Rappels en probabilité (utiles en statistique bayésienne)

Sophie Ancelet

[1]IRSN, PSE-SANTE/SESANE/LEPID, Fontenay-Aux-Roses, France (sophie.ancelet@irsn.fr)

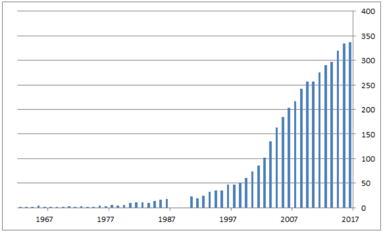
Ecole-chercheurs BioBayes 2019, Pornic, 18 Mars 2019

Course outline

IRSN INSTITUT DE RADIOPROTECTION ET DE SÛRETÉ NUCLÉAIRE

La statistique bayésienne

- Branche la plus récente de la statistique (même si historiquement la plus ancienne !)
- **Remarquable** accroissement de sa **popularité** sur les 20 dernières années
- **De plus en plus utilisée** en ingénierie mathématique et, plus généralement, en épidémiologie, environnement, alimentation, génétique, fiabilité, ...
  - $\nearrow$ Nombre d'articles portant sur "Bayesian statistics" pour 100000 articles publiés dans la base de données medline (littérature biomédicale) - Période 1963-2017



Le rôle important des probabilités

Les 3 étapes principales de l'analyse de données bayésienne

1. Proposer un modèle **probabiliste** (i.e., **une loi de probabilité jointe**) permettant de décrire un système aléatoire dans lequel **le hasard** intervient, en vue de l'expliquer et/ou de le prédire :
   - Ingrédients de base : les **variables aléatoires**
   - Statistique bayésienne : spécification de **lois de probabilité dites a priori** sur toutes les grandeurs inconnues du système
2. **Conditionnellement** à des réalisations observées du système d'intérêt (les données), calculer analytiquement ou, le plus souvent, **générer des valeurs aléatoires** selon la **loi de probabilité dite a posteriori** des quantités inconnues
   - Utilisation de **lois de probabilité conditionnelles** !
3. Evaluer le modèle : Le modèle s'ajuste t'il correctement aux données ? Permet-il de **prédire** des données plausibles ? Résultats sensibles aux hypothèses de modélisation ?

⇒ Pour bien comprendre et utiliser la statistique bayésienne, des connaissances de base sont requises **en théorie des probabilités** *(Chapitre 1, pages 3-69, Livre Biobayes)*

Incertitude, modélisation et statistique appliquée

Lindley (2006)

"Uncertainty is everywhere and you cannot escape from it."

- Quand le modélisateur-statisticien cherche à décrire, expliquer et/ou prédire un système biologique ou physique complexe, son **incertitude** peut être importante
  - Données de terrain et/ou expérimentales **peu nombreuses** et/ou **peu informatives**
  - **Fluctuations non contrôlables** du système étudié en l'état de connaissances actuelles
  - Incertitude sur certaines quantités inconnues **non observables**
- Encore **trop souvent négligée ou mal prise en compte** par le scientifique
  - $\Rightarrow$ Estimations et/ou prédictions potentiellement fausses !
- **Prendre en compte "proprement" l'incertitude** dans l'analyse d'un système aléatoire
  - $\Rightarrow$ Peut conduire à des estimations et/ou prédictions approximativement correctes !

IRSN INSTITUT DE RADIOPROTECTION ET DE SÛRETÉ NUCLÉAIRE

L'incertitude est partout ! (1/3)

Lancer d'un dé... Alea Jacta est !

- Quelle face s'affichera au terme du lancer d'un dé ?
- Lancers consécutifs ⇒ **Résultats différents régis par le hasard...**
  - Exemple : 10 lancers ⇒ 2 4 3 6 1 6 1 3 5 1

Mesurer le diamètre d'une pastèque

- Plusieurs mesures de la même pastèque ⇒ **Résultats différents** (qualité de l'appareil de mesure, rigueur des observateurs)
- Plusieurs mesures de pastèques issues d'un même champ ⇒ **Résultats différents** (variabilité du diamètre)

Temps mis pour se rendre à son travail le matin

- 5 jours consécutifs la semaine dernière ⇒ **Résultats différents**
  - 22 min, 35 min, 21 min, 30 min, 25 min
- **Aléa** lié par exemple à l'occurrence ou non d'embouteillages, d'incidents de voyageurs, de grèves,...

L'incertitude est partout ! (2/3)

Nombre de chromosomes dicentriques dans 1 cellule irradiée à 1 Gray

- Plusieurs mesures pour une même cellule ⇒ **Résultats différents** (rigueur des observateurs)
- Plusieurs mesures de cellules irradiées à la même dose de 1 Gray ⇒ **Résultats différents** (variabilité biologique et physique)

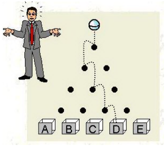
Collecter l'âge au décès de 100 mineurs d'uranium (décédés)

- **Variabilité inter-individuelle** de l'âge au dècès : styles de vie différents, différentes expositions professionnelles, environnementales, médicales aux rayonnements ionisants et autres pathogènes, prédispositions familiales, ...

Notions d'aléa et d'expérience aléatoire

Qu'appelle-t-on "aléa" ?

Aléa vient du latin *alea* qui signifie "jeu de dés". Il peut être vu comme la cause de la part imprévisible des résultats d'une expérience qui, même dans des conditions expérimentales supposées identiques, peut donner lieu à des résultats différents



Expérience aléatoire

Expérience plus ou moins complexe :

- dont on ne peut prévoir par avance le résultat $\Rightarrow$ Résultat **incertain**
- qui peut être répétée indéfiniment

Notions d'expérience aléatoire et d'événements (1/2)

Univers des possibles

Ensemble (fini, infini dénombrable, infini non dénombrable) de tous les résultats possibles d'une expérience aléatoire (généralement noté Ω)

Ex.	Expérience aléatoire	Ω
1	Jet d'un dé	$\{1, 2, 3, 4, 5, 6\}$
2	Comptage du nombre de chromosomes dicentriques dans 500 cellules irradiées à 1 Gray	$\mathbb{N}$
3	Mesure du diamètre d'une pastèque	$\mathbb{R}^+$
4	Temps de trajet maison-travail	$\mathbb{R}^+$

Notions d'expérience aléatoire et d'événements (2/2)

Evénément aléatoire

Assertion ou proposition logique relative au résultat d'une expérience aléatoire. On lui associe tous les résultats (sous-ensemble de Ω) de l'épreuve pour lesquels il est réalisé :

- Evénement certain : Ω / Evénement impossible : $\emptyset$
- Evénement simple (ou élémentaire)

Ex.	Un événement aléatoire	Un événement impossible
1	Obtenir 2	Obtenir 13
2	Aucun chromosome dicentrique	Nombre négatif de dicentriques
3	Diamètre supérieur à 40 cm	Diamètre négatif
4	Temps de trajet compris entre 5 et 10 minutes	Temps de trajet nul

$\Rightarrow$ Un événement aléatoire est un **événement incertain**

Evénement incertain : besoin d'une définition plus générale ?

Mais :

- Notre incertitude ne peut pas toujours être associée à un aléa
- Un événement incertain ne peut pas toujours être associé à une expérience aléatoire !

Vrai ou Faux ?

- E : "Le Portugal a une superficie inférieure ou égale à 80000 km^2"
- F : "La 5ème décimale du nombre Π est 9"
- G : "Il y a eu entre 300 et 400 décès par leucémie dans la cohorte actuelle des survivants des bombardements d'Hiroshima et Nagasaki"

IRSN INSTITUT DE RADIOPROTECTION ET DE SÛRETÉ NUCLÉAIRE

Evénement incertain : définition plus générale

Lindley, D. (2006) *Understanding uncertainty* John Wiley & Sons, INC, New-York

- *"An event is a statement whose truth is contemplated by a person."*
- Toute assertion ou proposition logique dont **la réalisation ou l'exactitude pose question** pour un individu donné
- Un événement est incertain pour un individu s'il ne sait pas s'il est vrai ou faux
- Deux éventualités forment l'univers Ω des possibles
    - Exemple F : $\Omega = \{$"La 5ème décimale du nombre Π est 9", "La 5ème décimale du nombre Π n'est pas 9"$\}$

IRSN
INSTITUT DE RADIOPROTECTION ET DE SÛRETÉ NUCLÉAIRE

Comment décrire et quantifier notre incertitude ?

Jeter un dé

- Evénement incertain : "Obtenir un nombre pair" $\rightarrow A = \{2, 4, 6\}$

A tout événement incertain, on souhaite attribuer un nombre, plus ou moins grand, permettant de quantifier un niveau de vraisemblance ou un degré de confiance en sa réalisation

$\Rightarrow$ La probabilité est **un** outil mathématique possible, basée sur des propriétés de cohérence et de rationalité

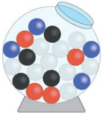
1. Introduction
2. L'incertitude
3. **Qu'est-ce qu'une probabilité ?**
4. Probabilité conditionnelle et indépendance stochastique
5. Variable aléatoire et loi de probabilité : le cas univarié
6. Variable aléatoire et loi de probabilité : le cas bivarié
7. Simulations
8. Quelques remarques importantes pour finir...

La probabilité par l'exemple

Quel niveau de vraisemblance accorderiez-vous aux événements incertains ci-dessous ?

Jet d'un dé non pipé

- A= "Obtenir un nombre pair"
- $\bar{A}$ = "Obtenir un nombre impair"
- B= "Obtenir un nombre inférieur à 8"
- C= "Obtenir 0"



Tirage dans une urne

Soit une urne contenant 10 **boules rouges à pois noirs**, 30 **boules rouges unies**, 30 **boules blanches à pois noirs** et 30 **boules blanches unies**. On tire au hasard une boule dans l'urne.

- D= "La boule tirée est blanche"
- E = "La boule tirée est rouge"
- F= "La boule tirée est blanche ou rouge"
- G= "La boule tirée est rouge ou à pois noirs"

Définition axiomatique de Kolmogorov

Soit un univers Ω définissant un ensemble fini d'événements élémentaires possibles , notés $\{\omega_1, \dots, \omega_n\}$, et $\mathcal{C}$ la tribu associée (cf. chapitre 1 page 12).

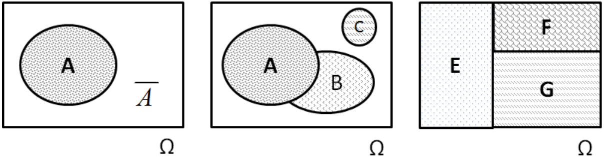
Axiomatique

Une probabilité $\mathbb{P}$ sur $(\Omega, \mathcal{C})$ est **une fonction** telle que :

- **Axiome 1** : A tout événement A de $\mathcal{C}$ est assigné un nombre réel $\mathbb{P}(A)$ tel que $0 \leq \mathbb{P}(A) \leq 1$;
- **Axiome 2** : La probabilité de l'événement certain Ω est donnée par : $\mathbb{P}(\Omega) = \sum_{i=1}^{n} \mathbb{P}(\omega_i) = 1$;
- Axiome 3 (dit règle d'additivité) : toute suite d'événements $A_1, A_2, \dots, A_n$ deux à deux **disjoints** satisfait :

$$\mathbb{P}(A_1 \cup A_2 \cup \ldots \cup A_n) = \Sigma_{i=1}^{n} \mathbb{P}(A_i). \tag{1}$$

Quelques règles importantes



1. $\mathbb{P}(\overline{A}) = 1 - \mathbb{P}(A)$ (En particulier, $\mathbb{P}(\emptyset) = 0$) ;
2. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$;
3. Si $A \subset B$, alors : $\mathbb{P}(A) \leq \mathbb{P}(B)$
4. $\mathbb{P}(A \cap C) = 0$ (événement impossible)
5. $\{E,\ F,\ G\}$= partition de $\Omega \Rightarrow \mathbb{P}(E \cup F \cup G) = \mathbb{P}(E) + \mathbb{P}(F) + \mathbb{P}(G) = \mathbb{P}(\Omega) = 1$.

On ne peut additionner des probabilités pour évaluer la probabilité de survenue d'un événément ou d'un autre que si les événements sont disjoints !

Définitions opérationnelles de la probabilité : la vision "objective"

Etant donné un jeu d'hypothèses fixées par le modélisateur, la probabilité d'un événement est déterminée **de manière unique**

La vision classique (héritée des jeux de hasard)

- Hypothèse : Ω **est fini** + tous les résultats élémentaires sont équiprobables
- Le calcul des probabilités n'est qu'une affaire de dénombrement !
- Probabilité d'un événement A : $P(A) = \frac{\text{Nombre de résultats favorables}}{\text{Nombre de résultats possibles}}$

La vision fréquentiste

- Ω **est fini, infini dénombrable ou continu**
- Idée : Répéter une expérience aléatoire à l'infini !
- Probabilité $\mathbb{P}(A)$ d'un événement A : **limite de la fréquence d'occurrence** de A quand le nombre d'expériences n est très grand
- Exemple (Jet de dé ; A="Obtenir un nombre pair") :
  3 5 6 2 3 4 3 2 2 6 5 2 6 3 1 2 1 3 3 6 (20 tirages successifs)

$$P_{20}(A) = \frac{5+1+4}{20} = 0.5$$

Définitions opérationnelles de la probabilité : la vision "subjective"

Quel degré de confiance - entre 0 et 1- accorderiez-vous à l'événement :

- E : "Le Portugal a une superficie inférieure ou égale à 80000 km^2" ?

Evénement incertain non répétable ! ⇒ Deuxième conception opérationnelle

La probabilité "subjective"

- Permet de quantifier **le degré de confiance** d'une personne, basée sur ses connaissances et son opinion, en la réalisation de n'importe quel événement incertain ⇒ **Pas d'unicité !**
- Obéit aux **axiomes de Kolmogorov** : rationalité !
- Peut **varier** d'une personne à l'autre et dans le temps en fonction des connaissances
- Liée à la notion de **pari** : Combien un individu rationnel est prêt à parier pour qu'un événement incertain d'intérêt se réalise ?
- **Concept clé de la statistique bayésienne** : on peut probabiliser des événements non répétables et tout ce qui est incertain !

La probabilité conditionnelle par l'exemple

- Evénement A : "Ce mineur d'uranium est décédé d'un cancer du poumon"
- Q1 : Quelle est la probabilité de l'événement incertain A ?

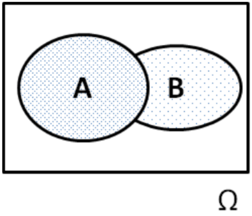
- Evénement B : "Ce mineur d'uranium était fumeur"
- Q2 : Quelle est la probabilité de l'événement incertain A (sachant B) ?

L'information complémentaire apportée par l'événement B a sans doute permis d'augmenter le degré de vraisemblance de l'événement A !

$\Rightarrow$ Concept de probabilité conditionnelle

IRSN INSTITUT DE RADIOPROTECTION ET DE SÛRETÉ NUCLÉAIRE

Qu'est-ce qu'une probabilité conditionnelle ?



Définition

Soit B un événement de probabilité non nulle. On appelle **probabilité conditionnelle** de l'événement A sachant B, notée $\mathbb{P}(A|B)$, le rapport suivant :

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \tag{2}$$

Remarque : Univers des possibles restreint à B !

Exemple : Cancer du poumon/Fumeur

Table de contingence : table des probabilités jointes (et marginales)

	Cancer du poumon	Pas cancer du poumon	Total_2
Fumeur	**0.012**	0.238	**0.250**
Non Fumeur	0.007	0.743	0.750
Total_1	**0.019**	0.981	1

- Probabilité de l'événement A :"Cancer du poumon" : $P(A) = 0.019$
- Probabilité de l'événement B :"Fumeur" : $P(B) = 0.250$
- Calculer $\mathbb{P}(A|B)$

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{0.012}{0.250} = 0.048$$

IRSN INSTITUT DE RADIOPROTECTION ET DE SÛRETÉ NUCLÉAIRE

La clé de voûte de la statistique bayésienne : la formule de Bayes

Soient deux événements A et B tels que $\mathbb{P}(A) > 0$. $\mathbb{P}(A|B)$ et $\mathbb{P}(B|A)$ sont reliées par :

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

Interprétation bayésienne : *Formule d'inversion des causes*

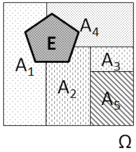
- Mécanisme d'inversion de relations de conditionnement probabiliste
- Evaluer la probabilité de $B|A$ avec B **la cause** et A **l'effet** quand on connaît la probabilité de $A|B$ et les probabilités marginales de A et B.

Exemple : Cancer du poumon/Fumeur

Calculer la probabilité de l'événement B (Cause : "Fumeur") sachant l'événement A (Effet : "Cancer du poumon")

$$\mathbb{P}(B|A) = \frac{0.048 \times 0.250}{0.019} = \frac{0.012}{0.019} = 0.63$$

La formule des probabilités totales



Soient A_1, A_2,..., A_n une partition de l'univers des possibles Ω. Alors, pour tout événement incertain $E \in \Omega$:

$$\mathbb{P}(E) = \sum_{i=1}^{n} \mathbb{P}(E \cap A_i) = \sum_{i=1}^{n} \mathbb{P}(E|A_i)\mathbb{P}(A_i)$$

$\Rightarrow$ Autre expression de la formule de Bayes

$$\mathbb{P}(A_j|E) = \frac{\mathbb{P}(E|A_j)\mathbb{P}(A_j)}{\sum_{i=1}^{n} \mathbb{P}(E|A_i)\mathbb{P}(A_i)}$$

Exemple : Cancer du poumon/Fumeur

- {"Fumeur","Non fumeur"} = Partition de Ω

$$\mathbb{P}(A) = \mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|C)\mathbb{P}(C) = 0.048 \times 0.250 + 0.009 \times 0.750 = 0.019$$

INSTITUT DE RADIOPROTECTION ET DE SÛRETÉ NUCLÉAIRE

Notion d'indépendence stochastique

Definition

Deux événements A et B sont indépendants si $\mathbb{P}(A|B) = \mathbb{P}(A)$

- Le fait de savoir que B est réalisé n'apporte aucune information sur A

Propriétés (Généralisables à n événements !)

- A indépendant de B $\iff$ B indépendant de A
- A et B indépendants $\iff$ $\mathbb{P}(A \cap B) = \mathbb{P}(A) \times \mathbb{P}(B)$

- Le modélisateur **suppose** souvent l'indépendance (conditionnelle) des événéments qu'il cherche à décrire ou prédire...

Pour quantifier la probabilité conjointe de deux (ou plus de deux) événements, on ne peut multiplier leur probabilité respective qu'après s'être assuré de l'indépendance de ces événements !

IRSN INSTITUT DE RADIOPROTECTION ET DE SÛRETÉ NUCLÉAIRE

La formule des probabilités composées

Que faire quand on souhaite quantifier la probabilité conjointe de deux (ou plus de deux) événements non indépendants stochastiquement ?

Formule des probabilités composées (Pour 2 événements)

Soient A_1 et A_2 deux événements tels que $P(A_1 \cap A_2) \neq 0$. Alors :

$$\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1)\mathbb{P}(A_2|A_1)$$

Formule généralisable à n événements de probabilité jointe non nulle.

Exemple : Cancer du poumon/Fumeur

$$\begin{aligned}
&\mathbb{P}(\text{"Fumeur"} \cap \text{"Cancer poumon"})\\
&= \mathbb{P}(\text{"Fumeur"}|\text{"Cancer poumon"})\mathbb{P}(\text{"Cancer du poumon"})\\
&= 0.63 \times 0.019 = 0.012
\end{aligned}$$

IRSN
INSTITUT DE RADIOPROTECTION ET DE SÛRETÉ NUCLÉAIRE

Attention : Dépendance stochastique ne veut pas dire causalité !

Exemple

Dépendance entre la pointure des enfants et leur niveau de langage mais pas de causalité ! Une troisième variable sous-jacente (appelée facteur de confusion) explique cette dépendance probabiliste : l'âge

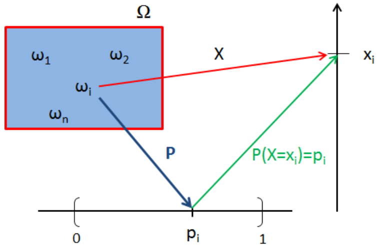
IRSN INSTITUT DE RADIOPROTECTION ET DE SÛRETÉ NUCLÉAIRE

Qu'est-ce qu'une variable aléatoire (v.a) ?

- Grandeur mathématique prenant différents états ou valeurs de **manière imprévisible** mais avec une certaine **régularité d'occurrence**
- La régularité en question est décrite à travers une **loi (ou distribution) de probabilité**.
- Ingrédient de base de tout modèle probabiliste
- **Application** assignant à chaque résultat élémentaire d'une expérience aléatoire une valeur, un état ou un intervalle de valeurs avec une certaine probabilité

x_i est la réalisation de la v.a. **X** pour l'événement élémentaire ω_i (seulement) :

$$\mathbb{P}(X = x_i) = \mathbb{P}(\{\omega \in \Omega | X(\omega) = x_i) = p_i$$



Variables aléatoires discrètes

- Variables prenant un **nombre fini ou infini mais dénombrable** de valeurs (v.a. discrètes valuées) ou d'états (v.a. discrètes catégorielles) possibles

Un joueur lance un dé non pipé

- X=1 si la face qui s'affiche est "un nombre pair"
- X= 0 si la face qui s'affiche est "un nombre impair"

Loi de probabilité d'une v.a. discrète

Fonction associant à toute valeur/état possible x_i (i=1,. . .,n) d'une v.a. X sa probabilité

$$p_i = P(X = x_i) \in [0, 1]$$

telle que $\sum_{i=1}^{n} p_i = 1$

X= v.a. binaire (2 états possibles $\{0, 1\}$)

- P(X=1)= 0.5
- P(X= 0)= 0.5

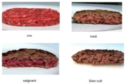
Exemple : Variables aléatoires catégorielles ordonnées

Variables discrètes prenant un **nombre fini d'états possibles naturellement ordonnés** (mais non associés à un nombre)

> Le choix d'une cuisson de viande rouge : cru/saignant/rosé/bien cuit

v.a. : Choix de cuisson du steak haché dans la population française

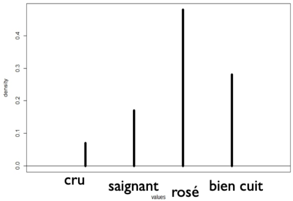
- Cru : 0.05
- Saignant : 0.10
- Rosé : 0.37
- Bien cuit : 0.28
- Ne mange pas de steak haché: 0.20

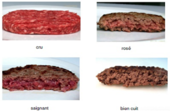


(1) recenser les états élémentaires exclusifs possibles

(2) affecter une probabilité à chaque état (valeur comprise entre 0 et 1)
(3) S'assurer que la somme totale des probabilités affectées soit égale à 1!

Loi de probabilité de la v.a. choix d' une cuisson de viande rouge





Notions d'espérance et de variance : cas d'une v.a. discrète valuée

- Espérance mathématique de X : $\mathbb{E}(X) = \mu = \sum_{i=1}^{n} x_i p_i$
- Variance de X : $Var(X) = \sum_{i=1}^{n} (x_i - \mu)^2 p_i$
- Ecart-type de X : $\sigma = \sqrt{Var(X)}$

Propriétés utiles

- $Var(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$
- $\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$ (a et b constantes)
- $Var(aX + b) = a^2 Var(X)$ (a et b constantes)

IRSN
INSTITUT DE RADIOPROTECTION ET DE SÛRETÉ NUCLÉAIRE

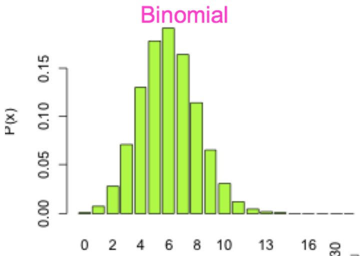
Quelques lois de probabilité discrètes(1/2)

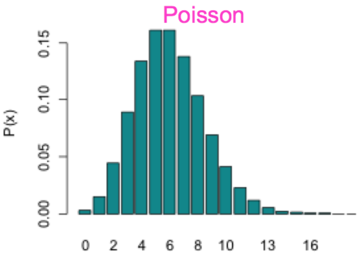
	Support	PMF$^{(a)}$	Param	E$^{(b)}$	Var$^{(c)}$
Bernoulli	$\{0, 1\}$	$P(X = 1) = p$	$p \in [0, 1]$	p	p(1-p)
Binomial	$k \in \{0, ..., n\}$	$\frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$	n $\in N$, p $\in [0, 1]$	np	np(1-p)
Poisson	$k \in N$	$e^{-\lambda} \frac{\lambda^k}{k!}$	$\lambda \in]0, +\infty[$	λ	λ
Geometric	$k \in N^*$	$(1-p)^{k-1} p$	$p \in [0, 1]$	$\frac{1}{p}$	$\frac{1-p}{p^2}$

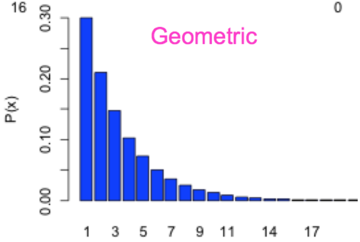
- (a) -> Loi de probabilité (PMF) ; (b) -> Espérance ; (c) -> Variance

Quel contexte d'utilisation ?

Quelques lois de probabilité discrètes (2/2)







Variables aléatoires continues

Variables prenant une **infinité non dénombrable de valeurs possibles** (e.g., toutes les valeurs dans un intervalle borné ou non de $\mathbb{R}$)

- Le log10 de la concentration initiale en *Escherichia Coli* dans un steak surgelé
- La température minimale de croissance d'une bactérie
- La dose cumulée annuelle de gaz radon inhalée par un mineur d'uranium dans le cadre de son activité professionnelle
- La taille d'un individu

Comment leur assigner une loi de probabilité ? (1/3)

Une montre tombe en panne. La position exacte de la grande aiguille au moment de l'arrêt est une variable aléatoire X continue. Les positions possibles sont l'ensemble des angles entre 0° et 360°. Il y a un nombre infini non dénombrable de possibilités et on veut leur attribuer une probabilité.

Comment spécifier la loi de probabilité de X?

Une solution possible: Au lieu d'attribuer à chaque valeur possible une probabilité, une probabilité est associée à **chaque intervalle de valeurs**!

Exemple: Pour tous réels a et b compris entre 0 et 360

$$P(a \leq X \leq b) = \frac{b-a}{360}$$

IRSN INSTITUT DE RADIOPROTECTION ET DE SÛRETÉ NUCLÉAIRE

Comment leur assigner une loi de probabilité ? (2/3)

☞ **Via la spécification de la fonction de répartition**

v.a. X: Position exacte de la grande aiguille d'une montre à l'arrêt

$$\boxed{F(x) = P(X \leq x) = \frac{x}{360}}$$

*(1) Associer une **probabilité F(x)** d'être inférieur ou égal à x à chaque réel x*

⟺ *Spécifier la fonction de répartition de X :*

*(fonction **croissante continue** à valeurs dans [0,1] tel que la limite en -∞ est 0 et en +∞ est 1)*

Comment leur assigner une loi de probabilité ? (3/3)

☞ **Via la spécification de la densité de probabilité**

v.a. X : Position exacte de la grande aiguille d'une montre à l'arrêt

Entre 0 et 360 degrés	$f(x) = \dfrac{1}{360}$

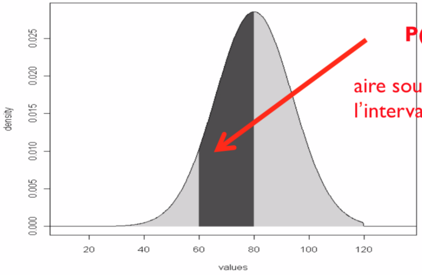
(1) définir un support de valeurs possibles

(2) Associer une ***densité de probabilité*** *f(x) à chaque point x du support.*
*(**fonction positive** dont l 'aire sous la courbe est égale à 1)*

☞ Densité de probabilité et fonction de répartition sont deux fonctions permettant de **caractériser** la loi de probabilité d' une v.a. continue X.

Lien entre fonction de répartition et densité de probabilité

Densité de probabilité f de la v.a. X
Rendement d'une culture de blé en plein champ



P(60≤X≤80)
=
aire sous la courbe f sur l'intervalle *[60, 80]*

$$P(a \leq X \leq b) = F(b) - F(a) = \int_a^b f(x)dx$$ ≃ somme de *a* à *b* de *f(x)*

Densité de probabilité f(x) au point x = Dérivée de la fonction de répartition F au point x.

Notions d'espérance et de variance : cas d'une v.a. continue

Espérance mathématique: Soit X une v.a. continue prenant ses valeurs sur un intervalle réel D, son espérance (si elle existe) est le nombre:

$$E(X) = \mu = \int_D x.f(x)dx$$

Somme → Intégrale

Coefficients de pondération

Variance et écart-type: Soit X une v.a. continue prenant ses valeurs sur un intervalle réel D, sa variance (si elle existe) est le nombre:

$$Var(X) = \sigma^2 = \int_D (x-\mu)^2 .f(x)dx$$

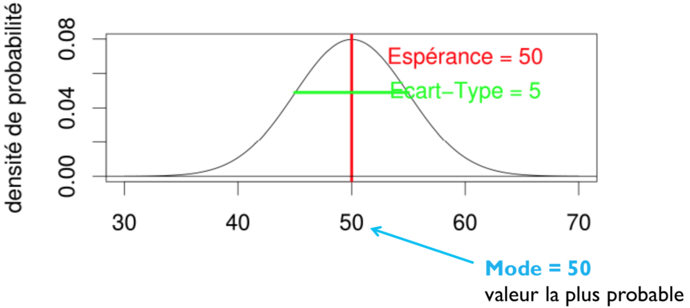
$$\sigma = \sqrt{Var(X)}$$

Quelques propriétés utiles:

- $Var(X)=E(X^2)-E(X)^2$
- $E(aX+b)=aE(X)+b$ (a et b constantes)
- $Var(aX+b)=a^2Var(X)$ (a et b constantes)

Caractéristiques d'une densité de probabilité (1/3)

Cas d'une densité de probabilité symétrique

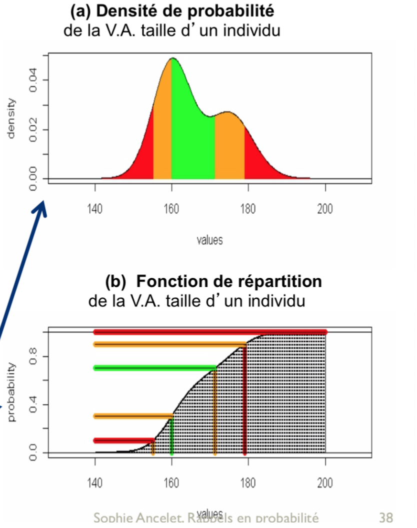


Caractéristiques d'une densité de probabilité (2/3)

Le α quantile (on parle aussi de percentile) d'une loi de probabilité est la valeur q_α telle que la probabilité qu'une v.a. X suivant cette distribution lui soit inférieure ou égale est égale à α.

$$P(X \leq q_\alpha) = \alpha$$

Représentation graphique des quantiles / percentiles à **10-30-70-90 %** à partir d'une densité de probabilité (a) et d'une fonction de répartition (b)

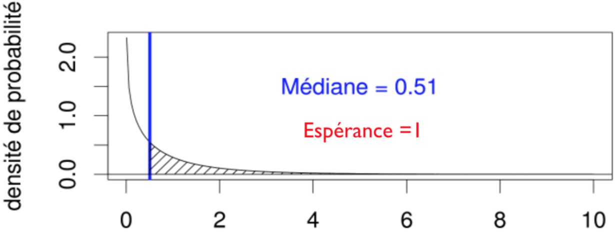


Caractéristiques d'une densité de probabilité (3/3)

- **Médiane=Quantile 50%** $P(X \leq q_{0.5}) = P(X \geq q_{0.5}) = 0.5$

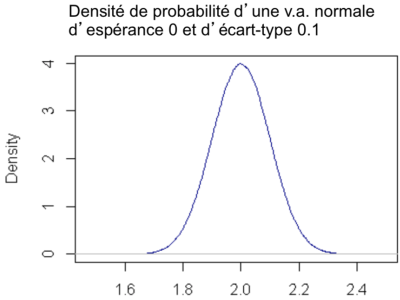
Ne pas confondre espérance et médiane!

Exemple: Cas dissymétrique vers la droite: espérance >médiane



Deux écueils à éviter

- Dans le cas d'une v.a. continue X:
  - On ne sait pas attribuer une probabilité à une valeur x donnée
  - La densité de probabilité en un point n'est pas nécessairement comprise entre 0 et 1 (c'est une dérivée!)



Quelques densités de probabilité continues (1/3)

	Support	PDF[a]	Param	E[b]	Var[c]
Normal	$x \in R$	$\frac{1}{\sigma\sqrt{2\pi}} exp^{-\frac{(x-\mu)^2}{2\sigma^2}}$	$\mu \in R,\ \sigma > 0$	μ	σ^2
Beta	$x \in [0,1]$	$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$	$\alpha > 0,\ \beta > 0$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
Expo*	$x \in R^+$	$\lambda exp^{-\lambda x}$	$\lambda > 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gamma	$x \in R^+$	$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} exp^{-\beta x}$	$\alpha > 0,\ \beta > 0$	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$
LogN**	$x \in R^+$	$\frac{1}{x\sigma\sqrt{2\pi}} exp^{-\frac{(ln(x)-\mu)^2}{2\sigma^2}}$	$\mu \in R,\ \sigma > 0$	$e^{\mu+\frac{\sigma^2}{2}}$	$(e^{\sigma^2} - 1)e^{(2\mu+\sigma^2)}$

Remarques :

- (a) -> Densité de probabilité ; (b) -> Espérance ; (c) -> Variance
- $*$ -> Loi exponentielle ; $**$ -> Loi log-normale

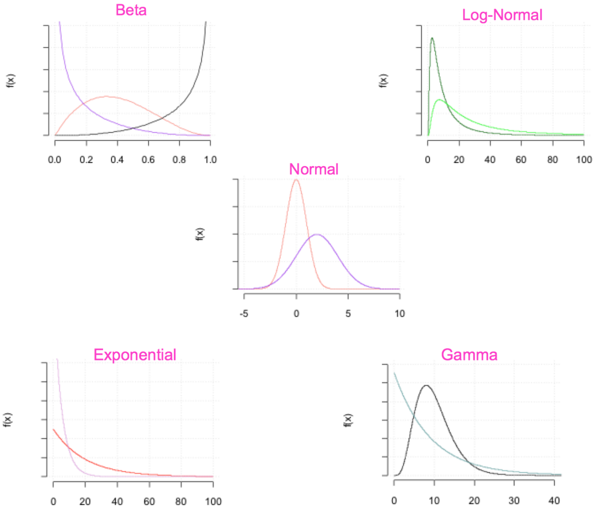
Quel contexte d'utilisation ?

IRSN INSTITUT DE RADIOPROTECTION ET DE SÛRETÉ NUCLÉAIRE

Quelques densités de probabilité continues (2/3)

- Si $X \sim LogN(\mu, \sigma^2)$ alors $ln(X) \sim N(\mu, \sigma^2)$
  - Moyenne dite géométrique de la loi lognormale : e^{μ}
  - Ecart-type dit géométrique de la loi lognormale : e^{σ}
- Si $X \sim Beta(a, b)$
  - a peut s'interpréter comme un nombre de succès, b comme un nombre d'échecs et $a + b$ comme une taille virtuelle d'échantillon
    - Choisir a=10 et b=5 fournit un état de connaissance a priori équivalent à un échantillon virtuel de taille 15 comprenant a=10 succès
  - $Y = y_{min} + (y_{max} - y_{min}) \times X$ suit une loi beta généralisée de paramètres a et b et de support donné par l'intervalle réel $[y_{min}, y_{max}]$

Quelques densités de probabilité continues (3/3)



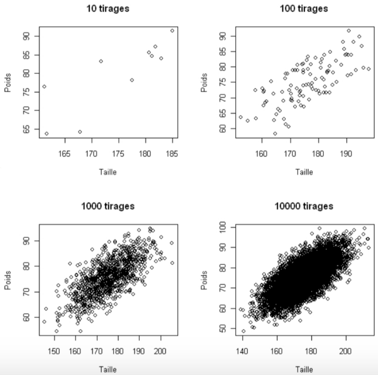
Oui, mais...

En pratique, le statisticien-modélisateur travaille avec de multiples variables aléatoires.... Comment faire ?

Variables aléatoires vectorielles

- Une v.a. peut être **vectorielle** :
  - masse et taille d'un individu
  - masse et âge d'un individu
  - sexe, IMC, âge d'un individu
  - concentrations de différentes espèces de bactéries

Nuage de points (simulés) pour le couple de v.a. masse vs taille d'un individu



Cas bivarié discret : Loi de probabilité jointe (1/2)

Exemple : Cancer du poumon/Fumeur

Soit (X,Y) la paire de variables aléatoires ("Statut de la maladie", "Statut fumeur")

	Cancer du poumon	Pas de cancer du poumon
Fumeur	0.012	0.238
Non Fumeur	0.007	0.743

Comment définir la loi de probabilité jointe du couple (X,Y) ?

1. Recenser tous les couples d'états élémentaires exclusifs possibles
2. Affecter une probabilité à chaque couple d'états (valeur comprise entre 0 et 1)
3. S'assurer que la somme totale des probabilités affectées soit égale à 1 !

IRSN
INSTITUT
DE RADIOPROTECTION
ET DE SÛRETÉ NUCLÉAIRE

Cas bivarié discret : Loi de probabilité jointe (2/2)

Soit (X,Y) une paire de variables aléatoires discrètes

Loi de probabilité jointe

Fonction associant à chaque couple de valeurs possibles (x_i, y_j) $(i = 1, \ldots, I, j = 1, \ldots, J)$ de (X,Y) sa probabilité :

$$p_{ij} = P(X = x_i \cap Y = y_j)$$

tel que $\sum_{i=1}^{I} \sum_{j=1}^{J} p_{ij} = 1$

Cas bivarié discret : : Lois de probabilité marginales (1/2)

Exemple : Cancer du poumon/Fumeur

Soit (X,Y) la paire de variables aléatoires ("Statut de la maladie", "Statut fumeur")
Quelles sont les lois de probabilités marginales des v.a. X et Y ?

	Cancer poumon	Pas cancer poumon	Marginale Y
Fumeur	0.012	0.238	0.25
Non fumeur	0.007	0.743	0.75
Marginale X	0.019	0.981	1

Soit (X,Y) une paire de variables aléatoires discrètes

Loi de probabilité marginale de X

Fonction associant à tous les états/valeurs possibles x_i de X sa probabilité :
$P(X = x_i) = \sum_{j=1}^{J} P(X = x_i \cap Y = y_j)$ (Loi des probabilités totales !)

Même principe pour la loi marginale de Y

IRSN INSTITUT DE RADIOPROTECTION ET DE SÛRETÉ NUCLÉAIRE

Cas bivarié discret : : Lois de probabilité marginales (2/2)

Attention ! Les lois marginales de X et Y ne suffisent pas à définir la loi jointe du couple (X,Y)

	Cancer poumon	Pas cancer poumon	**Marginale Y**
Fumeur	0.012 -0.01	0.238 +0.01	0.25
Non fumeur	0.007 +0.01	0.743 -0.01	0.75
Marginale X	0.019	0.981	1

... Sauf si on ajoute l'hypothèse d'indépendance : $p_{ij} = p_i p_j$!

IRSN INSTITUT DE RADIOPROTECTION ET DE SÛRETÉ NUCLÉAIRE

Cas bivarié discret : : Lois de probabilité conditionnelles (2/2)

Exemple : Cancer du poumon/Fumeur

Soit (X,Y) la paire de variables aléatoires ("Statut de la maladie", "Statut fumeur")
Quelle est la loi de probabilité conditionnelle de Y sachant X ?

	Cancer du poumon	Pas cancer du poumon
Fumeur	0.63	0.24
Non fumeur	0.37	0.76
Total	1.0	1.0

Soit (X,Y) une paire de variables aléatoires discrètes

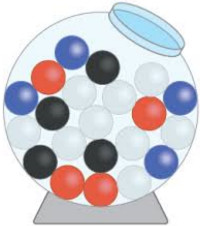
Loi conditionnelle de X sachant $Y = y_j$

Fonction associant à tous les états/valeurs possibles x_i de X sa probabilité :
$P(X = x_i | Y = y_j) = \frac{P(X=x_i \cap Y=y_j)}{P(Y=y_j)}$

Même principe pour la loi conditionnelle de Y sachant $X = x_i$

La loi multinomiale - une loi de probabilité discrète multivariée

Contexte d'utilisation : n expériences aléatoires indépendantes, m résultats possibles de probabilité p_i ($i \in 1, ..., m$).



Support	$n_i \in \{0, ..., n\}$ with $\sum_{i=1}^{m} n_i = n$
PMF[(a)]	$\frac{n!}{n_1! \ldots n_m!} p_1^{n_1} \ldots p_m^{n_m}$
Param	$n \in N^*$, $p_1, ..., p_m$ with $\sum_{i=1}^{m} p_i = 1$
$E(N_i)^{(b)}$	np_i
$Var(N_i)^{(c)}$	$np_i(1 - p_i)$
$Cov(N_i, N_j)^{(d)}$	$-np_i p_j$ if $i \neq j$

- (a) -> Loi de probabilité jointe
- (b) -> Espérance ; (c) -> Variance
- (d) -> Covariance

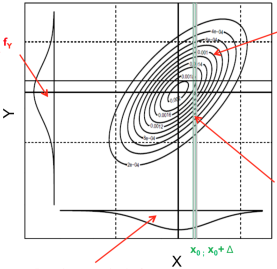
IRSN INSTITUT DE RADIOPROTECTION ET DE SÛRETÉ NUCLÉAIRE

Qu'en est-il dans le cas continu ? (1/2)

Illustration par l'exemple…

Soit un couple de v.a. continues (X,Y)={ Masse ,Taille } d'un individu

Densité marginale f_Y de la v.a. Y



Densité marginale f_X de la v.a. X

Densité jointe f_{XY} du couple (X,Y):

(1) Définir un support de valeurs possibles $D_X \times D_Y$
(2) Associer une densité de probabilité jointe $f_{XY}(x,y) \geq 0$ à chaque couple (x,y) du support
(3) S'assurer que le volume sous la surface représentative de f_{XY} **est égal à 1**

Densité conditionnelle de Y sachant X=x_0

(1) Zoomer sur la valeur X=x_0 (i.e. Δ tend vers 0)

Qu'en est-il dans le cas continu ? (2/2)

Plus formellement...

<u>Fonction de répartition du couple (X,Y):</u> Fonction associant **à toutes les valeurs possibles (x , y)** la probabilité:

$$F_{XY}(x,y) = P(X \leq x \cap Y \leq y)$$

<u>Densité jointe du couple (X,Y) (si elle existe) :</u> $f_{XY}(x,y) = \dfrac{\partial^2 F_{XY}}{\partial x \partial y}$

<u>Densité marginale de X:</u> $f_X(x) = \int f_{XY}(x,y)dy$

<u>Densité conditionnelle de Y sachant X=x_0:</u>

$$g(y|x_0) = \frac{f_{XY}(x_0, y)}{f_X(x_0)}$$

$\Rightarrow$ **Plus il y a de v.a., plus les intégrales à calculer sont complexes...**

La loi normale multivariée

Support	$x \in R^n$
PDF$^{(a)}$	$\frac{1}{(2\pi)^{n/2} det(\Sigma)^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$
Param	$\mu = (\mu_1, ..., \mu_n) \in R^n$, Σ
$E^{(b)}$	μ
$Var^{(c)}$	Σ

- (a) -> Densité de probabilité jointe
- (b) -> Espérance ; (c) -> Variance
- Σ : $n \times n$ matrice de variance-covariance, semi-définie positive

La loi de Dirichlet - une loi multivariée continue

Support	$p_k \in [0,1]^K$ with $\sum_{k=1}^{K} p_k = 1$
PDF$^{(a)}$	$\frac{\Gamma(\alpha_+)\prod_{k=1}^{K} p_k^{\alpha_k - 1}}{\prod_{k=1}^{K}\Gamma(\alpha_k)}$ with $\alpha_+ = \sum_{k=1}^{K}\alpha_k$
Param	α=$(\alpha_1, ..., \alpha_K)$ with $\alpha_k > 0$ for all k
$E^{(b)}(p_k)$	$\frac{\alpha_k}{\alpha_+}$
$Var^{(c)}(p_k)$	$\frac{\alpha_k(\alpha_+ - \alpha_k)}{\alpha_+^2(\alpha_+ + 1)}$

- (a) -> Densité de probabilité jointe
- (b) -> Espérance ; (c) -> Variance

- If $(p_1, ..., p_K) \sim Dirichlet(\alpha_1, ..., \alpha_K)$ then $p_k \sim Beta(\alpha_k, \alpha_+ - \alpha_k)$

Autres lois multivariées : Student multivariée et Wishart

IRSN INSTITUT DE RADIOPROTECTION ET DE SÛRETÉ NUCLÉAIRE

Avant d'assigner une loi de probabilité, à quoi faut il réfléchir ?

- Au type (discrète/continu) de la grandeur d'intérêt
- Au support de valeurs possibles de la grandeur d'intérêt
- Aux valeurs centrales de la grandeur d'intérêt
- Aux valeurs les plus probables de la grandeur d'intérêt
- A la symétrie (ou pas) de la distribution
- A l'épaisseur des queues
- A une transformation préalable de la v.a.

IRSN INSTITUT DE RADIOPROTECTION ET DE SÛRETÉ NUCLÉAIRE

Comment simuler des réalisations d'une variable aléatoire ? (1/2)

- La simulation de v.a. occupe **une place centrale** en **statistique bayésienne**
  - Le plus souvent, on cherche à simuler des réalisations de v.a. selon :
    - la **loi a posteriori** des paramètres d'un modèle probabiliste (cf. Eric et Samuel)
    - la **loi prédictive a posteriori** des grandeurs observables (cf. Eric, Samuel, David)
- Etre capable de simuler des réalisations d'une v.a.

  $\Longleftrightarrow$ Connaître sa loi de probabilité

Comment simuler des réalisations d'une variable aléatoire ? (2/2)

- Générer des nombres aléatoires suppose d'être capable de générer une suite de nombres dont il est très difficile de dériver des **propriétés déterministes**
- $\Rightarrow$ La plupart des logiciels de calculs modernes permettent de **générer des séquences de nombres dites pseudo-aléatoires**
  - Suite de nombres aléatoires "artificiels" obtenus, en pratique, à partir d'algorithmes récursifs ayant des propriétés déterministes

$$x_0(\text{état initial} = \text{graine})$$
$$x_{n+1} = w(x_n)$$

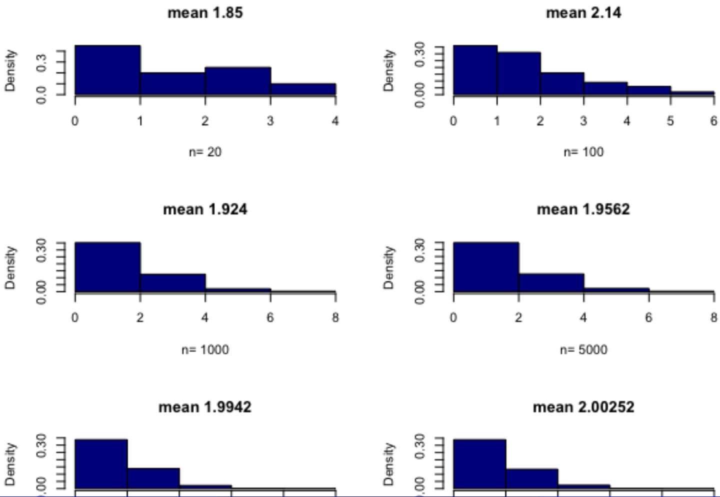
  - La suite générée est périodique : l'objectif est d'obtenir la plus longue période !
  - R, Python, SAS, Excel,...possèdent leur propre générateur de nombres pseudo-aléatoires

Simuler sous R : quelques commandes

Categorical	sample(x, size, replace, prob)
Bernoulli(π)	rbinom(n,size=1,prob=π)
Binomial(N,π)	rbinom(n,size=N,prob=π)
Poisson(λ)	rpois(n,lambda= λ)
Geometric(π)	rgeom(n,prob=π)
Normal(μ,σ)	rnorm(n, mean = μ, sd = σ)
Uniform(a,b)	runif(n,min=a, max=b)
Beta(a,b)	rbeta(n, shape1=a, shape2=b, ncp = 0)
Expo(λ)	rexp (n, rate=λ)
Gamma(a,b)	rgamma(n, shape=a, rate = b, scale = 1/rate)
Multinomial	rmultinom(n, size, prob)

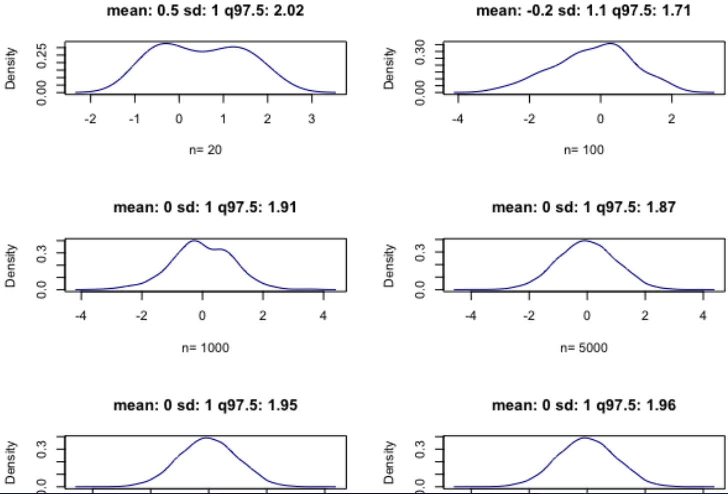
Exemple : La loi de Poisson

Combien de tirages aléatoires indépendants sont nécessaires pour "bien" représenter la loi de Poisson de paramètre $\lambda = 2$?



Exemple : La loi Normale

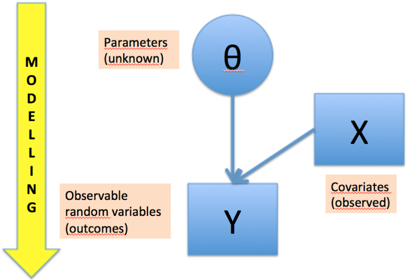
Combien de tirages aléatoires indépendants sont nécessaires pour "bien" représenter la loi Normale $\mathcal{N}(0,1)$?



Qu'est-ce qu'un modèle probabiliste ?

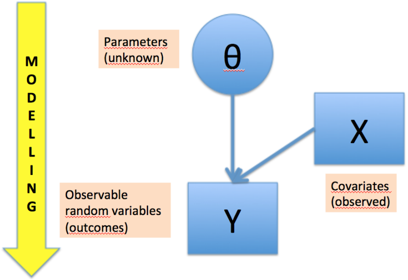
- Représentation mathématiques **approchée mais utile** d'un système aléatoire (biologique, physique,...) étant donné un état de connaissances du (des) modélisateur(s) (cf. cours Eric et David)
- Représentation limitée à un certain contexte mais qui doit être **suffisamment flexible** pour s'adapter à différents scenarii
    - **Modèle paramétrique** : un ensemble fini de paramètres inconnus θ permettent de décrire le système d'intérêt. (Contexte de l'école)

Comment construire un modèle probabiliste ?



- Identifier les variables aléatoires dont on souhaite décrire les fluctuations
- Identifier les quantités fixes inconnues, appelées (paramètres) et les quantités supposées connues (covariables, constantes)
- Specifier les relations de conditionnement entre les variables aléatoires et les paramètres
- Proposer des lois de probabilité pour quantifier ces relations
  - Exemple : **Le modèle d'observations** décrit les **fluctuations des variables aléatoires observables** Y sachant les paramètres θ et les covariables X

Ne pas confondre modélisation probabiliste et inférence statistique !

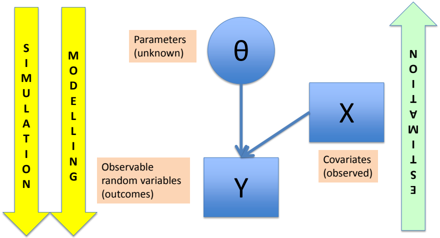


MODELISATION PROBABILISTE N'IMPLIQUE PAS DONNÉES !

Si les paramètres θ sont connus, le modèle probabiliste $\mathcal{P}_\theta$ permet **de simuler des réalisations "plausibles"** $\mathbf{y}_n=(y_1,\ldots,y_n)$ de la v.a Y **(appelée "variable réponse")**.

- **(Forward) Propagation d'incertitude** : Quantifier/décrire l'incertitude sur les variables réponse du modèle en propageant notamment l'incertitude inhérente aux paramètres d'entrée du modèle
- Simulations Monte-Carlo, prédictions, ...

Ne pas confondre modélisation probabiliste et inférence statistique !



INFERENCE STATISTIQUE IMPLIQUE DONNEES

$\Rightarrow$ Des réalisations $\mathbf{y}_n$=(y_1, ..., y_n) de la v.a. Y **sont observées** (échantillon de données).

Principal objectif : *Proposer* des valeurs "possibles" et quantifier l'incertitude d'estimation associée pour les paramètres inconnus θ (les causes) étant donné les données observées (conséquences) y_n=(y_1, ..., y_n) of Y

- **(Backward) Inverse uncertainty quantification** => Démarche inverse par rapport à la modélisation et la simulation
- Estimation

Peut-on réellement définir un modèle physique qui permette de **prédire exactement** quelle face de ce dé non pipé apparaîtra après un lancer, étant donné les conditions expérimentales initiales ?

IRSN INSTITUT DE RADIOPROTECTION ET DE SÛRETÉ NUCLÉAIRE

Peut-on réellement définir un modèle physique qui permette de **prédire exactement** quelle face de ce dé non pipé apparaîtra après un lancer, étant donné les conditions expérimentales initiales ?

Quand le mécanisme physique sous-jacent est trop complexe...

Variabilité naturelle : une première "interprétation" de l'incertitude

- Incertitude inévitable et inhérente à de nombreux systèmes biologiques, physiques, environnementaux,...
- Dûe à des fluctuations incontrôlables (aléa !) du système et de son environnement, en l'état actuel des connaissances
- Aussi appelée "Incertitude par essence", "Incertitude aléatoire" ou "Incertitude stochastique"
- $\Rightarrow$ Incertitude irréductible d'une grandeur (observable ou non)

Exemples

- **Erreur de mesure** (ubiquitaire en science observationnelle comme l'épidémiologie) peut être interprétée comme de la variabilité (due à un manque de précision de l'appareil de mesure, à la variabilité de rigueur des observateurs. . .)
- N'importe quelle **grandeur de terrain ou expérimentale** est généralement sujette à variabilité

Incertitude épistémique : une deuxième "interprétation" de l'incertitude

- Incertitude associée à un manque de connaissances concernant le système d'intérêt (comportement/caractéristiques) et/ou son environnement
  - Aussi appelée incertitude par ignorance
- ⇒ Incertitude réductible par l'apport de connaissances ou de données supplémentaires

Exemples

- Incertitude associée à l'estimation d'un paramètre inconnu
- Incertitude associée aux hypothèses de modélisation (Exemple : Forme des relations dose-réponse en épidémiologie)

IRSN
INSTITUT DE RADIOPROTECTION ET DE SÛRETÉ NUCLÉAIRE

Merci pour votre attention !