envisage les probabilités d'événements exclusifs et décrivant l'ensemble des possibles (cf. §9.2). Ceci n'est pas vrai par colonne puisqu'il s'agit des probabilités associées à une même valeur de R dans des circonstances différentes : les probabilités d'avoir une capacité respiratoire faible selon qu'on est plus ou moins fumeur n'ont pas de relation obligatoire à respecter. Même si dans le cas présent, la nature du phénomène entraîne cependant une série croissante de probabilités pour la première colonne, et décroissante pour la dernière colonne.

4.3.2 influence de S sur T ($T \mid S$)

De manière complètement similaire, on peut traduire l'influence de la fumée sur la toux. Une proposition est faite dans le tableau 13 . Les mêmes remarques s'appliquent.

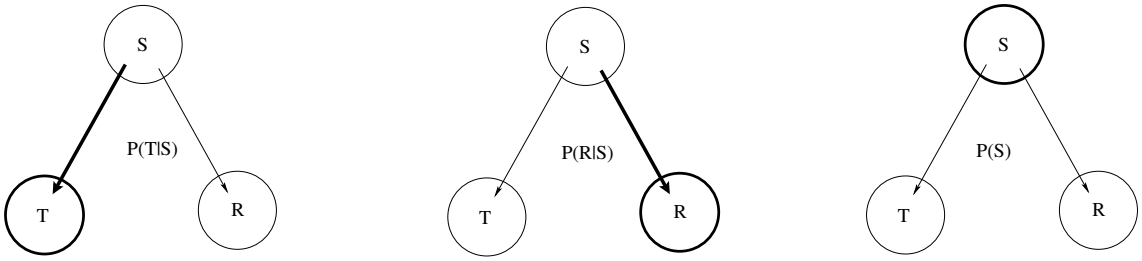
4.3.3 probabilité marginale de S

Pour définir complétement le réseau, il faut aussi donner la distribution de S. C'est ce qu'illustre la figure 8. Nous allons donc supposer aussi connaître la probabilité marginale de la variable S, consommation de tabac (tableau 14).

TAB. 13 – Probabilité conditionnelle de T sachant S
Les valeurs en gras sont les modes de chaque distribution.

$P(T\|S)$	$T=0$	$T=1$
$S=0$	**0.80**	0.20
$S=1-5$	**0.70**	0.30
$S=6-20$	**0.50**	**0.50**
$S=+20$	0.10	**0.90**

FIG. 8 – Association des distributions de probabilité d'un réseau divergent (aux différents parties du graphe : en gras)



TAB. 14 – Distribution marginale de S

	$P(S)$
$S=0$	0.40
$S=1-5$	0.20
$S=6-20$	0.20
$S=+20$	0.20

TAB. 15 – Distribution conjointe de S et T.

$P(T,S)$	$S=0$	$S=1-5$	$S=6-20$	$S=+20$
$T=0$	0.32	0.14	0.10	0.02
$T=1$	0.08	0.06	0.10	0.18

4.4 calcul de la probabilité conjointe

Dans le cas présent, nous n'allons pas expliciter la probabilité conjointe, mais cela pourrait se faire, car il est plus aisé de calculer directement la probabilité répondant à notre question sur la capacité respiratoire.

4.5 calcul de la probabilité conditionnelle

4.5.1 le problème posé

Nous venons d'établir par construction la relation $S \rightarrow T$ et la relation $S \rightarrow R$. Pour répondre à la question "*si je sais qu'un individu tousse, que puis-je dire de sa probabilité d'avoir une capacité respiratoire faible*", il faudrait que nous disposions de $T \rightarrow S$ et la relation $S \rightarrow T$ pour pouvoir imaginer les enchaîner. Il faut donc commencer par pratiquer l'inversion de $S \rightarrow T$ en $T \rightarrow S$; c'est l'opération déjà pratiquée en §3.4.4. Là encore, elle ne peut pas se faire avec la seule table de $P(T \mid S)$. En effet, la table $P(S|T)$ dépend de $P(T \mid S)$ et de $P(S)$ puisque le théorème de Bayes (cf. §9.4) indique que

$$P(S=i \mid T=j) \;=\; \frac{P(T=j \mid S=i)\,P(S=i)}{P(T=j)}. \qquad (1)$$

Pour s'en convaincre de nouveau, on peut aussi revenir au petit contre-exemple présenté dans le tableau 4 en §3.3.2.

4.5.2 information de T sur S, $(S \mid T)$

Bien que d'un point de vue modélisation ce soit complètement symétrique, il n'est pas opportun de libeller le titre de cette sous-section *influence de T sur S* puisque nous laisserions supposer que la toux exerce une influence sur la tendance à fumer, ce que nous ne croyons pas ! Par contre, on peut facilement admettre que la connaissance de la variable *toux* nous procure une information sur la variable *fumer*. Ce qui justifie le titre choisi.

Pour trouver les probabilités $P(S|T)$, il convient de calculer la probabilité conjointe $P(S,T)$. Pour cela, on applique la formule du théorème de Bayes (§9.4). Par exemple, en utilisant les valeurs des tableaux 13 et 14, $P(S=0,T=0) = P(S=0)\,P(T=0|S=0) = 0.4 \times 0.8 = 0.32$. On obtient finalement le tableau 15. On observera que cette fois-ci, c'est la somme de toute la table qui fait l'unité.

A partir de la probabilité conjointe, $P(S,T) = P(T|S)\,P(S)$, rien n'empêche de repasser aux lois conditionnelles qui nous intéressent. Par exemple si $T=0$, alors S prend la valeur 0 dans la proportion $0.32/(0.32+0.14+0.10+0.02) = 0.552$. C'est appliquer dans l'autre sens, $P(S,T) = P(S|T)\,P(T)$, le théorème de Bayes. Il en résulte le tableau 16 de $P(S \mid T)$. Cette fois-ci, par construction, les sommes par ligne font bien l'unité, signant une probabilité conditionnelle.

TAB. 16 – Distribution de S sachant T

$P(S \mid T)$	$S=0$	$S=1-5$	$S=6-20$	$S=+20$
$T=0$	0.552	0.241	0.172	0.034
$T=1$	0.190	0.143	0.238	0.429

TAB. 17 – Distribution de R sachant T.

$P(R \mid T)$	$R=f$	$R=m$	$R=F$
$T=0$	0.172	0.441	0.386
$T=1$	0.333	0.438	0.229

On pouvait bien entendu, appliquer directement la formule exprimée dans l'équation 1 de §4.5.1 ; elle conduit aux mêmes résultats.

4.5.3 information de T sur R, $(R \mid T)$

Pour atteindre notre but, c'est-à-dire trouver la table des probabilités $P(R \mid T)$, les deux tables $P(R \mid S)$ et $P(S \mid T)$, dont nous disposons maintenant, suffisent. En effet, par exemple, l'événement $(R=f \mid T=0)$ peut se décomposer en quatre événements disjoints :

$$\begin{aligned}(R=f, S=0 \mid T=0)\\ (R=f, S=1-5 \mid T=0)\\ (R=f, S=6-20 \mid T=0)\\ (R=f, S=+20 \mid T=0)\end{aligned}$$

En utilisant le théroème de Bayes, on peut écrire, par exemple, que

$$\begin{aligned}P(R=f, S=0 \mid T=0) &= P(R=f \mid S=0, T=0)\, P(S=0 \mid T=0)\\ &= P(R=f \mid S=0)\, P(S=0 \mid T=0)\end{aligned}$$

le passage de la première à la deuxième ligne est une conséquence du modèle posé : si on connait S, les informations sur T n'ont aucune influence. Donc il suffit d'additionner les probabilités suivantes (cf. tableaux 12 et 16) :

$$\begin{aligned}P(R=f \mid T=0) &= P(R=f \mid S=0)\, P(S=0 \mid T=0)\\ &+ P(R=f \mid S=1-5)\, P(S=1-5 \mid T=0)\\ &+ P(R=f \mid S=6-20)\, P(S=6-20 \mid T=0)\\ &+ P(R=f \mid S=+20)\, P(S=+20 \mid T=0)\\ &= 0.1 \times 0.552 + 0.2 \times 0.241 + 0.3 \times 0.172 + 0.5 \times 0.034\\ &= 0.172\end{aligned}$$

Globalement, il s'agit d'une multiplication matricielle (colonnes du tableau 12 par les lignes du tableau 16). Le résultat complet est donné par le tableau 17.

TAB. 18 – Distribution marginale de T.

$P(T)$	
$T=0$	0.58
$T=1$	0.42

TAB. 19 – Distribution marginale de R.

$P(R)$	$R=f$	$R=m$	$R=F$
	0.240	0.440	0.320

4.6 calcul des probabilités marginales

Le calcul des probabilités marginales, $P(R)$ et $P(T)$, est la dernière étape proposée en §2.2[23].

$P(T)$ peut se déduire directement de la probabilité conjointe de (T,S) calculée dans le tableau 15 simplement en sommant les probabilités par ligne, ce qui produit le tableau 18.

Pour $P(R)$, il suffit de reprendre la loi conditionnelle de $R \mid T$ et la marginale de T que nous venons de calculer pour leur appliquer le théorème de Bayes sur les événements élémentaires. Par exemple

$$\begin{aligned} P(R=f) &= P(R=f,T=0)+P(R=f,T=1) \\ &= P(R=f \mid T=0)\,P(T=0)+P(R=f \mid T=1)\,P(T=1) \\ &= 0.172\times 0.58+0.333\times 0.42 \\ &= 0.2396 \end{aligned}$$

On obtient ainsi la table 19.

4.7 avec des variables continues

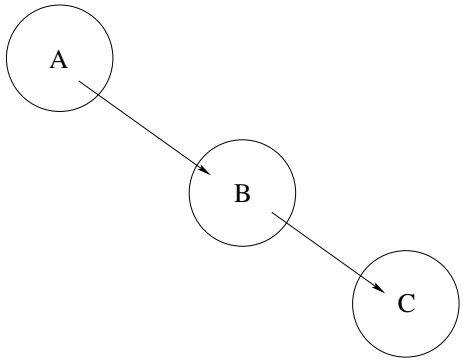
Comme nous l'avions fait dans le cas du réseau simple maillon (§3), il serait possible de poser une modélisation similaire avec 3 variables continues plutôt que 3 variables discrètes ; ou encore deux variables continues et une variable discrète... Mais nous considérons que la discussion sur le choix de la nature des variables aléatoires est reportée en §10.3 et nous ne nous concentrerons maintenant que sur le fonctionnement des réseaux bayésiens. Les réseaux suivants utiliseront des variables continues.

5 Réseau à 3 noeuds en chaîne : croissance de plante

La situation envisagée dans cette section est celle d'un réseau bayésien de trois noeuds enchaînés (cf. la figure 9). Elle correspond à la modélisation d'un processus séquentiel. A part la première, chaque variable est déterminée par une variable précédente. Certaines des variables de la chaîne seront observées, d'autres pas. Il s'agit aussi d'une situation très commune.

[23] La troisième probabilité marginale, $P(S)$, est elle donnée par lors de la construction du réseau.

FIG. 9 – Réseau bayésien en chaîne
Les variables s'influencent en cascade.



5.1 définition des variables

Supposons qu'on s'intéresse à la biomasse produite par une plante à trois stades phénologiques bien caractérisés et successifs. Nous les dénoterons respectivement A, B et C. On suppose que chacun est une variable aléatoire, mais qu'il existe une relation positive entre les stades successifs dans la mesure où si A est fort, alors B a plus de chance de l'être aussi. On admettra de ne pouvoir mesurer que la biomasse finale, c'est-à-dire C et on subodore que connaître la valeur prise par cette variable pourrait nous aider à préciser notre modélisation *a priori* de A qui est notre variable d'intérêt principal. B n'est donc qu'une variable intermédiaire pour notre modélisation.

L'idée de base est que si C est observé fort, alors A devrait l'être aussi. Mais de combien ? C'est à cette question que nous allons essayer de répondre par la construction d'un réseau bayésien.

Nous supposerons pour cet exemple que les trois variables sont continues. Comme elles sont de même nature (biomasse), il est raisonnable de les choisir de même type, mais il pourrait y avoir des situations où un type différent d'observations conduirait à un choix autre. Ici, on supposera que $A, B, C \in \mathbf{R}$. On aurait pu restreindre à $\mathbf{R}^+$ mais ceci nous compliquerait la spécification des distributions de probabilité. Nous allons supposer que chacune des trois variables suit une distribution normale.

5.2 définition du graphe

Pour tenter de formaliser ce qui vient d'être admis, on peut considérer les liaisons entre les trois variables A, B, C traduites par le graphe présenté en figure 9.

Les liaisons entre les variables sont symbolisées par des flèches pour indiquer le sens de l'influence[24] : l'effet va de A vers B puis de B vers C. Le graphe ne précise rien sur la nature des relations... Il permet quand même de raisonner. Par exemple, si nous connaissions B, la connaissance de C ne nous servirait de rien pour mieux caractériser A.

5.3 définition des relations

Maintenant que nous avons un petit référenciel de la spécification des réseaux bayésiens, nous pouvons nous servir du graphe pour déterminer quelles sont les distributions à proposer. Il

[24] en fait, cette présentation n'a rien de statistique (cf. §9.6)

y en a trois : la variable en tête de pont, A, dont il faut donner la marginale, puis la distribution conditionnelle de B par rapport à A, enfin celle de C par rapport à B. Par simplicité, nous resterons dans le cadre des distributions normales pour toutes ces distributions.

5.3.1 probabilité marginale de A : $[A]$

$$A \sim N\left(\mu_A, \sigma_A^2\right)$$

5.3.2 influence de A sur B : $[B \mid A]$

$$B \mid A \sim N\left(\mu_B + \beta_{B.A}\left(A - \mu_A\right), \sigma_{B.A}^2\right)$$

B sachant la valeur prise par A suit une loi normale dont l'espérance dépend linéairement de la valeur de A et dont la variance est constante. On a introduit $A - \mu_A$, et non pas seulement A, dans le terme linéaire en toute généralité puisque μ_A est une constante ; ainsi on tient compte du caractère plutôt élevé ou plutôt bas de A par rapport à sa valeur centrale. Cette précaution aura pour effet de donner à μ_B une interprétation simple comme on le verra au moment du calcul de la probabilité conjointe.

5.3.3 influence de B sur C : $[C \mid B]$

On développe le même type d'arguments que pour $C \mid B$:

$$C \mid B \sim N\left(\mu_C + \beta_{C.B}\left(B - \mu_B\right), \sigma_{C.B}^2\right)$$

5.4 probabilité conjointe

Il suffit simplement de faire le produit des trois densités :

$$[A = a, B = b, C = c] = [A = a]\,[B = b \mid A = a]\,[C = c \mid B = b]$$

$$\begin{aligned}
[A = a]\,[B = b \mid A = a]\,[C = c \mid B = b] &= \frac{1}{\sigma_A\sqrt{2\pi}} \exp\left(-\frac{\left(a - \mu_A\right)^2}{2\sigma_A^2}\right) \\
&\times \frac{1}{\sigma_{B.A}\sqrt{2\pi}} \exp\left(-\frac{\left(b - \left(\mu_B + \beta_{B.A}\left(a - \mu_A\right)\right)\right)^2}{2\sigma_{B.A}^2}\right) \\
&\times \frac{1}{\sigma_{C.B}\sqrt{2\pi}} \exp\left(-\frac{\left(c - \left(\mu_C + \beta_{C.B}\left(b - \mu_B\right)\right)\right)^2}{2\sigma_{C.B}^2}\right)
\end{aligned}$$

Quelques développements fastidieux de cette expression conduisent à la conclusion que le vecteur des trois variables aléatoires (A, B, C) est multinormal.

Et leurs paramètres sont donnés par :

$$\begin{aligned}
E\begin{pmatrix} A \\ B \\ C \end{pmatrix} &= \begin{pmatrix} \mu_A \\ \mu_B \\ \mu_C \end{pmatrix} \\
V\begin{pmatrix} A \\ B \\ C \end{pmatrix} &= \begin{pmatrix} \sigma_A^2 & \beta_{B.A}\sigma_A^2 & \beta_{B.A}\beta_{C.B}\sigma_A^2 \\ \beta_{B.A}\sigma_A^2 & \sigma_{B.A}^2 + \beta_{B.A}^2\sigma_A^2 & \beta_{C.B}\left(\sigma_{B.A}^2 + \beta_{B.A}^2\sigma_A^2\right) \\ \beta_{B.A}\beta_{C.B}\sigma_A^2 & \beta_{C.B}\left(\sigma_{B.A}^2 + \beta_{B.A}^2\sigma_A^2\right) & \sigma_{C.B}^2 + \beta_{C.B}^2\left(\sigma_{B.A}^2 + \beta_{B.A}^2\sigma_A^2\right) \end{pmatrix}
\end{aligned}$$

Les variances respectives des trois variables sont donc :

$$\begin{pmatrix} V(A) \\ V(B) \\ V(C) \end{pmatrix} = \begin{pmatrix} \sigma_A^2 \\ \sigma_{B.A}^2 + \beta_{B.A}\sigma_A^2 \\ \sigma_{C.B}^2 + \beta_{C.B}^2\left(\sigma_{B.A}^2 + \beta_{B.A}^2\sigma_A^2\right) \end{pmatrix}$$

ce qui peut apparaître très complexe, en fait on peut en faire une interprétation basée sur le réseau bayésien représenté en figure 9.

- A conserve sa variance de départ, c'est la moindre des choses puisqu'on lui a spécifié sa distribution marginale.
- B a comme variance $\sigma_{B.A}^2 + \beta_{B.A}^2\sigma_A^2$, c'est-à-dire sa variance conditionnelle plus $\beta_{B.A}^2\sigma_A^2$ que l'on peut interpréter comme la part supplémentaire de variation apportée par le déconditionnement de A. Ceci est d'autant plus clair que σ_A^2, la variance de A, est pondérée par $\beta_{B.A}^2$, le coefficient de régression de A pour B au niveau de l'espérance telle que nous l'avons définie.
- Et la même interprétation se retrouve pour C dont la variance est $\sigma_{C.B}^2+\beta_{C.B}^2\left(\sigma_{B.A}^2 + \beta_{B.A}^2\sigma_A^2\right)$. On y retrouve la variance conditionnelle $\sigma_{C.B}^2$ sommée à la part héritée de la variabilité de B $\left(\beta_{C.B}^2\left(\sigma_{B.A}^2 + \beta_{B.A}^2\sigma_A^2\right)\right)$ qui comprend la part héritée de la variabilité de A au travers de B.

On suit bien la progression par le graphe. Le même genre de raisonnement peut s'opérer pour les corrélations. Cette interprétation est à la base de la démonstration de cette formule dans la section suivante (§5.5).

5.5 calcul de la loi conjointe du réseau en chaîne

Il s'agit d'une démonstration du réseau à trois noeuds chaînés (figure 9) dont le résultat est énoncé en §5.4. Pour que les choses soient faciles, l'astuce consiste à introduire 2 nouvelles variables annexes que nous noterons $\tilde{B}$ et $\tilde{C}$; il s'agit en fait des résiduelles des régressions par rapport aux variables parentes. A, $\tilde{B}$ et $\tilde{C}$ sont indépendantes et les deux résiduelles sont définies par

$$\begin{aligned} B &= \beta_{B.A}A + \tilde{B} \\ C &= \beta_{C.B}B + \tilde{C} \\ &= \beta_{B.A}\beta_{C.B}A + \beta_{C.B}\tilde{B} + \tilde{C} \end{aligned}$$

Pour que la spécification du modèle soit bien identique à celle des lois conditionnelles, il faut donc poser

$$\begin{aligned} \tilde{B} &\sim N(\mu_B, \sigma_{B.A}) \\ \tilde{C} &\sim N(\mu_C, \sigma_{C.B}) \end{aligned}$$

A partir de cette nouvelle formulation, on retrouve bien la spécification proposée en §5.3 et il est très facile de calculer la loi conjointe du triplet (A, B, C) puisqu'on a la relation

$$\begin{pmatrix} A \\ B \\ C \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ \beta_{B.A} & 1 & 0 \\ \beta_{B.A}\beta_{C.B} & \beta_{B.C} & 1 \end{pmatrix} \begin{pmatrix} A \\ \tilde{B} \\ \tilde{C} \end{pmatrix}$$

On retrouve assez directement les résultats proposés en §5.4.

5.6 probabilité conditionnelle de la variable d'intérêt

La probabilité conjointe qui vient d'être calculée nous permet de répondre à la question initiale, celle de la distribution de A conditionnellement à la variable observée C.

Il suffit d'appliquer la théorie standard de la distribution multinormale.

$$A \mid C \sim N\left(\mu_A + \frac{\beta_{B.A}\beta_{C.B}\sigma_A^2}{\sigma_{C.B}^2 + \beta_{C.B}^2\left(\sigma_{B.A}^2 + \beta_{B.A}^2\sigma_A^2\right)}\left(C - \mu_C\right), \frac{\sigma_{C.B}^2 + \beta_{C.B}^2\sigma_{B.A}^2}{\sigma_{C.B}^2 + \beta_{C.B}^2\sigma_{B.A}^2 + \beta_{B.A}^2\beta_{C.B}^2\sigma_A^2}\sigma_A^2\right)$$

Bien que le noeud B n'intervienne qu'implicitement, son influence est déterminante aux deux niveaux de l'espérance et de la variance dans cette distribution. Ce qui est normal puisque c'est lui le point de passage obligé entre A et C. En particulier, il suffit qu'une des deux liaisons soit rompue ($\beta_{B.A} = 0$ ou $\beta_{C.B} = 0$) pour que la distribution conditionnelle se réduise à la distribution marginale définie par $N\left(\mu_A, \sigma_A^2\right)$.

6 Réseau à 3 noeuds convergent : accidents cardio-vasculaires

Pour terminer la série des réseaux de trois noeuds comportant deux relations, nous allons traiter du réseau convergent qui correspond aussi à une situation classique, celle de deux facteurs influençant une même variable. L'exemple proposé est celui de l'influence simultanée du tabac et de l'alcool sur les accidents cardio-vasculaires.

6.1 définition des variables

Pour mettre un peu de diversité dans les exemples, nous allons considérer différents types de variable : binaire pour la réponse ($A \in \{0, 1\}$) suivant que l'accident cardio-vasculaire s'est produit ou pas, continue pour la consommation d'alcool ($C \in \mathbf{R}^+$) et discrète pour la consommation de tabac ($T \in \{0, 1-5, 6-20, +20\}$). Nous allons aussi introduire une variable intermédiaire utile à notre modélisation : $p \in [0, 1]$ la probabilité de déclencher un accident ou son logit.

Le logit se définit par $l = \log \frac{p}{1-p}$. Cette transformation est couramment utilisée pour modéliser les probabilités. En fait, il y a complète équivalence entre les deux variables p et l : la transformation logit est biunivoque et $p = \frac{\exp(l)}{1+\exp(l)}$.

6.2 définition du graphe du réseau

La figure 10 propose à gauche la version directe du réseau et à droite celle faisant apparaître la variable cachée (comme p et l sont strictement équivalentes, on peut considérer qu'elles ne forment qu'un noeud). Nous allons bien entendu travailler avec le réseau comportant les quatre noeuds. Il s'agit d'un réseau divergent prolongé, mais il n'y a pas vraiment de structure à la prolongation et à la limite, nous pourrions ne pas expliciter p dans la modèlisation (cf. la remarque en §6.3.4), intrinsèquement il s'agit bien d'un réseau à trois noeuds.

6.3 définition des relations

Comme nous l'avons vu antérieurement, il s'agit de préciser les relations stochastiques qui existent entre chaque noeud et ses parents. Nous commencerons[25] par les noeuds sans parents : T et C.

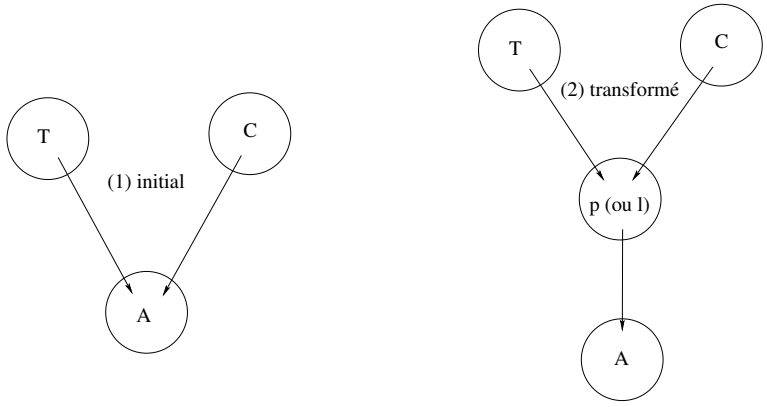
6.3.1 distribution marginale de T (consommation de Tabac)

C'est une variable que nous avions déjà rencontrée et nous allons reprendre la distribution proposée par le tableau 14, en changeant la notation : $P\left(S\right)$ devient $P\left(T\right)$.

[25] En fait l'ordre n'a aucune importance.

FIG. 10 – Réseau convergent augmenté

Pour expliquer les accidents cardio-vasculaires de manière plus synthétique, une variable cachée peut être introduite.



6.3.2 distribution marginale de C (consommation d'alCool)

Là encore, nous allons faire une petite innovation et considérer une variable hybride entre binaire et continue. Nous allons supposer qu'une proportion (50%) de la population considérée ne boit jamais d'alcool, et que l'autre partie consomme de l'alcool suivant une distribution demi-normale[26] centrée. La figure 11 illustre cette distribution.

6.3.3 distribution conditionnelle de l sachant T et C

Contrairement à p, limité au segment $[0, 1]$, la variable $l = logit\,(p)$ peut décrire toute la droite réelle, c'est pourquoi nous n'avons aucune difficulté à lui appliquer un modèle de régression linéaire en fonction des deux facteurs explicatifs que sont les noeuds T et C :

$$l \mid T, C \quad = \quad \mu_T + \beta C + E \tag{2}$$

Il y a 4 paramètres μ_T, un par classe de la variable T. En revanche la variable consommation d'alcool est supposée intervenir linéairement au travers du coefficient de régression β. E est une variable résiduelle, indépendante de T comme de C, traduisant le fait que la variabilité de l n'est pas complétement décrite par les facteurs explicatifs; nous la supposerons normale centrée d'écart-type σ_E.

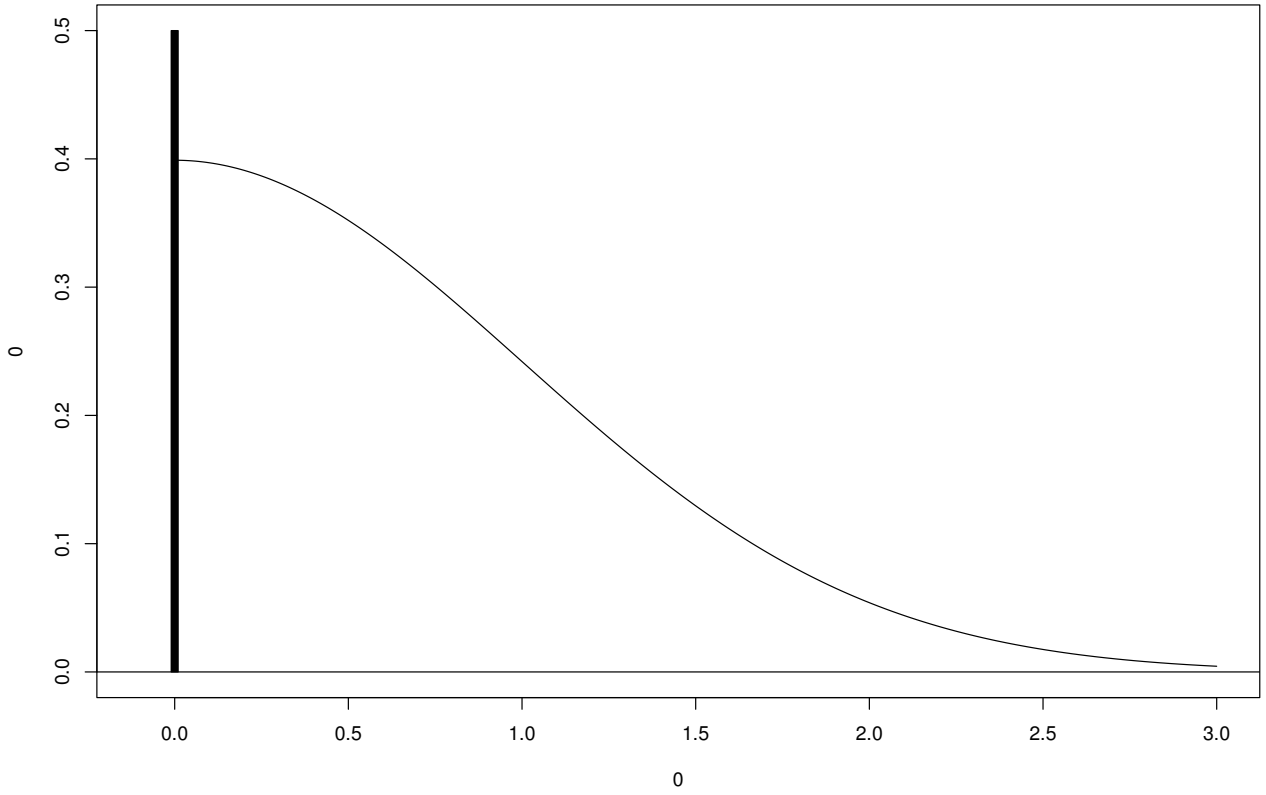
Spécifier l, c'est bien aussi spécifier p puisque

$$\begin{aligned} p &= \frac{\exp(l)}{1+\exp(l)} \\ &= \frac{\exp(\mu_T + \beta C + E)}{1+\exp(\mu_T + \beta C + E)} \end{aligned}$$

Ces paramètres s'interprètent assez naturellement. μ_1 correspond au risque d'accident cardio-vasculaire pour un niveau de consommation de tabac $T = 1$, c'est à dire aucune associée à aucune consommation d'alcool (puisqu'on doit mettre $C = 0$ pour le faire apparaître tout seul). β n'est rien d'autre que l'augmentation du risque par unité de consommation d'alcool ajouté à celui du tabac. On remarquera l'additivité supposée par ce modèle entre les deux sources du risque; cependant il faut faire remarquer que l'additivité est au niveau du logit, ce qui n'est pas vrai au niveau de la probabilité (cf. la petite application numérique du tableau 20).

[26] c'est-à-dire qu'on ne considère que la partie supérieure à l'espérance de la distribution normale.

FIG. 11 – Distribution marginale de la consommation d'alcool.
La barre verticale en zéro figure la probabilité de 0.5 de ne pas boire d'alcool parce qu'il s'agit d'une personne abstinente. La courbe à sa droite est celle d'une demi-normale ($\mu = 0, \sigma = 1$).



TAB. 20 – Application du modèle (2)
L'additivité postulée sur le logit ne se retrouve pas au niveau des probabilités correspondantes. Les valeurs des paramètres sont : $\mu_1 = -3$, $\mu_2 = -2$ et $\beta = 1$

	$C = 0$	$C = 1$
$T = 1$	$l = -3$ $p = 0.05$	$l = -2$ $p = 0.12$
$T = 2$	$l = -2$ $p = 0.12$	$l = -1$ $p = 0.27$

6.3.4 distribution conditionnelle de A sachant p

Connaissant la probabilité d'un accident, nous allons décrire son occurrence comme le résultat d'une binomiale de taille 1[27] et de paramètre p :

$$A \mid p \quad \sim \quad Bino\,(1,p)$$

Ce type de modélisation est souvent employé pour prendre en compte l'effet de facteurs continus (ici la consommation d'alcool) sur une variable d'intérêt binaire (l'occurrence d'un accident). On voit mal comment on pourrait procéder si on se livrait à une modélisation déterministe[28] !

Si nous n'avions pas voulu faire apparaître p, nous aurions pu directement le remplacer par son logit muni des paramètres de la régression linéaire :

$$A \quad \mid (T,C) \sim \quad Bino\left(1, \frac{\exp\left(\mu_T + \beta C + E\right)}{1+\exp\left(\mu_T + \beta C + E\right)}\right) \tag{3}$$

6.4 calcul de la distribution conjointe

Formellement, c'est le simple produit de toutes les distributions composant le réseau. Nous les rappelons en ne faisant pas apparaître la variable intermédiaire p (ou l) :

$$\begin{aligned}
[T=t] &= \left\{\begin{array}{ccc} 0 & \text{avec proba} & 0.4 \\ 1-5 & — & 0.2 \\ 6-20 & — & 0.2 \\ +20 & — & 0.2 \end{array}\right. \\
[C=c] &= \left\{\begin{array}{ccc} 0 & \text{avec proba} & 0.5 \\ c>0 & \text{suivant} & \frac{1}{2}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{c^2}{2}\right) \end{array}\right. \\
[E=e] &= \frac{1}{\sigma_E\sqrt{2\pi}}\exp\left(-\frac{e^2}{2\sigma_E^2}\right) \\
[A=a \mid T=t, C=c, E=e] &= \left\{\begin{array}{ccc} 0 & \text{avec proba} & 1-\left(1+\exp\left(\mu_t+\beta c+e\right)\right)^{-1} \\ 1 & \text{sinon} & \end{array}\right.
\end{aligned}$$

Ce qui est un peu compliqué à écrire avec ces mélanges de divers types de variables. Nous ne l'entreprendrons pas ici.

6.5 probabilité conditionnelle de la variable d'intérêt

Cette question ne se pose pas dans ce réseau bayésien puisque la réponse est déjà donnée au moment de la définition du réseau par (3).

6.6 probabilité marginale de la variable d'intérêt

Si la distribution condionnelle de la variable d'intérêt est par construction résolue, on peut se poser une autre question qui est celle de la probabilité marginale. Celle-ci représente ce qui arrive au niveau de la population totale, pour les différentes occurences des valeurs des covariables du modèle.

Dans le cas présent, il faut intégrer les covariables suivant leur distribution de probabilité. Nous y avons procédé numériquement (par simulation). Pour les valeurs $\mu_* = (-3,-2,-1,0)$;

[27] Les spécialistes nomment cette distribution une Bernouilli.

[28] Il est probable que le modélisateur déterministe se défendrait en disant qu'il ne faut pas prévoir l'accident sur un individu mais sur une population d'individus... ce qui revient à introduire la notion implicite de probabilité en se restreignant à l'espérance.

$\beta = 1$ et $\sigma_E = 0.2$, nous avons trouvé pour 10000 tirages de A, 6368 qui valaient 1. Comme ces tirages sont indépendants et de même distribution, A suit une loi binomiale dont le paramètre est voisin de 0.64.

7 Réseaux plus complexes

Il n'y a bien entendu aucune limitation au nombre de noeuds et aux flèches d'un réseau bayésien sinon celle des logiciels utilisés. Les réseaux précédents ont été limités à moins de 4 noeuds pour faciliter la compréhension et l'intuition du lecteur. Dans la pratique, la limite est la capacité à bien définir les lois de probabilité marginales et conditionnelles, ce qui devient toujours plus difficiles lorsque le nombre de variables s'accroit.

Les réseaux décrits dans la littérature scientifique se limitent en général à une dizaine de variables. Les applications réelles, en général présentées de manière très allusive, comportent jusqu'à une centaine de variables. Elles sont pour la plupart bâties sur des variables discrètes[29]. C'est d'ailleurs pour ce genre de variables que l'offre en logiciels est la plus abondante.

8 Quelques logiciels disponibles

Les logiciels existant traitant les réseaux bayésiens le font par une approche bayésienne (cf. le distinguo que nous avons tenu à faire en §10.5 et §10.6), ce n'est sans doute pas fortuit : l'approche bayésienne est à la fois élégante et puissante ! Les indications qui suivent ne sont pas au même niveau : nous pratiquons principalement WinBugs, et avons regardé les possiblités des autres plus ou moins en détail.

Une autre remarque intéressante est que la quasi-totalité des logiciels disponibles sont en pleine évolution : c'est un indicateur de la vitalité de ce domaine.

8.1 Bugs et associés

La version 0.5 de Bugs a été placée sur la toile en 1996. Elle contenait déjà la majorité des caractéristiques du noyau de modèles possibles. Cette version initiale écrite pour Unix et Dos a constamment été améliorée (interfaces graphiques et algorithmes internes) dans la version WinBugs (qui ne tourne que sous MSWindows). Jusque la fin 2004, le logiciel était disponible dans sa version compilée pour 0 dollars moyennant un enregistrement comme utilisateur pour obtenir un chiffre de décryptage. Actuellement, le logiciel est *open source*, c'est à dire que ces codes sont disponibles. Malheureusement, ils sont en *Pascal Component*, un langage de programmation qui n'est implanté que sous MSWindows. Un équivalent Linux est théoriquement disponible, mais il n'est pas clair comment le mettre en oeuvre.

Les modèles peuvent être définis de manière graphique par la création d'un dag avec des propriétés distributionnelles attachées aux noeuds. Celui-ci est traduit par une spécification scripturale du modèle qui sert d'entrée au calcul. Pour celui qui accepte d'en apprendre la syntaxe (simple), il paraît plus efficace d'écrire soi-même (pour modifications) le script décrivant le modèle. Les résultats du calcul des densités empiriques peuvent facilement être récupérés pour un traitement par des logiciels statistiques généraux comme R ou SAS. De plus, il est possible à l'intérieur d'une des sessions de ces logiciels de lancer WinBugs pour en récupérer les calculs.

Les modèles envisageables sont extrêment nombreux, il n'y a pas de limite informatique sur la construction du réseau qui peut contenir des noeuds discrets et continus, chacun suivant une distribution de probabilité choisie parmi un large ensemble. Des éléments de programmation dans la construction du modèle peuvent être introduits pour transformer les données et les variables par des fonctions mathématiques usuelles ; on dispose même d'un branchement logique élémentaire.

Le calcul des distributions *a posteriori* se fait par un algorithme MCMC dont la convergence doit être contrôlée par l'utilisateur. Si cela ne pose pas de problèmes pour les modèles simples, la mise en oeuvre de modèles complexes requiert un certain savoir-faire.

[29] On peut toujours transformer une variable continue en variable discrète en choisissant des limites de classes pour obtenir des catégories du genre : *rien, peu, pas beaucoup, moyennement, beaucoup, énormément.*

8.1.1 WinBugs

Toujours diffusée, la 1.4.1 est la dernière version compilée proposée.

8.1.2 OpenBugs

Théoriquement OpenBugs correspond à l'ensemble des déclinaisons de Bugs sur les différents systèmes d'exploitation. Pratiquement, cela ne marche que pour MSWindows. Il est raisonnablement possible de créer de nouvelles distributions grâce à des exemples (templates) en Pascal Component. Si ce n'est pas le cas, mieux vaut utiliser WinBugs puisque l'expérience montre qu'il est plus rapide en temps de d'exécution.

8.1.3 LinBugs

Version Linux de Winbugs (sans l'interface graphique, parfois nommée plus génériquement ClassicBugs), n'est pas fonctionnelle à notre connaissance.

8.1.4 Jags

Un clone indépendant de Bugs, écrit en C par Martyn Plummer pour tourner sous Linux.

8.2 paquets de R

8.2.1 Coda

C'est un paquet de R dédié au traitement des sorties numériques de WinBugs et Jags.

8.2.2 BRugs

Paquet pour disposer des fonctionnalités de ClassicBugs (pas d'interface graphique) sous R. Ne fonctionne pour l'instant que sous MSWindows.

8.2.3 Deal

Paquet de R orienté vers la recherche de réseaux bayésiens optimaux (le programme cherche à trouver quels sont les flèches utiles au graphe, seuls les noeuds étant prédéfinis !). Le contexte est restreint par la normalité pour les variables continues et la multinomialité pour les variables discrètes, et les distributions conjuguées associées. Contraintes semblables à celles de Hugin (cf. §8.3.4) : il comprend d'ailleurs une interface vers cette application.

8.2.4 Grappa

Paquet de R pour la propagation des probabilités dans des réseaux bayésiens ne comportant que des variables discrètes. Equivalent de Hugin (cf. §8.3.4) mais sans les variables continues et l'interface graphique.

8.3 autres produits

8.3.1 FBM

FBM pour Flexible Bayesian Model, logiciel promu par Ratford Neal, auteur de nombreux papiers fort intéressant sur les approches bayésiennes dans un contexte d'apprentissage et d'intelligence artificielle (réseaux de neurones).

8.3.2 Bayes Net Toolbox

BNT est une bibliothèque *open source* de fonctions MatLab. Les variables peuvent être discrètes ou continues mais avec une variété limitée de distributions. L'approche statistique utilisée n'est pas bayésienne mais classique.

8.3.3 BayesiaLab

Laboratoire d'étude et de manipulation de réseaux bayésiens dans un environnement graphique sophistiqué. Il inclut des possibilités de réseau dynamique[30]. Il ne travaille que sur des variables discrètes ou discrétisées, et permet l'application de nombreux algorithmes pour l'estimation des paramètres des tables de probabilité. C'est un produit commercial.

8.3.4 Hugin

C'est le produit commercial de référence (créé en 1989), robuste et simple à utiliser. L'inclusion de noeuds continus s'accompagne de quelques contraintes (normalité, ils ne peuvent être parents de noeuds discrets, s'ils ont des noeuds parents discrets, alors l'espérance et la variance dépendent tous les deux de la valeur prise par la variable discrète, s'ils ont des noeuds parents continus, alors ils en hérite comme une composante additive).

8.3.5 Netica

Début de commercialisation en 1995, Netica a pour objectifs le diagnostic, la prédiction et la simulation. D'abord orienté sur les variables discrètes, l'introduction de variables continues s'effectue par un additif qui ne les prend pas en compte comme telles. La mise à jour des probabilités à partir de données se fait par approche bayésienne. Permet de pratiquer des études de sensibilité.

8.4 programmation directe

C'est l'utilisation directe de langages de programmation de base[31] ou de programmation plus globale[32] (en général interprété). Etant donné la variété des logiciels dédiés cités ci-dessus, on peut s'étonner que des statisticiens pratiquent une programmation directe. Et bien, c'est le cas de la plupart des chercheurs dans ce domaine. Les raisons en sont que c'est le seul moyen de conserver complète liberté mais aussi que les algorithmes à mettre en place ne sont pas extrêmement sophistiqués. Néanmoins, nous le déconseillons aux néophytes des réseaux bayésiens et/ou de la programmation.

9 notions élémentaires sur les probabilités

9.1 variables aléatoires

Dans tous nos développements, nous avons pris soin de bien définir des variables aléatoires, supports de nos modélisation. Il faut sans doute revenir sur cette notion de base. Comme la dénomination l'indique, une variable aléatoire est une variable dont la valeur ne peut être déterminée de manière certaine à l'avance. Cependant la répartition des valeurs qu'elle prend n'est pas complètement imprévisible, ce qui explique qu'on puisse tirer des informations de l'observation de variables aléatoires.

Quelques exemples.

9.1.1 variables aléatoires discrètes

On dispose d'une urne contenant 100 boules blanches et 200 boules noires de même taille, indiscernables au toucher. Après avoir mélangé les boules un bon moment, en fermant les yeux, on sort au tâté une boule de l'urne. **La couleur de la boule** est une variable aléatoire discrète (et même binaire) qui prend les deux valeurs *blanche* ou *noire* avec les probabilités respectives

[30] Les variables du réseau varient aussi en fonction du temps.

[31] dits de 3ème génération comme Fortran, Pascal, C, C++, Perl,...

[32] dits de 5ème génération comme MatLab, R, APL,... qui comportent une panoplie élaborée de traitement de structure et épargne, par exemple, la gestion de la mémoire au programmeur.

de $\frac{1}{3}$ et $\frac{2}{3}$. Il est impossible de prédire de manière certaine la couleur qui va sortir, mais si on recommence l'opération 200 fois, on n'admettra pas de toujours sortir une boule noire, bien que ce soit tout à fait possible.

C'est dans de tels exemples que les notions de nombre de cas favorables et nombre de cas possibles dont le rapport définit la probabilité sont les plus appropriées. Les modèles d'urne, en augmentant le nombre de couleurs, en variant la composition de l'urne suivant les tirages effectués donnent lieu à de très nombreux (et parfois difficiles) calculs de probabilités.

On lance 3 dés et s'intéresse à **la somme des points obtenus**. On a ainsi généré une variable aléatoire discrète dont les valeurs varient de 3 à 18, et dont l'histogramme obtenu par un grand nombre de lancés se rapproche déjà de la distribution normale.

9.1.2 variables aléatoires continues

Dans une pièce rectangulaire, après avoir effectué 5 tours sur soi-même les yeux fermés, on lance en l'air sans les rouvrir une épingle. On s'intéresse à **l'aire du triangle défini par la tête de l'épingle et le côté** de la pièce opposé à l'unique porte. Il s'agit d'une variable aléatoire continue qui peut prendre des valeurs entre 0 et la moitié de l'aire de la pièce. Dans cet exemple géométrique, il est plus difficile de supputer quelle sera la répartition de la variable aléatoire ; en fait cela va dépendre de paramètres comme la dimension de la pièce mais aussi de la façon dont le lancer s'effectue.

Dans un grand champ de blé, on lance un cerceau le plus loin que l'on peut. On recueille ensuite toutes les tiges portant un épi qui se trouvent à l'intérieur du cerceau et on s'intéresse à la **moyenne arithmétique de leur hauteur**. On a à faire à une variable continue dont il est difficile de préciser la borne supérieure, même si on sait qu'elle est inférieure à 100 cm. Sans doute un agronome averti saurait-il donner des indications sur la répartition attendue de cette variable aléatoire. Notons qu'elle dépend de la variété de blé, mais aussi du nombre de tiges emprisonnées qui sera une variable aléatoire dépendant de la taille du cerceau.

9.1.3 variable continue ou variable discrète ?

La matière, et encore plus la manière dont nous la percevons, étant de nature discontinue, on pourrait bien arguer que toutes les variables (aléatoires ou non) sont discrètes. Si par exemple on s'intéresse à la taille d'un élève pris au hasard dans un établissement scolaire de 1000 élèves, on pourrait bien l'assimiler à une des au plus 1000 valeurs possibles. De la même manière qu'on considère un jet de dé comme discret, et non pas continu, parce qu'il prend des valeurs numériques. En fait, ce débat se pose peu en modélisation. La question n'est pas tant la nature exacte de la variable que celle de la qualité de l'approximation du modèle qu'on retiendra ; le modèle, on sait qu'il est faux quel qu'il soit. Et pour le cas des 1000 élèves, mieux vaut considérer une variable continue qu'une variable discrète car il sera plus aisé (et moins dispendieux en paramètres) de caractériser la répartition.

En fait, ce qui nous importe c'est de pouvoir caractériser l'occurrence des événements du phénomène étudié. C'est justement l'objet des probabilités

9.2 probabilité

La réponse à certaines questions peut être sans nuances : « oui » ou « non », « vrai » ou « faux », il n'y a pas alors de place pour le hasard. Mais plus souvent, parce que la question posée se situe à la frontière des connaissances, ou plus simplement parce que tous les déterminants ne sont pas connus, la réponse est incertaine. La théorie des probabilités est un outil

utile pour bien formaliser cette situation. **Dans une optique d'approche bayésienne, on admet que la probabilité de réalisation d'un événement est une caractéristique de notre information à son sujet. Elle doit être modifiée dès que cette information est modifiée : c'est l'objet du théorème de Bayes**. Ce théorème est présenté en §9.4. Dans cette section, nous nous bornons à rappeler les caractéristiques principales des probabilités.

Une probabilité se rapporte à un événement et caractérise la plus ou moins grande certitude de son occurrence. Plus la probabilité est grande, plus il sera *probable* que l'événement se produise. En général, il est commode de définir les événements au moyen de variables aléatoires pour une facilité de formalisme mathématique. Citons quelques événements pour lesquels on s'intéresse aux probabilités : arriver en retard à une réunion, sortir 421 avec trois dés, parier le tiercé gagnant dans l'ordre, perdre à la loterie, que le niveau d'un fleuve dépasse la hauteur de la digue, que des personnes succombent à une intoxication alimentaire d'un certain type,...

- Une probabilité est une valeur numérique comprise entre 0 et 1. La valeur 0 siginifie que l'événement est impossible ; la valeur 1 que l'événement est certain. Attention un zéro mathématique est très différent d'une valeur supérieure à zéro, si petite soit elle : une probabilité de 10^{-12}est > 0.
- La probabilité de l'union d'événements exclusifs (si l'un se produit l'autre ne peut se produire et réciproquement) est égale à la somme des probabilités des événements. La "proba" de tirer (1 ou 2) avec un dé est égale à la probabilité de tirer 1 plus la probabilité de tirer 2.
- Les probabilités d'événements complémentaires somment à 1. La probabilité de tirer (1 ou 2) ou (3 ou 4 ou 5 ou 6) en lançant un dé est égale à 1 car dans la modélisation, on suppose que le dé (i) retombera et (ii) retombera sur une de ses 6 faces.

Il existe certains débats philosophiques sur la manière dont on doit considérer les probabilités. S'agit-il d'une notion subjective traduisant un certain degré de confiance (et qui peut varier d'un individu à l'autre), ou d'une valeur inconnue que nous pourrions connaître objectivement, si nous étions capables de répéter l'expérience de son observation un nombre infini de fois. Mais le débat ne nous paraît pas très important. Ce qui compte, c'est de savoir manipuler correctement les probabilités une fois qu'on s'est mis d'accord sur leurs prémisses. Autrement dit, une fois un modèle admis, être capable d'en tirer tous les mêmes conclusions. Le choix d'un modèle restera toujours problématique.

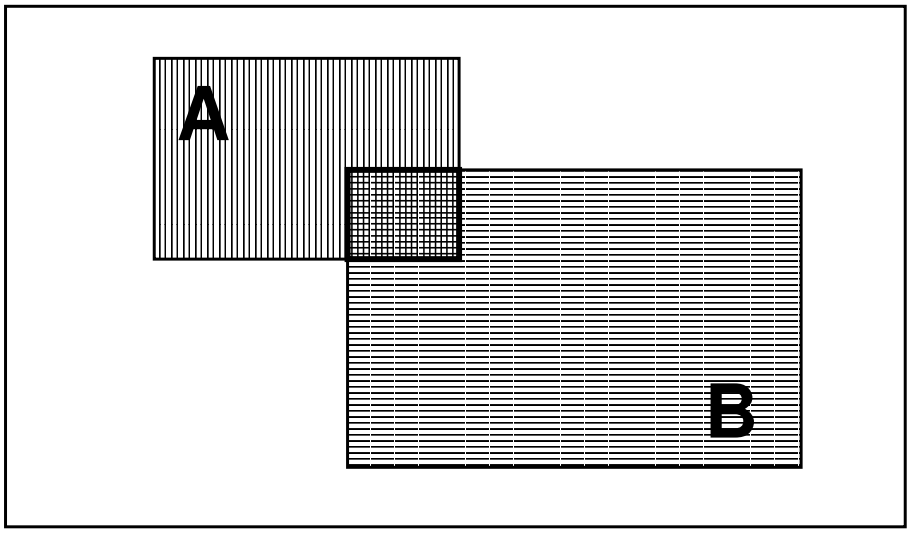
9.3 probabilité conditionnelle

On parle de probabilité conditionnelle lorsqu'on restreint les événements possibles. Par exemple, si on s'intéresse à la probabilité de tirer 3 avec un dé : la probabilité sera $\frac{1}{6}$, si on s'intéresse à la probabilité de tirer 3 sachant que le résultat du dé est inférieur à ou égal à 4, la probabilité devient $\frac{1}{4}$. Pratiquement, on restreint l'ensemble des possibles par une information supplémentaire. Si on note D la variable aléatoire associée au résulat du dé, on notera :

$$\begin{aligned} P(D=3) &= \frac{1}{6} \\ P(D=3 \mid D \leq 4) &= \frac{1}{4} \end{aligned}$$

La notation *barre verticale* se lit *conditionnellement à* ou *sachant que*, ou de manière plus usuelle *étant donné que*. Certains utilisent le signe / mais il a l'inconvénient d'être ambigü avec la division. Suivant la représentation par flèches utilisées dans les réseaux bayésiens, il aurait été tentant de noter $(D = 3 \leftarrow D \leq 4)$ mais l'influence fleure la causalité ce qui pour nous

FIG. 12 – Illustration du théorème de Bayes.
Symboliquement, la probabilité d'un événement est représentée par l'aire du sous-ensemble d'un rectangle associé. Ici A est le rectangle hachuré verticalement ; B est le rectangle hachuré horizontalement. Suivant le théorème de Bayes, $P(A \mid B)$ vaut le rapport des aires de $(A \cap B)$ et de B. On réalise par exemple que si $A = B$, alors$P(A \mid B) = 1$, ce qui est normal !



appartient au domaine (subjectif) de l'interprétation d'un modèle. De toute façon, la notation "|" est devenue classique.

9.4 théorème de Bayes

Le théorème de Bayes, ou théorème des probabilités conditionnelles, a prêté son nom aux réseaux bayésiens car la distribution conjointe des variables aléatoires est décrite au moyen de probabilités conditionnelles enchaînées. On peut l'exprimer sous plusieurs forme. Nous présentons une des plus simples.

Théorème de Bayes

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \tag{4}$$

ce qui s'énonce *la probabilité de A conditionnellement à B est égale au rapport des probabilités de $A \cap B$ et de B*.

L'illustration classique est celle du diagramme de Venn (Cf. Figure 12).

Une application symétrique du téorème permet de formuler que

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

On voit immédiatement que si A et B sont des événements exclusifs, $A \cap B = \emptyset$ et la "proba" conditionnelle est nulle. Ce qui est logique puisque dans ce cas savoir que B est réalisé implique bien que A ne peut pas l'être.

De ce théorème, on peut aussi retrouver la relation qui lie la probabilité de deux événements indépendants. En effet, si $P(A \mid B)$ ne dépend pas de B, on peut l'écrire $P(A)$ et on obtient

TAB. 21 – Distribution conjointe sous forme de probabilité

$P(S,T)$	$T=0$	$T=1$
$S=0$	0.70	0.05
$S=1$	0.05	0.20

TAB. 22 – Distribution conjointe sous forme d'effectifs

Effectifs	$T=0$	$T=1$
$S=0$	700	50
$S=1$	50	200

que $P(A \cap B) = P(A)\,P(B)$: si deux événements sont indépendants, la probabilité de leur intersection est égale au produit de leurs probabilités.

Dans le cas de variables aléatoires continues X, Y en utilisant la notation [] pour les densités de répartition, la formule du théorème de Bayes s'écrit

$$[X \mid Y] \quad = \quad \frac{[X,Y]}{[Y]}$$

une formulation plus symétrique est

$$[X \mid Y]\,[Y] = [X,Y] = [Y \mid X]\,[X]$$

De ce théorème, on peut aussi retrouver la rela

9.5 probabilités et effectifs dans une population

Nous avons toujours employé la notion de probabilité. Ce n'est pas forcément la plus aisée pour l'intuition ; certains préféreraient raisonner à partir des effectifs d'une population. Voyons comment se fait la correspondance entre les deux.

Soient deux variables binaires T et S. Supposons que nous y associions la distribution de probabilité conjointe du tableau 21. On notera que la somme de toutes les probabilités fait bien l'unité. On aurait pu tout aussi bien obtenir cette distribution par la répartition des individus d'une population de 1000 individus. C'est ce que propose le tableau 22. Cette fois-ci la somme égale la taille de la population, c'est-à-dire 1000. Bien entendu, 1000 nous permet de retrouver immédiatement les probabilités, mais nous aurions pu utiliser une population de taille quelconque.

Ce qui relie les deux présentations, c'est le raisonnement suivant : si on tire un indivudu au hasard[33] dans la population décrite par le tableau 22, alors le résultat est une variable aléatoire qui suit la distribution de probabilité proposée dans le tableau 21.

On utilise parfois la notion de fréquence relative qui consiste à diviser les effectifs de chaque classe par la taille de la population, ce qui redonne les valeurs associées aux probabilités.

L'inconvénient de la présentation par effectifs de population est qu'on ne dispose pas du même formalisme pour les variables continues. C'est la raison principale pour laquelle il nous

[33] En fait, la notion de *hasard* mérite d'être précisée. Il s'agit de munir chacun des individus de la population d'une même probabilité d'être tiré, soit dans le cas présent, de $\frac{1}{1000}$.

semble plus judicieux d'utiliser la formalisation des probabilités. Néanmoins, nous n'avons pas hésité à emprunter la représentation par effectifs pour développer l'intuition autour de certaines situations comme en §3.3.3.

9.6 causalité, corrélation, distribution conditionnelle

Quand on se place au niveau d'une application, il est beaucoup plus facile et naturel d'employer une perspective de causalité. C'est, par exemple, de ne pas énoncer *la distribution de Y sachant que $X = x$ est telle que l'espérance est linéaire en x*, mais plutôt *l'influence de X sur Y se traduit par une fonction linéaire de l'espérance...* Dans le cadre des réseaux bayésiens, c'est considérer que les flèches du réseau modélisent des relations de cause à effets. Ce qui est très parlant pour le futur client d'un éditeur de logiciel de traitement des réseaux bayésiens... mais n'est pas correct sauf si une hypothèse, supplémentaire au réseau bayésien, est ajoutée. Nous avons vu en effet à diverses reprises que l'on pouvait exprimer le même réseau en changeant certaines flèches. La question est identique à l'interprétation d'un coefficient de corrélation élevé entre deux variables : s'il est réellement significatif, trois possiblités (non exclusives) se présentent :

- la première variable influence la seconde,
- la seconde variable influence la première,
- les deux variables subissent l'influence d'une (ou plusieurs) autre(s) variable(s).

10 Considérations supplémentaires

10.1 dénombrements des dag (directed acyclic graphs)

10.1.1 avec trois noeuds ou moins

La figure 13 présente tous les réseaux de type *dag* de 1, 2 ou 3 noeuds. On remarquera que les symétries et l'interdiction des cycles limite sérieusement le nombre de cas de figures puisque nous avons seulement 6 graphes distincts pour 3 noeuds, alors qu'il y en a $2^{2\binom{3}{2}} = 2^6 = 64$ envisageables si on considère l'existence ou non des 6 flèches différentes entre 3 noeuds.

10.1.2 avec quatre noeuds

Le dénombrement des dag n'est pas une affaire simple pour les graphes comportant 4 noeuds, nous avons trouvé (mais sans complète certitude) 1 à zéro relation, 1 à une relation, 4 à deux relations, 9 à trois relations, 8 à quatre relations, 10 à cinq relations et 3 à six relations, ce qui ferait un nombre total de 36 parmi les $2^{2\binom{4}{2}} = 2^{12} = 4096$ envisageables ? (Cf. la figure 14)

10.2 réseaux équivalents

Il n'est pas utile de conserver dans un catalogue des réseaux qui permettent les mêmes modélisation. A cette fin, **on définira deux réseaux comme équivalents s'ils sont basés sur le même ensemble de noeuds et si les probabilités conjointes qui leur sont associables sont équivalentes dans le sens où on peut passer de l'une à l'autre et réciproquement**.

FIG. 13 – Réseaux bayésiens comportant moins de quatre noeuds.
Il y en a neuf. Le premier chiffre est le nombre de noeuds (ronds / variables), le second le nombre d'arcs (flèches / relations directes).

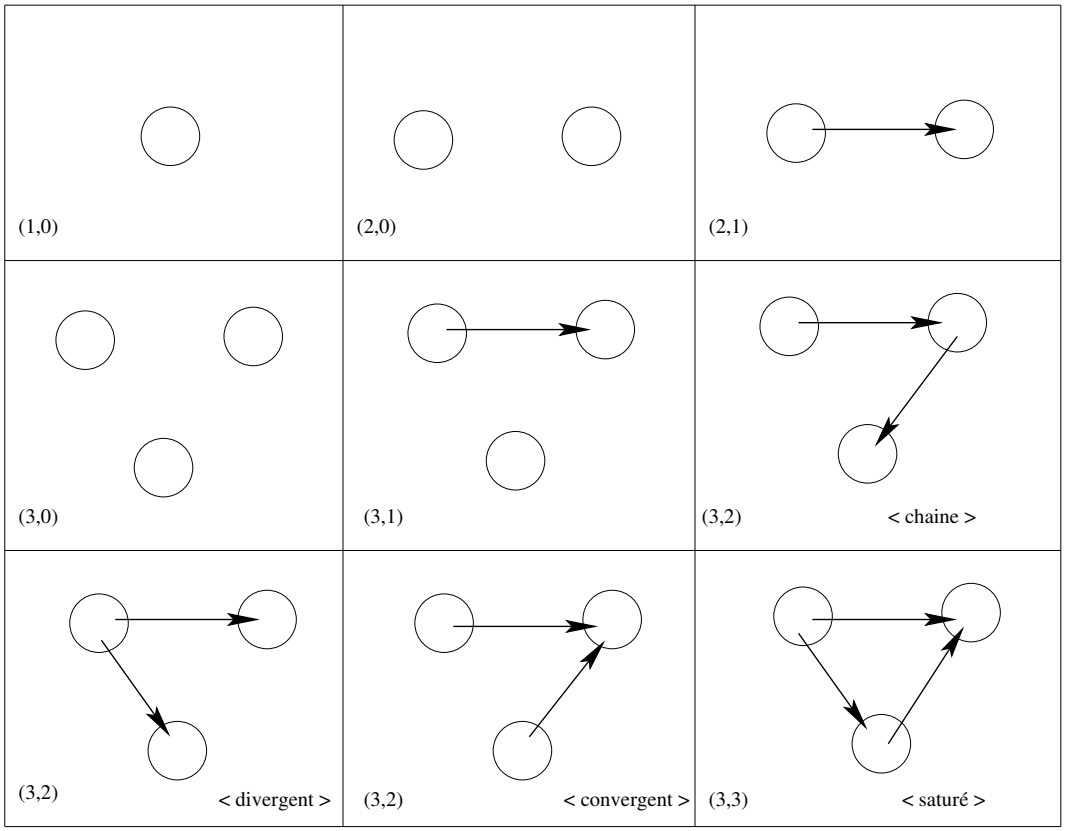
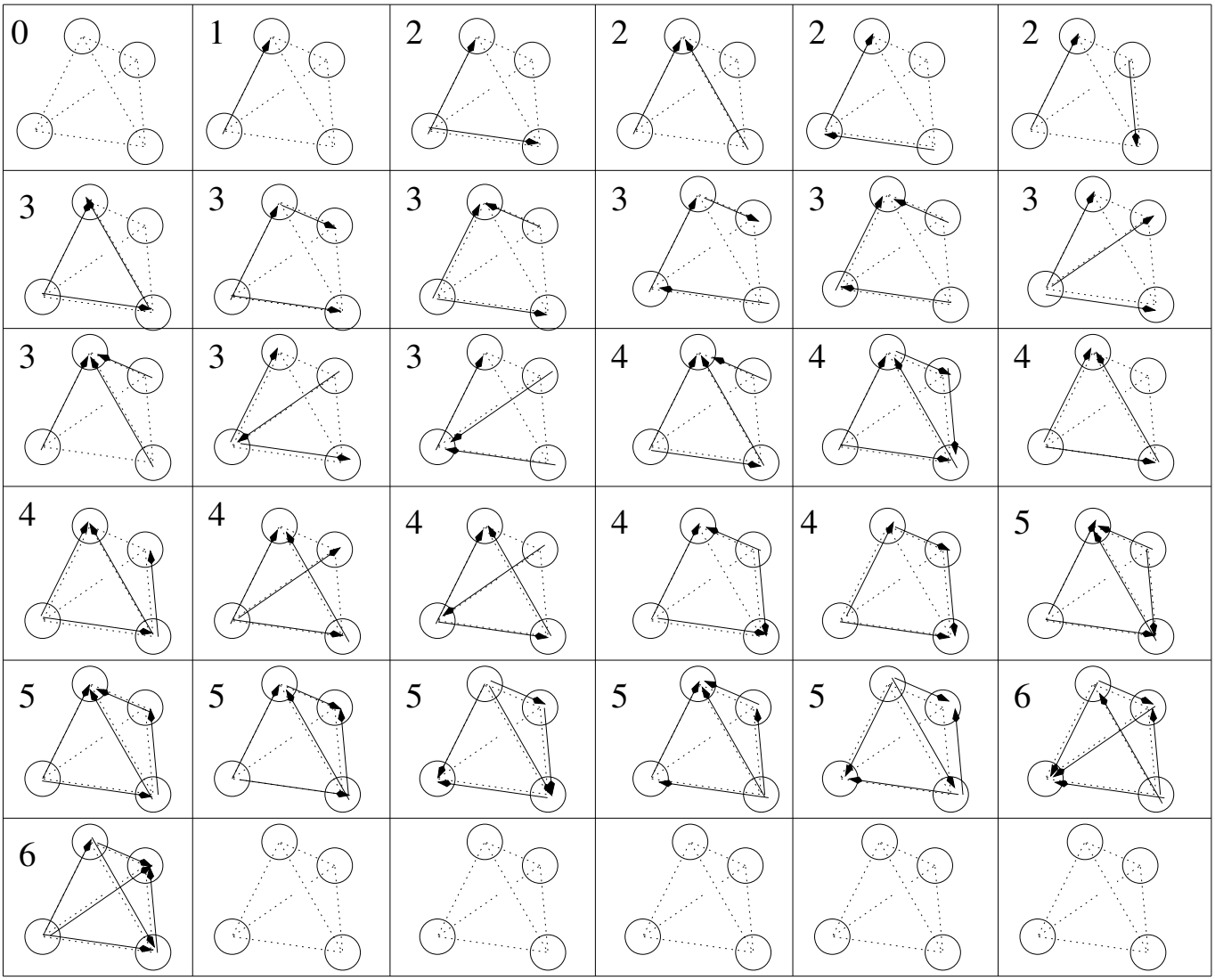


FIG. 14 – Réseaux bayésiens comportant quatre noeuds.
Nous en avons dénombré 31. C'est le nombre d'arcs est indiqué.



10.2.1 le maillon élémentaire

Dans le cas d'un réseau constitué d'un seul maillon, on peut montrer l'équivalence des deux réseaux possibles $(A \rightarrow B)$ et $(A \leftarrow B)$.

Dans le cas où les deux variables sont discrètes, on peut le démontrer en énumérant les nombres de paramètres libres de la modélisation. Notons N_A et N_B les nombres de modalités des deux variables. Alors :

- $[A, B]$ comporte $N_A N_B - 1$ paramètres libres pour le modèle le plus général,
- $[B]$ comporte $N_B - 1$ paramètres libres et $[A \mid B]$ en comporte $N_B (N_A - 1)$, le total fait bien $N_A N_B - 1$.
- $[A]$ comporte $N_A - 1$ paramètres libres et $[B \mid A]$ en comporte $N_A (N_B - 1)$, le total fait toujours $N_A N_B - 1$.

Dans le cas où les deux variables sont continues binormales (cf. §), on peut montrer que

$$\begin{array}{lcl} A \sim Normale & \text{et} & B \mid A \sim Normale \text{ d'espérance linéaire en } A \\ & \Leftrightarrow & \\ & \begin{pmatrix} A \\ B \end{pmatrix} \sim BiNormale & \\ & \Leftrightarrow & \\ B \sim Normale & \text{et} & A \mid B \sim Normale \text{ d'espérance linéaire en } B \end{array}$$

Ce qui montre l'équivalence des deux réseaux.

10.2.2 réseaux à trois noeuds

Mais tous les réseaux à 3 noeuds que nous avons vus ne sont pas équivalents. Si nous considérons les réseaux à trois noeuds ayant le même noeud central (c'est à dire la chaîne $A \rightarrow B \rightarrow C$), le divergent $(A \leftarrow B \rightarrow C)$ et le convergent $(A \rightarrow B \leftarrow C)$, un calcul de dimension paramétrique nous en persuade rapidement.

dans le cas discret

- chaîne : $[A], [B \mid A], [C \mid B]$ donnent $N_A - 1 + N_A (N_B - 1) + N_B (N_C - 1)$ ce qui produit $N_B (N_A + N_C - 1) - 1$.
- divergent : $[B], [A \mid B], [C \mid B]$ donnent $N_B - 1 + N_B (N_A - 1) + N_B (N_C - 1)$ ce qui produit $N_B (N_A + N_C - 1) - 1$.
- convergent : $[A], [C], [B \mid A, C]$ donnent $N_A - 1 + N_C - 1 + N_A N_C (N_B - 1)$ ce qui produit $N_A N_B N_C - (N_A - 1)(N_C - 1) - 1$.

Si les dimensions paramétriques sont différentes, les distributions conjointes le sont aussi.

`< continuer le développement en exhibant des paramétrisations pour` $N_A = 2, N_B = 3, N_C = 4$ `>>>`

dans le cas continu gaussien

`< à faire >`

10.2.3 réseaux quelconques

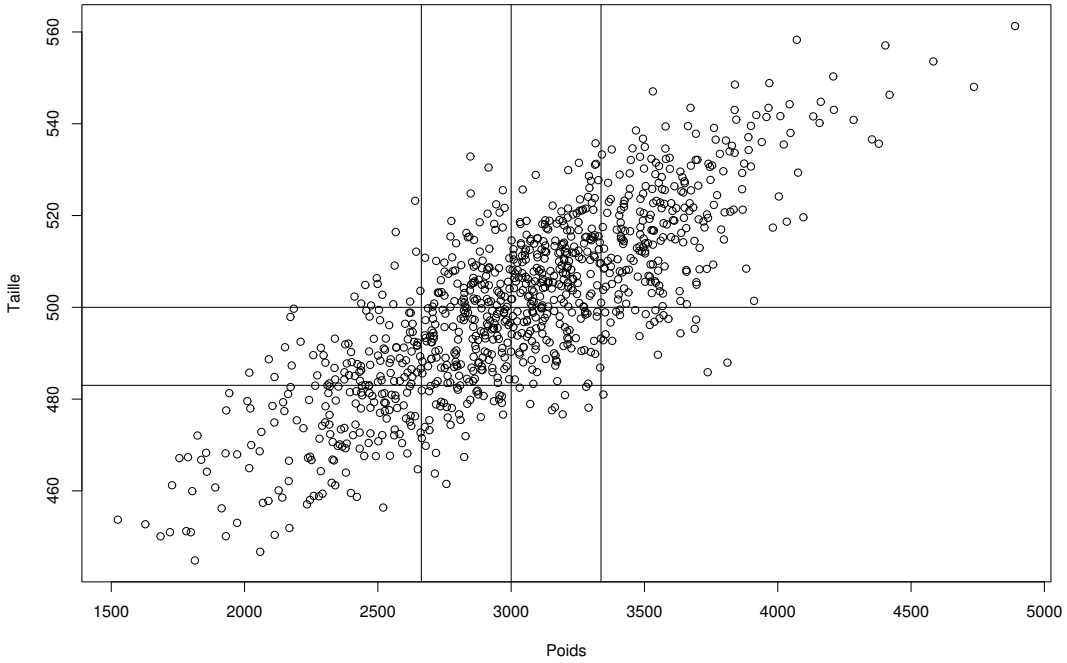
`< à réfléchir >`

10.3 Variables discrètes *versus* variables continues

10.3.1 l'envers du décor de l'illustration "une histoire de poids"

Les exemples binaire (§3.3), discret (§3.4) et continu (§3.5) sont en fait cohérents entre eux. La figure 15 dévoile le mécanisme. Nous avons supposé une distribution binormale des deux variables continues, la figure en

FIG. 15 – La distribution réelle de l'exemple Poids-Taille



TAB. 23 – Effectifs du découpage binaire

Effectifs	$P_b = 1$	$P_b = 2$	
$T_b = 1$	456	24	480
$T_b = 2$	291	229	520
	747	253	1000

présente un échantillon aléatoire de 1000 points. Les paramètres sont l'espérance, vecteur μ de dimension 2 et la variance, matrice Σ de dimension 2×2. Ce qu'on peut noter par :

$$\begin{pmatrix} P_c \\ T_c \end{pmatrix} \sim N\left(\mu = \begin{pmatrix} 3000 \\ 500 \end{pmatrix}, \Sigma = \begin{pmatrix} 250^2 & 0.8 \times 250 \times 20 \\ 0.8 \times 250 \times 20 & 20^2 \end{pmatrix}\right)$$

Plus explicitement, les espérances de P_c et T_c sont 3000 et 500, alors que leurs écarts-type sont 250 et 20, enfin leur corrélation vaut 0.8.

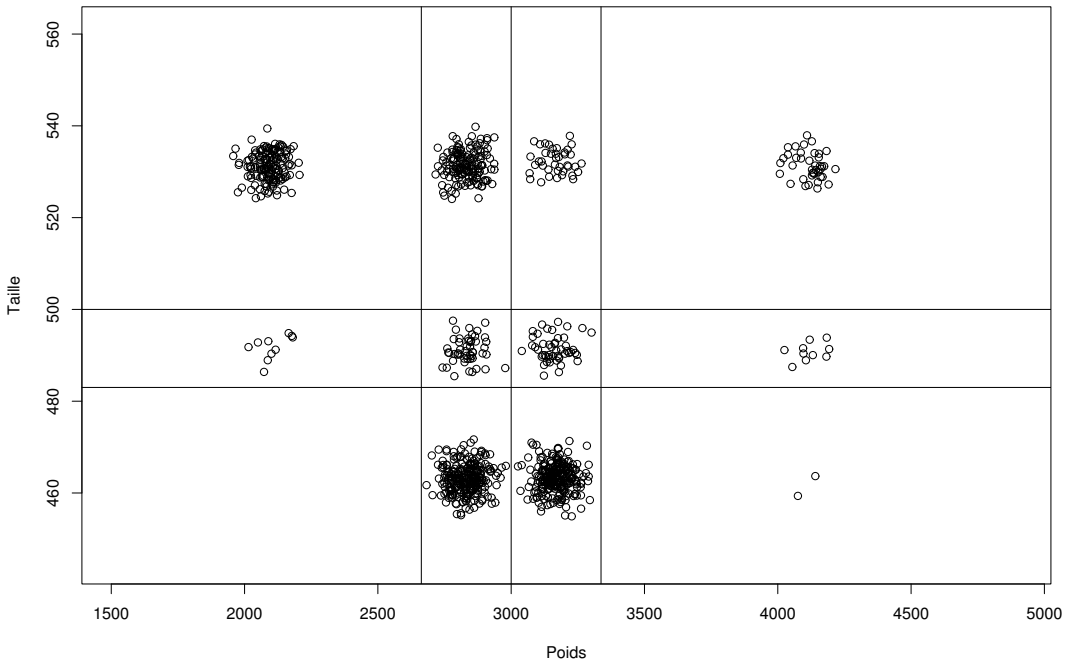
Nous avons ensuite délimité 12 régions dans le graphe (P_c, T_c) suivant les quantiles $(0.25, 0.50, 0.75)$ de P_c et $(0.2, 0.5)$ de T_c. Suivant ces régions, l'échantillon des 1000 points tirés nous a servi à générer une table d'effectifs 1000 de dimension 3×4 (cas discret : table 24), de laquelle l'équivalente pour le cas binaire (table 23) a été tirée. C'est à partir de ces tables que toutes les probabilités ont été construites, assurant la cohérence de l'ensemble. Le lecteur peut s'en convaincre en recomptant les effectifs des régions faiblement représentées.

Il est clair que l'exemple utilisé est finalement complétement adapté à celui de variables binormales. Ce n'est pas forcément le cas, si on imagine une répartion en différentes populations dont nos classes seraient de bonnes délimitations ; la figure 16 propose une illustration d'une telle répartition imaginaire.

TAB. 24 – Effectifs du découpage discret

Effectifs	$P_d = a$	$P_d = b$	$P_d = c$	$P_d = d$	
$T_d = 0$	134	39	8	1	(182)
$T_d = 1$	81	119	75	23	(298)
$T_d = 2$	14	88	189	229	(520)
	(229)	(246)	(272)	(253)	(1000)

FIG. 16 – Distribution imaginaire du maillon élémentaire



10.3.2 avantages et inconvénients des divers types de variables

Une question importante est bien entendu de savoir s'il vaut mieux modéliser avec des variables discrètes (éventuellement binaires si elles ne comportent que deux classes) ou continues. Il n'y a pas de réponse universelle. Notons néanmoins quelques arguments qui peuvent faire pencher la réponse d'un côté ou de l'autre.

- La nature des variables. Si une variable est la couleur de la fleur, ou le type de parasite, sa conversion en variable continue est sans doute assez incongrue. La question ne se pose vraiment que pour les variable de nature sous-jacente continue ou au moins ordonnée.
- La nature des données dont on peut disposer (si dans ce document, l'utilisation des données n'est pas du tout abordée, le plus souvent l'utilisation d'un modèle se fait en conjonction avec des données),
- Le nombre de données dont on dispose peut aussi guider le nombre de classes des variables discrètes. Si on veut distinguer beaucoup de graduations, la modélisation par variable continue est en général plus parcimonieuse... mais au prix de l'adoption d'une échelle, c'est-à-dire que chaque occurrence peut être placée sur une droite relativement à la position des autres.
- Si l'échelle ne se discute pas (comme dans le cas de notre exemple), c'est au contraire la question des limites de classes qui peut être difficile, et on préférera des variables continues.

10.4 Mélanges de variables continues et discrètes

Un réseau bayésien ne comporte pas forcément soit des variables discrètes soit des variables continues. Le mélange des deux types peut se faire de deux manières.

Tout d'abord, on peut avoir des variables en partie discrète, en partie continue. Un exemple simple est celui du niveau de contamination. On peut imaginer une probabilité de 0.8 de ne pas avoir de contamination et que les 0.2 restant soient distribués suivant une loi lognormale... C'est ce qui a été fait pour la variable C (consommation d'alcool) introduite en §6.3.2.

Aussi, on peut considérer que certaines variables sont discrètes et d'autres continues. C'est le cas de la variable continue P_c telle que définie en §3.5.1 et de la variable discrète T_d dont les trois valeurs possibles sont $\{0, 1, 2\}$ (cf. §3.4.1). On imagine facilement que pour chaque catégorie possible de la variable T_d, la loi conditionnelle de la variable continue, P_c, soit différente. Egalement, on admettra que pour chaque valeur de P_c, on dispose d'une répartition des probabilités sur les trois états possibles de T_d qui traduisent l'influence de la variable continue sur la variable discrète.

10.5 différents types de modéles probabilistes

modèles univariables Par modèles univariables, nous entendons des modèles dans lesquels on suppose que les observations sont découpées en sous-ensembles similaires[34] (aux données manquantes éventuelles près) que l'on peut considérer suivre une même distribution et dans lesquels se trouve une variable d'intérêt principal dont on cherche à expliquer le comportement. Ce sont les modèles de régression classique ; ils incluent les modèles d'analyse de la variance.

Exemple. On s'intéresse au rendement du blé en France, variable à expliquer. On dispose d'un certain nombre d'observations sur le rendement en quintaux par hectare. On dispose aussi d'un certain nombre de covariables[35] continues comme la profondeur du sol, la pluviométrie durant la montaison, les unités d'azote apportées à la parcelle,... et/ou d'un certain nombre de covariables discrètes comme la variété cultivée, le type de sol, l'année de production,... Alors un modèle de régression classique consiste à modéliser le rendement comme une variable aléatoire

[34] Ce sont les unités statistiques.

[35] Les covariables sont aussi dénommées *variables explicatives*.

ayant une fonction simple des covariables comme espérance et comme variance une fonction souvent plus simple (par exemple constante) des mêmes covariables.

Plus précisément si les rendements observés sont notés y_i pour $i = 1, ..., I$, que les covariables continues sont notées x_i^p, $p = 1, ..., P$ et les covariables discrètes le sont f_i^q, $q = 1, ..., Q$, alors pour les variables aléatoires Y_i représentant les observations y_i, on posera un modèle de la forme

$$[Y_i] \quad = \quad \Phi\left(x_i^*, f_i^*, \theta\right) \tag{5}$$

où $[Y_i]$ représente la densité de distribution de Y_i et θ le vecteur de paramètres du modèle ; un astérisque remplaçant un indice signifie l'ensemble des valeurs qu'il peut prendre[36]. On formule souvent le modèle par la seule[37] spécification des espérance et variance des variables aléatoires :

$$\begin{aligned} E\left(Y_i\right) &= \mu\left(x_i^*, f_i^*, \theta\right) \\ V\left(Y_i\right) &= \gamma\left(x_i^*, f_i^*, \theta\right) \end{aligned}$$

Dans les cas les plus fréquents d'indépendance entre unités statistiques, la matrice[38] de variance de l'ensemble des observations est diagonale. Et si la variance ne dépend pas des covariables, on a alors

$$V\left(Y_*\right) \quad = \quad \mathbf{I}\,\sigma^2$$

où $\mathbf{I}$ est la matrice identité[39] dont la taille est égale au nombre d'unités statistiques et σ^2 est la variance commune à toutes les observations.

modèles multivariables Les modèles multivariables sont les généralisations les plus directes des modèles univariables tels que définis précédemment. La différence essentielle est que la variable d'intérêt n'est plus unique : il y en a plusieurs. Il s'agit d'un vecteur[40] dont les composantes sont autant de variables d'intérêt univariables.

Exemple. On peut reprendre l'exemple précédent en ajoutant au rendement, deux variables d'intérêt comme le taux de protéine par rapport à la matière sèche et un indice à trois classes de la panifiabilité de la farine issue du blé. Ce sont tout simplement des modèles de régression multiviariable. Ils se traitent de manière très proche des modèles univariables.

Dans les cas les plus fréquents d'indépendance entre unités statistiques, la matrice de variance de l'ensemble des observations est bloc-diagonale[41]. Si la variance ne dépend pas des covariables, elle peut s'écrire sous la forme d'un produit de Kronecker[42] :

$$V\left(Y_*\right) \quad = \quad \mathbf{I} \otimes \Sigma$$

où Σ est une matrice de variance-covariance dont la taille est égale au nombre de variables d'intérêt.

[36] si u_i est un scalaire alors u_* est un vecteur dont les composantes sont $u_1, u_2, ..., u_i, ...$

[37] Ce qui laisse un flou sur la modélisation, en général implicitement levé par une hypothèse de normalité.

[38] On appelle matrice un ensemble de valeurs disposées dans un tableau ; le nombre de lignes et le nombre de colonnes de ce tableau sont les dimensions de la matrice.

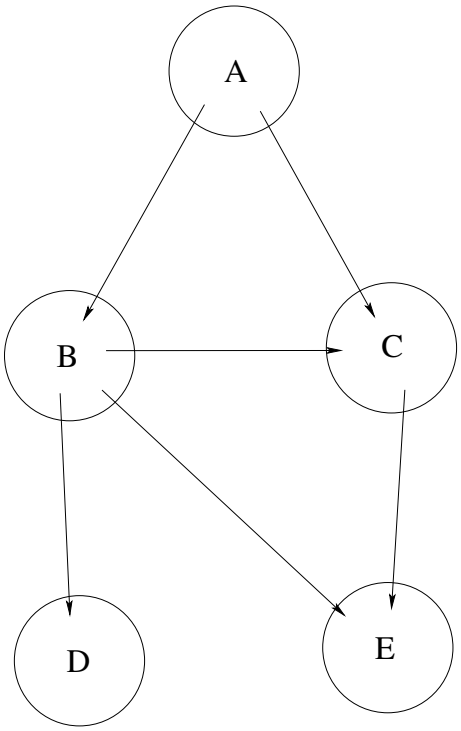
[39] Une matrice identité est une matrice carrée dont tous les éléments diagonaux sont égaux à 1, et tous les éléments non-diagonaux sont nuls. Elle est dite identité car l'application linéaire qui lui est associée laisse invariés tous les points de l'espace.

[40] Sont appelés vecteurs, des structures comportant une collection de variables scalaires, dites composantes du vecteur ; leur nombre est la dimension du vecteur. Par exemple, les deux coordonnées cartésiennes d'un point dans le plan forment un vecteur de dimension deux. Les vecteurs peuvent être assimilés à des matrices ne comportant qu'une seule ligne (ou une seule colonne).

[41] Une matrice est bloc diagonale si toutes ses composantes, non situées sur des blocs carrés centrés sur sa diagonale, sont nulles,

[42] Le produit de Kronecker de deux matrices, noté $\otimes$, est une généralisation du produit d'une matrice par un scalaire.

FIG. 17 – Exemple de réseau bayésien



réseaux bayésiens Comme les modèles multivariables comprennent comme cas particulier, les modèles univariables, les modèles exprimables sous forme de réseaux bayésiens comprennent comme cas particulier les modèles multivariables (et beaucoup d'autres). La différence principale est structurelle : la notion d'unité statistique n'est plus du tout centrale. En effet les réseaux bayésiens permettent de modéliser tout un ensemble de variables aléatoires (observées ou pas) par une densité de probabilité quelconque ! C'est donc *a priori* toute modélisation stochastique[43] mais on n'y trouve qu'une partie restreinte des modélisations envisageables (encore que le critère de celles qui y appartiennent est bien flou). Il s'agit, comme nous allons le voir par la suite, de distributions de probabilités conjointes[44] qui peuvent s'exprimer "*aisément*" à l'aide de distributions conditionnelles guidées par un réseau dont chaque variable est un noeud et les arcs représentent les conditionnements.

Exemple. Si nous considérons cinq variables aléatoires (A, B, C, D, E) et que leur densité conjointe globale présente la particularité de pouvoir s'écrire

$$[A, B, C, D, E] \quad = \quad [A]\,[B \mid A]\,[C \mid A, B]\,[D \mid B]\,[E \mid B, C] \tag{6}$$

où la notation $[Y \mid X]$ signifie loi conditionnelle de Y sachant X, c'est à dire la restriction de la distribution de Y lorsqu'on connait la valeur de X. Alors le réseau bayésien associé peut se représenter par le diagramme de la figure 17. On observe que les flèches arrivant à chacun des noeuds sont issues des variables conditionnantes dans l'expression ci-dessus. C'est bien le cas de la variable E pointée par B et C. Grâce à la formule (6) la spécification de la loi conjointe totale se fait par une spécification pour chacune des variables aléatoires[45], demandant pour chacune de considérer un sous-ensemble réduit de toutes les variables aléatoires.

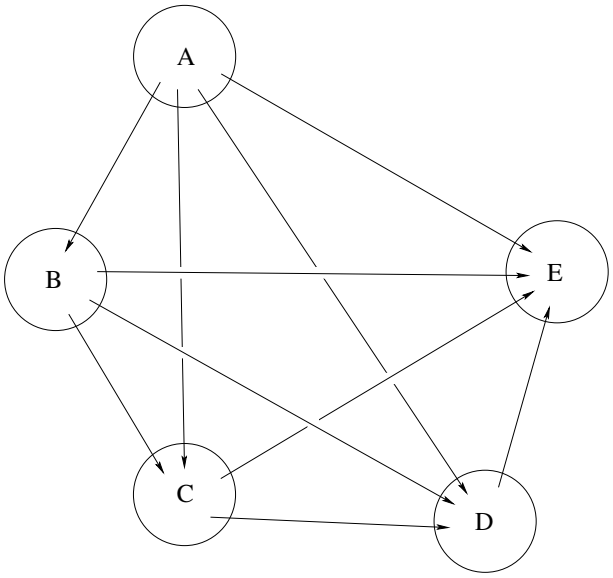
On pourrait vérifier que la forme $[A]\,[B \mid A]\,[C \mid A, B]\,[D \mid B]\,[E \mid B, C]$ ne permet pas, loin de là, de "couvrir" toutes les lois conjointes. Par exemple, D et A sont forcément indépendantes conditionnellement à B puisque

[43] Pour nous, probabiliste et stochastique sont synonymes.

[44] La probabilité conjointe de plusieurs variables aléatoires est la spécification de leur distribution simultanée (par opposition à probabilité ou distribution marginale).

[45] Nommément : $[A]$, $[B \mid A]$, $[C \mid A, B]$, $[D \mid B]$ et $[E \mid B, C]$.

FIG. 18 – Réseau bayésien hiérarchique complet



le réseau spécifie que l'influence de A passe par B : si donc on connaît B, une variation de A n'a aucune conséquence sur celle de D. Il y a donc forte simplification. On peut facilement dans le cas de variables discrètes dénombrer les dimensions paramétriques du modèle complet et du modèle restreint. Si on part d'une forme quelconque de la densité conjointe et qu'on tente d'écrire par produits de densité conditionnelles, on obtient par exemple

$$[A, B, C, D, E] \quad = \quad [A]\,[B \mid A]\,[C \mid A, B]\,[D \mid A, B, C]\,[E \mid A, B, C, D]$$

qui correspond au diagramme du réseau bayésien présenté en figure 18 . C'est déjà un réseau bien compliqué pour 5 variables ; en fait il comporte le maximum de flèches : tous les couples de variables sont reliées. Il serait complètement inextricable pour un nombre plus important bien que raisonnable de variables.

Contrairement aux deux précédents types de modélisation, la notion d'unité statistique n'est pas obligatoire. Une conséquence est que la matrice de variance des observations ne se met plus sous forme bloc diagonale.

$$V\,(Y_*) \quad = \quad \mathbf{\Sigma}$$

mais cette matrice générale possède cependant des propriétés particulières (faute de quoi l'exploitation statistique du modèle serait impossible). Par exemple dans le cas gaussien, son inverse a des zéros pour tous les couples de variables qui ne sont pas reliées par des flèches dans le réseau bayésien (cf. la définition des flèches en §2.1).

discussion La classification proposée ci-dessus en 3 types de modélisation comporte de nombreux et graves défauts. Tout d'abord les classes ne sont pas séparées mais emboîtées. Ensuite, beaucoup de modélisations standard ne sont pas clairement repérées. Ne citons pour exemples que les modèles liés aux mesures répétées, les séries chronologiques,... il y en a beaucoup d'autres. Mais notre but n'est pas d'établir une bonne classification des modélisations mais de faire apparaître la place qu'occupent les réseaux bayésiens. La première indication à retenir est qu'ils autorisent des modélisations beaucoup plus complexes que les modèles de régression. Le tableau 25 récapitule les différences principales entre les trois types de modèles.

10.6 différentes approches statistiques

Reprenons le modèle de régression univariable présenté autour de l'équation (5). Quatre types de variables mathématiques y sont distinguées :

TAB. 25 – Caractéristiques des trois types de modèles

	variables d'intérêt	covariables	structure de variance
modèles univariables	une seule	plusieurs	indépendance
modèles multivariables	plusieurs	plusieurs	indépendance par bloc
réseaux bayésiens	plusieurs	plusieurs	quelconque

1. Y_i : la variable d'intérêt considérée comme une variable aléatoire dont une réalisation est l'observation y_i[46]. Il peut s'agir d'une variable aléatoire continue (rendement d'une parcelle de blé) ou discrète (la qualité d'un gigot appréciée par dégustation),
2. x_i : les covariables continues dont certaines sont de nature aléatoire (mais observées ; exemple le taux d'azote dans le sol au moment du semis d'une culture), d'autres fixes (déterminées par le protocole de recueil des données : exemple la dose d'azote apportée au semis de la même culture). Dans le modèle elles sont considérées comme parfaitement connues, et donc fixées (par opposition à aléatoires),
3. f_i : les covariables discrètes dont certaines sont de nature aléatoire (mais observées ; exemple le sexe de l'agneau), d'autres fixes (déterminées par le protocole de recueil des données : le niveau de la ration alimentaire qui lui est attribué). Mais dans le modèle elles sont considérées comme parfaitement connues, et donc fixées (par opposition à aléatoires),
4. θ : les paramètres du modèle, quantités inconnues qu'il faudra estimer à partir des quantités connues pour spécifier numériquement le modèle.

Pour donner une vision plus concrète, le modèle de régression linéaire simple peut s'écrire, à la manière de (5), comme

$$Y_i \quad = \quad \alpha + \beta x_i + E_i$$

où $\alpha + \beta x_i$ est l'espérance (la tendance centrale) et E_i est le terme d'erreur.

Dans une formalisation qui permet de spécifier les distributions de probabilité, on notera

$$[Y_1, Y_2, ..., Y_I] \quad = \quad \prod_{i=1}^{I} \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(y_i - (\alpha + \beta x_i))^2}{2\sigma^2}\right)$$

ou de manière plus habituelle

$$Y_i \quad \sim \quad N\left(\alpha + \beta x_i, \sigma^2\right)$$

les variables Y_i étant indépendantes entre elles. Dans ce modèle, les Y_i sont les variables d'intérêt, il n'y a qu'une covariable x_i qui est continue et les paramètres sont au nombre de trois : α (terme constant), β (coefficient de régression à appliquer à la covariable x) et σ^2 (variance du terme erreur autour de la droite de régression).

approche statistique classique (fréquentiste) Dans une approche classique, on exploite la distribution de probabilité spécifiée sur les données que l'on note $[Y \mid X, F, \Theta]$ puisqu'elle dépend des paramètres et des valeurs prises par les covariables. C'est au travers d'elle que l'information est extraite des données. On estime (ou teste des hypothèses sur) les paramètres

[46] C'est le propre d'une modélisation statistique que de supposer que certaines observations sont la réalisation de variables aléatoires associées. C'est ainsi qu'on peut espérer déduire des informations sur les paramètres de leurs distributions.

à partir des valeurs connues des covariables, et des valeurs observées de la variable d'intérêt. Dans le cas du modèle de régression simple présenté, on cherchera à estimer les paramètres de la régression (α, β), sans doute à tester l'hypothèse de non effet de la covariable sur la variable d'intérêt $(\beta = 0)$.

approche statistique bayésienne Dans une approche bayésienne, toutes les quantités sont considérées comme des variables aléatoires avec deux catégories principales : les variables qui sont observées (considérées comme parfaitement connues dans le modèle) et celles qui ne le sont pas. Les observations et les covariables[47] font partie de la première catégorie, et les paramètres, de la seconde[48].

Pour toutes ces variables aléatoires, on spécifie des distributions de probabilités. On nomme traditionnellement *vraisemblance* la distribution sur l'ensemble des observations ; elle est conditionnelle aux paramètres et reprend donc la modélisation de l'approche fréquentiste. On nomme *a priori* la distribution de probabilité portée sur les paramètres. Grâce au théorème de Bayes (§9.4), l'ensemble des deux définit une distribution conjointe totale sur variables et paramètres.

Ce présupposé requiert de spécifier la distribution conjointe de toutes ces quantités, on la notera $[Y, X, F, \Theta]$. Pour la vraisemblance, on retrouve la distribution de probabilité spécifiée dans le cadre de l'approche classique $[Y \mid X, F, \Theta]$, c'est à dire la distribution conditionnelle des observations (Y) pour les covariables et les paramètres. Comme ce qui nous intéresse, c'est tirer des inférences sur les paramètres, les seules quantités qui nous sont inconnues, on va le faire en fonction de ce qui est connu, c'est à dire de leur distribution conditionnelle par rapport à ce qui nous est connu : Y, X, F et donc, dans nos notations, de $[\Theta \mid Y, X, F]$. L'application répétée du théorème de Bayes (cf. §9.4) conduit à

$$\begin{aligned}
[\Theta \mid Y, X, F] &= \frac{[Y, X, F, \Theta]}{[Y, X, F]} \\
&= \frac{[Y, \Theta \mid X, F]}{[Y \mid X, F]} \\
&= \frac{[Y \mid X, F, \Theta]\,[\Theta \mid X, F]}{[Y \mid X, F]}
\end{aligned}$$

Observant que $[Y \mid X, F]$ se déduit de $[Y \mid X, F, \Theta]\,[\Theta \mid X, F]$, il suffit d'adjoindre à la vraisemblance de l'approche classique, la distribution $[\Theta \mid X, F]$ que l'on dénomme *distribution a priori de* Θ. La distribution recherchée est $[\Theta \mid Y, X, F]$, dénommée *distribution a posteriori de* Θ, sous entendu de l'information apportée par les observations.

discussion L'approche bayésienne est plus exigente que l'approche classique puisqu'elle requiert, en plus de la vraisemblance utilisée dans l'approche classique, la distribution *a priori* des paramètres du modèle. C'est la base des principales critiques qui lui sont adressées : difficulté et subjectivité de la détermination de cette distribution supplémentaire. Ce document n'est pas le lieu pour entamer une polémique ancienne et loin d'être terminée. Nous sommes partisans de l'approche bayésienne et pensons au contraire que ce défaut est une qualité car il permet d'intégrer des informations *a priori* dans le traitement des données.

Si on reprend l'exemple de la régression linéaire, l'espérance d'une observation est $\alpha + \beta x_i$. Si on admet que les valeurs des deux paramètres sont n'importe quelle valeur, n'est-ce pas admettre que cette espérance peut varier de $-\infty$ à $+\infty$? Est-ce bien raisonnable lorsqu'il s'agit d'un rendement de blé ?

[47] On pourrait objecter que les covariables issues d'un protocole expérimental ne sont pas aléatoires mais (pratiquement) on peut montrer que cette supposition n'a aucun effet sur le traitement statistique car la partie utile de la modélisation sera toute conditionnelle aux valeurs des covariables, et (théoriquement) on peut alléguer que le processus de mise au point d'une expérimentation est un processus complexe aléatoire !

[48] mais on peut y ajouter les données manquantes, des prédictions souhaitées, des effets aléatoires, des erreurs de mesure...

Un dernier point mérite d'être ajouté ici. Malgré une apparente simplicité en termes de formules de distributions de probabilité, la mise en oeuvre de l'approche bayésienne recèle dans le cas général de formidables difficultés numériques qui n'ont pu être abordées qu'assez récemment par le développement de méthodes d'intégration basées sur des techniques MCMC[49] et la puissance suffisante des ordinateurs.

Références

[Jensen] Finn V. Jensen (2001). *Bayesian networks and decision graphs.* Statistics for Engineering and Information Science. Springer-Verlag.

[Neal] Radford M. Neal (1996). *Bayesian learning for neural networks.* Volume 118 of Lecture notes in statistics. Springer-Verlag.

remerciements

Une version de travail de ce document a bénéficié de la relecture éclairée d'Isabelle Albert (unité mét@risk de l'Inra).

[49] Les MCMC (pour Monte Carlo Markov Chain) sont des algorithmes itératifs simulant des chaînes de Markov pour obtenir des distributions de probabilités numériquement. Dans le cas bayésien, il s'agit de calculer les distributions *a posteriori*. Un des problèmes difficiles qu'ils posent est de vérifier la bonne convergence du calcul vers la loi limite.