

Chapitre 2 – La théorie des probabilités (1^{re} partie)

STT-1920 Méthodes statistiques

Département de mathématiques et de statistique



UNIVERSITÉ
LAVAL

Probabilités et statistique

Où en sommes-nous ?

Au chapitre 1, on a décrit des données : histogramme, moyenne $\bar{x}$, écart-type s , quantiles, diagramme en boîte...

Pour tirer des conclusions sur une population à partir d'un échantillon, il faut un modèle du hasard.

Probabilités

Outil mathématique formel servant à décrire et modéliser les comportements aléatoires.

On connaît les « paramètres » du phénomène et on décrit ce qu'on attend.

Exemple : Une génisse a autant de chances d'avoir une femelle qu'un mâle. Quelle est la probabilité d'avoir au moins une femelle parmi ses trois premiers veaux ?

Probabilités vs statistique

Statistique

Étude quantitative des attributs d'une « population » à partir d'observations.

On ne connaît pas les « paramètres » du phénomène : c'est ce qu'on tente d'estimer.

Exemple : Sur 200 vêlages observés, 112 ont donné une femelle. Peut-on encore croire que les deux sexes sont équiprobables ?

⇒ Le caractère **aléatoire** de l'échantillonnage fait que la statistique repose sur la théorie des probabilités. C'est l'objet de ce chapitre.

2.1 Expérience aléatoire et événements

Expérience aléatoire

Expérience aléatoire $\mathcal{E}$

Expérience dont le résultat dépend du hasard, ne pouvant être prédit avec certitude.

Espace échantillonnal Ω (ou ensemble fondamental)

Ensemble de tous les résultats possibles de l'expérience $\mathcal{E}$. Le nombre d'éléments de Ω est appelé le *cardinal* de Ω .

- Ω **discret** : si Ω est fini ou dénombrable (ex. : la liste des entiers)
- Ω **continu** : si Ω est non dénombrable (ex. : un intervalle de valeurs sur la droite réelle)

Exemple 1 : quel est l'espace échantillonnal ?

- $\mathcal{E}_1$: on note le sexe des trois premiers veaux d'une génisse.
 $\Rightarrow \Omega_1 = \{\text{FFF, FFM, FMF, MFF, FMM, MFM, MMF, MMM}\}$
- $\mathcal{E}_2$: on compte le nombre de lésions foliaires sur une parcelle de gazon de 1 m sur 2 m. $\Rightarrow \Omega_2 = \{0, 1, 2, \dots\}$
- $\mathcal{E}_3$: on mesure les précipitations (en mm) reçues à Québec en juillet.
 $\Rightarrow \Omega_3 = [0, \infty)$

Lesquels de ces espaces sont discrets ? continus ?

Événement

Événement

Sous-ensemble de l'espace échantillonnal Ω . On utilise une lettre majuscule pour l'identifier.

Retour aux expériences de l'exemple 1 :

- $\mathcal{E}_1 : A = \text{« au moins une femelle »} = \{FFF, FFM, FMF, MFF, FMM, MFM, MMF\}$
- $\mathcal{E}_2 : C = \text{« moins de 10 lésions »} = \{0, 1, \dots, 9\}$
- $\mathcal{E}_3 : D = \text{« plus de 150 mm de pluie »} = (150, \infty)$

Exemple 2 : une espèce envahissante

L'hydrocharis grenouillère est une plante aquatique flottante envahissante.

$\mathcal{E}$: à plusieurs points d'un lac,

- on note la présence de l'hydrocharis;
- on note la présence d'un autre macrophyte flottant;
- on mesure la profondeur du lac.

Espace échantillonnal ?

Quelques événements ?

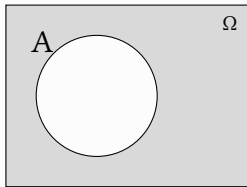


Christian Fischer, CC BY-SA 3.0, [https:](https://commons.wikimedia.org/w/index.php?curid=1042515)

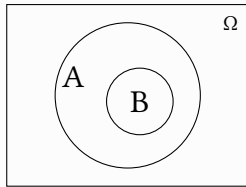
[//commons.wikimedia.org/w/index.php?curid=1042515](https://commons.wikimedia.org/w/index.php?curid=1042515)

Rappel des opérations sur les ensembles

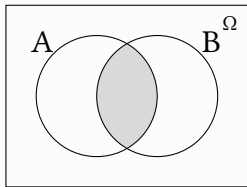
1) Complément de A : A^c



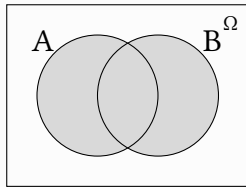
2) Inclusion de B dans A : $B \subset A$



3) Intersection de A et B : $A \cap B$



4) Union de A et B : $A \cup B$



Événements mutuellement exclusifs

Événements mutuellement exclusifs (ou disjoints)

Des événements sont **mutuellement exclusifs** si toutes leurs intersections possibles sont vides, c'est-à-dire lorsqu'**au plus un** de ces événements peut se produire à la fois.

En d'autres termes, A, B et C sont mutuellement exclusifs si

$$A \cap B = \emptyset, \quad A \cap C = \emptyset, \quad B \cap C = \emptyset \quad \text{et} \quad A \cap B \cap C = \emptyset.$$

QUIZ!

Les événements suivants sont-ils mutuellement exclusifs ?

- $\mathcal{E}_1$: sexe des trois premiers veaux.

A = « au moins une femelle », B = « trois mâles » et C = « exactement deux femelles ».

- $\mathcal{E}_3$: précipitations en juillet.

A = « plus de 150 mm », B = « moins de 175 mm » et C = « entre 200 et 300 mm ».

Interprétation d'une probabilité

La probabilité d'un événement A , notée $P(A)$, est un nombre entre 0 et 1 qui mesure ses chances de survenir.

Fréquence relative d'un événement

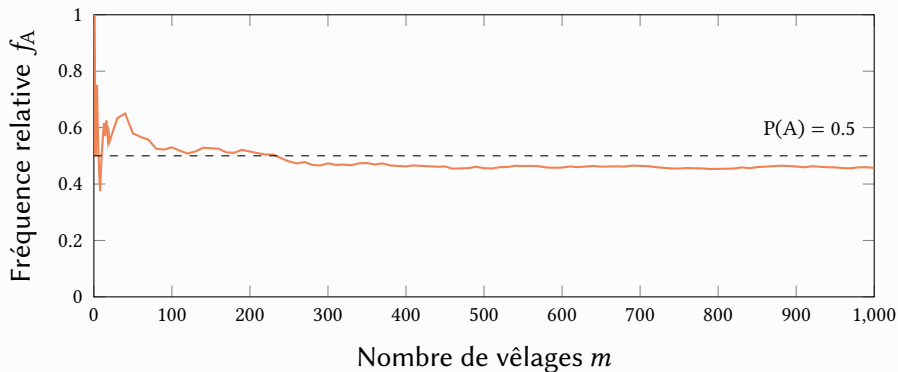
On répète m fois une expérience $\mathcal{E}$ et on note m_A le nombre de fois où l'événement A se produit. Alors $f_A = m_A/m$ est la **fréquence relative** de A .

Intuitivement, $P(A)$ est la fréquence relative de A lorsqu'on répète l'expérience un très grand nombre de fois :

$$P(A) = \lim_{m \rightarrow \infty} \frac{m_A}{m}.$$

La fréquence relative se stabilise

On simule 1000 vêlages et on suit la proportion de femelles obtenues.



2.2 Propriétés des probabilités

Définition. Soit une expérience aléatoire $\mathcal{E}$ et son espace échantillonnal Ω . Si un événement A est défini dans Ω , alors la **probabilité de A** , notée $P(A)$, est un nombre réel respectant les axiomes^a suivants :

1. $0 \leq P(A) \leq 1$
2. $P(\Omega) = 1$
3. Pour toute suite d'événements **mutuellement exclusifs** $A_1, A_2, \dots, A_k$ définis dans Ω , on a

$$P(A_1 \cup A_2 \cup \dots \cup A_k) = P(A_1) + P(A_2) + \dots + P(A_k).$$

4. L'axiome 3 reste vrai quand $k \rightarrow \infty$.

a. Point de départ d'un raisonnement considéré comme non démontrable, évident.
(Larousse)

Ce sont les *axiomes de Kolmogorov* (p. 29 des notes).

Spécification des probabilités

Comment peut-on déterminer des probabilités ?

- estimations basées sur des observations
(ex. : 85 % des grains d'un lot d'avoine germent)
- analyse des conditions expérimentales
(ex. : un dé régulier, une pièce équilibrée)
- modèles probabilistes et/ou hypothèses
(ex. : chaque vêlage donne une femelle avec probabilité $1/2$, indépendamment des autres)

Le cas équiprobable

Proposition. Si Ω est un ensemble fini et si tous les résultats possibles ont la même probabilité, alors pour tout événement A ,

$$P(A) = \frac{\text{cardinal}(A)}{\text{cardinal}(\Omega)}.$$

Retour à l'exemple 1 : si les deux sexes sont équiprobables à chaque vêlage, les 8 résultats de Ω_1 sont équiprobables. Donc

$$P(\text{« exactement deux femelles »}) = \frac{\text{cardinal}\{\text{FFM}, \text{FMF}, \text{MFF}\}}{8} = \frac{3}{8}.$$

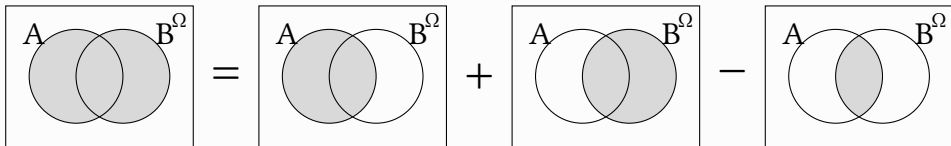
Résultats découlant des axiomes

Les résultats qui suivent sont des conséquences des axiomes.

- **Ensemble vide** : $P(\emptyset) = 0$
[Ω et $\emptyset$ sont mutuellement exclusifs, $\Omega \cup \emptyset = \Omega$ et $P(\Omega) = 1 \dots$]
- **Complément** : $P(A) = 1 - P(A^c)$
[A et A^c sont mutuellement exclusifs et $\Omega = A \cup A^c \dots$]
- **Formule de Poincaré** : $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- et pour trois événements :

$$\begin{aligned} P(A \cup B \cup C) = & P(A) + P(B) + P(C) \\ & - P(A \cap B) - P(A \cap C) - P(B \cap C) \\ & + P(A \cap B \cap C) \end{aligned}$$

Formule de Poincaré : pourquoi soustraire $P(A \cap B)$?



En additionnant $P(A)$ et $P(B)$, la zone $A \cap B$ est comptée **deux fois**.

Résultats utiles

Ces deux résultats (lois de De Morgan) sont souvent utiles :

$$P((A \cap B)^c) = P(A^c \cup B^c)$$

$$P((A \cup B)^c) = P(A^c \cap B^c)$$

Exemple : « ni l'espèce A ni l'espèce B n'est présente » = $A^c \cap B^c = (A \cup B)^c$.

Exemple 1 (suite)

En supposant qu'une génisse a autant de chances d'accoucher d'une femelle que d'un mâle à chaque gestation, quelle est la probabilité qu'elle ait **au moins une femelle** parmi ses 3 premiers petits ?

Indice : quel est le complément de « au moins une femelle » ?

Exemple 2 (suite)

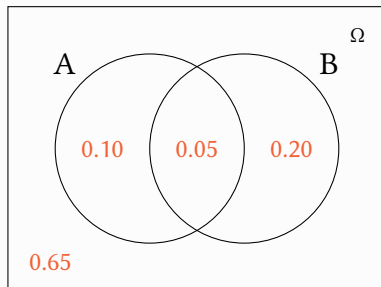
Les photographies aériennes du lac montrent que l'hydrocharis grenouillère (espèce A) est présente dans 15 % de la superficie du lac. Il n'y a qu'une autre espèce de macrophyte flottant (espèce B), présente dans 25 % de la superficie. Au total, 35 % de la superficie du lac est occupée par les macrophytes flottants.

- a) Quelle portion du lac est une zone d'eau libre, inoccupée par les macrophytes ?

- b) Si on choisit un site d'échantillonnage au hasard, quelle est la probabilité de tomber sur une zone bispécifique (les deux espèces présentes) ?

Exemple 2 : piste de solution

Remplir toutes les zones mutuellement exclusives d'un diagramme de Venn.
A = « l'hydrocharis est présente », B = « l'autre macrophyte est présent ».



2.3 Notions d'analyse combinatoire

Pourquoi compter ?

Dans le **cas équiprobable**, $P(A) = \text{cardinal}(A) / \text{cardinal}(\Omega)$: il suffit de **compter**.

Avec 3 veaux, on peut énumérer les 8 résultats à la main. Mais...

- avec 10 veaux, combien de séquences F/M possibles ?
- combien d'échantillons différents de 5 semences peut-on prélever dans un lot de 20 ?

⇒ L'**analyse combinatoire** regroupe les techniques de dénombrement utiles pour ces calculs (section 2.3 des notes). Trois outils suffisent pour ce cours : le principe de multiplication, les **permutations** et les **combinaisons**.

Principe fondamental du dénombrement

Principe fondamental du dénombrement

Si une procédure comporte k étapes successives, avec n_1 façons de réaliser la 1^{re} étape, puis n_2 façons de réaliser la 2^e (peu importe le choix fait à la 1^{re}), ..., puis n_k façons de réaliser la k^e , alors il y a en tout

$$n_1 \times n_2 \times \cdots \times n_k$$

façons de réaliser la procédure.

- Essai agronomique : 3 variétés de blé $\times$ 4 doses d'engrais $\times$ 2 régimes d'irrigation = 24 traitements.
- Exemple 1 : chaque veau est F ou M, donc $2 \times 2 \times 2 = 2^3 = 8$ résultats. Avec 10 veaux : $2^{10} = 1024$ résultats.

Permutations

Une **permutation** est un arrangement **ordonné** d'objets : l'ordre compte.

Nombre de permutations

- de n objets : $n! = n \times (n - 1) \times \cdots \times 2 \times 1$ (« n factoriel » ; $0! = 1$) ;
- de k objets choisis parmi n :

$$P_{n,k} = n \times (n - 1) \times \cdots \times (n - k + 1) = \frac{n!}{(n - k)!}.$$

- Ordre de passage de 5 échantillons sanguins dans un analyseur : $5! = 120$ ordres possibles.
- On attribue 3 traitements *différents* (T_1, T_2, T_3) à 3 des 8 parcelles d'un champ (une parcelle par traitement) : $P_{8,3} = 8 \times 7 \times 6 = 336$ façons.

Combinaisons

Une **combinaison** de k objets parmi n est un sous-ensemble de k objets : l'ordre ne compte pas.

Nombre de combinaisons (coefficient binomial)

$$\binom{n}{k} = C_{n,k} = \frac{n!}{k!(n-k)!} = \frac{P_{n,k}}{k!}$$

Chaque combinaison de k objets correspond à $k!$ permutations : on divise donc par $k!$.

- On choisit 3 des 8 parcelles pour recevoir *le même* engrais : $\binom{8}{3} = 56$ façons (et non 336).
- Exemple 1 : séquences de 3 veaux avec exactement 2 femelles = $\binom{3}{2} = 3$ (on choisit la position des 2 F).

Exemple 3 : un lot de semences

Un lot de 20 semences contient 4 semences infectées par un champignon pathogène. On en prélève 5 au hasard (sans remise) pour les analyser.

a) Combien d'échantillons de 5 semences différents peut-on prélever ?

$\binom{20}{5} = 15\,504$ échantillons, tous équiprobables.

b) Quelle est la probabilité qu'aucune semence infectée ne soit prélevée ?

Il faut choisir les 5 semences parmi les 16 saines : $\frac{\binom{16}{5}}{\binom{20}{5}} = \frac{4\,368}{15\,504} \approx 0.282$.

c) Quelle est la probabilité que l'échantillon contienne exactement une semence infectée ?

Indice : choisir 1 infectée parmi 4 et 4 saines parmi 16.

Dans R : `factorial(5)` donne 5! et `choose(20, 5)` donne $\binom{20}{5}$.

QUIZ!

Permutation ou combinaison? Combien de possibilités?

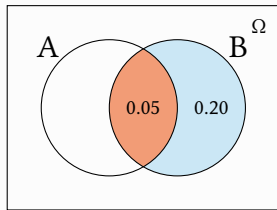
- On forme un comité de protection des animaux de 4 membres parmi 9 chercheurs.
- On dépose 6 échantillons d'ADN dans les 6 puits d'un gel d'électrophorèse. Combien d'ordres de dépôt?
- Un généticien choisit 2 lignées parmi 7 pour un croisement : l'une servira de mère, l'autre de père.
- Retour à l'exemple 1 : quelle est la probabilité qu'une génisse ait exactement 2 femelles parmi ses 10 premiers veaux?

2.4 Probabilité conditionnelle et indépendance

Une information partielle change les probabilités

Retour à l'exemple 2 : $P(A) = 0.15$, $P(B) = 0.25$ et $P(A \cap B) = 0.05$.

On choisit un site au hasard et on nous apprend que **l'espèce B y est présente**.
Quelle est maintenant la probabilité d'y trouver l'hydrocharis ?



On sait qu'on est dans B : **B devient le nouvel espace échantillonnal**. Quelle fraction de B est aussi dans A ?

$$\frac{P(A \cap B)}{P(B)} = \frac{0.05}{0.25} = 0.20.$$

L'information « B est présente » fait passer la probabilité de A de 0.15 à 0.20.

Probabilité conditionnelle

Définition. Soit A et B deux événements avec $P(B) > 0$. La **probabilité conditionnelle de A sachant B** est

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

$P(\cdot \mid B)$ respecte tous les axiomes et leurs conséquences : $0 \leq P(A \mid B) \leq 1$, $P(B \mid B) = 1$, $P(A^c \mid B) = 1 - P(A \mid B)$, etc.

Règle de multiplication (on isole l'intersection)

$$P(A \cap B) = P(A \mid B)P(B) = P(B \mid A)P(A)$$

Très utile quand l'expérience se déroule **en étapes successives**.

Attention : $P(A | B) \neq P(B | A)$!

Retour à l'exemple 2 :

$$P(A | B) = \frac{0.05}{0.25} = 0.20 \quad \text{mais} \quad P(B | A) = \frac{0.05}{0.15} \approx 0.33.$$

- 20 % des sites occupés par l'espèce B contiennent aussi l'hydrocharis;
- 33 % des sites occupés par l'hydrocharis contiennent aussi l'espèce B.

Autre exemple : presque toutes les personnes atteintes de la rougeole font de la fièvre : $P(\text{fièvre} | \text{rougeole}) \approx 1$. Mais parmi les personnes qui font de la fièvre, très peu ont la rougeole : $P(\text{rougeole} | \text{fièvre})$ est très petite.

Confondre les deux est une erreur **très fréquente**. On la retrouvera avec les tests diagnostiques (plus loin) et avec les tests d'hypothèses.

Exemple 3 (suite) : tirages successifs

Lot de 20 semences, dont 4 infectées. On tire 2 semences l'une après l'autre, sans remise. Soit I_1 = « la 1^{re} est infectée » et I_2 = « la 2^e est infectée ».

a) Que vaut $P(I_1)$? $4/20 = 0.2$

b) Que vaut $P(I_2 | I_1)$? Il reste 19 semences, dont 3 infectées : $3/19$.

c) Quelle est la probabilité que les deux semences soient infectées?

$$P(I_1 \cap I_2) = P(I_1) P(I_2 | I_1) = \frac{4}{20} \times \frac{3}{19} = \frac{12}{380} \approx 0.0316.$$

Vérification par dénombrement : $\binom{4}{2} / \binom{20}{2} = 6/190 \approx 0.0316$.

QUIZ!

Supposons $P(B) > 0$. Que vaut $P(A | B)$...(faites un diagramme de Venn!)

- si $B \subset A$?
- si $A \subset B$?
- si A et B sont mutuellement exclusifs?

Indépendance de deux événements

Je note le sexe d'un veau, puis je lance une pièce de monnaie. Soit les événements

- A = « le veau est une femelle » ;
- B = « la pièce indique face ».

Le fait que A se réalise ou non n'a aucun impact sur la probabilité de B !

On dira alors que A et B sont **indépendants**.

Indépendance de deux événements (section 2.4)

Deux événements A et B sont **indépendants** si et seulement si

$$P(A \cap B) = P(A) \cdot P(B).$$

Indépendance et probabilité conditionnelle

Si $P(B) > 0$, la règle de multiplication donne $P(A \cap B) = P(A | B) P(B)$. Donc

Reformulation

A et B sont indépendants $\iff P(A | B) = P(A)$.

Savoir que B s'est réalisé **ne change pas** la probabilité de A.

Plus de deux événements : A, B et C sont **(mutuellement) indépendants** si

$$P(A \cap B) = P(A) P(B), \quad P(A \cap C) = P(A) P(C), \quad P(B \cap C) = P(B) P(C)$$

$$\text{et } P(A \cap B \cap C) = P(A) P(B) P(C).$$

Exemple 1 : si les vêlages sont indépendants et que $P(F) = 1/2$ à chaque vêlage, alors $P(MMM) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}$. C'est ce qui justifie que les 8 résultats de Ω_1 sont équiprobables.

QUIZ!

Retour à l'exemple 2 : $P(A) = 0.15$, $P(B) = 0.25$, $P(A \cup B) = 0.35$.

- La présence de l'espèce A est-elle indépendante de celle de l'espèce B?
- S'agit-il d'événements mutuellement exclusifs?
- Deux événements mutuellement exclusifs, chacun de probabilité positive, peuvent-ils être indépendants?

Exemple 4 : un test de dépistage

Un test rapide sert à dépister une maladie M . Dans la population visée :

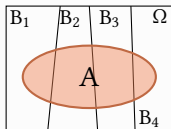
- **prévalence** : 2 % des personnes ont la maladie, $P(M) = 0.02$;
- **sensibilité** : le test est positif chez 95 % des malades, $P(T^+ | M) = 0.95$;
- **spécificité** : le test est négatif chez 90 % des non-malades, $P(T^- | M^c) = 0.90$.

Une personne choisie au hasard obtient un **résultat positif**. Quelle est la probabilité qu'elle soit réellement malade ?

C'est la **valeur prédictive positive** : $VPP = P(M | T^+)$.

Votre intuition ? Environ 95 % ? 90 % ? 50 % ? Moins ?

Partition et loi des probabilités totales



Partition : $B_1, \dots, B_k$ forment une **partition** de Ω s'ils sont mutuellement exclusifs et que leur union est Ω .

Ex. : $\{M, M^c\}$ (malade ou non).

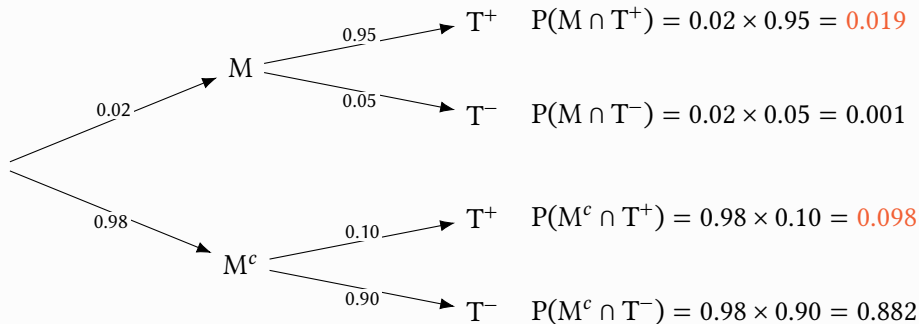
Loi des probabilités totales

Pour toute partition $B_1, \dots, B_k$ de Ω :

$$P(A) = \sum_{i=1}^k P(A \cap B_i) = \sum_{i=1}^k P(A \mid B_i) P(B_i).$$

En particulier, avec la partition $\{B, B^c\}$: $P(A) = P(A \mid B) P(B) + P(A \mid B^c) P(B^c)$.

Exemple 4 : arbre de probabilités



Chaque feuille : règle de multiplication (on multiplie le long des branches).

Loi des probabilités totales :

$$P(T^+) = P(M \cap T^+) + P(M^c \cap T^+) = 0.019 + 0.098 = 0.117.$$

Théorème de Bayes

Souvent, on connaît $P(A \mid B_j)$, mais on veut $P(B_j \mid A)$: il faut « renverser » le conditionnement.

Théorème de Bayes

Soit $B_1, \dots, B_k$ une partition de Ω et A un événement tel que $P(A) > 0$. Alors, pour chaque j ,

$$P(B_j \mid A) = \frac{P(A \mid B_j) P(B_j)}{\sum_{i=1}^k P(A \mid B_i) P(B_i)}.$$

Le numérateur vient de la règle de multiplication, le dénominateur de la loi des probabilités totales.

Exemple 4 : la valeur prédictive positive

Une personne obtient un test positif. Quelle est la probabilité qu'elle soit malade ? On applique le théorème de Bayes avec la partition $\{M, M^c\}$:

$$\begin{aligned} \text{VPP} = P(M \mid T^+) &= \frac{P(T^+ \mid M) P(M)}{P(T^+ \mid M) P(M) + P(T^+ \mid M^c) P(M^c)} \\ &= \frac{0.95 \times 0.02}{0.95 \times 0.02 + 0.10 \times 0.98} = \frac{0.019}{0.117} \approx 0.162. \end{aligned}$$

Seulement 16 % des personnes ayant un test positif sont réellement malades !

Attention !

La sensibilité $P(T^+ \mid M) = 0.95$ n'est pas la VPP $P(M \mid T^+) \approx 0.16$: on ne peut pas inverser le conditionnement sans tenir compte de la prévalence $P(M)$.

Pourquoi si peu ? Raisonons sur 10 000 personnes

	Test positif (T^+)	Test négatif (T^-)	Total
Malades (M)	190	10	200
Non-malades (M^c)	980	8 820	9 800
Total	1 170	8 830	10 000

Parmi les 1 170 tests positifs, seulement 190 viennent de personnes malades :
 $VPP = 190/1\,170 \approx 0.162$.

À retenir

Quand la maladie est **rare**, les faux positifs (980) sont beaucoup plus nombreux que les vrais positifs (190), même avec un bon test. La VPP dépend de la **prévalence**, pas seulement de la sensibilité et de la spécificité.

QUIZ!

Exemple 4 : prévalence 2 %, sensibilité 95 %, spécificité 90 %.

- Quelle est la **valeur prédictive négative**, $VPN = P(M^c | T^-)$?
- Le même test est utilisé dans une clinique spécialisée où 30 % des patients ont la maladie. Que devient la VPP?

2.5 Variables aléatoires et distributions

Exemple 5

Le résultat d'une expérience aléatoire est souvent un nombre réel, ou encore on assigne une valeur réelle aux résultats possibles d'une expérience.

$\mathcal{E}$: on note le sexe des trois premiers veaux d'une génisse.

$\Omega = \{FFF, FFM, FMF, MFF, FMM, MFM, MMF, MMM\}$.

Valeur associée à chaque résultat : le **nombre de femelles**. On note cette variable X .

$X = 0$ si MMM, $X = 1$ si FMM, MFM ou MMF, $X = 2$ si ..., $X = 3$ si FFF.

Variable aléatoire

Soit $\mathcal{E}$ une expérience aléatoire ayant comme espace échantillonnal Ω .

Variable aléatoire

Une **variable aléatoire** X est une quantité dont la valeur numérique dépend du résultat de l'expérience aléatoire : elle associe un nombre réel $X(e)$ à chaque résultat $e \in \Omega$.

On note habituellement une v.a. par une lettre majuscule et les valeurs prises par cette v.a. par une lettre minuscule. On écrira par exemple $P(X = x)$ pour la probabilité que la variable aléatoire X prenne la valeur x .

QUIZ!

Les variables aléatoires suivantes sont-elles discrètes ou continues?

- Précipitations en juillet à Québec (en mm)
- Diamètre à hauteur de poitrine d'un arbre choisi au hasard
- Biomasse de gazon sur une parcelle de 1 m sur 2 m
- Nombre de lésions foliaires sur le gazon d'une parcelle de 1 m sur 2 m
- Présence (1) ou absence (0) de composantes sériques dans un échantillon d'urine

Cas discret et continu

Type	Variables discrètes	Variables continues
Valeurs possibles	Dénombrables (souvent des entiers)	Un intervalle de la droite réelle
Description	Fonction de masse $p(x) = P(X = x)$	Fonction de densité $f(x)$
Représentation	Diagramme en bâtons	Courbe continue par morceaux
$P(2 \leq X < 5)$	Somme des bâtons de 2 à 4	Aire sous la densité entre 2 et 5
Propriétés	$0 \leq p(x) \leq 1, \sum_x p(x) = 1$	$f(x) \geq 0, \int_{-\infty}^{\infty} f(x) dx = 1$

Fonction de masse

Définition. Soit X une v.a. discrète. On appelle **fonction de masse** (ou distribution de probabilité) de X la fonction qui associe la valeur

$$p(x) = P(X = x)$$

à toute valeur x .

Une fonction $p(x)$ est une fonction de masse **si et seulement si**

1. $0 \leq p(x) \leq 1$ pour tout x ;
2. $\sum_x p(x) = 1$.

Retour à l'exemple 5

Trouvons la fonction de masse de X = nombre de femelles parmi les trois premiers veaux.

e	$X(e)$	$P(e)$
FFF		1/8
FFM		1/8
FMF		1/8
MFF		1/8
FMM		1/8
MFM		1/8
MMF		1/8
MMM		1/8

x	$p(x) = P(X = x)$
0	
1	
2	
3	

QUIZ!

Les fonctions suivantes sont-elles des fonctions de masse ?

- $p(x) = x/6$ pour $x \in \{1, 2, 3\}$ (et 0 ailleurs)
- $p(x) = (x - 1)/4$ pour $x \in \{0, 1, 2, 3\}$ (et 0 ailleurs)
- $p(x) = 0.3$ pour $x \in \{0, 1, 2\}$ (et 0 ailleurs)

Exemple 6

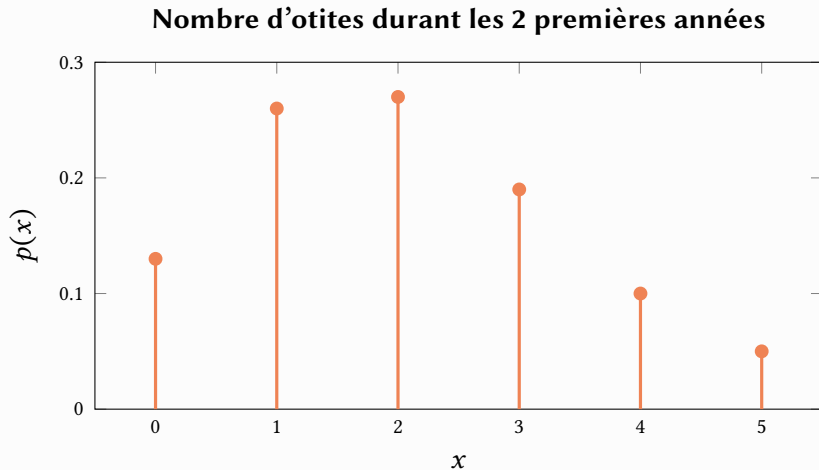
Soit X = le nombre d'otites durant les deux premières années de vie d'un bébé. Les valeurs possibles de X sont 0, 1, 2, ...

Supposons que la fonction de masse de X est donnée par le tableau ci-contre.

- a) Quelle est la probabilité de faire plus de 3 otites ?
- b) Quelle est la probabilité de faire au moins une otite ?

x	$p(x)$
0	0.13
1	0.26
2	0.27
3	0.19
4	0.10
5	0.05

Exemple 6 : représentation graphique



Aspects importants d'une distribution

Même si la fonction de masse (ou de densité) spécifie complètement la distribution d'une variable aléatoire, il est utile de résumer les aspects principaux de son comportement :

- **tendance centrale** → l'espérance μ ;
- **dispersion** → la variance σ^2 et l'écart type σ ;
- forme.

Ce sont les analogues « théoriques » de $\bar{x}$ et de s vus au chapitre 1.

Mesure de la tendance centrale

Rappel de la moyenne échantillonnale : $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$.

Définition. L'**espérance** (moyenne théorique) d'une variable aléatoire discrète X est donnée par

$$\mu = E(X) = \sum_x x \cdot p(x).$$

Interprétation : c'est la valeur moyenne que l'on obtiendrait si on réalisait l'expérience aléatoire un très grand nombre de fois, en notant la valeur de X à chaque fois.

Exemple 6 (suite)

En moyenne, combien d'otites un bébé fait-il durant ses deux premières années de vie ?

Variabilité d'une variable aléatoire

Deux variables aléatoires peuvent avoir la même espérance, mais l'une peut être beaucoup plus variable que l'autre.

Expérience : rouler 30 dés réguliers.

- X_1 = la valeur indiquée par le premier dé;
- X_2 = la moyenne des 30 dés.

On peut montrer que X_1 et X_2 ont la même espérance (3.5).

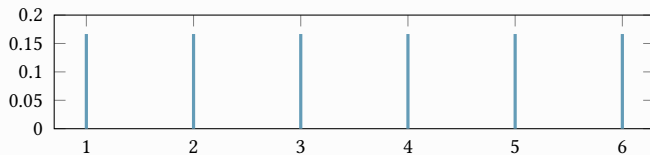
Cependant, X_2 aura tendance à prendre des valeurs très proches de 3.5, alors que X_1 peut prendre n'importe quelle valeur entre 1 et 6 avec la même probabilité.

Petite simulation

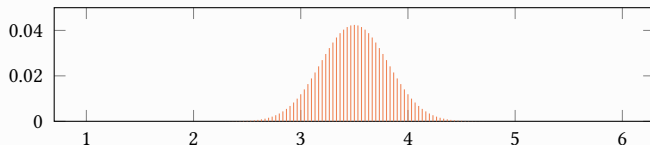
On a simulé 12 fois
l'expérience à l'aide de R :

	X_1	X_2
1	1	3.633
2	2	3.867
3	1	3.900
4	4	3.900
5	6	4.067
6	4	3.233
7	2	3.833
8	5	3.200
9	4	3.767
10	5	3.433
11	3	3.500
12	5	3.333

Fonction de masse de X_1



Fonction de masse de X_2



Mesure de dispersion

Rappel de la variance échantillonnale : $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$.

Définition. La **variance** d'une variable aléatoire discrète X est donnée par

$$\sigma^2 = \text{Var}(X) = E[(X - \mu)^2] = \sum_x (x - \mu)^2 \cdot p(x).$$

Une quantité très utile qui découle de la variance est l'**écart type** :
 $\sigma = \sqrt{\text{Var}(X)} = \sqrt{\sigma^2}$.

Interprétation : c'est la dispersion que l'on obtiendrait si on réalisait l'expérience aléatoire un très grand nombre de fois.

Une formule de calcul pratique

Plus généralement, pour une fonction g , l'espérance de $g(X)$ est

$$E[g(X)] = \sum_x g(x) \cdot p(x).$$

Par exemple, $E(X^2) = \sum_x x^2 \cdot p(x)$.

Proposition.

$$\sigma^2 = \text{Var}(X) = E(X^2) - \mu^2.$$

Cette formule est souvent plus rapide que la définition : on calcule $\sum_x x p(x)$ et $\sum_x x^2 p(x)$ dans le même tableau.

Exemple 6 (suite)

Calculez la variance et l'écart type du nombre d'otites durant les deux premières années de vie.

x	$p(x)$	$x \cdot p(x)$	$x^2 \cdot p(x)$
0	0.13		
1	0.26		
2	0.27		
3	0.19		
4	0.10		
5	0.05		
Total	1	$\mu =$	$E(X^2) =$

QUIZ!

Retour à l'exemple 2. On choisit un site au hasard dans le lac et on pose $X = 1$ si l'hydrocharis y est présente, $X = 0$ sinon. Ainsi, $P(X = 1) = 0.15$ et $P(X = 0) = 0.85$.

- Que vaut $E(X)$?
- Que vaut $\text{Var}(X)$?
- Et en général, si $P(X = 1) = p$ et $P(X = 0) = 1 - p$? (On retrouvera cette variable, dite de *Bernoulli*, dans la 2^e partie du chapitre.)

Fonction de densité (p. 47)

Soit X une variable aléatoire continue. L'ensemble des valeurs possibles de X est donc un intervalle de la droite réelle.

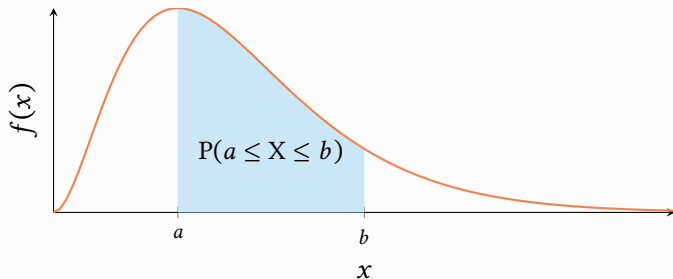
Définition. La **fonction de densité** $f(x)$ de X est une fonction telle que, pour tout $a \leq b$,

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

Une fonction $f(x)$ est une fonction de densité **si et seulement si**

1. $f(x) \geq 0$ pour tout x ;
2. $\int_{-\infty}^{\infty} f(x) dx = 1$.

Probabilité = aire sous la densité



$f(x)$ n'est **pas** une probabilité, mais son intégrale sur une région donne la probabilité que X prenne une valeur dans cette région.

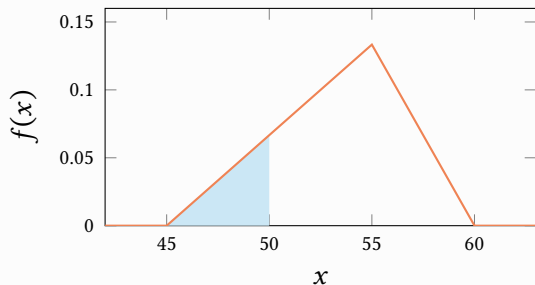
Comme $\int_a^a f(x) dx = 0$, on a $P(X = a) = 0$: pour une v.a. continue, $P(a \leq X \leq b) = P(a < X < b)$.

Exemple 7

Supposons que le poids de 1000 grains de blé (PMG) est une variable X mesurée en grammes, de fonction de densité suivante pour une certaine variété :

$$f(x) = \begin{cases} \frac{x - 45}{75} & \text{si } 45 \leq x < 55, \\ \frac{120 - 2x}{75} & \text{si } 55 \leq x < 60, \\ 0 & \text{sinon.} \end{cases}$$

Densité du PMG



- a) Quelle est la probabilité que le PMG soit inférieur à 50 g?
- b) Quelle est la probabilité que le PMG soit entre 50 g et 56 g?

QUIZ!

Vrai ou faux?

- Une fonction de densité peut prendre des valeurs supérieures à 1.
- Pour le PMG de l'exemple 7, $P(X = 50) = 0$.
- Pour le PMG de l'exemple 7, $P(X < 50) = P(X \leq 50)$.
- Pour le nombre d'otites de l'exemple 6, $P(X < 2) = P(X \leq 2)$.

Fonction de répartition

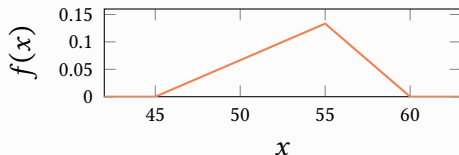
Définition. La **fonction de répartition** de X est la fonction

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

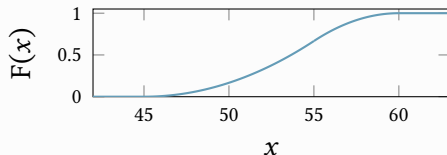
Cela revient à calculer l'aire à gauche de x sous la densité. On a aussi

$$P(a < X \leq b) = F(b) - F(a).$$

Densité du PMG



Répartition du PMG



Espérance et variance d'une v.a. continue

Même définitions que dans le cas discret, en remplaçant la somme par une intégrale et $p(x)$ par $f(x)$.

Définition.

$$\mu = E(X) = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

$$\sigma^2 = \text{Var}(X) = E[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 \cdot f(x) dx$$

μ et σ^2 s'interprètent de la même façon que dans le cas discret.

Exemple 7 (suite)

Quel est le PMG moyen de cette variété?

Résultats utiles (p. 50)

Soit X et Y des variables aléatoires (discrètes ou continues), et a et b des constantes.

Proposition.

1. $E(aX + b) = aE(X) + b$
2. $E(X + Y) = E(X) + E(Y)$
3. $\text{Var}(aX + b) = a^2 \text{Var}(X)$
4. $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$ si X et Y sont **indépendantes**.

Pourquoi b disparaît-il de la variance ? Ajouter une constante déplace la distribution sans changer sa dispersion.

Exemple 8

La température corporelle des rats de laboratoire est une variable aléatoire ayant une espérance de $37,6^{\circ}\text{C}$ et une variance de $1.44(^{\circ}\text{C})^2$.

a) En $^{\circ}\text{F}$, à combien s'élève la température moyenne des rats, sachant que $^{\circ}\text{F} = 1.8^{\circ}\text{C} + 32$?

b) Quelle est la variance de la température en $^{\circ}\text{F}$?

Exemple 9

Supposons que la masse d'une orange est une variable aléatoire dont l'espérance est de 204 g et la variance est de 256 g^2 . On crée des sacs de 7 oranges, dont les masses sont indépendantes.

a) Quelle sera l'espérance de la masse des sacs d'oranges ?

b) Quel sera l'écart type de la masse des sacs d'oranges ?

QUIZ!

Soit X la masse d'une orange. Pourquoi la masse d'un sac, $X_1 + X_2 + \dots + X_7$, n'a-t-elle **pas** la même variance que $7X$?

Réponses aux exemples

1. Ω_1 et Ω_2 sont discrets, Ω_3 est continu. $P(\text{au moins une femelle}) = 1 - P(\text{MMM}) = 1 - 1/8 = 7/8 = 0.875$.
Quiz (mutuellement exclusifs) : non dans les deux cas ($A \cap C = C \neq \emptyset$; $A \cap B = (150, 175) \neq \emptyset$).
2. a) $1 - 0.35 = 0.65$; b) $P(A \cap B) = 0.15 + 0.25 - 0.35 = 0.05$. $P(A | B) = 0.20$, $P(B | A) \approx 0.33$.
Quiz : non indépendants ($P(A)P(B) = 0.0375 \neq 0.05$, ou $P(A | B) = 0.20 \neq 0.15$); non mutuellement exclusifs ($P(A \cap B) > 0$); jamais : $P(A \cap B) = 0 \neq P(A)P(B) > 0$.
3. a) 15 504; b) ≈ 0.282 ; c) $\binom{4}{1} \binom{16}{4} / \binom{20}{5} = 7\,280 / 15\,504 \approx 0.470$; tirages successifs : ≈ 0.0316 .
Quiz (dénombrement) : combinaison $\binom{9}{4} = 126$; permutation $6! = 720$; permutation $P_{7,2} = 42$;
 $\binom{10}{2} / 2^{10} = 45 / 1024 \approx 0.044$.
Quiz ($P(A | B)$) : 1; $P(A)/P(B)$; 0.
4. $P(T^+) = 0.117$; $VPP = 0.019 / 0.117 \approx 0.162$.
Quiz : $VPN = 0.882 / 0.883 \approx 0.999$; prévalence de 30 % : $VPP = 0.285 / (0.285 + 0.070) = 0.285 / 0.355 \approx 0.803$.
5. $p(0) = 1/8$, $p(1) = 3/8$, $p(2) = 3/8$, $p(3) = 1/8$.
Quiz (discrète ou continue) : continue, continue, continue, discrète, discrète. Quiz (fonctions de masse) : oui; non ($p(0) < 0$); non (somme = 0.9).

Réponses aux exemples (suite)

6. a) $P(X > 3) = 0.10 + 0.05 = 0.15$; b) $P(X \geq 1) = 1 - 0.13 = 0.87$. $\mu = E(X) = 2.02$ otites.
 $E(X^2) = 5.90$, $\sigma^2 = 5.90 - 2.02^2 = 1.8196 \approx 1.82$, $\sigma \approx 1.35$ otite.
Quiz (Bernoulli) : $E(X) = 0.15$, $\text{Var}(X) = 0.15 \times 0.85 = 0.1275$; en général $E(X) = p$ et $\text{Var}(X) = p(1 - p)$.
7. a) $P(X < 50) = F(50) = 25/150 = 1/6 \approx 0.167$; b) $F(56) - F(50) = 59/75 - 1/6 = 0.62$. PMG moyen :
 $E(X) = 160/3 \approx 53.3$ g.
Quiz (vrai ou faux) : vrai (ex. : $f(x) = 2$ sur $[0, 0.5]$); vrai; vrai; faux ($P(X < 2) = 0.39$ mais $P(X \leq 2) = 0.66$).
8. a) $1.8 \times 37.6 + 32 = 99.68^\circ\text{F}$; b) $1.8^2 \times 1.44 = 4.6656 (^\circ\text{F})^2$.
9. a) $7 \times 204 = 1428$ g; b) $\text{Var} = 7 \times 256 = 1792 \text{ g}^2$, donc $\sigma = \sqrt{1792} \approx 42.3$ g.
Quiz : $\text{Var}(7X) = 49 \times 256 = 12\,544 \text{ g}^2$ ($\sigma = 112$ g). $7X$ suppose que les 7 oranges ont *la même* masse; avec 7 oranges indépendantes, les écarts à la moyenne se compensent en partie, d'où une variance plus petite.

Références

Ce document est rédigé avec Beamer ainsi qu'une classe .cls fournie par Jérôme Soucy.

Les diapositives sont adaptées des diapositives créées par Emmanuelle Reny-Nolin pour le cours STT-1920 ainsi que de celles des cours STT-1000 et STT-1900.

Référence : C. Bélisle, *STT-1920 Méthodes statistiques*, notes de cours, Université Laval, 2020, chapitre 2.