6-615-09 Analyse de décision
Notes de cours

Erick Delage

24 juillet 2018

Table des matières

Première partie

Évaluer une décision en présence d'incertitude

Chapitre 1

Rappel des Probabilités

Voici un résumé des définitions importantes en théories des probabilités :

— Expérience aléatoire Z : expérience dont le résultat dépend du hasard, e.g., rouler un dé sur une table

— Réalisation (ou éventualité) z : résultat possible de l'expérience aléatoire, e.g., la position du dé lorsqu'il s'arrête de bouger

— Événement A : une situation qui peut avoir lieu lors d'une réalisation, e.g., le côté supérieur du dé présente un chiffre pair

— Événement élémentaire : événement qui n'a lieu que pour une seule réalisation, e.g., le dé présente le chiffre 6

— Un ensemble d'événements probabilisables contient :

  — L'événement certain Ω qui inclut toutes les réalisations

  — L'événement nulle $\emptyset$ qui exclut toutes les réalisations

  — Tous les événements complémentaires : l'événement A implique l'existence de l'événement ≪A n'a pas lieu≫

  — Toutes les unions d'événements : l'existence des événements A et B impliquent l'existence de l'événement ≪A ou B a lieu≫

— Mesure de probabilité $\mathbb{P}(A)$: le potentiel de réalisation que l'on attribue à un événement

— Probabilité conditionnelle, $\mathbb{P}(A|B)$: le potentiel de réalisation que l'on attribue à un événement lorsqu'on connaît la nature d'un second événement

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \text{ et } B)}{\mathbb{P}(B)}$$

— Propriétés d'une mesure de probabilité :

  — Les probabilités sont non-négatives et plus petites ou égales à 1

  — Les probabilités s'additionnent : $\mathbb{P}(A \text{ ou } B) = \mathbb{P}(\text{A}) + \mathbb{P}(\text{B})$ si A et B sont mutuellement exclusifs

  — La somme des probabilités est 1 : $\mathbb{P}(\Omega) = 1$

  — Probabilité du complément : $\mathbb{P}(\bar{A}) = 1 - \mathbb{P}(A)$

  — Décomposition de la probabilité d'un événement :

$$\mathbb{P}(A) = \mathbb{P}(A|B_1)\mathbb{P}(B_1) + \mathbb{P}(A|B_2)\mathbb{P}(B_2) + ... + \mathbb{P}(A|B_n)\mathbb{P}(B_n) \ ,$$

où $B_1, B_2, ..., B_n$ sont des événements mutuellement exclusifs et collectivement exhaustifs

— Théorème de Bayes :

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}$$

— Variable aléatoire : valeur numérique qui dépend du résultat de l'expérience ; $Y : \Omega \to \mathbb{R}$

— Fonction de répartition : mesure la probabilité que la variable ne dépasse pas une valeur ; $F(y) = \mathbb{P}(Y <= y)$

— Loi de densité : mesure la vraisemblance que la variable prenne une valeur ; $f(y) = dF(y)/dy$

— Espérance mathématique : mesure la moyenne de la variable ; $E[Y] = \int yf(y)dy$

— Variance : mesure la diversité des valeurs possibles d'une variable ; $Var[Y] = E[(Y - E[Y])^2]$

— n-ième centile : valeur pour laquelle nous avons $n\%$ de chance que la variable aléatoire soit plus petite, i.e. $y : F(y) = n\%$

— Intervalle de confiance de $p\%$: Intervalle ayant $p\%$ de chance de contenir la variable aléatoire

Chapitre 2

Caractériser l'incertitude par une probabilité

Lorsque nous prenons une décision, il arrive souvent que certains paramètres qui influenceront le résultat de notre décision soient inconnus au moment de la prendre. On peut penser par exemple aux conditions climatiques qui nous font hésiter entre organiser un événement à l'extérieur ou à l'intérieur ; à la demande (ou son élasticité) pour un produit qui nous fait hésiter sur la quantité (ou le prix) d'un produit à produire ; au prix du pétrole qui influence le type de système de chauffage à installer ; la quantité de temps entre deux irruptions volcaniques qui nous fait hésiter à voyager dans les îles de Hawaï, etc. Dans ce cours, on dira qu'un paramètre est ≪incertain≫ lorsqu'il peut prendre différentes valeurs et que la vraisemblance de chacune est calculée par l'entremise de la théorie des probabilités. L'incertitude nous intéressera particulièrement lorsqu'elle donne lieu à la notion de risque : i.e., chaque action a le potentiel de mener à des conséquences positives autant qu'à des conséquences négatives (avec plus ou moins de chances).

Remarque 2.0.1. *Il peut arrivé que le paramètre inconnu au moment de prendre la décision ne puisse pas être considéré aléatoire, i.e. que l'on ne peut concevoir prendre des décisions basées sur une notion de vraisemblance. On dira alors que le paramètre est ≪ambiguë≫ (et non incertain). Ceci peut par exemple avoir lieu dans le cas d'un événement qui a eu lieu mais qui n'a pas encore été observé : e.g., la présence d'une vie intelligente sur une autre planète dans l'univers, l'offre faite sur la maison qui vous intéresse par un autre acheteur potentiel, la quantité de gazoline restant dans votre voiture lorsque la lumière d'avertissement s'allume, etc.*

2.1 Quelques définitions

La théorie des probabilités est employée principalement pour modéliser deux types de représentations de l'incertitude : l'approche fréquentiste et l'approche subjective.

Définition 2.1.1. *L'approche fréquentiste définit une probabilité comme étant la proportion de fois qu'un événement aura lieu si on fait une certaine expérience un nombre infini de fois.*

Cette approche nous permet d'analyser des expériences physiques autant que statistiques. Dans le premier cas, les conditions physiques de l'expérience nous permette de déduire des probabilités de chacun des événements possibles. Imaginons par exemple l'expérience de rouler un dé à six faces sur une table, ou bien prendre une balle à l'aveuglette dans une urne que l'on a vu emplir de balles de même tailles mais de différentes couleurs. Dans chacun des cas, la propriété d'équiprobabilité peut nous permettre de conclure de la probabilité de chacun des événements possibles. Une expérience statistique impliquera plutôt la notion de tendance sur un grand échantillon pris à partir d'une certaine population : e.g. proportion d'accidents de voiture impliquant de jeunes conducteurs entre 16 et 25 ans, cote d'écoute d'une émission de radio, proportion de la population avec droit de vote qui supporte un candidat comme premier ministre.

Définition 2.1.2. *L'approche subjective définit une probabilité comme étant le niveau de conviction du sujet par rapport au potentiel de réalisation d'un événement. Ces probabilités peuvent être mises à jour lorsqu'une nouvelle information est rendue disponible.*

Malgré que cette notion de probabilité ne peut-être facilement mesurée, n'étant perceptible qu'à travers l'interprétation du sujet, elle a la force de pouvoir être appliquée à toute déclaration non-confirmée, qu'elle porte sur un événement aléatoire ou ambiguë. Chacun d'entre nous peut tenter de quantifier le potentiel de réalisation des événements suivants alors qu'il nous serait impossible d'en obtenir une mesure fréquentiste par manque d'un modèle physique ou de notion de répétition d'une expérience spécifique : le professeur se présente au premier cours avec 50$ en poche, le soleil se lève demain, les Canadiens de Montréal se rendent en série.

Exemple 2.1.1. *Imaginons avoir devant nous une urne blanche contenant 100 balles. Après avoir observé dix balles retirées à l'aveuglette de l'urne l'une après l'autre en remettant à chaque fois la balle choisie dans l'urne, nous notons que des 10 balles aperçues, seulement 2 d'entre elles n'étaient pas rouge. Que peut-on déduire de la proportion de balles rouges dans l'urne ?*

- *Le Fréquentiste : Puisque j'ai vu 8 balles rouges, je sais que cette proportion n'est pas 0, autrement je ne peux rien affirmer avec conviction. J'ajouterais peut-être que je rejetterais plus facilement l'hypothèse que la proportion est 10% que celle que la proportion est 80% (voir la notion de test d'hypothèse sur Wikipedia).*
- *Le Subjectif : Originalement, je croyais que toutes les proportions étaient aussi vraisemblables les unes que les autres (i.e., probabilité égale sur chaque proportion). Ce que j'ai observé, me fait déduire qu'il y a plus de chance que cette proportion soit de 80%. J'irais même jusqu'à concevoir que les chances qu'il y ait 80 balles rouges sont de 33%.* [1] *En passant, si les mêmes observations avaient été faites mais avec une urne de couleur rouge, alors j'avoue que la probabilité serait plus élevée que 33%, potentiellement 90%.*

Il est à noter que puisque le fréquentiste est incapable de décrire l'aspect physique de cette expérience (i.e., comment le contenu de l'urne a été assemblé), il ne peut se prononcer sur

1. Ceci est une application du théorème de Baye's présenté à la section 2.3.

la probabilité d'une proportion. Selon lui, la probabilité devrait capturer la fréquence à laquelle une certaine proportion serait réalisée si l'on répétait un nombre infini de fois le processus d'assemblage de l'urne. De son côté, le subjectif peut raisonnablement bien identifier la probabilité de différentes proportions. Malheureusement, toute conclusion qu'il fera sera influencée par sa subjectivité : dans cet exemple, la couleur de l'urne fait croire au subjectif qu'il y a plus de chance que l'urne contienne beaucoup de balles rouges.

2.2 Règles de base

Lorsque nous utilisons la notion de probabilité pour justifier une décision qui est prise, celle-ci doit impliquer une mesure qui satisfait certaines règles de base importantes.

1. Les probabilités sont toutes non-négatives et plus petites que 1
2. La somme des probabilités d'événements mutuellement exclusifs et collectivement exhaustifs est de 1.
3. L'ensemble des probabilités conditionnelles doit être cohérent avec l'ensemble des probabilités marginales
$$P(A\&B) = P(A|B)P(B)$$
4. Décomposition de la probabilité d'un événement :
$$\mathbb{P}(A) = \mathbb{P}(A\&B_1) + \mathbb{P}(A\&B_2) + ... + \mathbb{P}(A\&B_n) \ ,$$
où $B_1, B_2, ..., B_n$ sont des événements mutuellement exclusifs et collectivement exhaustifs

Ces règles de base peuvent être expliquées par l'application de la logique mathématique dans le contexte de l'approche fréquentiste. La règle (1) indique qu'une proportion doit être entre 0 et 100%. La règle (1) indique que la somme des proportions liées à des événements qui ne peuvent pas avoir lieu en même temps et qui ensemble couvrent tout ce qui peut arriver doit donner 100%. La règle 3 dit que la proportion de fois que les événements A et B ont lieu en même temps doit être égale à la proportion de fois que B a lieu fois la proportion de fois que A a lieu parmi les fois où B a eu lieu. La dernière règle découle de

$$\begin{aligned}\mathbb{P}(A\&B_1)+&\mathbb{P}(A\&B_2) + ... + \mathbb{P}(A\&B_n)\\ &= \ \mathbb{P}(B_1|A)P(A) + \mathbb{P}(B_2|A)P(A) + ... + \mathbb{P}(B_n|A)P(A)\\ &= \ (\mathbb{P}(B_1|A) + \mathbb{P}(B_2|A) + ... + \mathbb{P}(B_n|A))P(A) \ = \ \mathbb{P}(A) \ .\end{aligned}$$

Il est de plus à noter qu'une conséquence importante des règles 1 et 2 est que des événements mutuellement exclusifs, collectivement exhaustifs et équiprobables ont nécessairement chacun une probabilité $1/\text{Nbr}$ d'événements d'avoir lieu.

Remarque 2.2.1. *L'approche subjective requiert tout autant le respect de ces règles potentiellement pour assurer que l'interprétation de la mesure de probabilité obtenue soit*

cohérente avec la notion de fréquence. En particulier, si un expert nous explique qu'il considère que le potentiel de succès d'un projet est de 80%, alors nous aimerions en comprendre qu'il estime que parmi une population de projet similaire, 80% d'entre eux serait réussis. Nous pourrons ainsi plus tard faire des calculs mélangeant estimation subjective pour A et estimation fréquentiste pour B, lorsque A et B sont jugés indépendants l'un de l'autre.

Prenons un instant pour introduire un exemple de type «Dutch book» servant de mise en garde pour le respect des règles de base.

Exemple 2.2.1. *Considérons un match des Canadiens contre Boston en série, i.e. qu'il ne peut terminer avec une nulle. Votre subjectivité vous pousse à formuler deux probabilités,* $\mathbb{P}(\textit{Canadiens gagnent}) = 0.6$ *et* $\mathbb{P}(\textit{Boston gagne}) = 0.5$*, qui ne respectent pas une des règles de base (laquelle ?). Considérant que vos décisions sont prises en calculant la valeur espérée, vous pourriez justifier la décision d'accepter les deux paris suivants :*

- *Pari A*
  - *Vous gagnez* $41\$$ *si Canadiens gagnent*
  - *Vous perdez* $60\$$ *si Canadiens perdent*
- *Pari B*
  - *Vous gagnez* $51\$$ *si Boston gagne*
  - *Vous perdez* $50\$$ *si Boston perd*

En effet, il pourrait vous sembler que la valeur espérée du Pari A est de 0.60$ et que le Pari B vaut 0.50$ alors qu'en prenant les deux paris vous vous assurez de perdre 9$ peu importe qui gagne.

$$\textit{Valeur espérée de A} = 41 \cdot 0.6 - 60 \cdot (1 - 0.6) = 0,60$$

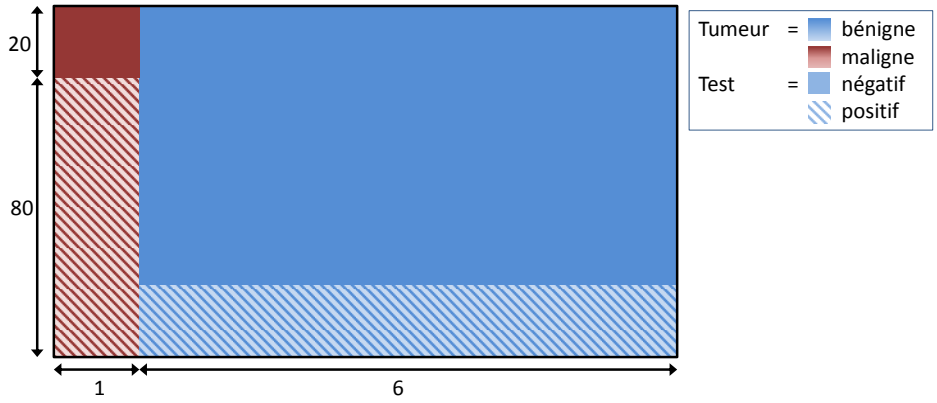
$$\textit{Valeur espérée de B} = 51 \cdot 0.5 - 50 \cdot (1 - 0.5) = 0,50$$

Ceci étant dit, il a été démontré en laboratoire qu'il arrive souvent à l'homme d'enfreindre certaines règles lorsqu'on lui demande d'estimer des probabilités.

> « Un certain test de biopsie a la réputation de bien diagnostiquer la nature bénigne ou maligne d'une tumeur 80% du temps : i.e., une tumeur maligne sera bien identifiée 80% du temps et une tumeur bénigne le sera aussi 80% du temps. En général, 6 fois plus de tumeurs sont bénignes que malignes. Le test vient d'indiquer positif donc que la tumeur serait maligne. Quelle est la probabilité que la tumeur soit réellement maligne ? »

Lorsque cette question est posée à des sujets en laboratoire, les réponses sont très diverses. Que répondriez-vous ? En fait, plusieurs vont répondre un nombre dans les alentours de 40% alors que la réponse logique dicte que cette probabilité devrait être de 80%. Cette erreur d'estimation est souvent associée au problème d'intégration des probabilités de base.

Expliquons la déduction logique qui mène à la valeur de 80% à l'aide de la figure suivante. Cette figure représente une probabilité à l'aide de la notion d'aire couverte. En effet, considérant qu'un point est pris au hasard de manière uniforme sur la surface de ce rectangle, la probabilité qu'un point soit pris dans une région spécifique serait égale à la proportion de l'aire total du rectangle couverte par cette région.



Notez que les quatre régions ont été dessinées en accord avec l'information disponible. La région bleu (uni+barrée) est bel et bien 6 fois plus grande que la région rouge, ce qui est en accord avec les proportions de tumeurs bénigne et maligne. De plus, tel qu'indiqué la proportion des tumeurs malignes identifiées comme maligne par le test est de 80%, et la proportion des tumeurs bénignes identifiée comme bénigne par le test est de 80%. Puisque le test nous indique que la tumeur est maligne, pour répondre à la question, il suffit de déterminer quelle proportion de l'espace associé au résultat positif est couvert par la région associée au fait que la tumeur soit maligne.

$$\begin{aligned}\text{Prop. tumeur maligne si test positif} &= \frac{\text{Prop. tumeur maligne \& test=positif}}{\text{Prop. test=positif}} \\ &= \frac{1/7 \times 8/10}{1/7 \times 8/10 + 6/7 \times 2/10} = 40\%\end{aligned}$$

Ce calcul peut être résumé à l'aide du théorème de Baye's.

2.3 Théorème de Baye's

Le théorème de Baye's est un théorème important du fait qu'il nous indique comment prendre en compte une nouvelle information à propos d'un événement qui nous intéressent.

Théorème 2.3.1. *Imaginons une situation dans laquelle nous avons formulé les probabilités associées à une série d'événements* $\{A_1, A_2, \cdots, A_n\}$ *mutuellement exclusif et collectivement exhaustif, et que nous venons tous juste de recevoir la nouvelle information* B *pour laquelle nous sommes en mesure de caractériser quelle était la probabilité conditionnelle que* B *se réalise sous chacun des événements* $\{A_1, A_2, \cdots, A_n\}$ *: i.e.* $P(B|A_1)$, $P(B|A_2), \cdots, P(B|A_n)$. *Sous ces conditions, l'on peut déduire la nouvelle probabilité de*

voir chacun des événements A_i se réaliser en appliquant la formule suivante :

$$\begin{aligned} P(A_i|B) &= \frac{P(B|A_i)P(A_i)}{P(B)} \\ &= \frac{P(B|A)P(A)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \cdots P(B|A_n)P(A_n)} \end{aligned}$$

Démonstration. Du fait que B est réalisable, autrement la question ne se pose pas, on peut assumer que $P(B) \neq 0$. La règle 3 nous dit que $P(A_i \& B) = P(A_i|B)P(B) = P(B|A_i)P(A_i)$. En divisant des deux côtés de l'égalité par $P(B)$ on obtient la première égalité mentionné. La règle 4 nous permet de reformuler le dénominateur sous la forme

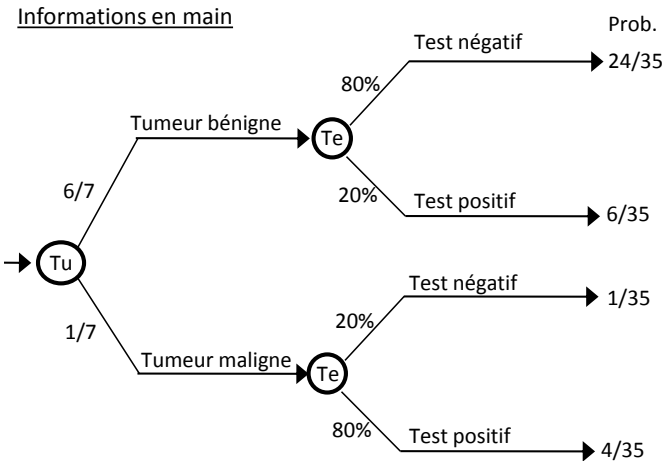
$$P(B) = P(B \& A_1) + P(B \& A_2) \ ,$$

alors que la règle 3 nous donne la forme du théorème. □

Exemple 2.3.1. *Revenons à l'exemple du test de biopsie où nous souhaitions calculer la probabilité que la tumeur soit maligne conditionnellement au fait que nous venons d'apprendre que le test est positif. Dans cet exemple, la nouvelle information est le résultat du test (événements B_1 = négatif et B_2 = positif alors que l'événement pour lequel nous avions initialement caractériser l'incertitude était la nature de la tumeur (événements A_1 = bénigne et A_2 = maligne). En appliquant la formule, nous obtenons :*

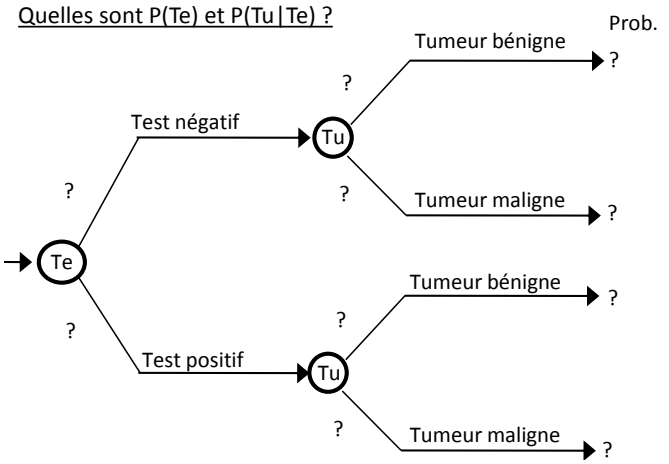
$$\begin{aligned} P(A_2|B_2) &= \frac{P(B_2|A_2)P(A_2)}{P(B_2)} \\ &= \frac{P(B_2|A_2)P(A_2)}{P(B_2|A_1)P(A_1) + P(B_2|A_2)P(A_2)} \\ &= \frac{80\% \cdot 1/7}{20\% \cdot 6/7 + 80\% \cdot 1/7} = 40\% \end{aligned}$$

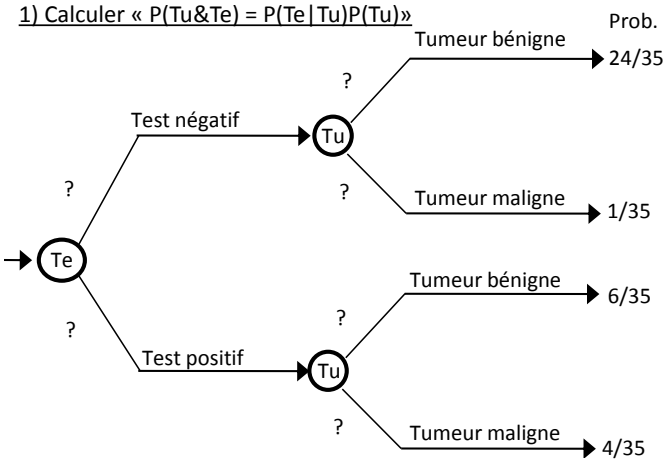
Exemple 2.3.2. *Le même exemple de biopsie peut être étudié en implémentant le théorème de Baye's par la confection de deux arbres de scénarios. D'abord, l'information présentée par l'énoncé peut être capturée par l'arbre de scénario suivant :*

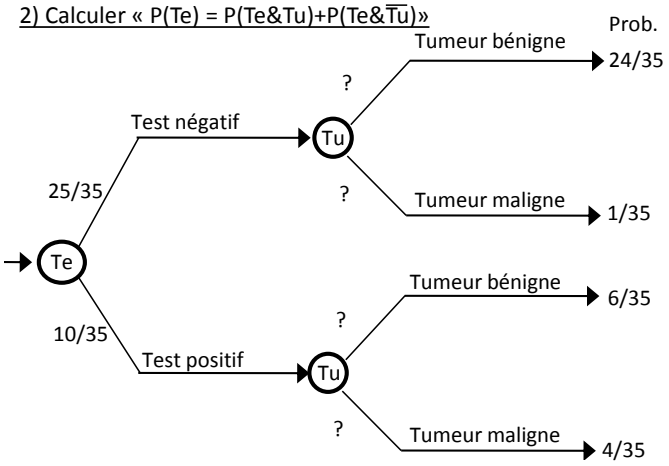


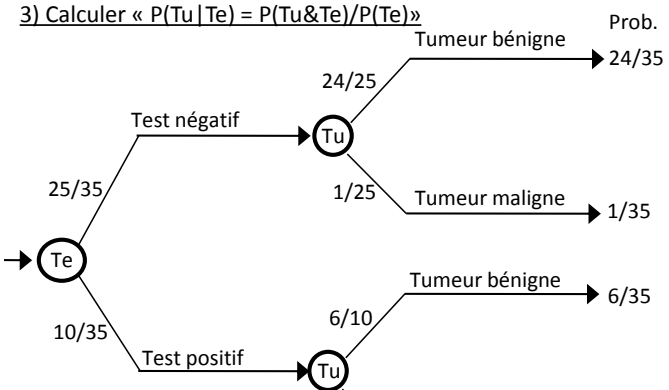
Dans cet arbre, nous pouvons suivre en suivant les noeuds de gauche à droite un ordre spécifique de réception d'information à propos de l'expérience impliquant identité de la tumeur et le résultat du test. En particulier l'ordre décrit que nous apprenons d'abord la nature de la tumeur (A_1 = bénigne, A_2 = maligne) et ensuite le résultat du test (B_1 = négatif, B_2 = positif. Chaque branche indique la probabilité conditionnelle liée à chaque étape de réalisation. Par exemple, la première branche du premier noeud nous dit que la probabilité que la tumeur soit bénigne $P(A_1)$ est de $6/7$, alors que la première branche du premier noeud de la deuxième série nous indique que la probabilité que le test soit négatif si la tumeur est bénigne $P(B_1|A_1)$ est de 80%. Dans la dernière colonne de l'arbre, nous avons calculé la probabilité de chacun des événements élémentaires, i.e. de chaque paire de résultats (e.g. $P(A_1\&B_1) = 24/35$). Puisque la probabilité que nous recherchons $P(A_2|B_2)$ n'est pas dans cet arbre, nous allons nous servir des probabilités d'événements élémentaires pour dériver les quantités qui devraient présenter dans un arbre équivalent qui contient une branche associée à la probabilité conditionnelle que nous cherchons.

Dans cet arbre (voir ci-dessous), il est d'abord possible d'inscrire les probabilités associées aux événements élémentaires dans la dernière colonne simplement en retrouvant ces événements dans l'arbre initial. Nous pouvons ensuite déterminer les probabilités des branches de l'arbre de gauche à droite en calculant la somme des événements élémentaires atteignables à partir de chaque noeud et en divisant cette somme par la probabilité d'atteindre le noeud. Une fois le nouvel arbre complété, nous retrouvons que la probabilité conditionnelle qui nous intéresse est de 40%. Vous aurez remarqué que le travail est plus ardu que par l'application de la formule, mais il donne le même résultat de manière plus intuitive et nous calcule du même coup une gamme de probabilités supplémentaires qui pourraient nous intéresser. Les détails de tous ces calculs sont présentés dans les figures suivantes.


Quelles sont P(Te) et P(Tu|Te) ?
Prob.
Te
Test négatif
Test positif
Tu
Tu
Tumeur bénigne
Tumeur maligne
Tumeur bénigne
Tumeur maligne
?


1) Calculer « P(Tu&Te) = P(Te|Tu)P(Tu)»
Prob.
Te
Test négatif
Test positif
Tu
Tu
Tumeur bénigne
Tumeur maligne
Tumeur bénigne
Tumeur maligne
24/35
1/35
6/35
4/35
?


2) Calculer « P(Te) = P(Te&Tu)+P(Te&$\overline{Tu}$)»
Prob.
Te
25/35
10/35
Test négatif
Test positif
Tu
Tu
Tumeur bénigne
Tumeur maligne
Tumeur bénigne
Tumeur maligne
24/35
1/35
6/35
4/35
?


3) Calculer « P(Tu|Te) = P(Tu&Te)/P(Te)»
Prob.
Te
25/35
10/35
Test négatif
Test positif
Tu
Tu
24/25
1/25
6/10
Tumeur bénigne
Tumeur maligne
Tumeur bénigne
Tumeur maligne
24/35
1/35
6/35

2.4 Indépendance et pertinence entre événement

On dit que l'événement B est indépendant de l'événement A, $A \perp B$, si connaître le statut de A n'affecte aucunement la probabilité que B ait lieu. En d'autre mots, $P(B) = P(B|A)$. La même définition est utilisée pour les variables aléatoires mais dans ce cas il faut le confirmer pour toute valeur que les deux variables aléatoires peuvent prendre. Voici quelques faits intéressants au sujet de la propriété d'indépendance :

— $\ll A$ indépendant de $B \gg$ implique que B est indépendant de A

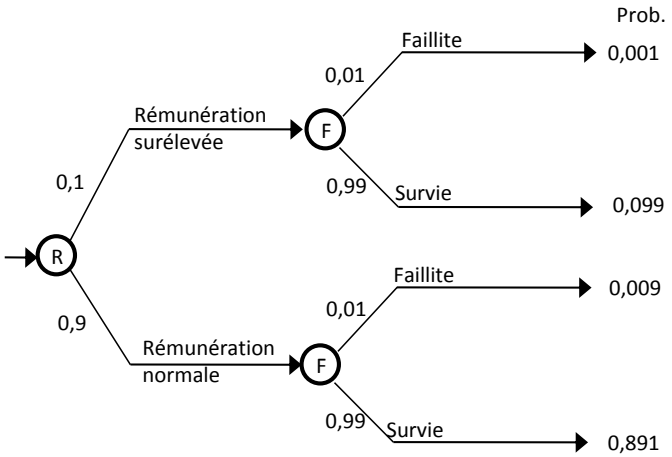
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B)P(A)}{P(B)} = P(A)$$

— A indépendant de B implique que $P(A\&B) = P(A)P(B)$, une définition un peu plus populaire

— Si A et B sont des variables aléatoires, alors A indépendant de B implique que $E[AB] = E[A]E[B]$, et donc que les deux variables ne sont pas corrélées : $E[(A - E[A])(B - E[B])] = E[AB] - 2E[A]E[B] + E[A]E[B] = 0$.

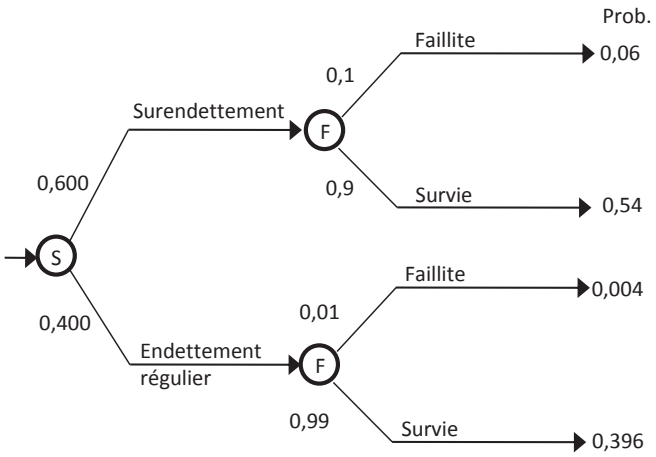
Dans le cas où deux événements ne sont pas indépendants, alors on dira que l'événement A est pertinent à l'événement B, $A \not\perp B$, (i.e. connaître le statut de A influence la probabilité que B ait lieu). Encore une fois, pour la raison évoquée plus haut «A pertinent à B» implique que «B est pertinent à A».

Exemple 2.4.1. *Voici un exemple d'indépendance présenté à l'aide d'un arbre de scénario. Êtes-vous capable de reconnaître si la faillite prochaine d'une entreprise est indépendante de la rémunération d'un CEO ?*

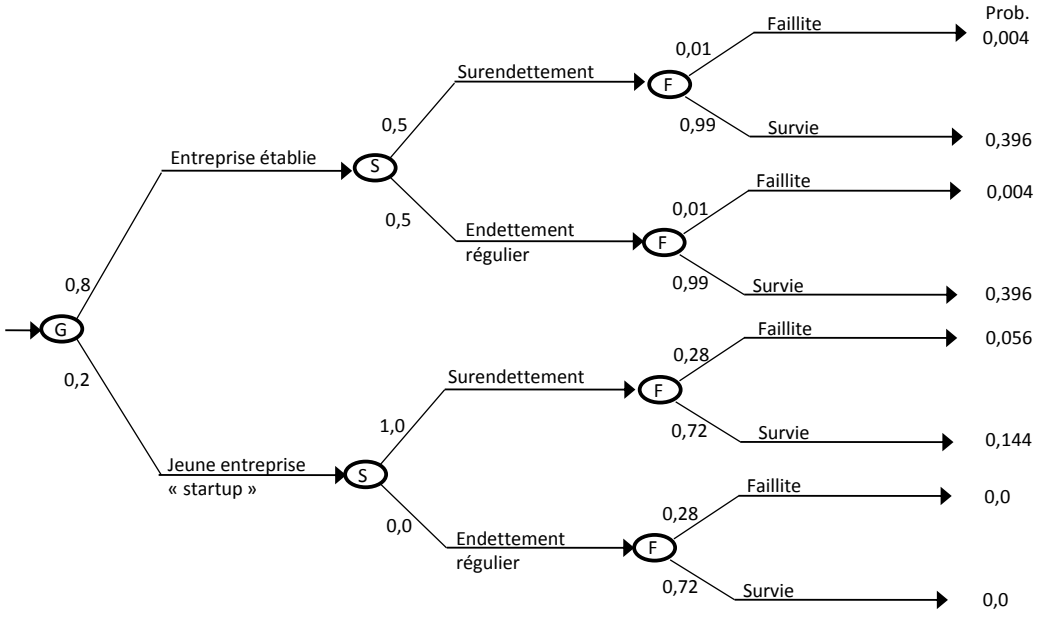


Lorsque deux événements sont pertinents l'un pour l'autre, il est tout de même possible qu'ils deviennent indépendants conditionnellement à la connaissance d'un autre événement explicatif, $A \perp B|C$.

Exemple 2.4.2. *Voici un exemple de pertinence présenté à l'aide d'un arbre de scénario. Êtes-vous capable de reconnaître si le surendettement d'une compagnie est pertinent à la faillite prochaine d'une entreprise ? Remarquez cependant que malgré que ces deux événements*



soit pertinents l'un pour l'autre, ils peuvent devenir indépendant conditionnellement à la connaissance du type d'entreprise (i.e. jeune entreprise ou entreprise établie).



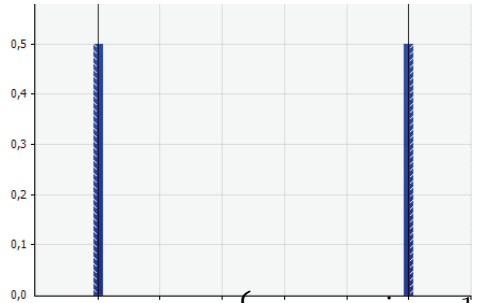
Exemple 2.4.3. *Voici un autre exemple où la relation d'indépendance conditionnelle n'implique pas la relation d'indépendance non-conditionnelle :* $A \perp B|C \not\Rightarrow A \perp B$. *Considérer trois lancer d'un dé à six faces, soit* A *la valeur du premier lancé,* C *la somme des deux premiers lancés,* B *la sommes des trois lancés. Il est clair que nous avons* $A \perp B|C$

puisque si nous connaissons la somme des deux premiers lancés, la valeur du premier lancé ne nous donne aucune information de plus sur la somme des trois lancés. Hors, $A \not\perp B$ *puisque connaître la valeur du premier lancé nous informe du résultat de la somme des trois : si* $A = 6$ *alors* $B > 8$*, etc.*

2.5 Choisir une forme analytique

Certaines formes analytiques sont des choix naturels pour représenter l'incertitude de certains types de processus physique. Lorsque c'est le cas, le choix d'un petit nombre de paramètres permet de définir une mesure de densité sur un espace continu. Nous verrons prochainement qu'il peut être assez facile de déterminer la valeur de ces paramètres à partir de données recueillies. N'oubliez pas cependant que vous êtes toujours responsable du choix de la forme que vous utilisée. Il peut s'avérer nécessaire d'effectuer une analyse de sensibilité pour confirmer que la décision suggérée n'est pas trop sensible au choix de forme de distribution. Dans ce qui suit, nous présentons certaines formes analytiques de distribution en donnant des exemples de situations dans lesquelles elles peuvent s'avérer utiles.

Loi de Bernoulli

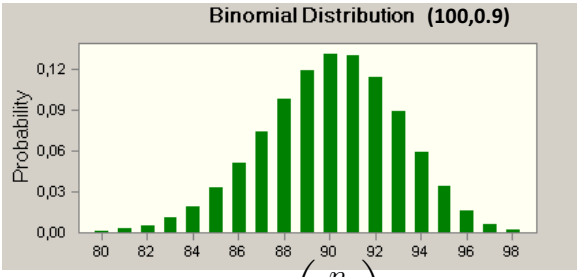


$$P(Z = z; p) = \begin{cases} p & \text{si } z = 1 \\ 1 - p & \text{si } z = 0 \end{cases}$$

— Représente le fait qu'un événement ait lieu ou non. Le paramètre p capture la probabilité d'avoir lieu

— Exemple : le prochain client achètera au moins un produit

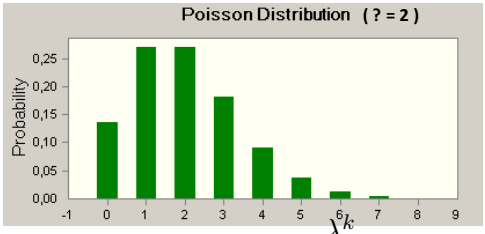
Loi Binomiale



$$P(Z = k; n, p) = \binom{n}{k} p^k (1-p)^{n-k}$$

— Représente le nombre de fois qu'un événement aléatoire aura lieu dans le contexte de n expériences, si cet événement a à chaque fois p probabilité d'avoir lieu
— Exemple : combien de clients achèteront un produit dans un lot de 100 visiteurs

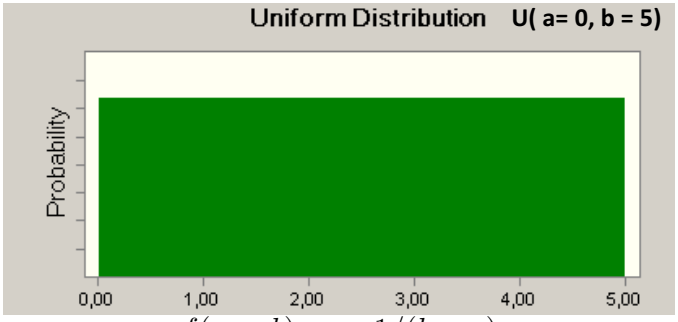
Loi de Poisson



$$P(Z = k; \lambda) = \frac{\lambda^k}{k!} e^{-\lambda}$$

— Représente la prob. que k événements aient lieu dans une période où la quantité moyenne est λ et le délai avant prochain événement est indépendant de ce qui s'est passé jusqu'à date
— Exemple : le nombre de clients se présentant à la boutique entre 5-6pm
— C'est la seule distribution qui satisfait ces critères

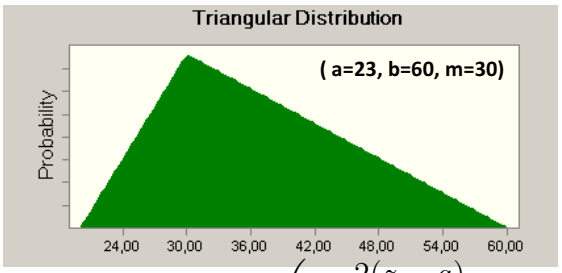
Loi Uniform



$$f(z; a, b) = 1/(b-a)$$

— Représente une incertitude complète par rapport à la position de Z sauf le fait que Z est entre a et b
— Exemple : une livraison aura lieu entre 10-11am

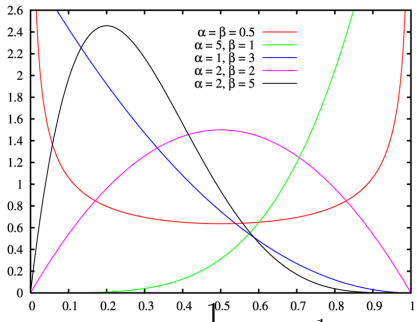
Loi Triangulaire



$$f(z; a, b, m) = \mathbb{1}\{z \in [a, b]\} \min\left(\frac{2(z-a)}{(m-a)(b-a)}, \frac{2(b-z)}{(b-m)(b-a)}\right)$$

— Utilisé couramment lorsque le minimum, le maximum et la valeur dominante de la variable Z sont connus (voir loi bêta)
— Exemple : le projet devrait se terminer entre 20 et 60 jours mais prévu pour 30 jours

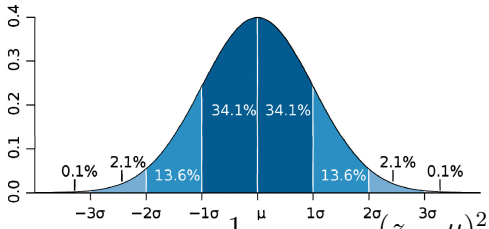
Loi Bêta



$$f(z;\alpha,\beta) = \frac{1}{B(\alpha,\beta)} z^{\alpha-1}(1-z)^{\beta-1}$$

— Utilisé lorsque le min = 0, le max = 1 et la valeur dominante = $\alpha/(\alpha+\beta)$
— Contrairement à la loi triangulaire, on peut représenter le niveau de dominance du mode (i.e., $\alpha + \beta$)
— Exemple : la proportion de mes clients qui aiment mon produit après en avoir interrogé quelques-uns

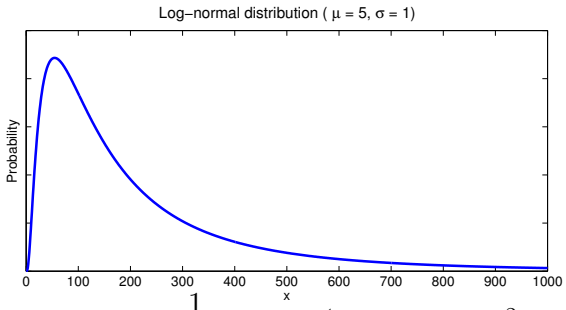
Loi Normale



$$f(z;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(z-\mu)^2}{2\sigma^2}\right)$$

— Selon le théorème central limite, la distribution d'une somme de variables aléatoires (même moyenne et variance) converge vers la loi normale
— Représente bien une population de scores, ou une somme de résultats aléatoires
— Maximise l'entropie lorsque $E[Z]$ et $E[Z^2]$ sont connus

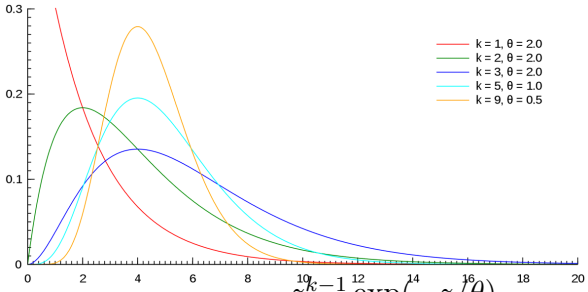
Loi Log-normale



$$f(z;\mu,\sigma) \quad = \quad \frac{1}{\sqrt{2\pi\sigma^2 z^2}} \exp\left(-(\ln(z)-\mu)^2/(2\sigma^2)\right)$$

— Le logarithme d'une variable log-normale suit la loi normale
— Représente bien les multiplications de valeurs aléatoires (théorème central limite)
— Ex. : l'accumulation de changements de valeur d'un titre
— Maximise l'entropie lorsque $E[\ln(Z)]$ et $E[\ln(Z)^2]$ sont connus

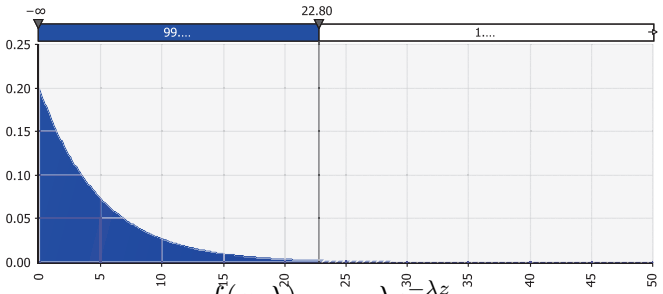
Loi Gamma



$$f(z;k,\theta) \quad = \quad \frac{z^{k-1}\exp(-z/\theta)}{\Gamma(k)\theta^k}$$

— Le délai avant le $k^{\text{ième}}$ événement, chacun suivant la loi exponentielle ($\lambda = 1/\theta$)
— Ex. : le temps d'engager une équipe de 6 personnes
— Maximise l'entropie lorsque $E[Z]$ et $E[\ln(Z)]$ sont connus

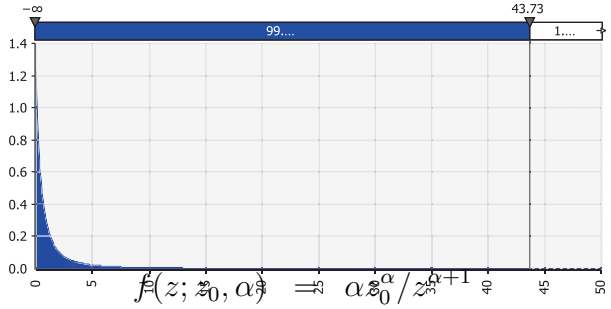
Loi Exponentielle



$$f(z;\lambda) = \lambda e^{-\lambda z}$$

— Représente la durée d'un délai lorsque celle-ci est indépendante du temps déjà passé à attendre (en moyenne $1/\lambda$)
— C'est la seule distribution qui satisfait ces critères
— Exemple : le temps d'attente avant le prochain client
— Maximise l'entropie lorsque $Z \in [0, \infty[$ et $E[Z]$ sont connus

Loi Pareto



$$f(z;z_0,\alpha) = \alpha z_0^{\alpha}/z^{\alpha+1}$$

— La sévérité des dommages reliés à une catastrophe
— Ex. : poursuite légale, risque environnemental, etc.
— Distribution dite «à queue épaisse»

Chapitre 3

Estimation d'une Probabilité par un Expert

Il est parfois impossible de déterminer par nous-même la valeur d'un paramètre qui influencera les retombés de nos décisions. Dans ce cas, il est intéressant de pouvoir avoir recours à l'opinion d'un expert. Dans le meilleur des cas, l'expert peut potentiellement connaître avec certitude la valeur de ce paramètre. Cependant, si ce n'est pas le cas, nous aimerions que l'expert nous donne son avis à propos des valeurs plausibles et de leurs probabilités respectives d'avoir lieu. Nous présenterons ci-dessous une liste de méthodes qui peuvent être utilisées pour obtenir cette information.

3.1 Probabilité d'un événement

D'abord, considérons le cas le plus simple, c'est à dire la validation d'un événement. Pensons par exemple à la prédiction des conditions météorologiques pour une date précise. Même les meilleurs experts météorologues sont incapables de prédire s'il pleuvra ou non à plus d'une journée d'avance (sauf potentiellement l'hiver). Pour cette raison, ils est plus raisonnable d'estimer des probabilités de précipitation. Malheureusement, dans des domaines plus distants de la météorologie, il peut être plus difficile pour un expert d'estimer une probabilité. Une solution possible à ce problème serait de demander à l'expert d'exprimer en mots à quel point il pense que l'événement aura lieu et de traduire les termes utilisés en probabilité tel qu'indiquer dans la table suivante. Que pensez-vous des probabilités suggérées ?

Réponse	Valeur
Improbable	0,1
Peu de chance	0,15
Je doute qu'il aie lieu	0,25
Il y a une chance que	0,5
C'est possible	0,5
Probable	0,65
Il y a de bonne chance que	0,75
Très probable	0,85
Quasiment certain	0,95

En fait, en étudiant cette table, on se rend compte rapidement que l'usage d'une expression comme «il y a peu de chance que» peu être associé à différentes probabilités dépendamment de la personne qui l'emploie ou même de l'événement qui est étudié. Comparez par exemple les deux expressions suivantes :

— Il y a peu de chance que j'arrive en retard au cours.

— Il y a peu de bonne chance qu'il y ai un tremblement de terre d'ici la fin du mois.

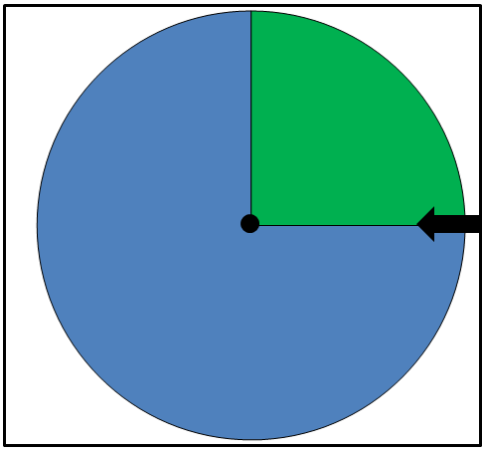
Une alternative serait de demander plutôt la question suivante :

> « Combien de fois plus probable l'événement «A a lieu» est-il comparé à l'événement «A n'a pas lieu» ? »

Selon la réponse obtenue, il est possible de déduire quelle probabilité utilisée. Par exemple, si A est k fois plus probable, alors $P(A) = k/(k+1)$ puisque

$$P(A) = kP(\bar{A}) \quad \Rightarrow \quad P(\bar{A}) + kP(\bar{A}) = 1$$

Dans une situation où l'expert est incapable de se commettre à une quantité, on peut avoir un résultat similaire en employant une «roue de fortune».



Après avoir choisi deux prix de différentes valeurs et qui devrait intéressé l'expert (e.g., un voyage en Europe et un certificat cadeau de 50$), présenter les deux paris suivants :

— Premier Pari :
  — Il gagne le «Grand Prix» si A a lieu
  — Il gagne le «Petit Prix» si A n'a pas lieu
— Deuxième Pari (roue de fortune) :
  — Il gagne le «Grand Prix» si l'aiguille tombe sur le vert
  — Il gagne le «Petit Prix» si l'aiguille tombe sur le bleu

Votre travail consiste à ajuster la taille de la région verte de la roue de fortune jusqu'à ce que l'expert soit indifférent entre les deux paris. En particulier, s'il choisit le deuxième pari alors réduisez la taille de la région verte et vice-versa. À la fin de l'exercice, la proportion de la région verte sur la roue devrait indiquer la probabilité subjective de l'événement A perçu par l'expert.

3.2 Loi de probabilité pour une variable aléatoire

Dans le cas d'une variable aléatoire, la même procédure peut-être répétée pour plusieurs valeurs de la variable afin de déterminer la probabilité que la variable soit plus petite que chacune des valeurs. Une fonction de répartition peut ensuite être estimée. Voici un résumé de cette procédure :

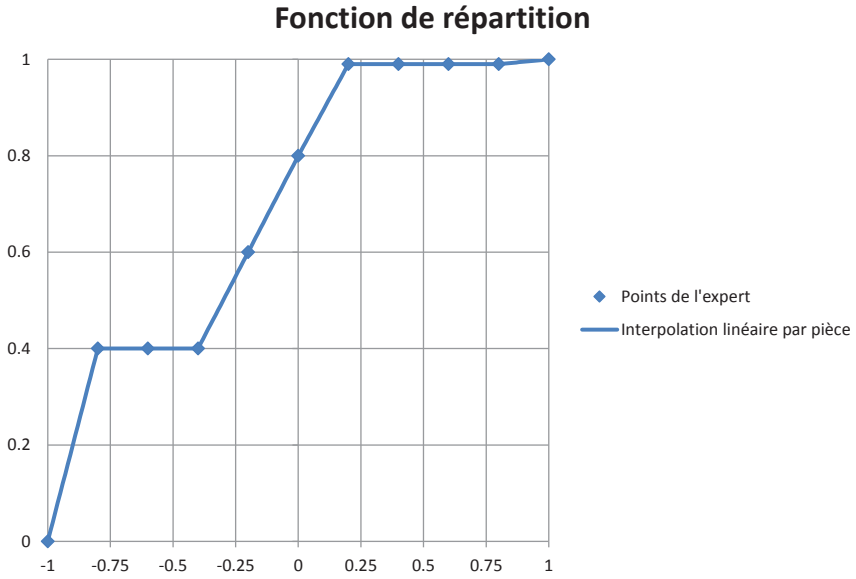
1. Déterminez la valeur minimale et maximale de la variable
2. Discrétisez l'intervalle $[\text{valeur min.}, \text{valeur max.}]$
3. Pour chaque valeur $a_1, a_2, ...$, utilisez la roue des probabilités pour déterminer les probabilités $P(Z \leq a_1)$, $P(Z \leq a_2)$, ...
4. Tracez une courbe continue et non-décroissante en passant par les points trouvés.
5. Confirmez le résultat avec l'expert

Pour estimer une fonction de répartition à partir des points obtenus, il est possible d'utiliser une interpolation linéaire, qui lie chacun des points par une droite, ou d'ajuster les paramètres d'une loi de densité paramétrique pour obtenir le meilleur «fit».

Exemple 3.2.1. *Imaginons que nous avons recueilli les points suivants de notre expert. Par exemple, celui-ci croit qu'il y a 40% de chance que la valeur du paramètre a soit plus petit que -0.8 et 80% de chance qu'il soit plus petit que 0.*

index	*a*	$P(X \leq a)$ *(i.e.* p_i*)*
1	*-1*	*0*
2	*-0.8*	*0.4*
3	*-0.6*	*0.4*
4	*-0.4*	*0.4*
5	*-0.2*	*0.6*
6	*0*	*0.8*
7	*0.2*	*0.99*
8	*0.4*	*0.99*
9	*0.6*	*0.99*
10	*0.8*	*0.99*
11	*1*	*1*

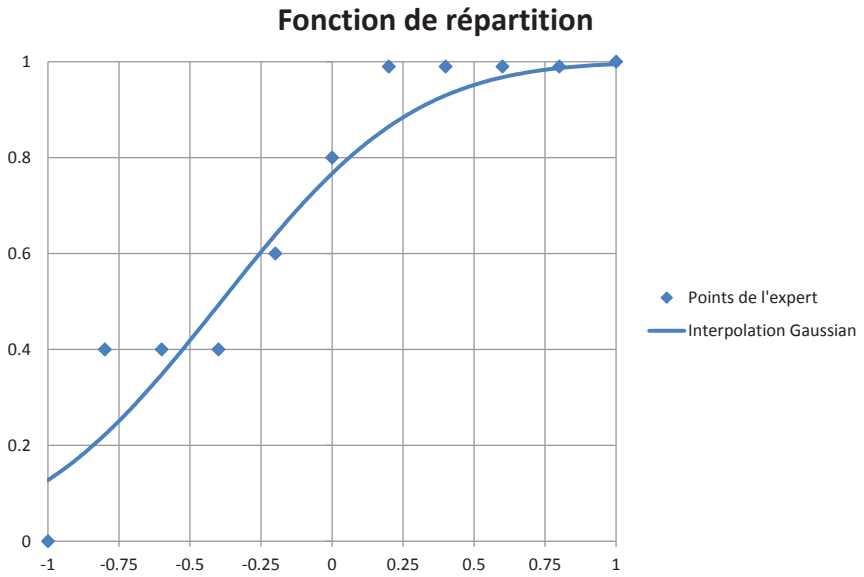
Une interpolation linéaire nous donnera le graphique suivant pour la fonction de répartition.



De son côté, pour obtenir une fonction de répartition de type loi normale, il est possible de résoudre le problème d'optimisation suivant :

$$\min_{\mu,\sigma} \sum_{i=1}^{11} (p_i - F_{\mu,\sigma}(a_i))^2 ,$$

où $F_{\mu,\sigma}(a)$ *est la fonction de répartition de la loi normale dont la moyenne est* μ *et l'écart type* σ *évaluée à la valeur* a*. La fonction de répartition obtenue prend la forme présentée dans la figure suivante. Il est à noter que cette fonction ne passe pas exactement sur les points estimés, ce qui pourrait indiquer qu'il y a eu erreur d'estimation s'il y a de réelles raisons de croire en la loi normale.*



Chacune des deux méthodes proposées à ses avantages respectifs. L'interpolation linéaire capture le fait que si nous savons rien de plus que le fait qu'il y a une probabilité $p_2 - p_1$ que la variable aléatoire prenne une valeur dans l'intervalle $]z_1, z_2]$ alors nous assumons que la variable a autant de chance de prendre n'importe laquelle des valeurs dans cet intervalle. Cette approche respecte donc un certain principe de simplicité : i.e. que souvent les hypothèses les plus simples sont les plus vraisemblables. D'un autre côté, dans certaines situations, il y a des raisons de croire que la variable aléatoire prend une forme paramétrique particulière (e.g. la loi normale). En ajustant, les paramètres de cette forme il est possible de déceler des erreurs d'estimation des points obtenus et de les corriger.

3.3 Loi de probabilité pour un ensemble de variable

Dans le cas d'un ensemble de variable, disons W, X, Y, Z, il est bien sûr possible d'estimer chacune des lois de probabilité marginales en utilisant la technique présentée ci-dessus. Cependant, à moins qu'on fasse l'hypothèse que chacune des variables soit indépendante des autres, il est nécessaire d'estimer les liens de pertinence entre chacune d'entre elles.

L'approche la plus générale permet d'aller estimer tous les liens de pertinence dans leurs moindres détails. Il en coûtera cependant beaucoup de travail. L'idée est de décomposer la loi de probabilité conjointe sur les variables en système de loi de probabilité conditionnelle univariée. En d'autres mots, on emploie la règle de décomposition des probabilité à répétition

pour obtenir :

$$\begin{aligned} P(W,X,Y,Z) &= P(W|X,Y,Z)\cdot P(X,Y,Z) \\ &= P(W|X,Y,Z)\cdot P(X|Y,Z)\cdot P(Y,Z) \\ &= P(W|X,Y,Z)\cdot P(X|Y,Z)\cdot P(Y|Z)\cdot P(Z) \end{aligned}$$

Il nous faut ensuite estimer chacune des lois conditionnelles. En fait, il est important de réaliser qu'une fonction de répartition doit être déterminée pour chaque réalisation potentielle des variables conditionnées. Voici un exemple dans lequel les variables aléatoires sont de type Bernoulli.

Exemple 3.3.1. *Imaginons quatre variables aléatoires W,X,Y et Z de type Bernoulli, i.e. que chacune ne prend qu'une valeur parmi zéro et un. Dans ce cas, si nous souhaitons modéliser tous les liens de pertinences, alors il est nécessaire de compléter les tableaux suivants :*

$P(Z=1)$
?

Z	$P(Y=1\|Z)$
0	*?*
1	*?*

Y	Z	$P(X=1\|Y,Z)$
0	*0*	*?*
0	*1*	*?*
1	*0*	*?*
1	*1*	*?*

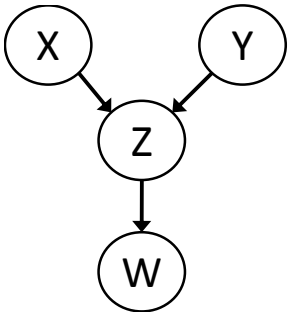
X	Y	Z	$P(W=1\|X,Y,Z)$	X	Y	Z	$P(W=1\|X,Y,Z)$
0	*0*	*0*	*?*	*1*	*0*	*0*	*?*
0	*0*	*1*	*?*	*1*	*0*	*1*	*?*
0	*1*	*0*	*?*	*1*	*1*	*0*	*?*
0	*1*	*1*	*?*	*1*	*1*	*1*	*?*

Dans ces tableaux, nous avons omis l'estimation des probabilités que chacune des variables prenne la valeur de zéro puisque celle-ci peut-être déduite directement : $P(Z=0)=1-P(Z=1)$, $P(Y=0|Z=0)=1-P(Y=1|Z=0)$, etc. Dans cet exemple, il est nécessaire d'estimer 15 probabilités distinctes.

Vous serez probablement d'avis que ce type de procédure d'estimation devient beaucoup trop encombrante dans un problème réel impliquant plus d'une trentaine de variables aléatoires, surtout si certaines d'entre elles sont associées à un espace continue de réalisations. Heureusement, la tâche devient plus facile lorsqu'il est possible de faire des hypothèse d'indépendance (potentiellement conditionnelle) entre certains ensemble de variables. Nous exploiterons la structure d'un diagramme d'influence pour représenter ces indépendances.

Théorème 3.3.1. *Dans un diagramme d'influence qui respecte la propriété de Markov globale, une variable est indépendante de toutes les variables qui ne sont pas en aval si l'on connaît ni plus ni moins que la valeur des variables qui l'influencent directement.*

Prenons l'exemple de diagramme d'influence suivant.



Si l'on assume que ce diagramme respecte la propriété de Markov globale, alors nous pouvons en conclure que $X \perp Y$ et que $W \perp X, Y|Z$. Nous avons donc que

$$\begin{aligned} P(W, X, Y, Z) &= P(W|X, Y, Z)P(Z|X, Y)P(Y|X)P(X) \\ &= P(W|Z)P(Z|X, Y)P(Y)P(X) \end{aligned}$$

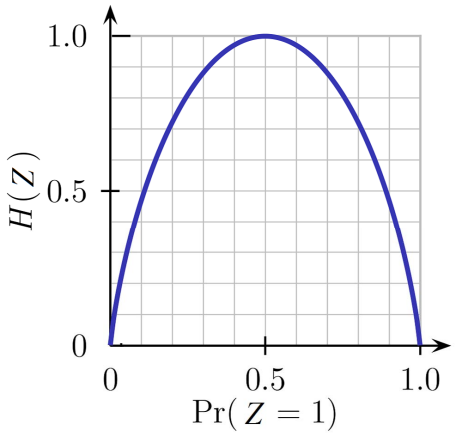
Dans le cas où les variables sont toutes de types Bernoulli, alors il nous suffit d'estimer 8 probabilités au lieu de 15. Pouvez-vous confirmer ce calcul ?

Remarque 3.3.1. *Dans un diagramme d'influence qui respecte la propriété de Markov globale, si l'on connaît plus que la valeur des variables qui influence directement une variable X, il est possible qu'une variable qui n'est pas en aval de cette variable soit pertinente à la connaissance de X. Par exemple, le diagramme ci-dessus capture seulement le fait que X et Y sont indépendant lorsqu'aucune autre information n'est disponible. On ne peut conclure que «X est indépendant de Y si on connaît Z». Au contraire, si Z est connu, la connaissance de X pourrait être pertinente à la connaissance de Y. Disons que la relation d'influence de Z et Y sur Z serait $Z = X \cdot Y$. Sachant que $Z = 0$, l'information supplémentaire que $X = 1$ nous indiquerait que Y doit être nécessairement égale à 0. Sans la connaissance de $X = 1$, Y pourrait être distribuée sur différentes valeurs.*

3.4 Qualité d'une loi de probabilité subjective

Il existe deux critères de qualité importants pour une loi de probabilité subjective. Premièrement, cette loi doit être informative, c'est-à-dire qu'elle doit considérer certaines réalisations plus plausibles que d'autres ou même être différentes de celles que nous aurions estimées nous-même. Par exemple, la probabilité d'un événement devrait s'éloigner de 50%, sinon l'expert ne nous informe pas quel événement est le plus probable. De plus, une densité estimé devrait être concentrée plutôt qu'uniforme. En fait, il est possible de mesurer la quantité d'information présentée dans une loi de probabilité en mesurant son entropie (moins d'entropie = plus d'information)

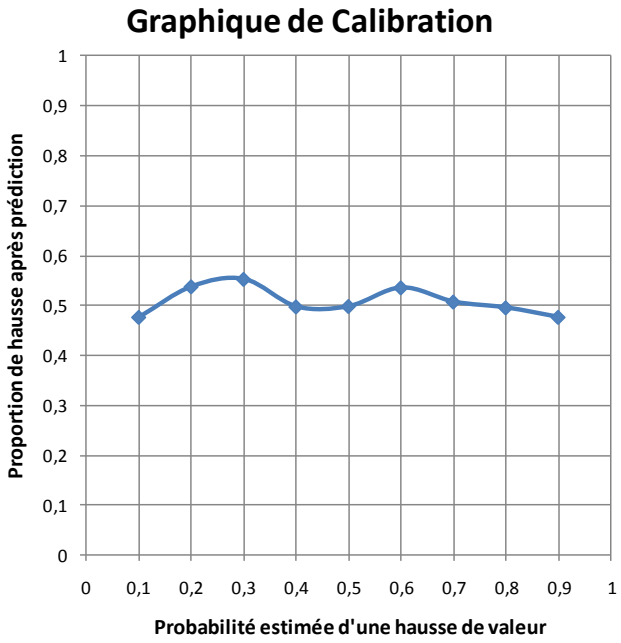
$$H(Z) := \sum_i P(Z = z_i) \log_2(1/P(Z = z_i))$$



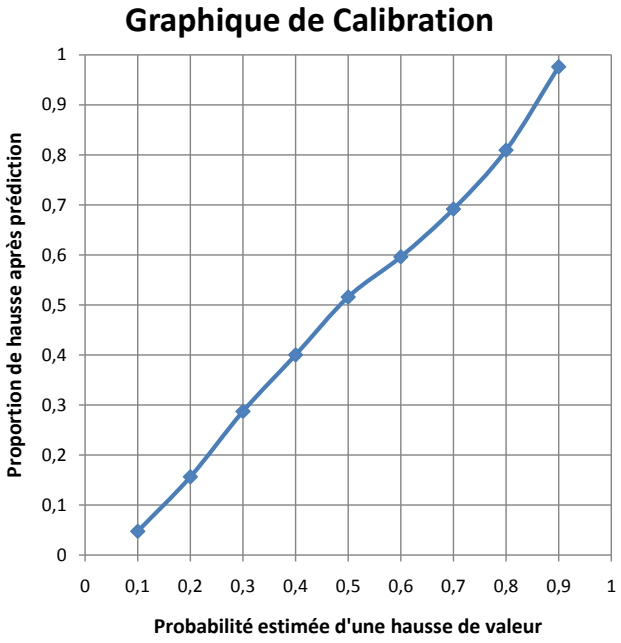
Deuxièmement, la loi de probabilité devrait représenter l'incertitude de manière authentique. En quelque sorte, nous aimerions vérifier que l'expert dit vrai. Nous ne pouvons cependant pas accuser l'expert si un événement n'a pas lieu malgré qu'il estimait la probabilité à plus de 50%, autrement l'expert ne se prononcerait tout simplement pas. Nous allons donc plutôt vérifier si les prédictions de l'expert sont bien «calibrées». C'est à dire qu'à long terme, parmi les événements que l'expert prévoyait avoir $p\%$ de chance d'avoir lieu, la proportion de ces prévisions pour lesquelles l'événement en question a réellement eu lieu approche la valeur de p.

Exemple 3.4.1. *Voici deux exemples de graphiques de calibration des prédictions de deux experts différents à qui nous avons demandé durant une certaine période d'estimer à chaque jour la probabilité que la valeur du titre de Microsoft prenne de la valeur le lendemain. On peut remarquer que le premier expert est plutôt mal calibré : parmi toutes les fois où il prévoyait 10% de chance d'une hausse, près de la moitié du temps le titre augmentait de valeur. Le deuxième expert était significativement meilleur puisque parmi les fois où il prévoyait 10% de chance d'une hausse de la valeur du titre, on n'a pu observer le titre augmenter de valeur que 5% des fois.*

Expert mal calibré



Expert mieux calibré



Chapitre 4

Caractériser une loi à partir de données

Avec l'ère du ≪Big Data≫, il devient de plus en plus intéressant de remplacer (ou de complémenter) l'information obtenue d'experts par l'acquisition de données. En effet, les experts ne sont pas toujours faciles à trouver ou peuvent vendre leur expertise à gros prix alors que il est de plus en plus facile de recueillir de l'information, de l'emmagasiner et de la traiter à l'aide de logiciels de fine pointe. Pensons par exemple à tout ce qui est lié au commerce électronique où des millions de clients visitent les sites de ventes chaque jour. Il est donc de plus en plus commun d'avoir accès à une multitude de données historiques. Les données peuvent aussi parfois être achetées comme dans le cas de la valeur des titres boursiers, la météo, ou même à travers l'usage de sondages. Comparativement à l'opinion d'un expert, l'usage de données peut donner une nature un plus objective aux conclusions qui sont obtenues.

4.1 Calibration de modèles stochastiques

Dans cette section, nous réviserons brièvement quelques méthodes de calibration utilisées en pratique. Nous considérerons d'abord des méthodes permettant d'inférer une loi de probabilité marginale, ainsi qu'une loi de probabilité conditionnelle pour modéliser les liens de pertinences entre variables.

4.1.1 Alignement d'histogramme

Considérons que nous souhaitons modéliser la distribution marginale $P(Z)$ à l'aide de données $\{z_1, z_2, ..., z_M\}$. Nous ferons l'hypothèse que $P(Z)$ prend la forme paramétrique $f(z;\theta)$ et que toutes les données ont été générées indépendamment selon la même loi. La méthode d'alignement d'histogramme recommande de diviser l'intervalle de valeur atteignable par Z en sillons $\{[a_1, b_1], [a_1, b_1], \ldots, [a_N, b_N]\}$ et d'identifier la valeur des paramètres de $f(z;\theta)$ qui minimisent la différence entre l'histogramme de $f(z;\theta)$ sur les sillons choisis

et l'histogramme empirique basé sur les données :

$$\underset{\theta}{\text{minimiser}} \sum_i (P_\theta(a_i \leq Z \leq b_i) - \hat{P}(a_i \leq Z \leq b_i))^2$$

où $P_\theta(\cdot)$ est la probabilité selon $f(z;\theta)$, $\hat{P}(\cdot)$ est la probabilité selon l'échantillon de données. Chaque terme de l'objectif calcule la distance entre la probabilité prévue par la forme paramétrique et celle prévue par la fréquence calculée dans les données. L'erreur totale calculée peut être utilisée pour confirmer que la forme paramétrique choisie est un bon «fit» pour ces données. S'il y a lieu d'hésiter entre plusieurs formes pour $f(z,\theta)$, alors il y a intérêt à choisir la forme qui mène à l'erreur la plus basse. Malheureusement, la méthode d'alignement d'histogramme s'adapte mal aux distributions multi-variées puisque dans ce cas il faut utiliser un histogramme construit sur le treillis d'un espace multi-dimensionnel.

Exemple 4.1.1. *Dans le cas où l'on assumerait que les données sont générées à partir de la loi normale, nous pourrions utiliser les sillons*

$$\{[-\infty,-10],[-10,-6],[-6,-2],[-2,2],[2,6],[6,10],[10,\infty]\}\ ,$$

dans ce cas le problème de calibration devient :

$$\underset{\mu,\sigma}{\text{minimiser}} \ \sum_i (\int_{a_i}^{b_i} (1/\sqrt{2\pi\sigma^2}) e^{-\frac{(x-\mu)^2}{2\sigma^2}} - N_i/M)^2$$

où N_i est le nombre d'échantillons observés dans le sillons $[a_i, b_i]$.

Remarque 4.1.1. *La statistique Chi-Square peut-être utilisée pour calculer la probabilité qu'une variable Chi-square avec une liberté de degré "nbr sillon - 1" prenne une valeur plus grande que la valeur d'erreur obtenue. Si cette probabilité (connue sous le nom de P-value) est plus petite que 0.05 on devrait s'inquiéter de la validité de l'hypothèse que les données ont été générées à partir de cette forme paramétrique.*

4.1.2 Maximisation de la vraisemblance

Similairement à la méthode décrite ci-dessus, nous considérerons que nous souhaitons modéliser la distribution marginale $P(Z)$ à l'aide de données $\{z_1, z_2, ..., z_M\}$. Nous ferons l'hypothèse que $P(Z)$ prend la forme paramétrique $f(z;\theta)$ et que toutes les données ont été générées indépendamment selon la même loi. La méthode de maximisation de la vraisemblance suggère de choisir les paramètres qui maximisent la vraisemblance des réalisations de Z exprimées dans les données. Spécifiquement, il nous faut résoudre le problème d'optimisation suivant :

$$\underset{\theta}{\text{maximiser}} \ \prod_{i=1}^{M} f(z_i;\theta)\ .$$

Notez que le terme $\prod_{i=1}^{M} f(z_i;\theta)$ capture la vraisemblance d'observer conjointement les valeurs z_1,z_2,...,z_M après M observations indépendantes. En pratique, le problème est équivalent au problème

$$\underset{\theta}{\text{maximiser}} \quad \sum_{i=1}^{M} \ln(f(z_i;\theta))$$

puisque la fonction $\ln(\cdot)$ est strictement croissante. Cette dernière forme est plus communément utilisée puisqu'elle est relativement plus facile à résoudre.

Exemple 4.1.2. *Imaginons que Z soit une épreuve de Bernoulli, dont p est la probabilité de succès. Dans ce cas, le problème de maximisation de vraisemblance prend la forme*

$$\underset{p}{\text{maximiser}} \prod_{i} p^{z_i}(1-p)^{1-z_i} \ ,$$

où chaque terme multiplié prend la valeur de $p^1(1-p)^{1-1} = p$ si l'observation était un succès, sinon la valeur $p^0(1-p)^{1-0} = 1-p$. En étudiant la forme du problème qui se décompose en une somme sur chacune des observations, on peut démontrer que le paramètre optimal est nécessairement $p^* = \frac{\sum_i z_i}{M}$. En d'autres mots, la probabilité de succès la plus vraisemblable est celle qui décrit la fréquence de succès observée. Le résultat peut être obtenu en confirmant que l'objectif du problème suivant est concave pour toutes les valeurs de p entre 0 et 1, et en vérifiant que sa dérivée première atteint la valeur de zéro au point p^*.

$$\underset{p}{\text{maximiser}} \sum_{i=1}^{M} \ln\left(p^{z_i}(1-p)^{1-z_i}\right).$$

$$f(p) = \sum_i \ln\left(p^{z_i}(1-p)^{1-z_i}\right) = \sum_i z_i \ln(p) + (1-z_i)\ln(1-p)$$

$$df(p)/dp = \sum_i \frac{z_i}{p} - \frac{1-z_i}{1-p} = \frac{1}{p(1-p)} \sum_i z_i - pz_i - p + pz_i = \frac{\sum_i (z_i - p)}{p(1-p)}$$

$$df(p)/dp = 0 \Leftrightarrow \sum_i (z_i - p) = 0 \Leftrightarrow (1/M) \sum_i z_i = p$$

Exemple 4.1.3. *Imaginons maintenant que Z prenne une valeur parmi les scénarios $\{s_1, s_2, ..., s_K\}$ et que chaque scénario s_j a une probabilité p_j d'avoir lieu ; la somme des p_j doit nécessairement donner 1. La variable aléatoire Z a donc la fonction de masse paramétrique suivante : $P(Z = z) = \prod_{j=1}^{K} p_j^{\mathbf{1}\{z=s_j\}}$. Notez que le produit de cette fonction ne sert qu'à retourner la valeur de la probabilité associée au scénario : par exemple, si le scénario est s_2 alors la fonction retourne $p_1^0 p_2^1 p_3^0 p_4^0 ... p_K^0 = p_2$. Pour cette forme paramétrique, le problème de maximisation de vraisemblance prend la forme*

$$\underset{p}{\text{maximiser}} \prod_{i=1}^{M} \prod_{j=1}^{K} p_j^{\mathbf{1}\{z_i=s_j\}} \ .$$

Similairement à ce qui est établit pour la loi d'une épreuve de Bernoulli, on peut démontrer que les paramètres optimaux sont nécessairement $p_j^* = \frac{\sum_i \mathbf{1}\{z_i = s_j\}}{M}$. *En d'autres mots, la probabilité du scénario* j *la plus vraisemblable est nulle autre que celle qui décrit la fréquence observée du scénario. Le résultat est plus difficile à obtenir mais en voici tout de même les détails. Il est obtenu en confirmant que l'objectif du problème d'optimisation*

$$\begin{aligned} \underset{p}{\text{maximiser}} \quad & \sum_i \sum_j \mathbb{1}\{z_i = s_j\} \ln(p_j) \\ \text{sous contraintes} \quad & \sum_j p_j = 1 \end{aligned}$$

est concave pour toutes les valeurs de p_j *entre 0 et 1, et en trouvant le point stationnaire du Lagrangien* $L(p, \lambda) = \sum_i \sum_j \mathbb{1}\{z_i = s_j\} \ln(p_j) + \lambda(1 - \sum_j p_j)$.

$$\begin{aligned} \partial L(p, \lambda)/\partial p_j &= \frac{\sum_i \mathbb{1}\{z_i = s_j\}}{p_j} - \lambda = 0 \\ \partial L(p, \lambda)/\partial \lambda &= 1 - \sum_j p_j = 0 \end{aligned}$$

Donc, nous devons avoir que $\lambda p_j = \sum_i \mathbb{1}\{z_i = s_j\}$ *et que* $\sum_j p_j = 1$. *En sommant la première expression sur tout les* j, *nous obtenons que* $\lambda \sum_j p_j = \sum_j \sum_i \mathbb{1}\{z_i = s_j\}$ *et ainsi* $\lambda = \sum_j \sum_i \mathbb{1}\{z_i = s_j\}$ *puisque la somme des* p_j *est de un. On en conclut que que* $\lambda = \sum_j \sum_i \mathbb{1}\{z_i = s_j\}$ *et que* $p_j = \frac{\sum_i \mathbf{1}\{z_i = s_j\}}{\sum_j \sum_i \mathbf{1}\{z_i = s_j\}}$. *En d'autres mots, que* λ *compte le nombre d'observations et que* p_j *est la fréquence du scénario* s_j *dans l'échantillon observé.*

Exemple 4.1.4. *Imaginons que* Z *soit une variable Gaussienne, dont* μ *est la moyenne et* σ *est l'écart type. Dans ce cas, le problème de maximisation de vraisemblance prend la forme*

$$\underset{p}{\text{maximiser}} \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(z_i - \mu)^2}{2\sigma^2}} \ .$$

En étudiant la forme du problème qui se décompose en une somme sur chacune des observations, on peut démontrer que les paramètres optimaux sont nécessairement $\mu^* = \frac{\sum_i z_i}{M}$ *et* $\sigma^* = \sqrt{\frac{\sum_i (z_i - \mu^*)^2}{M}}$. *En d'autres mots, la moyenne et l'écart type les plus vraisemblables sont les valeurs observées dans le jeu de données. Ce résultat peut être obtenu en confirmant que l'objectif du problème suivant est concave par rapport à* μ *et* σ, *et en vérifiant que sa dérivée*

première atteint la valeur de zéro au point μ^* et σ^*.

$$f(\mu,\sigma) = \sum_i \ln\left((1/\sqrt{2\pi\sigma^2})e^{-\frac{(z_i-\mu)^2}{2\sigma^2}}\right) = \sum_i -\frac{(z_i-\mu)^2}{2\sigma^2} - \ln(\sqrt{2\pi\sigma^2})$$

$$df(\mu,\sigma)/d\mu = \sum_i 2\frac{(z_i-\mu)}{2\sigma^2}$$

$$df(\mu,\sigma)/d\mu = 0 \Leftrightarrow \sum_i (z_i-\mu) = 0 \Leftrightarrow (1/M)\sum_i z_i = \mu$$

$$df(\mu,\sigma)/d\sigma = \sum_i 2\frac{(z_i-\mu)^2}{2\sigma^3} - \frac{1}{\sqrt{2\pi\sigma^2}}\frac{1/2}{\sqrt{2\pi\sigma^2}}4\pi\sigma = \sum_i \frac{(z_i-\mu)^2}{\sigma^3} - (1/\sigma)$$

$$df(\mu,\sigma)/d\sigma = 0 \Leftrightarrow \sigma^2 = (1/M)\sum_i (z_i-\mu)^2$$

4.1.3 Régression Linéaire

Nous souhaitons modéliser la probabilité conditionnelle $P(Z|X,Y)$ à l'aide de données

$$\{(x_1,y_1,z_1),(x_2,y_2,z_2),...,(x_M,y_M,z_M)\} \ .$$

L'approche par régression linéaire nous demande de faire l'hypothèse que la variable Z peut être expliquée par la structure suivante :

$$Z = \theta_0 + \theta_1 X + \theta_2 Y + \epsilon \ ,$$

où la variable aléatoire ϵ est considérée indépendante de X et Y et est considérée être «concentrée» autour de 0. En fait, nous assumerons que la notion de concentration sous-entend que la variance de ϵ est petite : $E[|\epsilon|^2] \approx 0$. En d'autres mots, on considère que la variable Z est une fonction affine (i.e. linéaire + constante) de X, Y et une perturbation ϵ. La méthode suggère donc de choisir les paramètres qui minimisent la concentration empirique de ϵ telle que mesurée avec les données :

$$\underset{\theta}{\text{minimiser}} \frac{1}{M}\sum_{i=1}^{M} \|\epsilon_i\|^2 \quad \equiv \quad \underset{\theta}{\text{minimiser}} \frac{1}{M}\sum_{i=1}^{M} \|z_i - \theta_0 - \theta_1 x_i - \theta_2 y_i\|^2 \ .$$

Cette optimisation peut-être faite de manière très efficace même lorsque le jeu de données est très large. Afin de compléter le modèle stochastique pour $P(Z|X,Y)$, il est nécessaire de faire une dernière étape ; i.e., modéliser l'aspect stochastique de la relation de pertinence entre X, Y et Z. Il faut en effet s'assurer de modéliser $P(\epsilon)$ à l'aide d'une méthode telle que celle par alignement des histogrammes ou celle par maximisation de la vraisemblance en utilisant comme donnée $\{\epsilon_1,\epsilon_2,...,\epsilon_M\}$ où $\epsilon_i = z_i - \theta_0 - \theta_1 x_i - \theta_2 y_i$.

Remarque 4.1.2. *Noter que malgré que la méthode se dénomme régression linéaire, il est possible de modéliser des relations d'influence non-linéaires. En effet, il s'agit simplement*

de faire la régression sur un ensemble augmenté de variables décrivant différent type d'influences. Il est possible par exemple de modéliser les influences quadratiques en utilisant comme modèle d'influence :

$$Z = \theta_0 + \theta_1 X + \theta_2 Y + +\theta_3 X^2 + \theta_4 XY + \theta_5 Y^2 + \epsilon \ .$$

Si l'on pense que la relation d'influence est différente lorsque X est positif versus lorsque X est négatif, on peut utiliser le modèle suivant :

$$Z = \theta_0 1_{X\geq 0} + \theta_1 X 1_{X\geq 0} + \theta_2 Y 1_{X\geq 0} + \theta_3 1_{X<0} + \theta_4 X 1_{X<0} + \theta_5 Y 1_{X<0} + \epsilon \ .$$

Cette dernière approche est en fait similaire à l'idée de faire deux régression après avoir séparé les données en deux groupes, celles où $x_i \geq 0$ et celles où $x_i < 0$. En générale, le modèle de pertinence peut prendre n'importe quelle forme du type :

$$Z = \sum_j \theta_j h_j(X, Y)$$

où chaque $h_j(X, Y)$ va extraire une certaine caractéristique de la paire (X, Y).

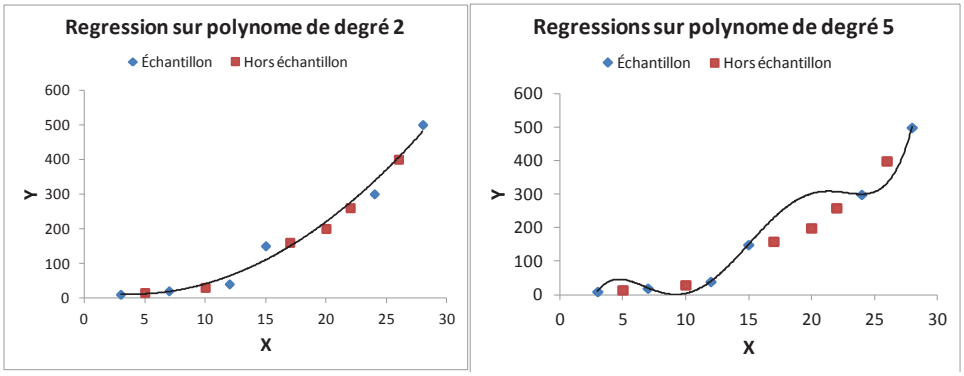
4.1.4 Attention au Surapprentissage

Lorsque le nombre de paramètres augmente, il est nécessaire de calibrer le modèle stochastique en utilisant une quantité plus grande de données. En règle générale, il est recommander d'utiliser au moins 10 fois plus de données qu'il y a de paramètres à calibrer. Une méthode communément employée est de valider la performance du choix de modèle et de ses paramètres sur des données neuves, i.e. qui n'ont pas été utilisées pour faire ces deux choix. Lorsqu'il y a peu de données disponibles, il est possible que cette performance s'améliore en réduisant la complexité du modèle (i.e. en fixant certains paramètres à zéro lors de l'optimisation).

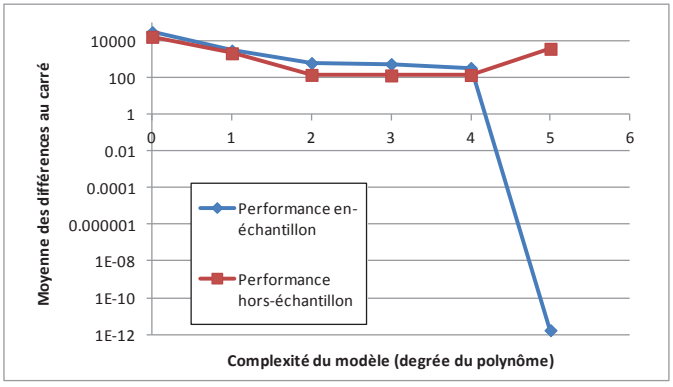
Exemple 4.1.5. *Prenons pour exemple, une régression sur une série de 6 données qui aurait pour but de modéliser la relation de pertinence entre X et Y. Une série de 6 modèles de relation ont été évaluer. Spécifiquement, le k-ième modèle capture la relation suivante :*

$$Y = \theta_0 + \sum_{i=1}^{k-1} \theta_i X^i \ .$$

Remarquez comment la complexité des modèles va en s'accroissant du premier au sixième. La figure suivante présente les deux régression obtenus lorsque la relation de pertinence prend la d'un polynôme de degré 2 ou de degré 5. Laquelle des deux régressions choisiriez-vous ? Si on limite l'analyse au données historiques utilisées pour la calibration, alors le polynôme de degré 5 semble être le meilleur choix. En fait, selon se polynôme la variable Y peut sensiblement être entièrement expliquée par la variable Y. Cependant, il est clair en



observant la qualité des prédictions pour les nouveaux échantillons que le modèle plus complexe est bien moins adéquat. En général, on devrait pouvoir observer la tendance suivante. En augmentant la complexité d'un modèle simple, on améliore normalement la performance de prédiction du modèle, cependant lorsque la complexité est trop grande le manque de données fait en sorte que l'on perd la puissance de généralisation. C'est ce qu'on appelle surapprentissage, ou «overfitting» en anglais. La figure suivante nous illustre ce phénomène dans le cas de la régression des 6 points sur un polynôme de degré croissant.



4.2 Approche Bayesienne

Nous avons vu que le théorème de Baye's nous indique comment prendre en compte d'une nouvelle information à propos de variables aléatoire qui nous intéressent et pour lesquelles nous avions déjà de l'information. Nous allons maintenant réutiliser ce théorème pour proposer une méthode appelée approche Bayesienne. Contrairement aux approches de calibration mentionnées plus haut, cette nouvelle approche nous permettra à tout moment d'avoir une mesure de la confiance que nous avons vis-à-vis de la nature du modèle stochastique qui se cache derrière les données.

Remarque 4.2.1. *Rappelez-vous que pour résoudre le problème lié au surapprentissage, il nous a fallu évaluer le modèle calibrer sur de nouvelles données. En d'autres mots, chaque jeu de données a le potentiel de nous suggérer un modèle stochastique différent. Pour un jeu de données suffisamment grand par rapport à la complexité du modèle stochastique, les modèles obtenus en changeant le jeu de données devraient être relativement similaires (en comparant les paramètres calibrés). L'approche Bayesienne nous livre cette information directement.*

Considérons donc une dernière fois que nous souhaitons modéliser $P(Z)$ à l'aide des données $\{z_1, z_2, ..., z_M\}$. Nous ferons l'hypothèse à nouveau que $P(Z)$ prend la forme paramétrique $f(z;\theta)$ et que toutes les données ont été générées indépendamment selon la même loi. Cette fois cependant, nous ferons l'hypothèse Bayesienne suivante : avant même de regarder nos données, nous sommes capables de représenter notre croyance subjective de la valeur de θ à l'aide d'une loi de probabilité : i.e. $f(\theta)$.

Cette hypothèse additionnelle implique deux faits importants. Premièrement, avant même de voir quelques données que ce soit, il est déjà possible d'évaluer la probabilité marginale de Z :

$$P(Z \in A) = \int P(Z \in A|\theta) f(\theta) d\theta = \int \int_A f(z;\theta) f(\theta) dz d\theta$$

Notez que pour évaluer cette probabilité nous avons deux niveaux d'incertitude à prendre en compte : l'incertitude vis-à-vis du paramètre de la loi de probabilité de Z et l'incertitude vis-à-vis de Z une fois que ce paramètre est connu. Deuxièmement, une fois que les données ont été observées, il est possible d'appliquer le théorème de Bayes pour déterminer notre connaissance à posteriori du paramètre θ et indirectement de la variable Z. Pour ce faire, il nous faut évaluer la loi conditionnelle de Z étant donné la connaissance des observations : $P(Z \in A|\mathcal{O})$, où $\mathcal{O} = \{z_1, z_2, ..., z_M\}$. Mathématiquement, le théorème de Baye's nous donne :

$$f(\theta|\mathcal{O}) = \frac{f(\mathcal{O}|\theta) f(\theta)}{\int f(\mathcal{O}|\theta) f(\theta) d\theta} .$$

Notez que tous les éléments de ce calcul sont connus. La loi $f(\theta|\mathcal{O})$ représente notre connaissance à postériori du paramètre θ. Plus nous aurons d'observation, plus cette loi sera informative (i.e. concentrée autour de notre meilleur estimation de θ). Pour évaluer la loi marginale à postériori de Z, nous suivons le raisonnement suivant :

$$\begin{aligned} P(Z \in A|\mathcal{O}) &= \int \int_A f(z|\theta, \mathcal{O}) f(\theta|\mathcal{O}) dz d\theta \\ &= \int \int_A f(z;\theta) f(\theta|\mathcal{O}) dz d\theta \end{aligned}$$

où nous exploitons que Z est indépendant de $\mathcal{O}$ si θ est connu.

Malheureusement, en général il est difficile de faire le suivi de $f(\theta|\mathcal{O})$ et d'évaluer des probabilités à partir de cette loi à cause de l'intégrale

$$\int f(\mathcal{O}|\theta) f(\theta) d\theta = \int \left(\prod_{i=1}^{M} f(z_i;\theta) \right) f(\theta) d\theta$$

qui est très difficile à évaluer analytiquement ou même numériquement. La méthode la plus efficace pour ce faire est appelée méthode de Monte-Carlo par chaînes de Markov.

La raison principale pour laquelle l'approche Bayesienne est considérée si importante est que dans des situations bien particulières (mais tout de même très présente en pratique), l'analyse de cette intégrale mène à une simplification bien particulière. En effet, si $f(\theta)$ est une loi de distribution dite «à priori conjuguée» pour $f(z;\theta)$, alors $f(\theta|\mathcal{O})$ prend alors la même forme que $f(\theta)$. Voici quelques exemples, de ce résultats.

Loi	Conjuguée	Suivi des paramètres
Bernoulli	bêta(α,β)	$\alpha' = \alpha + \sum_i z_i\ ,\ \beta' = \beta + \sum_i (1-z_i)$
Poisson	Gamma(k,θ)	$k' = k + \sum_i z_i\ ,\ \theta' = \theta/(n\theta+1)$
Exponentielle	Gamma(k,θ)	$k' = k + n\ ,\ \theta' = \theta/(1+\theta\sum_i z_i)$
...	...	...

Exemple 4.2.1. *Un sac contient 1000 jetons de poker de deux couleurs : rouge ou blanc. Parmi les dix derniers jetons retirés aléatoirement du sac, trois d'entre eux étaient rouges. Quelle est votre connaissance à postériori de la probabilité que le prochain jeton sorti du sac soit rouge ? Quelles sont les chances que la proportion de jeton rouge dans le sac soit de plus de 50% ?*

L'approche Bayesienne peut être appliqué en considérant que Z est la variable de Bernoulli décrivant si le prochain jeton est rouge ou non. Le paramètre de cette loi est p qui capture la probabilité de succès. L'hypothèse Bayesienne nous demande de représenter notre connaissance de p avant même de voir les observations. Pour pouvoir profiter des résultats analytiques présentés dans la table ci-dessus, nous utiliserons la loi bêta. Il est commun d'assumer qu'à priori toutes les valeurs de p sont équiprobables afin d'influencer le moins possible le résultat de l'analyse. Dans ce cas, il s'agit de considérer que $p \sim$ Bêta$(1,1)$, qui est équivalent à la loi uniforme sur l'intervalle $[0,1]$. Après l'observation des 10 tirs de jetons dont 3 étaient rouges, la connaissance à postériori de p est bêta avec $\alpha = 1+3$ et $\beta = 1+7$. La probabilité de succès peut être calculée est appliquant :

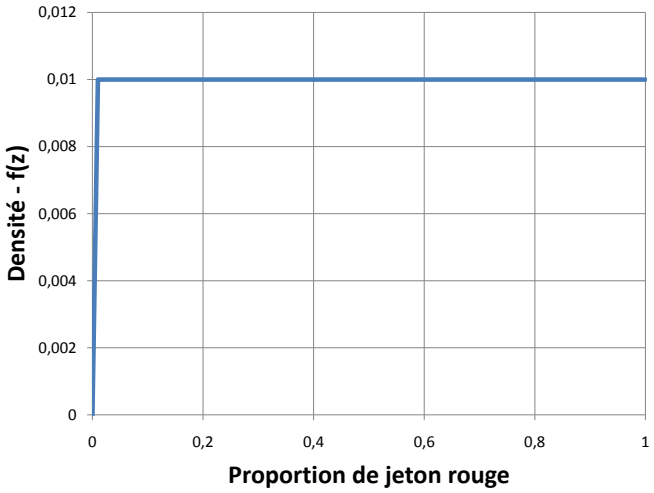
$$\begin{aligned} P(Z=1|\mathcal{O}) &= \int P(Z=1|p) f(p|\mathcal{O}) dp \\ &= \int \frac{p}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1} dp = \frac{\alpha}{\alpha+\beta} = 4/12 \end{aligned}$$

D'un autre côté la probabilité que la proportion de jeton rouge soit au-delà de 50% peut être calculée est résolvant :

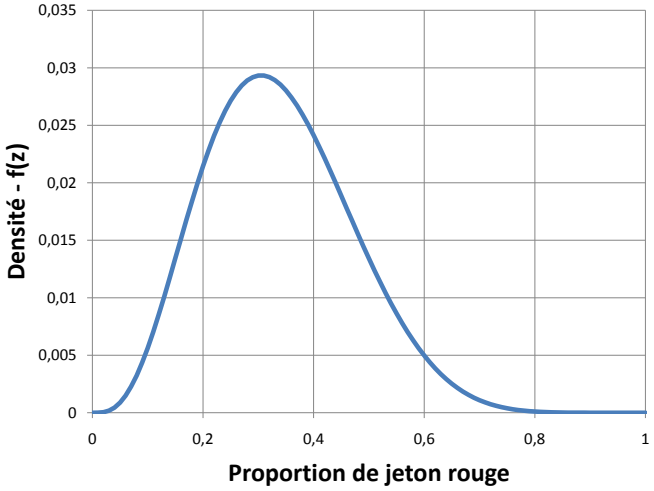
$$P(p \geq 50\%) = \int_{0.5}^{1} f(p|\mathcal{O}) dp = \int_{0.5}^{1} \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1} dp \approx 34.4\%$$

Les deux figures suivantes illustrent le type de connaissance capturé par la loi bêta lorsque $\alpha = \beta = 1$ et lorsque $\alpha = 4$ et $\beta = 8$.

Connaissance à priori de p



Connaissance à postériori de p



Chapitre 5

Simulation par la méthode de Monte-Carlo

Dans un problème avec incertitude, la valeur de notre objectif doit être considérée comme une variable aléatoire puisque nous ne pouvons prévoir exactement sa valeur. C'est souvent le cas pour des objectifs tels que le profit généré, le succès d'un projet, l'importance de l'impact environnemental, etc. La simulation Monte-Carlo permet d'étudier la nature stochastique de cette variable aléatoire. Nous pourrons ainsi avoir une meilleure idée de la valeur espérée du profit, la probabilité de succès, caractériser la loi de probabilité de l'objectif sous la forme d'un histogramme ou d'une fonction de répartition, et produire une estimation de statistiques capturant la nature des risques impliqués (variances, centiles, etc.). Alors que la simulation permettra de comparer différentes alternatives, elle ne permettra malheureusement pas directement l'optimisation de la décision.

Voici la procédure qui sera suivie dans une étude basée sur une simulation de Monte-Carlo.

1. Construire le diagramme d'influence
2. Bien décrire les relations impliquées dans chacune des boîtes
3. Fixer les alternatives à évaluer
   — Pour les problèmes dynamiques décider d'une «stratégie» à évaluer : i.e. : Choisir une fonction des variables d'influence
4. Générer M scénarios de la réalisation des variables aléatoires
   — Identifier les variables aléatoires sans influence et générer un scénario pour ces variables selon leur distribution
   — Pour chaque variable qui n'est influencée que par variables résolues, générer un scénario selon la distribution conditionnelle
   — Répéter jusqu'à ce que toutes les variables soient résolues
   — Évaluer la valeur de l'objectif pour ce scénario
5. Analyser la distribution empirique obtenue pour l'objectif
6. Estimer le niveau de confiance des statistiques pour possiblement augmenter le nombre de scénarios

Étant donné que les deux premières étapes ont été discutées dans les sections précédentes, nous nous attarderons sur l'étape de génération des scénarios et l'étape d'analyse des scénarios obtenus.

5.1 Génération de scénarios pseudo-aléatoires

La composante de base de toute simulation de Monte-Carlo réside dans la génération d'une variable de Bernoulli à probabilité 50%. C'est-à-dire que nous avons besoin d'un processus qui génère aléatoire la valeur de 1 ou 0 et pour lequel après chaque nombre généré nous ne pouvons que prédire que le prochain nombre tiré à 50% chance de prendre la valeur 1. Nous verrons en effet qu'à partir d'un tel outil, il est possible de générer des scénarios à partir de toute distribution imaginable. Malheureusement, il est difficile de développer un tel processus qui soit réellement aléatoire. Certains appareils simulent réellement cette variable à l'aide d'un phénomène physique (e.g. bruit thermal) mais ceux-ci sont plutôt coûteux. Nos ordinateurs simulent plutôt cette variable à l'aide d'un processus déterministe. On dira donc que nos simulations sont basées sur une génération pseudo-aléatoire des scénarios.

Remarque 5.1.1. *En fait, nos ordinateurs utilisent des registres à décalage à rétroaction linéaire (≪Linear Feedback Shift Register≫) pour générer une séquence de réalisation d'une variable de Bernoulli. Ce système génère une longue séquence prédéterminée de 0 et de 1 que l'on ne peut distinguer d'une séquence purement aléatoire. Puisque la séquence contient autant de 0 que de 1, on peut en déduire que chaque valeur a 50% de chance d'être égale à* 1*. Voici un exemple d'une séquence générée à partir de cet outil et qui sera réutilisé encore et encore pour générer de tels valeurs :*

$$1, 1, 0, 0, 0, 1, 0, 0, 1, 1, 0, 1, 0, 1, 1, 0$$

Il est à noter que si la séquence est trop courte, ou si je connais comment la séquence est générée, alors les nombres ne me sembleront plus réellement aléatoires d'où l'intérêt de développer des outils qui génèrent des séquences très longues et irrégulières avant de se répéter.

5.2 Génération uniforme sur Intervalle $[0, 1]$

Tel que mentionné plus haut, il est possible de générer un nombre aléatoire selon la distribution de son choix à partir d'un système qui produit des zéros et des uns avec une probabilité de 50% chacun. Commençons donc par la deuxième variable aléatoire la plus simple, une variable distribuée uniformément entre 0 et 1.

Comment générer une valeur uniforme entre 0 et 1 en utilisant une pièce de 25 cents sans biais :

1. Diviser l'intervalle choisi en deux sous-intervalles de même taille

2. Tirer le 25 cents. Si le résultat est «pile», alors poursuivre l'exercice avec l'intervalle de gauche sinon avec celui de droite.
3. Après k lancer, la grandeur de l'intervalle est de $1/2^k$. Répéter les étapes 1 et 2 jusqu'à ce que l'intervalle obtenu soit assez petit
4. Garder comme valeur aléatoire la valeur moyenne sur l'intervalle obtenu

La valeur obtenue à l'aide de cette procédure sera distribuée uniformément parmi les valeurs suivantes :

$$\{\frac{1}{2^{k+1}}, \frac{1}{2^{k+1}} + \frac{1}{2^k}, \frac{1}{2^{k+1}} + \frac{2}{2^k}, \frac{1}{2^{k+1}} + \frac{3}{2^k}, \cdots, 1 - \frac{1}{2^{k+1}}\}$$

Si k est assez grand (disons 50) il devient numériquement impossible de distinguer la distribution de ces valeurs d'une distribution uniforme sur l'intervalle.

5.3 Génération d'une variable discrète I

Il est possible de simuler une variable aléatoire prenant les valeurs $\{z_1, z_2, ..., z_N\}$ avec les probabilités $\{p_1, p_2, ..., p_N\}$ respectivement en générant une variable uniformément distribuée sur $[0, 1]$ puis en suivant les directives du tableau ci-dessous

Si $U \in [a, b]$	alors $Z =$
$[\ 0\ ,\ p_1\ ]$	z_1
$]\ p_1\ ,\ p_1 + p_2\ ]$	z_2
$]\ p_1 + p_2\ ,\ p_1 + p_2 + p_3\ ]$	z_3
...	...
$]\ p_1 + ... + p_{n-1}\ ,\ 1\ ]$	z_n

Nous profitons de cet exemple pour présenter le type de méthode employé pour prouver qu'une variable générée pseudo-aléatoirement respecte réellement la distribution qui nous intéresse. D'abord, le processus décrit plus haut peut être présenté sous une forme mathématique :

$$V = \sum_{i=1}^{N} z_i \mathbb{1}\left\{\sum_{j=1}^{i-1} p_j \leq U < \sum_{j=1}^{i} p_j\right\},$$

où la fonction indicatrice $\mathbb{1}\{C\}$ est une fonction qui retourne la valeur 1 si C est vrai et zéro si C est faut.

Nous cherchons donc à prouver que la distribution de V tel que générée est telle que prescrite. En voici la preuve :

$$\begin{aligned} P(V = z_i) &= P\left(\sum_{j=1}^{i-1} p_j \leq U \leq \sum_{j=1}^{i} p_i\right) = F_U\left(\sum_{j=1}^{i} p_j\right) - F_U\left(\sum_{j=1}^{i-1} p_j\right) \\ &= \sum_{j=1}^{i} p_j - \sum_{j=1}^{i-1} p_j = p_i. \end{aligned}$$

5.4 Génération d'une variable continue

Pour générer une variable aléatoire continue, deux méthodes sont souvent employées. La plus commune assume que nous connaissons la forme mathématique de la fonction de répartition de la variable alors que la seconde peut-être utilisée lorsque nous n'avons en mains que la fonction de densité. Cette dernière méthode est surtout utile pour les vecteurs aléatoires.

5.4.1 Méthode d'inversion

Considérons une situation dans laquelle la fonction de répartition prend la forme $F : \mathbb{R} \to [0,1]$, et est continue et strictement croissante. Il est possible de générer une variable V distribuée selon F, simplement en transformant une variable uniformément distribuée sur $[0,1]$ à l'aide de la fonction de répartition inverse F^{-1}. En particulier, nous aurons que $V = F^{-1}(U)$ où U est une variable uniforme sur $[0,1]$.

Il est possible de prouver que V suit la bonne distribution en étudiant quelle est la fonction de répartition de la variable V ainsi générée.

$$F_V(v) \;=\; P(F^{-1}(U) \leq v) = P(U \leq F(v)) = F(v)\ .$$

Exemple 5.4.1. *Pour générer une variable distribuée selon la loi exponentielle, dont* $F(z) = (1 - \exp(-z/\mu))\mathbb{1}\{z \geq 0\}$*, il suffit d'établir quelle est la fonction de répartition inverse de cette variable.*

$$F^{-1}(y) = -\mu \ln(1-y)$$

étant donné que

$$(1 - \exp(-(-\mu \ln(1-y))/\mu))\mathbb{1}\{-\mu \ln(1-y) \geq 0\} = (1-(1-y)) = y\ .$$

Nous avons donc que $V = F^{-1}(U) = -\mu \ln(1-U)$ *satisfait la loi exponentielle.*

5.4.2 Méthode du rejet

Considérons que nous souhaitons générer une variable selon la densité $f : \mathbb{R}^n \to \mathbb{R}$. Pour ce faire, nous allons nous servir de U, la variable simulée uniforme sur $[0,1]$ et d'une seconde variable simulée selon une seconde densité $f_Z : \mathbb{R}^n \to \mathbb{R}$ pour laquelle nous savons qu'il existe une constante c telle que $f(z) \leq c f_Z(z)\ ,\ \forall\ z \in \mathbb{R}^n$. La procédure est la suivante : (1) générer pseudo-aléatoirement une séquence de paires $(U_1, Z_1), (U_2, Z_2), ...$, où chaque U_i est indépendant et uniformément distribué sur $[0,1]$ et chaque V_i est générée i.i.d. selon f_V ; (2) retourner la première valeur Z_i pour laquelle $cU_i \leq f(Z_i)/f_Z(Z_i)$.

Nous pouvons démontrer que cette valeur Z_i est distribuée selon la densité f. Appelons i^* le 1^{er} index i pour lequel $cU_i \leq f(Z_i)/f_Z(Z_i)$. Nous allons confirmer que $V = Z_{i^*}$ suit la

bonne fonction de répartition lorsque $n = 1$.

$$\begin{aligned} P(V \leq v) &= P(Z_{i^*} \leq v) = P\left(Z \leq v \middle| cU \leq \frac{f(Z)}{f_Z(Z)}\right) \\ &= \frac{P(Z \leq v \ \& \ cU \leq f(Z)/f_Z(Z))}{P(cU \leq f(Z)/f_Z(Z))} \\ &\propto P(Z \leq v \ \& \ cU \leq f(Z)/f_Z(Z)) \\ &= \int_{-\infty}^{v} P(cU \leq f(Z)/f_Z(Z) | Z = y) f_Z(y) dy \\ &= \int_{-\infty}^{v} P\left(cU \leq \frac{f(y)}{f_Z(y)}\right) f_Z(y) dy = \int_{-\infty}^{v} (1/c) \frac{f(y)}{f_Z(y)} f_Z(y) dy \\ &= (1/c) F(v) \end{aligned}$$

Puisque $P(V \leq v)$ est proportionnel à $(1/c)F(v)$ et que $P(V \leq \infty) = 1 = (\alpha/c)F(\infty) = \alpha/c$, il est nécessaire que $\alpha = c$ et que $P(V \leq v) = F(v)$.

Un élément important de la méthode de rejet est le choix de c qui doit être le plus petit possible puisque le nombre de termes rejetés dépend de sa taille. En fait, il est possible de démontrer que nous devrons rejeter en moyenne $c - 1$ termes. D'abord, pour chaque paire nous pouvons déterminer la probabilité que la condition soit satisfaite :

$$\begin{aligned} P\left(cU \leq \frac{f(Z)}{f_Z(Z)}\right) &= \int_{-\infty}^{\infty} P\left(cU \leq \frac{f(y)}{f_Z(y)}\right) f_Z(y) dy \\ &= \int_{-\infty}^{\infty} (1/c) \frac{f(y)}{f_Z(y)} f_Z(y) dy = 1/c \end{aligned}$$

La variable i^* est donc distribuée selon une loi géométrique qui compte le nombre de lancés, à probabilité $1/c$ de succès, avant le premier succès. Donc, $E[i^*] = 1/(1/c) = c$.

Exemple 5.4.2. *Nous allons employer la méthode de rejet pour générer un point uniformément dans un cercle de rayon* 1. *La densité de cette loi peut être décrite comme suit :*

$$f_Z(z_1, z_2) = (1/\pi) \mathbb{1}\{z_1^2 + z_2^2 \leq 1\} \ .$$

La méthode nous requiert de mettre de l'avant un vecteur aléatoire de dimensions que nous savons simuler. Considérons Z uniformément distribué sur le carré $[-1, 1] \times [-1, 1]$. Un tel vecteur peut facilement être généré en simulant chacun de ses termes Z_1 et Z_2 à l'aide de deux variables uniforme sur l'intervalle $[0, 1]$. De plus, la loi de densité de Z est :

$$f_Z(z_1, z_2) = (1/4) \mathbb{1}\{-1 \leq z_1 \leq 1 \ \& \ -1 \leq z_2 \leq 1\} \ .$$

Il ne nous reste qu'à identifier la valeur de c qui respecte $f(Z) \leq f_Z(z)$ pour tout z. On choisit c :

$$c = \inf\{c | f \leq c f_Z\} = \inf\{c | (1/\pi \leq c/4\} = 4/\pi \ .$$

Pour simuler un scénario de uniformément dans un cercle de rayon 1*, on simulera donc d'abord une série de scénarios* (U, Z_1, Z_2) *jusqu'à ce que*

$$U \leq \frac{\pi}{4} \frac{(1/\pi)\mathbb{1}\{V_1^2 + V_2^2 \leq 1\}}{(1/4)\mathbb{1}\{-1 \leq V_1 \leq 1, -1 \leq V_2 \leq 1\}} = \mathbb{1}\{V_1^2 + V_2^2 \leq 1\} \ .$$

En d'autres mots, la première paire de coordonnées (Z_1, Z_2) *qui tombe dans la sphère est considérée être distribuée uniformément sur le cercle. En moyenne, on s'attend à devoir générer :* $1/c = \pi/4 = 1,27$ *paires avant d'obtenir un scénario.*

5.5 Génération d'un vecteur aléatoire

La méthode du rejet est un exemple de méthode qui peut être utilisée pour générer un scénario de vecteur aléatoire. Cependant, en général il sera plus commode d'exploiter un diagramme d'influence pour générer les termes du vecteur. Par exemple, si nous voulons simuler un vecteur de variables aléatoires dont la fonction de masse est $P(z_1, z_2, z_3, ..., z_n)$, il est toujours possible de décomposer la distribution conjointe sur une série de loi marginale conditionnelle. Il suffit de calculer les densités marginales et conditionnelles suivantes :

$$P_1(Z_1 = y) = \sum_{z_2, z_3, ..., z_n} P(y, z_2, z_3, ..., z_n)$$

$$P_2(Z_2 = y | z_1) = \sum_{z_3, ..., z_n} P(z_1, y, z_3, ..., z_n) \Big/ P(Z_1 = z_1)$$

...

$$P_n(Z_n = y | z_1, ..., z_{n-1}) = P(z_1, ..., z_{n-1}, y) / P(Z_1 = z_1, ..., Z_{n-1} = z_{n-1})$$

Dans cet exemple, nous commencerions par générer la variable Z_1. Une fois celle-ci obtenue, il est par la suite possible d'utiliser la loi conditionnelle pour Z_2 étant donné le scénario de Z_1 obtenu et ainsi de suite pour tous les termes du vecteurs.

Remarque 5.5.1. *Tel qu'observé dans les sections précédentes, il peut devenir assez difficile de déterminer les modèles stochastiques qui devraient être utilisé pour modéliser les liens de pertinence entre différentes variables aléatoires. Pourtant, pour générer une analyse réaliste des risques liés aux décisions, il est important de modéliser adéquatement la pertinence entre variables. Pour cette raison, plusieurs outils informatiques permettent maintenant d'approximer ces liens en imposant seulement qu'un certain niveau de corrélation entre variables soit respectée. En d'autres mots, l'outil requiert alors le détail des lois marginales de chaque variable aléatoire, ainsi qu'une mesure de corrélation entre chaque paire de variables :*

$$\rho_{Z_1, Z_2} = E[(Z_1 - \mu_1)(Z_2 - \mu_2)] / (\sigma_1 \sigma_2) \ .$$

Ce facteur prend une valeur sur $[-1, 1]$*, étant positif si les variables sont généralement du même côté de leur moyenne et négatif si elles sont de côtés opposés. Lors de la simulation, le niveau de corrélation est injecté à l'aide du concept de «copule».*

5.6 Analyse statistiques des conséquences

Dans cette section, nous nous attarderons à l'analyse des scénarios d'une conséquence qui nous intéresse obtenus à partir de simulation de Monte-Carlo. En particulier, après avoir généré un échantillon de M scénarios supposés être générés selon la distribution des réalisations futures, nous considérerons que la distributions empirique de ces valeurs est une approximation de la vraie distribution des conséquences possibles.

Définition 5.6.1. *La distribution empirique d'une variable aléatoire est obtenue en affectant un poids équiprobable sur chacune des valeurs d'un échantillon. La fonction de répartition empirique basée sur l'échantillon $Z_1, Z_2, ..., Z_M$ est donc :*

$$F_Z^M(z) = \frac{1}{M} \sum_{i=1}^{M} \mathbb{1}\{Z_i \leq z\}$$

En fait, $F_Z^M(z)$ est la proportion de fois qu'on observe que Z_i est plus petite ou égale à z dans l'échantillon

Il existe plusieurs mesures servant à quantifier le risque lié aux réalisations futures d'une variable qui nous intéresse.

— Valeur espérée : valeur moyenne atteinte par la distribution. La loi des grands nombres nous indique qu'à long terme la somme des rendements de nos décisions devrait tendre vers la somme des valeurs espérées.
— Écart type : écart moyen des valeurs par rapport à la valeur espérée. Cette valeur nous indique à quel point la réalisation pourrait être différente de la valeur espérée. Cette statistique est typiquement jugée proportionnelle au risque malgré que dans certains cas, l'écart pourrait être principalement causé par des rendements au-dessus de la moyenne.
— Étendue : intervalle contenant toutes les réalisation possibles de la variable. Celui-ci nous indique l'étendue de ce à quoi nous pouvons nous attendre.
— Probabilité d'atteinte d'un niveau : potentiel d'atteinte d'un niveau cible. Une décision qui a moins de chance d'atteindre une cible raisonnable devrait être considérée plus risquée.
— Médiane : cible qui a autant de chance d'être dépassée que d'être manquée

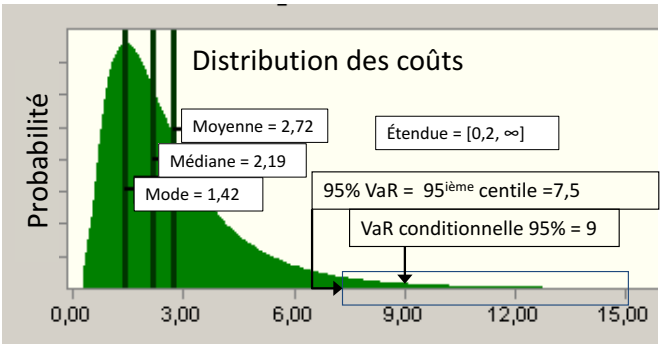
Récemment, trois statistiques sont devenus très populaires dans le monde de la gestion des risques.

Définition 5.6.2. *Le n-ième centile est la valeur la plus petite au dessous de laquelle nous sommes assurés que la variable aléatoire a plus que $n\%$ de chance de tomber. Si $F_Z(\cdot)$ est continu, le n-ième centile est la valeur z pour laquelle $F_Z(z) = n/100$.*

Définition 5.6.3. *La valeur à risque pour un niveau de confiance de α est la valeur y pour laquelle nous avons $\alpha \cdot 100\%$ de chance de faire un perte plus petite ou égale à y. Si $F_Z(\cdot)$ est continu et Z est un coût, $VaR_\alpha(Z)$ est la valeur z pour laquelle $F_Z(z) = \alpha$. Si Z représente un profit alors la définition est inversée : $VaR_\alpha(Z)$ est la valeur z pour laquelle $F_Z(-z) = 1 - \alpha$.*

Définition 5.6.4. *La valeur à risque conditionnelle (aussi appelée manque à gagner espéré) pour un niveau de confiance de α est la valeur espérée de la perte monétaire encourue conditionnellement à la réalisation d'un scénario parmi l'ensemble des $(1-\alpha)\cdot 100\%$ pires scénarios (intuitivement, c'est la perte espérée lorsque celle-ci dépasse la perte mesurée par la VaR_α). Si Z est un coût et $F_Z(\cdot)$ est continu, alors on peut la calculer par $E[Z|Z \geq VaR_\alpha(Z)] = E[Z \cdot \mathbb{1}\{Z \geq VaR_\alpha(Z)\}]/(1-\alpha)$. Alternativement, si Z est un profit et $F_Z(\cdot)$ est continu, alors la valeur à risque conditionnelle se calcule suivant $-E[Z|Z \leq -VaR_\alpha(Z)] = -E[Z \cdot \mathbb{1}\{Z \leq -VaR_\alpha(Z)\}]/(1-\alpha)$.*

La figure suivante présente un exemple de plusieurs de ces statistiques.



Exemple 5.6.1. *Pour la loi normale dont la moyenne est zéro et l'écart type est un, le 10-ième centile est d'environ $-1,28$, alors que le 90-ième centile est d'environ $1,28$. La valeur à risque de 90% est de $1,28$ peu importe que Z soit un profit ou une dépense. Finalement, la valeur à risque conditionnelle de 90% peut être calculée suivant*

$$CVaR-90\%[Z] = E[Z|Z \geq VaR-90\%(Z)] = \psi(\Phi^{-1}(90\%))/(1-90\%) \approx 1.75$$

si Z est une dépense, où $\Phi^{-1}(y)$ est la fonction de répartition inverse de la loi normale standard et où

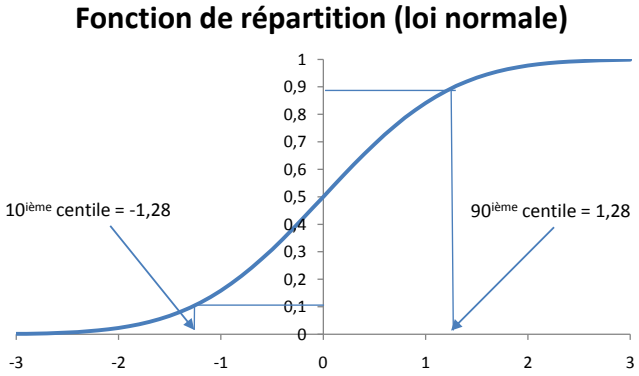
$$\psi(y) := \frac{1}{\sqrt{2\pi}} e^{\frac{-y^2}{2}},$$

est la fonction de densité de la loi normale standard. Dans le cas où Z est plutôt un profit, alors le calcul prend la forme :

$$CVaR-90\%[Z] = -E[Z|Z \leq -VaR-90\%(Z)] = \psi(\Phi^{-1}(90\%))/(1-90\%) \approx 1.75$$

La figure suivante illustre comment obtenir le centile à partir du graphe de la fonction de répartition.

En général, il peut être difficile de juger si les conséquences d'une décision sont mieux ou pires que les conséquences d'une autre décision. si l'une semble meilleure par rapport à la valeur espérée, l'autre pourrait être plus attrayante vis-à-vis de la valeur à risque. Il existe tout de même deux situations dans lesquelles une décision peut réellement sembler dominer une autre.



Définition 5.6.5. *L'alternative x_A domine de manière stochastique l'alternative x_B si pour tout niveau de performance, l'alternative x_A a plus de chance de surpasser ce niveau que l'alternative x_B vis-à-vis de cet objectif. Mathématiquement, si $g(x, Z)$ représente le profit, on dira que si*

$$P(g(x_A, Z) \geq \alpha) \geq P(g(x_B, Z) \geq \alpha)\,,\ \forall\, \alpha \in \mathbb{R}$$

ou de manière équivalente

$$P(g(x_A, Z) \leq \alpha) \leq P(g(x_B, Z) \leq \alpha)\,,\ \forall\, \alpha \in \mathbb{R}$$

alors x_A domine x_B stochastiquement selon la conséquence $g(\cdot, Z)$. Nous verrons comment cette observation peut-être faite en étudiant les deux fonctions de répartition.

Il existe une autre forme de dominance, que nous appellerons déterministe, qui est encore plus convaincante que la dominance stochastique. celle-ci est cependant plus difficile à établir.

Définition 5.6.6. *L'alternative x_A domine de manière déterministe l'alternative x_B si nous sommes garantis que les conséquences de x_A seront préférées aux conséquences de x_B peu importe la réalisation des paramètres incertains. Mathématiquement, si $g(x, Z)$ est un profit, on dira que*

$$P(g(x_A, Z) \geq g(x_B, Z)) = 1$$

implique que x_A domine x_B déterministiquement selon la conséquence $g(\cdots, Z)$.

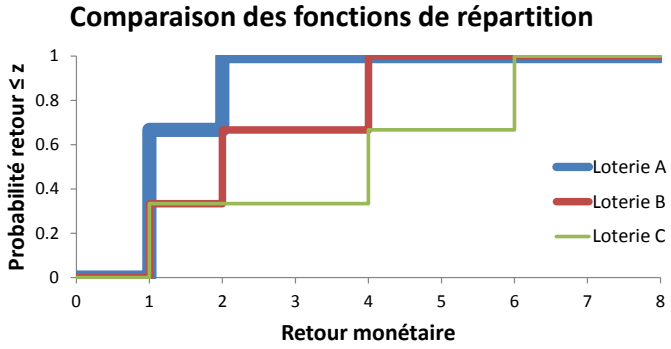
Cette forme de dominance implique la dominance stochastique. Elle ne peut que parfois être déduite en comparant les fonctions de répartitions : i.e. dans le cas où les conséquences les plus négatives de l'alternative x_A sont préférables aux conséquences les plus positives de l'alternative x_B.

Exemple 5.6.2. *Considérons trois loteries associé au lancé d'un dé à six faces sans biais (le même lancé du même dé est utilisé pour les trois loteries).*

- *Loterie A : (1-2) gagne 1$, (3-4) gagne 1$, (5-6) gagne 2$*
- *Loterie B : (1-2) gagne 1$, (3-4) gagne 2$, (5-6) gagne 4$*

— *Loterie C : (1-2) gagne 6$, (3-4) gagne 4$, (5-6) gagne 1$*

La figure suivante illustre les fonctions de répartition associées aux trois loteries.



D'abord, en comparant le résultat de chacune des loteries sous chacun des résultat du dé, nous réalisons que peut importe la réalisation du dé, la loterie B retourne toujours plus d'argent que la loterie A. C'est donc que nous sommes assurés que B est préférable à A ; A est dominé déterministiquement par B. En regardant les fonctions de répartition de ces deux loteries, on remarque aussi que pour tout niveau d'argent que l'on souhaiterait atteindre, la loterie B a plus de chance de nous satisfaire que la loterie A. Nous confirmons donc que B est aussi dominé stochastiquement par A. Remarquons de plus que la loterie C à plus de chance d'atteindre tout niveau de retour monétaire que ces deux concurrentes. Elle les domine donc stochastiquement toutes deux sans toutefois les dominer déterministiquement puisque lorsque le dé est de 5 ou 6, nous préférerions avoir jouer au loterie A ou B.

5.7 Calculer l'erreur d'estimation

Considérons les valeurs $Z_1, Z_2, ..., Z_M$ simulées des paramètres incertains du problème. Nous venons d'argumenter que les risques d'une décision devraient être évalués à partir de statistiques calculées selon la distribution empirique de la conséquence $g(x, Z)$ selon l'échantillon de scénarios. Puisque les valeurs de $g(x, Z_1), g(x, Z_1), \cdots, g(x, Z_M)$ ont été générées aléatoirement, nous devons nous attendre à ce que les statistiques estimées ne soient pas exactement égales aux vraies statistiques de $g(x, Z)$. En fait, si l'on refait la même analyse par simulation de Monte-Carlo à deux reprises, ou sur deux ordinateurs, nous n'obtiendrons jamais le même résultat. Il est donc important de se faire une idée de la grandeur de l'erreur d'estimation avant de tirer nos conclusions.

5.7.1 L'erreur d'une estimation de valeur espérée

L'estimation d'une valeur espérée $E[g(x, Z)]$ à partir de simulation Monte-Carlo est calculée selon

$$\hat{\theta}_M(x, Z) = \frac{1}{M} \sum_{i=1}^{M} g(x, Z_i)$$

Noter que la valeur estimée est aléatoire puisque chaque Z_i a été généré aléatoirement. Cette estimateur $\hat{\theta}_M(x, Z)$ a cependant plusieurs propriétés qui jouent en sa faveur.

1. Il est sans biais : $E[\hat{\theta}_M(x, Z)] = E[g(x, Z)]$. En d'autres mots, nous n'avons aucune raison de croire que la valeur sera sur-évaluée ou sous-évaluer.
2. Il est convergent : c'est à dire, que nous sommes assurés que plus nous utiliserons de valeurs simulées plus la valeur estimée sera proche de la vraie valeur. Mathématiquement, il est possible d'établir que la variance de l'estimation converge vers 0 : $Var[\hat{\theta}_M(x, Z)] = Var[g(x, Z)]/M \rightarrow 0$ lorsque $M \rightarrow \infty$. De plus, selon la loi forte des grands nombres, la convergence de l'écart type garantit que $\hat{\theta}_M$ converge presque certainement vers $E[g(x, Z)]$:

   $$P(\lim_{M\to\infty} |\hat{\theta}_M - E[g(x, Z)]| = 0) = 1 \ .$$

   En termes simple, cette convergence nous indique que toute personne qui utilise la moyenne empirique verra cette estimateur converger vers la vraie valeur espérée.
3. Il est asymptotiquement normal : par le théorème limite centrale la distribution de l'estimation converge vers la loi normale. Mathématiquement, nous avons que

   $$\frac{\hat{\theta}_M - E[g(x, Z)]}{\sigma/\sqrt{M}} \rightarrow \mathcal{N}(0, 1) \text{ lorsque } M \rightarrow \infty$$

   où $\sigma^2 = Var[g(x, Z)]$. Intuitivement, on peut dire, si M est assez grand, que l'estimateur est distribué selon la loi $\mathcal{N}(E[g(x, Z)], Var[g(x, Z)]/M)$.

La dernière propriété présentée nous fournit un moyen d'estimer où la vraie valeur de $E[g(x, Z)]$ peut potentiellement se trouver. En effet, sachant que $\hat{\theta}_M \approx \mathcal{N}(E[g(x, Z)], Var[g(x, Z)]/M)$, nous pouvons prévoir que pour tout $c \geq 0$ la probabilité que

$$-c\sqrt{\frac{Var[g(x, Z)]}{M}} \leq \hat{\theta}_M - E[g(x, Z)] \leq c\sqrt{\frac{Var[g(x, Z)]}{M}}$$

est de $\Phi(c) - \Phi(-c) = \Phi(c) - (1 - \Phi(c)) = 2\Phi(c) - 1$, où $\Phi(\cdot)$ est la fonction de répartition de la distribution normale dont la moyenne est 0 est l'écart type de 1. Par conséquent, nous sommes assurés avec une probabilité de $1 - \alpha$ que

$$-\Phi^{-1}(1 - \frac{\alpha}{2})\sqrt{\frac{Var[g(x, Z)]}{M}} \leq \hat{\theta}_M - E[g(x, Z)] \leq \Phi^{-1}(1 - \frac{\alpha}{2})\sqrt{\frac{Var[g(x, Z)]}{M}}$$

tout simplement en se basant sur la relation $1 - \alpha = 2\Phi(c) - 1$. Quelques manipulations algébriques nous permettent de reconnaître qu'avec une probabilité de $1 - \alpha$,

$$\hat{\theta}_M - \Phi^{-1}(1 - \frac{\alpha}{2})\sqrt{\frac{Var[g(x,Z)]}{M}} \leq E[g(x,Z)] \leq \hat{\theta}_M + \Phi^{-1}(1 - \frac{\alpha}{2})\sqrt{\frac{Var[g(x,Z)]}{M}} \ .$$

Si la variance de $g(x, Z)$ était connue, nous aurions donc en mains un intervalle qui devrait contenir, avec une probabilité de $1-\alpha$, la vraie valeur espérée de $g(x, Z)$. En pratique, cette variance n'est pas plus connue que la valeur espérée, hors il est commun de simplement remplacer cette variance par la variance empirique pour obtenir un estimé de l'intervalle en question.

On en conclut qu'à partir de l'estimateur empirique $\hat{\theta}_M$ de $E[g(x, Z)]$ et de l'estimateur empirique $\hat{\sigma}^2_M$ de $Var[g(x, Z)]$, il est possible d'utiliser l'intervalle

$$E[g(x,Z)] \in \left[\hat{\theta}_M - \Phi^{-1}(1 - \frac{\alpha}{2})\frac{\hat{\sigma}_M}{\sqrt{M}}, \hat{\theta}_M + \Phi^{-1}(1 - \frac{\alpha}{2})\frac{\hat{\sigma}_M}{\sqrt{M}}\right]$$

où $\Phi^{-1}(\cdot)$ est la fonction de répartition inverse de la distribution normale, comme intervalle confiance, à probabilité $1 - \alpha$, pour la vraie valeur espérée de $g(x, Z)$.

Remarque 5.7.1. *Il est à noter qu'une approche Bayesienne peut toujours être employée pour se faire une idée d'un tel intervalle de confiance. L'approche Bayesienne requiert cependant de faire des hypothèse au sujet de la forme de la distribution de $g(x, Z)$ (ex. la loi normale) ainsi qu'une hypothèse par rapport à la probabilité à priori liée à chacune des instances de cette forme (ex. la moyenne suit elle-même une loi normale). L'intervalle qui est obtenu dans cette section par l'application du théorème limite centrale ne requiert rien de plus que le fait que la variance de $g(x, Z)$ soit bornée.*

Remarque 5.7.2. *Il est aussi important de réaliser que contrairement à une approche Bayesienne qui considérerait la moyenne de $g(x, Z)$ comme aléatoire, l'approche utilisée dans cette section ne considère pas que la moyenne de $g(z, Z)$ est une variable aléatoire. La notion de probabilité et d'intervalle de confiance ne nous vient que du fait que les données sont générées aléatoirement. En fait, la conclusion tirée n'est pas que conditionnellement à l'observation des données la probabilité que $E[g(x, Z)]$ soit dans l'intervalle est $1 - \alpha$ (puisque $E[g(x, Z)]$ n'est pas aléatoire) mais plutôt que lorsque nous générons des données aléatoirement selon la loi de $g(x, Z)$ nous avons $1-\alpha$ probabilité que $E[g(x, Z)]$ se retrouve dans l'intervalle calculé à partir des données.*

5.7.2 L'erreur d'une estimation de centile

Appelons $\hat{z}^M_\alpha$ l'estimateur empirique du centile d'ordre α de la distribution de Z : i.e. le centile d'ordre α de la fonction empirique $F^M_Z(z)$. Cette estimation de statistique est à nouveau sujette à être erronée puisque F^M_Z est une version approximative et aléatoire de F_Z, puisqu'elle dépend d'un nombre limité de données obtenues aléatoirement. Nous considérerons

de plus que $Z_{(1)}, Z_{(2)}, ..., Z_{(M)}$ est la version ordonnée de l'échantillon $Z_1, Z_2, ..., Z_M$ et souhaitons déterminer un intervalle de confiance pour $z_\alpha := F_Z^{-1}(\alpha)$ à partir de $\hat{z}_\alpha^M = Z_{(\lceil M\alpha \rceil)}$,où $\lceil x \rceil$ est le plus petit entier supérieur à x. [1]

Théorème 5.7.1. *Si M est assez large, alors la vraie valeur de z_α a une probabilité plus large que $1-\beta$ d'être dans l'intervalle :*

$$Z_{(\lfloor M(\alpha-\Delta_\beta) \rfloor)} \leq z_\alpha < Z_{(\lceil M(\alpha+\Delta_\beta) \rceil + 1)}$$

où $\Delta_\beta = \Phi^{-1}(1-\frac{\beta}{2})\sqrt{\alpha(1-\alpha)}/\sqrt{M}$ et où $\Phi^{-1}(\cdot)$ est à nouveau la fonction de répartition inverse de la loi normale.

Nous donnerons un exemple d'application de ce théorème sous peu. Observons pour l'instant que ce théorème nous demande d'aller identifier des termes qui sont plus tard et plus tôt que $Z_{(\lceil M\alpha \rceil)}$ dans la liste $Z_{(1)}, Z_{(2)}, ..., Z_{(M)}$. Plus le centile α demandé est prêt de 0.5, plus loin devrons nous aller dans la liste alors que si M augmente (i.e. le nombre de données) des termes plus prêts de $Z_{(\lceil M\alpha \rceil)}$ seront utilisés.

Démonstration. Nous prouverons ce fait en étudiant les propriétés de la variable aléatoire i^* qui caractérise l'index du dernier élément plus petit ou égal au centile z_α dans la liste ordonnée des échantillons, i.e. $Z_{(i^*)} \leq z_\alpha < Z_{(i^*+1)}$. Pour tout M, si Z a une fonction de répartition continue à z_α alors de par la définition d'un centile, chaque échantillon Z_i a exactement α chance de tomber à gauche de la valeur z_α. Le nombre d'échantillons dans un ensemble de M échantillons tombant à gauche de la valeur z_α est donc distribué selon la loi binomiale(M, α) qui pour M assez large ressemble à $\mathcal{N}(M\alpha, M\alpha(1-\alpha))$. La variable aléatoire i^* correspond exactement à cette valeur. C'est donc qu'il y a une probabilité de $1-\beta$ que i^* tombe dans l'intervalle :

$$M\alpha - \Phi^{-1}(1-\beta/2)\sqrt{M\alpha(1-\alpha)} \leq i^* \leq M\alpha + \Phi^{-1}(1-\beta/2)\sqrt{M\alpha(1-\alpha)}\ .$$

C'est donc que nous avons une confiance de niveau $1-\beta$ que la valeur z_α se retrouve dans l'intervalle

$$Z_{(\lfloor M(\alpha-\Delta_\beta) \rfloor)} \leq z_\alpha \leq Z_{(\lceil M(\alpha+\Delta_\beta) \rceil + 1)}$$

où $\Delta_\beta = \Phi^{-1}(1-\beta/2)\sqrt{\alpha(1-\alpha)}/\sqrt{M}$. □

Exemple 5.7.1. *Supposons que la taille de l'échantillon est 10000, nous pouvons déterminer les index des valeurs de l'échantillon ordonné qui permettent de déterminer l'intervalle de confiance de 95% du 10ième centile. D'abord, considérant que $\beta = 0.05$, il est possible de vérifier que $\Phi^{-1}(1-\beta/2) = \Phi^{-1}(0.975) \approxeq 1.96$. L'erreur relative sur l'index de ce centile est donc*

$$\Delta = 1.96 \times \sqrt{0.1(1-0.1)}/\sqrt{M} = 0.00588\ .$$

Nous devons finalement aller retracer les échantillons qui sont situés aux index $\lfloor 10000 \times (0.1 - 0.00588) \rfloor = 941$ et $\lceil 10000 \times (0.1 + 0.00588) \rceil + 1 = 1060$ de notre liste ordonnée

1. On remarquera que lorsque Z représente un coût, notre définition de $\hat{z}_\alpha$ semble être un estimé conservateur puisque si $\frac{k-1}{M} < \alpha \leq \frac{k}{M}$ alors $\hat{z}_\alpha^M = Z_{(k)}$, i.e. la valeur la plus élevée des deux.

d'échantillon. Ceci est approximativement équivalent à dire que le 10ième centile de la distribution de Z est situé quelque part entre le 9.41ième et le 10.60ième centile de la distribution empirique observée dans les données. La valeur la plus vraisemblable de ce centile est bien entendu le 10ième centile de la distribution empirque.

Deuxième partie

Prise de décision sous incertitude

Jusqu'à présent nous nous sommes attardés sur la question de la modélisation de l'incertitude et de l'évaluation des risques associés à une décision. Lorsque vient le temps de prendre une décision, il est cependant souhaitable que la décision soit la meilleure parmi toutes les décision qui peuvent être implémentées. Dans les chapitres qui suivent, nous étudierons la question de l'optimisation d'une mesure qui décrit quel genre de compromis nous sommes prêts à faire entre les retours (monétaires) espérés et les risques (monétaires) auxquels nous seront potentiellement exposés.

Chapitre 6

Problèmes dynamiques

Tel que vu précédemment, la simulation par la méthode de Monte-Carlo semble être l'outil idéal pour comparer les risques associés à différentes alternatives. Malheureusement cet outil n'est plus adéquat, lorsque nous souhaitons ≪optimiser≫ des décisions qui s'échelonnent dans le temps. La difficulté est causée par le fait que lorsqu'un problème de décision implique un aspect dynamique, ou temporel (i.e. différentes décisions vont être appliquées à différent moment). il devient nécessaire de comparer des stratégies plutôt que des alternatives.

Définition 6.0.1. *Une stratégie est un plan qui décrit exactement quelle action entreprendre à tout moment et peu importe les circonstances. En d'autres mots, une stratégie est un document que vous donneriez à une personne en charge de la mise en oeuvre de votre plan. A tout moment, cette personne devrait être en mesure de se référer à ce document et savoir clairement ce qu'il faut faire ensuite.*

Par exemple, la décision d'acheter un billet de loto ne peut être considérée ≪bonne≫ si nous décidons plus tard de jeter le billet avant de savoir quels sont les numéros gagnants. La stratégie ≪acheter et garder précieusement jusqu'au tirage≫ est donc nécessairement meilleure que la stratégie ≪acheter et jeter la veille du tirage≫. En effet, dans un problème de décision dynamique les actions que nous mettons en oeuvre à chaque moment peuvent être différentes selon ce qui s'est passé jusqu'à ce moment-là. Lors de la prise de décision initiale, il est donc important de décrire précisément quelle sera la stratégie à adopter et vérifier qu'il s'agit réellement du meilleur plan d'action.

6.1 Les arbres de décision

Un ≪arbre de décision≫ permet de décrire clairement quelles sont les stratégies possibles et de les comparer entre elles de manière efficace. Il sera constitué des éléments suivants :

— Noeud de décision (carré) : ce type de noeud décrit une situation dans laquelle nous devrons implémenter une action. Chaque noeud de décision possède une branche pour chacune des actions implémentables dans la situation.

— Noeud d'événement aléatoire (cercle) : ce type de noeud décrit une situation dans laquelle nous observerons la réalisation d'un ou plusieurs éléments incertains. Chaque noeud d'événement aléatoire possède une branche pour chacun des événements possibles liés à cette réalisation. Les événements associés aux branches d'un même noeud doivent être mutuellement exclusifs et collectivement exhaustifs (un et seulement un des événements aura lieu).
— Noeud terminal : ce type de noeud décrit une situation dans laquelle nous pourrions nous retrouver lorsque viendra le temps d'évaluer la performance de nos actions.

Remarque 6.1.1. *L'ordre des noeuds dans l'arbre est crucial. Les noeuds d'événement et de décision doivent être insérés chronologiquement : i.e, un noeud de décision doit être placé après tous les noeuds des événements connus et devant tous les noeuds d'événements qui lui sont inconnus au moment d'implémenter la décision.*

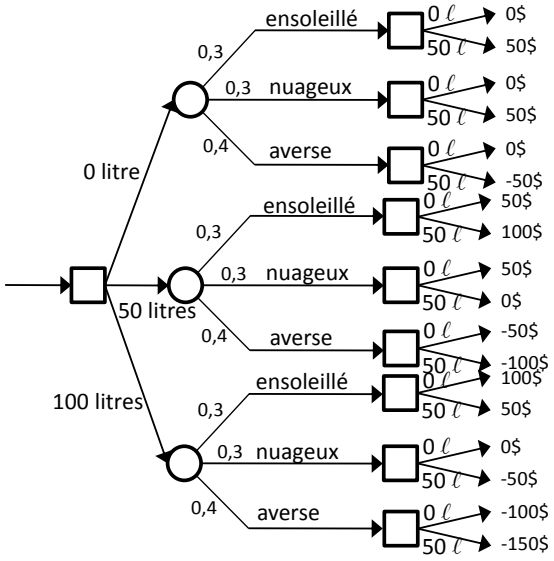
6.2 Résolution par programmation dynamique

Considérez la situation suivante :

> « Pour se faire un peu d'argent de poche, votre neveu vend des verres de limonade le samedi dans un stand installé le long d'une avenue passante de son quartier. Récemment, votre neveu est venu s'enquérir de votre aide dans le choix de la quantité de limonade qu'il devrait préparer pour samedi prochain. Il estime que les coûts de production seront de 1$/litre et a choisi de fixé le prix de vente à 2$/litre. Si la journée est ensoleillée, il prévoit qu'il y aura suffisamment de clientèle pour vendre 100ℓ de limonade. Autrement, il estime la demande à 50ℓ si la journée est nuageuse et 0ℓ s'il y a averse. Après consultation du journal et de bulletins télévisés, il estime les probabilités suivantes pour la météo de samedi comme suit :
>
> | | Ensoleillé | Nuageux | Averse |
> |---|---|---|---|
> | Probabilité | 30% | 30% | 40% |
>
> De plus, il croit avoir le temps de préparer jusqu'à 50 litres samedi matin, alors qu'il aura une meilleure idée de la météo. Il souhaite déterminer la production qui maximise son profit espéré. »

Remarque 6.2.1. *On appel «action de recours» une action qui est implémentée une fois que le statut d'un événement est connu. Cette action peut dépendre du statut de l'événement qui est observé. Il est important d'insérer le noeud représentant une action de recours chronologiquement dans l'arbre : i.e, après les noeuds des événements connus et devant les événements qui ne seront toujours pas connus. Dans le problème du stand de limonade, la possibilité de préparer 50 litres de limonade durant la journée des ventes constitue une action de recours.*

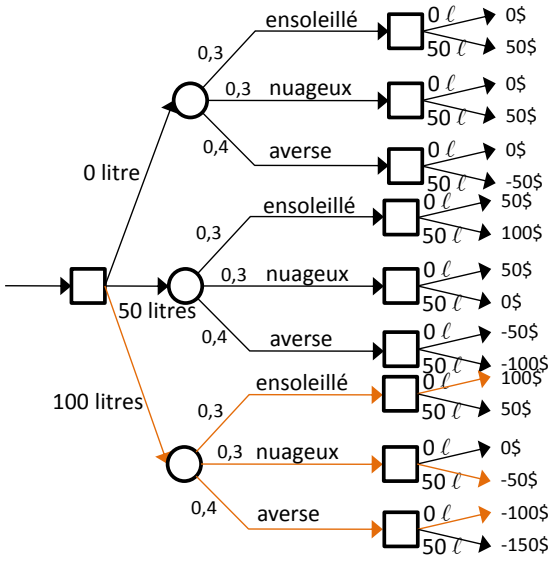
La figure suivante décrit l'arbre de décision associé à ce problème. Il est à noter qu'afin d'avoir un nombre de branches raisonnable à chacun des noeuds de décision, nous nous limiterons aux actions de produire 0, 50, ou 100 litres.



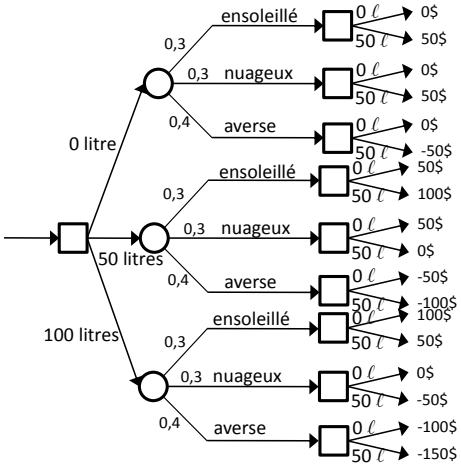
Dans cet arbre, il est possible d'identifier jusqu'à 24 stratégies différentes. Chacune d'entre elles correspond à faire le choix d'une des branches pour chacun des noeuds de décision de gauche à droite pour chacun des noeuds de décision atteignable. Par exemple, le plan d'action présenté dans la figure suivante décrit la stratégie : produire initialement 100 litres, observer la température samedi matin et produire 50 litres s'il fait nuageux, sinon produire 0 litre. Il ne s'agit bien sûr pas de la stratégie optimale.

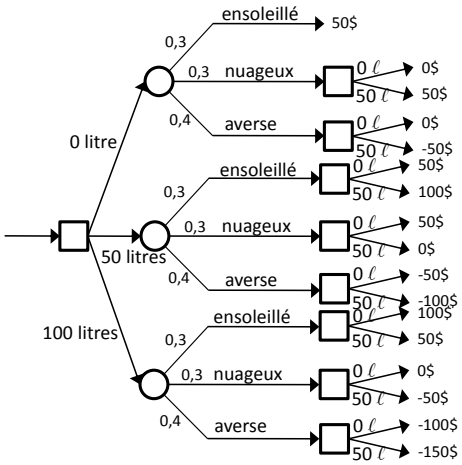
Lorsque l'objectif est de maximiser la valeur monétaire espérée, on peut trouver la stratégie optimale en utilisant une méthode appelée «programmation dynamique». La méthode s'applique comme suit :

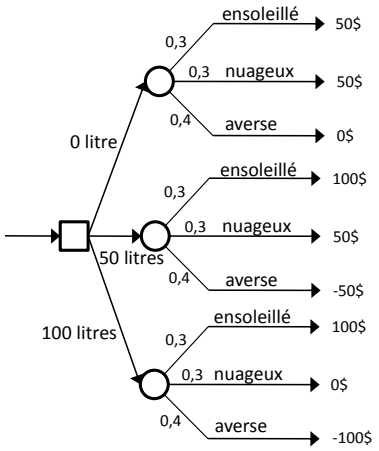
1. Choisir un noeud dont toutes les branches mènent à des noeuds terminaux.
   — S'il s'agit d'un noeud d'événement, prendre la moyenne pondérée des valeurs monétaires associées à chacune des branches
   — S'il s'agit d'un noeud de décision, prendre la valeur monétaire maximale parmi celles associées à chacune des branches
2. La valeur obtenue devient la valeur monétaire associée au noeud
3. Enlever chacune des branches du noeud considérant ainsi que le noeud est devenu un noeud terminal
4. Répéter les étapes ci-dessus dans l'ordre jusqu'à ce qu'il ne reste qu'un noeud terminal. La valeur de ce dernier noeud est égale à la valeur espérée de la stratégie optimale à employer dans ce problème. La stratégie optimale peut être identifiée en sélectionnant la décision considérée optimale (i.e. qui atteint la valeur la plus élevée) à chacun des noeuds de décision de l'arbre original.

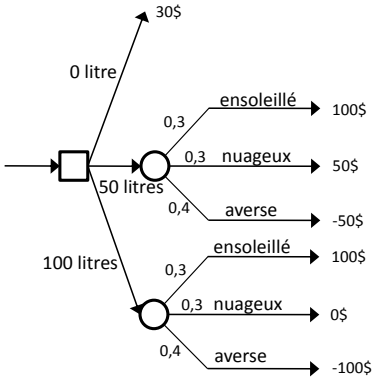


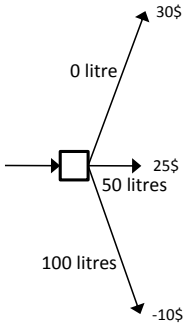
Les figures qui suivent illustrent comment appliquer la programmation dynamique au problème du stand de limonade.

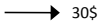










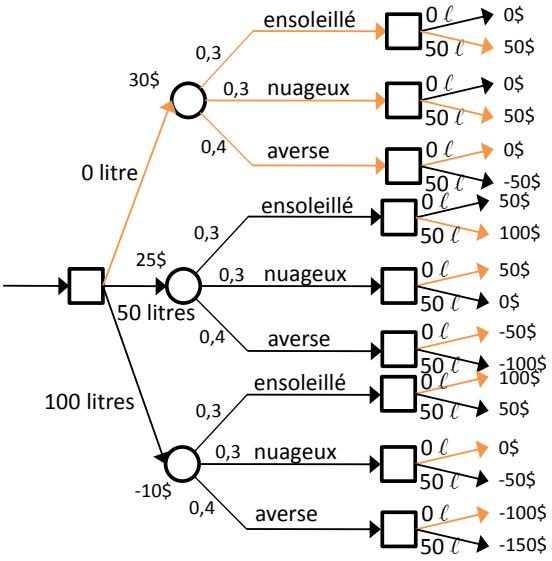


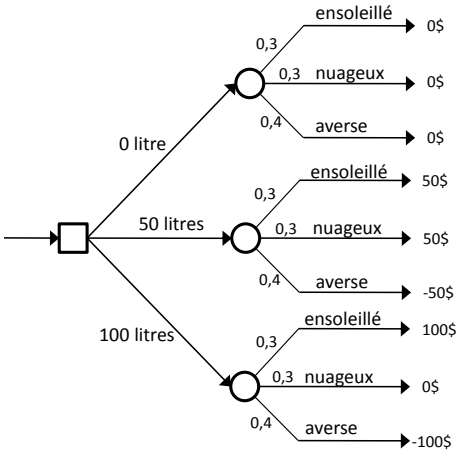
La valeur espérée de la stratégie optimale est de 30$. Cette valeur espérée peut-être atteinte en suivant la stratégie décrite dans la figure suivante : i.e. produire 0 litre initialement et produire 50 litres le samedi matin seulement s'il fait nuageux ou ensoleillé.

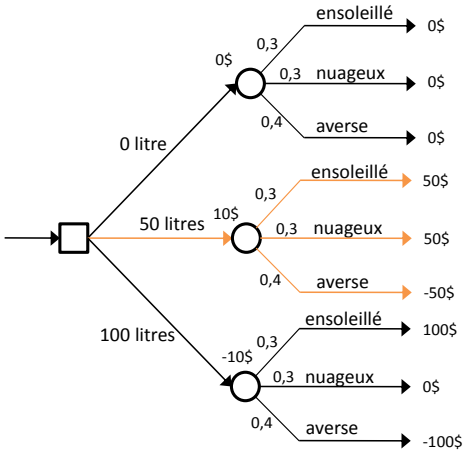
Considérez la nouvelle situation suivante :

> « Votre neveu vient de se faire demander s'il pourrait s'occuper de sa petite soeur samedi matin. S'il acceptait il ne pourrait préparer de limonade le matin même. Avant de répondre, il aimerait savoir quelle est la valeur de garder sa liberté tant qu'à l'usage de son temps. »

Il est possible d'évaluer la «valeur de la flexibilité» en se demandant quel est le montant que nous serions prêts à prendre pour avoir moins de flexibilité dans un problème donné. Pour répondre à la question de votre neveu, nous pouvons mesurer quel est l'impact sur la valeur espérée de la nouvelle stratégie optimale lorsque nous fixons l'action de recours concerné.







En calculant la différence entre les deux valeurs espérées obtenues, nous obtenons que la valeur pour votre neveu de garder sa liberté pour le samedi matin est de 20$. En effet, il faut ajouter au moins 20$ à chacun des noeuds terminaux pour que l'arbre de décision sans flexibilité mène à une stratégie optimale qui a la même valeur que la stratégie optimale qui profite de la flexibilité. Malgré que sa flexibilité vaille 20$, dans ce qui suit nous considérerons que votre neveu accepte tout de même de garder sa soeur.

Remarque 6.2.2.

Malheureusement, la grandeur d'un arbre de décision augmente très rapidement dans un problème réaliste. Si nous prenons l'exemple du stand de limonade, déjà pour planifier la production sur une semaine entière, il serait nécessaire d'évaluer les profits atteints pour au moins 2187 différents situations terminales possibles (i.e. noeuds terminaux). En effet, nous aurions au moins un noeud terminal pour chaque séquence de météo possible (ex. «AA-NAENE»), donc un total de $3^7 = 2187$ *scénarios. Pour un mois, le compte est de* $3^{30} > 10^{14}$

scénarios. Typiquement, un arbre de décision sera utilisé pour acquérir de l'intuition quant à la structure de la stratégie optimale en considérant une version simplifiée du problème. Cette intuition peut ensuite servir pour choisir une stratégie plus réaliste dont les risques seront évalués à l'aide d'une simulation Monte-Carlo.

6.3 Valeur de l'information

La valeur de l'information est le prix maximal que nous sommes prêts à payer pour connaître la valeur d'un paramètre incertain avant de prendre notre décision. Celle-ci est nécessairement positive puisque prendre une décision plus éclairée mènera toujours à une meilleure performance. De plus, la valeur de l'information ne peut être non-nulle seulement que si la connaissance du paramètre causerait la prise d'une action différente pour au moins une des réalisations.

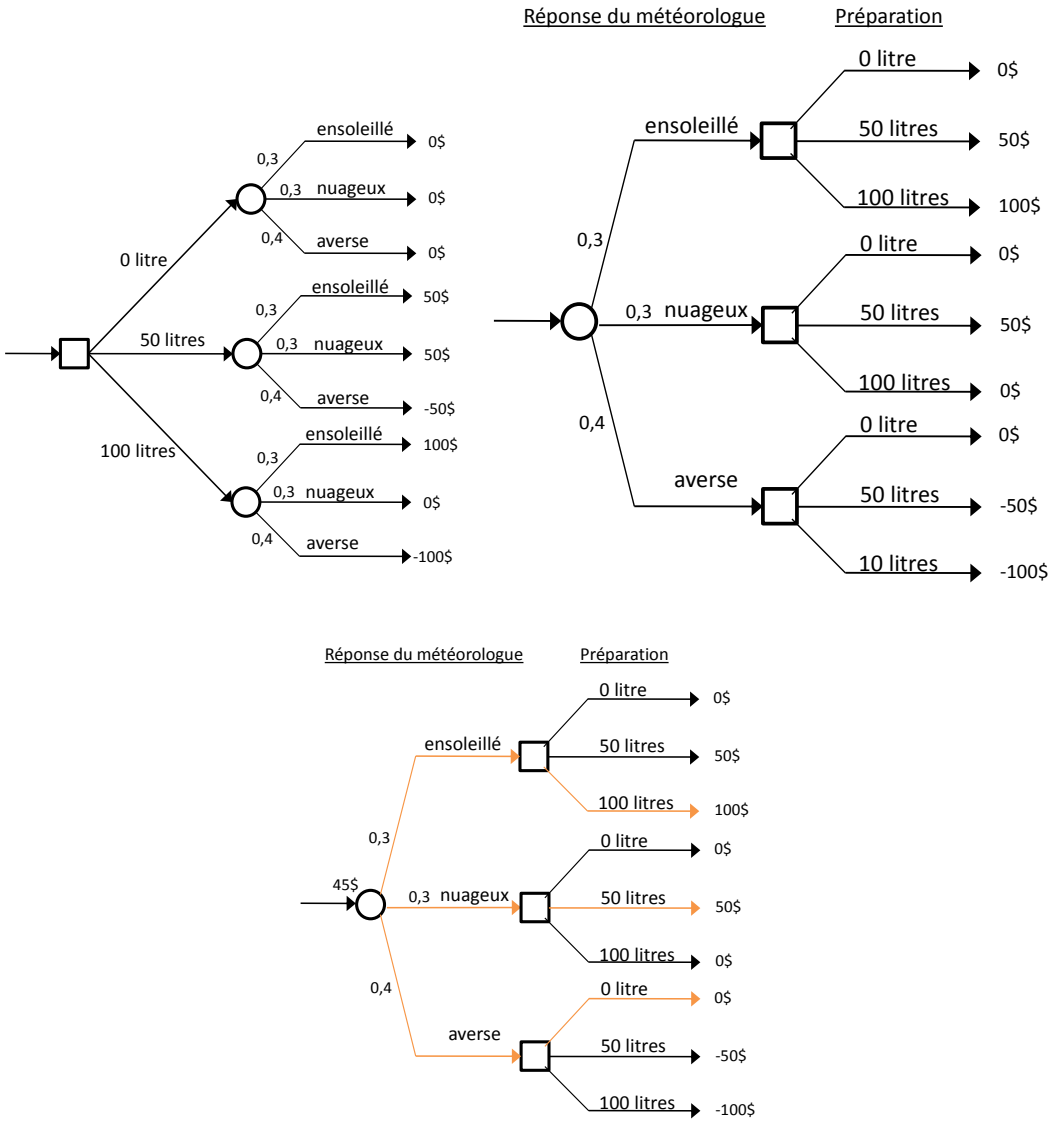
> « Jusqu'à combien votre neveu devrait-il être prêt à payer pour connaître la météo de samedi avant de faire sa production de limonade ? »

6.3.1 Valeur espérée de l'information parfaite

Nous considérerons d'abord la valeur espérée de l'information parfaite à propos d'un paramètre,c'est-à-dire quelle est la valeur espérée de connaître exactement la valeur du paramètre en question. La décision dépenderait alors de l'information reçue. Pour en faire l'analyse il s'agit de considérer la situation dans laquelle le paramètre est observé avant de prendre la décision initiale.

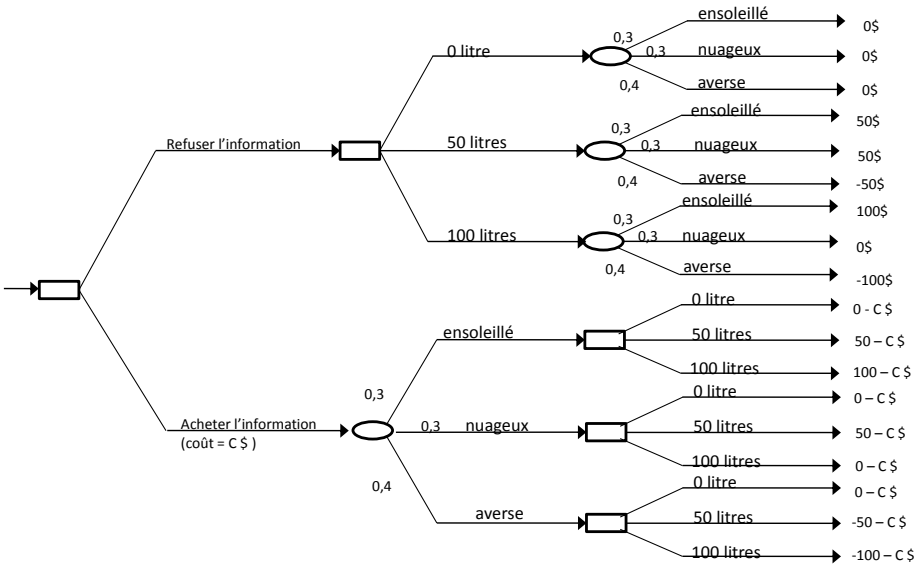
Alors que l'arbre de décision associé au problème où l'information n'est pas acquise prenait la forme de la figure suivante de gauche (rappelez-vous que votre neveu n'a plus d'action de recours puisqu'il gardera sa petite soeur). Si l'information était acquise alors l'arbre serait plutôt tel que la figure de droite le décrit. Il est à noter que l'ordre du noeud de production et du noeud de météo a été inversé.

En appliquant la programmation dynamique, nous obtenons la stratégie optimale décrite dans la figure suivante. La valeur espérée de cette stratégie est de 45$. C'est donc que nous payons moins de 45-10=35$ pour cette information (i.e. réduire la valeur de chaque noeud terminal de 35$), il est toujours préférable de l'acquérir.

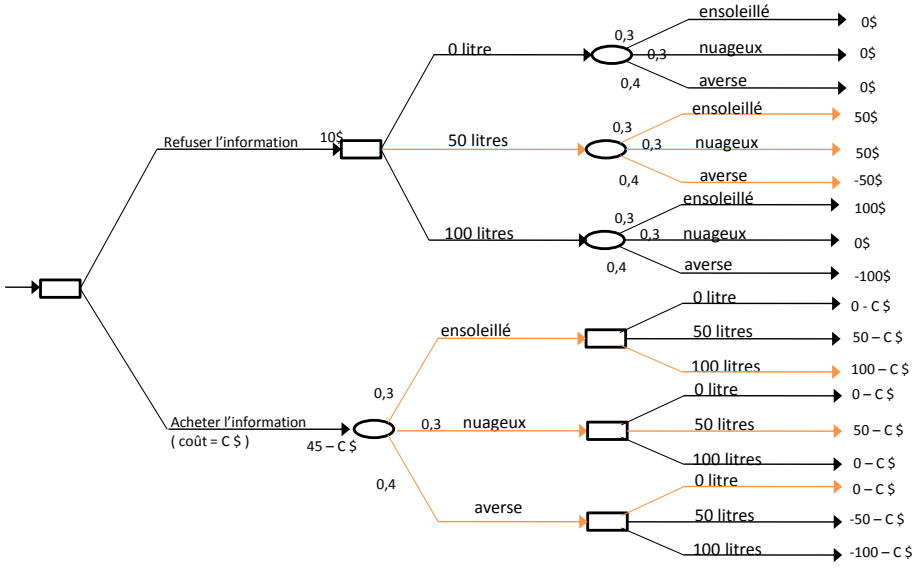


Remarque 6.3.1. *En fait, une analyse plus rigoureuse pourrait se baser sur l'arbre de décision suivant. Dans cet arbre, nous représentons le fait que la première décision à prendre est celle d'acheter ou non l'information. Si l'information est achetée, alors la météo sera connue avant de produire et nous devrons déduire le coût C d'achat de l'information à chacun des noeuds terminaux. Autrement, la production doit être faite et les conséquences liées à la météo sont ensuite observées.*

Lorsque nous appliquons la méthode de programmation dynamique en considérant C comme une constante, nous obtenons la solution partielle suivante. Pour conclure, il faudrait comparer la valeur des deux noeuds d'événement 10 et $45 - C$. Pour que l'information soit acquise par la stratégie optimale, il est nécessaire que $45 - C$ soit plus grand que 10. Nous en concluons que l'information n'est intéressante que si son coût est plus petit que 35$. La



valeur de l'information parfaite est donc 35$.



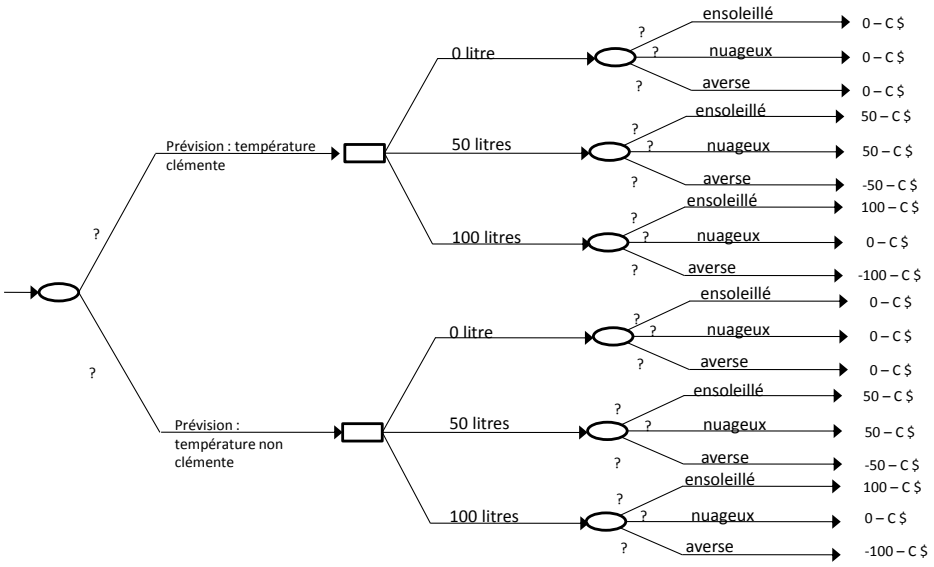
6.3.2 Valeur espérée de l'information imparfaite

Nous considérons maintenant une situation dans laquelle l'information qui nous vient ne nous renseigne pas parfaitement par rapport à la nature du paramètre qui affecte la performance de notre décision. L'information devra donc être utilisée pour rafraîchir nos connaissances des probabilités (à l'aide du théorème de Bayes). Nous appellerons le paramètre qui affecte la performance de nos actions M, comme pour «Météo», et considérerons que l'information qui sera obtenue est I, pour «Information». Puisque l'information est imparfaite,

il sera nécessaire d'identifier les différents scénarios (réalisations de I) pour l'information qui sera reçue, de quantifier la probabilité de chacune de ces réalisations pour I, ainsi que pour chacune de ces réalisations quelle est la loi de probabilité conditionnelle associée à M lorsque I est observé. En d'autres mots, caractériser quel type d'information nous pensons recevoir (et la probabilité de chacun de ces types), et quel effet l'information aura sur notre perception des chances liées à la réalisation de M à l'aide des probabilités. Nous illustrerons le tout à l'aide de notre problème de stand de limonade.

> ≪ Votre neveu compte consulter un expert météorologue. Celui-ci lui dira s'il croit que la température sera clémente ou non samedi prochain. Considérant les prévisions passées faites par ce météorologue, votre neveu cumule les statistiques présentées dans le tableau suivant :
>
	Information de l'expert I	
La réalité M	Température clémente I_1	Température non-clémente I_2
Ensoleillé M_1	$P(I_1\|M_3) = 1$	$P(I_2\|M_1) = 0$
Nuageux M_2	$P(I_1\|M_3) = \frac{1}{2}$	$P(I_2\|M_2) = \frac{1}{2}$
Averse M_3	$P(I_1\|M_3) = 0$	$P(I_2\|M_3) = 1$
>
> Quelle est donc la valeur d'obtenir l'opinion de ce météorologue ? ≫

Comme dans le cas de la valeur de l'information parfaite, il s'agit d'estimer quel est le prix le plus élevé pour lequel il est toujours préférable d'acquérir l'information. Ceci peut-être présenté à l'aide de l'arbre de décision suivant :

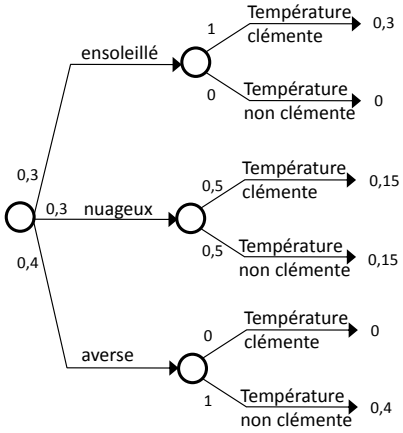


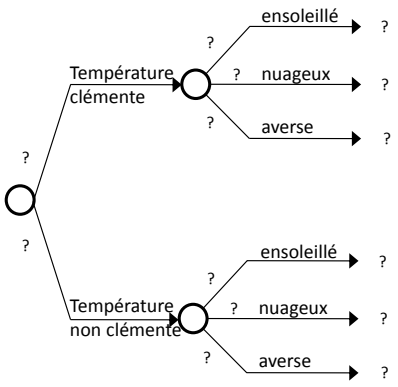
Pour compléter les valeurs manquantes de cet arbre, il est nécessaire d'appliquer le théorème de Bayes. Rappelons-nous que si $\{I_1, I_2, ...\}$ sont les différents types d'information que nous pourrions recevoir et si $\{M_1, M_2, \cdots\}$ sont les événements futurs qui affectent

directement la valeur de notre décision, et finalement si nous connaissons $P(I_j|M_i)$ pour chacune des paires (i, j), alors

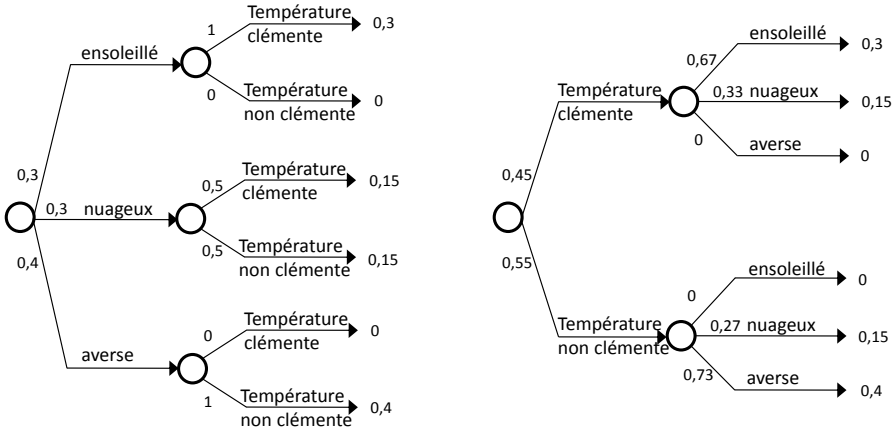
$$P(M_i|I_j) = \frac{P(I_j|M_i)P(M_i)}{P(I_j)} \quad ,$$

où $P(I_j) = \sum_i P(I_j|M_i)P(M_i)$. Nous pouvons résoudre ce calcul à l'aide des arbres de scénarios suivants :

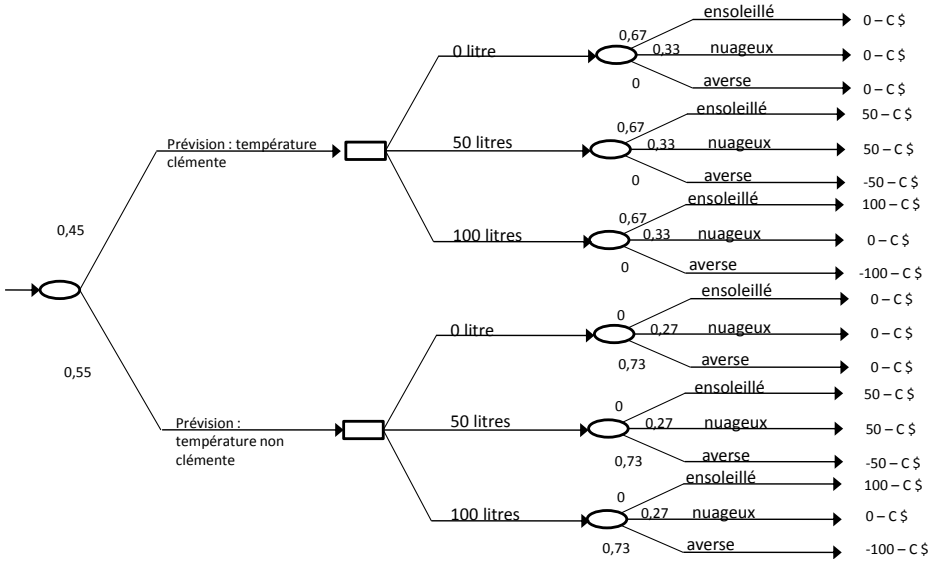




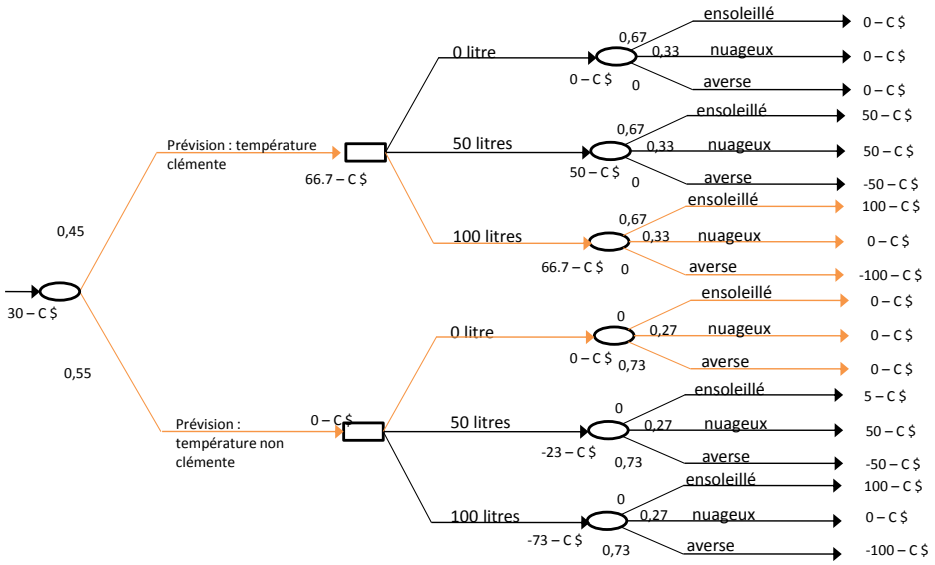
L'arbre de gauche présente l'information que nous connaissons et l'arbre de droite présente l'information que nous cherchons à connaître pour évaluer la performance atteignable si l'information était acquise. Suivant la méthode étudié précédemment nous obtenons :



Il s'agit maintenant de résoudre l'arbre de décision qui étudie la valeur du problème juste avant d'obtenir l'information considérant qu'un coût de C a été payé pour obtenir l'information.



La stratégie optimale et sa valeur est présentée dans l'arbre de décision suivant.



La valeur de l'information imparfaite est donc de $30 - 10 = 20\$$ puisque si C dépasse 20\$ alors la stratégie optimale qui utilise l'information génère moins de profit que celle qui se débrouille sans elle (et donc évite de débourser le montant C) (il serait donc profitable de ne pas acquérir cette information). Mathématiquement, nous cherchons le coût maximal pour lequel 30-C \$ est plus grand que 10\$.

Remarque 6.3.2. *Vous remarquez peut-être que la difficulté de l'exercice d'estimation de la valeur de l'information imparfaite réside principalement dans l'estimation des probabilités conditionnelles $P(M|I)$ et de $P(I)$. Celui-ci serait grandement simplifié si on nous donnait directement ces probabilités plutôt que de passer par $P(I|M)$ et l'application du théorème de Bayes. Malheureusement, en pratique l'estimation de ces probabilités doit se faire avec attention. Dans le cas par exemple où l'on utiliserait une méthode de sollicitation afin d'en formuler une estimation subjective (à l'aide de la méthode de la roue de fortune par exemple), il est crucial de vérifier que les valeurs de $P(M|I)$ et de $P(I)$ sont cohérentes avec les valeurs déjà connues pour $P(M)$: rappelez vous que $P(M) = \sum_i P(M|I_i)P(I_i)$. En terme simple, il faut s'assurer que notre caractérisation de la qualité de l'information de l'expert et des chances d'obtenir une information ou une autre soit cohérente avec notre connaissance de M. Dans le cas où l'on utilise des données historiques de prédictions faites par cet expert, il peut sembler opportun de directement estimer $P(I)$ et $P(M|I)$ à partir des proportions observées. Malheureusement, ceci mène typiquement à de l'incohérence dans l'analyse. Pensez par exemple à un cas où les données historiques recensent la qualité des prédictions météorologiques pour un été exceptionnel dans lequel il n'a jamais plu. Dans ce cas, il est clair qu'à chaque fois que le météorologue aurait prédit du beau temps, alors il aurait eu raison. Utiliser ces proportions pour estimer les valeurs de $P(M|I)$ mènerait à la conclusion que le météorologue a toujours raison lorsqu'il annonce du beau temps et*

toujours tort lorsqu'il annonce le contraire. Si nous savons à priori que la période qui nous intéresse n'est pas un été exceptionnel, alors nous ne pouvons pas utiliser ces proportions puisque $P(M)$ ne ressemble pas aux proportions présentes dans les données historiques. À ce titre, on pourrait penser qu'il est impossible de trouver des données historiques qui répliquent exactement les conditions de la période qui nous intéresse. Heureusement, en passant par $P(I|M)$ et le théorème de Bayes, il est au contraire toujours possible de se faire une idée cohérente de $P(I)$ et $P(M|I)$, à la condition que chacune des réalisations de M figure à multiples reprises (pour une meilleure estimation) dans nos données.