Probabilités et Statistique pour Physique

http://stat.epfl.ch

1 Introduction

slide 2

Probabilités

A quel genre de questions répond-on en probabilités ?

— Quelle est la chance que deux personnes soient nées le même jour dans cette classe ?

— Si un singe tape sur un clavier au hasard, combien de temps avant de voir 'epfl' et 'zzzz' ?

— Si on lance une pièce de monnaie 10000 fois, on aura environ 5000 piles, 5000 faces. Quel est l'ordre de grandeur de la déviation ? A quoi ressemble-t-elle ?

— Si on marche au hasard sur une grille infinie, quelle chance de revenir d'où on est parti ?

— Comment expliquer la trajectoire d'une particule de pollen dans l'eau ?

`http://stat.epfl.ch` slide 3

Statistiques

A quel genre de questions répond-on en statistiques ?

— On lance un dé à six faces 60000 fois. On obtient 9300 fois le nombre 5. Est-ce que le dé est truqué ?

— On veut savoir la proportion de gauchers et droitiers à Lausanne. Combien de personnes doit-on interroger pour avoir une estimation au % près, valide à 99% ?

— Est-ce qu'on peut améliorer cette estimation avec des informations d'internet ?

— Est-ce que boire du café améliore les notes ?

— Combien de collisions effectuer au LHC pour savoir si le boson de Higgs existe avec probabilité $1 - 10^{-7}$?

— On a une suite de mesures (avec erreurs) de la position d'un corps céleste. Comment prédire sa position dans 10 ans ?

`http://stat.epfl.ch` slide 4

Probabilités vs. statistiques ?

— Pour simplifier :

  — en **probabilités**, on connaît le modèle, on s'intéresse à ce qu'il génère ;

  — en **statistiques** on a des données, on s'intéresse au modèle sous-jacent (en vrai, c'est plus compliqué).

`http://stat.epfl.ch` slide 5

Quelques résultats intéressants (en vrac)

— Si on trace un histogramme de données, avec beaucoup de données, on obtient très souvent une gaussienne $\propto e^{-x^2/2}$.

— On revient toujours à la position si la grille est en 2D, pas toujours en 3D (sur $\mathbb{Z}^3$, à chaque saut, on choisit un des 6 voisins au hasard).

— Le mouvement du pollen dans l'eau a été expliqué par Einstein en 1905 et est relié à la variable aléatoire gaussienne.

— Pour déterminer si un produit a un effet, on peut faire un test simple avec du papier et une calculatrice.

CERN

Standard Model

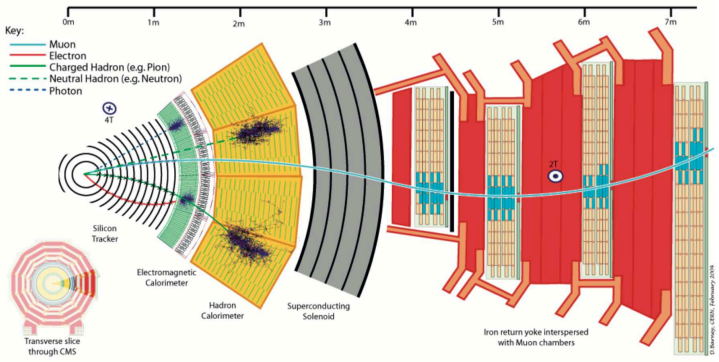


Figure 2.1: Illustration of the detection of particles at the CMS experiment (Barney, 2004). Each type of particle leaves its characteristic trace in the various subdetectors. This enables the identification of different particles and the measurement of their energies and trajectories. Copyright: CERN, for the benefit of the CMS Collaboration.

‘Standard Model’

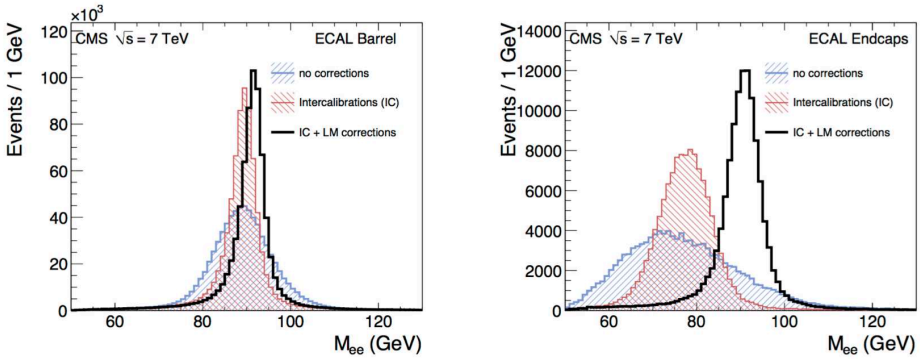
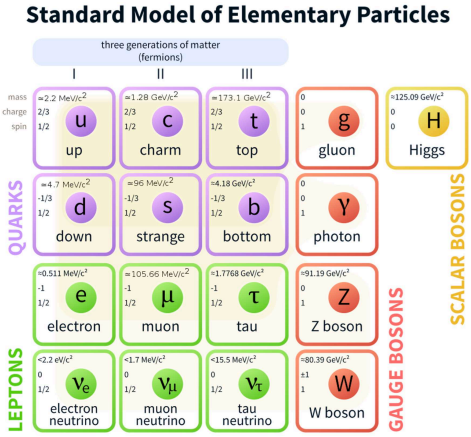


Figure 8: Reconstructed invariant mass from $Z \to e^+e^-$ decays, for single-channel corrections set to unity (blue), for final intercalibration (red), and for both final intercalibration and LM corrections (black), in the EB (left) and the EE (right).

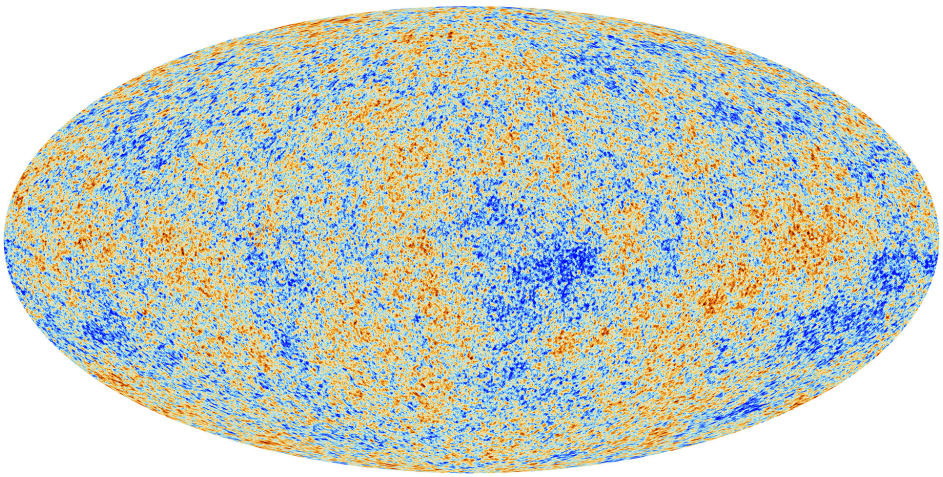
‘Standard Model’



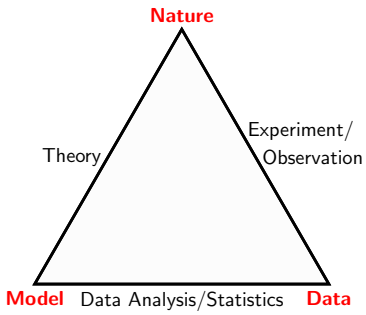
— Le top quark a été découvert en 1995.

— Le résultat des expériences menées pour le trouver était une variable $y = 17$, qui devrait avoir une loi de probabilité avec moyenne 6.7 si ce quark n'existait pas. Est-ce ceci donnait un preuve de son existance ?

Cosmic microwave background



Construction of knowledge : Recent past



— To try and understand **Nature**, we invent **theories** that lead to **models** (e.g., Mendelian genetics, quantum theory, fluid mechanics, ...).

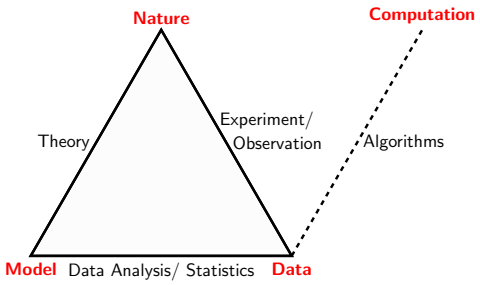
— We contrast the models with **observations**, preferably from **experiments** over which we have some control, to assess if the theory is adequate, or should be rejected or improved.

— The **data** are never measured exactly, so we usually need **Statistics** to assess whether differences between the data and model are due to measurement error, or whether the data conflict with the model, and therefore undermine or even falsify the theory.

— Data can only be used to falsify a theory—future data might be incompatible with it.

Big Data . . .

Construction of knowledge : Now-ish



— Sometimes we just want a good prediction :
  — How long will it take to drive to Geneva ?
  — Should I give this person life insurance ?
— Then we can use **algorithmic learning** — random forests, SVMs, deep learning, . . .— increasingly used in hard science also.

Best Jobs 2019 . . .

BEST JOBS OF 2019

RANKING	PROFESSION	ANNUAL MEDIAN SALARY	GROWTH OUTLOOK (TO 2026)
1	Data scientist	"$118,370 "	19%
2	Statistician	"$88,190 "	33%
3	University professor	"$78,470 "	15%
4	Occupational therapist	"$84,270 "	24%
5	Genetic counselor	"$80,370 "	29%
6	Medical services manager	"$99,730 "	20%
7	Information security analyst	"$98,350 "	28%
8	Mathematician	"$88,190 "	33%
9	Operations research analyst	"$83,390 "	27%
10	Actuary	"$102,880 "	22%

Worst Jobs 2019 . . .

AT THE BOTTOM OF THE LIST

RANKING	PROFESSION	ANNUAL MEDIAN SALARY	GROWTH OUTLOOK (TO 2026)
211	Broadcaster	"$66,880 "	0%
212	Advertising salesperson	"$51,740 "	-4%
213	Nuclear decontamination technician	"$42,030 "	17%
214	Disc jockey	"$31,990 "	-9%
215	Correctional officer	"$44,400 "	-7%
216	Enlisted military personnel	"$26,802 "	n/a
217	Retail salesperson	"$24,340 "	2%
218	Newspaper reporter	"$43,490 "	-9%
219	Logging worker	"$40,650 "	-13%
220	Taxi driver	"$25,980 "	5%

Matériel de cours

Les **probabilités** constituent à peu près les deux premiers tiers du cours, de bons livres sont

— Ross, S. M. (2007) *Initiation aux probabilités*. PPUR : Lausanne.

— Pfister, C.-E. (2014) *Théorie des probabilités*. PPUR : Lausanne.

— Dalang, R. C. et Conus, D. (2018) *Introduction à la théorie des probabilités*, deuxième édition. PPUR : Lausanne.

— Il y a beaucoup d'autres excellents livres d'introduction : regarder au RLC.

Les références en **statistiques** seront données ultérieurement.

Informations

- **Enseignant** : Professeur A. C. Davison
- **Cours** : Mardi 13.15–15.00, CM1
- **Exercices** : Lundi 9.15–10.00, CM1
- **Test (?)** : Mardi 5 novembre 2024, 13.15–15.00, sans aucune matière écrite
- **Moodle** avec notes de cours, exercices (y compris **Random Exercise Generator**), solutions, ...
- **Sorties de sécurité**
- **Respect** : https://www.epfl.ch/about/respect/

2 Probabilités

slide 19

2.1 Espace de Probabilité

slide 20

Les joueurs de cartes

Paul Cézanne, 1894–95, Musée d'Orsay, Paris

slide 21

Les événements

- Exemples :
  - On lance un dé et obtient 5.
  - On lance 3 dés et obtient 3 nombres distincts.
  - Il pleuvra à la gare de Lausanne le lundi du Jeûne fédéral.
  - Deux particules données vont se déintégrer en entrant en collision.
  - Il y a de la vie sur la planète Mars.
- Points communs :
  - un événement fait référence à une extension (développement) incertaine (e.g., le futur) d'une situation donnée (e.g., on s'apprête à lancer un dé) ;
  - dans cette extension, la question "l'événement a-t-il eu lieu ?" a une réponse Oui ou Non.

slide 22

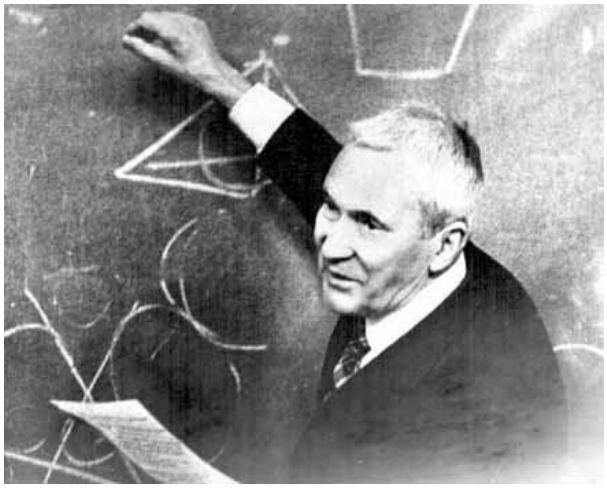
Calcul des événements

— Soient $\mathcal{A}$, $\mathcal{B}$, $\{\mathcal{A}_i\}_{i=1}^{\infty}$ des événements. Alors on peut aussi répondre **Oui** ou **Non** aux questions :
  — 'est-ce que $\mathcal{A}$ n'a pas eu lieu ?'
  — 'est-ce que $\mathcal{A}$ **et** $\mathcal{B}$ ont eu lieu ?'
  — 'est-ce que $\mathcal{A}$ **ou** $\mathcal{B}$ ont eu lieu ?'
  — 'est-ce que $\bigcup_{i=1}^{\infty} \mathcal{A}_i$ a eu lieu ?'
— Donc un modèle mathématique pour des événements doit pouvoir répondre à ces questions, et en plus les donner des probabilités.

Définition 1. *On définit les intersections et unions infinies ainsi :*

$$x \in \bigcup_{i=1}^{\infty} \mathcal{A}_i \iff \exists i : x \in \mathcal{A}_i, \quad x \in \bigcap_{i=1}^{\infty} \mathcal{A}_i \iff \forall i : x \in \mathcal{A}_i.$$

Andrey Nikolaevich Kolmogorov (1903–1987)



Grundbegriffe der Wahrscheinlichkeitsrechnung (1933)

Espace de probabilité

Définition 2. *Une* **expérience aléatoire** *est une 'expérience' dont le résultat est (ou peut être traîté comme) aléatoire.*

Une expérience aléatoire est modelisée par **un espace de probabilité**.

Définition 3. *Un* **espace de probabilité** $(\Omega, \mathcal{F}, \mathrm{P})$ *est un objet mathématique associé à une expérience aléatoire, constitué de :*

1. *un ensemble* Ω*,* ***l'ensemble fondamental (univers)****, qui contient tous les* ***résultats (épreuves, événements élémentaires)*** ω *possibles de l'expérience ;*
2. *une collection* $\mathcal{F}$ *de sous-ensembles de* Ω*. Ces sous-ensembles sont appelés* ***événements****, et* $\mathcal{F}$ *est appelé* ***l'espace des événements*** *;*
3. *une fonction* $\mathrm{P} : \mathcal{F} \mapsto [0, 1]$ *appelée* ***loi de probabilité****, qui associe une probabilité* $\mathrm{P}(A)$ *à chaque* $A \in \mathcal{F}$*.*

Ensemble fondamental Ω

— L'ensemble fondamental Ω est l'ensemble composé d'éléments représentant tous les résultats possibles d'une expérience aléatoire.

— Chaque élément $\omega \in \Omega$ est associé à un résultat différent.

— Ω est analogue à l'ensemble univers. Il peut être fini, dénombrable ou non dénombrable.

— Ω est non-vide. (Si $\Omega = \emptyset$, alors l'expérience est triviale.)

Exemple 4. *Décrire* Ω *quand je jête deux dès, un rouge et un vert.*

Pause pensées. *Décrire* Ω *pour un lancer d'une pièce avec deux faces.*

Note to Example 4

On peut écrire

$$\Omega = \{(r, g) : r, g = 1, \ldots, 6\} = \{\omega_1, \ldots, \omega_{36}\},$$

avec $\omega_1 = (1, 1), \omega_2 = (1, 2), \ldots, \omega_{35} = (5, 6), \omega_{36} = (6, 6)$, par exemple.

Espace des evénements $\mathcal{F}$

$\mathcal{F}$ est un ensemble de sous-ensembles de Ω qui représente les événements d'intérêt.

Définition 5. *Un* **espace des événements** $\mathcal{F}$ *est un ensemble de sous-ensembles de* Ω *tel que :*

$(\mathcal{F}1)$ $\mathcal{F}$ *est non vide ;*

$(\mathcal{F}2)$ *si* $\mathcal{A} \in \mathcal{F}$ *alors* $\mathcal{A}^c \in \mathcal{F}$ *;*

$(\mathcal{F}3)$ *si* $\{\mathcal{A}_i\}_{i=1}^{\infty}$ *sont tous des éléments de* $\mathcal{F}$*, alors* $\bigcup_{i=1}^{\infty} \mathcal{A}_i \in \mathcal{F}$.

$\mathcal{F}$ *est aussi appelée* **une tribu** *ou* **σ-algèbre**.

Soient $\mathcal{A}, \mathcal{B}, \{\mathcal{A}_i\}_{i=1}^{\infty}$ des éléments de $\mathcal{F}$. Alors on a (par exemple)

(a) $\bigcup_{i=1}^{n} \mathcal{A}_i \in \mathcal{F}$,

(b) $\Omega \in \mathcal{F}, \quad \emptyset \in \mathcal{F}$,

(c) $\mathcal{A} \cap \mathcal{B} \in \mathcal{F}$,

(d) $\bigcap_{i=1}^{n} \mathcal{A}_i \in \mathcal{F}$.

Pause pensées. *Trouver une tribu valide lorsque* $\Omega = \{0, 1\}$.

Use of these axioms

To prove (a)–(d), we argue as follows :

(a) Take $A_{n+1} = A_{n+2} = \cdots = A_n$, and apply $(\mathcal{F}3)$.

(b) If $\mathcal{F}$ is non-empty, then it has an element A, and by $(\mathcal{F}2)$ $A^c \in \mathcal{F}$, so $A \cup A^c = \Omega \in \mathcal{F}$ and $\Omega^c = \emptyset \in \mathcal{F}$.

(c) $A \cap B = (A^c \cup B^c)^c$, and sets operated on by union and complement remain in $\mathcal{F}$.

(d) We write $\bigcap_{i=1}^{n} A_i = ((\bigcap_{i=1}^{n} A_i)^c)^c = (\bigcup_{i=1}^{n} A_i^c)^c \in \mathcal{F}$.

Espace des evénements $\mathcal{F}$, II

— Si Ω est dénombrable, on prend souvent pour $\mathcal{F}$ l'ensemble de tous les sous-ensembles de Ω. C'est le plus grand (et le plus riche) espace des événements possibles pour Ω.

— Si Ω est non-dénombrable (e.g., $\mathbb{R}$), on ne peut pas faire ainsi—on arrive à des inconsistances mathématiques que l'on peut ignorer sans trop de problèmes pendant ce cours.

— On peut définir des tribus différents pour le même Ω.

— Le tribu choisi dépend non seulement de expérience aléatoire, mais aussi de ce que l'on est capable d'observer.

Exemple 6. *Je lance deux dés, un rouge et un vert.*
(a) Quel est ma tribu $\mathcal{F}_1$ *?*
(b) J'informe James seulement du total. Quel est sa tribu $\mathcal{F}_2$ *?*
(c) John regarde lui-même les dés, mais il est daltonien. Quel est sa tribu $\mathcal{F}_3$ *?*

Note to Example 6

(a) Since we see an outcome of the form $\omega = (r, g)$, we can reply to any question about the outcomes ; thus we take $\mathcal{F}_1$ to be the set of all possible subsets of $\Omega = \{(r, g) : r, g \in \{1, \ldots, 6\}\}$. If we take an arbitrary element $\mathcal{A} \in \mathcal{F}_1$, then each of the 36 ordered pairs (r, g) can either be in $\mathcal{A}$ or not, independently, so there are 2^{36} distinct sets $\mathcal{A}$, and thus $|\mathcal{F}_1| = 2^{36}$. In this case $\mathcal{F}_1$ is the so-called 'power set' of Ω.
(b) If I tell James only that the 'total is t' for $t = 2, \ldots, 12$, then he can reply to any question about the total, but nothing else. We can think of this as a mapping

$$t : \Omega \to \Omega_2 = \{\omega_2', \ldots, \omega_{12}'\},$$

where $t(r, g) = r + g$, and $\omega_i' = \{\omega \in \Omega : t(\omega) = i\}$, for $i = 2, \ldots, 12$. Thus Ω_2 has composite elementary outcomes

$$\omega_2' = \{(1,1)\}, \omega_3' = \{(1,2),(2,1)\}, \omega_4' = \{1,3),(2,2)(3,1)\}, \ldots, \omega_{11}' = \{(5,6),(6,5)\}, \omega_{12}' = \{(6,6)\}.$$

James can respond to questions about arbitrary combinations of the ω_i', so his sigma-algebra $\mathcal{F}_2$ is constructed by taking all possible subsets of Ω_2'. Now $|\Omega_2| = 11$, so by the argument in (a) we have $|\mathcal{F}_2| = 2^{11}$.
(c) Since John is colour-blind, he cannot tell the difference between $(1, 2)$ and $(2, 1)$, etc.. Thus $\mathcal{F}_3$ is made up of all possible complements and unions of the sets

$$\{(1,1)\}, \{(2,2)\}, \ldots, \{(6,6)\}, \{(1,2),(2,1)\}, \{(1,3),(3,1)\}, \ldots, \{(5,6),(6,5)\}.$$

John cannot answer 'yes' or 'no' to questions such as 'did $(1, 2)$ occur ?' ; for him this is not an event. There are 6 + 15 of such sets, so $|\mathcal{F}_3| = 2^{21}$, and obviously $\mathcal{F}_2 \subset \mathcal{F}_3 \subset \mathcal{F}_1$.
In cases (b) and (c) the tribus have less information than in (a) : they represent a coarsening of $\mathcal{F}_1$, so less precise questions can be answered.

Probabilité

— Dans les exemples élémentaires avec Ω fini, on choisit Ω de manière à ce que $\omega \in \Omega$ soit équiprobable :

$$\mathrm{P}(\{\omega\}) = \frac{1}{|\Omega|}, \quad \text{pour chaque } \omega \in \Omega.$$

Alors $\mathrm{P}(\mathcal{A}) = |\mathcal{A}|/|\Omega|$, pour tout $\mathcal{A} \subset \Omega$.

— Si on répète une expérience n fois de manière 'indépendantes', on peut montrer ('la loi des grands nombres') que

$$\frac{\#\text{occurrences de } \mathcal{A}}{n} \underset{n\to\infty}{\longrightarrow} \mathrm{P}(\mathcal{A}).$$

— En général : $\mathrm{P}(\mathcal{A})$ représente la confiance qu'on a dans le fait que $\mathcal{A}$ se produise.

— On attend que

  — Si $\mathcal{B}$ a toujours lieu quand $\mathcal{A}$ a lieu (c'est à dire $\mathcal{A} \subset \mathcal{B}$) alors $\mathrm{P}(\mathcal{B}) \geq \mathrm{P}(\mathcal{A})$.

  — Si $\mathcal{A}$ et $\mathcal{B}$ sont **disjoints** (n'ont jamais lieu simultanément, c'est-à-dire $\mathcal{A}$ **et** $\mathcal{B} = \mathcal{A} \cap \mathcal{B} = \emptyset$) alors $\mathrm{P}(\mathcal{A}$ **ou** $\mathcal{B}) = \mathrm{P}(\mathcal{A} \cup \mathcal{B}) = \mathrm{P}(\mathcal{A}) + \mathrm{P}(\mathcal{B})$.

  — En fait : on s'attend à ce que P soit un 'comptage de cas' généralisé.

Loi de probabilité P

Définition 7. *Une* **loi de probabilité** P *associe une probabilité à chaque élément de l'espace des événements* $\mathcal{F}$*, avec les propriétés suivantes :*

$(P1)$ *si* $\mathcal{A} \in \mathcal{F}$*, alors* $0 \leq \mathrm{P}(\mathcal{A}) \leq 1$ *;*

$(P2)$ $\mathrm{P}(\Omega) = 1$ *;*

$(P3)$ *si* $\{\mathcal{A}_i\}_{i=1}^{\infty}$ *sont disjoints deux à deux (c'est à dire que,* $\mathcal{A}_i \cap \mathcal{A}_j = \emptyset$*,* $i \neq j$*), alors* P *est* **σ-additive** *:*

$$\mathrm{P}\left(\bigcup_{i=1}^{\infty} \mathcal{A}_i\right) = \sum_{i=1}^{\infty} \mathrm{P}(\mathcal{A}_i).$$

Pause pensées. *Soient* $\Omega = \{0, 1\}$ *et* $\mathcal{F} = \{\emptyset, \{0\}, \{1\}, \Omega\}$*. Est-ce que*

$$\mathrm{P}(\{0\}) = 1/2, \quad \mathrm{P}(\{1\}) = 1/3,$$

donne une loi de probabilité P *valide ? Justifier par rapport aux axiomes.*

Propriétés de P

Théorème 8. *Soient* $\mathcal{A}, \mathcal{B}, \{\mathcal{A}_i\}_{i=1}^{\infty}$ *des événements de l'espace de probabilité* $(\Omega, \mathcal{F}, \mathrm{P})$*. Alors*

(a) $\mathrm{P}(\emptyset) = 0$ *;*

(b) $\mathrm{P}(\mathcal{A}^c) = 1 - \mathrm{P}(\mathcal{A})$ *;*

(c) $\mathrm{P}(\mathcal{A} \cup \mathcal{B}) = \mathrm{P}(\mathcal{A}) + \mathrm{P}(\mathcal{B}) - \mathrm{P}(\mathcal{A} \cap \mathcal{B})$*. Si* $\mathcal{A} \cap \mathcal{B} = \emptyset$*, alors*

$$\mathrm{P}(\mathcal{A} \cup \mathcal{B}) = \mathrm{P}(\mathcal{A}) + \mathrm{P}(\mathcal{B});$$

(d) si $\mathcal{A} \subset \mathcal{B}$*, alors* $\mathrm{P}(\mathcal{A}) \leq \mathrm{P}(\mathcal{B})$ *;*

(e) $\mathrm{P}\left(\bigcup_{i=1}^{\infty} \mathcal{A}_i\right) \leq \sum_{i=1}^{\infty} \mathrm{P}(\mathcal{A}_i)$ *(***inégalité de Boole***);*

(f) une mesure de probabilité est une fonction 'continue', i.e.,

$$\lim_{n\to\infty} \mathrm{P}\left(\bigcup_{i=1}^{n} \mathcal{A}_i\right) = \mathrm{P}\left(\lim_{n\to\infty} \bigcup_{i=1}^{n} \mathcal{A}_i\right), \quad \lim_{n\to\infty} \mathrm{P}\left(\bigcap_{i=1}^{n} \mathcal{A}_i\right) = \mathrm{P}\left(\lim_{n\to\infty} \bigcap_{i=1}^{n} \mathcal{A}_i\right).$$

Note to Theorem 8

(a) Since $\emptyset \cap \mathcal{A} = \emptyset$ for any $\mathcal{A} \in \mathcal{F}$, we can apply $(P3)$ to a finite number of sets, just by including an infinite number of $\emptyset$s. In particular, $\Omega = \Omega \cup \emptyset \cup \emptyset \cup \cdots$, and these are pairwise disjoint, so

$$1 = \mathrm{P}(\Omega) = \mathrm{P}(\Omega) + \mathrm{P}(\emptyset) + \mathrm{P}(\emptyset) + \cdots,$$

so since $\mathrm{P}(\emptyset) \geq 0$, we must have $\mathrm{P}(\emptyset) = 0$.
(b) Follows from $(P3)$ by setting $\mathcal{A}_1 = \mathcal{A}$, $\mathcal{A}_2 = \mathcal{A}^c$, $\mathcal{A}_3 = \mathcal{A}_4 = \cdots = \emptyset$.
(c) Follows from $(P3)$ by writing $\mathcal{A} \cup \mathcal{B} = (\mathcal{A} \cap \mathcal{B}^c) \cup (\mathcal{A} \cap \mathcal{B}) \cup (\mathcal{A}^c \cap \mathcal{B})$, which are pairwise disjoint.
(d) Follows by writing $\mathcal{B} = \mathcal{A} \cup (\mathcal{B} \cap \mathcal{A}^c)$, and noting that $\mathcal{B} \backslash \mathcal{A} = (\mathcal{B} \cap \mathcal{A}^c)$.
(e) Iteration : write $\bigcup_{i=1}^{\infty} \mathcal{A}_i = \mathcal{A}_1 \cup \bigcup_{i=2}^{\infty} \mathcal{A}_i = \mathcal{A}_1 \cup \mathcal{B}$, and note that by (c) we have

$$\mathrm{P}\left(\bigcup_{i=1}^{\infty} \mathcal{A}_i\right) = \mathrm{P}(\mathcal{A}_1 \cup \mathcal{B}) \leq \mathrm{P}(\mathcal{A}_1) + \mathrm{P}(\mathcal{B}) \leq \cdots.$$

(f) Let $\mathcal{B}_n = \bigcup_{i=1}^{n} \mathcal{A}_i$, and note that $\mathcal{B}_n \subset \mathcal{B}_{n+1}$ for every n, so $(\mathcal{B}_{i+1} \backslash \mathcal{B}_i) \cap (\mathcal{B}_{j+1} \backslash \mathcal{B}_j) = \emptyset$ when $i \neq j$ (draw picture). Therefore the $\mathcal{B}_{n+1} \backslash \mathcal{B}_n$ are pairwise disjoint for $n = 1, 2, \ldots$. Hence $(P3)$ gives

$$\begin{aligned}
\mathrm{P}\left(\bigcup_{i=1}^{\infty} \mathcal{A}_i\right) = \mathrm{P}(\mathcal{B}_\infty) &= \mathrm{P}(\mathcal{B}_1) + \sum_{i=2}^{\infty} \mathrm{P}(\mathcal{B}_i \backslash \mathcal{B}_{i-1}) \\
&= \mathrm{P}(\mathcal{B}_1) + \sum_{i=2}^{\infty} \{\mathrm{P}(\mathcal{B}_i) - \mathrm{P}(\mathcal{B}_{i-1})\}, \\
&= \lim_{n \to \infty} \left[\mathrm{P}(\mathcal{B}_1) + \sum_{i=2}^{n} \{\mathrm{P}(\mathcal{B}_i) - \mathrm{P}(\mathcal{B}_{i-1})\}\right], \\
&= \lim_{n \to \infty} \mathrm{P}(\mathcal{B}_n) = \lim_{n \to \infty} \mathrm{P}\left(\bigcup_{i=1}^{n} \mathcal{A}_i\right)
\end{aligned}$$

where we used the fact that $\mathrm{P}(\mathcal{B}_n)$ is increasing and bounded above and therefore has a limit.
The second part is an exercise.

Formules d'inclusion-exclusion

Lemme 9. *Soient $\mathcal{A}_1, \ldots, \mathcal{A}_n$ des événements de $(\Omega, \mathcal{F}, P)$, alors*

$$\begin{aligned}
\mathrm{P}(\mathcal{A}_1 \cup \mathcal{A}_2) &= \mathrm{P}(\mathcal{A}_1) + \mathrm{P}(\mathcal{A}_2) - \mathrm{P}(\mathcal{A}_1 \cap \mathcal{A}_2), \\
\mathrm{P}(\mathcal{A}_1 \cup \mathcal{A}_2 \cup \mathcal{A}_3) &= \mathrm{P}(\mathcal{A}_1) + \mathrm{P}(\mathcal{A}_2) + \mathrm{P}(\mathcal{A}_3) \\
&\quad - \mathrm{P}(\mathcal{A}_1 \cap \mathcal{A}_2) - \mathrm{P}(\mathcal{A}_1 \cap \mathcal{A}_3) - \mathrm{P}(\mathcal{A}_2 \cap \mathcal{A}_3) \\
&\quad + \mathrm{P}(\mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3), \\
&\vdots \\
\mathrm{P}\left(\bigcup_{i=1}^{n} \mathcal{A}_i\right) &= \sum_{r=1}^{n} (-1)^{r+1} \sum_{1 \leq i_1 < \cdots < i_r \leq n} \mathrm{P}(\mathcal{A}_{i_1} \cap \cdots \cap \mathcal{A}_{i_r}).
\end{aligned}$$

Pause pensées. *Combien de termes y a-t-il dans la formule générale ci-dessus ?*

Note to Lemma 9

We saw the first equality as part (c) of Theorem 8.
For the second, write $B = \mathcal{A}_2 \cup \mathcal{A}_3$, and note that

$$\begin{aligned} \mathrm{P}(\mathcal{A}_1 \cup \mathcal{A}_2 \cup \mathcal{A}_3) &= \mathrm{P}(\mathcal{A}_1) + \mathrm{P}(\mathcal{A}_2 \cup \mathcal{A}_3) - \mathrm{P}\{\mathcal{A}_1 \cap (\mathcal{A}_2 \cup \mathcal{A}_3)\} \\ &= \mathrm{P}(\mathcal{A}_1) + \mathrm{P}(\mathcal{A}_2 \cup \mathcal{A}_3) - \mathrm{P}\{(\mathcal{A}_1 \cap \mathcal{A}_2) \cup (\mathcal{A}_1 \cap \mathcal{A}_3)\} \\ &= \mathrm{P}(\mathcal{A}_1) + \mathrm{P}(\mathcal{A}_2) + \mathrm{P}(\mathcal{A}_3) - \mathrm{P}(\mathcal{A}_2 \cap \mathcal{A}_3) \\ &\quad -\mathrm{P}(\mathcal{A}_1 \cap \mathcal{A}_2) - \mathrm{P}(\mathcal{A}_1 \cap \mathcal{A}_3) + \mathrm{P}\{(\mathcal{A}_1 \cap \mathcal{A}_2) \cap (\mathcal{A}_1 \cap \mathcal{A}_3)\} \end{aligned}$$

which is what we want, since the last term is $\mathrm{P}(\mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3)$. The general formula follows by iterating this argument.

Exemple

Exemple 10. *Une urne contient 1000 tickets de loterie numérotés de 1 à 1000. On tire un ticket au hasard. Auparavant un artiste de foire a offert de payer \$3 à quiconque qui lui donne \$2, si le numéro du ticket est divisible par 2, 3, ou 5. Est ce que vous lui donneriez vos \$2 avant le tirage ? (Vous perdez votre argent si le ticket n'est pas divisible par 2, 3, ou 5.)*

Note to Example 10

Here we can write $\Omega = \{1, \ldots, 1000\}$, and let D_i be the event that the number is divisible by i. We want

$$\begin{aligned} \mathrm{P}(D_2 \cup D_3 \cup D_5) &= \mathrm{P}(D_2) + \mathrm{P}(D_3) + \mathrm{P}(D_5) - \mathrm{P}(D_2 \cap D_3) - \mathrm{P}(D_2 \cap D_5) - \mathrm{P}(D_3 \cap D_5) \\ &\quad +\mathrm{P}(D_2 \cap D_3 \cap D_5) \\ &= \mathrm{P}(D_2) + \mathrm{P}(D_3) + \mathrm{P}(D_5) - \mathrm{P}(D_6) - \mathrm{P}(D_{10}) - \mathrm{P}(D_{15}) + \mathrm{P}(D_{30}) \\ &= \frac{500 + 333 + 200 - 166 - 100 - 66 + 33}{1000} = \frac{367}{500} \doteq 0.734. \end{aligned}$$

So with probability 0.734 you gain 1 and with probability 0.266 you lose 2 : the average gain is $1 \times 0.734 + (-2) \times 0.266 = 0.202$: you will win on average if you play. The 'return on investment' is $0.202/2 \approx 0.10$, or 10%, which is very good compared to a bank.

2.2 Un Peu de Combinatoire

Motivation

— Souvent, on doit compter des cas. Comment ?

— Deux principes de base :

  — **addition** — si j'ai n chats blancs et m chats noirs, j'ai un total de $n+m$ chats ;

  — **multiplication** — si j'ai n chats et m chiens, il y a nm façons de promener un chat avec un chien.

— Stratégie : il n'y a pas vraiment de stratégie générale. Il faut connaître des exemples et s'adapter à la situation, essayer avec de petites valeurs pour comprendre.

— En combinatoire il y a beaucoup de factorielles, en particulier dans le nombre de manières d'arranger k objets identiques dans n boîtes,

$$C_n^k = \binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n(n-1)\cdots(n-k+1)}{k!}.$$

Exemple 11. *Quelle est la probabilité qu'il n'y a pas de coincidences d'anniversaires dans une classe de n étudiants (éliminant les hypothèses d'années bissextiles, des jumeaux, etc.) ?*

Note to Example 11

On a $\Omega = \{(i_1, \ldots, i_n) : i_1, \ldots, i_n \in \{1, \ldots, 365\}\}$, avec $|\Omega| = 365^n$, et

$$\mathcal{A}_n = \{(i_1, \ldots, i_n\} : \text{ tous les } i_j \text{ sont distincts } \}.$$

Ainsi

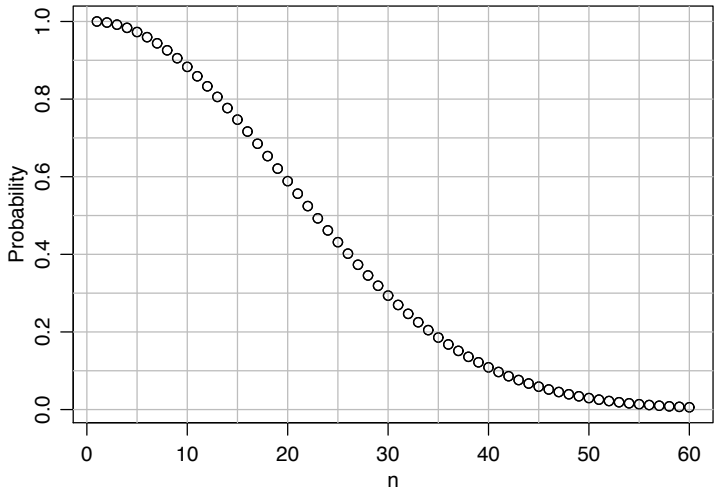
$$|\mathcal{A}_n| = 365 \times \cdots \times (365 - n + 1) = \frac{365!}{(365-n)!},$$

et

$$\mathrm{P}(\mathcal{A}_n) = \frac{|\mathcal{A}_n|}{|\Omega|} = \frac{365!}{(365-n)!} \div 365^n.$$

Le graphique montre les probabilités des $\mathcal{A}_n$ pour $n = 1, \ldots, 60$.

Anniversaires



Suite harmonique

Lemme 12. *La* **suite harmonique** *est*

$$S_n = 1 + \frac{1}{2} + \cdots + \frac{1}{n}.$$

Quand $n \to \infty$, *on a* $S_n = \ln n + \gamma + o(1)$, *où* γ *est une constante.*

Pour montrer que $S_n - \ln n$ converge, on peut montrer que c'est une suite monotone bornée en la représentant comme

$$\int_1^n \left(\frac{1}{\lfloor x \rfloor} - \frac{1}{x} \right) \mathrm{d}x.$$

Exemple 13. *Une princesse a un nombre* n *assez grand de prétendants, parmi lesquels se trouve un prince charmant. Elle les rencontre dans sa vie dans un ordre aléatoire. A chaque fois qu'elle rencontre un prétendant, elle peut soit l'épouser, soit passer au suivant. Elle peut le classer parmi les précédents. Quelle est la stratégie qui lui permet de maximiser sa probabilité d'épouser le prince charmant ?*

Note to Example 13

— L'espace Ω contient des $n!$ permutations de $\{1, \ldots, n\}$ possibles (les rangs absolus des prétendants, alors que la princesse peut ordonner seulement ceux qu'elle a déjà vu).

— La princesse décide de rejeter les k premiers prétendants et d'épouser ensuite le premier qui sera meilleur que tous les précédents. Soit C l'événement que ceci est le prince charmant. Alors $C = \bigcup_{i=k+1}^{n} C_i$, oú C_i est l'événement qu'elle choisi le prince charmant et qu'il soit à la place i dans l'ordre ; $\mathrm{P}(C) = \sum_{i=k+1}^{n} \mathrm{P}(C_i)$, car les C_i sont disjoints. De plus, $C_i = \bigcup_{j=1}^{k} C_{i,j}$, où $C_{i,j}$ est l'événement que le prince charmant soit à la place i et que le meilleur avant lui soit à la place j. Puisque les $C_{i,j}$ sont aussi disjoints, le **principe d'addition** donne

$$\mathrm{P}(C) = \sum_{i=k+1}^{n} \mathrm{P}(C_i) = \sum_{i=k+1}^{n} \sum_{j=1}^{k} \mathrm{P}(C_{i,j}).$$

— Combien des $n!$ permutations possibles donnent $C_{i,j}$? Il y a une façon de mettre le prince charmant à la place i, puis il y a $n-i$ places après lui que l'on peut remplir avec un choix parmi $n-1$ grenouilles ; il y a $(n-1)(n-2)\cdots\{n-1-(n-i-1)\} = (n-1)!/(i-1)!$ façons de ce faire. Pour les $i-1$ places avant le prince charmant, on veut que le meilleur grenouille soit à la place j, et que les $i-2$ autres soient dans n'importe quel ordre. Par le **principe de multiplication**,

$$1 \times \frac{(n-1)!}{(i-1)!} \times 1 \times (i-2)! = \frac{(n-1)!}{i-1}$$

permutations donnent $C_{i,j}$. Ainsi $\mathrm{P}(C_{i,j}) = \{(n-1)!/(i-1)\}/n! = 1/\{n(i-1)\}$, et

$$\mathrm{P}(C) = \sum_{i=k+1}^{n} \sum_{j=1}^{k} \frac{1}{n(i-1)} = \frac{k}{n} \sum_{i=k+1}^{n} \frac{1}{i-1} = \frac{k}{n}(S_{n-1} - S_{k-1}) \underset{k,n \text{ grands}}{\approx} \frac{k}{n} \ln(n/k).$$

— Avec cette stratégie, $\mathrm{P}(C) \approx -x \ln x$, avec $x = k/n$. La fonction $-x \ln x$ est maximisée pour $x = 1/e$, et alors $k \approx ne^{-1} \approx 0.368n$ et $\mathrm{P}(C) \approx e^{-1} \approx 0.368$.

2.3 Probabilité Conditionnelle

Probabilité conditionnelle

L'idée centrale du conditionnement est l'effet d'un fait sur une probabilité : comment change $\mathrm{P}(\mathcal{A})$ si je sais que $\mathcal{B}$ s'est passé ?

Définition 14. *Soient $\mathcal{A}, \mathcal{B}$ des événements de l'espace de probabilité $(\Omega, \mathcal{F}, \mathrm{P})$, tel que $\mathrm{P}(\mathcal{B}) > 0$. Alors la* **probabilité conditionnelle de $\mathcal{A}$ sachant $\mathcal{B}$** *est*

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

Si $\mathrm{P}(\mathcal{B}) = 0$, on adopte la convention $\mathrm{P}(\mathcal{A} \cap \mathcal{B}) = \mathrm{P}(\mathcal{A} \mid \mathcal{B})\mathrm{P}(\mathcal{B})$, des deux côtés on a la valeur zéro. Ainsi

$$\mathrm{P}(\mathcal{A}) = \mathrm{P}(\mathcal{A} \cap \mathcal{B}) + \mathrm{P}(\mathcal{A} \cap \mathcal{B}^c) = \mathrm{P}(\mathcal{A} \mid \mathcal{B})\mathrm{P}(\mathcal{B}) + \mathrm{P}(\mathcal{A} \mid \mathcal{B}^c)\mathrm{P}(\mathcal{B}^c)$$

même si $\mathrm{P}(\mathcal{B}) = 0$ ou $\mathrm{P}(\mathcal{B}^c) = 0$.

Exemple

Exemple 15. *On lance deux dés équilibrés, un rouge et un vert. Soient $\mathcal{A}$ et $\mathcal{B}$ les événements 'le total excède 8', et 'on a 6 sur le dé rouge'. Si on sait que $\mathcal{B}$ s'est produit, comment change $\mathrm{P}(\mathcal{A})$?*

Note to Example 15

Here $\Omega = \{(r, g) : r, g = 1, \ldots, 6\}$, and the outcomes (r, g) are equiprobable by symmetry. Now

$$\begin{aligned} \mathcal{A} &= \{(3,6),(4,6),(5,6),(6,6),(4,5),(5,5),(6,5),(5,4),(6,4),(6,3)\}, \\ \mathcal{B} &= \{(6,1),\ldots,(6,6)\}, \\ \mathcal{A}\cap\mathcal{B} &= \{(6,3),(6,4),(6,5),(6,6)\}, \end{aligned}$$

so $\mathrm{P}(\mathcal{A}) = 10/36$, $\mathrm{P}(\mathcal{B}) = 6/36$, $\mathrm{P}(\mathcal{A}\cap\mathcal{B}) = 4/36$, and

$$\mathrm{P}(\mathcal{A} \mid \mathcal{B}) = \frac{\mathrm{P}(\mathcal{A}\cap\mathcal{B})}{\mathrm{P}(\mathcal{B})} = \frac{4/36}{6/36} = 2/3.$$

Thus knowledge that $\mathcal{B}$ has occurred more than doubles the probability of $\mathcal{A}$.

Lois de probabilité conditionnelle

Théorème 16. *Soit $(\Omega, \mathcal{F}, \mathrm{P})$ un espace de probabilité, et soient $\mathcal{B} \in \mathcal{F}$ tel que $\mathrm{P}(\mathcal{B}) > 0$ et $Q(\mathcal{A}) = \mathrm{P}(\mathcal{A} \mid \mathcal{B})$. Alors $(\Omega, \mathcal{F}, Q)$ est un espace de probabilité. En particulier,*

(a) *si $\mathcal{A} \in \mathcal{F}$, alors $0 \le Q(\mathcal{A}) \le 1$;*

(b) $Q(\Omega) = 1$;

(c) *si $\{\mathcal{A}_i\}_{i=1}^{\infty}$ sont disjoints 2 à 2, alors*

$$Q\left(\bigcup_{i=1}^{\infty} \mathcal{A}_i\right) = \sum_{j=1}^{\infty} Q(\mathcal{A}_i).$$

Ainsi le conditionnement nous permet de construire beaucoup de lois de probabilités différentes, à partir d'une loi de probabilité donnée.

Note to Theorem 16

We just need to check the axioms. If $\mathcal{A} \in \mathcal{F}$, then

$$Q(\mathcal{A}) = \mathrm{P}(\mathcal{A} \mid \mathcal{B}) = \mathrm{P}(\mathcal{A} \cap \mathcal{B})/\mathrm{P}(\mathcal{B}) \in [0, 1],$$

because $\mathcal{A} \cap \mathcal{B} \subset \mathcal{B}$ and therefore $\mathrm{P}(\mathcal{A} \cap \mathcal{B}) \leq \mathrm{P}(\mathcal{B})$. Likewise

$$Q(\Omega) = \mathrm{P}(\Omega \cap \mathcal{B})/\mathrm{P}(\mathcal{B}) = \mathrm{P}(\mathcal{B})/\mathrm{P}(\mathcal{B}) = 1,$$

and finally,

$$Q\left(\bigcup_{i=1}^{\infty} \mathcal{A}_i\right) = \frac{\mathrm{P}(\bigcup_{i=1}^{\infty} \mathcal{A}_i \cap \mathcal{B})}{\mathrm{P}(\mathcal{B})} = \frac{\mathrm{P}\{\bigcup_{i=1}^{\infty} (\mathcal{A}_i \cap \mathcal{B})\}}{\mathrm{P}(\mathcal{B})} = \frac{\sum_{i=1}^{\infty} \mathrm{P}(\mathcal{A}_i \cap \mathcal{B})}{\mathrm{P}(\mathcal{B})} = \sum_{i=1}^{\infty} Q(\mathcal{A}_i),$$

using the properties of $\mathrm{P}(\cdot)$ and the fact that if $\mathcal{A}_1, \mathcal{A}_2, \ldots$ are pairwise disjoint, so are the $\mathcal{A}_j \cap \mathcal{B}, \ldots$.

Thomas Bayes (1702–1761)

Essay towards solving a problem in the doctrine of chances. (1763/4) Philosophical Transactions of the Royal Society of London.

Théorème de Bayes

Définition 17. *Une* **partition** $\mathcal{B}_1, \ldots, \mathcal{B}_K$ *d'un ensemble* Ω *a les propriétés suivantes :*

$$\bigcup_{k=1}^{K} \mathcal{B}_k = \Omega \quad \text{et} \quad \mathcal{B}_k \cap \mathcal{B}_{k'} = \emptyset \text{ pour tout } k \neq k',$$

c'est à dire les $\mathcal{B}_k$ *couvrent* Ω *et sont disjoints 2 à 2. Il est possible que* $K = \infty$.

Théorème 18 (Loi des probabilités totales). *Soient* $(\Omega, \mathcal{F}, \mathrm{P})$ *un espace de probabilité,* $\{\mathcal{B}_k\}_{k=1}^{K}$ *une partition de* Ω*, et* $\mathcal{A}, \mathcal{B}_1, \ldots, \mathcal{B}_K \in \mathcal{F}$*, alors*

$$\mathrm{P}(\mathcal{A}) = \sum_{k=1}^{K} \mathrm{P}(\mathcal{A} \cap \mathcal{B}_k) = \sum_{k=1}^{K} \mathrm{P}(\mathcal{A} \mid \mathcal{B}_k)\mathrm{P}(\mathcal{B}_k).$$

Théorème 19 (Bayes). *Supposons que les conditions ci-dessus soient vérifiées, et que* $\mathrm{P}(\mathcal{A}) > 0$*. Alors*

$$\mathrm{P}(\mathcal{B}_j \mid \mathcal{A}) = \frac{\mathrm{P}(\mathcal{A} \mid \mathcal{B}_j)\mathrm{P}(\mathcal{B}_j)}{\sum_{k=1}^{K} \mathrm{P}(\mathcal{A} \mid \mathcal{B}_k)\mathrm{P}(\mathcal{B}_k)}, \quad j = 1, \ldots, K.$$

Ces résultats sont aussi vrais si les $\mathcal{B}_k$ sont disjoints 2 à 2 et $\mathcal{A} \subset \bigcup_{k=1}^{K} \mathcal{B}_k \neq \Omega$.

Note to Theorems 18 and 19

Since the $\mathcal{B}_k$ are disjoint, so are the $\mathcal{A} \cap \mathcal{B}_k$. Thus

$$\mathrm{P}(\mathcal{A}) = \mathrm{P}\left\{\mathcal{A} \cap \bigcup_{k=1}^{K} \mathcal{B}_i\right\} = \mathrm{P}\left\{\bigcup_{k=1}^{K} (\mathcal{A} \cap \mathcal{B}_k)\right\} = \sum_{k=1}^{K} \mathrm{P}(\mathcal{A} \cap \mathcal{B}_k) = \sum_{k=1}^{K} \mathrm{P}(\mathcal{A} \mid \mathcal{B}_k)\mathrm{P}(\mathcal{B}_k).$$

For Bayes' theorem, we note that

$$\mathrm{P}(\mathcal{B}_j \mid \mathcal{A}) = \frac{\mathrm{P}(\mathcal{A} \cap \mathcal{B}_j)}{\mathrm{P}(\mathcal{A})} = \frac{\mathrm{P}(\mathcal{A} \mid \mathcal{B}_j)\mathrm{P}(\mathcal{B}_j)}{\mathrm{P}(\mathcal{A})} = \frac{\mathrm{P}(\mathcal{A} \mid \mathcal{B}_j)\mathrm{P}(\mathcal{B}_j)}{\sum_{k=1}^{K} \mathrm{P}(\mathcal{A} \mid \mathcal{B}_k)\mathrm{P}(\mathcal{B}_k)}$$

using Theorem 18.

Exemples

Exemple 20. *Pour dépister une maladie, on applique un test. Si le patient est effectivement atteint, le test donne un résultat positif dans* 99% *des cas. Mais il se peut aussi que le résultat du test soit positif alors que le patient est en bonne santé, et ceci se produit dans* 2% *des cas. Sachant qu'en moyenne un patient sur* 1000 *est atteint de la maladie, calculer la probabilité qu'un patient soit atteint sachant que son test a été positif. Comment améliorer ce resultat ?*

Note to Example 20

— Soit M l'événement "le patient est atteint", M^c l'événement complémentaire, et $\mathcal{A}$ l'événement "le résultat du test est positif". De la formule de Bayes on a

$$\begin{aligned} \mathrm{P}(M \mid \mathcal{A}) &= \frac{\mathrm{P}(\mathcal{A} \mid M)\mathrm{P}(M)}{\mathrm{P}(\mathcal{A} \mid M)\mathrm{P}(M) + \mathrm{P}(\mathcal{A} \mid M^c)\mathrm{P}(M^c)} \\ &= \frac{\frac{99}{100}\frac{1}{1000}}{\frac{99}{100}\frac{1}{1000} + \frac{2}{100}\frac{999}{1000}} \\ &= \frac{99}{2097} \approx 0.0472. \end{aligned}$$

— Ceci n'est pas très concluant. Pourquoi ? Si on test 100100 personnes dont 100 atteint de la maladie, on aura en moyenne $100000 \times 0.02 = 2000$ 'faux positifs' et 99 'vrais positifs'. Donc, des 2099 personnes avec des resultats positifs, la proportion de malades est $99/2099 = 0.0472 \approx 5\%$.

— Pour essayer de l'améliorer on répète le test. Pour $m \leq n$ soit $\mathcal{A}_m^n$ l'événement "lors de n tests indépendants sur le même patient, m ont donné un résultat positif". Alors, on a

$$\mathrm{P}(\mathcal{A}_m^n \mid M^c) = \binom{n}{m}\left(\frac{2}{100}\right)^m\left(\frac{98}{100}\right)^{n-m}, \quad \mathrm{P}(\mathcal{A}_m^n \mid M) = \binom{n}{m}\left(\frac{99}{100}\right)^m\left(\frac{1}{100}\right)^{n-m}.$$

Donc la probabilité que la patient soit atteint sachant $\mathcal{A}_m^n$ est

$$\begin{aligned} \mathrm{P}(M \mid \mathcal{A}_m^n) &= \frac{\mathrm{P}(\mathcal{A}_m^n \mid M)\mathrm{P}(M)}{\mathrm{P}(\mathcal{A}_m^n \mid M)\mathrm{P}(M) + \mathrm{P}(\mathcal{A}_m^n \mid M^c)\mathrm{P}(M^c)} \\ &= \frac{99^m}{99^m + 2^m 98^{n-m} \times 999} \\ &= \frac{1}{1 + 999 \times 98^{n-m}(2/99)^m}. \end{aligned}$$

De cette formule on peut conclure que par exemple $\mathrm{P}(M \mid \mathcal{A}_2^2) \approx 0.710$ et $\mathrm{P}(M \mid \mathcal{A}_3^3) \approx 0.992$.

Commentaires

— La formule de Bayes est très simple mais très utile, car elle permet une 'inversion du point de vue' dont on a souvent besoin en pratique.

— Parfois, quand on essaie de calculer les probabilités relatives des A_k sachant B, on peut utiliser

$$\mathrm{P}\left(\mathcal{A}_k \mid \mathcal{B}\right) \propto \mathrm{P}\left(\mathcal{B} \mid \mathcal{A}_k\right)\mathrm{P}\left(\mathcal{A}_k\right),$$

ce qui évite de calculer le dénominateur (qui peut être compliqué à calculer).

— Par exemple, si on un a priori sur diverses hypothèses H' possible, on met des probabilités sur les H' selon leur plausibilité a priori, et on fait une expérience dont les probabilités dépendent des données D de manière connue $\mathrm{P}\left(D \mid H'\right)$, on a

$$\mathrm{P}\left(H \mid D\right) = \frac{\mathrm{P}\left(D \mid H\right)\mathrm{P}\left(H\right)}{\mathrm{P}\left(D \mid H\right)\mathrm{P}\left(H\right) + \sum_{H' \neq H}\mathrm{P}\left(D \mid H'\right)\mathrm{P}\left(H'\right)}.$$

Conditionnement multiple

Lemme 21 ('Prediction decomposition'). *Soient $A_1, \ldots, A_n$ des événements d'un espace de probabilité. Alors*

$$\begin{aligned} \mathrm{P}(A_1 \cap A_2) &= \mathrm{P}(A_2 \mid A_1)\mathrm{P}(A_1) \\ \mathrm{P}(A_1 \cap A_2 \cap A_3) &= \mathrm{P}(A_3 \mid A_1 \cap A_2)\mathrm{P}(A_2 \mid A_1)\mathrm{P}(A_1) \\ &\vdots \\ \mathrm{P}(A_1 \cap \cdots \cap A_n) &= \prod_{i=2}^{n} \mathrm{P}(A_i \mid A_1 \cap \cdots \cap A_{i-1}) \times \mathrm{P}(A_1) \end{aligned}$$

Note to Lemma 21

Just iterate. For example, if we let $B = A_1 \cap A_2$ and note that $\mathrm{P}(B) = \mathrm{P}(A_2 \mid A_1)\mathrm{P}(A_1)$ by the definition of conditional probability, then

$$\mathrm{P}(A_1 \cap A_2 \cap A_3) = \mathrm{P}(A_3 \cap B) = \mathrm{P}(A_3 \mid B)\mathrm{P}(B) = \mathrm{P}(A_3 \mid A_1 \cap A_2)\mathrm{P}(A_2 \mid A_1)\mathrm{P}(A_1),$$

on using the definition of conditional probability, twice. For the general case, just extend this idea, by setting

$$\begin{aligned} \mathrm{P}(A_1 \cap \cdots \cap A_n) &= \mathrm{P}(A_n \mid A_1 \cap \cdots \cap A_{n-1})\mathrm{P}(A_1 \cap \cdots \cap A_{n-1}) \\ &= \mathrm{P}(A_n \mid A_1 \cap \cdots \cap A_{n-1})\mathrm{P}(A_{n-1} \mid A_1 \cap \cdots \cap A_{n-2})\mathrm{P}(A_1 \cap \cdots \cap A_{n-2}) \\ &\vdots \\ &= \prod_{i=2}^{n} \mathrm{P}(A_i \mid A_1 \cap \cdots \cap A_{i-1}) \times \mathrm{P}(A_1), \end{aligned}$$

as required.

'Matchings'

Exemple 22. *n hommes vont à un diner. Chacun laisse son chapeau au vestiaire. Lorsqu'ils repartent, ayant bien échantillioné du vin régional, ils choisissent leurs chapeaux de façon aléatoire.*
(a) Quelle est la probabilité que personne n'ait son chapeau ?
(b) Quelle est la probabilité qu'exactement r hommes choisissent leur propre chapeau ?
(c) Que se passe-t-il lorsque n est très grand ?

Note to Example 22

— This is an example of many types of matching problem, going back to Montmort (1708).

— The sample space here is the set Ω containing the $n!$ permutations of the numbers $\{1, \ldots, n\}$.

— Let A_i denote the event that the ith hat is on the ith head, and note that $\mathrm{P}(A_i) = 1/n$,

$$\mathrm{P}(A_i \cap A_j) = \mathrm{P}(A_i \mid A_j)\mathrm{P}(A_j) = \frac{1}{n-1} \times \frac{1}{n}, \ldots, \mathrm{P}(A_1 \cap \cdots \cap A_r) = \frac{(n-r)!}{n!},$$

using the prediction decomposition. (Equivalently, we can use the multiplication principle.) Thus the probability that r specific men have their hats is $(n-r)!/n!$.

— (a) Let $p_n(k)$ denote the probability that exactly k out of n men get the right hat, for $k = 0, \ldots, n$. We want

$$p_n(0) = \mathrm{P}(A_1^c \cap \cdots \cap A_n^c) = 1 - \mathrm{P}(A_1 \cup \cdots \cup A_n),$$

so we use the inclusion-exclusion formula to compute

$$\begin{aligned} \mathrm{P}(A_1 \cup \cdots \cup A_n) &= \sum_{r=1}^{n} (-1)^{r+1} \sum_{1 \le i_1 < \cdots < i_r \le n} \mathrm{P}(A_{i_1} \cap \cdots \cap A_{i_r}) \\ &= n \times n^{-1} - \binom{n}{2} \times \frac{(n-2)!}{n!} + \cdots + (-1)^{n+1} \times \binom{n}{n} \times \frac{(n-n)!}{n!} \\ &= \sum_{i=1}^{n} (-1)^{i+1}/i!. \end{aligned}$$

Hence $p_n(0) = 1 - \sum_{i=1}^{n} (-1)^{i+1}/i! = \sum_{i=0}^{n} (-1)^i/i!$.

— (b) The probability that men $1, \ldots, r$ have the right hats and no-one else does is

$$\begin{aligned} \mathrm{P}(A_1 \cap \cdots \cap A_r \cap A_{r+1}^c \cap \cdots \cap A_n^c) &= \mathrm{P}(A_1 \cap \cdots \cap A_r) \times \mathrm{P}(A_{r+1}^c \cap \cdots \cap A_n^c \mid A_1 \cap \cdots \cap A_r) \\ &= \frac{(n-r)!}{n!} \times p_{n-r}(0) \\ &= \frac{(n-r)!}{n!} \times \sum_{i=0}^{n-r} (-1)^i/i!, \end{aligned}$$

but since there are $\binom{n}{r}$ distinct ways of choosing r men to have the right hats from the total n men available, the overall probability (using the addition principle) is

$$p_n(r) = \frac{n!}{r!(n-r)!} \times \frac{(n-r)!}{n!} \times \sum_{i=0}^{n-r} (-1)^i/i! = \frac{1}{r!} \times \sum_{i=0}^{n-r} (-1)^i/i!.$$

— (c) Since $\sum_{i=0}^{k} (-1)^i/i! \to e^{-1}$ as $k \to \infty$, we have

$$\lim_{n \to \infty} p_n(r) = \frac{1}{r!} e^{-1}, \quad r \in \{0, 1, 2, \ldots\}.$$

Utilisation des partitions

La loi des probabilités totales permet d'attaquer aux problèmes complexes par un choix astucieux de partition.

Exemple 23. *Pour un modèle simple des journées pluvieuses, supposons que s'il fait beau aujourd'hui, alors la probabilité qu'il pleuvra demain est p, et que s'il pleut aujourd'hui, il fera beau demain avec probabilité q.*
(a) S'il pleut aujourd'hui, trouver la probabilité qu'il va encore pleuvoir pendant k jours.
(b) S'il pleut aujourd'hui, trouver la probabilité qu'il fera beau en k jours.

Note to Example 23

— Denote a rainy day by 1, and a dry day by 0, so $\Omega = \{\omega = x_1 x_2 \cdots : x_j \in \{0,1\}\}$, and define the events $D_k = \{\omega : x_k = 1\}$ and $W_k = \{\omega : x_k = 0\}$ for $k = 1, 2, \ldots$. We have $\mathrm{P}(W_k \mid D_{k-1}) = p$ and $\mathrm{P}(D_k \mid W_{k-1}) = q$, for each k.

— (a) Find $q_k = \mathrm{P}(W_k \cap \cdots \cap W_1 \mid W_0)$. The prediction decomposition (Lemma 21) gives

$$q_k = \mathrm{P}(W_1 \cap \cdots \cap W_k \mid W_0) = \prod_{i=2}^{k} \mathrm{P}(W_i \mid W_0 \cap W_1 \cap \cdots \cap W_{i-1}) \times \mathrm{P}(W_1 \mid W_0),$$

and

$$\mathrm{P}(W_i \mid W_0 \cap W_1 \cap \cdots \cap W_{i-1}) = \mathrm{P}(W_i \mid W_{i-1}) = 1 - q, \quad \mathrm{P}(W_1 \mid W_0) = 1 - q,$$

so $q_k = (1-q)^k$, for $k = 1, 2, 3, \ldots$.

— (b) Find $p_k = \mathrm{P}(D_k \mid W_0)$. Obviously $p_0 = 0$ and $p_1 = q$. The law of total probability using the partition D_1, W_1 gives

$$\begin{aligned} p_2 &= \mathrm{P}(D_2 \mid W_0) \\ &= \mathrm{P}(D_2 \mid W_1 \cap W_0)\mathrm{P}(W_1 \mid W_0) + \mathrm{P}(D_2 \mid D_1 \cap W_0)\mathrm{P}(D_1 \mid W_0) \\ &= q(1-q) + (1-p)q = q(2-p-q). \end{aligned}$$

We could proceed similarly to get p_3 etc., but this would be painful.

— For a general formula, we use the partition D_{k-1}, W_{k-1} and condition on the state of the $(k-1)$th day. This gives

$$\begin{aligned} p_k &= \mathrm{P}(D_k \mid W_0) \\ &= \mathrm{P}(D_k \mid W_{k-1} \cap W_0)\mathrm{P}(W_{k-1} \mid W_0) + \mathrm{P}(D_k \mid D_{k-1} \cap W_0)\mathrm{P}(D_{k-1} \mid W_0) \\ &= q(1-p_{k-1}) + (1-p)p_{k-1} \\ &= q + (1-p-q)p_{k-1}, \quad k = 2, \ldots. \end{aligned}$$

Multiply by u^k with $|u| < 1$ and sum over k to get a probability generating function

$$\begin{aligned} G(u) &= \sum_{k=1}^{\infty} u^k p_k \\ &= q\sum_{k=1}^{\infty} u^k + \sum_{k=1}^{\infty}(1-p-q)p_{k-1}u^k \\ &= \frac{qu}{1-u} + (1-p-q)uG(u), \end{aligned}$$

using the fact that $p_0 = 0$, giving

$$G(u) = \frac{qu}{(1-u)\{1-(1-p-q)u\}} = q \times \sum_{i=1}^{\infty} u^i \times \sum_{j=1}^{\infty} u^j (1-p-q)^j,$$

which converges for small enough u. For $k = 1, 2, \ldots$, the coefficient of u^k in $G(u)$ is

$$p_k = q\sum_{j=0}^{k-1}(1-p-q)^j = \frac{q\{1-(1-p-q)^k\}}{1-(1-p-q)} = \frac{q\{1-(1-p-q)^k\}}{p+q}.$$

This equals q when $k = 1$, $q(2-p-q)$ when $k = 2$, and tends to $q/(p+q)$ as $k \to \infty$.

2.4 Indépendance

Indépendance

Intuitivement, dire que 'A et B sont indépendants' signifie que la réalisation d'un des deux n'affecte pas la réalisation de l'autre. C'est à dire que, $\mathrm{P}(A \mid B) = \mathrm{P}(A)$, donc la connaissance de la réalisation de B laisse $\mathrm{P}(A)$ inchangée.

Définition 24. *Soit $(\Omega, \mathcal{F}, \mathrm{P})$ un espace de probabilité. Deux événements $A, B \in \mathcal{F}$ sont* **indépendants** *(on écrit $A \perp\!\!\!\perp B$) ssi*

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

Conformément à notre intuition, cela implique que

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

et par symétrie $\mathrm{P}(B \mid A) = \mathrm{P}(B)$.

Exemple 25. *Une famille a deux enfants.*
(a) On sait que le premier est un garçon. Donner la probabilité que le second soit un garçon.
(b) On sait qu'un des deux est un garçon. Donner la probabilité que l'autre soit un garçon.

Note to Example 25

The sample space can be written as $\Omega = \{BB, BG, GB, GG\}$, in an obvious notation, and the events that 'the ith child is a boy' are $B_1 = \{BB, BG\}$ and $B_2 = \{BB, GB\}$. Then

— (a) $\mathrm{P}(B_2 \mid B_1) = \mathrm{P}(B_1 \cap B_2)/\mathrm{P}(B_2) = \mathrm{P}(\{BB\})/\mathrm{P}(B_1) = 1/4 \div 1/2 = 1/2 = \mathrm{P}(B_2)$. Thus B_2 and B_1 are independent.

— (b) the event 'at least one child is a boy' is $C = B_1 \cup B_2 = \{BB, BG, GB\}$, and the event 'two boys' is $D = \{BB\}$, so now we seek $\mathrm{P}(D \mid C) = 1/4 \div 3/4 = 1/3 \neq \mathrm{P}(D)$. Thus D and C are not independent.

Note the importance of precise language : in (a) we know that a specific child is a boy, and in (b) we are told only that one of the two children is a boy. These different pieces of information change the probabilities, because the conditioning event is not the same.

Types d'indépendances

Définition 26. *(a) Les événements $A_1, \ldots, A_n$ sont* **(mutuellement) indépendants** *si pour tout ensemble d'indices $F \subset \{1, \ldots, n\}$, on a*

$$\mathrm{P}\left(\bigcap_{i \in F} A_i\right) = \prod_{i \in F} \mathrm{P}(A_i).$$

(b) Les événements $A_1, \ldots, A_n$ sont **indépendants 2 à 2** *si*

$$\mathrm{P}(A_i \cap A_j) = \mathrm{P}(A_i)\ \mathrm{P}(A_j), \quad 1 \leq i < j \leq n.$$

(c) Les événements $A_1, \ldots, A_n$ sont **conditionnellement indépendants sachant B** *si pour tout ensemble d'indices $F \subset \{1, \ldots, n\}$ on a*

$$\mathrm{P}\left(\bigcap_{i \in F} A_i \mid B\right) = \prod_{i \in F} \mathrm{P}(A_i \mid B).$$

Exemple

Exemple 27. *Une année donnée, la probabilité qu'un conducteur fasse une déclaration de sinistre à son assurance est μ, indépendamment des autres années. La probabilité pour une conductrice est de $\lambda < \mu$. Un assureur a le même nombre de conducteurs que de conductrices, et en sélectionne un(e) au hasard.*
(a) Donner la probabilité qu'il (elle) déclare un sinistre cette année ?
(b) Donner la probabilité qu'il (elle) déclare des sinistres durant 2 années consécutives ?
(c) Si la compagnie sélectionne une personne ayant fait une déclaration au hasard, donner la probabilité qu'elle fasse une déclaration l'année suivante ?
(d) Montrer que la connaissance qu'une déclaration de sinistre ait été faite une année augmente la probabilité d'une déclaration l'année suivante.

Note to Example 27

Let A_r denote the event that the selected driver has an accidents in r successive years, and M denote the event that this driver is male.
(a) Here the law of total probability gives

$$\mathrm{P}(A_1) = \mathrm{P}(A_1 \mid M)\mathrm{P}(M) + \mathrm{P}(A_1 \mid M^c)\mathrm{P}(M^c) = \mu \times \tfrac{1}{2} + \lambda \times \tfrac{1}{2} = (\mu + \lambda)/2.$$

(b) Independence of accidents from year to year, *for each driver individually*, gives

$$\mathrm{P}(A_2) = \mathrm{P}(A_2 \mid M)\mathrm{P}(M) + \mathrm{P}(A_2 \mid M^c)\mathrm{P}(M^c) = \mu^2 \times \tfrac{1}{2} + \lambda^2 \times \tfrac{1}{2} = (\mu^2 + \lambda^2)/2.$$

(c) Now we want

$$\mathrm{P}(A_2 \mid A_1) = \mathrm{P}(A_2 \cap A_1)/\mathrm{P}(A_1) = \mathrm{P}(A_2)/\mathrm{P}(A_1) = (\lambda^2 + \mu^2)/(\lambda + \mu).$$

(d) Note that $(\lambda^2 + \mu^2)/(\lambda + \mu) > (\lambda + \mu)/2$, because

$$2(\lambda^2 + \mu^2) - (\lambda + \mu)^2 = \lambda^2 + \mu^2 - 2\lambda\mu = (\lambda - \mu)^2 > 0.$$

Thus they would only be equal if $\lambda = \mu$, i.e., with no difference between the sexes.

Quelques remarques

— L'indépendance est un idée clé qui simplifie beaucoup des calculs de probabilité. En pratique, il est essentiel de vérifier si les événements sont indépendants, étant donné qu'une dépendance non détectée peut modifier grandement le resultat.

— L'indépendance mutuelle entraîne l'indépendance deux à deux, mais l'inverse est vrai seulement quand $n = 2$.

— L'indépendance mutuelle entraîne l'indépendance conditionnelle, mais l'inverse est vrai seulement si $B = \Omega$.

3 Variables Aléatoires

slide 54

3.1 Notions de Base

slide 55

Variables aléatoires

On considère souvent des quantités aléatoires numériques.

Exemple 28. *Deux dés équilibrés sont lancés indépendamment, un rouge et un vert. Soit X le total des faces supérieures, trouver ses valeurs possibles et leurs probabilités.*

Définition 29. *Soit $(\Omega, \mathcal{F}, \mathrm{P})$ un espace de probabilité. Une* **variable aléatoire (va)** *$X : \Omega \mapsto \mathbb{R}$ est une application mesurable de l'ensemble fondamental Ω dans $\mathbb{R}$.*

Par **mesurable** nous entendons que pour tout intervalle $[a, b] \subset \mathbb{R}$ (avec $a \leq b$)

$$\{\omega \in \Omega : X(\omega) \in [a, b]\} \in \mathcal{F},$$

et donc $\mathrm{P}(X \in [a, b])$ est définie. Sinon on dit que X est **non-mesurable** par rapport à $(\Omega, \mathcal{F}, \mathrm{P})$. Dorénavant on suppose que toutes nos variables aléatoires sont mesurable.

http://stat.epfl.ch

slide 56

Note to Example 28

X takes values in $S_X = \{2, \ldots, 12\}$, and so is clearly a discrete random variable. By symmetry the 36 points in Ω are equally likely, so, for example,

$$\mathrm{P}(X = 3) = \mathrm{P}(\{(1, 2), (2, 1)\}) = \frac{2}{36}.$$

Thus the probabilities $\mathrm{P}(X = x)$ for $x \in \{2, 3, 4 \ldots, 12\}$ are respectively

$$1/36,\ 2/36,\ 3/36,\ 4/36,\ 5/36,\ 6/36,\ 5/36,\ 4/36,\ 3/36,\ 2/36,\ 1/36.$$

http://stat.epfl.ch

note 1 of slide 56

Calcul des probabilités

Définition 30. *Le* **support** *de X est l'ensemble des valeurs prises par X,*

$$S_X = \{x \in \mathbb{R} : \exists\, \omega \in \Omega \text{ tel que } X(\omega) = x\}.$$

Si S_X est fini ou dénombrable, alors X est une **variable aléatoire discrète**.
On définit aussi

$$\mathcal{B}_x = \{\omega \in \Omega : X(\omega) = x\}, \quad x \in \mathbb{R}.$$

Notant que les $\mathcal{B}_x$ partitionnent Ω et $\mathcal{B}_x = \emptyset$ pour $x \notin S_X$, on a

$$\mathrm{P}(X = x) = \mathrm{P}(\mathcal{B}_x), \quad x \in \mathbb{R},$$

avec $\mathrm{P}(X = x) = 0$ pour $x \notin S_X$.

http://stat.epfl.ch

slide 57

Variable aléatoire de Bernoulli

Définition 31. *Pour un événement $\mathcal{A} \in \mathcal{F}$, la variable aléatoire*

$$I_{\mathcal{A}} \equiv I_{\mathcal{A}}(\omega) = \begin{cases} 1, & \omega \in \mathcal{A}, \\ 0, & \omega \notin \mathcal{A}, \end{cases}$$

s'appele une **variable indicatrice**, *une* **variable aléatoire de Bernoulli**, *ou un* **essai de Bernoulli**.

Exemple 32. *Supposons que n pièces identiques sont lancées indépendamment, soit $\mathcal{A}_i$ l'événement 'on obtient face pour la $i$ème pièce', et soit $I_i = I(\mathcal{A}_i)$ l'indicatrice de cet événement. Alors*

$$\mathrm{P}(I_i = 1) = \mathrm{P}(\mathcal{A}_i) = p, \quad \mathrm{P}(I_i = 0) = \mathrm{P}(\mathcal{A}_i^c) = 1 - p,$$

où p est la probabilité d'obtenir face. Si $n = 3$ et $X = I_1 + I_2 + I_3$, décrire Ω, S_X et les ensembles $\mathcal{B}_x$. Pour n générale, que représentent

$$X = I_1 + \cdots + I_n, \quad Y = I_1(1 - I_2)(1 - I_3), \quad Z = \sum_{j=2}^{n} I_{j-1}(1 - I_j)?$$

Note to Example 32

— When $n = 3$, $\Omega = \{000, 100, 010, 001, 110, 101, 011, 111\}$, where $\omega = i_1 i_2 i_3$ indicates the results for the three successive throws, with $i_j \in \{0, 1\}$. In this case

$$X(\omega) = i_1 + i_2 + i_3 \in S_X = \{0, 1, 2, 3\}$$

and $\mathcal{B}_x = \emptyset$ if $x \notin S_X$, while

$$\mathcal{B}_0 = \{000\}, \quad \mathcal{B}_1 = \{100, 010, 001\}, \quad \mathcal{B}_2 = \{110, 011, 101\}, \quad \mathcal{B}_3 = \{111\}.$$

As $\mathrm{P}(X = x) = \mathrm{P}(\mathcal{B}_x)$ we have

$$\begin{aligned}
\mathrm{P}(X = 0) &= \mathrm{P}(\{000\}) = (1-p)^3, \\
\mathrm{P}(X = 1) &= \mathrm{P}(\{100, 010, 001\}) \\
&= \mathrm{P}(\{100\}) + \mathrm{P}(\{010\}) + \mathrm{P}(\{001\}) \\
&= p(1-p)^2 + (1-p)p(1-p) + (1-p)^2 p \\
&= 3p(1-p)^2 \\
\mathrm{P}(X = 2) &= \mathrm{P}(\{110, 011, 101\}) \\
&= \mathrm{P}(\{110\}) + \mathrm{P}(\{011\}) + \mathrm{P}(\{101\}) \\
&= p^2(1-p) + (1-p)p^2 + p(1-p)p \\
&= 3p^2(1-p) \\
\mathrm{P}(X = 3) &= \mathrm{P}(\{111\}) = p^3,
\end{aligned}$$

where in each case we have used the independence of the three coins.

— In general, Ω comprises all distinct sequences $\omega = i_1 \cdots i_n$ in which $i_j \in \{0, 1\}$, and
  — X is the total number of heads (faces),
  — $Y = 1$ iff the sequence starts FPP (HTT), and
  — Z counts the number of times in ω a 1 is followed by a 0.

Jacob Bernoulli (1654–1705)

Ars Conjectandi, Basel (1713)

Loi d'une variable aléatoire

Définition 33. *La* **loi (de probabilité)** *de* X *est l'ensemble de probabilités* $\mathrm{P}(X \in [a,b])$ *pour tous intervalles* $[a,b] \subset \mathbb{R}$.

Définition 34. *La* **fonction de répartition** *de* X *est*

$$F_X(x) = \mathrm{P}(X \le x), \quad x \in \mathbb{R}.$$

La fonction F_X spécifie la loi de X, car

$$\mathrm{P}(a < X \le b) = F_X(b) - F_X(a),$$

et

$$\begin{aligned} \mathrm{P}(X \in [a,b]) &= \mathrm{P}\left(\bigcap_{n=1}^{\infty} \{X \le b\} \setminus \{X \le a - \tfrac{1}{n}\} \right) \\ &= \lim_{n\to\infty} \mathrm{P}\left(a - \tfrac{1}{n} < X \le b\right) \\ &= F_X(b) - \lim_{n\to\infty} F_X\left(a - \tfrac{1}{n}\right) \\ &= F_X(b) - F_X(a-). \end{aligned}$$

Exemples

Exemple 35. *La variable X telle que $S_X = \mathbb{N}$ et donc*

$$\mathrm{P}\left\{X \in \{1, 2, 3, \ldots\}\right\} = 1$$

et avec

$$\mathrm{P}(X = k) = 2^{-k}, \quad k \in \{1, 2, 3, \ldots\}$$

est une variable discrète, qui prend un nombre dénombrable de valeurs possibles. C'est un cas spécial de la loi **géométrique**,

$$\mathrm{P}(X = k) = (1-p)^{k-1}p, \quad k \in \{1, 2, 3 \ldots, \}, \quad 0 < p < 1.$$

Donner sa fonction de répartition.

Définition 36. *Une* **variable aléatoire continue** *est telle que $\mathrm{P}(X = x) = 0$ pour tout $x \in \mathbb{R}$.*

Exemple 37. *Construisons un nombre réel entre 0 et 1 à travers son expansion décimale de la manière suivante : pour chaque décimale (après la virgule), on lance un dé à 10 faces équilibré. La probabilité que l'on prenne* exactement *une valeur donnée, par exemple $0.314159\cdots$ est de fait nulle (même si on prend bien une valeur, la chance de prendre toute valeur donnée à l'avance est nulle). Quelle est sa fonction de répartition ?*

Note to Example 35

The support is $S = \mathbb{N}$, and for $x \geq 1$ we have

$$\mathrm{P}(X \leq x) = \sum_{r=1}^{\lfloor x \rfloor} p(1-p)^{r-1},$$

so we need to sum a geometric series with common ratio $1 - p$, giving

$$\mathrm{P}(X \leq x) = \frac{p\{1 - (1-p)^{\lfloor x \rfloor}\}}{1 - (1-p)} = 1 - (1-p)^{\lfloor x \rfloor}.$$

Thus

$$F_X(x) = \mathrm{P}(X \leq x) = \begin{cases} 0, & x < 1, \\ 1 - (1-p)^{\lfloor x \rfloor}, & x \geq 1. \end{cases}$$

Note to Example 37

— Let Ω be the set of infinite decimal expansions of real numbers in the interval $(0, 1)$, and for every $x \in (0, 1)$ we let $A_k(x) \subset \Omega$ be the elements of Ω with the same decimal expansion as x to the kth decimal place. To be formal about this, if $x = 0.d_1 \cdots d_k \cdots$, then let $x_k = 0.d_1 \cdots d_k$, so

$$A_k(x) = \{y \in [0, 1] : y_k = x_k\}.$$

— Clearly $\mathrm{P}\{A_k(x)\} = 10^{-k}$, since precisely k digits must be correct, each independently with probability $1/10$. But $\mathrm{P}(X = x) \leq \mathrm{P}\{A_k(x)\}$ for any k, so $\mathrm{P}(X = x) = 0$. That is, the probability of any particular infinite sequence (e.g., $0.314159 \cdots$) must be less than 10^{-k} for every positive k, and must therefore equal zero.

— However every such sequence must lie in $(0, 1)$, so the total probability attached to this interval is 1. Note that since $0.00 \cdots$ and $1.00 \cdots$ both have probability zero, they could be added to the interval without changing the distribution.

— Let $x = 0.d_1 d_2 \cdots$ denote the decimal expansion of a number, let $x_k = 0.d_1 \cdots d_k$ be that expansion truncated after the kth place, and let x_k^- and x_k^+ denote the largest numbers in the set Ω_k of such truncated expansions strictly less and strictly greater than x_k.
If $k = 2$, for example, then x_k can take values in $\Omega_2 = \{0.00, 0.01, \ldots, 0.98, 0.99\}$, and if $x = 0.314159 \cdots$, then $x_2 = 0.31$, $x_2^- = 0.30$ and $x_2^+ = 0.32$, so there are respectively 31 values of Ω_2 less than or equal to x_2^- and 33 values of Ω_2 less than or equal to x_2^+.
In general we therefore have

$$\mathrm{P}(X \leq x_k^-) \leq \mathrm{P}(X \leq x) \leq \mathrm{P}(X \leq x_k^+), \tag{1}$$

and by counting we have $\mathrm{P}(X \leq x_k^-) = x_k$ and $\mathrm{P}(X \leq x_k^+) = x_k + 2/10^k$.
But as $k \to \infty$, both x_k and $x_k + 2/10^{-k}$ converge to x, so letting $k \to \infty$ in (1) yields

$$\mathrm{P}(X \leq x) = F_X(x) = x, \quad x \in [0, 1].$$

— If $x_k = 0.0 \cdots 0$, then we set $x_k^- = x_k$, and if $x_k = 0.9 \cdots 9$, then we set $x_k^+ = 1.0 \cdots 0$ to ensure that (1) also holds at the edges of the interval $(0, 1)$.

Propriétés d'une fonction de répartition

Théorème 38. *Soit $(\Omega, \mathcal{F}, \mathrm{P})$ un espace de probabilité et $X : \Omega \mapsto \mathbb{R}$ une variable aléatoire. Sa fonction de répartition F_X satisfait :*

(a) $\lim_{x \to -\infty} F_X(x) = 0$;

(b) $\lim_{x \to \infty} F_X(x) = 1$;

(c) F_X *est non-décroissante, ainsi* $F_X(x) \leq F_X(y)$ *pour* $x \leq y$;

(d) F_X *est continue à droite, ainsi*

$$\lim_{t \downarrow 0} F_X(x+t) = F_X(x), \quad x \in \mathbb{R};$$

(e) $\mathrm{P}(X > x) = 1 - F_X(x)$;

(f) si $x < y$, *alors* $\mathrm{P}(x < X \leq y) = F_X(y) - F_X(x)$.

De plus on a les propriétés suivantes :

— si X est discrète, alors F_X est constante par intervalles, avec un nombre dénombrable de sauts ;

— si X est continue, alors F_X est continue.

Note to Theorem 38

(a) If not, there must be a blob of mass at $-\infty$, which is not allowed, as $X \in \mathbb{R}$.
(b) Ditto, for $+\infty$.
(c) If $y \geq x$, then $F(y) = F(x) + \mathrm{P}(x < X \leq y)$, so the difference is always non-negative.
(d) Now $F(x+t) = \mathrm{P}(X \leq x) + \mathrm{P}(x < X \leq x+t)$, and the second term here tends to zero, because any point in the interval $(x, x+t]$ at which there is positive probability must lie to the right of x.
(e) We have $\mathrm{P}(X > x) = 1 - \mathrm{P}(X \leq x) = 1 - F_X(x)$.
(f) We have $\mathrm{P}(x < X \leq y) = \mathrm{P}(X \leq y) - \mathrm{P}(X \leq y) = F_X(y) - F_X(x)$.

Variables aléatoires indépendantes

Définition 39. *Deux variables aléatoires X, Y sont dites* **indépendantes** *si les événements $\{X \in \mathcal{I}\}$ et $\{Y \in \mathcal{J}\}$ sont indépendants pour toute paire d'intervalles $\mathcal{I}$ et $\mathcal{J}$,*

$$\mathrm{P}\left(X \in \mathcal{I} \text{ et } Y \in \mathcal{J}\right) = \mathrm{P}\left(X \in \mathcal{I}\right) \times \mathrm{P}\left(\mathcal{Y} \in \mathcal{J}\right).$$

De même :

Définition 40. *Des variables aléatoires $X_1, \ldots, X_n, \ldots$ sont indépendantes si les événements $\{X_1 \in \mathcal{I}_1\}, \ldots, \{X_n \in \mathcal{I}_n\}, \ldots$ sont indépendants pour tous intervalles $\mathcal{I}_1, \ldots, \mathcal{I}_n, \ldots$.*

Comme on le verra, les variables indépendantes sont très importantes.

Lemme 41. *Si $X_1, \ldots, X_n, \ldots$ sont indépendantes et $g_1, \ldots, g_n, \ldots : \mathbb{R} \to \mathbb{R}$ des fonctions 'raisonnables', alors $g_1(X_1), \ldots, g_n(X_n), \ldots$ sont indépendantes.*

La preuve est laissée en exercice (par 'raisonnable' on entend n'importe quelle fonction qui peut être décrite par un nombre fini ou dénombrable d'opérations ; par exemple toute fonction continue ou avec un nombre dénombrable de discontinuités est raisonnable).

Remarques

— S'il n'y pas de risque de confusion on écrit S, f et F plutôt que S_X, f_X et F_X.

— On specifie la loi d'un variable aléatoire de manière équivalente en disant :

  — X suit une loi binomiale avec paramètres n et p ; ou

  — $X \sim B(n,p)$; ou

  — en donnant la fonction de masse de X ; ou

  — en donnant la fonction de répartition de X.

— Si $X_1, \ldots, X_n$ sont des variables indépendantes et identiquement distribuées avec la même loi F, nous écrivons

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} F.$$

— On va souvent ignorer l'espace de probabilité $(\Omega, \mathcal{F}, \mathrm{P})$ sous-jacent quand on a affaire à une variable aléatoire X. Nous penserons plutôt en termes de X, $F(x)$, et $f(x)$. On peut légitimer cet 'oubli' mathématiquement, car $(\Omega, \mathcal{F}, \mathrm{P})$ génère un espace de probabilité avec $\Omega_X = \mathbb{R}$, une tribu $\mathcal{F}_X$ généré sur les sous-ensembles de Ω_X, et une fonction de probabilité qui corréspond à F_X.

3.2 Variables Aléatoires Discrètes

Variable aléatoire discrète

Définition 42. *La* **fonction de masse (densité)** *d'une variable aléatoire discrète X est*

$$f(x) = \mathrm{P}(X = x) = F(x) - F(x-), \quad x \in \mathbb{R}.$$

Elle a deux propriétés clés :

(i) $f(x) \geq 0$, *et* $f(x) > 0$ *si et seulement si* $x \in S$;

(ii) *la probabilité totale* $\sum_{x \in S} f(x) = 1$.

— En anglais la fonction de masse est appelée **probability mass function (PMF)** ou **probability density function (PDF)**.

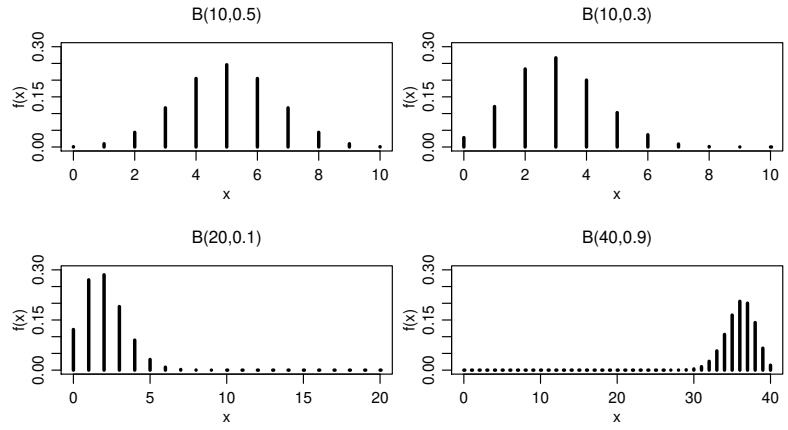
Définition 43. *Une variable aléatoire* **binomiale** *X a pour fonction de masse*

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x \in \{0, 1, \ldots, n\}, \quad n \in \mathbb{N}, \quad 0 \leq p \leq 1.$$

On écrit $X \sim B(n,p)$, et appelle n le **dénominateur** *et p la* **probabilité de succès**. *Avec $n = 1$, c'est une variable de Bernoulli.*

Pause pensées. *Esquissez la fonction de répartition de $X \sim B(1, 0.5)$.*

Fonctions de masse binomiale



Lois géométrique et binomiale négative

Considérons une suite d'essais de Bernoulli indépendants tous ayant probabilité p de succès.

Définition 44. *Une variable aléatoire* **géométrique** X *a pour fm*

$$f(x) = p(1-p)^{x-1}, \quad x \in \{1, 2, \ldots\}, \quad 0 \leq p \leq 1.$$

On note $X \sim \text{Geom}(p)$, *et on appelle* p *la* **probabilité de succès**.

Elle modélise le nombre d'essais jusqu'au prémier succès.

Proposition 45 (Manque de mémoire). *Si* $X \sim \text{Geom}(p)$, *alors*

$$\text{P}(X > n + m \mid X > m) = \text{P}(X > n).$$

Définition 46. *Une variable aléatoire* **binomiale negative** X *de paramètres* n *et* p *a pour fonction de masse*

$$f(x) = \binom{x-1}{n-1} p^n (1-p)^{x-n}, \quad x \in \{n, n+1, n+2, \ldots\}, \quad 0 \leq p \leq 1.$$

On note $X \sim \text{NegBin}(n, p)$. *Lorsque* $n = 1$, $X \sim \text{Geom}(p)$.

Elle modélise le nombre d'essais jusqu'au $n$ème succès.

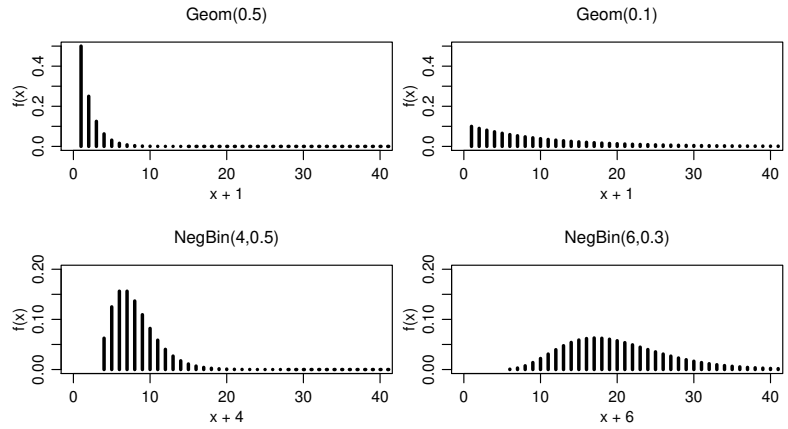
Note to Proposition 45

Since $\text{P}(X > n) = (1-p)^n$, we seek

$$\text{P}(X > n + m \mid X > m) = (1-p)^{m+n}/(1-p)^m = (1-p)^n = \text{P}(X > n).$$

Thus we see that there is a 'lack of memory' : knowing that $X > m$ does not change the probability that we have to wait at least another n trials before seeing the event.

FMs géométrique et binomiale négative



Loi binomiale négative : version alternative

Parfois on écrit les variables géométriques et binomiale negatives sous une forme plus générale, prenant $Y = X - n$, et alors la fonction de masse est

$$f_Y(y) = \frac{\Gamma(y+\alpha)}{\Gamma(\alpha)y!}p^\alpha(1-p)^y, \quad y \in \{0,1,2,\ldots\}, \quad 0 \leq p \leq 1, \alpha > 0,$$

où

$$\Gamma(\alpha) = \int_0^\infty u^{\alpha-1}e^{-u}\,du, \quad \alpha > 0,$$

est la **fonction Gamma**. Ses propriétés principales sont :

$$\begin{aligned}
\Gamma(1) &= 1; \\
\Gamma(\alpha+1) &= \alpha\Gamma(\alpha), \quad \alpha > 0; \\
\Gamma(n) &= (n-1)!, \quad n = 1,2,3,\ldots; \\
\Gamma(\tfrac{1}{2}) &= \sqrt{\pi}.
\end{aligned}$$

Lois uniforme, hypergéométrique et Poisson

Définition 47. *Un dé equilibré à n faces a*

$$\mathrm{P}(X = x) = f(x) = \frac{1}{n}, \quad x \in \{1, \ldots, n\},$$

et la loi de X est **(discrète) uniforme**.

Définition 48. *On tire sans remise un échantillon de m boules d'une urne contenant b blanches et n noires. Soit X le nombre de boules blanches tirées. Alors*

$$\mathrm{P}(X = x) = f(x) = \frac{\binom{b}{x}\binom{n}{m-x}}{\binom{b+n}{m}}, \quad x \in \{\max(0, m-n), \ldots, \min(b, m)\},$$

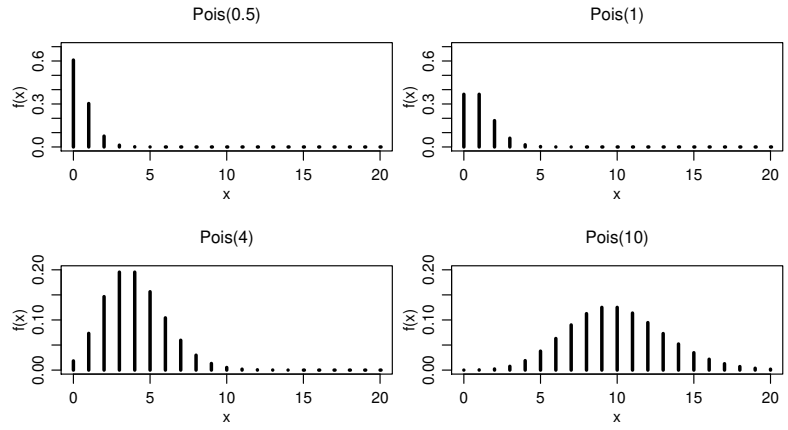
et la loi de X est **hypergéométrique**.

Définition 49. *Une variable aléatoire de* **Poisson**, $X \sim \mathrm{Pois}(\lambda)$, *a*

$$\mathrm{P}(X = x) = f(x) = \frac{\lambda^x}{x!} e^{-\lambda}, \quad x \in \{0, 1, \ldots\}, \quad \lambda > 0.$$

On verra que c'est la limite des variables $X_n \sim B(n, p_n)$ quand $n \to \infty$ avec $np_n \to \lambda > 0$.

Fonctions de masse de Poisson



Espérance

Définition 50. *Soit X une variable aléatoire discrète pour laquelle $\sum_{x\in S}|x|f(x)<\infty$, où S est le support de f.* **L'espérance** *de X est*

$$\mathrm{E}(X)=\sum xP(X=x)=\sum_{x\in S}xf(x).$$

— Si $\mathrm{E}(|X|)=\sum_{x\in S}|x|f(x)$ n'est pas finie, alors $\mathrm{E}(X)$ n'est pas bien défini.
— $\mathrm{E}(X)$ est parfois appelée la "moyenne de X", mais nous utiliserons le mot "moyenne" que pour des quantités empiriques.

Exemple 51. *Calculer l'espérance de $X\sim B(n,p)$.*

Pause pensées. *Calculer l'espérance du résultat d'un lancer d'un dé equilibré avec n faces.*

Note to Example 51

For any random variable with finite support, $\mathrm{E}(|X|)<\infty$, so the expectation is well-defined :

$$\begin{aligned}\mathrm{E}(X) &= \sum_{x=0}^{n} x\binom{n}{x}p^x(1-p)^{n-x}\\ &= \sum_{x=1}^{n}\frac{n\times(n-1)!}{(x-1)!\{n-1-(x-1)\}!}p\times p^{x-1}(1-p)^{(n-1)-(x-1)}\\ &= np\sum_{y=0}^{n-1}\binom{n-1}{y}p^y(1-p)^{n-1-y}=np,\end{aligned}$$

where we have set $y=x-1$.

Fonction d'une variable aléatoire discrète

Des fonctions (mesurables) à valeurs réelles de variables aléatoires sont elles-même des variables aléatoires, elles ont donc aussi des fonctions de masse et de répartition.

Théorème 52. *Si X est une variable aléatoire et $Y=g(X)$, alors Y a pour fonction de masse*

$$f_Y(y)=\sum_{x:g(x)=y}f_X(x)=\sum_{x\in g^{-1}(y)}f_X(x),$$

où $g^{-1}(y)=\{x:g(x)=y\}$.

Exemple 53. *Calculer la fonction de masse de $Y=I(X\geq 1)$ lorsque $X\sim\mathrm{Pois}(\lambda)$.*

Théorème 54. *Soit X une variable aléatoire de fonction de masse f_X et de support S_X, et soit g une fonction à valeurs réelles de $\mathbb{R}$. Alors si $\sum_{x\in S_X}|g(x)|f_X(x)<\infty$, on définit*

$$\mathrm{E}\{g(X)\}=\sum_{x\in S_X}g(x)f_X(x).$$

Exemple 55. *Soit $X\sim\mathrm{Pois}(\lambda)$, calculer $\mathrm{E}(X)$ et $\mathrm{E}\{X(X-1)\cdots(X-r+1)\}$.*

Note to Theorem 52

For $y \in S_Y$ we have

$$f_Y(y) = \mathrm{P}(Y = y) = \sum_{x:g(x)=y} \mathrm{P}(X = x) = \sum_{x:g(x)=y} f_X(x) = \sum_{x \in g^{-1}(y)} f_X(x)$$

Note to Example 53

Here $Y = I(X \geq 1)$ takes values in $S_Y = \{0, 1\}$, and

$$g^{-1}(0) = \{x : I(x \geq 1) = 0\} = \{0\}, \quad g^{-1}(1) = \{x : I(x \geq 1) = 1\} = \{1, 2, \ldots\},$$

so

$$\begin{aligned} f_Y(0) &= \sum_{x \in g^{-1}(0)} f_X(x) = f_X(0) = e^{-\lambda}, \\ f_Y(1) &= \sum_{x \in g^{-1}(1)} f_X(x) = \sum_{x=1}^{\infty} \frac{\lambda^x}{x!} e^{-\lambda} = 1 - e^{-\lambda}. \end{aligned}$$

Note to Theorem 54

Write $Y = g(X)$, then

$$\mathrm{E}(Y) = \sum_y y f_Y(y) = \sum_y y \sum_{x:g(x)=y} f_X(x) = \sum_y \sum_{x:g(x)=y} g(x) f_X(x) = \sum_x g(x) f_X(x),$$

as required, the interchange in the summation being justified by the absolute convergence.

Note to Example 55

Note that

$$\mathrm{E}\{X(X-1)\cdots(X-r+1)\} = \sum_{x=0}^{\infty} x(x-1)\cdots(x-r+1)\frac{\lambda^x}{x!}e^{-\lambda} = \lambda^r \sum_{x-r=0}^{\infty} \frac{\lambda^{x-r}}{(x-r)!}e^{-\lambda} = \lambda^r,$$

which yields $\mathrm{E}(X) = \lambda$ and $\mathrm{E}\{X(X-1)\} = \lambda^2$.

Propriétés de l'espérance

Théorème 56. *Soient X et Y des variables aléatoires d'espérance finie, et soit $a, b, c \in \mathbb{R}$ des constantes. Alors*

(a) $\mathrm{E}(\cdot)$ *est une application linéaire,*

$$\mathrm{E}(aX + bY + c) = a\mathrm{E}(X) + b\mathrm{E}(Y) + c;$$

(b) si $g(X)$ *et* $h(X)$ *ont des espérances finies, alors*

$$\mathrm{E}\{g(X) + h(X)\} = \mathrm{E}\{g(X)\} + \mathrm{E}\{h(X)\};$$

(c) si $\mathrm{P}(X = b) = 1$, *alors* $\mathrm{E}(X) = b$ *;*

(d) si $\mathrm{P}(a < X \leq b) = 1$, *alors* $a < \mathrm{E}(X) \leq b$ *;*

(e) $\{\mathrm{E}(X)\}^2 \leq \mathrm{E}(X^2)$ *;*

(f) si $X \perp\!\!\!\perp Y$ *et* $g(X)$ *et* $h(Y)$ *ont des espérances finies, alors*

$$\mathrm{E}\{g(X)h(Y)\} = \mathrm{E}\{g(X)\}\mathrm{E}\{h(Y)\}.$$

Note to Theorem 56

(a) We need to show absolute convergence :

$$\begin{aligned}\sum_{x,y}|ax+by+c|f(x,y) &\leq \sum_{x,y}(|a||x|+|b||y|+|c|)f(x,y)\\ &= |a|\sum_{x,y}|x|f(x,y)+|b|\sum_{x,y}|y|f(x,y)+|c|\sum_{x,y}f(x,y)<\infty,\end{aligned}$$

because $\mathrm{E}(|X|), \mathrm{E}(|Y|) < \infty$, and applying linearity of the summation, noting that (for example)

$$\begin{aligned}\mathrm{E}(aX) &= a\sum_{x,y}xf(x,y) = a\sum_x x\sum_y f(x,y) = a\sum_x x\mathrm{P}(X=x)) = a\sum_x xf_X(x) = a\mathrm{E}(X),\\ \mathrm{E}(c) &= \sum_{x,y}cf(x,y) = c\sum_{x,y}f(x,y) = c\times 1.\end{aligned}$$

(b) Follows the argument in (a), after noting that $|g(x)+h(x)| \leq |g(x)|+|h(x)|$.
(c) Here $f(b) = \mathrm{P}(X=b) = 1$, so $\mathrm{E}(X) = bf(b) = b$ by definition.
(d) Now $f(x) = 0$ for $x \notin (a,b]$, so $\mathrm{E}(X) = \sum_x xf(x) \leq \sum_x bf(x) = b$ and similarly $\mathrm{E}(X) > a$.
(e) Note that

$$0 \leq \mathrm{E}\left\{(X-a)^2\right\} = \mathrm{E}\left\{X^2-2aX+a^2\right\} = \mathrm{E}(X^2)-2a\mathrm{E}(X)+a^2,$$

which yields that $(-2\mathrm{E}(X))^2 - 4\mathrm{E}(X^2) \leq 0$, giving the inequality.
(f) If $X \perp\!\!\!\perp Y$, then $f(x,y) = f_{X,Y}(x,y) = f_X(x)f_Y(y)$ for all $x, y \in \mathbb{R}$, so in particular $S_{X,Y} \subset \mathbb{R}^2$ is the Cartesian product $S_X \times S_Y$. Hence

$$\begin{aligned}\mathrm{E}\{g(X)h(Y)\} &= \sum_{(x,y)\in S_{X,Y}} g(x)h(y)f(x,y)\\ &= \sum_{(x,y)\in S_{X,Y}} g(x)h(y)f_X(x)f_Y(y)\\ &= \sum_{x\in S_X, y\in S_Y} g(x)f_X(x)\times h(y)f_Y(y)\\ &= \sum_{x\in S_X} g(x)f_X(x)\sum_{y\in S_Y} h(y)f_Y(y) = \mathrm{E}\{g(X)\}\mathrm{E}\{h(Y)\}.\end{aligned}$$

Exemples

Exemple 57. *Dans l'exemple 22, soit X le nombre d'hommes qui s'en vont avec le correct chapeau. Montrer que $\mathrm{E}(X) = 1$, pour tout n.*

Exemple 58. *Soit $I_A, I_B, \ldots$ les indicatrices des événements $A, B, \ldots$. Montrer que*

$$I_{A\cap B} = I_A I_B, \quad I_{A\cup B} = 1-(1-I_A)(1-I_B), \quad \mathrm{E}(I_A) = \mathrm{P}(A).$$

et en déduire la formule d'inclusion-exclusion

$$\mathrm{P}\left(\bigcup_{i=1}^n A_i\right) = \sum_{r=1}^n (-1)^{r+1}\sum_{1\leq i_1<\cdots<i_r\leq n}\mathrm{P}(A_{i_1}\cap\cdots\cap A_{i_r}).$$

Note to Example 57

— Since $X = I_1 + \cdots + I_n$, we have

$$\mathrm{E}(X) = \mathrm{E}(I_1 + \cdots + I_n) = \sum_{i=1}^{n} \mathrm{E}(I_i) = n\mathrm{E}(I_1) = np,$$

the first equality by definition of X, the second by linearity of expectation, and the third by symmetry. Independence was not necessary here.

— For the hats, the variables are not independent ; in the case $n = 2$, for example, either $I_1 = I_2 = 1$ or $I_1 = I_2 = 0$. The probability that the first man gets his hat is $1/n$, so $\mathrm{E}(X) = n \times 1/n = 1$ for any n. There is just one person likely to get the correct hat whether two people are present or whether 1000 people are present, which seems surprising.

Note to Example 58

First show the first expression by checking that they are the same, then expand the others algebraically, before taking expectations and using linearity of the expectation operation.

Moments d'une distribution

Définition 59. *Si X a une fonction de masse $f(x)$ telle que $\sum_x |x|^r f(x) < \infty$, alors*

(a) le $r$ème **moment** *de X est $\mathrm{E}(X^r)$;*

(b) le $r$ème **moment centré** *de X est $\mathrm{E}[\{X - \mathrm{E}(X)\}^r]$;*

(c) le $r$ème **moment factoriel** *de X est $\mathrm{E}\{X(X-1)\cdots(X-r+1)\}$;*

(d) la **variance** *de X est $\mathrm{var}(X) = \mathrm{E}[\{X - \mathrm{E}(X)\}^2]$.*

Remarque 60. *L'espérance et la variance sont les moments les plus importants, car elles mesurent la localisation et la dispersion de X. Elles sont analogues en mécanique au* **centre de gravité** *et au* **moment d'inertie** *d'un objet dont la masse est distribuée selon f.*

Théorème 61. *Soit X une variable aléatoire dont la variance existe, et soient a, b des constantes. Alors*

$$\begin{aligned} \mathrm{var}(X) &= \mathrm{E}(X^2) - \mathrm{E}(X)^2 = \mathrm{E}\{X(X-1)\} + \mathrm{E}(X) - \mathrm{E}(X)^2; \\ \mathrm{var}(aX+b) &= a^2\mathrm{var}(X); \\ \mathrm{var}(X) = 0 &\Rightarrow X \text{ est constante de probabilité 1.} \end{aligned}$$

Exemple 62. *Calculer la variance d'une variable aléatoire Poissonienne.*

Note to Theorem 61

(a) Just expand, use linearity of $\mathrm{E}(\cdot)$, and simplify.
(b) Ditto.
(c) If we write $\mathrm{E}(X) = \mu$ and

$$\mathrm{var}(X) = \mathrm{E}[\{X - \mathrm{E}(X)\}^2] = \mathrm{E}[\{X - \mu\}^2] = \sum_x f(x)(x-\mu)^2 = 0,$$

then for each $x \in S_X$, either $x = \mu$ or $f(x) = 0$. Suppose that $f(a), f(b) > 0$ and $a \neq b$. Then if $\mathrm{var}(X) = 0$, we must have $a = \mu = b$, which is a contradiction. Therefore $f(x) > 0$ for a unique value of x, and then we must have $f(x) = 1$, so $\mathrm{P}(X = x) = 1$ and $(x - \mu)^2 = 0$; thus $\mathrm{P}(X = \mu) = f(\mu) = 1$.

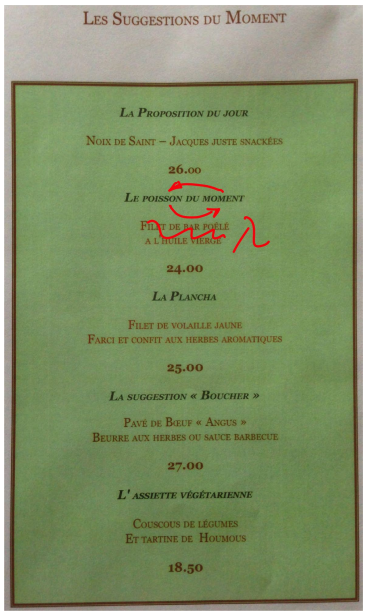
Note to Example 62

By recalling Example 55, we find

$$\mathrm{var}(X) = \mathrm{E}\{X(X-1)\} + \mathrm{E}(X) - \mathrm{E}(X)^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

Poisson du moment

Moment du Poisson



Loi conditionnelle

Définition 63. *Soit $(\Omega, \mathcal{F}, \mathrm{P})$ un espace de probabilité, sur lequel on définit une variable aléatoire X, et soit $\mathcal{B} \in \mathcal{F}$ avec $\mathrm{P}(\mathcal{B}) > 0$. Alors la* **fonction de masse conditionnelle** *de X sachant $\mathcal{B}$ est*

$$f(x \mid \mathcal{B}) = \mathrm{P}(X = x \mid \mathcal{B}) = \frac{\mathrm{P}(\mathcal{B} \cap \{X = x\})}{\mathrm{P}(\mathcal{B})} = \frac{f(x)\mathrm{P}(\mathcal{B} \mid X = x)}{\mathrm{P}(\mathcal{B})}, \quad x \in \mathbb{R}.$$

Si $\mathcal{B} = \{X \in \mathcal{A}\}$ pour $\mathcal{A} \subset \mathbb{R}$, alors $\mathrm{P}(\mathcal{B} \mid X = x) = I(x \in \mathcal{A})$ et

$$f(x \mid \mathcal{B}) = \mathrm{P}(X = x \mid X \in \mathcal{A}) = \frac{f(x)I(x \in \mathcal{A})}{\mathrm{P}(X \in \mathcal{A})}, \quad x \in \mathbb{R}.$$

*Si $\sum_x |g(x)| f(x \mid \mathcal{B}) < \infty$, alors l'*espérance conditionelle *de $g(X)$ est*

$$\mathrm{E}\{g(X) \mid \mathcal{B}\} = \sum_x g(x) f(x \mid \mathcal{B}).$$

Proposition 64. *Soient X une variable aléatoire avec $\mathrm{E}(|X|) < \infty$ et $\{\mathcal{B}_i\}_{i=1}^{\infty}$ une partition de Ω avec $\mathrm{P}(\mathcal{B}_i) > 0$ pour tout i, alors,*

$$\mathrm{E}(X) = \sum_{i=1}^{\infty} \mathrm{E}(X \mid \mathcal{B}_i)\mathrm{P}(\mathcal{B}_i).$$

Note to Proposition 64

Note that $\mathrm{E}(X)$ is well-defined, because $\mathrm{E}(|X|) < \infty$. Now Theorem 18 yields

$$f(x) = \mathrm{P}(X = x) = \sum_i \mathrm{P}(X = x \mid \mathcal{B}_i)\mathrm{P}(\mathcal{B}_i) = \sum_i f(x \mid \mathcal{B}_i)\mathrm{P}(\mathcal{B}_i),$$

so

$$\mathrm{E}(|X|) = \sum_x |x| f(x) = \sum_x |x| \sum_i f(x \mid \mathcal{B}_i)\mathrm{P}(\mathcal{B}_i) = \sum_i \mathrm{P}(\mathcal{B}_i) \sum_x |x| f(x \mid \mathcal{B}_i) < \infty, \tag{2}$$

where the interchange of summations is justified by the absolute convergence. Clearly

$$\sum_x |x| f(x \mid \mathcal{B}_i) = \mathrm{E}(|X| \mid \mathcal{B}_i) < \infty,$$

because all the terms in (2) are positive and must therefore be finite. Thus each $\mathrm{E}(X \mid \mathcal{B}_i)$ is well-defined. Hence

$$\mathrm{E}(X) = \sum_x x f(x) = \sum_x x \sum_i f(x \mid \mathcal{B}_i)\mathrm{P}(\mathcal{B}_i) = \sum_i \mathrm{P}(\mathcal{B}_i) \sum_x x f(x \mid \mathcal{B}_i) = \sum_i \mathrm{E}(X \mid \mathcal{B}_i)\mathrm{P}(\mathcal{B}_i),$$

as required, where the interchange of summations is again justified by the absolute convergence. Of course this is also true if the partition has a finite number of events.

Exemple

Exemple 65. *Calculer la fonction de masse et espérance conditionnelle de* $X \sim \mathrm{Poiss}(\lambda)$ *sachant que* $X \geq 1$.

Note to Example 65

Here $\mathcal{B} = \{X \in \mathcal{A}\}$ with $\mathcal{A} = \{1, 2, \ldots\}$, so $\mathrm{P}(\mathcal{B}) = 1 - \mathrm{P}(X = 0) = 1 - e^{-\lambda}$, and thus

$$f(x \mid \mathcal{B}) = \frac{f(x) I(x \in \mathcal{Z})}{\mathrm{P}(X \in \mathcal{A})} = \frac{\lambda^x e^{-\lambda}}{x!(1 - e^{-\lambda})}, \quad x = 1, 2, \ldots.$$

The easiest way to get the mean is to use Proposition 64, which gives

$$\lambda = \mathrm{E}(X \mid \mathcal{B})\mathrm{P}(\mathcal{B}) + \mathrm{E}(X \mid \mathcal{B}^c)\mathrm{P}(\mathcal{B}^c) = \mathrm{E}(X \mid \mathcal{B}) \times (1 - e^{-\lambda}) + 0 \times \mathrm{P}(\mathcal{B}^c),$$

and thus $\mathrm{E}(X \mid \mathcal{B}) = \lambda/(1 - e^{-\lambda})$.

Quelle distribution ?

On a rencontré plusieurs lois discrètes : Bernoulli, binomiale, géométrique, binomiale negative, hypergéométrique, Poisson—comment choisir ? Voici un petit algorithme pour aider votre raisonnement :

Est-que X se base sur des 'essais' (0/1) indépendants avec une même probabilité p, ou sur un tirage au sort d'une population finie, 'avec remise' ?

— **Oui**, est-ce que le nombre total d'essais n est fixé, et donc $X \in \{0, \ldots, n\}$?

  — **Oui** : loi **binomiale**, $X \sim B(n, p)$ (et donc loi **Bernoulli** si $n = 1$)

    — Si $n \approx \infty$ ou $n \gg np$, on peut utiliser la loi de **Poisson**, $X \sim \text{Pois}(np)$

  — **Non** : alors $X \in \{n, n+1, \ldots\}$, et on utilise la loi **géométrique** (si X est le nombre d'essais jusqu'à un seul succès) ou **binomiale negative** (si X est le nombre d'essais jusqu'au dernier de plusieurs succès)

— **Non** : si le tirage au sort est 'sans remise' d'une population finie, alors $X \sim$ **hypergéométrique**

Il y a beaucoup de lois, et quand il y a un choix on choisit souvent de manière empirique. La carte suivante provient de Leemis and McQueston (2008, American Statistician) ...

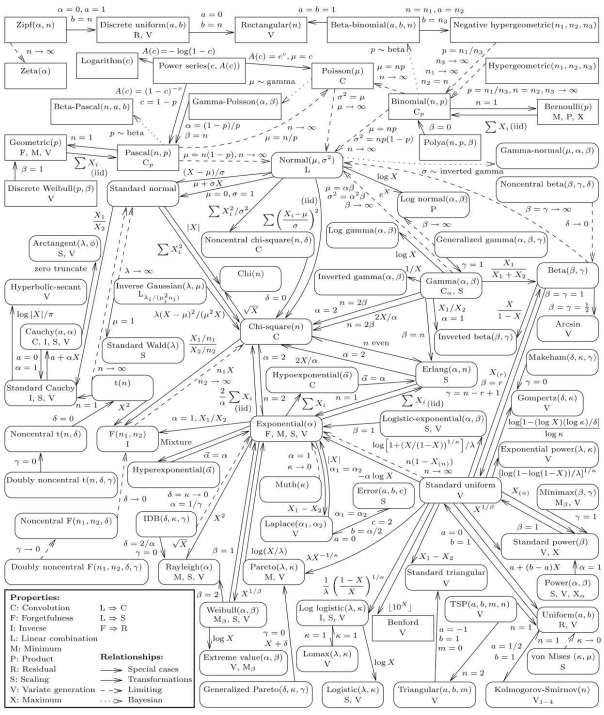


Figure 1. Univariate distribution relationships.

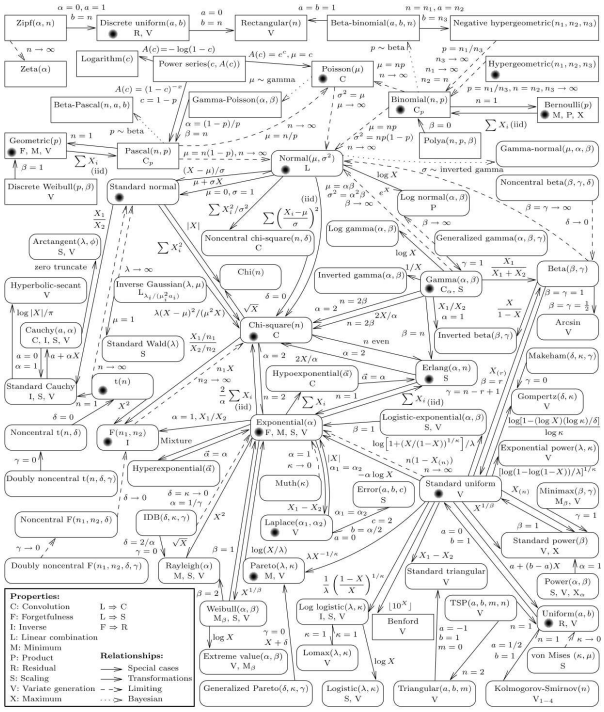


Figure 1. Univariate distribution relationships.

3.3 Variables Aléatoires Continues

slide 85

Variable aléatoire continue

Définition 66. *Une* **variable aléatoire continue** X *est telle que sa fonction de répartition* $F(x)$ *est continue pour tout* $x \in \mathbb{R}$.

— Ceci implique que

$$\mathrm{P}(X = x) = F(x) - \lim_{n\to\infty} F\left(x - \tfrac{1}{n}\right) = 0, \quad x \in \mathbb{R},$$

et donc est cohérent avec notre définition précédante. Si on répète des tirages indépendants de même loi que X, on n'aura jamais deux fois la même valeur.

— La plupart des variables aléatoires dans ce cours seront soit discrètes, soit continues.

— Toutefois on rencontre des **variables mixtes**, telle que la quantité de pluie qui tombera à la gare de Lausanne demain.

Exemple 67. *Une* **variable exponentielle** X *de paramètre* $\lambda > 0$ *a fonction de répartition*

$$F(x) = \begin{cases} 0, & x \le 0, \\ 1 - \exp(-\lambda x), & x > 0. \end{cases}$$

On écrit $X \sim \exp(\lambda)$.

Variable aléatoire absolument continue

Définition 68. *Une variable aléatoire X est dite* **absolument continue** *ou* **à densité** *s'il existe une fonction, dite* **fonction de densité (continue)**, *$f : \mathbb{R} \to [0, \infty)$ telle que*

$$\mathrm{P}\{X \in [a, b]\} = \int_a^b f(x)\,\mathrm{d}x, \quad a \leq b.$$

— On a

$$F(x) = \mathrm{P}(X \leq x) = \int_{-\infty}^{x} f(u)\,\mathrm{d}u,$$

et, quand la derivée existe,

$$f(x) = \frac{\mathrm{d}F(x)}{\mathrm{d}x}.$$

— On doit toujours avoir $f \geq 0$ et $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$.

— $f(x)$ n'est pas une probabilité, c'est la limite $\lim_{h\to 0} \mathrm{P}(X \in [x - h/2, x + h/2])/h$; la 'probabilité' correspondante est $f(x)\mathrm{d}x$, qui vaut formellement 0.

— Dans ce cours, les variables aléatoires continues seront supposées absolument continues.

— Les variables continues qui ne sont pas à densité ont des exemples pathologiques.

Exemples

Définition 69. *La variable aléatoire* **uniforme** *$U \sim U(a, b)$ a pour densité*

$$f(u) = \begin{cases} \frac{1}{b-a}, & a < u < b, \\ 0, & \textit{sinon}, \end{cases} \qquad a < b.$$

Définition 70. *La variable aléatoire* **Pareto** *X a pour fonction de répartition*

$$F(x) = \begin{cases} 0, & x < \beta, \\ 1 - \left(\frac{\beta}{x}\right)^{\alpha}, & x \geq \beta. \end{cases}, \quad \alpha, \beta > 0.$$

Définition 71. *La variable aléatoire* **Laplace** *X a pour densité*

$$f(x) = \frac{\lambda}{2} e^{-\lambda|x-\eta|}, \quad x \in \mathbb{R}, \quad \eta \in \mathbb{R}, \lambda > 0.$$

Définition 72. *La variable aléatoire* **gaussienne (ou normale)** *$X \sim \mathcal{N}(\mu, \sigma^2)$ a pour densité*

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}, \quad x \in \mathbb{R}, \quad \mu \in \mathbb{R}, \sigma > 0.$$

La variable $Z \sim \mathcal{N}(0, 1)$, **'standard normal'**, *a pour densité $\phi(z) = (2\pi)^{-1/2} e^{-z^2/2}$, $(z \in \mathbb{R})$ et fonction de répartition $\Phi(z) = \int_{-\infty}^{z} \phi(x)\,\mathrm{d}x$.*

La fonction $\Phi(z)$

z	0	1	2	3	4	5	6	7	8	9
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56750	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1.0	.84134	.84375	.84614	.84850	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92786	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
2.0	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
2.3	.98928	.98956	.98983	.99010	.99036	.99061	.99086	.99111	.99134	.99158
2.4	.99180	.99202	.99224	.99245	.99266	.99286	.99305	.99324	.99343	.99361
2.5	.99379	.99396	.99413	.99430	.99446	.99461	.99477	.99492	.99506	.99520
2.6	.99534	.99547	.99560	.99573	.99585	.99598	.99609	.99621	.99632	.99643
2.7	.99653	.99664	.99674	.99683	.99693	.99702	.99711	.99720	.99728	.99736
2.8	.99744	.99752	.99760	.99767	.99774	.99781	.99788	.99795	.99801	.99807
2.9	.99813	.99819	.99825	.99831	.99836	.99841	.99846	.99851	.99856	.99861
3.0	.99865	.99869	.99874	.99878	.99882	.99886	.99889	.99893	.99896	.99900

La fonction `pnorm` du logiciel `R` pourrait être utile : $\Phi(z) =$ `pnorm(z)`.

Moments et quantiles

Définition 73. *Soient $g(x)$ une fonction à valeurs réelles, et X une variable aléatoire continue de densité $f(x)$. Alors si $\mathrm{E}\{|g(X)|\} < \infty$, on définit l'espérance de $g(X)$ comme*

$$\mathrm{E}\{g(X)\} = \int_{-\infty}^{\infty} g(x) f(x)\, dx.$$

En particulier l'espérance et la variance de X sont

$$\begin{aligned} \mathrm{E}(X) &= \int_{-\infty}^{\infty} x f(x)\, dx, \\ \mathrm{var}(X) &= \int_{-\infty}^{\infty} \{x - \mathrm{E}(X)\}^2 f(x)\, dx = \mathrm{E}(X^2) - \mathrm{E}(X)^2. \end{aligned}$$

Définition 74. *Soit $0 < p < 1$. On définit le $p$ième quantile d'une fonction de répartition F par*

$$x_p = \inf\{x : F(x) \geq p\}.$$

Pour la plupart des variables aléatoires continues, x_p est unique et vaut $x_p = F^{-1}(p)$, où F^{-1} est la fonction inverse de F. Ainsi x_p est la valeur pour laquelle $\mathrm{P}(X \leq x_p) = p$. En particulier, on appelle le 0.5ème quantile la médiane de F.

Exemple 75. *Calculer les espérances, variances et quantiles des lois (a) $U(a, b)$, (b) Pareto.*

Note to Example 75

(a) We need $\mathrm{E}(U^r)$ for $r = 1, 2$, and this is $\frac{1}{r+1}(b^{r+1} - a^{r+1})/(b - a)$. Hence $\mathrm{E}(X) = \frac{1}{2}(b^2 - a^2)/(b - a) = (b + a)/2$, as expected. For the variance, note that

$$\mathrm{E}(X^2) - \mathrm{E}(X)^2 = \tfrac{1}{3}\frac{b^3 - a^3}{b - a} - (b + a)^2/4 = \tfrac{1}{3}(b^2 + ab + a^2) - (b^2 + 2ab + a^2)/4 = (b - a)^2/12.$$

For the quantiles, we solve $p = F(x_p) = \mathrm{P}(U \leq x_p) = (x_p - a)/(b - 1)$, for $a < x_p < b$, so $x_p = a + (b - a)p$, and in particular the median is $x_{0.5} = a + (b - a)/2 = (a + b)/2$.
(b) First we compute the density function,

$$f(x) = \frac{\mathrm{d}F(x)}{\mathrm{d}x} = \begin{cases} 0, & x < \beta, \\ \alpha\beta^\alpha x^{-\alpha-1}, & x \geq \beta. \end{cases}$$

Hence

$$\mathrm{E}(|X^r|) = \mathrm{E}(X^r) = \alpha \int_\beta^\infty \beta^\alpha x^{r-\alpha-1}\, dx = \frac{\alpha\beta^r}{\alpha - r}, \quad \alpha > r.$$

If $\alpha \leq r$ then the moment does not exist. In particular, $\mathrm{E}(X) < \infty$ only if $\alpha > 1$, and $\mathrm{var}(X) < \infty$ only if $\alpha > 2$. If they are finite,

$$\mathrm{E}(X) = \frac{\alpha\beta}{\alpha - 1}, \quad \mathrm{var}(X) = \frac{\beta^2\alpha}{(\alpha - 1)^2(\alpha - 2)}.$$

For the quantiles, we must solve $p = 1 - (\beta/x_p)^\alpha$, which gives $x_p = \beta(1 - p)^{-1/\alpha}$, for $0 < p < 1$.

Densités conditionelles

On peut aussi calculer les fonctions de répartitions et densités conditionelles : pour des ensembles $\mathcal{A} \subset \mathbb{R}$ telles que $\mathrm{P}(\mathcal{A}) > 0$ on a

$$F(x \mid X \in \mathcal{A}) = \mathrm{P}(X \leq x \mid X \in \mathcal{A}) = \frac{\mathrm{P}(X \leq x \cap X \in \mathcal{A})}{\mathrm{P}(X \in \mathcal{A})} = \frac{\int_{-\infty}^{x} f(y) I(y \in \mathcal{A})\, dy}{\mathrm{P}(X \in \mathcal{A})}$$

et

$$f(x \mid X \in \mathcal{A}) = \frac{f(x)}{\mathrm{P}(X \in \mathcal{A})} I(x \in \mathcal{A}).$$

Donc

$$\mathrm{E}\{g(X) \mid X \in \mathcal{A}\} = \frac{\mathrm{E}\left\{g(X)\ I(X \in \mathcal{A})\right\}}{\mathrm{P}(X \in \mathcal{A})},$$

Exemple 76. *Soit $X \sim \exp(\lambda)$. Trouver la PDF de X, sachant que (a) $X < u$, (b) $X > u$.*

Pause pensées. *Soit $X \sim U(0,1)$ et $0 < a < b < 1$. Trouver la PDF de X, sachant que $a < X < b$.*

Note to Example 76

(a) Here $\mathcal{A}_1 = (0, u)$, and $\mathrm{P}(X \in \mathcal{A}_1) = 1 - \exp(-u\lambda)$. Hence

$$F(x \mid X \in \mathcal{A}_1) = \begin{cases} 0, & x < 0, \\ \frac{1-\exp(-\lambda x)}{1-\exp(-u\lambda)}, & 0 < x < u, \\ 1, & x \geq u. \end{cases}$$

and the density is obtained by differentiation, giving

$$f(x \mid X \in \mathcal{A}_1) = \begin{cases} \frac{\lambda \exp(-\lambda x)}{1-\exp(-u\lambda)}, & 0 < x < u, \\ 0, & \text{sinon.} \end{cases}$$

(b) In this case the lack of memory property comes into play. $\mathcal{A}_2 = (u, \infty)$, and $\mathrm{P}(X \in \mathcal{A}_2) = \exp(-u\lambda)$. Hence

$$F(x \mid X \in \mathcal{A}_2) = \begin{cases} 0, & x < u, \\ \frac{\exp(-u\lambda)-\exp(-\lambda x)}{\exp(-u\lambda)}, & x \geq u, \end{cases}$$

and the formula here reduces to $1 - \exp\{-(x-u)\lambda\}$, $x > u$, so

$$f(x \mid X \in \mathcal{A}_2) = \begin{cases} \lambda \exp\{-\lambda(x-u)\}, & x > u \\ 0, & \text{sinon.} \end{cases}$$

Thus the distribution of $X - u$ given that $X > u$ is the same as the distribution of X : a lack of memory property.

Fonctions d'une variable continue

On considère souvent $Y = g(X)$, où $g : \mathbb{R} \to \mathbb{R}$ est connue, et on veut calculer F_Y et f_Y à partir de F_X et f_X.

Exemple 77. *Soit $Y = -\log(1-U)$, où $U \sim U(0,1)$. Calculer $F_Y(y)$ et discuter. Calculer aussi la densité et la fonction de répartition de $W = -\log U$. Expliquer.*

Exemple 78. *Soit $Y = Z^2$, où $Z \sim \mathcal{N}(0,1)$. Calculer f_Y.*

On peut résumer nos résultats sur les transformations de variables par ce théorème :

Théorème 79. *Soient $Y = g(X)$, $\mathcal{B}_y = (-\infty, y]$ et $g^{-1}(\mathcal{B}_y) = \{x \in \mathbb{R} : g(x) \leq y\}$, alors*

$$F_Y(y) = \mathrm{P}(Y \leq y) = \begin{cases} \int_{g^{-1}(\mathcal{B}_y)} f_X(x)\,dx, & X \text{ continue}, \\ \sum_{x \in g^{-1}(\mathcal{B}_y)} f_X(x), & X \text{ discrète}. \end{cases}$$

Si X est continue et g est monotone avec inverse g^{-1} différentiable, alors $Y = g(X)$ a densité

$$f_Y(y) = \left| \frac{dg^{-1}(y)}{dy} \right| f_X\{g^{-1}(y)\}, \quad y \in \mathbb{R}.$$

Pause pensées. *Donner g, $\mathcal{B}_y$ et $g^{-1}(\mathcal{B}_y)$ pour les exemples 77 et 78.*

Note to Example 77

Now $0 < 1 - U < 1$, so $\log(1-U)$ is well-defined and $Y = -\log(1-U) > 0$. Hence

$$\mathrm{P}(Y \leq y) = \mathrm{P}\{-\log(1-U) \leq y\} = \mathrm{P}\{U \leq 1 - \exp(-y)\} = 1 - \exp(-y), \quad y > 0,$$

which is the exponential CDF ; note that the transformation here is monotone. Thus Y has an exponential distribution.
For $W = -\log U$, we have

$$\begin{aligned} \mathrm{P}(W \leq w) &= \mathrm{P}\{-\log(U) \leq w\} \\ &= \mathrm{P}\{\log U \geq -w\} \\ &= \mathrm{P}(U \geq e^{-w}) \\ &= 1 - \mathrm{P}(U < e^{-w}) = 1 - e^{-w}, \quad w > 0, \end{aligned}$$

where the $<$ can become an $\leq$ because there is no probability at individual points in $\mathbb{R}$.
Hence W also has an exponential distribution. This is obvious, because if $U \sim U(0,1)$, then $1 - U \sim U(0,1)$ also.

Note to Example 78

— $Y = Z^2$ can only take positive values, so $F_Y(y) = \mathrm{P}(Y \leq y) = 0$ for $y < 0$, and then $f_Y(y) = 0$.

— When $y \geq 0$, $Y \leq y$ if and only if $Z^2 \leq y$ and this occurs if and only if $-\sqrt{y} \leq Z \leq \sqrt{y}$. Thus

$$\begin{aligned} \mathrm{P}(Y \leq y) &= \mathrm{P}(-\sqrt{y} \leq X \leq \sqrt{y}) \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}) \\ &= \Phi(\sqrt{y}) - \Phi(-\sqrt{y}). \end{aligned}$$

Differentiation with respect to y and using the chain rule then gives

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y}\left\{\Phi(\sqrt{y}) - \Phi(-\sqrt{y})\right\} = \tfrac{1}{2}y^{-1/2}\phi(\sqrt{y}) - \left(-\tfrac{1}{2}\right)y^{-1/2}\phi(-\sqrt{y}) = \frac{1}{\sqrt{2\pi y}}e^{-y/2}, \quad y > 0,$$

because $\Phi(x) = \int_{-\infty}^{x} \phi(z)\,\mathrm{d}z$ with $\phi(z) = (2\pi)^{-1/2}e^{-z^2/2}$, for $z \in \mathbb{R}$.

— Hence

$$f_Y(y) = \begin{cases} \frac{1}{\sqrt{2\pi y}}e^{-y/2}, & y > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Note to Theorem 79

We have

$$\mathrm{P}(Y \in \mathcal{B}) = \mathrm{P}\{g(X) \in \mathcal{B}\} = \mathrm{P}\{X \in g^{-1}(\mathcal{B})\},$$

because $X \in g^{-1}(\mathcal{B})$ if and only if $g(X) \in g\{g^{-1}(\mathcal{B})\} = \mathcal{B}$.
To find $F_Y(y)$ we take $\mathcal{B}_y = (-\infty, y]$, giving

$$F_Y(y) = \mathrm{P}(Y \leq y) = \mathrm{P}\{g(X) \in \mathcal{B}_y\} = \mathrm{P}\{X \in g^{-1}(\mathcal{B}_y)\},$$

which is the formula in the theorem.
When g is monotone increasing with (monotone increasing) inverse g^{-1}, we have $g^{-1}\{(-\infty, y]\} = (-\infty, g^{-1}(y)]$, and hence

$$F_Y(y) = \mathrm{P}\{Y \in \mathcal{B}_y\} = \mathrm{P}\{X \in g^{-1}(\mathcal{B}_y)\} = \mathrm{P}\{X \leq g^{-1}(y)\} = F_X\{g^{-1}(y)\}, \quad y \in \mathbb{R}.$$

In the case of a continuous random variable X, differentiation gives

$$f_Y(y) = \frac{dg^{-1}(y)}{dy} f_X\{g^{-1}(y)\}, \quad y \in \mathbb{R}.$$

When g is monotone decreasing with (monotone decreasing) inverse g^{-1}, we have $g^{-1}\{(-\infty, y]\} = [g^{-1}(y), \infty)$, and hence

$$F_Y(y) = \mathrm{P}\{Y \in \mathcal{B}_y\} = \mathrm{P}\{X \in g^{-1}(\mathcal{B}_y)\} = \mathrm{P}\{X \geq g^{-1}(y)\}, \quad y \in \mathbb{R}.$$

In the case of a continuous density, $F_Y(y) = \mathrm{P}\{X \geq g^{-1}(y)\} = 1 - F_X\{g^{-1}(y)\}$ and differentiation gives

$$f_Y(y) = -\frac{dg^{-1}(y)}{dy} f_X\{g^{-1}(y)\}, \quad y \in \mathbb{R};$$

note that $-dg^{-1}(y)/dy \geq 0$, because $g^{-1}(y)$ is monotone decreasing.
Thus in both cases we can write

$$f_Y(y) = \left| \frac{dg^{-1}(y)}{dy} \right| f_X\{g^{-1}(y)\}, \quad y \in \mathbb{R}.$$

X discrète ou continue ?

	Discrète	Continue
Support S_X	dénombrable	contient un intervalle $(x_-, x_+) \subset \mathbb{R}$
f_X	fonction de masse sans unité $0 \le f_X(x) \le 1$ $\sum_{x\in\mathbb{R}} f_X(x) = 1$	fonction de densité unité $[x]^{-1}$ $0 \le f_X(x)$ $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$
$F_X(a) = \mathrm{P}(X \le a)$	$\sum_{x\le a} f_X(x)$	$\int_{-\infty}^{a} f_X(x)\,dx$
$\mathrm{P}(X \in \mathcal{A})$	$\sum_{x\in\mathcal{A}} f_X(x)$	$\int_{\mathcal{A}} f_X(x)\,dx$
$\mathrm{P}(a < X \le b)$	$\sum_{\{x:a<x\le b\}} f_X(x)$	$\int_a^b f_X(x)\,dx$
$\mathrm{P}(X = a)$	$f_X(a) \ge 0$	$\int_a^a f_X(x)\,dx = 0$
$\mathrm{E}\{g(X)\}$ (si bien définie)	$\sum_{x\in\mathbb{R}} g(x) f_X(x)$	$\int_{-\infty}^{\infty} g(x) f_X(x)\,dx$

3.4 Fonctions Génératrices

Fonctions génératrices

Définition 80. *La fonction génératrice des moments et la fonction génératrice des cumulants d'une variable aléatoire X sont définies comme*

$$M_X(t) = \mathrm{E}\left(e^{tX}\right), \quad K_X(t) = \log M_X(t), \quad t \in \mathcal{T},$$

pour $\mathcal{T} = \{t \in \mathbb{R} : M_X(t) < \infty\}$. Noter que $0 \in \mathcal{T}$, car $M_X(0) = 1$.

— M_X est aussi appelé la **transformée de Laplace** de f_X.

— M_X et K_X résume toutes les propriétés de X, on peut écrire

$$M_X(t) = \mathrm{E}\left(\sum_{r=0}^{\infty} \frac{t^r X^r}{r!}\right) = \sum_{r=0}^{\infty} \frac{t^r}{r!}\mathrm{E}(X^r), \quad K_X(t) = \sum_{r=1}^{\infty} \frac{t^r}{r!}\kappa_r,$$

d'où on obtient les moments $\mathrm{E}(X^r)$ et les **cumulants** κ_r par différentiation.

Exemple 81. *Calculer $M_X(t)$ et les cumulants de (a) $X \sim \mathrm{Pois}(\lambda)$, (b) $X \sim \mathcal{N}(\mu, \sigma^2)$.*

Pause pensées. *Vérifier que*

$$\mathrm{E}(X) = K_X'(t)\big|_{t=0} = M_X'(t)\big|_{t=0}, \quad \mathrm{var}(X) = K_X''(t)\big|_{t=0} = M_X''(t)\big|_{t=0} - \left\{M_X'(t)\big|_{t=0}\right\}^2.$$

Note on the relation between M_X and K_X

— First note that $M_X(t) = 1$ when $t = 1$, since $\mathrm{E}(e^{tX}) = \mathrm{E}(1) = 1$; thus $0 \in \mathcal{T}$ for any X.

— Now suppose that $\mathcal{T}$ contains an open set $(-a, a)$ for some $a > 0$, and let $\mu_r = \mathrm{E}(X^r)$ be a shorthand notation for the rth moment of X. Now

$$K_X(t) = \sum_{r=1}^{\infty} \frac{t^r \kappa_r}{r!} = \log M_X(t) = \log\left(\sum_{r=0}^{\infty} \frac{t^r \mu_r}{r!}\right) = \log(1+b) = b - b^2/2 + b^3/3 + \cdots,$$

where $b = t\mu_1 + t^2\mu_2/2! + t^3\mu_3/3! + \cdots$. If we expand and compare coefficients of $t, t^2, t^3, \ldots$ in the two expansions we get

$$\kappa_1 = \mu_1, \quad \kappa_2 = \mu_2 - \mu_1^2, \quad \kappa_3 = \mu_3 - 3\mu_2\mu_1 + 2\mu_1^3, \quad \kappa_4 = \mu_4 - 4\mu_3\mu_1 + 6\mu_2\mu_1^2 - 3\mu_1^4, \quad \ldots,$$

so $\kappa_1 = \mathrm{E}(X)$, $\kappa_2 = \mathrm{var}(X)$, $\kappa_3 = \mathrm{E}\{(X - \mu_1)^3\}$.

Note to Example 81

— (a) We have

$$M_X(t) = \sum_{x=0}^{\infty} e^{xt} \frac{\lambda^x}{x!} e^{-\lambda} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e^t)^x}{x!} = \exp\{\lambda(e^t - 1)\}, \quad t \in \mathbb{R},$$

so $K_X(t) = \lambda(e^t - 1)$, $\mathcal{T} = \mathbb{R}$. Thus $\kappa_r = \lambda$ for all $r = 1, 2, \ldots$.

— (b) We first consider $Z \sim \mathcal{N}(0, 1)$ and compute

$$\mathrm{E}(e^{tZ}) = \int_{-\infty}^{\infty} e^{tz} \times \phi(z)\, dz = \int_{-\infty}^{\infty} e^{tz} \times \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz.$$

The fact that the $\mathcal{N}(\mu, \sigma^2)$ density integrates to 1 for any real μ and positive σ^2 , i.e.,

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}\, dx = 1, \quad \mu \in \mathbb{R}, \sigma > 0$$

implies, on expanding the exponent and re-arranging the result, that

$$\int_{-\infty}^{\infty} \frac{1}{(2\pi)^{1/2}} \exp\left\{-x^2/(2\sigma^2) + x\mu/\sigma^2\right\}\, dx = \sigma \exp\left\{\mu^2/(2\sigma^2)\right\}, \quad \mu \in \mathbb{R}, \sigma \in \mathbb{R}_+.$$

If we take $\sigma = 1, \mu = t$, the left-hand side is M_Z, and the right is $e^{t^2/2}$, valid for any $t \in \mathbb{R}$. Hence $M_Z(t) = \exp(t^2/2)$, with $t \in \mathcal{T}$.

— Now note that

$$\begin{aligned} \mathrm{E}(e^{tX}) &= \mathrm{E}[\exp\{t(\mu + \sigma Z)\}] \\ &= \exp(t\mu)\mathrm{E}[\exp\{(t\sigma)Z\}] \\ &= \exp\{t\mu + (t\sigma)^2/2\} \\ &= \exp(t\mu + t^2\sigma^2/2), \quad t \in \mathbb{R}. \end{aligned}$$

Hence $K(X(t) = t\mu + t^2\sigma^2/2$, and we see that $\mathcal{T} = \mathbb{R}$. Thus $\kappa_1 = \mathrm{E}(X) = \mu$, $\kappa_2 = \mathrm{var}(X) = \sigma^2$, and $\kappa_r = 0$ for all $r = 3, \ldots$.

Fonctions génératrices, II

Proposition 82. *Soient $X_1, \ldots, X_n$ des variables indépendantes avec fonctions génératrices de cumulants $K_1, \ldots, K_n$ définies sur $\mathcal{T}_1, \ldots, \mathcal{T}_n$, et $a, b_1, \ldots, b_n$ des constantes, alors*

$$K_{a+b_1X_1+\cdots+b_nX_n}(t) = at + \sum_{j=1}^{n} K_j(b_j t), \quad t \in \bigcap_{j=1}^{n} b_j^{-1}\mathcal{T}_j.$$

Théorème 83 (Analyse III/IV). *Il existe une bijection entre les fonctions de répartition $F_X(x)$ et les fonctions génératrices des moments $M_X(t)$ (et donc aussi avec les fonctions génératrices des cumulants $K_X(t)$).*

Exemple 84. *Soient $X_j \overset{\text{ind}}{\sim} \mathcal{N}(\mu_j, \sigma_j^2)$, pour $j = 1, \ldots, n$, et $a, b_1, \ldots, b_n$ des constantes, trouver la loi de $W = a + b_1X_1 + \cdots + b_nX_n$.*

— Si $\mathcal{T} = \{0\}$, alors M_X n'est pas utile, mais on peut définir la **fonction caractéristique**

$$\chi_X(t) = \mathrm{E}\left(e^{\mathrm{i}tX}\right), \quad t \in \mathbb{R},$$

aussi appelée la **transformée de Fourier** de f_X. On a alors besoin des outils de l'analyse complexe (théorème de Cauchy, résidus, ...).

— Dans ce cours, on utilisera les fonction génératrices de moments et de cumulants.

Note to Proposition 82

First note that

$$M_{a+b_1X_1+\cdots+b_nX_n}(t) = \mathrm{E}\left\{e^{t(a+b_1X_1+\cdots+b_nX_n)}\right\} = \mathrm{E}\left(e^{at}e^{tb_1X_1}\cdots e^{tb_nX_n}\right)$$

and independence of the X_j implies that this equals

$$\mathrm{E}\left(e^{at}e^{tb_1X_1}\cdots e^{tb_nX_n}\right) = e^{at}\prod_{j=1}^{n}\mathrm{E}\left(e^{tb_jX_j}\right) = e^{at}\prod_{j=1}^{n}M_{X_j}(b_jt).$$

Hence on taking logs we have

$$K_{a+b_1X_1+\cdots+b_nX_n}(t) = at + \sum_{j=1}^{n}K_{X_j}(b_jt),$$

provided all the cumulant-generating functions are well-defined, i.e., in the intersection of values of t for which $b_jt \in \mathcal{T}_j$ for $j = 1, \ldots, n$, i.e., provided $t \in \bigcap_{j=1}^{n} b_j^{-1}\mathcal{T}_j$.

Note to Example 84

Using the independence of the X_j and the fact that

$$M_{X_j}(t) = \exp(t\mu_j + t^2\sigma_j^2/2), \quad t \in \mathbb{R},$$

we see that

$$K_W(t) = t\left(a + \sum_{j=1}^{n} b_j\mu_j\right) + \frac{t^2}{2}\left(\sum_{j=1}^{n} b_j^2\sigma_j^2\right), \quad t \in \mathbb{R},$$

so Theorem 83 gives $W \sim \mathcal{N}(a + \sum_j b_j\mu_j, \sum_j b_j^2\sigma_j^2)$.

3.5 Vecteurs Aléatoires

Loi conjointe

— On veut décrire des variables aléatoires dépendantes.

— Si on a deux variables aléatoires X et Y, l'ensemble des valeurs à considérer pour la paire est $\mathbb{R} \times \mathbb{R} = \mathbb{R}^2$. Donc la loi du couple est spécifié par

$$\mathrm{P}\{(X,Y) \in \mathcal{I} \times \mathcal{J}\} = \mathrm{P}(X \in \mathcal{I} \text{ et } Y \in \mathcal{J}), \quad \mathcal{I}, \mathcal{J} \subset \mathbb{R}.$$

— Si les variables sont **indépendantes**, on a

$$\mathrm{P}(X \in \mathcal{I} \text{ et } Y \in \mathcal{J}) = \mathrm{P}(X \in \mathcal{I})\,\mathrm{P}(Y \in \mathcal{J})$$

et la loi conjointe n'est autre que le produit des lois (dites **marginales**) de X et de Y, i.e., $\mathrm{P}(X \in \mathcal{I})$, et $\mathrm{P}(Y \in \mathcal{J})$, pour tous les intervalles $\mathcal{I}, \mathcal{J} \subset \mathbb{R}$.

— Pour discussion générale on utilisera $X = (X_1, \ldots, X_n)$, mais souvent dans les exemples on aura $n = 2$, et pour alleger la notation on utilisera (X, Y), (S, T), etc.

Définition 85. *La* **loi conjointe** *des variables aléatoires $X_1, \ldots, X_n$ est donnée par les probabilités*

$$\mathrm{P}\{(X_1, \ldots, X_n) \in \mathcal{I}_1 \times \cdots \times \mathcal{I}_n\}$$

pour tous les intervalles $\mathcal{I}_1, \ldots, \mathcal{I}_n \subset \mathbb{R}$.

Vecteurs aléatoires discrets

Définition 86. *Un* **vecteur aléatoire discret** $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ *est tel qu'il existe une liste* $S_X = \{x_1, x_2, \ldots\}$ *finie ou dénombrable telle que* $\mathrm{P}(X \in S_X) = 1$. *On appelle* **fonction de densité discrète** *ou* **fonction de masse**

$$f_X(x) = \mathrm{P}(X = x) = \mathrm{P}(X_1 = x_1, \ldots, X_n = x_n).$$

Exemple 87. *Si on a* k *points* $x_1, \ldots, x_k \in \mathbb{R}^n$, *on peut considérer le vecteur aléatoire* **uniforme** *avec*

$$f_X(x) = \begin{cases} \frac{1}{k}, & x \in \{x_1, \ldots, x_k\}, \\ 0, & \text{sinon.} \end{cases}$$

Exemple 88. *Si on a* n *variables aléatoires* $X_1, \ldots, X_n$ *discrètes indépendantes avec fonctions de masse* $f_{X_1}, \ldots, f_{X_n}$ *prenant leurs valeurs dans des listes finies ou dénombrables* $S_1, \ldots, S_n$, *alors* $X = (X_1, \ldots, X_n)$ *est un vecteur aléatoire discret avec fonction de masse*

$$f_X(x) = f_{X_1}(x_1) \cdots f_{X_n}(x_n), \quad x = (x_1, \ldots, x_n) \in S_1 \times \cdots \times S_n.$$

Pause pensées. *Soient* $X_1 \sim \mathrm{Pois}(\lambda) \perp\!\!\!\perp X_2 \sim B(2, p)$, *trouver la fonction de masse de* $X = (X_1, X_2)$.

Modèle d'Ising

Un modèle pour le ferromagnétisme, défini sur une grille $\mathcal{G}$ de variables aléatoires $X_j \in \{-1, +1\}$. Donc $S = \{-1, +1\}^{|\mathcal{G}|}$ a $2^{|\mathcal{G}|}$ éléments, que l'on appelle des **configurations**, parmi lesquelles l'exemple ci-dessus. La fonction de masse est

$$f_X(x) = Z(\beta, \mu)^{-1} \exp\left\{\beta\left(\sum_{i \sim j} X_i X_j + \mu \sum_j h_j X_j\right)\right\},$$

où $i \sim j$ ssi X_i et X_j sont adjacentes, les h_j correspondent à un champ magnétique externe, β est la **température inverse**, et la '**partition function**'

$$Z(\beta, \mu) = \sum \exp\left\{\beta\left(\sum_{i \sim j} X_i X_j + \mu \sum_j h_j X_j\right)\right\}$$

normalise la fonction de masse ; ici la somme est sur les $2^{|\mathcal{G}|}$ configurations possibles.

Vecteurs aléatoires continues

Définition 89. *Un* **vecteur aléatoire à densité** *est un vecteur aléatoire* $X = (X_1, \ldots, X_n)$ *tel qu'il existe une fonction de densité* $f_{X_1,\ldots,X_n} : \mathbb{R}^n \to [0, \infty)$ *telle que*

$$\mathrm{P}\left\{(X_1, \ldots, X_n) \in \mathcal{I}_1 \times \cdots \times \mathcal{I}_n\right\} = \int_{\mathcal{I}_1} \cdots \int_{\mathcal{I}_n} f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n.$$

— On peut montrer que pour tout ensemble 'raisonnable' $D \subset \mathbb{R}^n$ on a

$$\mathrm{P}\left\{(X_1, \ldots, X_n) \in D\right\} = \int_D \cdots \int f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n.$$

— Comme avant, on a en particulier, que pour ϵ petit,

$$f_{X_1,\ldots,X_n}(x_1, \ldots x_n) \approx \frac{1}{\epsilon^n} \mathrm{P}\left(\bigcap_{j=1}^{n} \left\{X_j \in \left[x_j - \tfrac{\epsilon}{2}, x_j + \tfrac{\epsilon}{2}\right]\right\}\right).$$

— Si $X_1, \ldots, X_n$ sont des variables aléatoires indépendantes à densité, alors $(X_1, \ldots, X_n)$ est aussi à densité, et

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n).$$

Loi marginale

Si on s'intéresse à une seule composante de $(X_1, \ldots, X_n)$, on peut calculer sa fonction de densité individuelle, appelée **marginale**.

Définition 90. *Pour un vecteur aléatoire discret* $X = (X_1, \ldots, X_n)$ *la* **fonction de masse (ou parfois de 'densité') marginale** *de* X_j *est*

$$f_{X_j}(x) = \sum_{(x_1,\ldots,x_n) \in S_X : x_j = x} f_X(x_1, \ldots, x_n).$$

— X_j est une variable aléatoire discrète, donc $f_{X_j}(x) > 0$ seulement pour un nombre fini ou dénombrable de x.

Définition 91. *Pour un vecteur aléatoire continu à densité* $f_{X_1,\ldots,X_n}$, *la* **densité marginale** *de* X_j *est*

$$f_{X_j}(x) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)\, \mathrm{d}x_1 \cdots \mathrm{d}x_{j-1} \mathrm{d}x_{j+1} \cdots \mathrm{d}x_n.$$

Exemple 92. *Calculer les fonctions de densité marginales* f_X *et* f_Y *quand*

$$f_{X,Y}(x, y) \equiv f(x, y) = ce^{-x-y} I(y > x) I(x > 0).$$

Note to Examples 92 and 94

— Clearly the only interesting cases are when $x, y > 0$. In this case the marginal density of X is

$$f_X(x) = c \int_{y=x}^{\infty} e^{-x-y}\, dy = ce^{-2x}, \quad x > 0,$$

and obviously this integrates to unity only if $c = 2$. The marginal density of Y is thus

$$f_Y(y) = 2 \int_{x=0}^{y} e^{-x-y}\, dx = 2e^{-y}(1 - e^{-y}), \quad y > 0,$$

and its integral is $2(1 - 1/2) = 1$, so this is also a valid density function.

— For the conditional densities (Example 94), we have

$$f(y \mid x) = 2e^{-x-y}/(2e^{-2x}) = e^{x-y}, \quad y > x,$$

and

$$f(x \mid y) = 2e^{-x-y}/\{2e^{-y}(1 - e^{-y})\} = e^{-x}/(1 - e^{-y}), \quad 0 < x < y.$$

It is easy to check that both integrate to unity.

Loi conditionelle

Si l'on s'intéresse au comportement de certains composantes de $X_1, \ldots, X_n$ en sachant les valeurs d'autres, on a besoin de la loi **conditionelle**.

Définition 93. *La* **densité/fonction de masse conditionnelle** *de* $X_{-j} = (X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_n)$ *sachant* X_j *est*

$$f_{X_{-j}|X_j}(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n \mid x_j) = \frac{f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)}{f_{X_j}(x_j)}, \quad x \in \mathbb{R}^n,$$

en supposant que $f_{X_j}(x_j) > 0$. *Si* X *est discrète, alors*

$$f_X(x) = \mathrm{P}(X = x), \quad f_{X_{-j}|X_j}(x_{-j} \mid x_j) = \mathrm{P}(X_{-j} = x_{-j} \mid X_j = x_j), \quad x \in S_X.$$

— Des définitions analogues existent pour les fonctions de répartition conditionnelles $F_{X|Y}(x \mid y)$, $F_{Y|X}(y \mid x)$.

— Dans le cas continu on conditionne sur un événement de probabilité zéro, ce qui peut mener au paradoxe de Borel–Kolmogorov. C'est rarement un problème en pratique.

Exemple 94. *Trouver les lois conditionelles de* X *et* Y *dans l'exemple 92.*

Espérance conditionelle

Définition 95. *Si* $\mathrm{E}\{|g(X)| \mid X_j = x_j'\} < \infty$, *alors* **l'espérance conditionnelle de** $g(X)$ **sachant** X_{-j} *est définie comme*

$$\mathrm{E}\{g(X) \mid X_j = x_j'\} = \begin{cases} \sum_{(x_1,\ldots,x_n):x_j=x_j'} g(x) f_{X_{-j}|X_j}(x_{-j} \mid x_j'), & X \text{ discrète}, \\ \int_{(x_1,\ldots,x_n):x_j=x_j'} g(x) f_{X_{-j}|X_j}(x_{-j} \mid x_j') \mathrm{d}x_{-j}, & X \text{ continue}, \end{cases}$$

où l'on écrit $x_{-j} = (x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n)$, $\mathrm{d}x_{-j} = \mathrm{d}x_1 \cdots \mathrm{d}x_{j-1} \mathrm{d}x_{j+1} \cdots \mathrm{d}x_n$, *etc.*

— Ainsi l'espérance et la variance conditionelles de $X_1 \mid X_2 = x_2$ (par exemple) sont

$$\mathrm{E}(X_1 \mid X_2 = x_2), \quad \mathrm{var}(X_1 \mid X_2 = x_2) = \mathrm{E}\left[\{X_1 - \mathrm{E}(X_1 \mid X_2 = x_2)\}^2 \mid X_2 = x_2\right].$$

Théorème 96. *Si les espérances requises existent, alors*

$$\begin{aligned} \mathrm{E}\{g(X)\} &= \mathrm{E}_{X_j}\left[\mathrm{E}\{g(X) \mid X_j\}\right], \\ \mathrm{var}\{g(X)\} &= \mathrm{E}_{X_j}\left[\mathrm{var}\{g(X) \mid X_j\}\right] + \mathrm{var}_{X_j}\left[\mathrm{E}\{g(X) \mid X_j\}\right]. \end{aligned}$$

où E_{X_j} *et* var_{X_j} *représentent l'espérance et la variance par rapport à la loi marginale de* X_j.

Exemple 97. *Dans l'exemple 94, trouver* $\mathrm{E}(Y \mid X = x)$, $\mathrm{var}(Y \mid X = x)$, $\mathrm{E}(Y)$ *et* $\mathrm{var}(Y)$.

Note to Theorem 96

We prove this only in the continuous case, as the discrete case is similar. If the expectations exist and using compact notation, we have

$$\begin{aligned}
\mathrm{E}\{g(X)\} &= \int g(x) f_X(x) \mathrm{d}x \\
&= \int g(x) f_{X_{-j}|X_j}(x_{-j} \mid x_j) f_{X_j}(x_j) \mathrm{d}x \\
&= \int_{x_j} \int g(x) f_{X_{-j}|X_j}(x_{-j} \mid x_j) \mathrm{d}x_{-j} f_{X_j}(x_j) \mathrm{d}x_j, \\
&= \mathrm{E}_{X_j}\left[\mathrm{E}\{g(X) \mid X_j\}\right],
\end{aligned}$$

as required. For the variance, write $\mathrm{E}\{g(X)\} = \mu$ and $\mathrm{E}\{g(X) \mid X_j\} = \mu_j$ for compactness, noting that $\mathrm{E}_{X_j}(\mu_j) = \mu$. Then

$$\begin{aligned}
\mathrm{var}\{g(X)\} &= \mathrm{E}\left[\{g(X) - \mu\}^2\right] \\
&= \mathrm{E}\left[\{g(X) - \mu_j + \mu_j - \mu\}^2\right] \\
&= \mathrm{E}\left[\{g(X) - \mu_j\}^2 + 2(\mu_j - \mu)\{g(X) - \mu_j\} + (\mu_j - \mu)^2\right].
\end{aligned}$$

The first and third terms here can be written

$$\begin{aligned}
\mathrm{E}_{X_j}\left(\mathrm{E}_{X_{-j}|X_j}\left[\{g(X) - \mu_j\}^2 \mid X_j\right]\right) &= \mathrm{E}_{X_j}\left[\mathrm{var}\{g(X) \mid X_j\}\right], \\
\mathrm{E}_{X_j}\left[\mathrm{E}_{X_{-j}|X_j}\left\{(\mu_j - \mu)^2\right\}\right] &= \mathrm{E}_{X_j}\left\{(\mu_j - \mu)^2\right\} = \mathrm{var}_{X_j}\left[\mathrm{E}\{g(X) \mid X_j\}\right],
\end{aligned}$$

and the middle term equals zero, because

$$\begin{aligned}
\mathrm{E}_{X_j}\left(\mathrm{E}_{X_{-j}|X_j}\left[(\mu_j - \mu)\{g(X) - \mu_j\}\right]\right) &= \mathrm{E}_{X_j}\left[(\mu_j - \mu)\mathrm{E}_{X_{-j}|X_j}\{g(X) - \mu_j\}\right] \\
&= \mathrm{E}_{X_j}\{(\mu_j - \mu)(\mu_j - \mu_j)\} \\
&= \mathrm{E}_{X_j}\{(\mu_j - \mu) \times 0\} = 0.
\end{aligned}$$

This gives the required result.

Note to Example 97

Here we have $n = 2$, $X_j = X$, $g(X) = X_{-j} = Y$. We saw in Example 94 that the conditional density of Y given X is

$$f(y \mid x) = e^{x-y}, \quad y > x,$$

so

$$\mathrm{E}(Y \mid X = x) = \int_x^\infty y e^{x-y}\,\mathrm{d}y = \int_0^\infty (x+u)e^{-u}\,\mathrm{d}u = x + 1, \quad x > 0,$$

and

$$\mathrm{E}(Y^2 \mid X = x) = \int_x^\infty y^2 e^{x-y}\,\mathrm{d}y = \int_0^\infty (x+u)^2 e^{-u}\,\mathrm{d}u = x^2 + 2x + 2, \quad x > 0,$$

so $\mathrm{var}(Y \mid X = x) = \mathrm{E}(Y^2 \mid X = x) - \mathrm{E}(Y \mid X = x)^2 = 1$.
Hence

$$\mathrm{E}(Y) = \mathrm{E}_X\{\mathrm{E}(Y \mid X = x)\} = \mathrm{E}(X + 1) = 1 + \int_0^\infty x 2e^{-2x}\,\mathrm{d}x = 1 + \tfrac{1}{2} = 3/2$$

and

$$\mathrm{var}(Y) = \mathrm{E}_X\{\mathrm{var}(Y \mid X = x)\} + \mathrm{var}_X\{\mathrm{E}(Y \mid X = x)\} = \mathrm{E}_X(1) + \mathrm{var}_X(1 + X) = 1 + \mathrm{var}(X) = 5/4.$$

Lois mixtes

On rencontre parfois des lois avec certains X_j discrètes et d'autres continues.

Exemple 98. *Une grande compagnie d'assurance observe que la loi du nombre de sinistres X pendant une année pour ses clients ne suit pas une loi de Poisson. Pour modéliser X, on suppose que, pour chaque client, le nombre de sinistres X pendant une année suit une loi de Poisson $\mathrm{Pois}(y)$, mais que $Y \sim \mathrm{Gamma}(\alpha, \lambda)$. Le nombre moyen de sinistres pour un client avec $Y = y$ est alors $\mathrm{E}(X \mid Y = y) = y$, car certain clients sont plus susceptibles d'être sinistrés que d'autres. Trouver la loi conjointe de (X, Y), la loi marginale de X, et la loi conditionelle de Y sachant $X = x$.*

Note to Example 98

— The joint density is

$$f_{X|Y}(x \mid y) \times f_Y(y) = \frac{y^x}{x!}\exp(-y) \times \frac{\lambda^\alpha y^{\alpha-1}}{\Gamma(\alpha)}\exp(-\lambda y), \quad x \in \{0,1,\ldots,\}, y > 0,$$

for $\lambda, \alpha > 0$. Thus the marginal probability mass function of X is

$$\begin{aligned}
f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)\ dy \\
&= \frac{\lambda^\alpha}{x!\Gamma(\alpha)} \int_0^\infty y^{x+\alpha-1}\exp\{-(\lambda+1)y\}\ dy \\
&= \frac{\lambda^\alpha}{x!\Gamma(\alpha)}(\lambda+1)^{-(x+\alpha)} \int_0^\infty u^{x+\alpha-1}\exp(-u)\ du \quad \text{with } u = (\lambda+1)y \\
&= \frac{\lambda^\alpha}{x!\Gamma(\alpha)}(\lambda+1)^{-(x+\alpha)}\Gamma(x+\alpha), \\
&= \frac{\Gamma(x+\alpha)}{x!\Gamma(\alpha)}\left(\frac{\lambda}{\lambda+1}\right)^\alpha\left(\frac{1}{\lambda+1}\right)^x, \quad x = 0,1,\ldots,
\end{aligned}$$

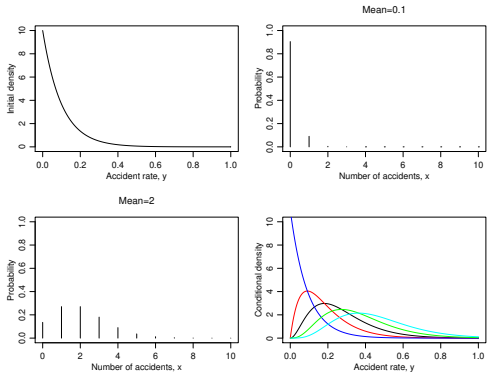
which is negative binomial with parameters $p = \lambda/(\lambda+1)$ and α, in the form given on slide 70.

— The conditional density of Y given that $X = x$ is

$$\begin{aligned}
f_{Y|X}(y \mid x) &= \frac{f_{X,Y}(x,y)}{f_X(x)} \\
&= \frac{f_{X|Y}(x \mid y) f_Y(y)}{f_X(x)} \\
&= \frac{\frac{y^x}{x!}\exp(-y) \times \frac{\lambda^\alpha y^{\alpha-1}}{\Gamma(\alpha)}\exp(-\lambda y)}{\frac{\lambda^\alpha}{x!\Gamma(\alpha)}(\lambda+1)^{-(x+\alpha)}\Gamma(x+\alpha)} \\
&= \frac{(\lambda+1)^{x+\alpha}y^{x+\alpha-1}}{\Gamma(x+\alpha)}\exp\{-y(\lambda+1)\}, \quad y > 0.
\end{aligned}$$

This is gamma with shape parameter $\alpha + x$ and scale parameter $1 + \lambda$.
Hence observing that a customer has x accidents updates our estimate for his/her value of y from the initial mean $\mathrm{E}(Y) = \alpha/\lambda$ to the posterior mean $\mathrm{E}(Y \mid X = x) = (\alpha + x)/(\lambda + 1)$. This is plotted for $x = 0, \ldots, 4$, $\alpha = 1$, $\lambda = 10$ in the figure.

Assurance et apprentissage



Le graphique montre comment la connaissance du nombre d'accidents change la loi pour le taux d'accidents y pour un assuré. En haut à gauche : la densité $f_Y(y)$. En haut à droite : la fonction de masse conditionnelle $f_{X|Y}(x \mid y = 0.1)$ pour un bon conducteur. En bas à gauche : la fonction de masse conditionnelle $f_{X|Y}(x \mid y = 2)$ pour un mauvais conducteur. En bas à droite : les densités conditionnelles $f_{Y|X}(y \mid x)$ avec $x = 0$ (bleu), 1 (rouge), 2 (noir), 3 (vert), 4 (cyan) .

Transformations de variables continues

Théorème 99. *Soit $X \in \mathbb{R}^n$ un vecteur aléatoire continu de densité f_X et la fonction $Y = g(X) \in \mathbb{R}^n$ définie par $Y_j = g_j(X)$ un difféomorphisme (fonction dérivable inversible) :*
(a) le système d'équations $y_1 = g_1(x), \ldots, y_n = g_n(x)$ a solution unique

$$x_1 = h_1(y), \ldots, x_n = h_n(y), \quad y \in \mathbb{R}^n;$$

(b) $g_1, \ldots, g_n$ ont des dérivées continues et

$$J(x) = \det \begin{pmatrix} \frac{\partial g_1}{\partial x_1} & \cdots & \frac{\partial g_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_n}{\partial x_1} & \cdots & \frac{\partial g_n}{\partial x_n} \end{pmatrix} \neq 0 \quad \textit{quand } f_X(x) > 0.$$

Alors

$$f_Y(y) = f_X(x) \times |J(x)|^{-1}\big|_{x_1 = h_1(y), \ldots, x_n = h_n(y)}, \quad y \in \mathbb{R}^n.$$

Exemple 100. *Soient $X_1, X_2 \overset{\text{iid}}{\sim} \exp(\lambda)$, trouver la loi conjointe de*

$$Y_1 = X_1 + X_2, \quad Y_2 = X_1/(X_1 + X_2).$$

Note to Example 100

We write

$$f(x_1, x_2) = \lambda^2 \exp\{-\lambda(x_1 + x_2)\} I(x_1 > 0) I(x_2 > 0).$$

With $Y_1 = X_1 + X_2 > 0$ and $Y_2 = X_1/(X_1 + X_2) \in (0, 1)$, we have

$$y_1 = g_1(x_1, x_2) = x_1 + x_2 > 0, \quad y_2 = g_2(x_1, x_2) = x_1/(x_1 + x_2) \in (0, 1),$$

and the corresponding inverse transformation is

$$x_1 = h_1(y_1, y_2) = y_1 y_2, \quad x_2 = h_2(y_1, y_2) = y_1(1 - y_2), \quad x_1, x_2 > 0.$$

Clearly these transformations satisfy the conditions of Theorem 99. We can either compute

$$J = \begin{vmatrix} 1 & 1 \\ \frac{x_2}{(x_1+x_2)^2} & -\frac{x_1}{(x_1+x_2)^2} \end{vmatrix} = \left| -\frac{(x_1 + x_2)}{(x_1 + x_2)^2} \right| = 1/y_1 > 0,$$

or (maybe better),

$$J^{-1} = \begin{vmatrix} \frac{\partial h_1}{\partial y_1} & \frac{\partial h_1}{\partial y_2} \\ \frac{\partial h_2}{\partial y_1} & \frac{\partial h_2}{\partial y_2} \end{vmatrix} = y_1 > 0.$$

Thus

$$\begin{aligned} f(y_1, y_2) &= \lambda^2 \exp\{-\lambda(x_1 + x_2)\} I(x_1 > 0) I(x_2 > 0) |J^{-1}| \big|_{x_1 = y_1 y_2, x_2 = y_1(1-y_2)} \\ &= y_1 \lambda^2 \exp(-\lambda y_1) I(y_1 y_2 > 0) I\{y_1(1 - y_2) > 0\}, \\ &= y_1 \lambda^2 \exp(-\lambda y_1) I(y_1 > 0) \times I(0 < y_2 < 1) \\ &= f_{Y_1}(y_1) \times f_{Y_2}(y_2). \end{aligned}$$

Integration over y_2 shows that the marginal density of Y_1 is $y_1 \lambda^2 \exp(-\lambda y_1) I(y_1 > 0)$, and so $Y_1 \sim \text{Gamma}(\alpha = 2, \lambda)$ and $Y_2 \sim U(0, 1)$, independently.

Moments conjoints

Définition 101. *Soit $X = (X_1, \ldots, X_n)$ un vecteur aléatoire et $g : \mathbb{R}^n \to \mathbb{R}$. Alors si $\mathrm{E}\{|g(X)|\} < \infty$, on peut définir* **l'espérance** *de $g(X)$ comme*

$$\mathrm{E}\{g(X)\} = \begin{cases} \sum_{x \in S_X} g(x) f_X(x), & \text{cas discret,} \\ \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x) f_X(x)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n, & \text{cas continu.} \end{cases}$$

En particulier on définit **les moments conjoints** *et* **les moments centraux conjoints** *par*

$$\mathrm{E}(X_i^r X_j^s), \quad \mathrm{E}\left[\{X_i - \mathrm{E}(X_i)\}^r \{X_j - \mathrm{E}(X_j)\}^s\right], \quad r, s \in \mathbb{N}.$$

Le plus important est la **covariance** *de X_i et X_j,*

$$\mathrm{cov}(X_i, X_j) = \mathrm{E}\left[\{X_i - \mathrm{E}(X_i)\}\{X_j - \mathrm{E}(X_j)\}\right] = \mathrm{E}(X_i X_j) - \mathrm{E}(X_i)\mathrm{E}(X_j).$$

On définit le **vecteur de l'espérance** *et la* **matrice de (co)variance**,

$$\mathrm{E}(X)_{n\times 1} = \begin{pmatrix} \mathrm{E}(X_1) \\ \vdots \\ \mathrm{E}(X_n) \end{pmatrix}, \quad \mathrm{cov}(X)_{n\times n} = \begin{pmatrix} \mathrm{cov}(X_1, X_1) & \cdots & \mathrm{cov}(X_1, X_n) \\ \vdots & \ddots & \vdots \\ \mathrm{cov}(X_n, X_1) & \cdots & \mathrm{cov}(X_n, X_n) \end{pmatrix}.$$

Parfois on note $\mathrm{E}(X) = \mu_{n\times 1}$, $\mathrm{cov}(X) = \Omega_{n\times n}$, où $\mathrm{E}(X_j) = \mu_j$ et $\mathrm{cov}(X_i, X_j) = \omega_{ij}$, pour $i, j \in \{1, \ldots, n\}$.

Propriétés de la covariance

Lemme 102. *Voici quelques propriétés utiles de la covariance :*

(a) $X_i \perp\!\!\!\perp X_j \quad \Rightarrow \quad \mathrm{cov}(X_i, X_j) = 0$, *mais le contraire est faux ;*

(b) $\mathrm{cov}(X_i, X_j) = \mathrm{cov}(X_j, X_i)$, *donc la matrice Ω est symétrique ;*

(c) $\mathrm{cov}(X_i, X_i) = \mathrm{var}(X_i)$, *donc* $\mathrm{diag}(\Omega) = (\mathrm{var}(X_1), \ldots, \mathrm{var}(X_n))$ *;*

(d) soient $a_{m\times 1}$, $B_{m\times n}$ des constantes et $W = a + BX$, alors

$$\mathrm{E}(W) = a + B\mu, \quad \mathrm{var}(W) = B\Omega B^{\mathrm{T}};$$

(e) si $b \in \mathbb{R}^n$, alors $\mathrm{var}(b^{\mathrm{T}} X) = b^{\mathrm{T}} \Omega b \geq 0$, donc Ω est semi-définie positive (et symétrique) ;

(f) si $X_{n\times 1}$ et $Y_{m\times 1}$ sont des vecteurs aléatoires, on peut écrire

$$\mathrm{cov}(X, Y)_{n\times m} = \mathrm{E}\left[\{X - \mathrm{E}(X)\}\{Y - \mathrm{E}(Y)\}^{\mathrm{T}}\right],$$

et alors pour les vecteurs $a_{p\times 1}$, $c_{q\times 1}$ et matrices $B_{p\times n}$, $D_{q\times m}$ de constantes,

$$\mathrm{cov}(a + BX, c + DY) = B\mathrm{cov}(X, Y)D^{\mathrm{T}}.$$

Note to Lemma 102

These are all easy calculations.

Exemple

Exemple 103. *On tire au hasard et sans remise m boules d'un sac qui contient b boules blanches et n boules noires. Soient $I = (I_1, \ldots, I_m)^{\mathrm{T}}$ le vecteur qui contient les variables indicatrices des événements B_j, "le $j$ième boule est blanche". Trouver $\mathrm{E}(I)$ et $\mathrm{var}(I)$ et en déduire $\mathrm{E}(W)$ et $\mathrm{var}(W)$, où W est le nombre de boules blanches parmi les m boules tirées.*

Note to Example 103

— We must have $m \in \{1, \ldots, b+n\}$, since we can't draw more balls than are in the bag.

— $I_1, \ldots, I_m$ are clearly negatively dependent, because if the first ball is white, then it is less likely that another ball in the sample will also be white.

— Let B_j denote the event that the jth ball is white, let $I = (I_1, \ldots, I_m)^{\mathrm{T}} \in \mathbb{R}^m$ denote the vector of indicator variables, and let 1_m denote the $m \times 1$ vector of ones.

— We have $\mathrm{E}(I_j) = \mathrm{P}(B_j) = b/(b+n)$, so $\mathrm{E}(I) = b/(b+n) \times 1_m$. If this seems surprising (after all, this is the jth ball, so maybe the marginal probability is influenced by what went before?), consider the following thought experiment. We take the m balls from the bag *without looking at them*, put them into another bag, and empty the new bag onto the table, so we see the balls simultaneously. The sample is exactly the same as before, so the probabilities must be the same, but clearly every ball we see has the same probability of being white, and this must be $b/(b+n)$.

— To compute the variance matrix, we need $\mathrm{var}(I_j)$ and $\mathrm{cov}(I_j, I_k)$, for $j \neq k$. Since I_j is an indicator variable, it has variance $p(1-p)$ with $p = b/(b+n)$, i.e.,

$$\mathrm{var}(I_j) = \omega_{jj} = \frac{b}{b+n}\left(1 - \frac{b}{b+n}\right) = \frac{bn}{(b+n)^2}, \quad j = 1, \ldots, n.$$

Also,

$$\begin{aligned} \mathrm{cov}(I_j, I_k) &= \mathrm{E}(I_j I_k) - \mathrm{E}(I_j)\mathrm{E}(I_k) = \mathrm{P}(B_j \cap B_k) - \mathrm{P}(B_j)\mathrm{P}(B_k) \\ &= \mathrm{P}(B_k \mid B_j)\mathrm{P}(B_j) - \mathrm{P}(B_j)\mathrm{P}(B_k) = \frac{b-1}{b+n-1} \times \frac{b}{b+n} - \left(\frac{b}{b+n}\right)^2, \end{aligned}$$

which gives

$$\mathrm{cov}(I_j, I_k) = \omega_{jk} = -\frac{bn}{(b+n)^2(b+n-1)}, \quad j \neq k;$$

as anticipated, this is negative. Hence

$$\mathrm{E}(I) = \frac{b}{b+n} 1_m, \quad \mathrm{cov}(I) = \frac{bn}{(b+n)^2} \begin{pmatrix} 1 & -\frac{1}{b+n-1} & \cdots & -\frac{1}{b+n-1} \\ -\frac{1}{b+n-1} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ -\frac{1}{b+n-1} & \cdots & -\frac{1}{b+n-1} & 1 \end{pmatrix}.$$

— To compute $\mathrm{E}(W)$ and $\mathrm{var}(W)$, we can use the fact that $W = 1_m^{\mathrm{T}} I$, so

$$\begin{aligned} \mathrm{E}(W) &= 1_m^{\mathrm{T}} \mathrm{E}(I) = 1_m^{\mathrm{T}} \frac{b}{b+n} \times 1_m = \frac{mb}{b+n}, \\ \mathrm{var}(W) &= 1_m^{\mathrm{T}} \Omega 1_m = \sum_{j=1}^{m} \mathrm{var}(I_j) + \sum_{j \neq k} \mathrm{cov}(I_j, I_k) \\ &= m\mathrm{var}(I_1) + m(m-1)\mathrm{cov}(I_1, I_2) = \frac{mbn(b+n-m)}{(b+n)^2(b+n-1)}. \end{aligned}$$

— If $m = 1$, we obtain the variance of a single Bernoulli variable with probability $b/(b+n)$. If $m = b+n$ then $\mathrm{var}(W) = 0$ because $W = b$ with probability one when we take all the balls.

— If m is fixed and $b, n \to \infty$ in such a way that $b/(b+n) \to p \in (0,1)$, the covariance tends to zero, because removing one ball from a very large number barely influences the outcome of another ball. Moreover in this case $\mathrm{var}(W) \to mp(1-p)$, corresponding to the binomial variance, because the negative dependence disappears when the number of balls is huge.

Fonctions génératrices

Définition 104. *La* **fonction génératrice des moments** *d'un vecteur aléatoire* X *est*

$$M_X(t) = \mathrm{E}(e^{t^{\mathrm{T}}X}) = \mathrm{E}(e^{\sum_{i=1}^n t_i X_i}), \quad t \in \mathcal{T} = \{t \in \mathbb{R}^n : M_X(t) < \infty\},$$

et sa **fonction génératrice de cumulants** *est* $K_X(t) = \log M_X(t)$, $t \in \mathcal{T}$.

On a

— $0 \in \mathcal{T}$, donc $M_X(0) = 1$ et $K_X(0) = 0$;

— si $\mathcal{T}$ contient un ouvert, alors

$$\mu = \mathrm{E}(X) = K_X'(0) = \left.\frac{\partial K_X(t)}{\partial t}\right|_{t=0}, \quad \Omega = \mathrm{var}(X) = \left.\frac{\partial^2 K_X(t)}{\partial t \partial t^{\mathrm{T}}}\right|_{t=0};$$

— si $\mathcal{A}, \mathcal{B} \subset \{1, \ldots, n\}$ et $\mathcal{A} \cap \mathcal{B} = \emptyset$, et on écrit $X_{\mathcal{A}}$ pour le sous-vecteur de X contenant $\{X_j : j \in \mathcal{A}\}$, etc., alors $X_{\mathcal{A}}$ et $X_{\mathcal{B}}$ sont indépendantes ssi

$$M_X(t) = \mathrm{E}(e^{t_{\mathcal{A}}^{\mathrm{T}} X_{\mathcal{A}} + t_{\mathcal{B}}^{\mathrm{T}} X_{\mathcal{B}}}) = M_{X_{\mathcal{A}}}(t_{\mathcal{A}}) M_{X_{\mathcal{B}}}(t_{\mathcal{B}}), \quad t \in \mathcal{T};$$

— la fonction génératrice des moments de $X_{\mathcal{A}}$ égale $M_X(t)$ évaluée avec $t_{\mathcal{B}} = 0$;

— il y a une injection entre les M_Xs et les lois de probabilités.

Loi gaussienne multivariée

Définition 105. *On dit que le vecteur aléatoire* $X = (X_1, \ldots, X_n)^{\mathrm{T}}$ *a une* **distribution gaussienne (ou normale) multivariée** *s'il existe un vecteur* $\mu = (\mu_1, \ldots, \mu_n)^{\mathrm{T}} \in \mathbb{R}^n$ *et une matrice* $\Omega \in \mathbb{R}^{n \times n}$ *symétrique composée d'éléments* ω_{jk} *tels que*

$$u^{\mathrm{T}} X \sim \mathcal{N}(u^{\mathrm{T}}\mu, u^{\mathrm{T}}\Omega u), \quad u \in \mathbb{R}^n;$$

on écrit alors $X \sim \mathcal{N}_n(\mu, \Omega)$.

Théorème 106. *(a) On a*

$$\mathrm{E}(X_j) = \mu_j, \quad \mathrm{var}(X_j) = \omega_{jj}, \quad \mathrm{cov}(X_j, X_k) = \omega_{jk}, \quad j \neq k.$$

(b) La fonction génératrice de moments de X *est* $M_X(u) = \exp(u^{\mathrm{T}}\mu + \frac{1}{2}u^{\mathrm{T}}\Omega u)$, $u \in \mathbb{R}^n$.
(c) Soient $\mathcal{A}, \mathcal{B} \subset \{1, \ldots, n\}$ *avec* $\mathcal{A} \cap \mathcal{B} = \emptyset$ *alors*

$$X_{\mathcal{A}} \perp\!\!\!\perp X_{\mathcal{B}} \quad \Leftrightarrow \quad \Omega_{\mathcal{A},\mathcal{B}} = 0,$$

où $\Omega_{\mathcal{A},\mathcal{B}}$ *contient* ω_{ij} *pour* $i \in \mathcal{A}$, $j \in \mathcal{B}$.
(d) Soient $X_1, \ldots, X_n \overset{\mathrm{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, *alors* $X = (X_1, \ldots, X_n)^{\mathrm{T}} \sim \mathcal{N}_n(\mu 1_n, \sigma^2 I_n)$.
(e) Les combinaisons linéaires de variables gaussiennes satisfont

$$a_{m\times 1} + B_{m\times n} X \sim \mathcal{N}_m(a + B\mu, B\Omega B^{\mathrm{T}}).$$

Note to Theorem 106

(a) Let e_j denote the n-vector with 1 in the jth place and zeros everywhere else. Then $X_j = e_j^{\mathrm{T}} X \sim N(\mu_j, \omega_{jj})$, giving the mean and variance of X_j.
Now $\mathrm{var}(X_j + X_k) = \mathrm{var}(X_j) + \mathrm{var}(X_k) + 2\mathrm{cov}(X_j, X_k)$, and

$$X_j + X_k = (e_j + e_k)^{\mathrm{T}} X \sim \mathcal{N}(\mu_j + \mu_k, \omega_{jj} + \omega_{kk} + 2\omega_{jk}),$$

which implies that $\mathrm{cov}(X_j, X_k) = \omega_{jk} = \omega_{kj}$.
(b) Since $u^{\mathrm{T}} X \sim \mathcal{N}(u^{\mathrm{T}}\mu, u^{\mathrm{T}}\Omega u)$, its MGF is $M_{u^{\mathrm{T}}X}(t) = \mathrm{E}(e^{tu^{\mathrm{T}}X}) = \exp(tu^{\mathrm{T}}\mu + \frac{1}{2}t^2 u^{\mathrm{T}}\Omega u)$. The MGF of X is $M_X(u) = \mathrm{E}(e^{u^{\mathrm{T}}X}) = M_{u^{\mathrm{T}}X}(1) = \exp(u^{\mathrm{T}}\mu + \frac{1}{2}u^{\mathrm{T}}\Omega u)$, for any $u \in \mathbb{R}^p$, as stated.
(c) Without loss of generality, let $X_{\mathcal{A}} = (X_1, \ldots, X_q)^{\mathrm{T}}$, for $1 \leq q < p$, and partition $t^{\mathrm{T}} = (t_{\mathcal{A}}^{\mathrm{T}}, t_{\mathcal{B}}^{\mathrm{T}})$, $\mu^{\mathrm{T}} = (\mu_{\mathcal{A}}^{\mathrm{T}}, \mu_{\mathcal{B}}^{\mathrm{T}})$, etc. Also without loss of generality suppose that $\mathcal{A} \cup \mathcal{B} = \{1, \ldots, n\}$, since otherwise we can just set $t_j = 0$ for $j \notin \mathcal{A} \cup \mathcal{B}$. Then, using matrix algebra, the joint CGF of X can be written as

$$K_X(t) = t^{\mathrm{T}}\mu + \tfrac{1}{2}t^{\mathrm{T}}\Omega t = t_{\mathcal{A}}^{\mathrm{T}}\mu_{\mathcal{A}} + t_{\mathcal{B}}^{\mathrm{T}}\mu_{\mathcal{B}} + \tfrac{1}{2}t_{\mathcal{A}}^{\mathrm{T}}\Omega_{\mathcal{A}\mathcal{A}}t_{\mathcal{A}} + \tfrac{1}{2}t_{\mathcal{B}}^{\mathrm{T}}\Omega_{\mathcal{B}\mathcal{B}}t_{\mathcal{B}} + t_{\mathcal{A}}^{\mathrm{T}}\Omega_{\mathcal{A}\mathcal{B}}t_{\mathcal{B}}.$$

This equals the sum of the CGFs of $X_{\mathcal{A}}$ and $X_{\mathcal{B}}$, i.e.,

$$K_{X_{\mathcal{A}}}(t) + K_{X_{\mathcal{B}}}(t) = t_{\mathcal{A}}^{\mathrm{T}}\mu_{\mathcal{A}} + \tfrac{1}{2}t_{\mathcal{A}}^{\mathrm{T}}\Omega_{\mathcal{A}\mathcal{A}}t_{\mathcal{A}} + t_{\mathcal{B}}^{\mathrm{T}}\mu_{\mathcal{B}} + +\tfrac{1}{2}t_{\mathcal{B}}^{\mathrm{T}}\Omega_{\mathcal{B}\mathcal{B}}t_{\mathcal{B}}$$

if and only if the final term of $K_X(t)$ equals zero for all t, which occurs if and only if $\Omega_{\mathcal{A}\mathcal{B}} = 0$. Hence the elements of the variance matrix corresponding to $\mathrm{cov}(X_r, X_s)$ must equal zero for any $r \in \mathcal{A}$ and $s \notin \mathcal{A}$, as required. Clearly this also holds if $\mathcal{A} \cup \mathcal{B} \neq \{1, \ldots, p\}$.
(d) Each X_j has mean μ and variance σ^2, and since they are independent, $\mathrm{cov}(X_j, X_k) = 0$ for $j \neq k$. If $u \in \mathbb{R}^n$, then $u^{\mathrm{T}}X$ is a linear combination of normal variables, with mean and variance

$$\sum_{j=1}^{n} u_j\mu = u^{\mathrm{T}}\mu 1_n, \quad \sum_{j=1}^{n} u_j^2\sigma^2 = u^{\mathrm{T}}\sigma^2 I_n u,$$

so $X \sim \mathcal{N}_n(\mu 1_n, \sigma^2 I_n)$, as required.
(e) The MGF of $a + BX$ equals

$$\begin{aligned}
\mathrm{E}\left[\exp\{t^{\mathrm{T}}(a + BX)\}\right] &= \mathrm{E}\left[\exp\{t^{\mathrm{T}}a + (B^{\mathrm{T}}t)^{\mathrm{T}}X)\}\right] \\
&= e^{t^{\mathrm{T}}a} M_X(B^{\mathrm{T}}t) \\
&= \exp\{t^{\mathrm{T}}a + (B^{\mathrm{T}}t)^{\mathrm{T}}\mu + \tfrac{1}{2}(B^{\mathrm{T}}t)^{\mathrm{T}}\Omega(B^{\mathrm{T}}t)\} \\
&= \exp\left\{t^{\mathrm{T}}(a + B\mu) + \tfrac{1}{2}t^{\mathrm{T}}(B\Omega B^{\mathrm{T}})t\right\},
\end{aligned}$$

which is the MGF of the $\mathcal{N}_m(a + B\mu, B\Omega B^{\mathrm{T}})$ distribution.

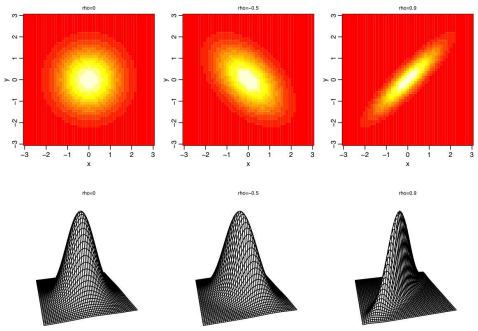
Loi gaussienne multivariée, II

Théorème 107. *$X \sim \mathcal{N}_n(\mu, \Omega)$ a une fonction de densité sur $\mathbb{R}^n$ ssi Ω est de rang n, et alors*

$$f_X(x; \mu, \Omega) = \frac{1}{(2\pi)^{n/2}|\Omega|^{1/2}} \exp\left\{-\tfrac{1}{2}(x-\mu)^{\mathrm{T}}\Omega^{-1}(x-\mu)\right\}, \quad x \in \mathbb{R}^n. \tag{3}$$

Si rang$(\Omega) = m < n$, alors X est une combinaison linéaire de variables gaussiennes ayant une fonction de densité sur un sous-ensemble linéaire de $\mathbb{R}^n$ de dimension m.

La densité normale avec $n = 2$, $\mu_1 = \mu_2 = 0$, $\omega_{11} = \omega_{22} = 1$, et $\omega_{12} = 0, -0.5, 0.9$:



Note to Theorem 107

— Since Ω is symmetric and positive semi-definite, the spectral theorem tells us that we may write $\Omega = ADA^{\mathrm{T}}$, where $D = \mathrm{diag}(d_1, \ldots, d_n)$ contains the (real) eigenvalues of Ω, with $d_1 \geq \cdots \geq d_n \geq 0$, and A is a $n \times n$ orthogonal matrix, i.e., $A^{\mathrm{T}}A = AA^{\mathrm{T}} = I_n$ and $|A| = 1$. The columns $A_1, \ldots, A_n$ of A are the eigenvectors corresponding to the respective eigenvalues,

$$\Omega = ADA^{\mathrm{T}} = \sum_{j=1}^{n} d_j a_j a_j^{\mathrm{T}},$$

with $|\Omega| = |ADA^{\mathrm{T}}| = |A| \times |D| \times |A^{\mathrm{T}}| = |D|$ and $\Omega^{-1} = AD^{-1}A^{\mathrm{T}}$ if the inverse exists.

— Now let $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ $Z = (Z_1, \ldots, Z_n)^{\mathrm{T}}$, and $u \in \mathbb{R}^n$, set and consider

$$u^{\mathrm{T}}(\mu + AD^{1/2}Z) = u^{\mathrm{T}}\mu + \sum_{j=1}^{n} Z_j u^{\mathrm{T}} a_j d_j^{1/2}.$$

This is a linear combination of normal variables, so it has a normal distribution, with mean $u^{\mathrm{T}}\mu$ and variance

$$\mathrm{var}\left(u^{\mathrm{T}}\mu + \sum_{j=1}^{n} Z_j u^{\mathrm{T}} a_j d_j^{1/2}\right) = \sum_{j=1}^{n} d_j (u^{\mathrm{T}} a_j)^2 \mathrm{var}(Z_j) = u^{\mathrm{T}} \left(\sum_{j=1}^{n} d_j a_j a_j^{\mathrm{T}}\right) u = u^{\mathrm{T}}\Omega u,$$

so we can write $X = \mu + AD^{1/2}Z \sim N_n(\mu, \Omega)$, according to Definition 105.

— If Ω has rank n, then $d_n > 0$. The change of variables $z \mapsto x = \mu + AD^{1/2}z$ has Jacobian

$$\left|\frac{\partial x}{\partial z}\right| = |AD^{1/2}| = |A||D|^{1/2} = 1 \times |D|^{1/2} = |\Omega|^{1/2} > 0.$$

Moreover $z = D^{-1/2}A^{\mathrm{T}}(x - \mu)$, and therefore $z^{\mathrm{T}}z = (x-\mu)^{\mathrm{T}}\Omega^{-1}(x-\mu)$. Hence using Theorem 99 and the joint density of Z, $f_Z(z) = (2\pi)^{-n/2}\exp(-\sum_{j=1}^{n} z_j^2/2)$,

$$f_X(x) = f_Z(z)|_{z=D^{-1/2}A^{\mathrm{T}}(x-\mu)} \left|\frac{\partial z}{\partial x}\right| = (2\pi)^{-n/2} \exp\left(-\frac{z^{\mathrm{T}}z}{2}\right)\Big|_{z=D^{-1/2}A^{\mathrm{T}}(x-\mu)} |\Omega|^{-1/2},$$

which reduces to (3). If $d_n = 0$, then the Jacobian is zero, so the transformation $z \mapsto x$ is singular and X does not have a density on $\mathbb{R}^n$.

— Now suppose that $d_m > d_{m+1} = 0$, so just m eigenvalues of Ω are positive. Then

$$X = \mu + \sum_{j=1}^{m} Z_j a_j d_j^{1/2} \in \mathcal{S} = \mu + \mathrm{span}(a_1, \ldots, a_m),$$

where $\mathcal{S}$ is a hyperplane of dimension m passing through μ and generated by the vectors $a_1, \ldots, a_m$. In this case X has an m-dimensional Gaussian density on $\mathcal{S}$, but places no probability elsewhere.

— For example, suppose that $n = 2$, $\mu = (\mu_1, \mu_2)^{\mathrm{T}}$, $a_1 = (1,0)^{\mathrm{T}}$ and $a_2 = (0,1)^{\mathrm{T}}$. If $m = 2$, then $d_1, d_2 > 0$, and X can lie anywhere in $\mathbb{R}^2$, whereas if $m = 1$, then $d_1 > 0$ but $d_2 = 0$, and X can only take values in the line parallel to the x-axis that passes through μ, within which $X_1 \sim \mathcal{N}(\mu_1, d_1)$ and $X_2 = \mu_2$. If $m = 0$, then $X = \mu$ with probability one (it lies in a zero-dimensional subset of $\mathbb{R}^2$).

Distributions marginales et conditionnelles

Théorème 108. *Soit $X \sim \mathcal{N}_n(\mu_{n\times 1}, \Omega_{n\times n})$, ou $|\Omega| > 0$, et soit $\mathcal{A}, \mathcal{B} \subset \{1, \ldots, n\}$ avec $|\mathcal{A}| = s < n, |\mathcal{B}| = t < n$ et $\mathcal{A} \cap \mathcal{B} = \emptyset$. Soient $\mu_{\mathcal{A}}$, $\Omega_{\mathcal{A}}$ et $\Omega_{\mathcal{AB}}$ respectivement le $s \times 1$ sous-vecteur de μ et les $s \times s$ et $s \times t$ sous-matrices de Ω conformées avec $\mathcal{A}$, $\mathcal{A} \times \mathcal{A}$, and $\mathcal{A} \times \mathcal{B}$. Alors*
(a) la loi marginale de $X_{\mathcal{A}}$ est normale, $X_{\mathcal{A}} \sim \mathcal{N}_s(\mu_{\mathcal{A}}, \Omega_{\mathcal{A}})$; et
(b) la loi conditionelle de $X_{\mathcal{A}}$ sachant $X_{\mathcal{B}} = x_{\mathcal{B}}$ est normale,

$$X_{\mathcal{A}} \mid X_{\mathcal{B}} = x_{\mathcal{B}} \sim \mathcal{N}_s \left\{ \mu_{\mathcal{A}} + \Omega_{\mathcal{AB}}\Omega_{\mathcal{B}}^{-1}(x_{\mathcal{B}} - \mu_{\mathcal{B}}), \Omega_{\mathcal{A}} - \Omega_{\mathcal{AB}}\Omega_{\mathcal{B}}^{-1}\Omega_{\mathcal{BA}} \right\}.$$

Proof of Theorem 108

Without loss of generality we can permute the elements of X so that the components of $X_{\mathcal{A}}$ appear before those of $X_{\mathcal{B}}$ and write $X^{\mathrm{T}} = (X_{\mathcal{A}}^{\mathrm{T}}, X_{\mathcal{B}}^{\mathrm{T}})$. Partition the vectors t, μ, and the matrix Ω conformally with X, using obvious notation, and note that $\Omega_{\mathcal{BA}^{\mathrm{T}}} = \Omega_{\mathcal{AB}}$ by symmetry of Ω.
(a) The CGF of X is

$$\begin{aligned} K_X(t) &= \log \mathrm{E}\{\exp(t^{\mathrm{T}}X)\} \\ &= t^{\mathrm{T}}\mu + \tfrac{1}{2}t^{\mathrm{T}}\Omega t \\ &= \begin{pmatrix} t_{\mathcal{A}} \\ t_{\mathcal{B}} \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} \mu_{\mathcal{A}} \\ \mu_{\mathcal{B}} \end{pmatrix} + \tfrac{1}{2} \begin{pmatrix} t_{\mathcal{A}} \\ t_{\mathcal{B}} \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} \Omega_{\mathcal{A}} & \Omega_{\mathcal{AB}} \\ \Omega_{\mathcal{BA}} & \Omega_{\mathcal{B}} \end{pmatrix} \begin{pmatrix} t_{\mathcal{A}} \\ t_{\mathcal{B}} \end{pmatrix} \\ &= t_{\mathcal{A}}^{\mathrm{T}}\mu_{\mathcal{A}} + t_{\mathcal{B}}^{\mathrm{T}}\mu_{\mathcal{B}} + \tfrac{1}{2}\left\{t_{\mathcal{A}}^{\mathrm{T}}\Omega_{\mathcal{A}}t_{\mathcal{A}} + 2t_{\mathcal{A}}^{\mathrm{T}}\Omega_{\mathcal{AB}}t_{\mathcal{B}} + t_{\mathcal{B}}^{\mathrm{T}}\Omega_{\mathcal{B}}t_{\mathcal{B}}\right\}. \end{aligned}$$

We obtain the marginal CGF of $X_{\mathcal{A}}$ by setting $t_{\mathcal{B}} = 0$, which thus gives $\log \mathrm{E}\{\exp(t_{\mathcal{A}}^{\mathrm{T}}X_{\mathcal{A}})\}$, and this is obviously the CGF of the stated normal distribution.
(b) Consider $W = X_{\mathcal{A}} - \Omega_{\mathcal{AB}}\Omega_{\mathcal{B}}^{-1}X_{\mathcal{B}}$. This is a linear combination of normals and so is normal by Theorem 106(e), with mean vector and variance matrix

$$\mu_{\mathcal{A}} - \Omega_{\mathcal{AB}}\Omega_{\mathcal{B}}^{-1}\mu_{\mathcal{B}}, \quad \Omega_{\mathcal{A}} - \Omega_{\mathcal{AB}}\Omega_{\mathcal{B}}^{-1}\Omega_{\mathcal{BA}},$$

and as $\mathrm{cov}(X_{\mathcal{B}}, W) = 0$ and they are normally distributed, $W \perp\!\!\!\perp X_{\mathcal{B}}$. Now

$$X_{\mathcal{A}} = W + \Omega_{\mathcal{AB}}\Omega_{\mathcal{B}}^{-1}X_{\mathcal{B}},$$

and as W and $X_{\mathcal{B}}$ are independent, the distribution of W is unchanged by conditioning on the event $X_{\mathcal{B}} = x_{\mathcal{B}}$. The conditional mean of $X_{\mathcal{A}}$ is therefore

$$\mathrm{E}(X_{\mathcal{A}} \mid X_{\mathcal{B}} = x_{\mathcal{B}}) = \mathrm{E}(W + \Omega_{\mathcal{AB}}\Omega_{\mathcal{B}}^{-1}X_{\mathcal{B}} \mid X_{\mathcal{B}} = x_{\mathcal{B}}) = \mathrm{E}(W) + \Omega_{\mathcal{AB}}\Omega_{\mathcal{B}}^{-1}x_{\mathcal{B}} = \mu_{\mathcal{A}} + \Omega_{\mathcal{AB}}\Omega_{\mathcal{B}}^{-1}(x_{\mathcal{B}} - \mu_{\mathcal{B}})$$

as required. Likewise

$$\mathrm{var}(X_{\mathcal{A}} \mid X_{\mathcal{B}} = x_{\mathcal{B}}) = \mathrm{var}(W + \Omega_{\mathcal{AB}}\Omega_{\mathcal{B}}^{-1}X_{\mathcal{B}} \mid X_{\mathcal{B}} = x_{\mathcal{B}}) = \mathrm{var}(W) = \Omega_{\mathcal{A}} - \Omega_{\mathcal{AB}}\Omega_{\mathcal{B}}^{-1}\Omega_{\mathcal{BA}},$$

because the term involving $X_{\mathcal{B}}$ is conditionally constant. This gives the required result.

Exemple

Exemple 109. *Soit (X_1, X_2) la paire (hauteur (cm), poids (kg)) pour une population de personnes agée de vingt ans. Pour modéliser ceci, on prend*

$$\mu = \begin{pmatrix} 180 \\ 70 \end{pmatrix}, \quad \Omega = \begin{pmatrix} 225 & 90 \\ 90 & 100 \end{pmatrix}.$$

(a) Trouver les lois marginales de X_1 et de X_2, et la corrélation

$$\mathrm{corr}(X_1, X_2) = \frac{\mathrm{cov}(X_1, X_2)}{\{\mathrm{var}(X_1)\mathrm{var}(X_2)\}^{1/2}}.$$

(b) Est-ce que les lois marginales déterminent la loi conjointe ?
(c) Trouver la loi conditionelle de X_2 sachant que $X_1 = x_1$.

Note to Example 109

(a) The marginal distributions are $X_1 \sim \mathcal{N}(180, 225)$ and $X_2 \sim \mathcal{N}(70, 100)$. The correlation is

$$\frac{\omega_{12}}{\sqrt{\omega_{11}\omega_{22}}} = \frac{90}{\sqrt{225 \times 100}} = \frac{90}{150} = 0.6.$$

(b) Clearly not, because they don't tell me the correlation.
(c) For this we have

$$X_{\mathcal{A}} \mid X_{\mathcal{B}} = x_{\mathcal{B}} \sim \mathcal{N}_q \left\{ \mu_{\mathcal{A}} + \Omega_{\mathcal{A}\mathcal{B}}\Omega_{\mathcal{B}}^{-1}(x_{\mathcal{B}} - \mu_{\mathcal{B}}), \Omega_{\mathcal{A}} - \Omega_{\mathcal{A}\mathcal{B}}\Omega_{\mathcal{B}}^{-1}\Omega_{\mathcal{B}\mathcal{A}} \right\}.$$

where $X_{\mathcal{A}} = X_2$, $X_{\mathcal{B}} = X_1$, so

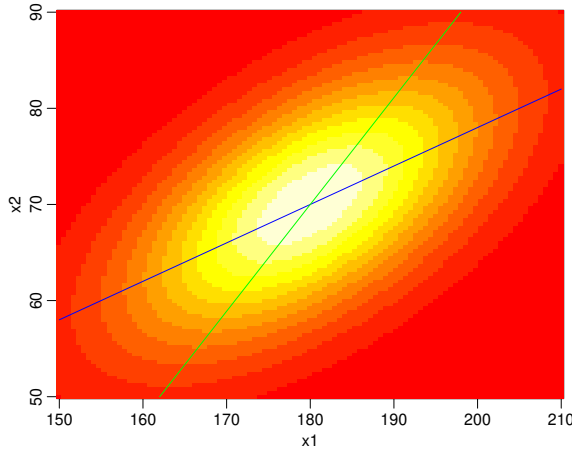
$$\begin{aligned} \mu_{\mathcal{A}} + \Omega_{\mathcal{A}\mathcal{B}}\Omega_{\mathcal{B}}^{-1}(x_{\mathcal{B}} - \mu_{\mathcal{B}}) &= \mu_2 + \omega_{21}\omega_{11}^{-1}(x_1 - \mu_1) = 70 + 0.4(x_1 - 180), \\ \Omega_{\mathcal{A}} - \Omega_{\mathcal{A}\mathcal{B}}\Omega_{\mathcal{B}}^{-1}\Omega_{\mathcal{B}\mathcal{A}} &= 100 - 90^2/225 = 64. \end{aligned}$$

Thus $X_2 \mid X_1 = x_1 \sim \mathcal{N}\{70 + 0.4(x_1 - 180), 64\}$: larger height leads to larger weight, on average.
A similar computation gives

$$X_1 \mid X_2 = x_2 \sim \mathcal{N}\{180 + 0.9(x_2 - 70), 144\}.$$

Loi normale bivariée

La densité normale bivariée pour (X_1, X_2) =(hauteur, poids), ainsi que les droites $\mathrm{E}(X_2 \mid X_1 = x_1) = 70 + 0.4(x_1 - 180)$ (bleu) et $\mathrm{E}(X_1 \mid X_2 = x_2) = 180 + 0.9(x_2 - 70)$ (vert).



3.6 Lois Associées à la loi Normale

Rappel : Loi normale

Définition 110. *Soit $X \sim \mathcal{N}(\mu, \sigma^2)$, avec $\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}_+$, alors*

$$f_X(x; \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}, \quad x \in \mathbb{R}.$$

Propriétés :

— la variable normale standard $Z = (X - \mu)/\sigma$ a moyenne 0 et variance 1, $Z \sim \mathcal{N}(0, 1)$;

— la densité $f_Z(z) = \phi(z) = (2\pi)^{-1/2} e^{-z^2/2}$, pour $z \in \mathbb{R}$, est symétrique autour de $z = 0$, avec

$$\mathrm{P}(Z \leq z) = \Phi(z) = \int_{-\infty}^{z} \phi(x)\,\mathrm{d}x,$$

et $\phi'(z) = -z\phi(z)$, $\phi''(z) = (z^2 - 1)\phi(z)$, etc. ;

— la symétrie implique que $\Phi(z) = 1 - \Phi(-z)$, pour tout $z \in \mathbb{R}$;

— les quantiles z_p de Z satisfont $z_p = \Phi^{-1}(p)$ et (par symétrie) $z_{1-p} = -z_p$.

— Les fonctions `pnorm` et `qnorm` du logiciel `R` pourrait être utile : $\Phi(z) =$ `pnorm(z)`, et $z_p =$ `qnorm(p)` :

p	0.8	0.9	0.95	0.975	0.99	0.995
z_p	0.84	1.28	1.64	1.96	2.33	2.58

La fonction $\Phi(z)$

z	0	1	2	3	4	5	6	7	8	9
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56750	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1.0	.84134	.84375	.84614	.84850	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92786	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
2.0	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
2.3	.98928	.98956	.98983	.99010	.99036	.99061	.99086	.99111	.99134	.99158
2.4	.99180	.99202	.99224	.99245	.99266	.99286	.99305	.99324	.99343	.99361
2.5	.99379	.99396	.99413	.99430	.99446	.99461	.99477	.99492	.99506	.99520
2.6	.99534	.99547	.99560	.99573	.99585	.99598	.99609	.99621	.99632	.99643
2.7	.99653	.99664	.99674	.99683	.99693	.99702	.99711	.99720	.99728	.99736
2.8	.99744	.99752	.99760	.99767	.99774	.99781	.99788	.99795	.99801	.99807
2.9	.99813	.99819	.99825	.99831	.99836	.99841	.99846	.99851	.99856	.99861
3.0	.99865	.99869	.99874	.99878	.99882	.99886	.99889	.99893	.99896	.99900

Loi khi-deux

Définition 111. *Soient* $Z_1, \ldots, Z_\nu \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, *alors* $W = Z_1^2 + \cdots + Z_\nu^2$ *suit la* **loi khi-deux (ou khi-carré) avec** ν **degrés de liberté**, *où* $\nu \in \mathbb{N}$. *Nous notons* $W \sim \chi^2_\nu$, *le* p*-quantile de* W *est* $\chi^2_\nu(p)$, *et la fonction de densité est*

$$f_W(w) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} w^{\nu/2-1} e^{-w/2}, \quad w > 0,\ \nu = 1, 2, \ldots,$$

où la **fonction gamma** $\Gamma(a) = \int_0^\infty u^{a-1} e^{-u}\, du$, *définit pour* $a > 0$, *satisfait* $\Gamma(1) = 1$, $\Gamma(1/2) = \sqrt{\pi}$, $\Gamma(a+1) = a\Gamma(a)$, *et* $\Gamma(n+1) = n!$ *pour* $n = 1, 2, \ldots$.

Propriétés :

— $\mathrm{E}(W) = \nu$, $\mathrm{var}(W) = 2\nu$;

— soient $W_1 \sim \chi^2_{\nu_1}$ et $W_2 \sim \chi^2_{\nu_2}$ indépendantes, alors $W_1 + W_2 \sim \chi^2_{\nu_1+\nu_2}$;

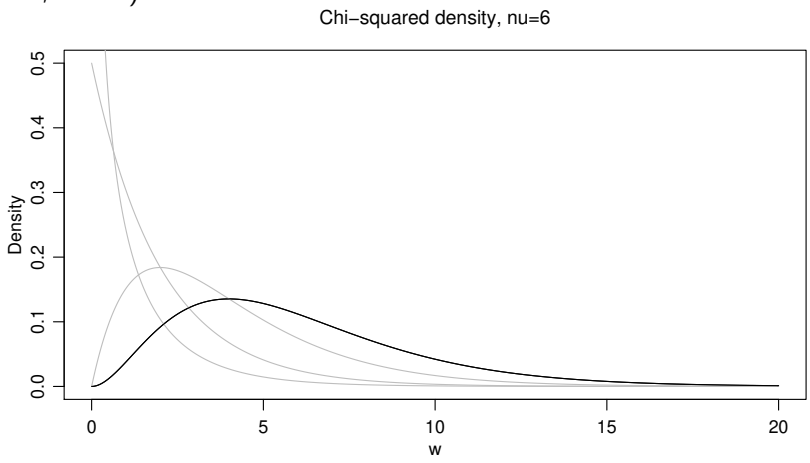
— $W \sim \chi^2_\nu$ implique que W suit la loi de gamma, dont la densité est

$$f(x) = \frac{\beta^\alpha x^{\alpha-1}}{\Gamma(\alpha)} e^{-\beta x}, \quad x > 0, \quad \alpha, \beta > 0,$$

avec paramètres $\alpha = \nu/2$ et $\beta = 1/2$.

Densité χ^2_ν

Ici $\nu = 1, 2, 4, 6$ (noir, $\nu = 6$) :



Loi de Student t

Définition 112. *Soient $Z \sim \mathcal{N}(0,1)$ et $W \sim \chi^2_\nu$ indépendantes, alors $T = Z/(W/\nu)^{1/2}$ suit la* **loi de Student avec ν degrés de liberté**. *Nous notons $T \sim t_\nu$, et on écrit $t_\nu(p)$ pour son p-quantile. Sa densité est*

$$f_T(t) = \frac{\Gamma\{(\nu+1)/2\}}{\sqrt{\nu\pi}\Gamma(\nu/2)} \frac{1}{(1+t^2/\nu)^{(\nu+1)/2}}, \quad -\infty < t < \infty, \ \nu = 1, 2, \ldots.$$

Propriétés :

— son espérance et sa variance existent seulement pour $\nu \geq 2$ et $\nu \geq 3$ respectivement, et alors

$$\mathrm{E}(T) = 0, \quad \mathrm{var}(T) = \frac{\nu}{\nu - 2};$$

— avec $\nu = 1$ on retrouve la **loi de Cauchy**, dont la densité est

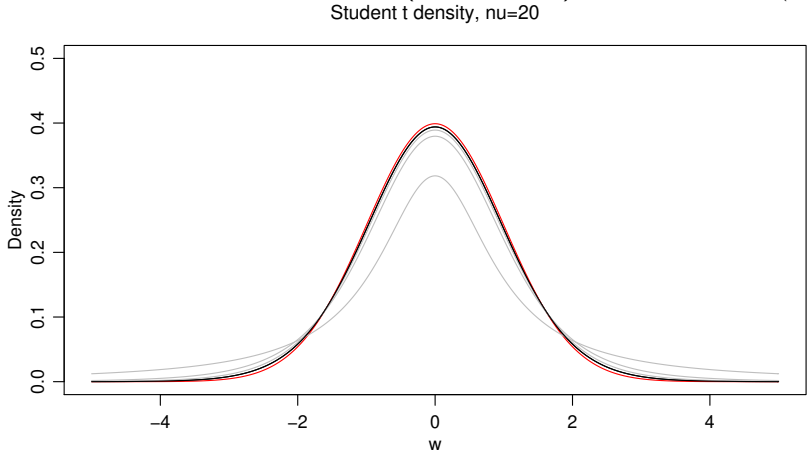
$$\frac{1}{\pi(1+t^2)} \quad -\infty < t < \infty,$$

c'est un exemple classique de loi sans moments ;

— quand $\nu \to \infty$, la loi de Student approche la loi $\mathcal{N}(0,1)$.

Densité Student

Densités de la loi de Student, avec $\nu = 1, 5, 10, 20$ (noir, $\nu = 20$), et la densité $\mathcal{N}(0,1)$ (rouge).



Loi de F

Définition 113. *Soient* $W_1, W_2 \overset{\text{ind}}{\sim} \chi^2_{\nu_1}, \chi^2_{\nu_2}$, *alors*

$$F = \frac{W_1/\nu_1}{W_2/\nu_2}$$

suit la **loi F avec ν_1 et ν_2 degrés de liberté** *: écrivons* $F \sim F_{\nu_1,\nu_2}$. *Sa fonction de densité est*

$$f_F(u) = \frac{\Gamma\left(\frac{1}{2}\nu_1 + \frac{1}{2}\nu_2\right) \nu_1^{\nu_1/2} \nu_2^{\nu_2/2}}{\Gamma\left(\frac{1}{2}\nu_1\right)\Gamma\left(\frac{1}{2}\nu_2\right)} \frac{u^{\frac{1}{2}\nu_1 - 1}}{(\nu_2 + \nu_1 u)^{(\nu_1+\nu_2)/2}}, \quad u > 0,\ \nu_1, \nu_2 = 1, 2, \ldots,$$

et son p-quantile est noté $F_{\nu_1,\nu_2}(p)$.

Logiciels

— Valeurs des quantiles des lois $\mathcal{N}(\mu, \sigma^2)$, χ^2_ν, t_ν, F_{ν_1,ν_2} sont disponible dans des tables.

— Des logiciels tels que R (voir http://www.r-project.org/) ont ces fonctions.

— On peut aussi simuler des valeurs de telles variables (et bien d'autres).

— Exemples :

```
R : Copyright 2005, The R Foundation for Statistical Computing
Version 2.2.1  (2005-12-20 r36812)
...
> qnorm(0.025)    # this is a comment; access normal quantiles
[1] -1.959964     # the [1] means this is the first  element of a vector
> ?qnorm          # help on use of function qnorm()
> qchisq(0.025,df=3)  # chi-squared quantiles, nu=3
[1] 0.2157953
> qt(0.025,df=3)     # t quantiles, nu=3
[1] -3.182446
> qf(0.025,df1=3,df2=4)  # F quantiles, nu1=3, nu2=4
[1] 0.06622087
```

4 Approximation et Convergence

slide 126

Motivation

Il est souvent difficile de calculer la probabilité p exacte d'un événement d'interêt, et on doit l'**approcher**. On pourrait :

— essayer de borner p ;

— utiliser une approximation analytique, souvent par appel aux lois des grands nombres et au théorème central limite ;

— utiliser une approximation numérique, souvent par des méthodes de Monte Carlo.

Les dernières approches utilisent la notion de la **convergence** des suites de variables aléatoires, que nous allons étudier dans ce chapitre.

4.1 Inégalités

slide 128

Inégalités

Théorème 114. *Soient X une variable aléatoire, $a > 0$ constante, h une fonction non-négative et g une fonction convexe, alors*

$$\begin{aligned} \mathrm{P}\{h(X) \geq a\} &\leq \mathrm{E}\{h(X)\}/a, \quad \textit{(inégalité de base)} \\ \mathrm{P}(|X| \geq a) &\leq \mathrm{E}(|X|)/a, \quad \textit{(inégalité de Markov)} \\ \mathrm{P}(|X| \geq a) &\leq \mathrm{E}(X^2)/a^2, \quad \textit{(inégalité de Chebyshov)} \\ \mathrm{E}\{g(X)\} &\geq g\{\mathrm{E}(X)\}. \quad \textit{(inégalité de Jensen)} \end{aligned}$$

— Ces inégalités sont plus utiles pour des calculs théoriques que pour obtenir des bornes pratiques.

— D'autres inégalités peuvent donner des bornes beaucoup plus petites, mais ne sont pas toujours applicables.

Exemple 115. *On test une méthode de classification, dont la probabilité d'une classification correcte est p, sur n cas indépendants. Soient $Y_1, \ldots, Y_n$ les indicatrices des classifications correctes, et $\overline{Y}$ leur moyenne. Pour $\varepsilon = 0.2$ et $n = 100$, borner*

$$\mathrm{P}(|\overline{Y} - p| > \varepsilon).$$

Note to Theorem 114

(a) If $h(x) \geq 0$, then for any $a > 0$, $h(x) \geq h(x)I\{h(x) \geq a\} \geq aI\{h(x) \geq a\}$. Therefore

$$\mathrm{E}\{h(X)\} \geq \mathrm{E}[h(X)I\{h(X) \geq a\}] \geq \mathrm{E}[aI\{h(X) \geq a\}] = a\mathrm{P}\{h(X) \geq a\},$$

and division by $a > 0$ gives the result.
(b) Note that $h(x) = |x|$ is a non-negative function on $\mathbb{R}$, and apply (a).
(c) Note that $h(x) = x^2$ is a non-negative function on $\mathbb{R}$, and that $\mathrm{P}(X^2 \geq a^2) = \mathrm{P}(|X| \geq a)$.
(d) A convex function has the property that, for all y, there exists a value $b(y)$ such that $g(x) \geq g(y) + b(y)(x - y)$ for all x. If $g(x)$ is differentiable, then we can take $b(y) = g'(y)$. To prove this result, we take $y = \mathrm{E}(X)$, and then have

$$g(X) \geq g\{\mathrm{E}(X)\} + b\{\mathrm{E}(X)\}\{X - \mathrm{E}(X)\},$$

and taking expectations of this gives the result.

Note to Example 115

We note that $\sum_{j=1}^{n} Y_j \sim B(n, p)$, so has mean np and variance $np(1-p)$, write $X = \overline{Y} - p$, and note that Chebyshov's inequality gives

$$\mathrm{P}(|\overline{Y} - p| \geq \varepsilon) \leq \mathrm{E}\{(\overline{Y} - p)^2\}/\varepsilon^2 = \mathrm{var}(\overline{Y})/\varepsilon^2 = p(1-p)/(n\varepsilon^2) \leq \tfrac{1}{2}(1 - \tfrac{1}{2})/(100 \times 0.2^2) = 1/16.$$

4.2 Convergence

slide 130

Convergence des variables aléatoires

Définition 116. *Soient $X, X_1, X_2, \ldots$ des variables aléatoires ayant pour fonction de répartition $F, F_1, F_2, \ldots$. Alors*
(a) X_n converge vers X **en moyenne quadratique**, *$X_n \xrightarrow{2} X$, si*

$$\lim_{n\to\infty} \mathrm{E}\{(X_n - X)^2\} = 0, \quad \textit{où} \quad \mathrm{E}(X_n^2), \mathrm{E}(X^2) < \infty;$$

(b) X_n converge vers X **en probabilité**, *$X_n \xrightarrow{P} X$, si pour tout $\varepsilon > 0$,*

$$\lim_{n\to\infty} \mathrm{P}(|X_n - X| > \varepsilon) = 0;$$

(c) X_n converge vers X **en distribution** *(ou* **en loi***), $X_n \xrightarrow{D} X$, si*

$$\lim_{n\to\infty} F_n(x) = F(x) \quad \textit{en tout } x \textit{ où } F(x) \textit{ est continue.}$$

— Si $X_n \xrightarrow{2} X$ ou $X_n \xrightarrow{P} X$, alors il faut que $X_1, X_2, \ldots, X$ soient tous definits par rapport à un seul espace de probabilité. Ceci n'est pas le cas pour $X_n \xrightarrow{D} X$, qui ne concerne que les probabilités. Ce dernier est donc plus faible que les autres.

Liens entre modes de convergence

— Ces modes de convergence sont reliées entre elles comme suit :

$$X_n \xrightarrow{2} X \quad \Rightarrow \quad X_n \xrightarrow{P} X \quad \Rightarrow \quad X_n \xrightarrow{D} X.$$

Toutes autres implications sont fausses en général.

— Les modes les plus importants dans ce cours sont $\xrightarrow{D}$ et $\xrightarrow{P}$, car on souhaite souvent approcher des probabilités, et $\xrightarrow{D}$ nous donne un moyen pour ce faire.

Exemple 117. *Dans l'Exemple 22, soit N_n le nombre de points fixes dans une permutation au hasard de n objets (chapeaux), on a vu que pour $r \in \{0, 1, 2, \ldots\}$ on a*

$$\mathrm{P}(N_n = r) = p_n(r) = \frac{1}{r!} \sum_{k=0}^{n-r} \frac{(-1)^k}{k!} \to \frac{e^{-1}}{r!}, \quad n \to \infty.$$

Donc $\mathrm{P}(N_n \leq x) \to \mathrm{P}(N \leq x)$, où $N \sim \mathrm{Pois}(1)$, pour tout $x \in \mathbb{R}$. Ainsi $N_n \xrightarrow{D} N$.

Exemple 118. *Soient $X_1, X_2, \ldots$ des variables aléatoires indépendantes, d'espérances μ et variances σ^2 finies, alors $\overline{X}_n = n^{-1}(X_1 + \cdots + X_n) \xrightarrow{2} \mu$. Donc $\overline{X}_n \xrightarrow{P} \mu$ et $\overline{X}_n \xrightarrow{D} \mu$.*

Note to Example 118

Let $X = (X_1, \ldots, X_n)^{\mathrm{T}}$, which this has mean vector $\mu 1_n$ and variance matrix $\Omega = \sigma^2 I_n$; the covariances equal zero because the variables are independent. Now $\overline{X}_n = a^{\mathrm{T}} X$, with $a^{\mathrm{T}} = n^{-1} 1_n$, so

$$\begin{aligned} \mathrm{E}(\overline{X}_n) &= \mathrm{E}(a^{\mathrm{T}} X) = a^{\mathrm{T}} \mathrm{E}(X) = n^{-1} 1_n \mu 1_n = \mu, \\ \mathrm{var}(\overline{X}_n) &= a^{\mathrm{T}} \Omega a = n^{-1} 1_n^{\mathrm{T}} \sigma^2 I_n n^{-1} 1_n = \sigma^2 / n. \end{aligned}$$

Therefore

$$\mathrm{E}\{(\overline{X}_n - \mu)^2\} = \mathrm{var}(\overline{X}_n) = \sigma^2/n \to 0, \quad n \to \infty,$$

which implies that $\overline{X} \xrightarrow{2} \mu$, as required. The other two convergence results follow directly.

Théorème de continuité

Théorème 119 (Continuité). *Soient $\{X_n\}$, X des variables aléatoires avec fonctions de répartitions $\{F_n\}$, F, dont les FGMs $M_n(t)$, $M(t)$ existent et $\lim_{n\to\infty} M_n(t) = M(t)$ pour $0 \leq |t| < a$ avec $a > 0$, alors $X_n \xrightarrow{D} X$.*

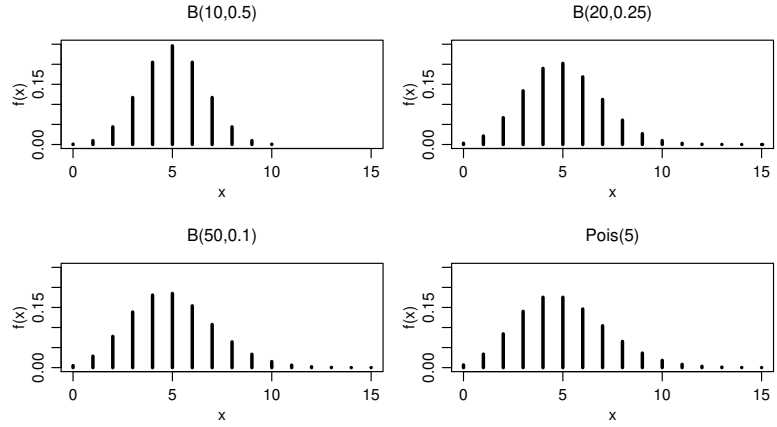
— Ici on peut remplacer $M_n(t)$ et $M(t)$ par les fonctions génératrices des cumulants

$$K_n(t) = \log M_n(t), \quad K(t) = \log M(t).$$

Théorème 120 (Loi des petits nombres). *Soient $X_n \sim B(n, p_n)$ et $np_n \to \lambda > 0$ lorsque $n \to \infty$, alors la fonction de masse limite de X_n est $\mathrm{Pois}(\lambda)$, c'est à dire que $X_n \xrightarrow{D} X$, où $X \sim \mathrm{Pois}(\lambda)$.*

Exemple 121. *Soient $X \sim \mathrm{Geom}(p)$ et $Y_p = pX$, montrer que $\mathrm{E}(Y_p) = 1$, et donner sa loi limite lorsque $p \to 0$.*

Loi des petits nombres



Fonctions de masse de trois lois binomiales et de la loi de Poisson, toutes avec espérance 5.

Note to Theorem 120

The MGF of X_n is

$$M_{X_n}(t) = \mathrm{E}(e^{tX_n}) = \sum_{x=0}^{n} e^{tx} \binom{n}{x} p_n^x (1-p_n)^{n-x} = \left(1 - p_n + p_n e^t\right)^n, \quad t \in \mathcal{T} = \mathbb{R}.$$

Let $p_n = \lambda_n / n$, where $\lambda_n \to \lambda$, and note that as $n \to \infty$,

$$\left(1 - p_n + p_n e^t\right)^n = \left(1 + \frac{\lambda_n (e^t - 1)}{n}\right)^n \to \exp\left\{\lambda (e^t - 1)\right\}, \quad t \in \mathcal{T}.$$

Hence $M_{X_n}(t) \to M_X(t)$, which is the MGF of a Poisson random variable with parameter λ. The continuity theorem implies that $X_n \xrightarrow{D} X$ as $n \to \infty$.

Note to Example 121

— Obviously $\mathrm{E}(Y_p) = p\mathrm{E}(X) = p/p = 1$, for any p. Now $Y_p \in \{p, 2p, 3p, \ldots\}$, so the spacing of the trials shrinks as $p \to 0$, though the mean time to the first success is unchanged.

— The MGF of Y_p is

$$\begin{aligned}\mathrm{E}(e^{tpX}) &= \sum_{x=1}^{\infty} e^{tpx} p(1-p)^{x-1} \\ &= \frac{pe^{tp}}{1-(1-p)e^{tp}} = \frac{1}{p^{-1}e^{-tp} - (1-p)/p} = \frac{1}{1 + (e^{-tp}-1)/p} \to \frac{1}{1-t}, \quad p \to 0,\end{aligned}$$

which is the MGF of $Y \sim \exp(1)$, because

$$\mathrm{E}(e^{tY}) = \int_0^{\infty} e^{ty} \times e^{-y}\,\mathrm{d}y = \int_0^{\infty} e^{-y(1-t)}\,\mathrm{d}y = \frac{1}{1-t}, \quad 1-t > 0;$$

thus here $\mathcal{T} = \{t : t < 1\}$. Hence $Y_p \xrightarrow{D} Y$ as $p \to 0$.

— Hence for very small p the (rescaled) waiting time to the first success in a sequence of independent trials is approximately an exponential variable, i.e., $\mathrm{P}(pX \le t) \approx \mathrm{P}(Y \le t)$, for $t > 0$.

Combinaison de suites convergentes

Théorème 122 (Combinaison de suites convergentes). *Soient x_0, y_0 des constantes, $X, Y, \{X_n\}, \{Y_n\}$ des variables aléatoires, et h une fonction continue en x_0. Alors*

$$\begin{aligned} X_n \xrightarrow{D} x_0 &\Rightarrow X_n \xrightarrow{P} x_0, \\ X_n \xrightarrow{P} x_0 &\Rightarrow h(X_n) \xrightarrow{P} h(x_0), \\ X_n \xrightarrow{D} X \text{ et } Y_n \xrightarrow{P} y_0 &\Rightarrow X_n + Y_n \xrightarrow{D} X + y_0, \quad X_n Y_n \xrightarrow{D} X y_0. \end{aligned}$$

La deux dernières lignes, le **continuous mapping theorem** *et le* **lemme de Slutsky**, *sont très utiles lors des applications statistiques.*

Exemple 123. *Soient $X_1, \ldots, X_n \overset{\text{iid}}{\sim} (\mu_X, \sigma_X^2)$, $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} (\mu_Y, \sigma_Y^2)$, $\mu_X \neq 0$, $\sigma_X^2, \sigma_Y^2 < \infty$, montrer que $R_n \xrightarrow{P} \mu_Y/\mu_X$ lorsque $n \to \infty$, où*

$$R_n = \overline{Y}/\overline{X}, \quad \overline{Y} = n^{-1}\sum_{j=1}^{n} Y_j, \quad \overline{X} = n^{-1}\sum_{j=1}^{n} X_j.$$

Exemple 124. *Soient $X_1, \ldots, X_n \overset{\text{iid}}{\sim} (\mu, \sigma^2)$ et $0 < \sigma^2 < \infty$, montrer que $S^2 = (n-1)^{-1}\sum_{j=1}^{n}(X_j - \overline{X})^2 \xrightarrow{P} \sigma^2$ lorsque $n \to \infty$.*

Note to Example 123

As $\sigma_X^2 < \infty$, by Example 118, $\overline{X} \xrightarrow{2} \mu_X$, and likewise $\overline{Y} \xrightarrow{2} \mu_Y$. Hence $\overline{X} \xrightarrow{P} \mu_X$, and since the function $h(x) = 1/x$ is continuous at $\mu_X \neq 0$, it must be true using line 2 of Theorem 122 that $1/\overline{X} \xrightarrow{P} 1/\mu_X$, a constant. Therefore we have by line 3 of Theorem 122 that

$$R_n = \overline{Y} \times 1/\overline{X} \xrightarrow{D} \mu_Y \times 1/\mu_X,$$

and as this is a constant, line 1 of Theorem 122 implies that $R_n \xrightarrow{P} \mu_Y \times 1/\mu_X$, as required.

Note to Example 124

A little algebra gives

$$S^2 = \frac{n}{n-1}\left(n^{-1}\sum_{j=1}^{n} X_j^2 - \overline{X}^2\right).$$

As $\sigma^2 < \infty$, it follows that $|\mu| < \infty$, and the weak law of large numbers (see below) gives

$$\overline{X} \xrightarrow{P} \mu, \quad Y_n = n^{-1}\sum_{j=1}^{n} X_j^2 \xrightarrow{P} \mathrm{E}(X^2) = \sigma^2 + \mu^2, \quad n \to \infty.$$

As $h(x) = x^2$ is continuous everywhere, $\overline{X}^2 \xrightarrow{P} \mu^2$ by the second line of Theorem 122. Moreover $Y_n \xrightarrow{D} \sigma^2 + \mu^2$, so $Y_n - \overline{X}^2 \xrightarrow{D} \sigma^2 + \mu^2 - \mu^2 = \sigma^2$ by Slutsky's lemma. But as σ^2 is constant, $Y_n - \overline{X}^2 \xrightarrow{P} \sigma^2$. As $n/(n-1) \xrightarrow{D} 1$, another application of Slutsky's lemma and the first line of Theorem 122 give $S^2 \xrightarrow{P} \sigma^2$.

4.3 Lois des Grands Nombres

Lois des grands nombres

Notre première partie de résultats limites est en rapport avec le comportement des moyennes de variable aléatoires indépendantes.

Théorème 125 (Loi faible des grands nombres). *Soient $X_1, X_2, \ldots$ des variables aléatoires indépendantes et identiquement distribuées, d'espérance finie μ. Alors leur moyenne*

$$\overline{X}_n = n^{-1}(X_1 + \cdots + X_n) \xrightarrow{P} \mu;$$

c'est à dire, pour tout $\varepsilon > 0$,

$$\mathrm{P}(|\overline{X}_n - \mu| > \varepsilon) \to 0, \quad n \to \infty.$$

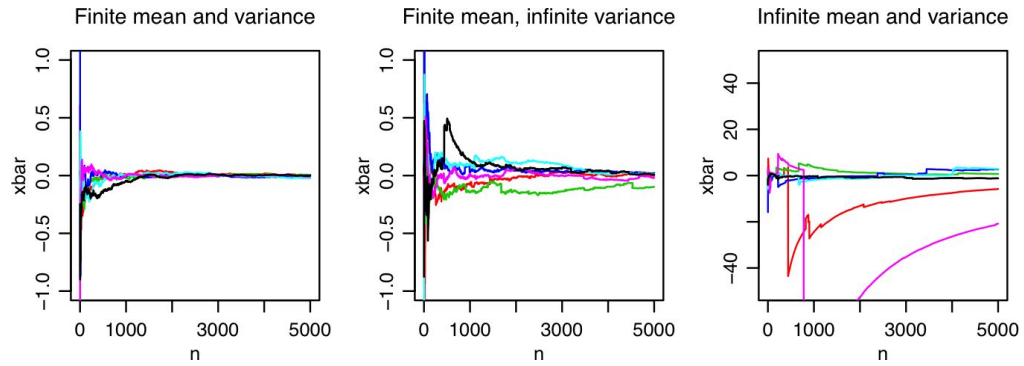
— Ainsi, sous de légères conditions, les moyennes d'échantillons de taille importante convergent vers l'espérance de la distribution dont l'échantillon est issu.

— Dans le cas où les X_i sont des essais de Bernoulli indépendants, nous arrivons enfin à notre notion primitive de probabilité comme limite de fréquences relatives. Le cercle est clos.

Loi faible de grands nombres

— Les graphes ci-dessous montre le comportement de $\overline{X}_n$ quand X_i a deux moments finies (à gauche), seul $\mathrm{E}(|X_i|) < \infty$ (centre), $\mathrm{E}(X_i)$ n'existe pas (et donc $\mathrm{var}(X)$ n'existe pas non plus) (à droite).

— Quand $\mathrm{E}(X_i)$ n'existe pas, la possibilité de valeurs enormes de X_i implique que $\overline{X}_n$ ne peut pas converger.



Remarques

— La loi faible est facile à démontrer si $\mathrm{var}(X_j) = \sigma^2 < \infty$. Dans ce cas, l'exemple 118 nous donne le resultat plus fort que $\overline{X}_n \xrightarrow{2} \mu$, ce qui implique que $\overline{X} \xrightarrow{P} \mu$.

— Une preuve n'est pas trop difficile en utilisant le théorème de la continuité.

— Le même résultat s'applique à de nombreuses statistiques qui peuvent être représentées comme des moyennes, comme par exemple les fonctions de moyennes et les quantiles empiriques.

— Soient $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$, où F est une fonction de répartition continue, et soit $x_p = F^{-1}(p)$ le p quantile de F. En remarquant que la $p$ième quantile empirique

$$X_{(\lceil np \rceil)} \leq x_p \quad \Leftrightarrow \quad \sum_{j=1}^{n} I(X_j \leq x_p) \geq \lceil np \rceil$$

et en appliquant la loi faible à la somme de droite, on a $X_{(\lceil np \rceil)} \xrightarrow{P} x_p$.

Note : sanitized proof of the weak law of large numbers

— La fonction génératrice des cumulants de $\overline{X}_n$ est

$$K_{\overline{X}_n}(t) = \sum_{j=1}^{n} K_{X_j}(tn^{-1}) = nK_{X_j}(tn^{-1}),$$

où (développement de Taylor de 1ère ordre) $K_{X_j}(t) = t\mu + o(t)$ quand $t \to 0$. Ainsi

$$K_{\overline{X}_n}(t) = n\left\{\left(tn^{-1}\mu\right) + o(t/n)\right\} = t\mu + no(t/n) \to t\mu, \quad n \to \infty,$$

soit la FGC d'une variable aléatoire constante, $X = \mu$ avec probabilité 1. Donc le résultat suit par le théorème de la continuité.

— Cette version est 'sanitized' car en générale on doit utiliser la fonction caractéristique.

Note : Convergence of empirical quantiles

— Let $r = \lceil np \rceil$ and note that $r = np + a$, where $a \in [0, 1)$.

— We want to show that $\mathrm{P}(|X_{(r)} - x_p| > \varepsilon) \to 0$ for all $\varepsilon > 0$ as $n \to \infty$. Consider for example

$$\mathrm{P}(X_{(r)} - x_p > \varepsilon) = \mathrm{P}(X_{(r)} > x_p + \varepsilon),$$

and write $F(x_p + \varepsilon) = q > p$, so

$$X_{(r)} > x_p + \varepsilon \quad \Longleftrightarrow \quad \sum_{j=1}^{n} I(X_j \le x_q) < np + a \quad \Longleftrightarrow \quad \overline{X}_n = n^{-1}\sum_{j=1}^{n} I(X_j \le x_q) < p + a/n.$$

Hence

$$\mathrm{P}(X_{(r)} > x_p + \varepsilon) \le \mathrm{P}(\overline{X}_n - q < p - q + 1/n) \to 0,$$

because $\overline{X}_n \xrightarrow{P} q$ implies that $\overline{X}_n - q \xrightarrow{P} 0$ and $p - q + 1/n \to p - q < 0$.

— A similar argument shows that $\mathrm{P}(X_{(r)} - x_p < -\varepsilon) \to 0$, so $\mathrm{P}(|X_{(r)} - x_p| > \varepsilon) \to 0$, i.e., $X_{(\lceil np \rceil)} \xrightarrow{P} x_p$.

Loi forte des grands nombres

En fait, un résultat plus fort est vrai :

Théorème 126 (Loi forte des grands nombres). *Sous les conditions du théorème précédent,*

$$\mathrm{P}\left(\lim_{n\to\infty} \overline{X}_n = \mu\right) = 1.$$

— Ceci est plus fort dans le sens que pour tout $\varepsilon > 0$, la loi faible permet à l'événement $|\overline{X}_n - \mu| > \varepsilon$ de se produire un nombre infini de fois, avec cependant des probabilités de moins en moins petites. La loi forte implique que cet événement peut se produire seulement un nombre fini de fois.

— Les lois faibles et fortes restent valables sous certains types de dépendances parmi les X_j.

4.4 Théorème Central Limite

Théorème central limite

— Comment se comporte $\overline{X}_n - \mu$, quand $n \to \infty$ et on fait un 'zoom', c'est à dire on élargit la différence $\overline{X}_n - \mu$ à la 'bonne vitesse' ?

— On a vu que $\mathrm{E}(\overline{X}_n) = \mu$ et $\mathrm{var}(\overline{X}_n) = \sigma^2/n$, et donc pour tout n

$$Z_n = \frac{\overline{X}_n - \mathrm{E}(\overline{X}_n)}{\mathrm{var}(\overline{X}_n)^{1/2}} = \frac{\overline{X}_n - \mu}{\sqrt{\sigma^2/n}} = n^{1/2}\frac{\overline{X}_n - \mu}{\sigma}$$

satisfait $\mathrm{E}(Z_n) = 0$, $\mathrm{var}(Z_n) = 1$, suggérant de multiplier $\overline{X}_n - \mu$ par $n^{1/2}$ ...

Théorème 127 (**Théorème centrale limite, Central Limit Theorem**). *Soient $X_1, X_2, \ldots$ des variables aléatoires indépendantes et identiquement distribuées d'espérance μ et de variance $0 < \sigma^2 < \infty$, alors,*

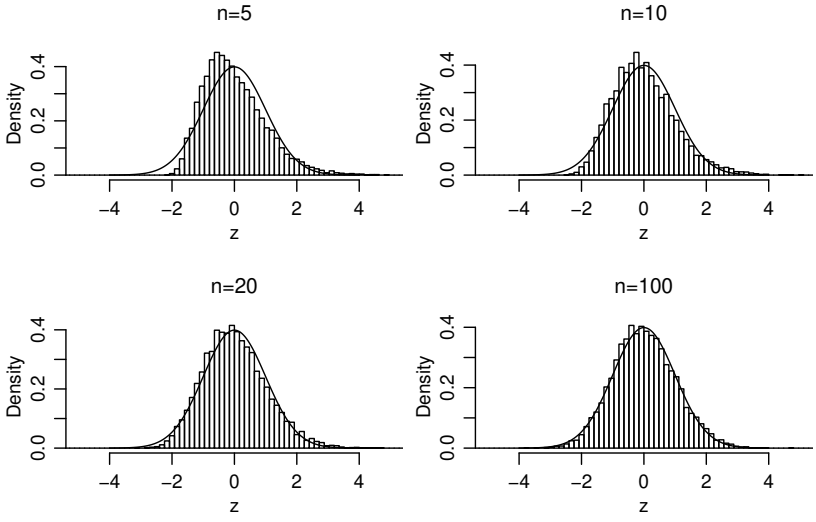
$$Z_n = n^{1/2}\frac{\overline{X}_n - \mu}{\sigma} \xrightarrow{D} Z \sim \mathcal{N}(0,1), \quad n \to \infty.$$

Ceci implique que pour n grand, on a

$$\overline{X}_n = \mu + \sigma n^{-1/2} Z_n \dot{\sim} \mu + \sigma n^{-1/2} Z, \quad Z \sim \mathcal{N}(0,1).$$

Illustration du TCL

Voici le TCL pour $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \exp(1)$; les histogrammes montrent comment les densités empiriques de Z_n s'approchent à la densité de Z.



Note to Theorem 127

Puisque les X_i sont indépendantes et identiquement distribuées, la fonction génératrice des cumulants de

$$Z_n = \left(\overline{X} - \mu\right) / (\sigma^2/n)^{1/2} = \sum_{j=1}^{n} (n^{-1/2}/\sigma) X_j - n^{1/2}\frac{\mu}{\sigma}$$

est la combinaison linéaire

$$K_{Z_n}(t) = \sum_{j=1}^{n} K_{X_j}(tn^{-1/2}/\sigma) - tn^{1/2}\frac{\mu}{\sigma} = nK_{X_j}(tn^{-1/2}/\sigma) - tn^{1/2}\frac{\mu}{\sigma},$$

où l'existence de μ et $\sigma^2 \in (0, \infty)$ implique qu'il existe un $\delta > 0$ tel que pour $|t| < \delta$, on a

$$K_{X_j}(t) = t\mu + t^2\sigma^2/2 + o(t^2).$$

Puisque $no(n^{-1}) \to 0$ quand $n \to \infty$, on a alors

$$\begin{aligned} K_{Z_n}(t) &= n\left\{tn^{-1/2}/\sigma \times \mu + (tn^{-1/2}/\sigma)^2 \times \sigma^2/2 + o(t^2/n\sigma)\right\} - tn^{1/2}\frac{\mu}{\sigma} \\ &= t^2/2 + no(t^2/n\sigma) \\ &\to t^2/2, \quad n \to \infty, \end{aligned}$$

soit la FGC de $Z \sim \mathcal{N}(0,1)$. Le résultat suit par le théorème de continuité.

Utilisation du TCL

Le TCL est utilisé pour approcher des probabilités impliquant des sommes de variables aléatoires indépendantes. Sous les conditions précédentes, on a

$$\mathrm{E}\left(\sum_{j=1}^{n} X_j\right) = n\mu, \quad \mathrm{var}\left(\sum_{j=1}^{n} X_j\right) = n\sigma^2,$$

donc

$$\frac{\sum_{j=1}^{n} X_j - n\mu}{\sqrt{n\sigma^2}} = \frac{n(\overline{X} - \mu)}{\sqrt{n\sigma^2}} = \frac{n^{1/2}(\overline{X} - \mu)}{\sigma} = Z_n$$

peut être approximé par une variable normale :

$$\mathrm{P}\left(\sum_{j=1}^{n} X_j \leq x\right) = \mathrm{P}\left\{\frac{\sum_{j=1}^{n} X_j - n\mu}{\sqrt{n\sigma^2}} \leq \frac{x - n\mu}{(n\sigma^2)^{1/2}}\right\} \doteq \Phi\left\{\frac{x - n\mu}{(n\sigma^2)^{1/2}}\right\}.$$

Exemple

Exemple 128. *Un livre de 640 pages a un nombre d'erreurs aléatoires à chaque page. Si le nombre d'erreurs par page suit une loi de Poisson d'espérance $\lambda = 0.1$, quelle est la probabilité que le livre contienne moins de 50 erreurs ?*

Quand $\sum_{j=1}^{n} X_j$ prend des valeurs entières, on peut obtenir une meilleure approximation en utilisant une correction de la continuité :

$$\mathrm{P}\left(\sum_{j=1}^{n} X_j \leq x\right) \doteq \Phi\left\{\frac{x + \frac{1}{2} - n\mu}{(n\sigma^2)^{1/2}}\right\};$$

ceci peut être important quand la loi de $\sum_{j=1}^{n} X_j$ est assez discrète.

Note to Example 128

We take $\mu = \sigma^2 = 0.1$ and $n = 640$. The expected number of errors is $n\mu = 640\lambda = 64$, and the variance is $n\sigma^2 = 64$, as the variable is Poisson. Thus we seek

$$\mathrm{P}\left(\sum_{j=1}^{n} X_j \leq 49\right) = \mathrm{P}\left(\frac{\sum_{j=1}^{n} X_j - 64}{\sqrt{64}} \leq \frac{49 - 64}{\sqrt{64}}\right) \doteq \Phi(-15/8) = 0.03.$$

The true number is 0.031.

4.5 Méthode Delta

slide 146

La méthode delta

On a souvent besoin de la loi approximative d'une fonction lisse d'une moyenne.

Théorème 129. *Soient $X_1, X_2, \ldots$ des variables aléatoires indépendantes d'espérance μ et de variance $0 < \sigma^2 < \infty$, et soit $g'(\mu) \neq 0$, où g' est la dérivée de g. Alors*

$$\frac{g(\overline{X}_n) - g(\mu)}{\{g'(\mu)^2\sigma^2/n\}^{1/2}} \xrightarrow{D} N(0,1), \quad n \to \infty.$$

Ceci implique que pour n grand, on a $g(\overline{X}_n) \dot{\sim} N\left\{g(\mu), g'(\mu)^2\sigma^2/n\right\}$. On verra plus tard que $S^2 \xrightarrow{P} \sigma^2$, où $S^2 = (n-1)^{-1}\sum_j (X_j - \overline{X}_n)^2$, et donc

$$g(\overline{X}_n) \dot{\sim} N\left\{g(\mu), g'(\overline{X}_n)^2 S^2/n\right\},$$

car $S^2 g'(\overline{X}_n)^2 \xrightarrow{P} \sigma^2 g'(\mu)^2$, et on utilise le lemme de Slutsky (voir Théorème 122).

Exemple 130. *Si $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \exp(\lambda)$, trouver la loi approximative de $\log \overline{X}_n$.*

Note to Theorem 129

We note first that

$$Z_n = \frac{\overline{X}_n - \mu}{\sqrt{\sigma^2/n}} \xrightarrow{D} Z \sim \mathcal{N}(0,1),$$

and therefore that we may write $\overline{X}_n = \mu + \sigma Z_n/\sqrt{n}$. Now

$$g(\overline{X}_n) = g(\mu + \sigma Z_n/\sqrt{n}) = g(\mu) + g'(\mu)\sigma Z_n/\sqrt{n} + o(n^{-1/2}),$$

and this implies that $\mathrm{E}\{g(\overline{X}_n)\} \doteq g(\mu) + o(n^{-1/2})$ and that

$$\mathrm{var}\{g(\overline{X}_n)\} \doteq g'(\mu)^2\sigma^2/n + o(n^{-3/2}).$$

We can also write

$$\frac{g(\overline{X}_n) - g(\mu)}{\{g'(\mu)^2\sigma^2/n\}^{1/2}} = \frac{g(\mu) + g'(\mu)\sigma Z_n/\sqrt{n} + o(n^{-1}) - g(\mu)}{\{g'(\mu)^2\sigma^2/n\}^{1/2}} = Z_n + o(n^{-1/2}) \xrightarrow{D} Z,$$

as $n \to \infty$, which proves the result. Slutsky's lemma is needed only to replace $g'(\mu)^2\sigma^2$ by $g'(\overline{X}_n)S^2 \xrightarrow{P} g'(\mu)^2\sigma^2$.

Note to Example 130

In this example, the X_i have mean $\mu = 1/\lambda$ and variance $\sigma^2 = 1/\lambda^2$, we take $g(x) = \log x$ and $g'(x) = 1/x$. Hence the mean and variance of $\log \overline{X}_n$ are

$$g(\mu) = \log(1/\lambda) = -\log\lambda, \quad g'(\mu)^2\sigma^2/n = 1/(1/\lambda)^2 \times (1/\lambda^2)/n = n^{-1}.$$

This is called a variance-stabilising transformation, as $\mathrm{var}(\log \overline{X}_n)$ does not depend on λ.

Quantiles de l'échantillon

Définition 131. *Soient* $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$, *et* $0 < p < 1$. *Alors le* p **quantile empirique** *des* X_j *est* $X_{(\lceil np \rceil)}$, *où les* **statistiques d'ordre** *sont les valeurs ordonnées de* $X_1, \ldots, X_n$, *soit*

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n-1)} \leq X_{(n)}.$$

Théorème 132. *(Loi asymptotique des statistiques d'ordre) Soient* $0 < p < 1$, $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$, *et* $x_p = F^{-1}(p)$. *Alors si* $f(x_p) > 0$,

$$\frac{X_{(\lceil np \rceil)} - x_p}{[p(1-p)/\{nf(x_p)^2\}]^{1/2}} \xrightarrow{D} N(0,1), \quad n \to \infty.$$

Ceci implique que

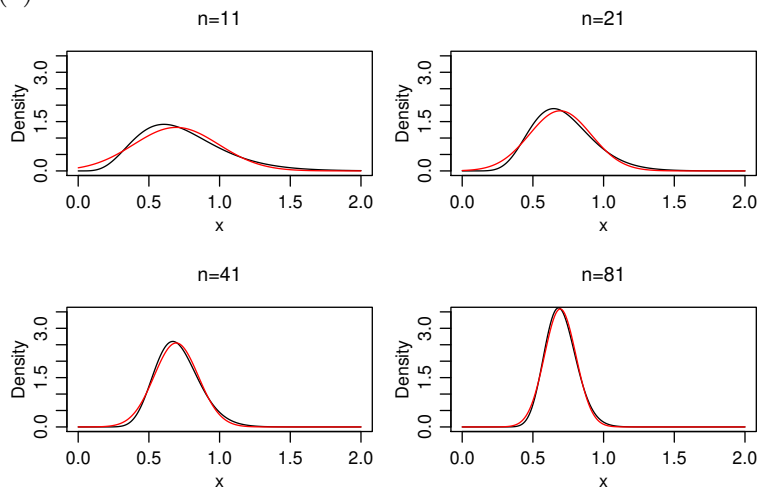
$$X_{(\lceil np \rceil)} \quad \dot{\sim} \quad N\left\{x_p, \frac{p(1-p)}{nf(x_p)^2}\right\}.$$

— Pour prouver ceci, on remarque que $X_{(r)} \leq x$ ssi $S_n = \sum_{j=1}^n I(X_j \leq x) \geq r$, et on applique le TCL à S_n.

Exemple 133. *Démontrer que la loi approximative de la médiane d'un échantillon normal de taille* n *est* $\mathcal{N}\{\mu, \pi\sigma^2/(2n)\}$.

Loi de la médiane

Ce graphique compare les densités exactes (noir) et approchées (rouge) de la médiane $X_{(\lceil n/2 \rceil)}$ pour $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \exp(1)$:



Note to Example 133

We first note that $F^{-1}(1/2)$ where $F(x) = \Phi\{(x - \mu/\sigma\}$ gives $(x - \mu)/\sigma = 0$ and this means that $x_{1/2} = \mu$. Then note that

$$f(\mu) = (2\pi\sigma^2)^{-1/2} \exp\left\{-(\mu - \mu)^2/2\sigma^2\right\} = (2\pi\sigma^2)^{-1/2},$$

so the asymptotic variance of the median is

$$\frac{1}{4} \div \frac{1}{2\pi\sigma^2} \times n^{-1} = \frac{\pi\sigma^2}{2n},$$

which proves the required result.

Note to Theorem 132

— We write $r = \lceil np \rceil = np + a$, where $a \in [0, 1)$, write $x = x_p + \sigma a_n z$, where σ and a_n are to be determined, and note that

$$X_{(r)} \leq x \quad \Longleftrightarrow \quad S_n = \sum_{j=1}^{n} I(X_j \leq x) \geq r,$$

where $S_n \sim B\{n, F(x)\}$ so

$$\begin{aligned} \mathrm{P}(X_{(r)} \leq y) &= \mathrm{P}\left(\frac{X_{(r)} - x_p}{a_n\sigma} \leq z\right) \\ &= \mathrm{P}(S_n \geq r) \\ &= \mathrm{P}(S_n \geq np + a) \\ &= \mathrm{P}\left\{\frac{S_n - nF(x)}{[nF(x)\{1 - F(x)\}]^{1/2}} \geq \frac{np + a - nF(x)}{[nF(x)\{1 - F(x)\}]^{1/2}}\right\}. \end{aligned}$$

— Provided F has a positive density f at x_p,

$$F(x) = F(x_p) + \sigma a_n z f(x_p) + o(a_n) = p + \sigma a_n z f(x_p) + o(a_n),$$

and if $a_n \to 0$ as $n \to \infty$ we see that $F(x)\{1 - F(x)\} \to F(x_p)\{1 - F(x_p)\} = p(1 - p)$. Hence if we set $a_n = n^{-1/2}$ then

$$\frac{np + a - nF(x)}{[nF(x)\{1 - F(x)\}]^{1/2}} = \frac{np + 1 - np - n\sigma a_n z + o(na_n)}{[nF(x)\{1 - F(x)\}]^{1/2}} \to -\frac{\sigma z f(x_p)}{\sqrt{p(1 - p)}},$$

and if we set $\sigma^2 = p(1 - p)/f(x_p)^2$, then this equals $-z$.

— Applying the CLT to the binomial variable S_n as $n \to \infty$, we obtain

$$\mathrm{P}\left(\frac{X_{(r)} - x_p}{a_n\sigma} \leq z\right) = \mathrm{P}\left\{\frac{S_n - nF(x)}{[nF(x)\{1 - F(x)\}]^{1/2}} \geq \frac{np + a - nF(x)}{[nF(x)\{1 - F(x)\}]^{1/2}}\right\} \to 1 - \Phi(-z) = \Phi(z),$$

which is the desired result.

5 Statistique

slide 150

5.1 Notions de Base

slide 151

Que est-ce c'est la statistique ?

— La statistique est la **science de l'apprentissage à partir des données**.

— Points clés :

  — le **contexte** du problème est important—il est essentiel de savoir comment les données ont été récoltées et ce qu'elles représentent ;

  — la **variabilité** (des données) et **l'incertitude** qui en résultent sont représentées avec des modèles probabilistes ;

  — on essaie de **quantifier l'incertitude** quand on tire des conclusions, et

  — de **prendre en compte l'incertitude** quand on choisit des actions suite à une étude.

`http://stat.epfl.ch`

slide 152

Probabilités et données

— Pour connecter les données et les probabilités, on suppose que nos observations y sont aléatoires :

  — soit en imposant un **mécanisme aléatoire**, tel que la **randomisation** d'une expérience ou d'un sondage ;

  — soit en supposant qu'elles sont issues d'une **expérience aléatoire**, p. ex., je suppose que le retard R de mon bus suit une loi $\exp(\lambda)$, puis j'essaie d'estimer $\mathrm{P}(R > 5)$ à partir des observations $r_1, \ldots, r_n$, car je veux arriver à l'heure pour un cours ...

— Souvent, on veut étudier le comportement d'une variable y dans

  — une **population** — l'ensemble sur lequel porte notre investigation — dont on a tiré

  — un **échantillon** $y_1, \ldots, y_n$,

  — et on suppose que cet échantillon est une réalisation des variables aléatoires $Y_1, \ldots, Y_n$ provenant d'un **modèle probabiliste**, F.

— Le probabiliste pose son modèle F, et utilise les lois de probabilité pour **déduire** les propriétés de Y — il en est certain, si son raisonnement est juste !

— Le statisticien fait la démarche opposée : il veut utiliser des données y pour **inférer** des propriétés du modèle F — il en est incertain, car y est finie et il est rarement certain que ses hypothèses sont correctes.

`http://stat.epfl.ch`

slide 153

Nomenclature

Définition 134. *On appelle* **modèle statistique** *une loi (ou une famille des lois) de probabilité construite pour une étude statistique. On appelle* **paramètre (statistique)**, *toute fonction (constante) d'une loi (et on le note généralement avec des lettres gréques). Un modèle determiné par un paramètre de dimension finie est dit* **paramétrique**, *sinon il est dit* **nonparamétrique**.

Définition 135. *On appelle* **statistique** $S = s(Y)$, *toute fonction qui ne dépend que des données* Y. *Ceci comprend les fonctions telle que la moyenne, mais aussi les graphes.*

Définition 136. *La* **loi d'échantionnage (sampling distribution)** *d'une statistique* $S = s(Y)$ *est sa loi de probabilité quand* Y *est généré par un modèle statistique.*

Définition 137. *On appelle* **échantillon aléatoire (random sample)** $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} f$ *ou une réalisation* $y_1, \ldots, y_n$ *de telles* $Y_1, \ldots, Y_n$.

Pause pensées. *Soient* $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, 1)$, *avec* μ *inconnu, et* $S = \overline{Y} = n^{-1} \sum_{j=1}^{n} Y_j$.

(a) Ce modèle est-il nonparamétrique ?

(b) Quelle est la loi d'échantionnage de S *?*

(c) Comment vos réponses changent-elles si $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} (\mu, 1)$ *?*

(d) $\overline{Y}$ *et* $\overline{Y} - \mu$ *sont-elles des statistiques ?*

Note to pause pensées

(a) No, it is parametric, because it is entirely determined by the scalar μ.

(b) S is a linear combination of normal variables, so $S \sim \mathcal{N}(\mu, 1/n)$.

(c) In this case the model is nonparametric, because knowing the mean and variance (the only two properties that are constrained) is not enough to tell me about the model.

(d) $\overline{Y}$ can be calculated from the data, but $\overline{Y} - \mu$ cannot, so only the first is a statistic.

La démarche statistique

Les principles étapes sont :

- **formulation** du sujet de recherche, d'hypothèse(s) ;
- **recherche de données** pertinentes, menant à
  - la **planification** d'une expérience (développement théorique du problème, élaboration du plan expérimental), mise en oeuvre du plan et réception des données ;
  - (si des expériences ne sont pas possibles), une **étude observationnelle**, où la récolte des données n'est pas sous le contrôle de l'investigateur ;
- **analyse des données**, souvent divisée en
  - analyse exploratoire,
  - inférence ;
- **interprétation des résultats** et conclusions pratiques (décisions/actions). Celles-ci sont souvent moins claires suite aux études observationnelles, à cause des **variables confondantes**.

Cette suite d'étapes correspond à une situation idéale qu'on atteint assez rarement en pratique.

Planification

— On souhaite arriver à des conclusions aussi précises que possible, mais il existe toujours des contraintes pratiques.

— Il n'est pas toujours possible de donner une réponse satisfaisante avec les ressources disponibles. De plus une expérience peut échouer pour des raisons imprévisibles.

— La **planification d'expérience (design of experiments/planning of investigations)** est un sujet important pour les expériences agricoles et industrielles, le contrôle de la qualité des processus, les essais cliniques, la gestion de la publicité sur le web, etc.

Exemple 138. *On va mesurer une variable sur n individus indépendants, dont n_1 vont recevoir un traitement, T, et $n_2 = n - n_1$ un placebo, P. Si l'on suppose que la variance des mesures est la même pour tous les individus, comment choisir n_1 pour minimiser la variance de la différence des moyennes,*

$$D = \overline{X}_T - \overline{X}_P = \frac{1}{n_1}\sum_{j=1}^{n_1} X_{T,j} - \frac{1}{n_2}\sum_{j=1}^{n_2} X_{P,j}?$$

Note to Example 138

— Suppose that for the treatment group $X_j \overset{\text{iid}}{\sim} (\mu+\delta, \sigma^2)$, and for the placebo group, $X_j \overset{\text{iid}}{\sim} (\mu, \sigma^2)$. Then if we write $n_1 = nt$, $n_2 = n(1-t)$, for $t \in (0,1)$, we have

$$\begin{aligned} \mathrm{E}(D) &= \mathrm{E}(\overline{X}_T) - \mathrm{E}(\overline{X}_P) = (\mu+\delta) - \mu = \delta, \\ \mathrm{var}(D) &= \mathrm{var}(\overline{X}_T) + \mathrm{var}(\overline{X}_P) = \sigma^2/n_1 + \sigma^2/n_2 = n^{-1}\sigma^2\{1/t + 1/(1-t)\}, \end{aligned}$$

and the variance is minimised over t when $1/t^2 = 1/(1-t)^2$, i.e., $t = 1/2$. Hence we should take the groups so that $n_1 \approx n_2$: then $\mathrm{var}(D) \approx 4\sigma^2/n$.

— If (using the CLT) we use the approximation $D \dot{\sim} \mathcal{N}(\delta, 4\sigma^2/n)$, then with $Z \sim \mathcal{N}(0,1)$ we can write

$$D \doteq \delta + \frac{2\sigma}{n^{1/2}} Z = \frac{2\sigma}{n^{1/2}}\left(\frac{n^{1/2}\delta}{\sigma} + Z\right),$$

so there will be little chance of detecting a treatment effect if $n^{1/2}|\delta|/\sigma$ is small compared to the variation of Z. Thus to increase our chance of detecting a difference of a given size δ, we must either increase n or make measurements more precise by decreasing σ.

Analyse de données

L'analyse de données est souvent décrite comme comprenant deux phases :

— **Phase 1 : l'analyse exploratoire** ("statistique descriptive") a recours principalement à des méthodes simples, flexibles, souvent graphiques. Elle permet d'étudier la structure des données et de détecter des structures spécifiques (tendances, formes, observations atypiques).

— Exemples :

  — dans quel intervalle la majorité de vos tailles se situe-t-elle ?

  — est-ce que vos tailles et vos poids sont associées ?

  — y-a-t il des personnes "extraordinaires" ?

— Cette phase n'utilise pas des idées probabilistes de façon explicite, elle suggère des hypothèses de travail et des modèles pouvant être formalisés et vérifiés dans la Phase 2 (en principe pas avec les mêmes données !)

— **Phase 2 : l'inférence statistique** conduit à des conclusions statistiques en utilisant des notions probabilistes — des méthodes de test, d'estimation et de prévision.

5.2 Quelques Statistiques

Types de variables

— Une variable peut être **quantitative** ou **qualitative**.

— Une **variable quantitative** peut être **discrète** (souvent entière) ou **continue** :

  — variables quantitatives discrètes : p. ex., nombre d'enfants dans une famille ;

  — variables quantitatives continues : p. ex., poids en kilos.

— Une **variable qualitative** (catégorielle) peut être **nominale** (non-ordonnée) ou **ordinale** (ordonnée).

  — variables qualitatives nominales : p. ex., le groupe sanguin (A, B, AB, O)

  — variables qualitatives ordinales : p. ex., le repas au Vinci (bon, passable, mauvais) ;

  Parfois on convertit des variables quantitatives en variables catégorielles : p. ex., la taille (S, M, L, XL, XXL).

Pause pensées. *Lesquelles des lois suivantes pourraient être appropriées pour les exemples ci-dessus : Bernouilli, binomiale, géometrique, binomiale négative, Poisson, exponentielle, gaussienne, Pareto, . . . ?*

Loi multinomiale

Définition 139. *La variable aléatoire* $(X_1, \ldots, X_K)$ *suit une loi* **multinomiale de dénominateur** m **et probabilités** $(p_1, \ldots, p_K)$ *si sa fonction de masse est*

$$f(x_1, \ldots, x_K) = \frac{m!}{x_1! \times \cdots \times x_K!} p_1^{x_1} p_2^{x_2} \cdots p_K^{x_K}, \quad x_1, \ldots, x_K \in \{0, \ldots, m\}, \sum_{k=1}^{K} x_k = m,$$

où $m \in \mathbb{N}$ *et* $p_1 \ldots, p_K \in [0, 1]$, *avec* $p_1 + \cdots + p_K = 1$.

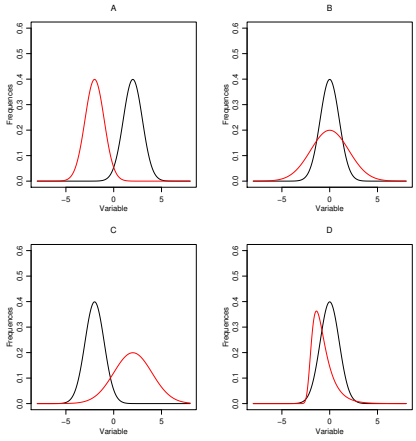
Cette loi apparaît comme la loi du nombre d'individus dans les catégories $\{1, \ldots, K\}$ quand m individus indépendants appartiennent aux catégories avec des probabilités $\{p_1, \ldots, p_K\}$. Elle généralise la loi binomiale à $K > 2$ catégories.

Caractéristiques principales des données

Pour des **variables quantitatives**, on s'intéresse généralement aux caractéristiques suivantes :

1. la **tendance centrale** qui informe sur le "milieu" (la position, le centre), par exemple la moyenne et la médiane ;
2. la **dispersion** qui renseigne sur la variabilité de la distribution autour du centre, par exemple l'étendue, l'écart-type et l'étendue interquartile ;
3. la **symétrie** ou **asymétrie** par rapport au centre ;
4. le nombre de **modes** ("bosses") ;
5. la présence éventuelle de **valeurs aberrantes (outliers)**, qui pourraient provenir d'erreurs de mesures (et donc sont à supprimer), mais pourraient aussi être les données les plus intéressantes, si elles sont correctes.

Formes des densités



A : Densités semblables mais pas le même centre
B : Même centre, dispersions différentes
C : Dispersions et centres différents
D : Asymétrie de la densité rouge

Quantiles

Définition 140. *Soient $y_1, \ldots, y_n$ un groupe d'observations, alors leur* **statistiques d'ordre** *sont les valeurs ordonnées*

$$y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)},$$

et pour $p \in (0,1)$ leur **$p$ème quantile** *est $y_{(\lceil np \rceil)}$. La* **médiane (empirique/de l'échantillon)** *est $y_{(\lceil n/2 \rceil)}$ et les* **quartiles** *sont $y_{(\lceil n/4 \rceil)}$ et $y_{(\lceil 3n/4 \rceil)}$, où $\lceil x \rceil$ dénote le plus petit entier tel que $\lceil x \rceil \geq x$.*

Parfois on parle de **pourcentile** : le p-quantile est le $100p$-pourcentile.

Les quantiles sont utiles car :

— ils sont faciles à calculer ;

— ils suggèrent la forme d'une loi sous-jacente ;

— ils peuvent donner des statistiques qui résistent mieux aux valeurs aberrantes que la moyenne, etc.

Caractéristiques numériques

Indicateurs de tendance centrale (mesures de position) :

— la **moyenne (arithmétique) (average)**,

$$\overline{y} = n^{-1} \sum_{j=1}^{n} y_j;$$

— la **médiane (median)**, $y_{(\lceil n/2 \rceil)}$.

Indicateurs de dispersion :

— **l'écart-type (standard deviation)**,

$$s = \left\{ \frac{1}{n-1} \sum_{j=1}^{n} (y_j - \overline{y})^2 \right\}^{1/2} = \left\{ \frac{1}{n-1} \left(\sum_{j=1}^{n} y_j^2 - n\,\overline{y}^2 \right) \right\}^{1/2},$$

où s^2 est la **variance de l'échantillon** (on verra plus tard pourquoi on divise par $n-1$) ;

— **l'étendue/l'écart-type interquartile (interquartile range, IQR)**,

$$\mathrm{IQR}(y) = y_{(\lceil 3n/4 \rceil)} - y_{(\lceil n/4 \rceil)};$$

— **l'étendue (range)**, $y_{(n)} - y_{(1)} = \max(y_1, \ldots, y_n) - \min(y_1, \ldots, y_n)$.

Mesures de corrélation

— On veut souvent mesurer la dépendance de paires de données $(x_1, y_1), \ldots, (x_n, y_n)$ pour n individus. La corrélation entre deux variables aléatoires (X, Y) est

$$\mathrm{corr}(X, Y) = \frac{\mathrm{cov}(X, Y)}{\{\mathrm{var}(X)\mathrm{var}(Y)\}^{1/2}}.$$

— La **coefficient de la corrélation (empirique) (correlation coefficient)**,

$$r_{xy} = \frac{n^{-1}\sum_{j=1}^n (x_j - \overline{x})(y_j - \overline{y})}{\left\{n^{-1}\sum_{j=1}^n (x_j - \overline{x})^2 \times n^{-1}\sum_{j=1}^n (y_j - \overline{y})^2\right\}^{1/2}},$$

satisfait

(a) $-1 \leq r_{xy} \leq 1$;

(b) si $r_{xy} = \pm 1$, alors les (x_j, y_j) sur une droite, de pente positive si $r_{xy} = 1$, et de pente negative si $r_{xy} = -1$;

(c) si $r_{xy} = 0$ il n'y a **pas de dépendance LINEAIRE** !

(d) si $(x_j, y_j) \mapsto (a + bx_j, c + dy_j)$, alors $r_{xy} \mapsto \mathrm{sign}(bd) r_{xy}$.

Limitations de la corrélation

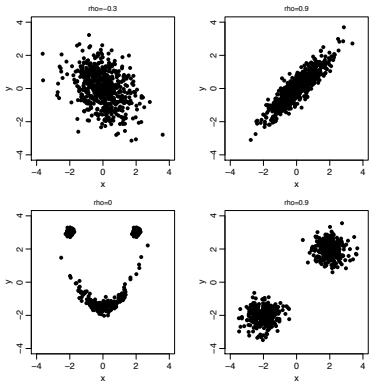
A noter :

— une corrélation entre deux variables n'implique pas une causalité entre elles ;

— r_{xy} mesure la dépendance linéaire (panneaux supérieurs) ;

— on peut avoir $r_{xy} \approx 0$, mais dépendance forte mais non-linéaire (en bas à gauche) ;

— une corrélation pourrait être forte mais **specieuse**, comme en bas à droite, ou deux sous-groupes, chacun sans corrélation, sont combinés.



Robustesse et résistance

— Toute analyse statistique se base sur les hypothèses du modèle statistique,
  — dont certaines sont **primaires**, tel que "les données sont une réalisation de variables indépendantes" ou "la variance est finie",
  — et certaines sont **secondaires**, tel que "la loi est gaussienne".
— L'invalidité des hypothèses secondaires ne devrait pas trop changer les conclusions, mais celle des hypothèses primaires peut tout changer.
— Une analyse/méthode **robuste** 'ne dépend pas trop' des hypothèses.
— Une statistique **résistante** 'ne change pas trop' quand on perturbe les données.

Définition 141 (Non-rigoreuse). *Le* **point de rupture (breakdown point)** *d'une statistique $s(y)$ est la plus petite fraction de contamination qui déstabilise complètement la statistique, dans le sens où $s(y) \to \pm\infty$.*

— Evidement un point de rupture est au plus 50%, car si $(50+\varepsilon)$% des données sont corrompues, il n'est plus clair quelle partie des données n'est pas contaminée.

Pause pensées. *Donner les points de rupture pour les caractéristiques numériques données dans la slide 164.*

Histogramme

Définition 142 (Histogram). *Soient $Y_1, \dots, Y_n \overset{\text{iid}}{\sim} f$ et supposons que les intervalles $\{\mathcal{I}_r\}_{r\in\mathcal{R}}$ partitionnent $\mathbb{R}$, i.e., $\bigcup_{r\in\mathcal{R}} \mathcal{I}_r = \mathbb{R}$, $\mathcal{I}_r \cap \mathcal{I}_s = \emptyset$ pour $r \neq s$; souvent $\mathcal{R} = \mathbb{Z}$ et $\mathcal{I}_r = [a + (r-1)h, a + rh)$ pour $a \in \mathbb{R}, h > 0$. Alors l'**histogramme** des Y_j est*

$$\widehat{f}_{\mathcal{I}}(y) = \sum_{r\in\mathcal{R}} \frac{I(y \in \mathcal{I}_r)}{|\mathcal{I}_r|} \frac{1}{n} \sum_{j=1}^{n} I(Y_j \in \mathcal{I}_r), \quad y \in \mathbb{R},$$

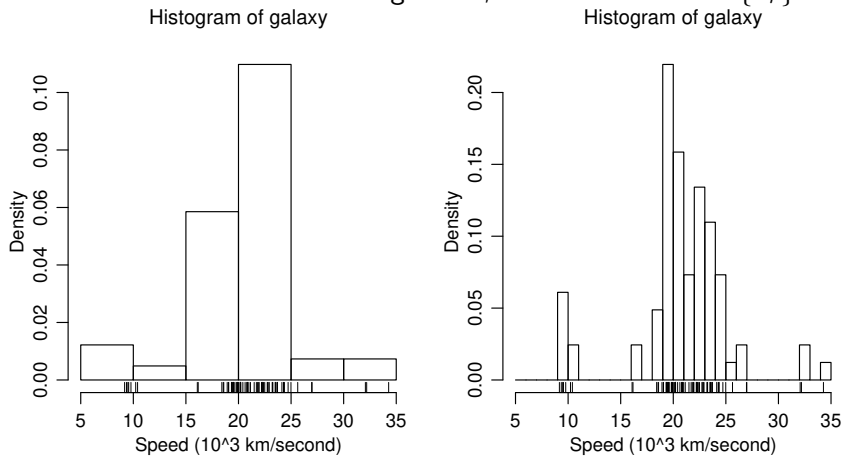
où $|\mathcal{I}_r|$ dénote la longueur de $\mathcal{I}_r$.

— Les vitesses (km/s) avec lesquelles 82 galaxies de la région Corona Borealis sont en train de diverger de notre galaxie. On suppose que l'erreur de mesure est inférieure à 50 km/s.
— Ces vitesses proviennent-elles d'une densité multimodale ?

9172	9350	9483	9558	9775	10227	10406	16084	16170	18419
18552	18600	18927	19052	19070	19330	19343	19349	19440	19473
19529	19541	19547	19663	19846	19856	19863	19914	19918	19973
19989	20166	20175	20179	20196	20215	20221	20415	20629	20795
20821	20846	20875	20986	21137	21492	21701	21814	21921	21960
22185	22209	22242	22249	22314	22374	22495	22746	22747	22888
22914	23206	23241	23263	23484	23538	23542	23666	23706	23711
24129	24285	24289	24366	24717	24990	25633	26960	26995	32065
32789	34279								

Histogramme, II

Histogrammes pour les données des vitesses des galaxies, avec deux choix de $\{\mathcal{I}_r\}$.



Boxplot, I

Garçons										
140	145	160	190	155	165	150	190	195	138	160
155	153	145	170	175	175	170	180	135	170	157
130	185	190	155	170	155	215	150	145	155	155
150	155	150	180	160	135	160	130	155	150	148
155	150	140	180	190	145	150	164	140	142	136
123	155									

Filles										
140	120	130	138	121	125	116	145	150	112	125
130	120	130	131	120	118	125	135	125	118	122
115	102	115	150	110	116	108	95	125	133	110
150	108									

— Ci-dessus le poids (en *pounds*) de 92 étudiants d'une école américaine.

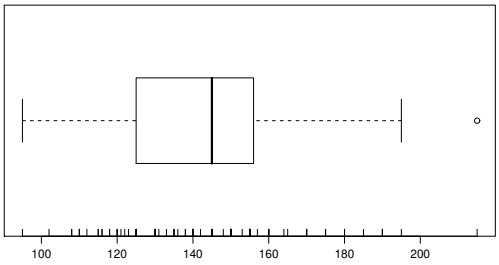
— Le **"five-number summary"** est la liste des cinq valeurs

$$y_{(1)}, \quad y_{(\lceil n/4 \rceil)}, \quad y_{(\lceil n/2 \rceil)}, \quad y_{(\lceil 3n/4 \rceil)}, \quad y_{(n)},$$

donne un résumé numérique simple et pratique des données.

— Cette liste est à la base de la **boîte à moustache (boxplot)**.

Boxplot, II



— Pour les poids, le "five-number summary" est $95, 125, 145, 156, 215$, et donc

$$\begin{aligned} \mathrm{IQR}(y) &= y_{(\lceil 3n/4 \rceil)} - y_{(\lceil n/4 \rceil)} = 156 - 125 = 31, \\ C &= 1.5 \times \mathrm{IQR}(y) = 1.5 \times 31 = 46.5, \\ y_{(\lceil n/4 \rceil)} - C &= 125 - 46.5 = 78.5, \\ y_{(\lceil 3n/4 \rceil)} + C &= 156 + 46.5 = 202.5. \end{aligned}$$

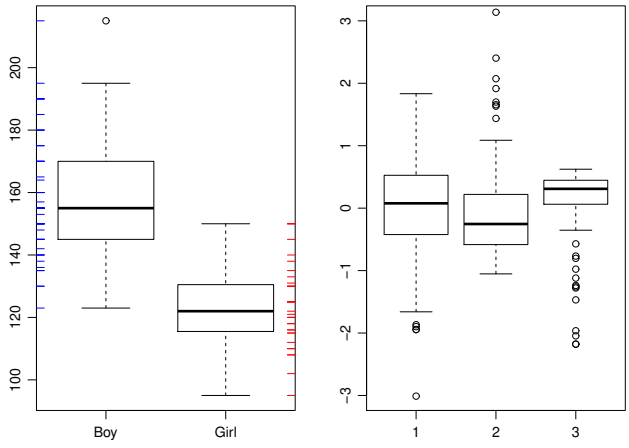
— Les limites de la moustache sont les y_i les plus extrêmes qui se trouvent à l'intérieur de l'intervalle dont les limites sont $y_{(\lceil n/4 \rceil)} - C$ et $y_{(\lceil 3n/4 \rceil)} + C$.

— Les y_i à l'extérieur de la moustache sont montrés individuellement.

Boxplot, III

— Le boxplot est utile pour la comparaison de groupes d'observations.

— Boxplots du poids des étudiants selon le sexe, et de trois groupes d'observations simulées :



Q-Q plot, I

— Lorsqu'on souhaite utiliser une loi F comme modèle pour un jeu de données $y_1, \ldots, y_n$, une comparaison graphique directe utile est le **quantile-quantile plot (Q-Q plot)** : on représente les statistiques d'ordre

$$y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)},$$

en fonction des quantiles théoriques de F, c'est-à-dire

$$F^{-1}\{1/(n+1)\}, F^{-1}\{2/(n+1)\}, \ldots, F^{-1}\{n/(n+1)\}.$$

— Si les points du Q-Q plot sont proches d'une droite, cela signifie que les observations pourront être modélisées par F.

— Les valeurs aberrantes apparaissent comme des points isolés.

— Souvent F a des paramètres inconnus, et alors on replace F par une loi standardisée.

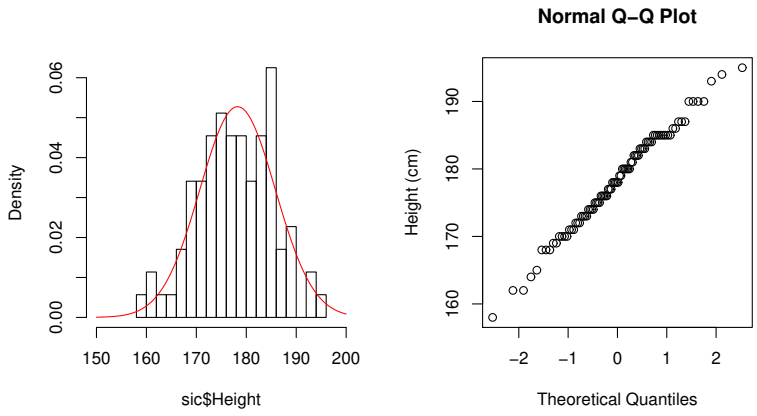
— Pour les comparaisons avec la loi normale, on utilise

$$\Phi^{-1}\{1/(n+1)\}, \Phi^{-1}\{2/(n+1)\}, \ldots, \Phi^{-1}\{n/(n+1)\},$$

et la pente et l'ordonnée à l'origine pour $x = 0$ donnent des estimations de σ et μ.

Q-Q plot, II

Histogramme et Q-Q plot normal des tailles de 88 étudiants :



5.3 Estimation Ponctuelle

Motivation

— Souvent on veut estimer un paramètre d'un modèle statistique à partir de $y_1, \ldots, y_n$.

— Rappel : un paramètre est une fonction d'une loi de probabilité.

Définition 143. *Soient Y une variable aléatoire (généralement un vecteur) issue d'une loi de probabilité F et $\theta = \theta(F)$ un paramètre (peut-être aussi un vecteur). Alors un* **estimateur (estimator)** *$T = t(Y)$ est une fonction de Y que l'on utilise pour estimer θ. On appelle* **estimation (estimate)** *une valeur spécifique $t = t(y)$ de $T = t(Y)$.*

Souvent on écrit $\widehat{\theta}$, $\tilde{\theta}, \ldots$ pour des estimateurs/estimations de θ.

Exemple 144. *Soient $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \exp(\lambda)$, alors donner des estimateurs pour λ, $\mu = 1/\lambda$, $\theta = \exp(-\lambda)$.*

— Ces estimateurs sont **ponctuels** : une seule valeur, sans indication de son incertitude.

— Evidemment on a plusieurs estimateurs possibles, quelles propriétés veut-on d'un estimateur, et comment choisir entre eux ?

— D'abord décrivons deux méthodes générales d'estimation.

Note to Example 144

We know that $\mathrm{E}(Y) = 1/\lambda$, so to estimate λ we might take

$$\tilde{\lambda}_1 = n^{-1} \sum_{j=1}^{n} 1/Y_j, \quad \tilde{\lambda}_2 = 1/\overline{Y},$$

to estimate $\mu = 1/\lambda = \mathrm{E}(Y)$ it seems natural to take $\tilde{\mu} = \overline{Y}$, and to estimate $\theta = \mathrm{P}(Y > 1)$ we might take

$$\tilde{\theta}_1 = n^{-1} \sum_{j=1}^{n} I(Y_j > 1), \quad \tilde{\theta}_2 = \exp(-\tilde{\lambda}_2).$$

Méthode des moments

Définition 145. *Soient $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} F$, où F est déterminé par $\theta_{p\times 1}$, et les moments de Y_j sont $\mathrm{E}(Y_1^r) = \mu_r(\theta)$, pour $r = 1, \ldots, p$. Alors la* **méthode des moments** *trouve les estimateurs $\tilde{\theta}$ comme solution aux equations*

$$n^{-1} \sum_{j=1}^{n} Y_j^r = \mu_r(\theta), \quad r = 1, \ldots, p.$$

— On a besoin d'autant de moments finis que de paramètres inconnus.

— Ce sont des estimateurs simples et généralement consistants mais dont les variances peuvent être grandes.

Exemple 146. *Soient $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \exp(\lambda)$, estimer λ par la méthode des moments.*

Exemple 147. *Soient $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, estimer μ et σ^2 par la méthode des moments.*

Exemple 146

We compute $\mathrm{E}(Y) = 1/\lambda$, and solve the equation $\overline{Y} = 1/\lambda$ to get $\tilde{\lambda} = 1/\overline{Y}$.

Exemple 147

The population equations corresponding to the first two moments are

$$\mathrm{E}(Y) = \mu, \quad \mathrm{E}(Y^2) = \mathrm{var}(Y) + \mathrm{E}(Y)^2 = \sigma^2 + \mu^2,$$

and the corresponding sample versions are

$$\overline{Y} = \tilde{\mu}, \quad n^{-1}\sum_{j=1}^{n} Y_j^2 = \tilde{\sigma}^2 + \tilde{\mu}^2.$$

Solving these gives

$$\tilde{\mu} = \overline{Y}, \quad \tilde{\sigma}^2 = n^{-1}\left(\sum_{j=1}^{n} Y_j^2 - n\,\overline{Y}^2\right) = n^{-1}\sum_{j=1}^{n}(Y_i - \overline{Y})^2.$$

Méthode du maximum de vraisemblance

Celle-ci est une méthode d'estimation plus générale, qui n'a pas besoin des moments.

Définition 148. *Soient $y \equiv y_1, \ldots, y_n$ une réalisation d'un échantillon aléatoire $Y \equiv Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} f(y;\theta)$, ou $\theta \in \Theta$, alors la* **vraisemblance (likelihood)** *et la* **log-vraisemblance (log-likelihood)** *pour θ sont*

$$L(\theta) = f(y;\theta) = \prod_{j=1}^{n} f(y_j;\theta), \quad \ell(\theta) = \log L(\theta), \theta \in \Theta.$$

Définition 149. *L'***estimateur du maximum de vraisemblance (maximum likelihood estimator, MLE)** *$\widehat{\theta}$ satisfait $L(\widehat{\theta}) \geq L(\theta), \quad \theta \in \Theta$.*

Définition 150. *L'***information observée (observed information)** *$\jmath(\theta)$ et l'***information espérée, information de Fisher (expected information, Fisher information)** *$\imath(\theta)$ sont*

$$\jmath(\theta) = -\frac{\mathrm{d}^2\ell(\theta)}{\mathrm{d}\theta^2}, \quad \imath(\theta) = \mathrm{E}\{\jmath(\theta)\} = \mathrm{E}\left\{-\frac{\mathrm{d}^2\ell(\theta)}{\mathrm{d}\theta^2}\right\}.$$

Elles mesurent la courbure de $-\ell(\theta)$: plus $\jmath(\theta)$ et $\imath(\theta)$ sont grandes, plus $\ell(\theta)$ est concentrée.

Calcul de $\widehat{\theta}$

Exemple 151. *Soient* $y_1, \ldots, y_n \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$, *calculer* $\ell(\theta)$, $\widehat{\theta}$, $\text{var}(\widehat{\theta})$, $\jmath(\theta)$, *et* $\imath(\theta)$.

— On facilite les calculs en maximisant $\ell(\theta)$ plutôt que $L(\theta)$.

— Le démarche pratique est souvent :

  — calculer $\ell(\theta) = \log L(\theta)$;

  — (dans la plupart des applications, numériquement) calculer $\widehat{\theta}$ tel que $\text{d}\ell(\theta)/\text{d}\theta = 0$;

  — verifier qu'il s'agit bien d'un maximum ;

  — (dans la plupart des applications, numériquement) calculer $\jmath(\widehat{\theta})$.

Note to Example 151

— The likelihood is

$$L(\theta) = f(y;\theta) = \prod_{j=1}^{n} f(y_j;\theta) = \prod_{j=1}^{n} \theta^{y_j}(1-\theta)^{1-y_j} = \theta^s(1-\theta)^{n-s}, \quad 0 \le \theta \le 1,$$

where $s = \sum y_j$ and we have used the fact that the observations are independent. Therefore

$$\ell(\theta) = s\log\theta + (n-s)\log(1-\theta), \quad 0 \le \theta \le 1.$$

— Differentiation yields

$$\frac{d\ell(\theta)}{d\theta} = \frac{s}{\theta} - \frac{n-s}{1-\theta}, \quad \frac{d^2\ell(\theta)}{d\theta^2} = -\frac{s}{\theta^2} - \frac{n-s}{(1-\theta)^2}.$$

Setting $d\ell(\theta)/d\theta = 0$ gives just one solution, $\widehat{\theta} = s/n = \overline{y}$, and since the second derivative is always negative, this is clearly the maximum.

— Clearly

$$\jmath(\theta) = -\frac{d^2\ell(\theta)}{d\theta^2} = \frac{s}{\theta^2} + \frac{n-s}{(1-\theta)^2}.$$

Now treating $\widehat{\theta}$ as a random variable, $\widehat{\theta} = S/n$, where $S \sim B(n,\theta)$, we see that since $\text{E}(S) = n\theta$ and $\text{var}(S) = n\theta(1-\theta)$, we have after a little algebra that

$$\text{var}(\widehat{\theta}) = \frac{\theta(1-\theta)}{n}, \quad \imath(\theta) = \text{E}\{\jmath(\theta)\} = \frac{n}{\theta(1-\theta)}, \quad 0 < \theta < 1.$$

Note that $\text{var}(\widehat{\theta}) = 1/\imath(\theta)$.

Qu'est-ce que c'est un bon estimateur ?

— Dorénavant on supposera que $Y \equiv Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} F$ et que l'on veuille estimer $\theta = \theta(F)$ avec un estimateur $\tilde{\theta} \equiv \tilde{\theta}(Y_1, \ldots, Y_n)$.

— On a deux types de critère pour rechercher les bons estimateurs :

  — **asymptotique** — comment se comporte $\tilde{\theta}$ quand $n \to \infty$?

  — **échantillon fini (finite-sample)** — comment comparer $\tilde{\theta}_1$ et $\tilde{\theta}_2$ pour n fixé ?

— La consistance est un critère asymptotique clé : $\tilde{\theta}$ se rapproche-t-il de θ quand $n \to \infty$?

Définition 152. *Un estimateur $\tilde{\theta}$ de θ est appellé* **consistant** *si $\tilde{\theta} \xrightarrow{P} \theta$ quand $n \to \infty$.*

Exemple 153. *Discuter la consistance des estimateurs de l'exemple 144.*

— Consistance d'un estimateur n'est pas assez, car

$$\tilde{\theta} \xrightarrow{P} \theta \quad \Rightarrow \quad \tilde{\theta} + 10^6/\sqrt{\log\log n} \xrightarrow{P} \theta, \quad n \to \infty,$$

mais ce dernier, bien que consistant, est inutile. Consistance est une propriété nécessaire, mais pas suffisante, d'un bon estimateur.

— Evidemment on aimerait que la distance $\tilde{\theta} - \theta$ soit petite.

Note to Example 153

Now $Y_j \overset{\text{iid}}{\sim} (1/\lambda, 1/\lambda^2)$, so $\overline{Y} \xrightarrow{P} 1/\lambda > 0$ by the weak law of large numbers. The functions $h(x) = 1/x$ and $h(x) = \exp(-x)$ are continuous for all $x > 0$, so Theorem 116, second line, implies that

$$\tilde{\lambda}_2 = 1/\overline{Y} \xrightarrow{P} 1/(1/\lambda) = \lambda, \quad \tilde{\theta}_2 = \exp(-\tilde{\lambda}_2) \xrightarrow{P} \exp(-\lambda) = \tilde{\theta}, \quad n \to \infty.$$

The weak law also gives that

$$\tilde{\theta}_1 = n^{-1} \sum_{j=1}^{n} I(Y_j > 1) \xrightarrow{P} \mathrm{E}\{I(Y_j > 1)\} = \mathrm{P}(Y_j > 1) = \exp(-\lambda) = \theta, \quad n \to \infty,$$

so we have two estimators of θ.
On the other hand we have $\mathrm{E}(\tilde{\lambda}_1) = \infty$, so it seems unlikely that this estimator could be consistent.

Erreur quadratique moyenne

Définition 154. *Le* **biais (bias)** *de* $\tilde{\theta}$ *est* $b(\tilde{\theta};\theta) = \mathrm{E}(\tilde{\theta}) - \theta$.

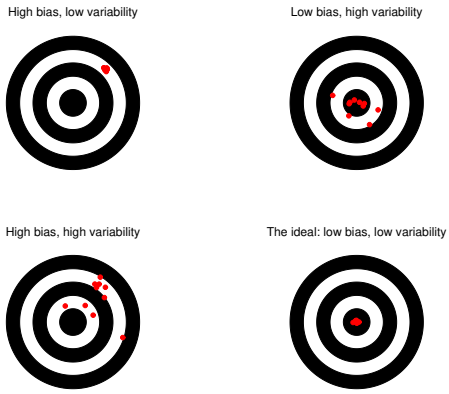
Définition 155. *L'*__erreur quadratique moyenne (mean squared error, MSE)__ *de* $\tilde{\theta}$ *est*

$$\mathrm{MSE}(\tilde{\theta};\theta) = \mathrm{E}\left\{(\tilde{\theta} - \theta)^2\right\}.$$

Lemme 156. *On peut écrire* $\mathrm{MSE}(\tilde{\theta};\theta) = b(\tilde{\theta};\theta)^2 + \mathrm{var}(\tilde{\theta})$.

— Le biais est une propriété de l'estimateur $\tilde{\theta}$, et cette propriété pourrait varier en fonction de θ.
— Interprétation du biais :
  — si pour tout θ, $b(\tilde{\theta};\theta) < 0$, alors en moyenne $\tilde{\theta}$ sous-estime θ ;
  — si pour tout θ, $b(\tilde{\theta};\theta) > 0$, alors en moyenne $\tilde{\theta}$ sur-estime θ ;
  — si pour tout θ, $b(\tilde{\theta};\theta) = 0$, alors $\widehat{\theta}$ est __non biaisé (unbiased)__..
— Un indicateur de la qualité de $\tilde{\theta}$ est l'absence d'un écart systématique entre $\widehat{\theta}$ et θ : $b(\tilde{\theta};\theta) \approx 0$.
— Un indicateur encore plus important est $\mathrm{MSE}(\tilde{\theta};\theta)$, qui mesure aussi la variabilité de $\tilde{\theta}$.

Biais et variance



— θ = "bulle centrale", supposé être la vraie valeur
— $\widehat{\theta}$ = fléchette rouge tirée sur la bulle centrale, valeur estimée à l'aide des données

Exemple 157. *Soient* $Y_1, \ldots, Y_n \overset{\mathrm{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. *Trouver le biais et la variance de* $\tilde{\mu} = \overline{Y}$ *et le biais de* $\tilde{\sigma}^2 = n^{-1}\sum_j (Y_j - \overline{Y})^2$.

Note to Example 157

— We've already seen that

$$\mathrm{E}(\overline{Y}) = \mu, \quad \mathrm{var}(\overline{Y}) = \sigma^2/n,$$

so the bias of $\overline{Y}$ as an estimator of μ is $\mathrm{E}(\overline{Y}) - \mu = 0$.

— To find the expectation of $n\tilde{\sigma}^2 = n^{-1}\sum_j (Y_j - \overline{Y})^2$, note that

$$\begin{aligned}
\sum_{j=1}^n (Y_j - \overline{Y})^2 &= \sum_{j=1}^n \left\{Y_j - \mu - (\overline{Y} - \mu)\right\}^2 \\
&= \sum_{j=1}^n (Y_j - \mu)^2 - 2\sum_{j=1}^n (Y_j - \mu)(\overline{Y} - \mu) + \sum_{j=1}^n (\overline{Y} - \mu)^2 \\
&= \sum_{j=1}^n (Y_j - \mu)^2 - 2n(\overline{Y} - \mu)^2 + n(\overline{Y} - \mu)^2 \\
&= \sum_{j=1}^n (Y_j - \mu)^2 - n(\overline{Y} - \mu)^2.
\end{aligned}$$

As

$$\begin{aligned}
\mathrm{E}\left(n\tilde{\sigma}^2\right) &= n\mathrm{E}\left\{(Y_j - \mu)^2\right\} - n\mathrm{E}\left\{(\overline{Y} - \mu)^2\right\} \\
&= n\sigma^2 - n\sigma^2/n \\
&= (n-1)\sigma^2,
\end{aligned}$$

we see that $\tilde{\sigma}^2$ has expected value $(n-1)\sigma^2/n$. Therefore the bias of $\tilde{\sigma}^2$ is $(n-1)\sigma^2/n - \sigma^2 = -\sigma^2/n$: the estimator is biased downwards.

Efficacité

Définition 158. *Soient $\tilde{\theta}_1$ et $\tilde{\theta}_2$ deux estimateurs sans biais du même paramètre θ. Alors*

$$\begin{aligned}
\mathrm{MSE}(\tilde{\theta}_1;\theta) &= \mathrm{var}(\tilde{\theta}_1) + b(\tilde{\theta}_1;\theta)^2 = \mathrm{var}(\tilde{\theta}_1) \\
\mathrm{MSE}(\tilde{\theta}_2;\theta) &= \mathrm{var}(\tilde{\theta}_2) + b(\tilde{\theta}_2;\theta)^2 = \mathrm{var}(\tilde{\theta}_2),
\end{aligned}$$

et on dit que $\tilde{\theta}_1$ est **plus efficace (more efficient)** *que $\tilde{\theta}_2$ si*

$$\mathrm{var}(\tilde{\theta}_1) \leq \mathrm{var}(\tilde{\theta}_2), \quad \theta \in \Theta.$$

On préfère alors $\tilde{\theta}_1$.

Exemple 159. *Soient $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, avec n grand. Trouver les propriétés de la médiane M et de la moyenne $\overline{Y}$. Lequel est préférable ? Et si des valeurs abérrantes peuvent apparaître ?*

Note to Example 159

— We've already seen that

$$\mathrm{E}(\overline{Y}) = \mu, \quad \mathrm{var}(\overline{Y}) = \sigma^2/n,$$

so the bias of $\overline{Y}$ as an estimator of μ is $\mathrm{E}(\overline{Y}) - \mu = 0$.

— Results from Example 133 give that for large n,

$$\mathrm{E}(M) \doteq \mu, \quad \mathrm{var}(M) = \frac{\pi\sigma^2}{2n},$$

so both estimators are (approximately) unbiased (in fact exactly unbiased), but

$$\frac{\mathrm{var}(M)}{\mathrm{var}(\overline{Y})} = \frac{\pi}{2} > 1,$$

so M is less efficient than $\overline{Y}$: its variance is larger. However if there are outliers, we have seen that the median M is little changed, whereas the average $\overline{Y}$ can be badly affected. Our choice between these estimators will depend on how much we fear that our data will be contaminated by bad values.

Commentaires

— La méthode de maximum de vraisemblance est très générale, pouvant être étendue à énormément de situations avec des données et/ou des modèles complexes.

— Sous des conditions mathématiques, notamment que les données $y_1, \ldots, y_n$ sont issues du modèle $f(y;\theta)$, on peut démontrer que l'estimateur de maximum de vraisemblance $\widehat{\theta}$ a de bonnes propriétés : pour n grand, $\mathrm{E}(\widehat{\theta}) \doteq \theta$, et que $\mathrm{var}(\widehat{\theta})$ est minimale — on ne peut pas mieux estimer θ.

— En réalité on n'est jamais certain du modèle, et on doit souvent sacrificier de l'efficacité (petite variance sous un modèle idéal) pour la robustesse (bonne estimation même s'il y a des mauvaises observations, ou si le modèle assumé n'est pas juste).

— En générale on obtient le MLE par un algorithme iterative, et on peut souvent utiliser la méthode des moments pour trouver une valeur initiale de θ.

Efficiency and the Cramèr–Rao lower bound

Définition 160. *If $\tilde{\theta}_1$ and $\tilde{\theta}_2$ are estimators of scalar θ, then the* **relative efficiency** *of $\tilde{\theta}_1$ compared to $\tilde{\theta}_2$ can be defined as*

$$\frac{\mathrm{MSE}(\tilde{\theta}_2;\theta)}{\mathrm{MSE}(\tilde{\theta}_1;\theta)}.$$

In large samples the squared bias is often negligible compared to the variance, and we define the **asymptotic relative efficiency** *as $\mathrm{var}(\tilde{\theta}_2)/\mathrm{var}(\tilde{\theta}_1)$. Similar expressions apply if the parameter has dimension d.*

— Under mild conditions a scalar estimator $\tilde{\theta}$ based on $Y \sim f(y;\theta)$ satisfies the **Cramèr–Rao lower bound**,

$$\mathrm{var}(\tilde{\theta}) \geq \frac{\{1+\nabla b(\tilde{\theta};\theta)\}^2}{\imath(\theta)},$$

where $\imath(\theta)$ is the **Fisher (or expected) information**. This applies for any sample size n, but

— as $n \to \infty$ the lower bound $\to 1/\imath(\theta)$, the asymptotic variance of the maximum likelihood estimator, which hence is most efficient in large samples; and

— a similar result applies for vector θ.

Bartlett identities

— For data $Y \sim f(y;\theta)$ we define the **log likelihood function** $\ell(\theta) = \log f(Y;\theta)$ and $d \times 1$ **score vector** $U(\theta) = \nabla\ell(\theta)$.

— If we can differentiate with respect to θ under the integral sign, we get the **Bartlett identities** :

$$\begin{aligned}
1 &= \int f(y;\theta)\,\mathrm{d}y,\\
0 &= \int \nabla \log f(y;\theta) \times f(y;\theta)\,\mathrm{d}y,\\
0 &= \int \nabla^2 \log f(y;\theta) \times f(y;\theta)\,\mathrm{d}y + \int \nabla \log f(y;\theta)\,\nabla^{\mathrm{T}} \log f(y;\theta) \times f(y;\theta)\,\mathrm{d}y,\\
0 &= \cdots
\end{aligned}$$

giving the moments of $U(\theta)$, i.e.,

$$\mathrm{E}\{U(\theta)\} = 0, \quad \mathrm{var}\{U(\theta)\} = \mathrm{E}\left\{\nabla\ell(\theta)\nabla^{\mathrm{T}}\ell(\theta)\right\} = \mathrm{E}\left\{-\nabla^2\ell(\theta)\right\} = \imath(\theta), \quad \ldots$$

— If $Y \equiv Y_1, \ldots, Y_n \overset{\mathrm{iid}}{\sim} f(y;\theta)$, then

$$U(\theta) = \sum_{j=1}^{n} \nabla \log f(Y_j;\theta)$$

is a sum of IID components to which a central limit theorem applies.

Note : Bartlett identities

— The first is true for any θ, and provided we can exchange the order of integration and differentiation we have

$$0 = \nabla \int f(y;\theta)\,\mathrm{d}y = \int \nabla f(y;\theta)\,\mathrm{d}y = \int \nabla f(y;\theta)\frac{f(y;\theta)}{f(y;\theta)}\,\mathrm{d}y = \int \nabla \log f(y;\theta)\, f(y;\theta)\,\mathrm{d}y.$$

— The second stems from a second differentiation and applying the chain rule to the terms in the final integral here; likewise for the third and higher-order ones, which give higher-order moments of $U(\theta)$.

— For independent data $Y_1, \ldots, Y_n$ we have $U(\theta) = \sum_{j=1}^n U_j(\theta)$, where the $U_j = \nabla \log f(Y_j;\theta)$ are independent, so using the Bartlett identities for the individual densities $f_j(y_j;\theta)$ we have

$$\mathrm{var}\{U(\theta)\} = \sum_{j=1}^n \mathrm{var}\{U_j(\theta)\} = \sum_{j=1}^n \mathrm{E}\{U_j(\theta)U_j^{\mathrm{T}}(\theta)\} = \sum_{j=1}^n -\mathrm{E}\{\nabla^{\mathrm{T}} U_j(\theta)\} = -\mathrm{E}\left\{\nabla^{\mathrm{T}} U(\theta)\right\}$$

and this equals $\mathrm{E}\left\{-\nabla^2 \ell(\theta)\right\} = \imath(\theta)$, and this in turn equals $n\imath_1(\theta)$.

http://stat.epfl.ch note 1 of slide 186

Note : CRLB

— We have

$$\mathrm{E}(\tilde{\theta}) = \int \tilde{\theta}(y) f(y;\theta)\,\mathrm{d}y = \theta + b(\tilde{\theta};\theta),$$

and differentiation with respect to θ gives (setting $b'(\theta) = \mathrm{d}b(\tilde{\theta};\theta)/\mathrm{d}\theta$)

$$1 + b'(\theta) = \int \tilde{\theta}(y)\nabla f(y;\theta)\,\mathrm{d}y = \int \tilde{\theta}(y)\nabla \ell(\theta) f(y;\theta)\,\mathrm{d}y = \mathrm{E}\{\tilde{\theta}U(\theta)\} = \mathrm{cov}\{\tilde{\theta}, U(\theta)\},$$

because $U(\theta)$ has mean zero. Hence the definition of correlation gives

$$\mathrm{cov}\{\tilde{\theta}, U(\theta)\}^2 = \{1 + b'(\theta)\}^2 \le \mathrm{var}(\tilde{\theta})\mathrm{var}\{U(\theta)\} = \mathrm{var}(\tilde{\theta})\imath(\theta),$$

which gives the result.

— If the bias is of order n^{-1}, so too is its derivative, so in large samples we obtain

$$\mathrm{var}(\tilde{\theta}) \ge \imath(\theta)^{-1},$$

and of course this also holds for any n if $\tilde{\theta}$ is unbiased, as in that case $b(\tilde{\theta};\theta) \equiv 0$. Thus if the Bartlett identities hold, the variance of any unbiased estimator is bounded below by $\imath(\theta)^{-1}$, for any n. It turns out that this is the asymptotic variance of the MLE $\widehat{\theta}$, so the latter is (asymptotically) minimum variance and unbiased.

http://stat.epfl.ch note 2 of slide 186

Loi limite du MLE

Théorème 161. *Soient $Y_1, \ldots, Y_n$ un échantillon aléatoire issu d'une densité paramétrique $f(y;\theta)$, et soit $\widehat{\theta}$ le MLE de θ. Si f satisfait des* **conditions de régularité** *(voir ci-après), alors*

$$\jmath(\widehat{\theta})^{1/2}(\widehat{\theta} - \theta) \xrightarrow{D} \mathcal{N}(0,1) \quad n \to \infty,$$

où $\jmath(\theta) = -\nabla^2 \ell(\theta)$. Donc pour n grand,

$$\widehat{\theta} \quad \dot{\sim} \quad \mathcal{N}\left\{\theta, \jmath(\widehat{\theta})^{-1}\right\}.$$

Note to Theorem 161

— We write $U(\theta) = \nabla \ell(\theta) = \partial \ell(\theta)/\partial \theta$ and suppose that the MLE $\widehat{\theta}$ satisfies $U(\widehat{\theta}) = 0$.

— Then

$$\mathrm{E}\{U(\theta)\} = 0, \quad \mathrm{var}\{U(\theta)\} = \imath(\theta) = \sum_{j=1}^{n} \mathrm{E}\left\{-\nabla^2 \ell(\theta)\right\} = n\imath_1(\theta),$$

where $\imath_1(\theta)$ is the Fisher information for a single observation, so the central limit theorem gives

$$Z_n = n^{-1/2} U(\theta) \xrightarrow{D} \mathcal{N}\{0, \imath_1(\theta)\}$$

if the Bartlett identities are satisfied.

— Now

$$0 = U(\widehat{\theta}) = U(\theta) + \int_0^1 \nabla^2 \ell\{\theta + t(\widehat{\theta} - \theta)\} (\widehat{\theta} - \theta) \, \mathrm{d}t,$$

and some reorganisation of this gives

$$n^{-1/2} \imath_1(\theta)^{-1/2} U(\theta) = n^{1/2} \imath_1(\theta)^{-1/2} (\widehat{\theta} - \theta) \int_0^1 n^{-1} \nabla^2 \ell\{\theta + n^{-1/2} t n^{1/2} (\widehat{\theta} - \theta)\} \, \mathrm{d}t.$$

The left-hand side of this expression is $Z_n \xrightarrow{D} \mathcal{N}(0,1)$ and the expression is exact, so the same must be true of the right-hand side. Under mild conditions on the second log-likelihood derivative in a neighbourhood of θ the integral will converge in probability to $-\imath_1(\theta)$, and therefore

$$n^{1/2} \imath_1(\theta)^{1/2} (\widehat{\theta} - \theta) \xrightarrow{D} \mathcal{N}(0,1), \quad n \to \infty,$$

leading to the finite-sample approximation

$$\widehat{\theta} \stackrel{\cdot}{\sim} \mathcal{N}\left\{\theta, \imath(\theta)^{-1}\right\}.$$

Since $n^{-1} \jmath(\theta) \xrightarrow{P} \imath_1(\theta)$ and $\widehat{\theta} \xrightarrow{P} \theta$, we can replace $\imath(\theta)$ by $\jmath(\widehat{\theta})$, giving the practically useful approximation

$$\widehat{\theta} \stackrel{\cdot}{\sim} \mathcal{N}\left\{\theta, \jmath(\widehat{\theta})^{-1}\right\}.$$

— Above we have not stated formal regularity conditions, but inspection of the proof suggests that for the argument to work

  — the parameter value must be interior to a continuous parameter space, so a limiting normal distribution is possible;

  — the MLE $\widehat{\theta}$ must be a consistent estimator (so in particular no two values of θ can give the same density for Y);

  — the Bartlett identities must hold;

  — the log likelihood must 'behave well' close to the true parameter θ, so that the integral converges to $-\imath(\theta)$.

Régularité

— Les conditions de régularité sont satisfaites dans la grande majorité des cas rencontrés en pratique.

— Les cas où elles sont fausses sont souvent quand
  — un des paramètres est discrêt ;
  — le support de $f(y;\theta)$ dépend de θ ;
  — le vrai θ se trouve sur une borne des valeurs possibles.

— Voici un exemple où elles ne sont pas vérifiées.

Exemple 162. *Soient $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} U(0,\theta)$, trouver la vraisemblance $L(\theta)$ et le MLE $\widehat{\theta}$. Montrer que la loi limite de $n(\theta - \widehat{\theta})/\theta$ quand $n \to \infty$ est $\exp(1)$. Discuter.*

Note to Example 162

— Recall that the density and distribution function of Y_j are

$$f(x;\theta) = \theta^{-1} I(0 < x < \theta), \quad F(x) = x/\theta, \quad 0 < x < \theta.$$

— wing to the independence, we have

$$L(\theta) = \prod \theta^{-1} I(y_j < \theta) = \theta^{-n} I(\max y_j < \theta), \quad \theta > 0,$$

and therefore $\widehat{\theta} = M = \max Y_j$, whose distribution is

$$\mathrm{P}(M \le x) = \prod_{j=1}^{n} \mathrm{P}(Y_j \le x) = (x/\theta)^n, \quad 0 < x < \theta.$$

— Now

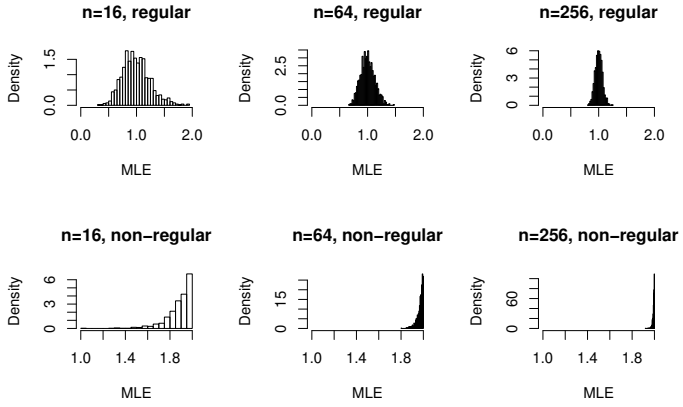
$$\mathrm{P}\left\{ n(\theta - \widehat{\theta})/\theta \le x \right\} = \mathrm{P}(\widehat{\theta} \ge \theta - x\theta/n) = 1 - \{(\theta - x\theta/n)/\theta\}^n \to 1 - \exp(-x),$$

as required. Note that :
  — the scaling needed to get a limiting distribution is much faster here than in the usual case ;
  — the limit is not normal.

— Here the Bartlett identities do not hold because the support of the density, $(0,\theta)$, depends on θ.

Exemple

Comparaison des lois de $\widehat{\theta}$ dans un cas regulier (panneaux dessus, avec déviation standard $\propto n^{-1/2}$) et dans un cas non-regulier (Exemple 162, panneaux dessous, avec déviation standard $\propto n^{-1}$). Dans d'autres cas non-reguliers il pourrait arriver que la loi n'est pas sympa (comme ici) et/ou que la convergence soit plus lente que dans des cas reguliers.



5.4 Estimation par Intervalle

Les intervalles de confiance

Un élément clé de la statistique est de donner une idée de l'incertitude d'un constat.
Soit θ un paramètre inconnu, et soit $\tilde{\theta} = 1$ une estimation de θ basée sur $y_1, \ldots, y_n$:

— alors si $n = 10^5$ on est beaucoup plus sûr que $\theta \approx \tilde{\theta}$ que si $n = 10$;

— pour exprimer ceci on aimerait donner un intervalle qui serait plus large quand $n = 10$ que quand $n = 10^5$, pour expliciter l'incertitude liée à $\tilde{\theta}$;

Définition 163. *Soient $Y \equiv Y_1, \ldots, Y_n$ des données issues d'une loi paramétrique F de paramètre θ scalaire. Un* **intervalle de confiance (confidence interval)** (L, U) **pour** θ *est une statistique sous forme d'intervalle qui contient θ avec un probabilité spécifiée. Cette probabilité s'appelle le* **niveau** *de l'intervalle.*

Noter que

— les limites L, U sont des fonctions des données $Y_1, \ldots, Y_n$ et non pas des inconnus ;

— un intervalle de confiance **bilatéral**, de la forme (L, U) est le plus souvent utilisé, mais

— un intervalle de confiance **unilatéral**, de la forme $(-\infty, U)$ ou (L, ∞) est parfois utile.

Interprétation d'un IC

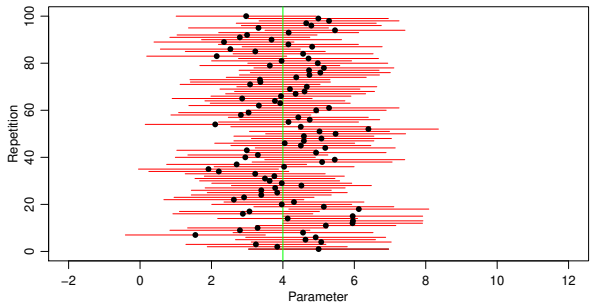
— $(L.U)$ est un intervalle aléatoire qui contient θ avec probabilité specifiée, $1-\alpha$.

— On imagine une suite infinie de répétitions de l'expérience qui a donné (L, U)

— L'IC que l'on a calculé est un des ICs possibles, et on peut considérer qu'il a été choisi au hasard parmi eux.

— Nous ne savons pas si notre IC contient θ, mais cet événement a une probabilité $1-\alpha$.

— Pour illustrer ce raisonnement, ici le paramètre θ (vert) est contenu (ou pas) dans des réalisations de l'IC (rouge), les $\tilde{\theta}$ sont les points noirs :



Des IC approximatifs

Dans le plupart des cas, on construit des ICs approximatifs, basés sur des estimateurs dont on a besoin d'estimer les variances.

Définition 164. *Soient $\tilde{\theta} = \tilde{\theta}(Y_1, \ldots, Y_n)$ un estimateur de θ, $\tau_n^2 = \mathrm{var}(\tilde{\theta})$ sa variance, et $V = v(Y_1, \ldots, Y_n)$ un estimateur de τ_n^2. Alors on appelle $V^{1/2}$ (également sa réalisation $v^{1/2}$) un* **écart-type** *de $\tilde{\theta}$.*

Théorème 165. *Soient $\tilde{\theta}$ un estimateur et $V^{1/2}$ son écart-type se basant sur un échantillon aléatoire de taille n, avec*

$$\frac{\tilde{\theta}-\theta}{\tau_n} \xrightarrow{D} Z, \quad \frac{V}{\tau_n^2} \xrightarrow{P} 1, \quad n \to \infty,$$

où $Z \sim \mathcal{N}(0,1)$. Alors par le théorème 122 on a

$$\frac{\tilde{\theta}-\theta}{V^{1/2}} = \frac{\tilde{\theta}-\theta}{\tau_n} \times \frac{\tau_n}{V^{1/2}} \xrightarrow{D} Z, \quad n \to \infty.$$

Implication : En construisant un IC par le TCL, on peut remplacer τ_n par $V^{1/2}$.

Exemple 166. *Soient $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathrm{Bernouilli}(\theta)$, trouver un écart-type pour le MLE de θ.*

Note to Example 166

— We saw in Example 151 that the MLE is $\widehat{\theta} = n^{-1}\sum_{j=1}^{n} Y_j = n^{-1}S$, where $S \in \mathcal{B}(n,\theta)$ has a binomial distribution. Therefore

$$\mathrm{var}(\widehat{\theta}) = \mathrm{var}(S/n) = n^{-2}\mathrm{var}(S) = n^{-2}n\theta(1-\theta) = \theta(1-\theta)/n = \tau_n^2.$$

This is estimated by $V = \widehat{\theta}(1-\widehat{\theta})/n$, and, since $\widehat{\theta} \xrightarrow{P} \theta$, by the central limit theorem, and the function $h(u) = u(1-u)$ is continuous for $u \in (0,1)$,

$$\frac{V}{\tau_n^2} = \frac{\widehat{\theta}(1-\widehat{\theta})}{\theta(1-\theta)} \xrightarrow{P} 1, \quad n \to \infty, \quad 0 < \theta < 1.$$

Hence a standard error for $\widehat{\theta}$ is $V^{1/2} = \{\widehat{\theta}(1-\widehat{\theta})/n\}^{1/2}$.

— The cases $\theta = 0, 1$ are not interesting, as in those cases $\widehat{\theta} = 0, 1$ with probability 1.

Construction d'un IC approximatif

— En général on construit des ICs approximatifs à l'aide du théorème central limite.

— Rappelons que la plupart des statistiques se basant sur les moyennes (implicites ou explicites) des variables $Y = (Y_1, \ldots, Y_n)$ ont des lois normales pour n grand.

— Si $\tilde{\theta} = t(Y)$ est un estimateur de θ avec écart-type $\sqrt{V}$, et si

$$\tilde{\theta} \;\dot{\sim}\; N(\theta, V),$$

alors $(\tilde{\theta} - \theta)/\sqrt{V} \dot{\sim} N(0,1)$. Soit z_α le α quantile de la loi $\mathcal{N}(0,1)$, alors

$$\mathrm{P}\left\{z_{\alpha_U} < (\tilde{\theta} - \theta)/\sqrt{V} \leq z_{1-\alpha_L}\right\} \doteq \Phi(z_{1-\alpha_L}) - \Phi(z_{\alpha_U}) = 1 - \alpha_L - \alpha_U,$$

impliquant qu'un IC (approximatif) bilatéral de niveau $(1 - \alpha_L - \alpha_U)$ pour θ est

$$(L, U) = (\tilde{\theta} - \sqrt{V}z_{1-\alpha_L}, \tilde{\theta} - \sqrt{V}z_{\alpha_U}).$$

— Par défaut on prend $\alpha_L = \alpha_U = \alpha/2$ (intervalle symétrique) avec α une valeur conventionnelle telle que 0.1, 0.05, 0.01, alors les IC correspondants ont niveaux de confiances 90%, 95%, 99%, c'est-à-dire probabilités $0.9, 0.95, 0.99, \ldots$ de contenir θ.

— Pour un IC unilatéral de niveau $(1-\alpha)$ on prend un IC bilatéral avec $\alpha_L = \alpha_U = \alpha$ et on remplace la limite dont on n'a pas besoin par $\pm\infty$.

Exemples

Exemple 167. *Soient* $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathrm{Bernoulli}(\theta)$*, donner un IC approximatif de niveau 95% pour* θ*.*

Exemple 168. *Soient* $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} U(0,\theta)$ *avec* θ *inconnu et* $\overline{Y} = n^{-1}\sum Y_j$*. Utiliser le théorème central limite pour trouver un IC de niveau 90% pour* θ*.*

Note to Example 167

In this case we saw in Example 166 that $\widehat{\theta} = S/n$ and $V = \widehat{\theta}(1-\widehat{\theta})/n$, $\alpha/2 = 0.025$, $z_{\alpha/2} = -1.96$, $z_{1-\alpha/2} = 1.96$, which gives

$$(L, U) = (\widehat{\theta} - 1.96\sqrt{\widehat{\theta}(1-\widehat{\theta})/n}, \widehat{\theta} + 1.96\sqrt{\widehat{\theta}(1-\widehat{\theta})/n}) \approx \widehat{\theta} \pm 2\sqrt{\widehat{\theta}(1-\widehat{\theta})/n}.$$

Note to Example 168

— For large n, $\overline{Y} \dot{\sim} \mathcal{N}\{\theta/2, \theta^2/(12n)\}$ using the CLT, and we note that the moment estimator of θ is the solution to the equation

$$\overline{Y} = \theta/2,$$

which yields $\tilde{\theta} = 2\overline{Y}$, and this has variance $\mathrm{var}(2\overline{Y}) = 4\mathrm{var}(\overline{Y}) = 4\theta^2/(12n) = \theta^2/(3n)$. Therefore $V = \tilde{\theta}^2/(3n)$ and the standard error is $V^{1/2} = \tilde{\theta}/(3n)^{1/2}$.

— By default $\alpha_U = \alpha_L = \alpha/2 = 0.1/2 = 0.05$, and $z_{1-\alpha_L} = z_{0.95} = -z_{0.05} = -z_{\alpha_U} = 1.645$.

— Hence

$$(L, U) = (\tilde{\theta} - z_{1-\alpha_L}\tilde{\theta}/(3n)^{1/2}, \tilde{\theta} - z_{\alpha_U}\tilde{\theta}/(3n)^{1/2}) = 2\overline{Y} \pm 1.645 \times 2\overline{Y}/(3n)^{1/2}.$$

— Alternatively we can argue that as $(3n)^{1/2}(2\overline{Y}/\theta - 1) \dot{\sim} \mathcal{N}(0,1)$ and $\alpha_U = \alpha_L = \alpha/2 = 0.1/2 = 0.05$,

$$\mathrm{P}\{z_{0.05} \le (3n)^{1/2}(2\overline{Y}/\theta - 1) \le z_{0.95}\} = 0.90,$$

giving

$$L = \frac{2\overline{Y}}{1 + z_{0.95}/(3n)^{1/2}}, \quad U = \frac{2\overline{Y}}{1 + z_{0.05}/(3n)^{1/2}};$$

note that for large n these are $L \approx 2\overline{Y}\{1 - z_{0.95}/(3n)^{1/2}\}$ and $U \approx 2\overline{Y}\{1 - z_{0.05}/(3n)^{1/2}\}$, with $z_{0.95} = -z_{0.05} = 1.645$, as above.

CIs basés sur le MLE

Le théorème 161 dit que (sous des conditions de régularité) et pour n grand,

$$\widehat{\theta} \quad \dot{\sim} \quad \mathcal{N}\left\{\theta, \jmath(\widehat{\theta})^{-1}\right\}.$$

Ainsi un IC pour θ de niveau approximative de $(1-\alpha)$ est

$$(L, U) = (\widehat{\theta} - \jmath(\widehat{\theta})^{-1/2} z_{1-\alpha/2}, \widehat{\theta} + \jmath(\widehat{\theta})^{-1/2} z_{1-\alpha/2}).$$

On peut montrer que pour n grand (et un modèle régulier) aucun estimateur a une variance plus petite que $\widehat{\theta}$, ce qui implique que des ICs sont aussi minces que possible.

Découverte du top quark (Abe et al., 1995, PRL)

Voici deux extraits de l'article annonçant la découverte du top quark :

TABLE I. Number of lepton + jet events in the 67 pb^{-1} data sample along with the numbers of SVX tags observed and the estimated background. Based on the excess number of tags in events with ≥3 jets, we expect an additional 0.5 and 5 tags from $t\bar{t}$ decay in the 1- and 2-jet bins, respectively.

N_{jet}	Observed events	Observed SVX tags	Background tags expected
1	6578	40	50 ± 12
2	1026	34	21.2 ± 6.5
3	164	17	5.2 ± 1.7
≥4	39	10	1.5 ± 0.4

The numbers of SVX tags in the 1-jet and 2-jet samples are consistent with the expected background plus a small $t\bar{t}$ contribution (Table I and Fig. 1). However, for the $W+ \geq$3-jet signal region, 27 tags are observed compared to a predicted background of 6.7 ± 2.1 tags [8]. The probability of the background fluctuating to ≥27 is calculated to be 2×10^{-5} (see Table II) using the procedure outlined in Ref. [1] (see [9]). The 27 tagged jets are in 21 events; the six events with two tagged jets can be compared with four expected for the top + background hypothesis and ≤1 for background alone. Figure 1 also shows the decay lifetime distribution

Démarche du test

— On a une **hypothèse nulle** à tester :

H_0 : le top quark n'existe pas.

Ceci semble contre-intuitif, mais puisque l'on ne peut pas prouver une hypothèse, on tente de réfuter son contraire — une '**preuve par contradiction' stochastique**.

— On obtient des données, $y_{\text{obs}} = 27$ événements observés sur les 3-jet, 4-jet, ...

— On compare y_{obs} avec sa distribution P_0 calculée en supposant que H_0 est vraie.

— Dans ce cas, P_0 est $\mathrm{Poiss}(\lambda_0 = 6.7)$ et répresente le bruit de fond sous H_0.

— On calcule la p-**valeur (P-value)**

$$p_{\text{obs}} = \mathrm{P}_0(Y \geq y_{\text{obs}}) = \sum_{y=y_{\text{obs}}}^{\infty} \frac{\lambda_0^y}{y!} e^{-\lambda_0} = 3 \times 10^{-9},$$

alors

— soit H_0 est vraie mais un événement très (très !) rare s'est passé,

— soit H_0 est fausse et le top quark existe.

— Dans ce cas Abe et al. ont annoncé la découverte du top quark, mais s'ils avaient obtenu $p_{\text{obs}} \approx 0.001$, peut-être ils auraient décidé que on ne peut pas (encore) rejeter H_0, ou de ne pas publier ...('file drawer problem').

Les éléments d'un test

— Une **hypothèse nulle** H_0 à tester.

— Une **statistique de test** T, choisie de telle sorte que de grandes valeurs de T suggèrent que H_0 est fausse, et dont la valeur observée est t_{obs}.

— Un **p-valeur** p_{obs} donnant la probabilité d'observer l'évènement $T \geq t_{\text{obs}}$ sous H_0 :

$$p_{\text{obs}} = \mathrm{P}_0(T \geq t_{\text{obs}}),$$

où la **distribution nulle** $\mathrm{P}_0(\cdot)$ indique une probabilité calculée sous H_0.

— Plus p_{obs} est petite, plus on va douter que H_0 est vraie.

— Si H_0 est vraie et T est continue, alors on peut considérer que p_{obs} est une réalisation d'une variable aléatoire $P \sim U(0,1)$, et dans ce cas

$$\mathrm{P}_0(P \leq p_{\text{obs}}) = p_{\text{obs}}.$$

— Si je décide que H_0 est fausse, alors qu'elle est vraie, je fais une erreur dont la probabilité (sous H_0) est p_{obs} — ainsi mon incertitude est quantifiée, car je connais la probabilité de déclarer une fausse découverte, un "**faux positif**".

Note : Why is a P-value uniform ?

— Let T be a test statistic whose distribution is $F_0(t)$ when the null hypothesis is true. Then the corresponding P-value is

$$\mathrm{P}_0(T \geq t_{\text{obs}}) = 1 - F_0(t_{\text{obs}}),$$

and if the value of t_{obs} is a realisation of T_{obs} (because the null hypothesis is true), then we can write the random value of p_{obs} seen in repetitions of the experiment as

$$P_{\text{obs}} = 1 - F_0(T_{\text{obs}}),$$

or equivalently $T_{\text{obs}} = F_0^{-1}(1 - P_{\text{obs}})$. Hence for $x \in [0,1]$,

$$\begin{aligned}
\mathrm{P}_0(P_{\text{obs}} \leq x) &= \mathrm{P}_0\left\{1 - F_0(T_{\text{obs}}) \leq x\right\} \\
&= \mathrm{P}_0\left\{1 - x \leq F_0(T_{\text{obs}})\right\} \\
&= \mathrm{P}_0\left\{T_{\text{obs}} \geq F_0^{-1}(1-x)\right\} \\
&= 1 - F_0\left\{F_0^{-1}(1-x)\right\} \\
&= x,
\end{aligned}$$

which shows that $P_{\text{obs}} \sim U(0,1)$.

— The above proof works for any continuous T_{obs}, but is only approximate if T_{obs} is discrete (e.g., has a Poisson distribution). In such cases P_{obs} can only take a finite or countable number of values known as the **achievable significance levels**.

Le niveau de significativité

— Pour décider si H_0 doit être considérée comme fausse, on utilise souvent des **niveaux de significativité α** conventionnels, tels que $\alpha = 0.05, 0.01, 0.001$.

— Par exemple, si on a choisit $\alpha = 0.05$ (5%), on

rejette H_0 au niveau de significativité 0.05 (5%) ssi $p_{\text{obs}} < 0.05$.

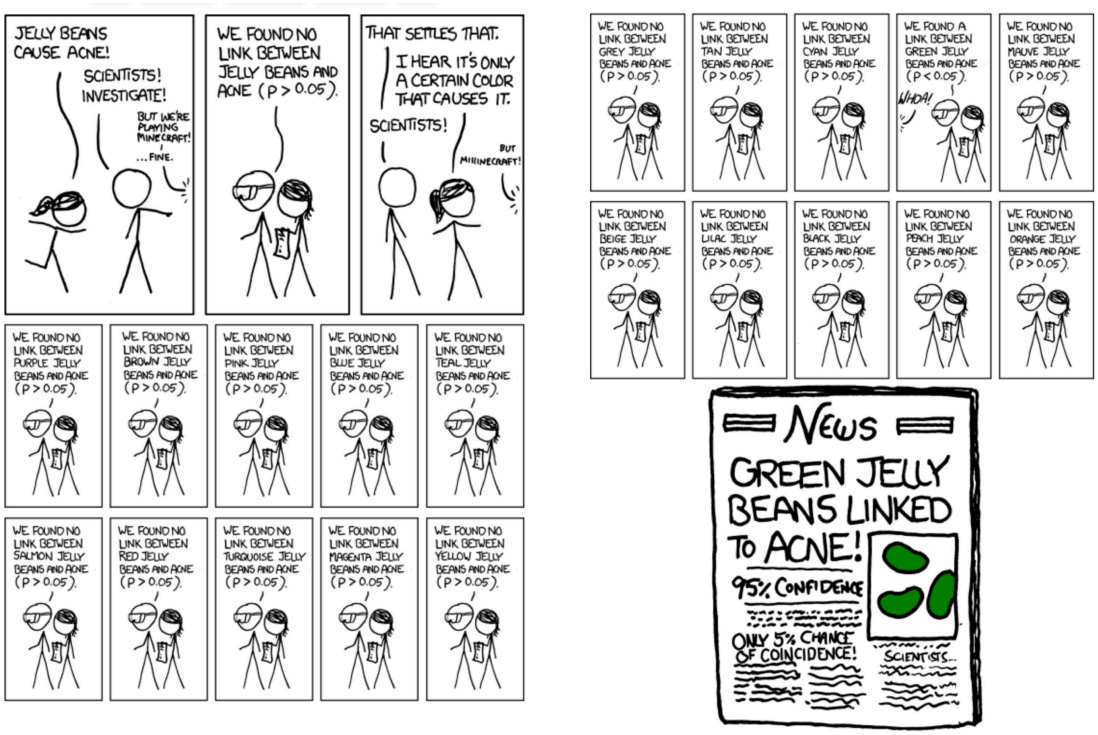
— On peut formuler ceci en termes d'une décision de rejeter ou pas H_0 sur la base du test, et on peut alors faire deux décisions correctes et deux erreurs différentes :

	Etat de la nature	
Décision	H_0 vraie	H_0 fausse
Non-rejet de H_0	Vrai négatif	**Faux négatif (erreur du type II)**
Rejet de H_0	**Faux positif (erreur du type I)**	Vrai positif

— Si H_0 est vraie et $p_{\text{obs}} < 0.05$, alors un évènement de probabilité 0.05 s'est produit.

— Si on fait N tests indépendants avec niveau de significativité α, et toutes les hypothèses nulles sont vraies, on doit alors attendre $N\alpha$ faux positifs.

Les tests multiples



Nomenclature

— Si on choisit de rejeter H_0 quand $p_{\text{obs}} < \alpha$, on pourrait également rejeter H_0 quand $t_{\text{obs}} > t_{1-\alpha}$, où $t_{1-\alpha}$ est le $(1-\alpha)$-quantile de P_0, c'est-à-dire

$$\text{rejeter } H_0 \quad \Leftrightarrow \quad p_{\text{obs}} < \alpha \quad \Leftrightarrow \quad t_{\text{obs}} > t_{1-\alpha}, \quad 0 < \alpha < 1,$$

alors on appelle $t_{1-\alpha}$ la **valeur critique de niveau α** du test.

— Pour aller plus loin, on doit préciser la situation quand H_0 est fausse. On parle alors de l'**hypothèse alternative, H_1**. Dans le cas du top quark, on a

$$H_1 : Y \sim \text{Poiss}(\lambda) \quad \text{avec} \quad \lambda > \lambda_0.$$

— Avec une H_1 specifiée, on peut définir la **probabilité d'un vrai positif/puissance (power)**,

$$\beta = \mathrm{P}(\text{rejeter } H_0 \text{ au niveau de significativité } \alpha \text{ quand } H_1 \text{ est vraie}) = \mathrm{P}_1(T \geq t_{1-\alpha}),$$

qui est une function de la **probabilité d'un faux positive/seuil (size)**,

$$\alpha = \mathrm{P}_0(T \geq t_{1-\alpha}), \quad 0 < \alpha < 1.$$

— Puisque $t_{1-\alpha}$ est fonction de α on pourrait également écrire

$$\beta(t) = \mathrm{P}_1(T \geq t), \quad \alpha(t) = \mathrm{P}_0(T \geq t).$$

Vrais et faux positifs

— Soient $H_0 : T \sim \mathcal{N}(0,1)$ et $H_1 : T \sim \mathcal{N}(\mu, 1)$, avec $\mu > 0$, et soit $t \in \mathbb{R}$.

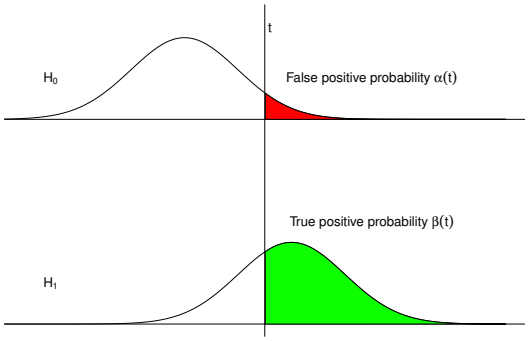
— Si on rejette H_0 si et seulement si $T \geq t$, alors

  — on rejette H_0 par erreur (faux positif) avec probabilité

$$\alpha(t) = \mathrm{P}_0(T \geq t) = 1 - \Phi(t) = \Phi(-t),$$

  — on rejette H_0 avec raison (vrai positif) avec probabilité

$$\beta(t) = \mathrm{P}_1(T \geq t) = \mathrm{P}_1(T - \mu \geq t - \mu) = 1 - \Phi(t - \mu) = \Phi(\mu - t).$$



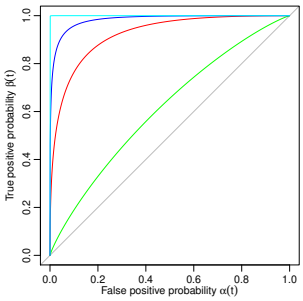
Courbe ROC

Définition 169. *La* **courbe ROC (receiver operating characteristic curve)** *d'un test est le graphe de* $\mathrm{P}_1(T \geq t)$ *comme fonction de* $\mathrm{P}_0(T \geq t)$ *pour* $t \in \mathbb{R}$.

Dans l'exemple précédent, $\alpha = \Phi(-t)$, et donc $t = -\Phi^{-1}(\alpha) = -z_\alpha$, et de façon équivalente on graphe

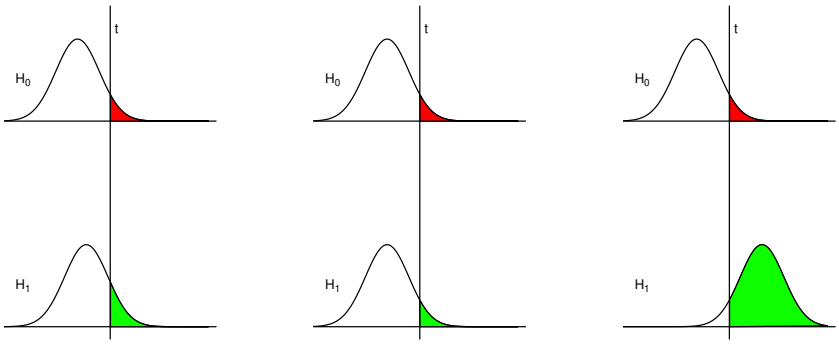
$$\beta(t) = \Phi(\mu + z_\alpha) \equiv \beta(\alpha) \text{ contre } \alpha \in (0, 1).$$

Pause pensées. *Voici la courbe ROC pour cet exemple avec* $\mu = 2$ *(rouge). On montre aussi les courbes ROC pour* $\mu = 0, 0.4, 3, 6$. *Laquelle est laquelle ?*



Example, II

— Si vous avez besoin d'aide, voici les densités dans trois des cas :



Tests optimaux

— Par 'optimal' on entend le test ayant la plus grande probabilité de rejeter H_0 si elle est fausse, c'est-à-dire les plus grande probabilité

$$\beta(\alpha) = \mathrm{P}_1(T \geq t_{1-\alpha})$$

pour un niveau de significativité $\alpha = \mathrm{P}_0(T \geq t_{1-\alpha})$ donné.

— Une hypothèse est dite **simple** si elle détermine entièrement la loi des données.

— On appelle **région de rejet au niveau α** d'un test l'ensemble $\mathcal{Y}_\alpha = \{y : t(y) \geq t_{1-\alpha}\}$.

Lemme 170 (Neyman–Pearson). *Soient $f_0(y)$, $f_1(y)$ les densités de Y sous des hypothèses nulle et alternative simples. Alors s'il existe, l'ensemble*

$$\mathcal{Y}_1 = \{y : f_1(y)/f_0(y) \geq t\}$$

tel que $\mathrm{P}_0(Y \in \mathcal{Y}_1) = \alpha$ maximise $\mathrm{P}_1(Y \in \mathcal{Y}_1)$, parmi tous les $\mathcal{Y}_1'$ tel que $\mathrm{P}_0(Y \in \mathcal{Y}_1') \leq \alpha$. Pour maximiser $\beta(\alpha)$ pour une α donnée, il faut donc rejeter H_0 ssi $Y \in \mathcal{Y}_1$.

Exemple 171. *Dans l'exemple du top quark, soient $H_0 : \lambda = \lambda_0$ et $H_1 : \lambda = \lambda_1 > \lambda_0$. Trouver le test optimal.*

Note to Lemma 170

— To simplify (and abuse) the notation, write $f_0(\mathcal{Y}) = \int_{\mathcal{Y}} f_0(y)\,\mathrm{d}y = \mathrm{P}_0(Y \in \mathcal{Y})$, etc.

— Suppose that $f_0(\mathcal{Y}) = \alpha$, i.e.,

$$\mathcal{Y} = \{y : f_1(y)/f_0(y) \geq t_{1-\alpha}\} = \{y : f_1(y) \geq t_{1-\alpha} f_0(y)\},$$

and let $\mathcal{Y}'$ be any other critical region such that $f_0(\mathcal{Y}') \leq \alpha$. Then for any density f,

$$f(\mathcal{Y}) - f(\mathcal{Y}') = f(\mathcal{Y} \cap \mathcal{Y}') + f(\mathcal{Y} \cap \overline{\mathcal{Y}'}) - f(\mathcal{Y}' \cap \mathcal{Y}) - f(\mathcal{Y}' \cap \overline{\mathcal{Y}}) = f(\mathcal{Y} \cap \overline{\mathcal{Y}'}) - f(\mathcal{Y}' \cap \overline{\mathcal{Y}}),$$

where $\overline{\mathcal{Y}}$ denotes the complement of $\mathcal{Y}$.

— Note that $f_0(\mathcal{Y}') \leq f_0(\mathcal{Y}) = \alpha$, so $f_0(\mathcal{Y}) - f_0(\mathcal{Y}') = f_0(\mathcal{Y} \cap \overline{\mathcal{Y}'}) - f_0(\mathcal{Y}' \cap \overline{\mathcal{Y}}) \geq 0$.

— If $y \in \overline{\mathcal{Y}}$, then $t_{1-\alpha} f_0(y) > f_1(y)$, while $f_1(y) \geq t_{1-\alpha} f_0(y)$ if $y \in \mathcal{Y}$. Hence

$$f_1(\mathcal{Y}) - f_1(\mathcal{Y}') = f_1(\mathcal{Y} \cap \overline{\mathcal{Y}'}) - f_1(\mathcal{Y}' \cap \overline{\mathcal{Y}}) \geq t_{1-\alpha} \left\{ f_0(\mathcal{Y} \cap \overline{\mathcal{Y}'}) - f_0(\mathcal{Y}' \cap \overline{\mathcal{Y}}) \right\} \geq 0.$$

Thus $f_1(\mathcal{Y}) = \mathrm{P}_1(Y \in \mathcal{Y}) \geq f_1(\mathcal{Y}') = \mathrm{P}_1(Y \in \mathcal{Y}')$, i.e., the power of $\mathcal{Y}$ is at least that of $\mathcal{Y}'$, and the result is established.

Note to Example 171

— Here

$$f_0(y) = \frac{\lambda_0^y}{y!}e^{-\lambda_0}, \quad f_1(y) = \frac{\lambda_1^y}{y!}e^{-\lambda_1},$$

so

$$\frac{f_1(y)}{f_0(y)} = \left(\frac{\lambda_1}{\lambda_0}\right)^y \exp(\lambda_0 - \lambda_1) = \exp\left\{y\log(\lambda_1/\lambda_0) + \lambda_0 - \lambda_1\right\},$$

and since $\lambda_1 > \lambda_0$ (and both λ_0 and λ_1 are known because the hypotheses are simple),

$$\exp\left\{y\log(\lambda_1/\lambda_0) + \lambda_0 - \lambda_1\right\} \geq t \quad \Leftrightarrow \quad y \geq y_{1-\alpha} = \frac{\log t - (\lambda_0 - \lambda_1)}{\log(\lambda_1/\lambda_0)},$$

where the right-hand size is known for a specified t.

— Hence the optimal critical region is of the form $\{Y : Y \geq y_{1-\alpha}\}$, where $y_{1-\alpha}$ is chosen as the $(1-\alpha)$-quantile of the distribution of Y under H_0, i.e.,

$$\alpha = \mathrm{P}_0(Y \geq y_{1-\alpha}) = \sum_{y=y_{1-\alpha}}^{\infty} \frac{\lambda_0^y}{y!}e^{-\lambda_0}.$$

This is precisely the form of the test used by Abe et al. (and of course is the obvious, essentially the only, choice in this case).

— As the Poisson distribution is discrete, $y_{1-\alpha}$ can only be found exactly for certain values of α. In fact, for $y = 1, \ldots,$ and $\lambda = 6.7$, we have $\mathrm{P}(Y \geq y)$ equal to the achievable significance levels

$$0.999,\ 0.991,\ 0.963,\ 0.901,\ 0.798,\ 0.659,\ 0.505,\ 0.357,\ 0.233,\ 0.140,\ 0.079,\ 0.041,\ \ldots$$

so we cannot choose $\alpha = 0.05$. One (unsatisfactory) solution to this is to randomise, but it's better just to compute p_{obs} and see if it is smaller than the α of interest.

5.6 Inférence Bayesienne

slide 208

Inférence bayésienne

Jusqu'à ici nous avons supposé que toute information à propos de θ provient des données y. Mais si on des connaissances a priori sur θ sous forme d'une **densité a priori** (anglais **prior density**)

$$\pi(\theta),$$

on peut trouver la **densité a posteriori** (anglais **posterior density**) pour θ, sachant les données y,

$$\pi(\theta \mid y) = \frac{f(y \mid \theta)\pi(\theta)}{f(y)},$$

par le théorème de Bayes. On peut baser $\pi(\theta)$ sur

— des données séparées de y ;

— une notion 'objective' de ce qu'il est 'raisonnable' de croire à propos de θ ;

— une notion 'subjective' de ce que 'je' crois à propos de θ.

On considèrera $\pi(\theta)$ après discussion de la mechanisme bayésienne.

Rappel : Théorème de Bayes

Soient $B_1, \ldots, B_k$ une partition de l'espace des échantillons E, et soit A un évènement quelconque de l'espace des échantillons. Alors

$$\begin{aligned} \mathrm{P}(B_i \mid A) &= \frac{\mathrm{P}(A \cap B_i)}{\mathrm{P}(A)} \\ &= \frac{\mathrm{P}(A \mid B_i)\mathrm{P}(B_i)}{\mathrm{P}(A)} \\ &= \frac{\mathrm{P}(A \mid B_i)\mathrm{P}(B_i)}{\sum_{j=1}^{k} \mathrm{P}(A \mid B_j)\mathrm{P}(B_j)}. \end{aligned}$$

Interprétation : la connaissance de la réalisation de l'évènement A met à jour les probabilités des évènements $B_1, \ldots, B_k$:

$$\mathrm{P}(B_1), \ldots, \mathrm{P}(B_k) \quad \longrightarrow \quad \mathrm{P}(B_1 \mid A), \ldots, \mathrm{P}(B_k \mid A).$$

Application du théorème de Bayes

On suppose que le paramètre θ a pour densité $\pi(\theta)$, et que la densité conditionelle de Y sachant θ, est $f(y \mid \theta)$. La densité conjointe est

$$f(y, \theta) = f(y \mid \theta)\pi(\theta),$$

et par le théorème de Bayes la densité conditionelle de θ sachant que $Y = y$ est

$$\pi(\theta \mid y) = \frac{f(y \mid \theta)\pi(\theta)}{f(y)},$$

où

$$f(y) = \int f(y \mid \theta)\pi(\theta)\, d\theta$$

est la densité marginale des données Y.

Mise à jour bayésienne

D'où l'utilisation du théorème de Bayes pour mettre à jour la densité a priori de θ en une densité a posteriori de θ :

$$\pi(\theta) \quad \stackrel{y}{\longrightarrow} \quad \pi(\theta \mid y),$$

ou de manière équivalente

$$\text{incertitude a priori} \quad \stackrel{\text{données}}{\longrightarrow} \quad \text{incertitude a posteriori.}$$

Nous utilisons $\pi(\theta), \pi(\theta \mid y)$ (plutôt que $f(\theta), f(\theta \mid y)$) pour expliciter que ces lois dépendent des informations extérieures aux données.

La densité Beta(a, b)

Définition 172. *La densité beta(a, b) pour $\theta \in (0, 1)$ a la forme*

$$\pi(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)}, \quad 0 < \theta < 1, \quad a, b > 0,$$

où a et b sont les paramètres, $B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a + b)$ est la fonction beta, et

$$\Gamma(a) = \int_0^\infty u^{a-1} e^{-u}\, du, \quad a > 0,$$

est la fonction gamma.

Noter que $a = b = 1$ donne la densité $U(0, 1)$.

Exemple 173. *Montrer que si $\theta \sim$ Beta(a, b), alors*

$$\mathrm{E}(\theta) = \frac{a}{a+b}, \quad \mathrm{var}(\theta) = \frac{ab}{(a+b+1)(a+b)^2}.$$

Exemple 174. *Calculer la densité a posteriori de θ pour une suite d'essais de Bernoulli, si la densité a priori est Beta(a, b).*

Note to Example 173

Since π is a density function, we have

$$\int_0^1 \pi(\theta)\, d\theta = 1,$$

and therefore

$$\int_0^1 \theta^{a-1}(1-\theta)^{b-1}\, d\theta = B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}, \quad a, b > 0.$$

This implies that

$$\mathrm{E}(\theta^r) = \int_0^1 \theta^r \pi(\theta)\, d\theta = \frac{1}{B(a,b)} \int_0^1 \theta^{r+a-1}(1-\theta)^{b-1}\, d\theta = \frac{B(a+r,b)}{B(a,b)} = \frac{\Gamma(a+r)\Gamma(a+b)}{\Gamma(a)\Gamma(a+b+r)},$$

and since $\Gamma(a + 1)/\Gamma(a) = a$, for $a > 0$, we have

$$\mathrm{E}(\theta) = \frac{a}{a+b}, \quad \mathrm{E}(\theta^2) = \frac{a(a+1)}{(a+b)(a+b+1)}, \quad \mathrm{var}(\theta) = \frac{ab}{(a+b+1)(a+b)^2}.$$

Note to Example 174

Suppose that conditional on θ, the data $y_1, \ldots, y_n$ are a random sample from the Bernoulli distribution, for which $\mathrm{P}(Y_j = 1) = \theta$ and $\mathrm{P}(Y_j = 0) = 1 - \theta$, where $0 < \theta < 1$. The likelihood is

$$L(\theta) = f(y \mid \theta) = \prod_{j=1}^{n} \theta^{y_j}(1-\theta)^{1-y_j} = \theta^s(1-\theta)^{n-s}, \quad 0 < \theta < 1,$$

where $s = \sum y_j$.
The posterior density of θ conditional on the data and using the beta prior density is given by Bayes' theorem, and is

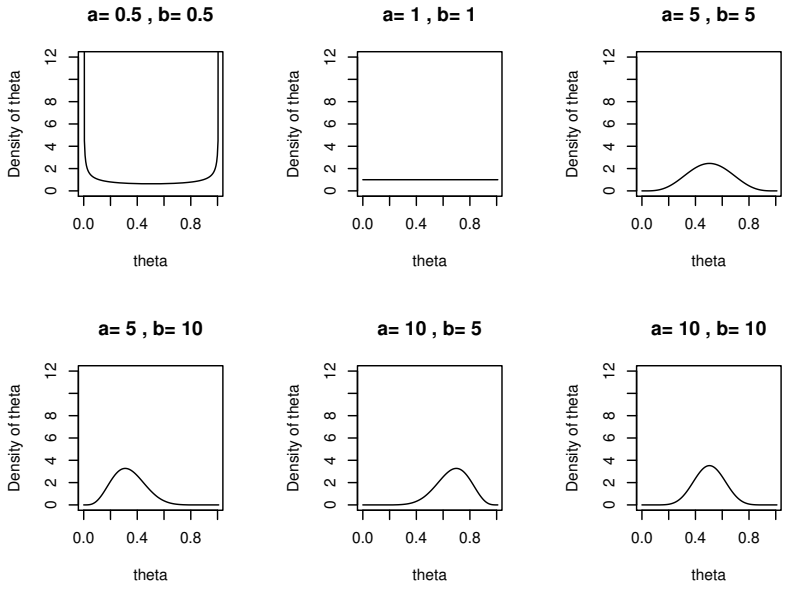
$$\begin{aligned} \pi(\theta \mid y) &= \frac{\theta^{s+a-1}(1-\theta)^{n-s+b-1}/B(a,b)}{\int_0^1 \theta^{s+a-1}(1-\theta)^{n-s+b-1}\, d\theta / B(a,b)} \\ &\propto \theta^{s+a-1}(1-\theta)^{n-s+b-1}, \quad 0 < \theta < 1. \end{aligned} \tag{4}$$

As this has unit integral for all positive a and b, the constant normalizing (4) must be $B(a+s, b+n-s)$. Therefore

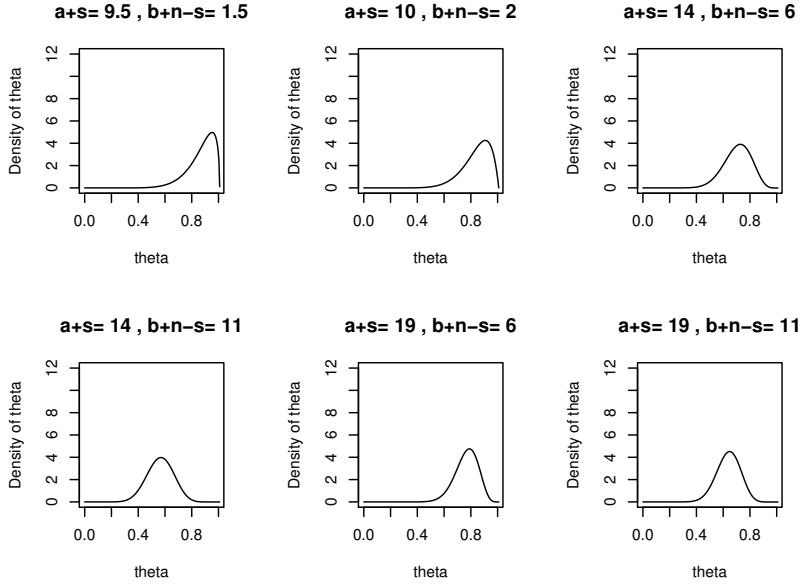
$$\pi(\theta \mid y) = \frac{1}{B(a+s, b+n-s)} \theta^{s+a-1}(1-\theta)^{n-s+b-1}, \quad 0 < \theta < 1.$$

Thus the posterior density of θ has the same form as the prior : acquiring data has the effect of updating (a, b) to $(a+s, b+n-s)$. As the mean of the $B(a,b)$ density is $a/(a+b)$, the posterior mean is $(s+a)/(n+a+b)$, and this is roughly s/n in large samples. Hence the prior density inserts information equivalent to having seen a sample of $a+b$ observations, of which a were successes. If we were very sure that $\theta \doteq 1/2$, for example, we might take $a = b$ very large, giving a prior density tightly concentrated around $\theta = 1/2$, whereas taking smaller values of a and b would increase the prior uncertainty.

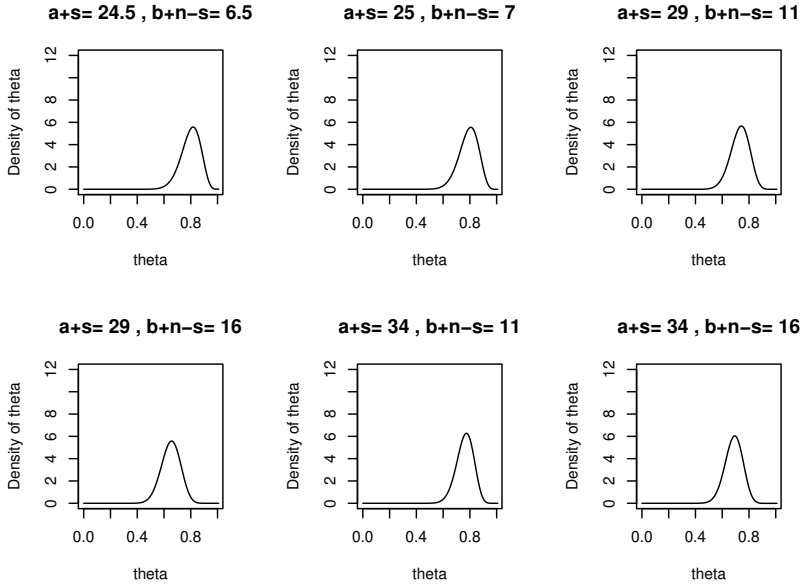
Densités a priori



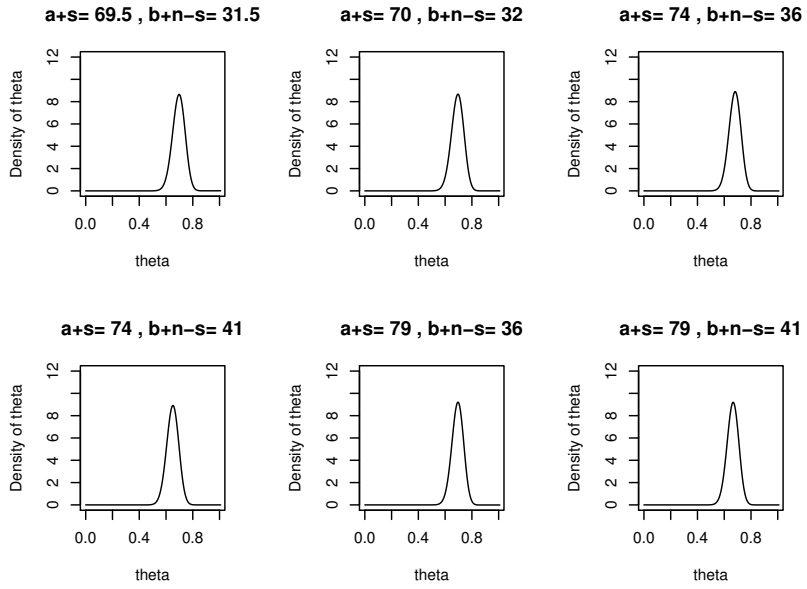
Densités a posteriori avec $n = 10$, $s = 9$



Densités a posteriori avec $n = 30$, $s = 24$



Densités a posteriori avec $n = 100$, $s = 69$



Interprétation de $\pi(\theta \mid y)$

— $\pi(\theta \mid y)$ contient notre 'information' au sujet de θ ayant vu les données y, quand notre 'information initiale' de θ est résumée dans la densité $\pi(\theta)$.

— La densité contient toute cette information, mais il est parfois utile d'extraire des résumés, tel que **l'espérance a posteriori** ou la **variance a posteriori**,

$$\mathrm{E}(\theta \mid y), \quad \mathrm{var}(\theta \mid y),$$

ou l'**estimation maximum a posteriori (estimation MAP)**, c'est à dire $\tilde{\theta}$ tel que

$$\pi(\tilde{\theta} \mid y) \geq \pi(\theta \mid y), \quad \forall \theta.$$

Exemple 175. *Calculer l'espérance et la variance a posteriori de θ, et son estimation MAP, pour l'exemple précédent.*

Note to Example 175

We saw earlier that the beta density with parameters a, b has

$$\mathrm{E}(\theta) = \frac{a}{a+b}, \quad \mathrm{E}(\theta^2) = \frac{a(a+1)}{(a+b)(a+b+1)}, \quad \mathrm{var}(\theta) = \frac{ab}{(a+b+1)(a+b)^2},$$

so since the posterior density is beta with parameters $a+s$, $b+n-s$, the posterior mean and variance are

$$\mathrm{E}(\theta \mid y) = \frac{a+s}{a+b+n}, \quad \mathrm{var}(\theta \mid y) = \frac{(a+s)(b+n-s)}{(a+b+n+1)(a+b+n)^2}.$$

The MAP estimate is obtained by maximising the density

$$\pi(\theta \mid y) = \frac{1}{B(a+s, b+n-s)} \theta^{s+a-1}(1-\theta)^{n-s+b-1}, \quad 0 < \theta < 1.$$

as a function of θ. Taking logs and differentiating gives that $\pi(\theta \mid y)$ is maximised by

$$\tilde{\theta} = \frac{a+s-1}{n+a+b-2}.$$

Les intervalles de crédibilité

— L'équivalent de l'IC à $(1-\alpha)$ pour θ, est **l'intervalle de crédibilité de niveau $(1-\alpha)$ de θ** obtenu en utilisant les quantiles $\alpha/2$ et $(1-\alpha/2)$ de $\pi(\theta \mid y)$.

— En prenant $\alpha = 0.05, a = b = 0.5$, on obtient

	$n=10$	$n=30$	$n=100$	$\widehat{\theta} \pm 1.96\jmath(\widehat{\theta})^{-1/2}$
Lower	0.619	0.633	0.595	0.599
Upper	0.989	0.912	0.774	0.781

Ici $\widehat{\theta}$ est le MLE de θ, et $\jmath(\widehat{\theta})$ est l'information observée.

— a, b n'ont que peu d'influence pour des grands échantillons, car les données contiennent alors beaucoup d'information sur θ.

Fonctions de perte

Pour construire un estimateur basé sur les données y, on considère que le choix d'estimation correspond à une décision, et on cherche à minimiser la perte potentielle.

Définition 176. *Soit $Y \sim f(y;\theta)$, alors une* **fonction de perte** $L(\tilde{\theta};\theta)$ *est une fonction non-négative de $\tilde{\theta}(y)$ et de θ. La* **perte moyenne a posteriori** *est*

$$\mathrm{E}\left\{L(\tilde{\theta};\theta) \mid y\right\} = \int L\{\tilde{\theta}(y);\theta\}\pi(\theta \mid y)\,\mathrm{d}\theta.$$

Exemple 177. *Si je cherche à estimer θ avec $\tilde{\theta}(y)$ en minimisant $\mathrm{E}\left\{L(\tilde{\theta};\theta) \mid y\right\}$ par rapport à $\tilde{\theta}$, montrer qu'avec*

$$L(\tilde{\theta};\theta) = (\tilde{\theta}-\theta)^2, \quad RL(\tilde{\theta};\theta) = |\tilde{\theta}-\theta|,$$

j'ai respectivement $\tilde{\theta} = \mathrm{E}(\theta \mid y)$ et $\tilde{\theta}$ la médiane de $\pi(\theta \mid y)$.

Cette idée est utile aussi quand on veut baser une décision sur les données : on construit $L(\tilde{\theta};\theta)$ pour représenter la perte quand on observe y et y base la décision, mais l'état de réalité est θ.

Note to Example 177

— For $L_1(\tilde{\theta};\theta) = (\tilde{\theta}-\theta)^2$, we have on setting $m(y) = \mathrm{E}(\theta \mid y)$ and using a little algebra that

$$\begin{aligned}\mathrm{E}\left\{L_1(\tilde{\theta};\theta) \mid y\right\} &= \mathrm{E}\left[\{\tilde{\theta}-m(y)+m(y)-\theta\}^2 \mid y\right] \\ &= \mathrm{E}\left[\{\tilde{\theta}-m(y)\}^2 \mid y\right] + 2\mathrm{E}\left[\{\tilde{\theta}-m(y)\}\{m(y)-\theta\} \mid y\right] + \mathrm{E}\left[\{m(y)-\theta\}^2 \mid y\right] \\ &= \{\tilde{\theta}-m(y)\}^2 + \mathrm{var}(\theta \mid y)\,:\end{aligned}$$

the first term is constant with respect to the posterior distribution of θ because $\tilde{\theta}$ and $m(y)$ do not depend on θ, but only on the variable y, which is fixed by the conditioning; the second term is

$$\mathrm{E}\left[\{\tilde{\theta}-m(y)\}\{m(y)-\theta\} \mid y\right] = \{\tilde{\theta}-m(y)\}E\left\{m(y)-\theta \mid y\right\} = 0;$$

and the third term is just the conditional variance of θ, given y.

Therefore we minimise $\mathrm{E}\left\{L_1(\tilde{\theta};\theta) \mid y\right\}$ by choosing $\tilde{\theta} = m(y)$ for all y, since this ensures that $\{\tilde{\theta}-m(y)\}^2$ is identically zero, and $\mathrm{var}(\theta \mid y)$ does not depend on $\tilde{\theta}$.

— For $L_2(\tilde{\theta};\theta) = |\tilde{\theta}-\theta|$, we have

$$\begin{aligned}\mathrm{E}\left\{L_2(\tilde{\theta};\theta) \mid y\right\} &= \mathrm{E}\left\{(\tilde{\theta}-\theta)I(\tilde{\theta}>\theta) \mid y\right\} + \mathrm{E}\left\{(\theta-\tilde{\theta})I(\tilde{\theta}<\theta) \mid y\right\} \\ &= \int_{-\infty}^{\tilde{\theta}} (\tilde{\theta}-\theta)\pi(\theta \mid y)\,d\theta + \int_{\tilde{\theta}}^{\infty} (\theta-\tilde{\theta})\pi(\theta \mid y)\,d\theta,\end{aligned}$$

and differentiation of this with respect to $\tilde{\theta}$ gives

$$\int_{-\infty}^{\tilde{\theta}} \pi(\theta \mid y)\,d\theta - \int_{\tilde{\theta}}^{\infty} \pi(\theta \mid y)\,d\theta$$

This equals zero when the two probabilities are the same, and then we must have

$$\int_{-\infty}^{\tilde{\theta}} \pi(\theta \mid y)\,d\theta = \int_{\tilde{\theta}}^{\infty} \pi(\theta \mid y)\,d\theta = \tfrac{1}{2},$$

so $\tilde{\theta} \equiv \tilde{\theta}(y)$ is the median of the posterior density $\pi(\theta \mid y)$.

Densités conjuguées

Des combinaisons particulières de données et de densités a priori engendrent des densités a posteriori de la même forme que celles a priori. Exemple :

$$\theta \sim \text{Beta}(a,b) \quad \xrightarrow{s,n} \quad \theta \mid x \sim \text{Beta}(a+s, b+n-s),$$

où les données $s \sim B(n,\theta)$.
La densité beta est dite **conjuguée** avec la binomial. C'est une idée très utile, car souvent on peut éviter de devoir intégrer. Ainsi :

Si l'on reconnaît $\pi(\theta \mid y)$, pas besoin d'intégrer !

Exemple 178. *Soient $Y_1, \ldots, Y_n \mid \mu \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ et $\mu \sim \mathcal{N}(\mu_0, \tau^2)$, ou σ^2 et τ^2 sont connus. Calculer la loi a posteriori de $\mu \mid Y_1, \ldots, Y_n$, sans faire d'intégration.*

Exemple 179. *Si*

$$y_{n\times 1} \mid \beta_{p\times 1}, \sigma^2 \sim \mathcal{N}_n(X\beta, \sigma^2 I_n), \quad \beta \mid \sigma^2 \sim \mathcal{N}_p(\beta_0, \sigma^2 R/\lambda), \quad \sigma^2 \propto 1/\sigma^2 \quad \lambda > 0,$$

démontrer que l'estimation MAP de β quand $\beta_0 = 0$ et $R = I_p$ est l'estimateur ridge de β.

Ce dernier exemple illustrate comment une loi a priori peut servir pour la régularisation d'un estimateur.

Note to Example 178

— The $\mathcal{N}(B, A)$ density $(2\pi A)^{-1/2} \exp\left\{-(x-B)^2/(2A)\right\}$ has exponent

$$x^2\left(-\frac{1}{2A}\right) + x\left(\frac{B}{A}\right) - \tfrac{1}{2}B^2/A,$$

so we can read off A and B from the coefficients of x^2 and x.

— We seek a density for μ, so any terms not involving μ can be treated as constants. Now

$$\begin{aligned}\pi(\mu \mid y) &\propto f(y \mid \mu) \times \pi(\mu) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_j-\mu)^2\right\} \times \frac{1}{\sqrt{2\pi\tau^2}} \exp\left\{-\frac{1}{2\tau^2}(\mu-\mu_0)^2\right\},\end{aligned}$$

whose exponent factorises as

$$\mu^2\left\{-\tfrac{1}{2}\left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right)\right\} + \mu\left(\frac{\sum y_j}{\sigma^2}+\frac{\mu_0}{\tau^2}\right) + \text{const},$$

and on comparing with the expression above, we have

$$A = \left(n/\sigma^2 + 1/\tau^2\right)^{-1}, \quad B/A = \sum y_j/\sigma^2 + \mu_0/\tau^2.$$

Thus

$$\mu \mid y \sim \mathcal{N}\left(\frac{\sum y_j/\sigma^2 + \mu_0/\tau^2}{n/\sigma^2 + 1/\tau^2}, \frac{1}{n/\sigma^2 + 1/\tau^2}\right).$$

Note that as $n \to \infty$ or $\tau^2 \to \infty$, this becomes $\mathcal{N}(\overline{y}, \sigma^2/n)$.

Note to Example 179

— We use the same argument as in the previous example, noting that

$$\pi(\beta, \sigma^2 \mid y) \propto f(y \mid \beta, \sigma^2)\pi(\beta \mid \sigma^2)\pi(\sigma^2),$$

and that the logarithm of the right-hand side equals

$$-\frac{1}{2\sigma^2}(y - X\beta)^{\mathrm{T}}(y - X\beta) - \frac{n}{2}\log\sigma^2 - \frac{\lambda}{2\sigma^2}(\beta - \beta_0)^{\mathrm{T}}R^{-1}(\beta - \beta_0) - \frac{p}{2}\log(\sigma^2/\lambda) - \log\sigma^2$$

plus some constants, and in terms of β this gives constants plus

$$-\frac{1}{2\sigma^2}\left(\beta^{\mathrm{T}}X^{\mathrm{T}}X\beta - 2y^{\mathrm{T}}X\beta + \lambda\beta^{\mathrm{T}}R^{-1}\beta - 2\lambda\beta_0^{\mathrm{T}}R^{-1}\beta\right).$$

— Hence in terms of the previous example we have

$$A^{-1} = (X^{\mathrm{T}}X + \lambda R^{-1})/\sigma^2, \quad A^{-1}B = (y^{\mathrm{T}}X\beta + \lambda\beta_0^{\mathrm{T}}R^{-1})/\sigma^2,$$

i.e.,

$$\beta \mid y, \sigma^2 \sim \mathcal{N}_p\left\{(X^{\mathrm{T}}X + \lambda R^{-1})^{-1}(X^{\mathrm{T}}y + \lambda R^{-1}\beta_0), \sigma^2(X^{\mathrm{T}}X + \lambda R^{-1})^{-1}\right\}.$$

This is maximised at its mean, and if $\beta_0 = 0$ and $R = I_p$ we obtain MAP estimator

$$(X^{\mathrm{T}}X + \lambda I_p)^{-1}X^{\mathrm{T}}y,$$

which is also the ridge estimator.

Prédiction d'une future variable aléatoire Z

Est-ce que le prochain résultat sera pile ($Z = 0$) ou face ($Z = 1$) ?
Utiliser le théorème de Bayes pour calculer la densité a posteriori de Z sachant $Y = y$:

$$\mathrm{P}(Z = z \mid Y = y) = \frac{\mathrm{P}(Z = z, Y = y)}{\mathrm{P}(Y = y)} = \frac{\int f(z, y \mid \theta)\pi(\theta)\, d\theta}{\int f(y \mid \theta)\pi(\theta)\, d\theta}.$$

Exemple 180. *Calculer la loi a posteriori pour un autre essai de Bernoulli, indépendant des précédents.*

Rappel : $B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a + b)$, and $\Gamma(a + 1) = a\Gamma(a)$, $a, b > 0$.

Note to Example 180

Let θ be the unknown probability of a head and let $Z = 1$ indicate the event that the next toss yields a head. Conditional on θ, $\mathrm{P}(Z = 1 \mid y, \theta) = \theta$ independent of the data y so far. If the prior density for θ is beta with parameters a and b, then

$$\begin{aligned}
\mathrm{P}(Z = 1 \mid y) &= \int_0^1 \mathrm{P}(Z = 1 \mid \theta, y)\pi(\theta \mid y)\, d\theta \\
&= \int_0^1 \theta\, \frac{\theta^{a+s-1}(1-\theta)^{b+n-r-1}}{B(a+s, b+n-s)}\, d\theta \\
&= \frac{B(a+s+1, b+n-s)}{B(a+s, b+n-s)} = \frac{a+s}{a+b+n},
\end{aligned}$$

on using results for beta functions. As $n, s \to \infty$, this tends to the sample proportion of heads s/n, so the prior information is drowned by the sample.

L'approche bayésienne

- On traite chaque inconnu (paramètre θ, prédicat Z, ...) comme une variable aléatoire, donner lui une distribution (en utilisant souvent l'indépendance), et calculer sa distribution a posteriori sachant les données, en utilisant le théorème de Bayes.
- On pait en devant construire un modèle plus élaboré, avec de l'information a priori, mais on gagne en pouvant traiter tous les inconnus sur le même base—paramètres, données, valeurs manquantes, prédicats, etc.—et donc on n'a qu'à appliquer les lois de probabilité, basant l'inférence sur ce que l'on a observé.
- **Questions philosophique :**
  - Est ce justifié d'incorporer les connaissances a priori de cette manière ?
  - D'où proviennent-elles ?
  - Souvent on choisit les lois a priori pour des raisons pratiques (e.g., lois conjugées) plutôt que philosophiques.
- **Question pratique :**
  - Comment faire tous les intégrales dont on a besoin ?
  - Souvent on utilise les méthodes de Monte Carlo, qui construisent les chaînes de Markov dont les lois limites sont les lois a posteriori $\pi(\theta \mid y)$. C'est une histoire pour un autre jour ...