Probabilité et statistiques - Cours 2

2021

Plan

Paysage actuel : • Ω : l'univers, l'ensemble des tirages possibles.

• $P : \Omega \to [0,1]$: probabilité de chaque évènement.

• $X : \Omega \to \mathbb{R}$: variable aléatoire qui mesure une grandeur sur les tirages.

• Ce qui nous intéresse : $P(X \leq x)$ pour $x \in \mathbb{R}$.

- Cas discret -

$$P(X \leq x) = \sum_{\substack{\omega \text{ avec} \\ X(\omega) \leq x}} P(\omega)$$

- Cas continu - (avec densité...)

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

Ces valeurs là sont appelées la loi de X

Quelques lois classiques se retrouvent souvent dans les phénomènes observés.

La loi uniforme discrète

Paramètre : r, entier

Ω fini et chaque élément de Ω a même probabilité.

L'ensemble des valeurs possibles est $\{1, 2, \ldots, r\}$ et chaque valeur reçoit la même probabilité $\frac{1}{r}$.

Exemple

Pour le lancer de dés, on définit la variable aléatoire X donnant le nombre de points obtenus. Alors X suit une loi uniforme discrète de paramètre $r = 6$.

• <u>Loi uniforme discrète</u> : $\Omega = \{\omega_1, \dots, \omega_r\}$ (ou $\{1, \dots, r\}$).

$\forall i \quad P(\omega_i) = \frac{1}{r}$ et $r = |\Omega|$.

Pour A un sous-ensemble de Ω, la probabilité de A est :

$$P_r(A) = \sum_{\omega \in A} P(\omega) = \sum_{\omega \in A} \frac{1}{r} = |A| \cdot \frac{1}{r} = \frac{|A|}{|\Omega|}$$

$$\boxed{P_R(A) = \frac{\text{“nbre de cas favorables”}}{\text{“nbre de cas possibles”}}} \quad \begin{array}{l} (= |A|) \\ (= |\Omega|). \end{array}$$

<u>Ex</u> : Je tire une carte dans un jeu de 32 cartes

P_R (reine ou un pique) = ?

La loi uniforme discrète

Proposition

Pour une v.a. X suivant la loi uniforme discrète de paramètre r, on a :

Par: $\Omega = \{\omega_1, \dots, \omega_r\}$
$X : \omega_i \longrightarrow i$

$$E(X) = \frac{r+1}{2}$$

et

$$V(X) = \frac{r^2-1}{12}$$

Exemple

Pour le lancer de dés : $E(X) = \frac{7}{2}$ et $V(X) = \frac{35}{12}$.

La loi de Bernoulli $\mathcal{B}(p)$

Paramètre : p, réel entre 0 et 1

C'est la loi **la plus simple**. La variable aléatoire qui suit cette loi prend deux valeurs : 0 ou 1.

$$P(X = 1) = p \qquad P(X = 0) = 1 - p.$$

Formule générale : $P(X = x) = p^x(1-p)^{1-x}$

La loi de Bernoulli $\mathcal{B}(p)$

$X \rightsquigarrow \mathcal{B}(p)$ est utilisée comme fonction indicatrice d'un évènement donné au cours d'une expérience aléatoire.

Exemple

Être infecté au cours d'une épidémie. X prend la valeur 1 si cela se produit, 0 sinon.

La v.a. X sera alors une variable de comptage lors de la répétition de l'évènement.

Le processus de Bernoulli et la loi binomiale $\mathcal{B}(n, p)$

Le *processus de Bernoulli* consiste en la répétition de l'expérience aléatoire de Bernoulli, chaque répétition a une probabilité de succès p.

Les répétitions sont indépendantes.

Le processus est donc modélisé par une suite de v.a.i.i.d. $X_1, X_2, \dots, X_n$ chacune de loi $\mathcal{B}(p)$.

Le processus de Bernoulli et la loi binomiale $\mathcal{B}(n, p)$

La loi binomiale est la loi de la variable aléatoire X correspondant au *nombre de succès au cours de n répétitions* du processus.

Exemple

Sondage : en sélectionnant au hasard n individus dans une grande population, on peut estimer la proportion p d'individus ayant un caractère donné par le nombre d'individus d'intérêt parmi les n individus sélectionnés.

Le processus de Bernoulli et la loi binomiale $\mathcal{B}(n, p)$

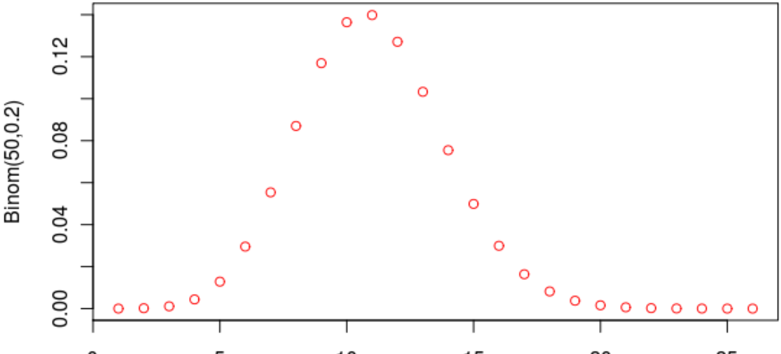
Paramètres : n et p

L'ensemble des valeurs possibles est $\{1, 2, \ldots, n\}$.

0,

$$P(X = x) = p(x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

Le processus de Bernoulli et la loi binomiale $\mathcal{B}(n, p)$



Le processus de Bernoulli et la loi binomiale $\mathcal{B}(n, p)$

Proposition

Soit $X_1, X_2, \ldots, X_n$ une suite de v.a.i.i.d. de loi $\mathcal{B}(p)$, alors $S_n = \sum_{i=1}^{n} X_i$ suit une loi $\mathcal{B}(n, p)$.

Remarque : Si $X \rightsquigarrow \mathcal{B}(n_1, p)$ et $Y \rightsquigarrow \mathcal{B}(n_2, p)$ alors $X + Y \rightsquigarrow \mathcal{B}(n_1 + n_2, p)$.

Proposition

Soit $X \rightsquigarrow \mathcal{B}(n, p)$ alors $E(X) = np$ et $V(X) = np(1 - p)$.

- Autre exemple d'une loi discrète : la loi géométrique -

$(X_i)_{i\geq 1}$ une suite de variables aléatoires indépendantes suivant une loi de Bernoulli de paramètre p.

Y la variable aléatoire qui donne l'indice du premier succès $(= \min\{i : X_i = 1\})$.

$Y : \Omega \to \mathbb{N}^*$.

$P(Y=k) = p\,(1-p)^k$

Théo : $\mathbb{E}(Y) = 1/p$

$\mathrm{Var}(Y) = \frac{1-p}{p^2}$

Ex : premier pile dans une suite de lancer de pièce.

Le processus et la loi de Poisson $\mathcal{P}(\lambda)$

On considère une série d'évènements donnés sur une échelle de temps.

Exemple

L'arrivée des appels téléphoniques sur un standard

Pour un temps $t > 0$ fixé, à partir d'une certaine origine des temps, on définit la variable aléatoire $X(t)$ comme le nombre d'occurrences de l'évènement dans l'intervalle $]0, t]$.

Par commodité, on note $p_k(t) = P(X(t) = k)$, pour $k \in \mathbb{N}$.

Le processus et la loi de Poisson $\mathcal{P}(\lambda)$

Definition

On dit qu'on a un *processus de Poisson* si :

1. il y a une invariance temporelle : $p_k(t)$ ne dépend pas de l'origine des temps, mais uniquement de la longueur t de l'intervalle, pour tout k et pour tout t;
2. il y a indépendance des nombres d'occurrences pour deux intervalles distincts;
3. pour un très petit intervalle, la probabilité d'avoir deux occurrences au moins est négligeable devant la probabilité d'avoir une occurrence exactement et celle-ci est proportionnelle à la longueur de l'intervalle.

Le processus et la loi de Poisson $\mathcal{P}(\lambda)$

Plus formellement :

$$p_1(h) = \lambda h + o(h)$$

et

$$\sum_{k=2}^{+\infty} p_k(h) = o(h)$$

où $o(h)$ désigne une fonction telle que $\lim\limits_{h \to 0} \dfrac{o(h)}{h} = 0$.

Sous ces hypothèses, on démontre que : $\forall k \in \mathbb{N}, p_k(t) = \dfrac{e^{-\lambda t}(\lambda t)^k}{k!}$

Le processus et la loi de Poisson $\mathcal{P}(\lambda)$

Definition

La loi de Poisson est la loi du nombre d'occurrences dans une unité de temps, donc pour $t = 1$.

La v.a. X suit donc une loi de Poisson $\mathcal{P}(\lambda)$ si sa fonction de probabilité est $p(k) = \frac{e^{-\lambda}\lambda^k}{k!}$ pour $k \in \mathbb{N}$.

Remarque : Sachant que $\displaystyle\sum_{k=0}^{+\infty} \frac{\lambda^k}{k!} = e^{\lambda}$, on a bien $\displaystyle\sum_{k=0}^{+\infty} p(k) = 1$.

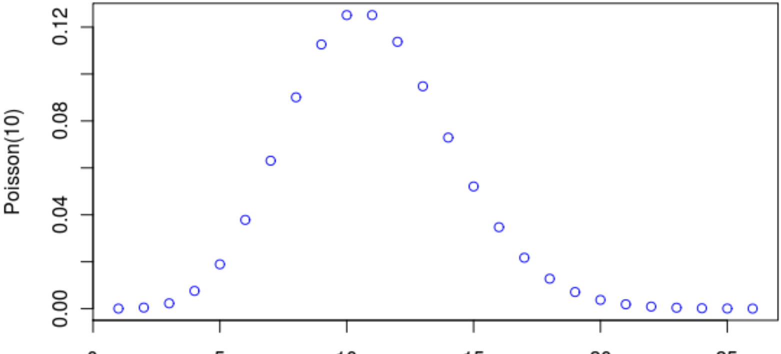
Le processus et la loi de Poisson $\mathcal{P}(\lambda)$

Proposition

Si $X \rightsquigarrow \mathcal{P}(\lambda)$ alors $E(X) = \lambda$ et $V(X) = \lambda$.

λ représente donc le nombre moyen d'occurrences dans une unité de temps pour le processus.

Le processus et la loi de Poisson $\mathcal{P}(\lambda)$



Le processus et la loi de Poisson $\mathcal{P}(\lambda)$

x

Remarques : La loi du temps entre deux occurrences successives est la loi exponentielle

Proposition (Propriété additive)

Si $X_1 \rightsquigarrow \mathcal{P}(\lambda_1)$ *et* $X_2 \rightsquigarrow \mathcal{P}(\lambda_2)$ *sont deux v.a. indépendantes, alors* $X_1 + X_2 \rightsquigarrow \mathcal{P}(\lambda_1 + \lambda_2)$.

Approximation de la loi binomiale par une loi de Poisson

Si on choisit une unité de temps suffisamment petite pour que la probabilité d'avoir plus d'une occurrence devienne négligeable, le processus de Poisson se rapproche d'un processus de Bernoulli par discrêtisation de l'écoulement continu en unités successives.

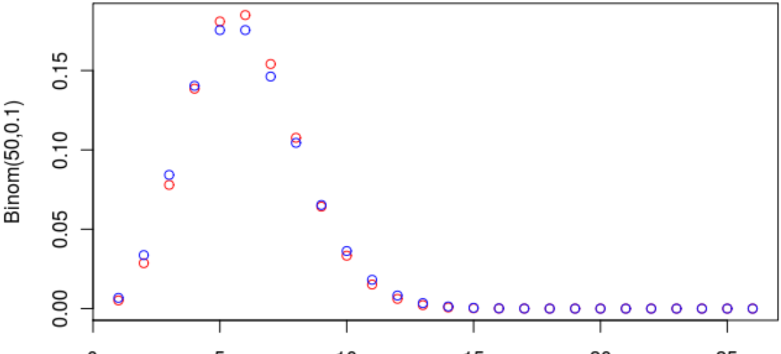
On montre que $\displaystyle\lim_{n \to +\infty, p \to 0, np \to \lambda} \mathcal{B}(n, p) = \mathcal{P}(\lambda)$.

Approximation de la loi binomiale par une loi de Poisson

Intérêt pratique : approcher la loi binomiale lorsque l'évènement "succès" est rare (p petit) avec un grand nombre de répétitions.

Si $n \geq 50$ et $p \leq 0.1$, l'approximation de $\mathcal{B}(n,p)$ par $\mathcal{P}(np)$ est satisfaisante.

Approximation de la loi binomiale par une loi de Poisson



Approximation de la loi binomiale par une loi de Poisson

Exemple : La probabilité pour qu'un réacteur d'avion d'un certain type connaisse une panne avant sa première révision est $1/1000$. Sachant qu'une companie d'aviation possède sur ses avions 100 réacteurs de ce type, calculons la probabilité qu'elle ne rencontre pas plus de 2 problèmes avec ces réacteurs avant la première révision.
Le nombre de réacteurs à problèmes est une v.a. X de loi $\mathcal{B}(100, 0.001)$ qui peut être approchée par une loi $\mathcal{P}(0.1)$. Donc :

$$P(X \leq 2) \simeq e^{-0.1} + \frac{e^{-0.1} \times 0.1}{1!} + \frac{e^{-0.1} \times (0.1)^2}{2!} \simeq 0.99985.$$

(0.9998496 donné par R)

La loi continue uniforme $\mathcal{U}([a,b])$

Definition

On dit que la v.a. X suit une loi uniforme sur l'intervalle $[a,b]$ si sa densité est constante sur $[a,b]$ et nulle en dehors.

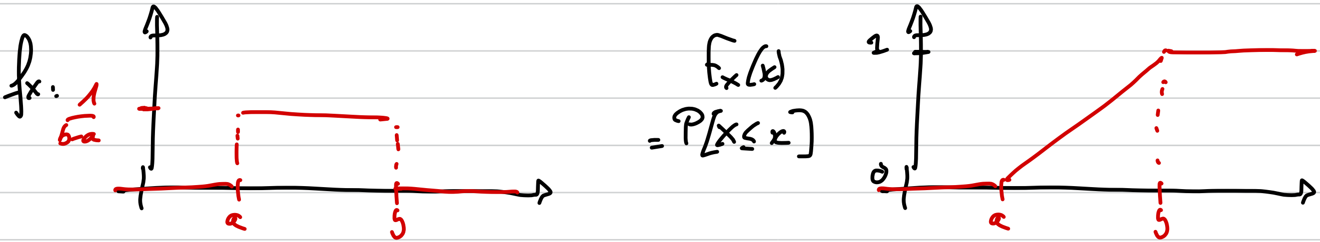
$$f(x) = \begin{cases} \frac{1}{b-a} & \text{si } a \leq x \leq b \\ 0 & \text{sinon.} \end{cases}$$

Fonction de répartition :

$$F(x) = \begin{cases} 0 & \text{si } x < a \\ \frac{x-a}{b-a} & \text{si } a \leq x \leq b \\ 1 & \text{sinon.} \end{cases}$$

Exemple d'utilisation :

$X : \Omega \longrightarrow [a,b]$ qui retourne un réel de $[a,b]$ choisi aléatoirement et uniformément.



$$F_X(x) = P(X \leq x) = \int_a^x \frac{dt}{b-a} \quad \left(= \int_a^x f_X(t)\,dt \right) = \frac{x-a}{b-a}$$

La loi continue uniforme $\mathcal{U}([a, b])$

Definition

On dit que la v.a. X suit une loi uniforme sur l'intervalle $[a, b]$ si sa densité est constante sur $[a, b]$ et nulle en dehors.

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{si } a \leq x \leq b \\ 0 & \text{sinon.} \end{cases}$$

Fonction de répartition :

$$F(x) = \begin{cases} 0 & \text{si } x < a \\ \frac{x-a}{b-a} & \text{si } a \leq x \leq b \\ 1 & \text{sinon.} \end{cases}$$

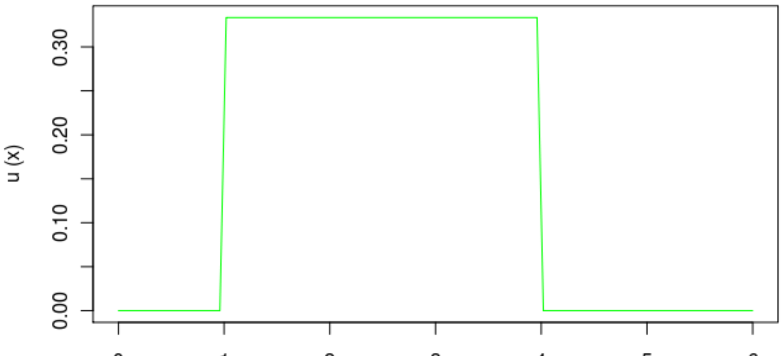
La loi continue uniforme $\mathcal{U}([a, b])$

Proposition

Soit $X \rightsquigarrow \mathcal{U}([a, b])$*. Alors* $E(X) = \frac{a+b}{2}$ *et* $V(X) = \frac{(b-a)^2}{12}$*.*

La loi $\mathcal{U}([0, 1])$ permet de générer des nombres au hasard. On peut simuler une loi quelconque à partir de ces nombres au hasard.

La loi continue uniforme $\mathcal{U}([a, b])$



La loi exponentielle $\mathcal{E}(\lambda)$

λ

La variable aléatoire X mesure le temps qui s'écoule entre deux occurrences successives d'un processus de Poisson. Elle suit alors une *loi exponentielle*.

La probabilité qu'il n'y ait aucune occurrence dans un intervalle de temps t est $p_0(t) = e^{-\lambda t}$.

D'où : $P(X > t) = e^{-\lambda t}$ et donc

$$P(X \leq t) = F(t) = \begin{cases} 1 - e^{-\lambda t} & \text{si } t \geq 0 \\ 0 & \text{si } t < 0 \end{cases}$$

La loi exponentielle $\mathcal{E}(\lambda)$

On obtient la fonction de densité par dérivation :

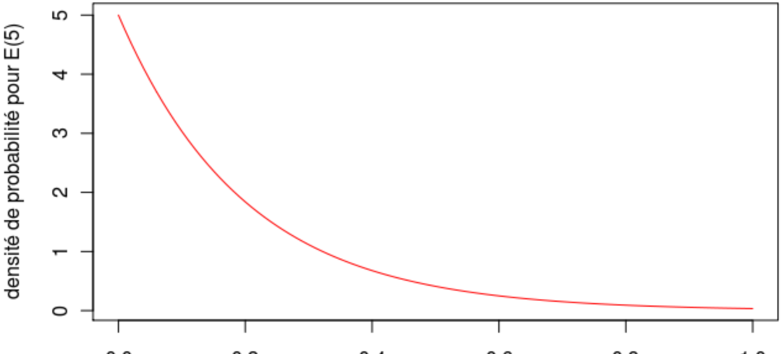
$$f(t) = \begin{cases} \lambda e^{-\lambda t} & \text{si } t \geq 0 \\ 0 & \text{si } t < 0 \end{cases}$$

Proposition

Si $X \rightsquigarrow \mathcal{E}(\lambda)$, alors $E(X) = \frac{1}{\lambda}$ et $V(X) = \frac{1}{\lambda^2}$.

Logiquement, puisque λ est le nombre moyen d'occurrences par unité de temps, $1/\lambda$ est la durée moyenne entre deux occurrences successives.

La loi exponentielle $\mathcal{E}(\lambda)$



La loi de Gauss ou loi normale $\mathcal{N}(\mu, \sigma^2)$

C'est la loi de probabilité fondamentale de la statistique, en raison du Théorème central limite.

La variable aléatoire X suit une loi de Gauss, ou loi normale, notée $\mathcal{N}(\mu, \sigma^2)$ si elle a pour densité définie sur $\mathbb{R}$:

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

$E(X) = \mu$

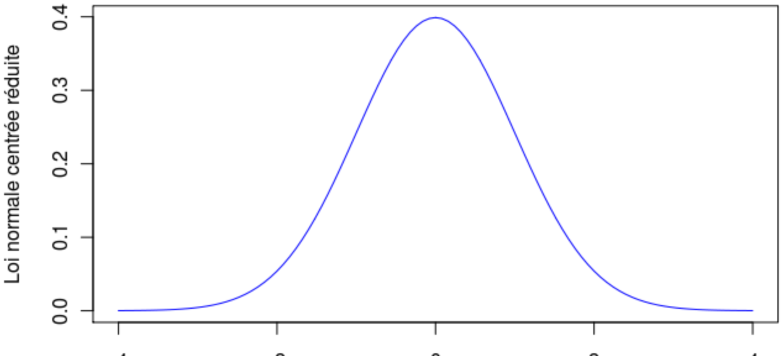
$VAR(X) = \sigma^2$

Pour $\mu = 0$ et $\sigma^2 = 1$, on parle de loi de Gauss centrée réduite $\mathcal{N}(0, 1)$, et sa fonction de répartition est

$$(F_X(x) =)\ \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{z^2}{2}} dz.$$

La loi de Gauss ou loi normale $\mathcal{N}(\mu, \sigma^2)$

loi classique de nombreux phénomènes naturels (relevé biologique, mesure sociale ...)



La loi de Gauss ou loi normale $\mathcal{N}(\mu, \sigma^2)$

Proposition

Si $X \rightsquigarrow \mathcal{N}(\mu, \sigma^2)$ alors $Z = \frac{X-\mu}{\sigma} \rightsquigarrow \mathcal{N}(0,1)$. Inversement, si $Z \rightsquigarrow \mathcal{N}(0,1)$, alors $X = \mu + \sigma Z \rightsquigarrow \mathcal{N}(\mu, \sigma^2)$.

On en déduit que toute fonction linéaire d'une variable aléatoire gaussienne est une v.a. gaussienne.

La loi de Gauss ou loi normale $\mathcal{N}(\mu, \sigma^2)$

Exemple : $X \rightsquigarrow \mathcal{N}(10, 4)$. Calculons $P(X \leq 13)$.

$$P(X \leq 13) = P\left(Z \leq \frac{13-10}{2}\right) = P(Z \leq 1.5) = \Phi(1.5) = 0.9332.$$

par lecture de la table de la loi normale centrée réduite.

Lecture inverse : on veut connaître le quantile d'ordre 0.95 de la loi de X. Pour Z on lit : 1.645.

$$P(Z \leq 1.645) = P\left(\frac{X-10}{2} \leq 1.645\right) = P(X \leq 10+1.645\times 2) = 0.95$$

le quantile est donc 13.29

La loi de Gauss ou loi normale $\mathcal{N}(\mu, \sigma^2)$

Proposition

Toute combinaison linéeaire de variables aléatoires gaussiennes indépendantes *est une variable aléatoire gaussienne.*

La loi de Gauss ou loi normale $\mathcal{N}(\mu, \sigma^2)$

Quelques valeurs clés de la loi de Gauss :

$$\begin{array}{ll} \Phi(0.975) = 1.96 & P(\mu - 1.96\sigma < X < \mu + 1.96\sigma) = 0.95 \\ \Phi(0.995) = 2.57 & P(\mu - 2.57\sigma < X < \mu + 2.57\sigma) = 0.99 \\ \Phi(0.9995) = 3.30 & P(\mu - 3.30\sigma < X < \mu + 3.30\sigma) = 0.999 \end{array}$$

A retenir : environ 95% des observations sont à plus ou moins deux écarts-types.

La loi de log-normale $\mathcal{LN}(\mu, \sigma^2)$

Elle représente un bon modèle pour des variables > 0, avec une distribution asymétrique avec allongement vers les valeurs élevées.
On dit que $X \rightsquigarrow \mathcal{LN}(\mu, \sigma^2)$ si $\ln X \rightsquigarrow \mathcal{N}(\mu, \sigma^2)$.

Fonction de densité :

$$f(x) = \begin{cases} \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(\ln x - \mu)^2\right) & \text{si } x > 0 \\ 0 & \text{sinon} \end{cases}$$

La loi de log-normale $\mathcal{LN}(\mu, \sigma^2)$

$\textcircled{x}$

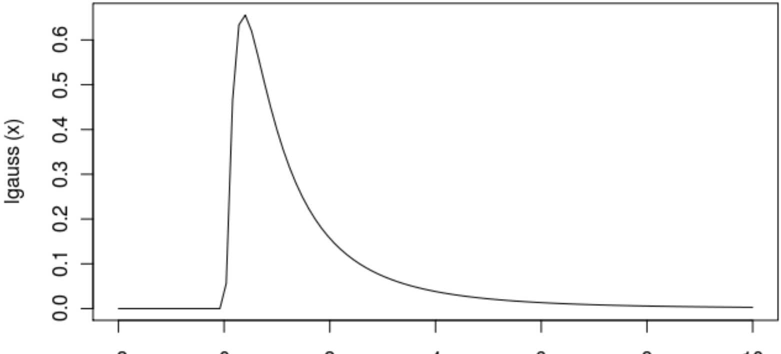
Comme $P(X \leq x) = P(\ln X \leq \ln x)$, les quantiles restent en correspondance.

Proposition

Si $X \rightsquigarrow \mathcal{LN}(\mu, \sigma^2)$*, alors* $E(X) = e^{\mu + \frac{\sigma^2}{2}}$ *et* $V(X) = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1)$.

La loi de log-normale $\mathcal{LN}(\mu, \sigma^2)$

X



Génération de nombres issus d'une loi

Il n'est pas toujours possible d'étudier de façon analytique le comportement de modèles, d'estimateurs ou de statistiques de tests en raison de leur complexité.

On a alors recours à de la simulation d'échantillons.

Génération de nombres issus d'une loi

Cas continu : F strictement croissante, on connaît F^{-1}.

Soit $U \rightsquigarrow \mathcal{U}([0,1])$. On considère $X = F^{-1}(U)$. On a :

$$P(X \leq x) = P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F(X)$$

puisque pour la loi uniforme sur $[0,1]$ on a $P(U \leq u) = u$. Donc X suit la loi de F.

À partir d'une suite de nombres au hasard $u_1, \ldots, u_n$, on peut obtenir une suite $x_1, \ldots, x_n$ issus de la loi F par la transformation $x_i = F^{-1}(u_i)$.

Génération de nombres issus d'une loi

Cas discrêt : F non inversible. Mais dans le cas où le nombre de valeurs possibles $a_1 < \ldots < a_r$ est restreint, on peut adapter la méthode :

Si $u_i \in [0, F(a_1)[$ alors générer a_1
Si $u_i \in [F(a_1), F(a_2)[$ alors générer a_2
$\ldots$
Si $u_i \in [F(a_{r-1}), 1]$ alors générer a_r

Plan

Phénomènes et échantillons aléatoires

On se penche sur l'étude d'observations répétées issues d'un certain phénomène de nature aléatoire.

Definition

On appelle *échantillon aléatoire de taille n*, ou $n-$échantillon, une suite de v.a.i.i.d.. Cette loi est appelée la *loi mère* de l'échantillon.

↳ variables aléatoires indépendantes identiquement distribuées (ie. même loi)

Phénomènes et échantillons aléatoires

La loi mère correspond à la distribution de la population.

Le statut de v.a.i.i.d. exige que le phénomène soit invariant au cours du temps et que ces observations n'exercent aucune influence entre elles. Il s'agit bien souvent d'une profession de foi...

On distingue la notion d'échantillon aléatoire $X_1, \ldots, X_n$ (résultats *a priori*) et d'échantillon réalisé $x_1, \ldots, x_n$ correspondant aux valeurs observées après expérience (résultats *a posteriori*).

Phénomènes et échantillons aléatoires

Definition (Statistique)

Soit $X_1, \ldots, X_n$ un $n-$échantillon. On appelle *statistique* sur cette échantillon toute variable aléatoire de la forme $T_n = h(X_1, \ldots, X_n)$.

On peut concrêtiser la loi d'une statistique en imaginant une simulation en très grand nombre d'échantillons de taille n, en calculant pour chacun d'entre eux la valeur prise par la statistique et en étudiant la distribution de ces valeurs (distribution d'échantillonnage de la statistique sur l'univers de tous les échantillons possibles).

Moyenne, variance, moments empiriques

Definition (Moyenne de l'échantillon)

On appelle *moyenne de l'échantillon*, ou *moyenne empirique*, la statistique notée $\overline{X}$, définie par :

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Moyenne, variance, moments empiriques

Proposition

Soient μ et σ^2 respectivement la moyenne théorique et la variance théorique de la loi mère. Alors on a :

$$E(\overline{X}) = \mu \qquad V(\overline{X}) = \frac{\sigma^2}{n}$$

Moyenne, variance, moments empiriques

tcd

Definition (Variance empirique)

On appelle *variance empirique*, la statistique notée $\tilde{S}^2$, définie par :

$$\tilde{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Proposition

Soient μ et σ^2 respectivement la moyenne théorique et la variance théorique de la loi mère. Alors on a :

$$E(\tilde{S}^2) = \frac{n-1}{n}\sigma^2.$$

Moyenne, variance, moments empiriques

+ tard

La statistique $\tilde{S}^2$ est un estimateur *biaisé* de σ^2.

Definition (Variance de l'échantillon)

On appelle *variance de l'échantillon* la statistique

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

La statistique S^2 est un estimateur sans biais de σ^2 car $E(S^2) = \sigma^2$.

Moyenne, variance, moments empiriques

Definition (Écart-type de l'échantillon)

$\sqrt{S^2}$

Definition (Écart-type empirique)

$\sqrt{\tilde{S}^2}$

Moyenne, variance, moments empiriques

X

Cas particulier de la loi Gaussienne :

Proposition

Si la loi mère est $\mathcal{N}(\mu, \sigma^2)$, alors

1. $\overline{X}$ *est gaussienne de loi* $\mathcal{N}(\mu, \frac{\sigma^2}{n})$,
2. $\overline{X}$ *et* S^2 *sont des variables aléatoires indépendantes.*

Moyenne, variance, moments empiriques

Definition (Moment empirique d'ordre r)

$$M_r = \frac{1}{n}\sum_{i=1}^{n} X_i^r$$

et on a $E(M_r) = \mu_r$

Definition (Moment empirique centré d'ordre r)

$$M'_r = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^r$$

Moyenne, variance, moments empiriques

x

Proposition (Formule de décentrage de la variance empirique)

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2$$

Moyenne, variance, moments empiriques

On considère un $n-$échantillon de couples de v.a. $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Definition (Moment croisé empirique d'ordre (p, q))

$$\frac{1}{n}\sum_{i=1}^{n} X_i^p Y_i^q$$

Definition (Moment croisé centré empirique d'ordre (p, q))

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^p(Y_i - \overline{Y})^q$$

Moyenne, variance, moments empiriques

X

Definition (Covariance empirique)

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$$

Remarque Comme pour la variance, on introduit le facteur $\frac{1}{n-1}$ au lieu de $\frac{1}{n}$ pour éliminer le biais vis-à-vis de la covariance théorique.

Moyenne, variance, moments empiriques

Definition (Corrélation linéaire empirique)

$$\frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

Loi du Chi-deux

+tard

Definition

Soit $Z_1, \ldots, Z_\nu$ une suite de v.a.i.i.d. de loi mère $\mathcal{N}(0,1)$. Alors la v.a. $T = \sum_{i=1}^{\nu} Z_i^2$ suit une loi appelée *loi du Chi-deux* à ν degrés de liberté, notée $\chi^2(\nu)$.

Proposition

$E(T) = \nu$ *et* $V(T) = 2\nu$.

Loi du Chi-deux

+ tard

Proposition

Si $T_1 \rightsquigarrow \chi^2(\nu_1)$ *et* $T_2 \rightsquigarrow \chi^2(\nu_2)$*, avec* T_1 *et* T_2 *indépendantes, alors* $T_1 + T_2 \rightsquigarrow \chi^2(\nu_1 + \nu_2)$*.*

Théorème

Soit un $n-$*échantillon* $X_1, \ldots, X_n$ *de loi* $\mathcal{N}(\mu, \sigma^2)$*. On a*

$$\frac{(n-1)S^2}{\sigma^2} \rightsquigarrow \chi^2(n-1).$$

On en déduit : $E(S^2) = \sigma^2$ et $V(S^2) = \frac{2\sigma^4}{n-1}$.

Loi de Student

+tard

Definition

Soient Z et Q deux variables aléatoires indépendantes de lois $Z \rightsquigarrow \mathcal{N}(0,1)$ et $Q \rightsquigarrow \chi^2(\nu)$. Alors la v.a. $T = \dfrac{Z}{\sqrt{\frac{Q}{\nu}}}$ suit une loi appelée loi de Student à ν degrés de liberté, notée $t(\nu)$.

Loi de Student

+ tard

Proposition

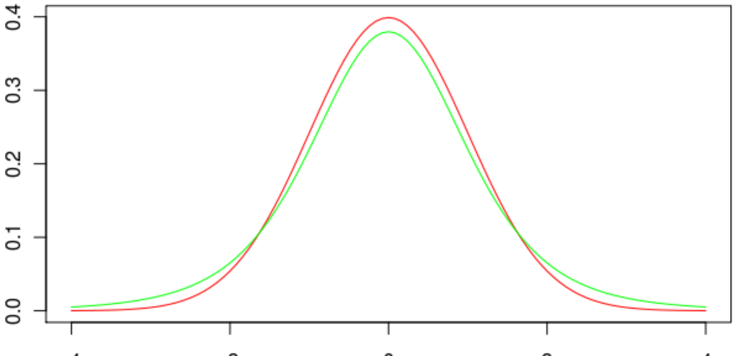
Soit $T \rightsquigarrow t(\nu)$. *Alors* $E(t) = 0$ *si* $\nu \geq 2$ *et* $V(T) = \frac{\nu}{\nu-2}$ *si* $\nu \geq 3$.

L'allure de la loi de student est similaire à une Gaussienne centrée réduite avec un étalement un peu plus fort. Cette différence s'estompe lorque ν augmente (et devient négligeable pour $\nu > 200$).

Pour $\nu = 1$, on a une loi de Cauchy.

Loi de Student

+tard



Loi de Student

+tard

Théorème (Application fondamentale)

Soit $X_1, \dots, X_n$ un $n-$échantillon de loi mère $\mathcal{N}(\mu, \sigma^2)$. Alors :

$$\frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \rightsquigarrow t(n-1).$$

On prend $Z = \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ et $Q = \frac{(n-1)S^2}{\sigma^2}$.

Rq : si on met σ à la place de S, on a $\mathcal{N}(0,1)$.

Loi de Student

+tard

Ce résultat met en évidence la modification apportée à la loi $\mathcal{N}(0, 1)$ de la v.a. Z lorsque l'on substitue à l'écart-type théorique σ de la loi mère, l'écart-type de l'échantillon, S.

On comprend au passage qu'en introduisant un terme aléatoire supplémentaire, on provoque un étalement plus grand.

Loi de Fisher-Snedecor

$\otimes$

Definition

Soit $U \rightsquigarrow \chi^2(\nu_1)$ et $V \rightsquigarrow \chi^2(\nu_2)$ deux v.a. indépendantes. Alors la v.a. $F = \dfrac{\frac{U}{\nu_1}}{\frac{V}{\nu_2}}$ suit une loi appelée loi de Fisher-Snedecor à ν_1 degrés de liberté au numérateur et ν_2 degrés de liberté au dénominateur, notée $F(\nu_1, \nu_2)$.

Proposition

Si $\nu_2 \geq 3$, la moyenne théorique est $\nu_2/(\nu_2 - 2)$ et si $\nu_2 \geq 5$ la variance théorique est $\frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)}$.

Loi de Fisher-Snedecor

$\star$

Proposition

Si $H \rightsquigarrow F(\nu_1, \nu_2)$ alors $\frac{1}{H} \rightsquigarrow F(\nu_2, \nu_1)$.

Application : la loi du rapport des variances $\frac{S_1^2}{S_2^2}$ de deux échantillons indépendants de tailles respectives n_1 et n_2 issus de deux lois mères gaussiennes de même variance σ^2. On a :

$$\frac{(n_1-1)S_1^2}{\sigma^2} \rightsquigarrow \chi^2(n_1-1) \qquad \frac{(n_2-1)S_2^2}{\sigma^2} \rightsquigarrow \chi^2(n_2-1)$$

d'où $\frac{S_1^2}{S_2^2} \rightsquigarrow F(n_1-1, n_2-1)$.

Les modes de convergence

On considère une suite infinie de v.a. $\{X_1, \ldots, X_n, \ldots\}$, notée $\{X_n\}$.

Definition (Convergence en loi)

La suite de v.a. $\{X_n\}$ converge en loi vers la v.a. X si on a, en toute x où sa fonction de répartition F_X est continue :

$$\lim_{n \to \infty} F_{X_n}(x) = F_X(x)$$

et on note $X_n \xrightarrow[n \to \infty]{\mathcal{L}} X$. La loi de X est appelée *loi limite ou asymptotique* de la suite $\{X_n\}$.

Celle qui va nous intéresser...

Les modes de convergence

Application : Calcul de la probabilité d'un évènement sur X_n lorsque n devient assez grand.

Definition (Convergence en probabilité)

La suite de v.a. $\{X_n\}$ converge en probabilité (ou faiblement) vers la v.a. X si on a

$$\forall \epsilon > 0, \ \lim_{n \to \infty} P(|X_n - X| < \epsilon) = 1$$

et on note $X_n \xrightarrow[n \to \infty]{p} X$.

Les modes de convergence

Definition (Convergence forte ou presque sûre)

La suite de v.a. $\{X_n\}$ converge presque sûrement (ou fortement, ou avec probabilité 1) vers la v.a. X si on a

$$\forall \epsilon > 0, \lim_{n\to\infty} P(\sup_{m\geq n}\{|X_m - X|\} < \epsilon) = 1$$

et on note $X_n \xrightarrow[n\to\infty]{ps} X$.

cv forte $\Rightarrow$ cv faible $\Rightarrow$ cv en loi.

Les modes de convergence

X

Proposition

Soient $\{X_n\}$ et $\{Y_n\}$ deux suites telles que $X_n \xrightarrow[n\to\infty]{ps} X$ et $Y_n \xrightarrow[n\to\infty]{ps} Y$.

1. *Si g est une fonction continue sur $\mathbb{R}$, alors $g(X_n) \xrightarrow[n\to\infty]{ps} g(X)$;*
2. *Si f est une fonction continue sur $\mathbb{R}^2$, alors $f(X_n, Y_n) \xrightarrow[n\to\infty]{ps} f(X, Y)$.*

Loi forte des grands nombres

Théorème (Loi forte des grands nombres)

Soit $\{X_n\}$ une suite de v.a.i.i.d. de moyenne théorique μ et de variance théorique σ^2. Alors la suite des moyennes empiriques $\{\overline{X_n}\}$ converge presque sûrement vers μ :

$$\overline{X_n} \xrightarrow[n\to\infty]{ps} \mu.$$

Cette loi garantit que la moyenne empirique se rapproche de plus en plus de la moyenne de la loi dont est issu l'échantillon quand on augmente la taille de l'échantillon. $\overline{X_n}$ peut ainsi prétendre à estimer μ.

Théorème central limite

Théorème (TCL)

Soit $\{X_n\}$ une suite de v.a.i.i.d. de moyenne théorique μ et de variance théorique σ^2. Alors la suite $\frac{\overline{X_n}-\mu}{\frac{\sigma}{\sqrt{n}}}$ converge en loi vers une v.a. normale centrée réduite :

$$\frac{\overline{X_n}-\mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

The théorème ...

Théorème central limite

Application fondamentale : Soit $\overline{X_n}$ la moyenne empirique d'un $n-$échantillon aléeatoire de loi mère quelconque, de moyenne théorique μ et de variance théorique σ^2. Alors, pour n assez grand, $\overline{X_n}$ suit approximativement une loi $\mathcal{N}(\mu, \frac{\sigma^2}{n})$.

Dans presque tous les cas, $n \geq 30$ suffit pour obtenir des approximations de probabilités à 10^{-2} près.

Application du Théorème Central Limite :

Dans une usine de bonbons, la contenance des sachets est de 100 bonbons en moyenne avec un écart type de 3.

Loi du nombre de bonbons par sachet : ?

On tire 30 sachets et on fait la moyenne du nbre de bonbon.

Quelle est la probabilité d'avoir entre 99 et 101 en moyenne ?

X_i : nbre de bonbons dans le sachet i.

$\bar{X} = \frac{1}{30} \sum_{i=1}^{30} X_i$

Par le TCL $\frac{\bar{X} - 100}{3/\sqrt{30}} \sim \mathcal{N}(0,1)$ environ.

$\rightsquigarrow \mathcal{N}(0,1)$. (approximativement)

$$P(\bar{X} < 99) = P\left(\frac{\bar{X}-100}{3/\sqrt{30}} < \frac{99-100}{3/\sqrt{30}}\right) = P\left(\frac{\bar{X}-100}{3/\sqrt{30}} < -1.82\right)$$

$$\simeq 0.035 \quad \text{(on lit dans les tables de } \mathcal{N}(0,1)\text{)}.$$

$$P(\bar{X} > 101) = P\left(\frac{\bar{X}-100}{3\sqrt{30}} > \frac{101-100}{3\sqrt{30}}\right) = P\left(\frac{\bar{X}-100}{3\sqrt{30}} > 1.82\right).$$

$$\simeq 0.035$$

Donc avec 93 % de chance, sur les 30 sachets il y a en moyenne entre 99 et 101 bonbons.

Cas particulier : processus de Bernoulli

Soit S_n le nombre total de succès au cours de n répétitions. Comme $E(S_n) = np$ et $V(S_n) = np(1-p)$, on a pour la fréquence relative $\frac{S_n}{n}$ une moyenne théorique p et une variance théorique $\frac{p(1-p)}{n}$.

Du TCL on déduit, que pour n suffisamment grand $\frac{S_n}{n}$ suit approximativement une loi normale $\mathcal{N}(p, \frac{p(1-p)}{n})$, ou encore, que S_n suit approximativement une loi normale $\mathcal{N}(np, np(1-p))$.

C'est l'approximation de la loi binomiale $\mathcal{B}(n, p)$ par la loi normale $\mathcal{N}(np, np(1-p))$.

Cas particulier : processus de Bernoulli

En pratique, cette approximation est satisfaisante pour $np \geq 5$ et $n(1-p) \geq 5$.

Comme on passe d'une loi discrête à une loi continue, on introduit une *correction de continuité* : $X \rightsquigarrow \mathcal{B}(n,p)$ alors
$P(X = k) \simeq P(k - \frac{1}{2} < U < k + \frac{1}{2})$ où $U \rightsquigarrow \mathcal{N}(np, np(1-p))$

Cas particulier : processus de Bernoulli

Exemple : Soit $X \rightsquigarrow \mathcal{B}(20, 0.3)$. On a $np = 6 \geq 5$ et $n(1-p) = 14 \geq 5$. Considérons $P(X = 8)$ et $P(X \leq 8)$:

$$\begin{array}{rcll} P(X=8) & \simeq & P(7.5 < U < 8.5) & \text{où } U \rightsquigarrow \mathcal{N}(6, 4.2) \\ & \simeq & P\left(\frac{7.5-6}{\sqrt{4.2}} < Z < \frac{8.5-6}{\sqrt{4.2}}\right) & \text{où } Z \rightsquigarrow \mathcal{N}(0,1) \\ & \simeq & P\left(0.73 < Z < 1.22\right) & = 0.8888 - 0.7673 = 0.1215 \end{array}$$

La valeur exacte dans la table est 0.1144. (des lois bin.)

$$P(X \leq 8) \simeq P(U < 8.5) \simeq P(Z < 1.22) = 0.8888.$$

La valeur exacte dans la table est 0.8866. (des lois bin.)