Pourquoi un module de proba-stats ?

Éveiller l'esprit critique

- comprendre et analyser des données chiffrées
- produire des "chiffres"

Modélisation mathématique

- module spécifique (M3202) - cf tout à l'heure
- maths partout derrière les nouvelles technologies
- ici modélisation aléatoire Blockchain logique
- les bases et trois exemples

Modeliser l'aléatoire

Vocabulaire

- aléatoire = dû au hasard (= stochastique)
- modèle = idéal mathématique pour décrire la réalité

La **théorie des probabilités** est une théorie mathématique destinée à maîtriser et quantifier le hasard

Bases de la théorie des probabilités

Deux ingrédients

- Ω : "univers" = ensemble de tous les possibles
- $\mathbb{P}$: "(mesure de)probabilité" = application qui associe à chaque événement E associe la probabilité que E se produise, notée $\mathbb{P}(E)$

Un vocabulaire spécifique

- $\mathbb{P}(E) = 0$: E est impossible
- $\mathbb{P}(E) = 1$: E est certain
- $A \cap B = \emptyset$: A et B sont incompatibles
- $\mathbb{P}(A \cap B) = \mathbb{P}(A) \times \mathbb{P}(B)$: A et B sont indépendants
- $\frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}$: probabilité conditionnelle de B sachant A, notée $\mathbb{P}_A(B)$

Bases de la théorie des probabilités (suite et fin)

Trois règles

- $0 \leq \mathbb{P}(E) \leq 1$
- si $A \cap B = \emptyset$ alors $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$
- si A et B sont indépendants alors $\mathbb{P}(A \cap B) = \mathbb{P}(A) \times \mathbb{P}(B)$

Conséquences

- $\mathbb{P}(\emptyset) = 0$ et $\mathbb{P}(\Omega) = 1$
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$
- $\mathbb{P}(A \cap B) = \mathbb{P}(A) \times \mathbb{P}_A(B)$
- si $\Omega = A_1 \cup A_2 \cup \cdots \cup A_n$ avec les (A_i) 2 à 2 incompatibles, alors
  $\mathbb{P}(B) = \mathbb{P}(A_1)\mathbb{P}_{A_1}(B) + \mathbb{P}(A_2)\mathbb{P}_{A_2}(B) + \cdots + \mathbb{P}(A_n)\mathbb{P}_{A_n}(B)$

On en déduit la fameuse FORMULE DE BAYES !

Module M3201 - Probabilités et Statistiques - Semestre 3

- 7 semaines avec 1 interro durant la semaine du 1er octobre ($Int1$)
- 1 DST durant la semaine du 5 novembre ($DST1$)
- 7 semaines avec 1 interro durant la semaine ?? ($Int2$)
- 1 DST durant la semaine du 14 janvier ($DST2$)

$\Rightarrow$ note du Module $= \frac{1}{8}(Int1 + Int2 + 3 \times DST1 + 3 \times DST2)$

ATTENTION : cela va passer TRÈS vite !!

Methode 100% gagnante ;-)

- AMPHI : slides et/ou tableau
  Je suis présent.e$\Rightarrow$ je comprends et je prends des notes $\Rightarrow$ je retiens
- TD ou TP : fiche distribuée
  Je suis actif.ve $\Rightarrow$ je comprends $\Rightarrow$ j'apprends et je sais refaire
- HOME : Je travaille cours+TD+TP et je refais exples+exos
- J'ai une question ?
  AE, chargés de TD, B2-01, email...
- Je suis absent.e à une interro?
  $\rightarrow$ je préviens mon chargé de TD et je justifie auprès du secrétariat; rattrapage en fin de semestre.

4 chapitres

1. Variables aléatoires et simulations
2. Théorèmes limites et intervalles de confiance
3. Tests statistiques
4. Chaînes de Markov

ATTENTION : cela va passer TRÈS vite !!

Pourquoi un module de proba-stats ?

1. Pour gagner au loto ?
2. Pour ennuyer les étudiants ?
3. Pour éveiller l'esprit critique ?
4. Pour comprendre la modélisation mathématique ?

Loi uniforme sur [0, 1]

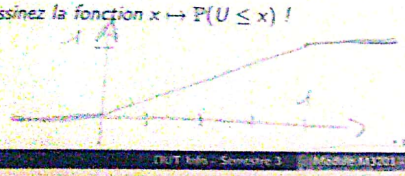
Définition

On dit que la variable aléatoire U suit la *loi uniforme sur* $[0,1]$ si

$$\forall x \in [0,1],\ \mathbb{P}(U \leq x) = x$$

et $\mathbb{P}(U \leq x) = 0$ si $x < 0$, $\mathbb{P}(U \leq x) = 1$ si $x > 1$

Dessinez la fonction $x \mapsto \mathbb{P}(U \leq x)$!



Loi uniforme sur [0, 1] (suite)

Propriétés

1. $\mathbb{E}(U) = \text{Med}(U) = 1/2$, *l'espérance et la médiane valent 1/2*
2. $\text{Var}(U) = 1/12$, *la variance vaut 1/12*
3. on a souvent besoin d'une famille $(U_1, U_2, \ldots, U_n)$ de v.a. i.i.d. de loi uniforme sur $[0,1]$, autrement dit

$$\forall x_1, \ldots, x_n \in [0,1],\ \mathbb{P}(U_1 \leq x_1, \ldots, U_n \leq x_n) = x_1 x_2 \cdots x_n$$

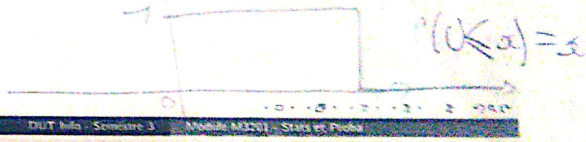
1) est évident :-)... mais 2) ne l'est pas :-(
dans 3) "i.i.d." signifie *indépendantes et identiquement distribuées*

Loi uniforme sur [0, 1] (suite)

Remarque/Vocabulaire

- la fonction $x \in \mathbb{R} \mapsto \mathbb{P}(U \leq x)$ s'appelle *fonction de répartition* de U
- cette fonction est croissante, limite 0 en $-\infty$ et limite 1 en $+\infty$
- pour tout x dans $\mathbb{R}$, $\mathbb{P}(U \leq x) = \int_{-\infty}^{x} f(t)\,dt$ avec une fonction f bien choisie (*à vous de bien choisir !*)
- la fonction f s'appelle *densité* de U

Faites un dessin !



Loi uniforme sur [a, b]

a et b sont des réels tels que $a < b$

Remarque

Si U prend ses valeurs "au hasard" dans $[0,1]$ alors $a + (b-a)U$ prend ses valeurs "au hasard" dans $[a,b]$. Et la réciproque est également vraie!

On en déduit

- un principe de simulation à partir d'une v.a. unif. sur $[0,1]$
- une propriété/définition:

$$V \text{ suit une loi unif sur } [a,b] \Leftrightarrow \forall y \in [a,b],\ \mathbb{P}(V \leq y) = \frac{y-a}{b-a}$$

Loi normale $N(0,1)$

Définition

On dit que la variable aléatoire N suit la *loi normale* $N(0,1)$ si

$$\forall x \in \mathbb{R},\ \mathbb{P}(N \leq x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt$$

ne pas apprendre

Remarque/Vocabulaire

- la fonction $x \in \mathbb{R} \mapsto \mathbb{P}(N \leq x)$ s'appelle *fonction de répartition* de N
- cette fonction est croissante, limite 0 en $-\infty$ et limite 1 en $+\infty$
- la fonction $t \mapsto f(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$ s'appelle *densité* de N

Faites un dessin !

densité

Loi normale $N(0,1)$ (suite)

Propriétés

1. $\mathbb{E}(N) = \text{Med}(N) = 0$ *l'espérance et la médiane valent 0*
2. $\text{Var}(N) = 1$ *la variance vaut 1*
3. il n'existe pas de formule pour calculer $\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt$, on utilise donc des valeurs approchées (*table, logiciel,...*)

1) est évident ;-)... 2) ne l'est pas ;-(

Loi normale $N(m, \sigma^2)$

Soient $m \in \mathbb{R}$ et $\sigma \in]0, +\infty[$

Définition

On dit que la variable aléatoire Z suit la *loi normale* $N(m, \sigma^2)$ si

$$\frac{Z - m}{\sigma} \text{ suit la loi } N(0,1)$$

On en déduit

- un principe de simulation à partir d'une loi $N(0,1)$
- la densité d'une v.a. de loi $N(m, \sigma^2)$: $t \mapsto \frac{1}{\sigma\sqrt{2\pi}} e^{-(t-m)^2/(2\sigma^2)}$

Loi normale $N(m, \sigma^2)$ (suite)

Soient $m \in \mathbb{R}$ et $\sigma \in]0, +\infty[$

Propriété fondamentale

Si Z suit la loi normale $N(m, \sigma^2)$ alors

$$\mathbb{E}(Z) = m \text{ et } \text{Var}(Z) = \sigma^2$$

Savez-vous pourquoi?...

$\mathbb{E}(Z) = \mathbb{E}(m + \sigma N)$
$= m + (\sigma \mathbb{E}(N))$
$= m + \sigma \times 0 = m$

Réponse...

En utilisant les propriétés suivantes de l'espérance :

Soient X et Y des variables aléatoires et soit $c \in \mathbb{R}$

$$\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y) \text{ et } \mathbb{E}(cX) = c\mathbb{E}(X)$$

et de la variance :

Soit X une variable aléatoire et soit $c \in \mathbb{R}$

$$\mathrm{Var}(X+c) = \mathrm{Var}(X) \text{ et } \mathrm{Var}(cX) = c^2\mathrm{Var}(X)$$

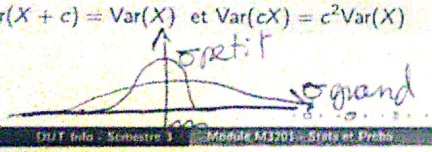


Image d'une loi uniforme sur [0, 1]

Soit U une v.a. de loi uniforme sur $[0,1]$
Soit $F : \mathbb{R} \to [0,1]$ une fonction de répartition (càd une fonction croissante avec limites 0 en $-\infty$ et 1 en $+\infty$) t.q. il existe une fonction $G : [0,1] \to \mathbb{R}$ vérifiant $\forall x \in [0,1], \forall y \in \mathbb{R},\ G(x) \leq y \Leftrightarrow x \leq F(y)$

Recette

- on pose $X = G(U)$
- alors pour tout $y \in \mathbb{R}$, $\mathbb{P}(X \leq y) = F(y)$, autrement dit
  X *a pour fonction de répartition* F

En effet, $\mathbb{P}(X \leq y) = \mathbb{P}(G(U) \leq y) = \mathbb{P}(U \leq F(y)) = F(y)$

Bernoulli à partir d'une loi uniforme sur [0, 1]

Soit U une v.a. de loi uniforme sur $[0,1]$
Soit p un nombre réel entre 0 et 1

Recette

- on pose $B = 1\{U \leq p\}$, càd
  $$B = 1 \text{ si } U \leq p \text{ ; } B = 0 \text{ sinon}$$
- alors B suit une *loi de Bernoulli de paramètre* p, càd
  $$\mathbb{P}(B=1) = p \text{ et } \mathbb{P}(B=0) = 1-p$$

$\mathbb{E}(B) = p$

Principe de simulation à partir d'une loi uniforme sur [0, 1]

- *loi normale* $N(0,1)$: il suffit d'inverser (numériquement) la fonction
  $$x \in \mathbb{R} \mapsto \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt$$
- *loi exponentielle de paramètre* λ : il suffit d'inverser la fonction
  $$x \in \mathbb{R}^+ \mapsto 1 - e^{-\lambda x}$$
- *loi de khi-deux* : il suffit d'inverser (numériquement) la fonction
  $$x \in \mathbb{R}^+ \mapsto \int_0^x c_n\, t^{\frac{n}{2}-1} e^{-\frac{t}{2}}\, dt$$

Géométrique à partir d'une loi uniforme sur [0, 1]

Soit $U_1, U_2, \ldots$ une suite de v.a. i.i.d. uniformes sur $[0,1]$
Soit p est un nombre réel entre 0 et 1

Recette

- on pose Ge = premier indice k tel que $U_k \leq p$, càd

$$\{Ge = k\} \Leftrightarrow \{U_1 > p, U_2 > p, \ldots, U_{k-1} > p \text{ et } U_k \leq p\}$$

- alors Ge suit une *loi géométrique de paramètre p*, càd

$$\mathbb{P}(Ge = k) = p(1-p)^{k-1}, \ \forall k \in \mathbb{N}^*$$

Récap'

- lois uniformes sur $[0,1]$ et sur $[a,b]$
- lois normales $N(0,1)$ et $N(m,\sigma^2)$
- simulations
  - principe à partir d'une loi uniforme sur $[0,1]$
  - lois classiques déjà programmées

Bien d'autres recettes de simulations encore...

Binomiale à partir de lois de Bernoulli :

Recette

- on pose $Z = B_1 + B_2 + \cdots + B_n$ où $(B_1, \ldots, B_n)$ est une famille i.i.d. de Bernoulli de paramètre p
- alors Z suit une *loi binomiale de paramètres* (n,p)

Khi-deux à partir de lois normales $N(0,1)$:

Recette

- on pose $K = (N_1)^2 + (N_2)^2 + \cdots + (N_k)^2$ où $(N_1, \ldots, N_k)$ est une famille i.i.d. de lois normales $N(0,1)$
- alors K suit une *loi de khi-deux à k degrés de liberté*

Binomiale
On répète n fois une exp
à deux issues
On compte le nb de fois qu
un event s'est produit

Deux exemples de séries entières

1) Soit G la fonction $G : x \in]-1,+1[\mapsto G(x) = \sum_{n\in\mathbb{N}} x^n$.

On sait (calculs...) que, pour tout $x \in]-1,+1[$,

$$G(x) = \sum_{n\in\mathbb{N}} x^n = \frac{1}{1-x}$$

et on sait aussi calculer la dérivée $x \mapsto G'(x) = \sum_{n\in\mathbb{N}^*} nx^{n-1} = \frac{1}{(1-x)^2}$

2) Soit H la fonction: $H : x \in \mathbb{R} \mapsto H(x) = \sum_{n\in\mathbb{N}} \frac{x^n}{n!}$.

Il reste à calculer combien vaut cette somme et à dériver H...

Retour sur la loi de Poisson

Rappel: La loi de Poisson de paramètre λ ($\lambda > 0$) est définie par

$$\mathbb{P}(X = n) = e^{-\lambda}\frac{\lambda^n}{n!}\ ,\ \forall n \in \mathbb{N}$$

On sait maintenant que $\sum_{n\in\mathbb{N}} e^{-\lambda}\frac{\lambda^n}{n!} = e^{-\lambda}H(\lambda) = e^{-\lambda}e^{\lambda} = 1$ (ouf!)

On est aussi capable de calculer l'espérance de X:

$$\begin{aligned}\mathbb{E}(X) &= \sum_{n\in\mathbb{N}} n\mathbb{P}(X=n) = \sum_{n\in\mathbb{N}} n\, e^{-\lambda}\frac{\lambda^n}{n!}\\ &= \dots\\ &= \dots\\ &= \lambda\end{aligned}$$

Application: calcul de la série $\sum_{n\in\mathbb{N}} \frac{x^n}{n!}$

Soit H la fonction donnée par la série entière

$$H : x \in \mathbb{R} \mapsto H(x) = \sum_{n\in\mathbb{N}} \frac{x^n}{n!}$$

D'après le théorème sur les séries entières, H est dérivable et sa dérivée est donnée par

$$H' : x \in \mathbb{R} \mapsto H'(x) = \sum_{n\in\mathbb{N}^*} \frac{nx^{n-1}}{n!} = \sum_{n=1}^{\infty} \frac{x^{n-1}}{(n-1)!} = \sum_{n=0}^{\infty} \frac{x^n}{n!} = \sum_{n\in\mathbb{N}} \frac{x^n}{n!}$$

Ainsi $H(x) = H'(x)$, $\forall x \in \mathbb{R}$. Ceci implique que $H(x) = C\exp(x)$, $\forall x \in \mathbb{R}$, avec C une constante réelle. Comme de plus, $H(0) = 1$, on en déduit que

$$\forall x \in \mathbb{R},\ H(x) = \sum_{n\in\mathbb{N}} \frac{x^n}{n!} = \exp(x) = e^x$$

Retour sur la loi géométrique

Rappel: La loi géométrique de paramètre p ($0 < p < 1$) est définie par

$$\mathbb{P}(X = n) = p(1-p)^{n-1}\ ,\ \forall n \in \mathbb{N}^*$$

On sait maintenant que $\sum_{n\in\mathbb{N}^*} p(1-p)^{n-1} = p\,G(1-p) = 1$ (ouf!)

On est aussi capable de calculer l'espérance de X:

$$\mathbb{E}(X) = \sum_{n\in\mathbb{N}} n\,\mathbb{P}(X=n) = \sum_{n\in\mathbb{N}^*} np(1-p)^{n-1} = p\,G'(1-p)$$

Ainsi

$$\mathbb{E}(X) = p\frac{1}{(1-(1-p))^2} = \frac{1}{p}$$

Distribution

Soit X une variable aléatoire discrète à valeurs dans $\mathbb{N}$

Définition

On appelle distribution de X (ou aussi loi de X) la suite

$$(\mathbb{P}(X = k))_{k\in\mathbb{N}}$$

$\mathbb{P}(X = k)$

- $\approx$ fréquence d'apparition de k parmi les valeurs prises par X
- $\approx$ densité pour les variables continues
- $\approx$ fréquence pour les variables statistiques (hauteurs des rectangles dans les histogrammes)

Propriété fondamentale : $\sum_{k\in\mathbb{N}} \mathbb{P}(X = k) = 1$

Exemples de variables discrètes avec une infinité de valeurs

Exemple 1 : Soit X le nombre de personnes connectées à un serveur internet à un instant donné. On peut choisir comme modèle

$$\mathbb{P}(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}, \text{ pour tout } k \in \mathbb{N}$$

$X \sim$ loi de Poisson de paramètre λ ($\lambda > 0$ fixé)

Exemple 2 : On tire des objets avec remise dans une population comportant une proportion p d'un type particulier,
X : rang du premier objet du type considéré obtenu

$$\mathbb{P}(X = k) = p(1-p)^{k-1}, \text{ pour tout } k \in \mathbb{N}^*$$

$X \sim$ loi géométrique de paramètre p

Au fait, "$\sum_{k\in\mathbb{N}} \mathbb{P}(X = k) = 1$" ça veut dire quoi ???

Définition

Soit $(a_n)_n$ une suite de nombres réels. On dit que la série de terme général (a_n) converge si la suite

$$(\sum_{k=0}^{n} a_k)_{n\in\mathbb{N}} \text{ admet une limite dans } \mathbb{R}$$

On note cette limite $\sum_{k\in\mathbb{N}} a_k$ ou encore $\sum_{n\in\mathbb{N}} a_n$

Remarque : si $a_k > 0$ pour tout entier k, alors la suite $(\sum_{k=0}^{n} a_k)_n$ est strictement croissante et donc, soit elle converge vers un nombre réel, soit elle tend vers $+\infty$.

Séries entières

Définition

On appelle série entière toute fonction de la forme $x \mapsto \sum_{n\in\mathbb{N}}(a_n x^n)$: série de terme général $a_n x^n$, où $(a_n)_n$ est une suite de nombres réels

Théorème

Soit $(a_n)_n$ est une suite de nombres réels et soit R un réel > 0.
Si la série $\sum_{n\in\mathbb{N}}(|a_n|R^n)$ est convergente ($< +\infty$)
alors la fonction $F : x \in [-R, +R] \mapsto F(x) = \sum_{n\in\mathbb{N}}(a_n x^n)$ est bien définie, elle est continue sur $[-R, +R]$ et est dérivable sur $]-R, +R[$ avec

$$\forall x \in]-R, +R[,\ F'(x) = \sum_{n\in\mathbb{N}^*}(n a_n x^{n-1}).$$

Densité

Soient X une v.a. (de loi) continue et F sa fonction de répartition.

Définition

Si il existe une fonction $f : \mathbb{R} \to \mathbb{R}^+$ telle que

$$\forall t \in \mathbb{R}, \quad \int_{-\infty}^{t} f(x)dx = F(t) = \mathbb{P}(X \leq t)$$

alors on appelle f la densité de X

Deux propriétés fondamentales : (faire des dessins!)

- "$f(x)dx$" représente la densité d'occupation de X,

$$"f(x)dx" = \mathbb{P}(x \leq X \leq x + dx)$$

- $\int_{-\infty}^{+\infty} f(x)dx = 1$

Retour aux exemples

Exemple 1 : une variable aléatoire de loi $N(0,1)$ a pour densité (définition)

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \ \forall x \in \mathbb{R}$$

Exemple 2 : (loi exponentielle) T admet la densité (on dérive F_T)

$$f(x) = 0 \text{ si } x < 0 \ ; \ \lambda e^{-\lambda x} \text{ si } x \in [0, +\infty[$$

Exemple 3 : (jeu de fléchettes) D admet la densité (on dérive F_D)

$$f(x) = 2x/R^2 \text{ si } x \in [0, R] \ ; \ 0 \text{ sinon}$$

Exemple 4 : une variable aléatoire de loi normale $N(m, \sigma^2)$ a pour densité à partir de la densité $N(0,1)$)

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}, \ \forall x \in \mathbb{R}$$

Règles de calcul

Soit X une v.a. de loi continue qui admet une densité. On note F sa fonction de répartition et f sa densité.

- $F(b) - F(a) = \mathbb{P}(a < X \leq b) = \int_a^b f(x)\,dx$;
- F est continue sur $\mathbb{R}$ et $\forall t \in \mathbb{R}$, $\mathbb{P}(X = t) = 0$
- conséquence: $\mathbb{P}(X \leq t) = \mathbb{P}(X < t)$ et $\mathbb{P}(X \geq t) = \mathbb{P}(X > t)$
- si F est dérivable sur I alors $\forall x \in I$, $F'(x) = f(x)$

Remarque : pour décrire la loi de X, on donne indifféremment sa fonction de répartition ou sa densité ou son nom en précisant la valeur des paramètres.

DUT Info - Semestre 3 Module M3201 - Proba-Stats

Variables aléatoires discrètes

Exemple : on lance 3 fois une pièce équilibrée et
X = nombre de fois où on a obtenu Pile

- ensemble des valeurs (possibles) de $X = \{0, 1, 2, 3\}$
- X est une variable discrète de loi binomiale $B(3, 1/2)$
- en fait la loi de X est donnée par $(\mathbb{P}(X = k))_{k \in \{0,1,2,3\}}$ (calcul...)

$$\mathbb{P}(X = 0) = \mathbb{P}(X = 3) = 1/8 \ ; \ \mathbb{P}(X = 1) = \mathbb{P}(X = 2) = 3/8$$

Remarque : $\mathbb{P}(X = 0) + \mathbb{P}(X = 1) + \mathbb{P}(X = 2) + \mathbb{P}(X = 3) = 1$

Variables aléatoires

C'est quoi une variable aléatoire?
↪ C'est une application X: expérience aléatoire $\mapsto$ nombre réel

Exemples :
- dans un parc informatique, X = la durée de vie d'une machine
- lors d'un match de rugby, X= nbre d'essais marqués
- dans un lancer de fléchette, X = distance au centre de la cible

Quel est le comportement de X ?
↪ Il est donné par la loi de X c'est-à-dire
"la probabilité que X prenne telle ou telle valeur"

Exemple 3 : jeu de fléchettes avec cible de rayon R, on suppose que le joueur atteint toujours la cible mais ne sait pas viser.

$$D = dist(\text{centre de la cible, point d'impact de la fléchette})$$

Alors pour tout $0 \leq r \leq R$,

$$\mathbb{P}(D \leq r) = \text{proba d'atteindre le "petit" disque de rayon } r$$
$$= k \times \pi r^2 \text{ avec } k \text{ une constante}$$

Comme $1 = \mathbb{P}(D \leq R) = k\pi R^2$, on déduit $k = 1/(\pi R^2)$ et donc

$$\mathbb{P}(D \leq r) = \frac{1}{\pi R^2} \times \pi r^2 = (r/R)^2$$

Variables aléatoires continues

Définition
Si l'ensemble {valeurs possibles de X} est fini ou $\subset \mathbb{N}$, on dit que (la loi de) X est discrète, sinon on dit que (la loi de) X est continue.

Exemple 1 : variable aléatoire X de loi $N(0,1)$

$$\forall t \in \mathbb{R}\ ,\ \mathbb{P}(X \leq t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx \ :\ \text{pas de calcul exact}$$

Exemple 2 : variable aléatoire T (p.ex. durée de vie d'atomes radio-actifs ou de composants électroniques) de loi exponentielle $Exp(\lambda)$

$$\forall t \in \mathbb{R}^+\ ,\ \mathbb{P}(T > t) = e^{-\lambda t} \ :\ \text{se calcule facilement}$$

Fonction de répartition

Définition
La fonction de répartition de la variable aléatoire X est la fonction

$$F : t \in \mathbb{R} \mapsto F(t) = \mathbb{P}(X \leq t) \in [0,1]$$

Remarques :
- $\mathbb{P}(X > t) = 1 - \mathbb{P}(X \leq t) = 1 - F(t)$
- la table de loi $N(0,1)$ donne $\mathbb{P}(X \leq t)$ pour certains $t \geq 0$
- d'autres tables donnent t tel que $\mathbb{P}(X > t) = \alpha$ pour certains α

attention ! bien lire les notices

Le fameux TCL (ou th. d'approximation gaussienne)

Situation : (X_n) suite de v.a. i.i.d. avec $\mathbb{E}(X_i) = m$ et $\mathrm{Var}(X_i) = \sigma^2$.
On note

$$S = X_1 + \cdots + X_n$$

Théorème de la limite centrale (TLC ou TCL)

$$\frac{S_n - nm}{\sigma\sqrt{n}} = \frac{(S_n/n) - m}{\sigma/\sqrt{n}} \underset{n\to+\infty}{\longrightarrow} \text{v.a. } N(0,1)$$

(convergence "en loi")

Règles d'approximation gaussienne

Si $(X_n)_n$ suite de v.a. i.i.d. avec $\mathbb{E}(X_i) = m$ et $\mathrm{Var}(X_i) = \sigma^2$ alors le TCL fournit les approximations en loi suivantes :

- pour $S = X_1 + X_2 + \cdots + X_n$

$$S \underset{n\to+\infty}{\approx} \text{v.a. de loi normale } N(nm, n\sigma^2)$$

- pour $\bar{X} = \frac{S}{n} = \frac{1}{n}(X_1 + \cdots + X_n)$

$$\bar{X} = S/n \underset{n\to+\infty}{\approx} \text{v.a. de loi normale } N(m, \sigma^2/n)$$

- pour $Z = \frac{S-nm}{\sigma\sqrt{n}} = \frac{(S/n)-m}{\sigma/\sqrt{n}}$

$$Z \underset{n\to+\infty}{\approx} \text{v.a. de loi normale } N(0,1)$$

Cas particulier : loi binomiale

Soit S une v.a. de loi binomiale $B(n,p)$. On écrit

$$S = B_1 + B_2 + \ldots + B_n$$

avec $B_1, \ldots, B_n$ v.a. i.i.d. Bernoulli tq $\mathbb{E}(B_i) = p$ et $\mathrm{Var}(B_i) = p(1-p)$.
Si n est "grand", le TCL s'applique et donne

Règle d'approximation gaussienne pour S de loi $B(n,p)$

Si n est grand (en pratique $n > 30$),

$$\text{loi de } S \underset{n\to+\infty}{\approx} N(np, np(1-p))$$

$$\text{loi de } S/n \underset{n\to+\infty}{\approx} N(p, p(1-p)/n)$$

Application : calculs "tractable" pour une variable binomiale $B(n,p)$ quand n est grand.

Approximation par la loi de Poisson quand p est très petit

Mise en garde : Si S une v.a. de loi binomiale $B(n,p)$ avec
- n grand
- p très petit

de telle sorte que np soit de l'ordre de quelques unités, alors l'approximation gaussienne n'est pas valable mais on a

Règle d'approximation poissonienne pour S de loi $B(n,p)$

Si n est grand ($n > 30$), p petit avec $np < 10$, alors

$$\text{loi de } S \underset{n\to+\infty}{\approx} \text{loi de Poisson de paramètre } np$$

autrement dit, $\forall k \in \mathbb{N}$, $\mathbb{P}(S = k) \approx e^{-np}\frac{(np)^k}{k!}$

...mateur

(X_n) suite de variables i.i.d. de loi quelconque avec paramètre θ inconnu
Rappel 1: i.i.d. ça veut dire indépendantes, identiquement distribuées
Rappel 2: un échantillon est une suite finie de variables i.i.d.

Exemples de lois avec un paramètre:
- loi de Bernoulli de paramètre $\theta = p = \mathbb{P}(X_1 = 1)$
- loi quelconque connue ou inconnue, $\theta = \mathbb{E}(X_1)$ ou $\theta = \mathrm{Var}(X_1)$
- loi exponentielle, $\theta =$ paramètre t.q. $\mathbb{P}(X_1 > t) = e^{-\theta t}$
- loi uniforme sur l'intervalle $[0, \theta]$

Définition
On appelle estimateur de θ toute fonction de la suite $(X_n)_n$ qui fournit une valeur approchée de θ

Moyenne empirique

Soit (X_n) une suite de variables i.i.d. de loi inconnue
paramètre inconnu à estimer: $\theta = m = \mathbb{E}(X_1)$

D'après la loi des grands nombres
avec probabilité 1,

$$\frac{1}{n}(X_1 + \cdots + X_n) \underset{n\to+\infty}{\longrightarrow} m$$

On note $\overline{X}(n) = \frac{1}{n}(X_1 + \cdots + X_n)$: moyenne empirique

La moyenne empirique observée sur l'échantillon nous informe sur la valeur du paramètre θ qui est inconnue.

Applications: tests de qualité, marketing, études économiques...

Variance empirique

(X_n) suite de variables i.i.d. de loi inconnue
paramètre inconnu à estimer: $\theta = \sigma^2 = \mathrm{Var}(X_1) = \mathbb{E}(X_1^2) - m^2$

D'après la loi des grands nombres
avec probabilité 1,

$$\frac{1}{n}(X_1^2 + \cdots + X_n^2) - (\bar{X}(n))^2 \underset{n\to+\infty}{\longrightarrow} \sigma^2$$

On note $\bar{V}(n) = \frac{1}{n}(X_1^2 + \cdots + X_n^2) - (\bar{X}(n))^2$: variance empirique

La variance empirique observée sur l'échantillon nous informe sur la valeur de σ^2 qui est inconnue.

Estimateur consistent

Définition
On dit d'un estimateur $F((X_n)_n)$ de θ qu'il est consistent si $\lim_{n\to+\infty} F((X_n)_n) = \theta$

Exemples:
- la moyenne empirique est un estimateur consistent de l'espérance
- la variance empirique est un estimateur consistent de la variance
- $max(X_1, \ldots, X_n)$ est un estimateur consistent de θ pour une loi uniforme sur $[0, \theta]$

Estimateur sans biais

Définition

On dit d'un estimateur $F((X_n)_n)$ de θ qu'il est sans biais si $\mathbb{E}F((X_n)) = \theta$

Exemples fondamentaux :

- la moyenne empirique est toujours un estimateur sans biais de l'espérance car ...
- la variance empirique n'est pas un estimateur sans biais de la variance car ...
- $\frac{n}{n-1}\bar{V}(n)$ est un estimateur sans biais de la variance car ...

Un premier exemple

Sondage effectué sur $n = 1000$ personnes:
600 sont favorables à l'introduction d'ours dans les Pyrénées

La question

Que peut-on dire de
p:= % de personnes favorables dans la population totale ?

↪ on modélise les réponses par un échantillon $X_1, ..., X_n$ de Bernoulli $B(p)$, avec p paramètre inconnu à estimer

$$\hat{p} := \frac{1}{n}(X_1 + \cdots + X_n) \approx p$$

Comme $\hat{p}$ est un bon estimateur de p (il est consistent et sans biais), une estimation ponctuelle de p est donnée par la valeur de $\hat{p}$, ici 60%

Un deuxième exemple

Contrôle fiscal chez Bemazone:
sur $n = 100$ jours, le chiffre d'affaire quotidien vaut en moyenne 50 kE

La question

Que peut-on dire de
m:= chiffre d'affaire quotidien moyen de Bemazone ?

↪ on modélise les CA relevés par un échantillon $X_1, ..., X_n$ de loi inconnue avec espérance m paramètre inconnu à estimer

$$\hat{m} := \frac{1}{n}(X_1 + \cdots + X_n) \approx m$$

Comme $\hat{m}$ est un bon estimateur de m (il est consistent et sans biais), une estimation ponctuelle de m est donnée par la valeur de $\hat{m}$, ici 50 kE

Intervalle de confiance

(X_n) suite de variables i.i.d. de loi quelconque avec paramètre θ inconnu

La question

On veut maintenant fournir un intervalle I dépendant des données observées tel que, pour une confiance $1-\alpha$ fixée,

$$\mathbb{P}(\theta \in I) = 1 - \alpha$$

Remarques

- dans la formule ci-dessus, c'est I qui est aléatoire, pas θ
- généralement, on choisit une "grande" confiance, p.ex. $1 - \alpha = 95\%$
- l'intervalle I peut être choisi centré sur une estimation ponctuelle $\hat{\theta}$ de θ, de la forme $I = [\hat{\theta} - \varepsilon, \hat{\theta} + \varepsilon]$: intervalle de confiance bilatère
- ou choisi décentré de la forme $I = [\hat{\theta} - \varepsilon, +\infty[$ ou encore $I =]-\infty, \hat{\theta} + \varepsilon]$: intervalle de confiance unilatère

Espérance

Définition/Mode de calcul

Si X est une variable aléatoire à valeurs dans $\mathbb{N}$,

$$\mathbb{E}(X) = \sum_{n \in \mathbb{N}} n\,\mathbb{P}(X = n)$$

Si X est une variable aléatoire avec une densité de probabilité f,

$$\mathbb{E}(X) = \int_{\mathbb{R}} x\,f(x)\,dx$$

Dans tous les cas, $\mathbb{E}(X)$="valeur moyenne des valeurs prises par X"

Propriété : l'espérance est linéaire

$$\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y) \quad \text{et} \quad \mathbb{E}(\alpha X) = \alpha \mathbb{E}(X)$$

Inégalité de Markov : $\forall \alpha > 0 \quad P(|X - E(X)| \geq \alpha) \leq \frac{Var(X)}{\alpha^2}$

Variance

Moyenne des carrés - carré de la moyenne

Définition/Mode de calcul

$$\mathrm{Var}(X) = \mathbb{E}((X - \mathbb{E}(X))^2) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$$

$\mathrm{Var}(X)$ mesure la dispersion de X autour de sa valeur moyenne

Propriétés

- $\mathrm{Var}(X)$ est toujours ≥ 0 (=0 *ssi* X est constante)
- $\mathrm{Var}(\alpha X) = \alpha^2 \mathrm{Var}(X)$
- $\mathrm{Var}(X+Y) \neq \mathrm{Var}(X) + \mathrm{Var}(Y)$ en général, égalité si X et Y sont indépendantes

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

Espérance et variance des lois classiques

loi	espérance	variance
Bernoulli $B(p)$	p	$p(1-p)$
Binomiale $B(n,p)$	np	$np(1-p)$
Poisson $P(\lambda)$	λ	λ
Uniforme sur $[a,b]$	$(a+b)/2$	$(b-a)^2/12$
Normale $N(m,\sigma^2)$	m	σ^2
Exponentielle $Exp(\lambda)$	$1/\lambda$	$(1/\lambda)^2$

Echantillon

Définition

On appelle échantillon de taille n de la variable X toute famille $(X_1, X_2, \cdots, X_n)$ de variables aléatoires i.i.d. de même loi que X.

- ça veut dire quoi *i.i.d.* ?
  $\longrightarrow$ i.i.d. signifie "indépendantes et identiquement distribuées"
- ça veut dire quoi *indépendantes* ?
  $\longrightarrow$ $X_1, X_2, \ldots, X_n$ sont indépendantes si pour tous intervalles $I_1, I_2, \ldots, I_n$ dans $\mathbb{R}$,
  $$\mathbb{P}(X_1 \in I_1, \cdots, X_n \in I_n) = \mathbb{P}(X_1 \in I_1) \times \cdots \times \mathbb{P}(X_n \in I_n)$$
- dans quelle(s) situation(s) rencontre-t-on un *échantillon* ?
  $\longrightarrow$ à chaque fois que l'on répète une même expérience dans les mêmes conditions et de façon indépendante

Somme et moyenne empirique d'un échantillon

moyenne Empirique

Proposition

Si $(X_1, \dots, X_n)$ est un échantillon de la variable X alors

$$\mathbb{E}(X_1 + \cdots + X_n) = n\,\mathbb{E}(X) \quad \text{et} \quad \mathbb{E}\left(\frac{X_1 + \cdots + X_n}{n}\right) = \mathbb{E}(X)$$

$$\mathrm{Var}(X_1 + \cdots + X_n) = n\,\mathrm{Var}(X) \quad \text{et} \quad \mathrm{Var}\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{\mathrm{Var}(X)}{n}$$

Application : soit $(B_1, \dots, B_n)$ un échantillon de Bernoulli $B(p)$
$\Rightarrow$ $S = B_1 + \cdots + B_n$ est une v.a. de loi binomiale $B(n, p)$
$\Rightarrow$ $\mathbb{E}(S) = np$ et $\mathrm{Var}(S) = np(1-p)$
$\Rightarrow$ $\mathbb{E}(S/n) = p$ et $\mathrm{Var}(S/n) = p(1-p)/n$

Loi des grands nombres

(X_n) suite de variables i.i.d. telle que $\sigma^2 = \mathrm{Var}(X_i) < +\infty$
On note $m = \mathbb{E}(X_1)$ et $S_n = X_1 + \cdots + X_n$

Loi des grands nombres (théorème)

$$\frac{S_n}{n} = \frac{1}{n}(X_1 + \cdots + X_n) \underset{n \to +\infty}{\longrightarrow} m \quad \text{avec probabilité 1}$$

autrement dit "la moyenne empirique tend vers la moyenne théorique"

Ce résultat est FONDAMENTAL en statistiques !

Idée de la preuve: ...

Cas des variables de Bernoulli

(X_n) suite de variables de Bernoulli i.i.d. avec $p = \mathbb{P}(X_1 = 1)$

D'après la loi des grands nombres

$$\frac{1}{n}(X_1 + \cdots + X_n) \underset{n \to +\infty}{\longrightarrow} p \quad \text{avec probabilité 1}$$

autrement dit "la fréquence empirique tend vers la fréquence théorique"

Application: sondages d'opinion
La fréquence observée sur l'échantillon nous informe sur la valeur de la fréquence théorique qui est inconnue

Cas des variables de loi inconnue

(X_n) suite de variables i.i.d. de loi inconnue
Valeurs de $m = \mathbb{E}(X_1)$ et $\sigma^2 = \mathrm{Var}(X_1) = \mathbb{E}(X_1^2) - m^2$??

D'après la loi des grands nombres

avec probabilité 1,

$$\bar{X}(n) := \frac{1}{n}(X_1 + \cdots + X_n) \underset{n \to +\infty}{\longrightarrow} m$$

$$\bar{V}(n) := \frac{1}{n}(X_1^2 + \cdots + X_n^2) - (\bar{X}(n))^2 \underset{n \to +\infty}{\longrightarrow} \sigma^2$$

Application: tests de qualité, marketing, études économiques,...
La moyenne empirique et la variance empirique observées sur l'échantillon nous informent sur les valeurs de m et σ^2 qui sont inconnues