Fonctions de densité et de répartition : propriétés

- Propriétés de la **fonction de densité** :
  - $f_X(x) \geq 0$ pour tout $x \in \mathbb{R}$;
  - $\int_{-\infty}^{\infty} f_X(x)\mathrm{d}x = 1$.
- Si l'on pose $a = b$, on a
$$\Pr(X = a) = \int_a^a f_X(x)\mathrm{d}x = 0.$$
- La **fonction de répartition**, F_X, vérifie
$$F_X(a) = \Pr(X \leq a) = \Pr(X < a) = \int_{-\infty}^{a} f_X(x)\mathrm{d}x, \quad a \in \mathbb{R}.$$
- On a, pour tout $a, b \in \mathbb{R}$ tels que $a < b$,
$$\Pr(a < X \leq b) = F_X(b) - F_X(a) = \Pr(a < X < b).$$
- On a
$$f_X(x) = \frac{\mathrm{d}}{\mathrm{d}x} F_X(x) = F'_X(x), \quad x \in \mathbb{R}.$$

Quelques lois continues

- **Loi uniforme** : $X \sim U(a, b)$, pour $a < b$, de densité
$$f_X(x) = \begin{cases} 1/(b-a) & \text{si } a \leq x \leq b, \\ 0 & \text{sinon.} \end{cases}$$
- **Loi exponentielle** : $X \sim \exp(\lambda)$, pour $\lambda > 0$, de densité
$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{si } x \geq 0, \\ 0 & \text{sinon.} \end{cases}$$
- **Loi normale** : $X \sim \mathcal{N}(\mu, \sigma^2)$, pour $\mu \in \mathbb{R}, \sigma > 0$, de densité
$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}, \quad x \in \mathbb{R}.$$

  Si $X \sim \mathcal{N}(\mu, \sigma^2)$, alors $Z = (X - \mu)/\sigma \sim \mathcal{N}(0, 1)$ (“standardisation”).
  Notations : $f_Z(z) = \phi(z)$ et $F_Z(z) = \Phi(z)$.

Quelques lois continues

Exemple

Exemple Le M1 passe toutes les 5.5 minutes. Si j'arrive à un moment choisi au hasard, quelle est la probabilité que je doive attendre (a) plus de 3 minutes ? (b) moins de 2 minutes ? (c) entre 1 et 4 minutes ?

Exemple

Exemple La probabilité qu'il pleuve pendant la journée est de 0.2. S'il pleut, la quantité de pluie journalière suit une loi exponentielle de parametre $\lambda = 0.05$ mm^{-1}. Trouver (a) la probabilité qu'il tombe au plus 5mm demain, (b) la probabilité qu'il tombe au moins 2mm demain.

Exemples

Exemple La quantité annuelle de pluie dans une certaine région est une variable aléatoire normale de moyenne $\mu = 140$ cm et de variance $\sigma^2 = 16$ cm^2. Quelle est la probabilité qu'il tombe entre 135 et 150 cm ?

2.2.3 Variables aléatoires conjointes

Variables aléatoires conjointes / simultanées

Soient X et Y deux variables aléatoires définies sur le même ensemble Ω. La **fonction de répartition conjointe (ou simultanée)** de X et Y est définie par

$$F_{X,Y}(x,y) = \Pr(X \leq x,\ Y \leq y), \qquad x, y \in \mathbb{R}.$$

- **Cas discret** (i.e., X et Y sont discrètes) : la loi de probabilité conjointe de X et Y est parfaitement déterminée si l'on connaît leur **fonction de masse conjointe**, i.e.,
$$f_{X,Y}(x_i, y_j) = \Pr(X = x_i, Y = y_j)$$
pour tous les couples (x_i, y_j) possibles.
- **Cas continu** (i.e., X et Y sont continues) : la loi de probabilité conjointe de X et Y est parfaitement déterminée si l'on connaît leur **fonction de densité conjointe**, définie (si elle existe) par
$$f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y}, \qquad x, y \in \mathbb{R}.$$

Cas discret : propriétés

- Propriétés de la fonction de masse conjointe :
  - $0 \leq f_{X,Y}(x_i, y_j) \leq 1$, $i, j = 1, 2, \ldots$
  - $f_{X,Y}(x,y) = 0$, pour toutes les autres valeurs de x et y.
  - $\sum_{i,j} f_{X,Y}(x_i, y_j) = 1$.
- La fonction de répartition conjointe vérifie

$$F_{X,Y}(x,y) = \sum_{\{(i,j):\ x_i \leq x, y_j \leq y\}} f_{X,Y}(x_i, y_j), \quad x, y \in \mathbb{R}.$$

Cas continu : propriétés

- Propriétés de la densité conjointe :
  - $f_{X,Y}(x,y) \geq 0, \quad x, y \in \mathbb{R}$.
  - $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(u,v) \mathrm{d}v \mathrm{d}u = 1$.
- La fonction de répartition conjointe vérifie

$$F_{X,Y}(x,y) = \Pr(X \leq x, Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(u,v) \mathrm{d}v \mathrm{d}u, \quad x, y \in \mathbb{R}$$

- On a, pour tout $a_1, a_2, b_1, b_2 \in \mathbb{R}$ tels que $a_1 < b_1$ et $a_2 < b_2$,

$$\Pr(a_1 < X \leq b_1,\ a_2 < Y \leq b_2) = \int_{a_1}^{b_1} \int_{a_2}^{b_2} f_{X,Y}(u,v) \mathrm{d}v \mathrm{d}u.$$

Lois marginales

Définition: Soient X, Y deux variables aléatoires ayant pour densité (ou fonction de masse) conjointe $f_{X,Y}$. Les **densités marginales** du couple (X, Y) sont respectivement les densités de X et Y, i.e., f_X et f_Y. De même, les **fonctions de répartition marginales** du couple (X, Y) sont respectivement les fonctions de répartition de X et Y, i.e., F_X et F_Y.

Dans le cas des densités, on a

- **cas discret** : $f_X(x_i) = \sum_j f_{X,Y}(x_i, y_j), \quad f_Y(y_j) = \sum_i f_{X,Y}(x_i, y_j)$;
- **cas continu** : $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) \mathrm{d}y, \quad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) \mathrm{d}x$.

Concernant les fonctions de répartition, on a

- **cas discret** : $F_X(x) = \sum_{\{i: x_i \leq x\}} f_X(x_i), \quad F_Y(y) = \sum_{\{j: y_j \leq y\}} f_Y(y_j)$;
- **cas continu** : $F_X(x) = \int_{-\infty}^{x} f_X(u) \, \mathrm{d}u, \quad F_Y(y) = \int_{-\infty}^{y} f_Y(v) \, \mathrm{d}v$.

Exemple X, Y prennent les valeurs $(1,2), (1,4), (2,3), (3,2), (3,4)$ avec probabilités égales. Trouver les lois marginales de X et de Y.

Solution 113 et 115

Exemple X, Y prennent les valeurs $(1,2), (1,4), (2,3), (3,2), (3,4)$ avec probabilités égales. Trouver les lois marginales de X et de Y.

Indépendance

Définition: Deux variables aléatoires X et Y sont **indépendantes** si

$$\Pr(X \leq x,\ Y \leq y) = \Pr(X \leq x) \times \Pr(Y \leq y), \quad \forall x, y \in \mathbb{R}.$$

Dans ce cas on écrit $X \perp\!\!\!\perp Y$.

- Donc $X \perp\!\!\!\perp Y \iff \forall x, y \in \mathbb{R} : F_{X,Y}(x,y) = F_X(x)F_Y(y)$
- si $X \perp\!\!\!\perp Y$ et f_X, f_Y sont connues, on peut obtenir $f_{X,Y}$. Ceci est **faux** pour des variables dépendantes
- si $X \perp\!\!\!\perp Y$, alors $g(X) \perp\!\!\!\perp h(Y)$ pour toutes fonctions g, h 'raisonnables'
- Pour des variables aléatoires **discrètes**

  $\forall x, y \in \mathbb{R} : f_{X,Y}(x,y) = f_X(x) \times f_Y(y) \iff \forall x, y \in \mathbb{R} : F_{X,Y}(x,y) = F_X(x) \times F_Y(y)$

- Pour des variables aléatoires **continues** $\Longrightarrow$ est vrai et pour montrer une **dépendance** il suffit de trouver x, y auxquels $f_{X,Y}$, f_X et f_Y sont continues et $f_{X,Y}(x,y) \neq f_X(x) \times f_Y(y)$

Exemple Les variables aléatoires X, Y de l'exemple précédant sont-elles indépendantes ?

Cas continu

La fonction de répartition conjointe est

$$\Pr(X \leq x,\ Y \leq y) = F_{X,Y}(x,y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f_{X,Y}(u,v)\,\mathrm{d}u\,\mathrm{d}v.$$

Propriétés :

- $f_{X,Y}(x,y) \geq 0$ pour tout $(x,y) \in \mathbb{R}^2$
- $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(u,v)\,\mathrm{d}u\,\mathrm{d}v = 1$
- $f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x\, \partial y}$
- $\Pr(a_1 < X \leq b_1,\ a_2 < Y \leq b_2) = \int_{a_2}^{b_2} \int_{a_1}^{b_1} f_{X,Y}(u,v)\,\mathrm{d}u\,\mathrm{d}v$
- Plus généralement, pour $A \subseteq \mathbb{R}^2$ 'raisonnable'

$$\Pr((X,Y) \in A) = \int_A f_{X,Y}(u,v)\mathrm{d}u\mathrm{d}v$$

Exemple Soient $X \sim U[0,1]$ et $Y \sim U[0,2]$ indépendantes. Trouver $\Pr(X > Y)$.

Noter : $Y' = 2X \sim U[0,2]$ mais $\Pr(X > Y') = 0$; X et Y' sont dépendantes !

Solution 116

Densité conditionnelle

Définition: La **densité conditionnelle** de X sachant $Y = y$ (tel que $f_Y(y) > 0$) est définie par

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}, \qquad x \in \mathbb{R}.$$

Si X et Y sont indépendantes, on a

$$f_{X|Y}(x \mid y) = f_X(x), \quad f_{Y|X}(y \mid x) = f_Y(y), \quad \text{pour tout } x \text{ et } y \in \mathbb{R}.$$

(mathématiquement, c'est pour 'presque' tout x, y)

Exemple Soient X et Y de densité conjointe

$$f_{X,Y}(x,y) = \begin{cases} x + y & \text{si} \quad 0 < x < 1, 0 < y < 1, \\ 0 & \text{sinon.} \end{cases}$$

Trouver les densités marginales de X et Y, et la densité conditionnelle $f_{X|Y}$. Les deux variables sont-elles indépendantes ?

Solution Exemple 118

2.3 Valeurs caractéristiques

Mesure de tendance centrale

Définition: L'**espérance** d'une variable aléatoire X est

$$\mathbb{E}(X) = \begin{cases} \sum_i x_i f_X(x_i), & X \text{ discrète}, \\ \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x, & X \text{ continue}, \end{cases}$$

si la somme/intégrale converge

Propriétés :

- Interprétation 1 : espérance $\equiv$ centre de gravité d'un ensemble de masses
- Interprétation 2 : espérance $\equiv$ moyenne pondérée par des masses
- si $X_1, \dots, X_n$ sont des variables aléatoires et $a, b_1, \dots, b_n$ des constantes, alors
$$\mathbb{E}\left(a + \sum_{i=1}^n b_i X_i\right) = a + \sum_{i=1}^n b_i \mathbb{E}(X_i)$$
- pour g fonction 'raisonnable', $\mathbb{E}\{g(X)\} = \begin{cases} \sum_i g(x_i) f_X(x_i), & X \text{ discrète} \\ \int_{-\infty}^{\infty} g(x) f_X(x)\mathrm{d}x, & X \text{ continue} \end{cases}$
- si X, Y sont indépendantes et g, h des fonctions 'raisonnables', alors
$$\mathbb{E}\{g(X)h(Y)\} = \mathbb{E}\{g(X)\}\mathbb{E}\{h(Y)\}$$

Exemples

Exemple Pour $X \sim \mathcal{B}(m, p)$, trouver $\mathbb{E}(X)$.

Exemple Pour $X \sim \mathrm{Poiss}(\lambda)$, trouver $\mathbb{E}(X)$ et $\mathbb{E}\{X(X-1)\}$.

Exemples

Exemple Soit $X \sim \mathcal{N}(\mu, \sigma^2)$, trouver $\mathbb{E}(X)$.

Mesure de dispersion

Définition: La **variance** d'une variable aléatoire X est définie comme

$$\text{var}(X) = \mathbb{E}[\{X - \mathbb{E}(X)\}^2] = \cdots = \mathbb{E}(X^2) - \mathbb{E}(X)^2$$

Propriétés :

- Interprétation physique : variance $\equiv$ moment d'inertie relatif au centre de masse
- $\text{var}(X) \geq 0$, et $\text{var}(X) = 0$ implique que X est constante
- la **déviation standard** de X est définie comme $\mathrm{sd}(X) = \sqrt{\text{var}(X)} \geq 0$
- si a, b sont des constantes, alors $\text{var}(a + bX) = b^2 \text{var}(X)$
- si $X_1, \dots, X_n$ sont indépendants et $a, b_1, \dots, b_n$ des constantes, alors
$$\text{var}\left(a + \sum_{i=1}^n b_i X_i\right) = \sum_{i=1}^n b_i^2 \text{var}(X_i)$$

Exemple Si $X \sim \mathrm{Poiss}(\lambda)$, montrer que $\text{var}(X) = \lambda$.

Exemple Si $X \sim \mathcal{B}(m, p)$, montrer que $\text{var}(X) = m\,p(1-p)$.

Exemple Si $X \sim \mathcal{N}(\mu, \sigma^2)$, montrer que $\text{var}(X) = \sigma^2$.

Exemples : variance

Exemple Si $X \sim \mathrm{Poiss}(\lambda)$, montrer que $\text{var}(X) = \lambda$.

Exemple Si $X \sim \mathcal{B}(m, p)$, montrer que $\text{var}(X) = m\,p(1-p)$.

Exemple Si $X \sim \mathcal{N}(\mu, \sigma^2)$, montrer que $\text{var}(X) = \sigma^2$.

Covariance

Définition: La **covariance** des variables aléatoires X, Y est

$$\text{cov}(X, Y) = \mathbb{E}\left[\{X - \mathbb{E}(X)\}\{Y - \mathbb{E}(Y)\}\right] = \cdots = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

Interprétation : C'est une mesure de dépendance **linéaire** entre X et Y

Propriétés :

- la covariance dépend des unités dont on mesure X, Y
- $\text{cov}(X, Y) = \text{cov}(Y, X)$
- $\text{cov}(X, X) = \text{var}(X)$
- $\text{cov}(X+Y, Z+W) = \text{cov}(X, Z) + \text{cov}(Y, Z) + \text{cov}(X, W) + \text{cov}(Y, W)$
- si a, b, c, d sont des constantes, alors
$\text{cov}(a\,X + b,\, c\,Y + d) = ac\,\text{cov}(X, Y)$
- $\text{var}(X \pm Y) = \text{var}(X) + \text{var}(Y) \pm 2\text{cov}(X, Y)$
- si X et Y sont indépendantes, alors $\text{cov}(X, Y) = 0$. Mais attention, l'inverse **n'est pas vraie** en général !

Exemple

Exemple (voir diapositive 118) Soient X et Y de densité conjointe

$$f_{X,Y}(x,y) = \begin{cases} x+y & \text{si} \quad 0<x<1,\ 0<y<1, \\ 0 & \text{sinon.} \end{cases}$$

Trouver $\mathrm{Var}(X)$, $\mathrm{Var}(Y)$, et $\mathrm{Cov}(X,Y)$.

Corrélation

Définition: La **corrélation** de X et Y est

$$\rho_{X,Y} = \rho(X,Y) = \mathrm{corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)\mathrm{var}(Y)}}$$

(zéro si une des variances est zéro).

Propriétés :

- $\rho_{X,Y}$ mesure la dépendance **linéaire** (et seulement linéaire !) entre X et Y
- $\rho(a+bX, c+dY) = \mathrm{sign}(bd)\rho(X,Y)$
- $\mathrm{corr}(X,Y) = \mathrm{corr}(Y,X)$
- $\mathrm{corr}(X,X) = 1$ (si X n'est pas constante)
- $\mathrm{corr}(X,-X) = -1$ (si X n'est pas constante)
- $-1 \leq \mathrm{corr}(X,Y) \leq 1$ (inegalité de Cauchy–Schwarz)
- si X et Y sont indépendantes, alors $\mathrm{corr}(X,Y) = 0$, mais la réciproque est faux !
- corrélation $\neq$ causalité !

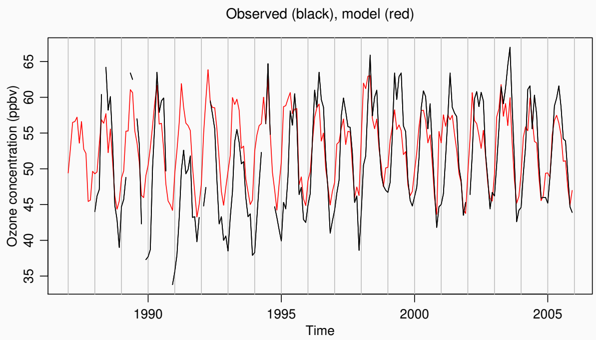
Corrélation empirique

Version empirique (si $\Pr((X = x_i, Y = y_i) = 1/n$ pour $i = 1, \dots, n$)

$$\frac{n^{-1}\sum_{j=1}^{n}(x_j-\bar{x})(y_j-\bar{y})}{\left\{n^{-1}\sum_{j=1}^{n}(x_j-\bar{x})^2 \times n^{-1}\sum_{j=1}^{n}(y_j-\bar{y})^2\right\}^{1/2}},$$

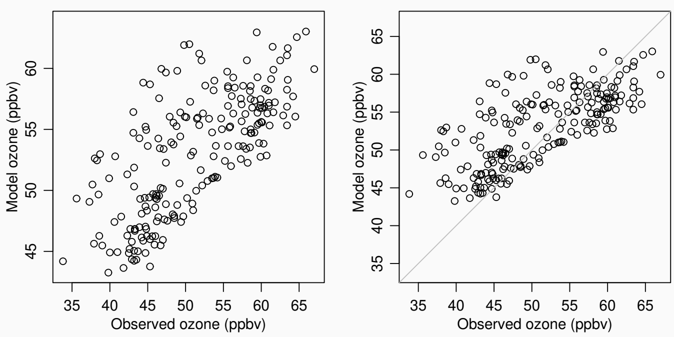
Exemple : ozone atmosphérique

Prof. Isabelle Bey (SIE) : observations de la concentration d'ozone au Jungfraujoch de janvier 1987 à décembre 2005 (quelques valeurs manquantes), et résultats d'une modélisation.



La modélisation vous paraît-elle bonne ?

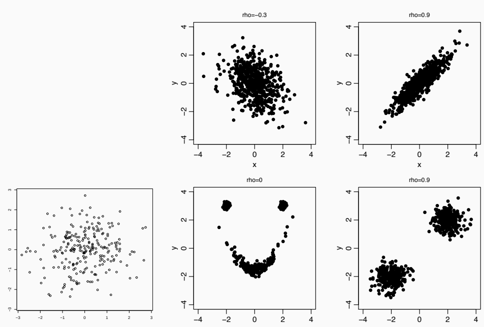
Exemple : ozone atmosphérique



La corrélation empirique est $\rho = 0.707$.

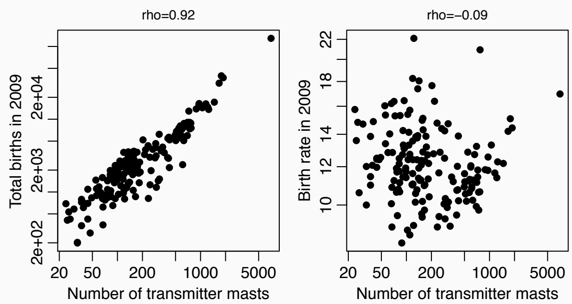
Limitations de la corrélation

- ρ mesure la dépendance linéaire (panneaux supérieurs)
- On peut avoir $\rho \approx 0$, mais dépendance forte mais non-linéaire (en bas au milieu)
- Une corrélation pourrait être forte mais **specieuse**, comme en bas à droite, ou deux sous-groupes, chacun sans corrélation, sont combinés
- Une corrélation entre deux variables n'implique pas une causalité entre elles



Corrélation $\neq$ causalité

Deux variables peuvent être très corrélées sans lien de causalité. Le graphique à gauche ici montre une corrélation forte entre le nombre de naissances et les mâts de communication dans les villes anglaises . . .



Danger !

- Les espérances/variances/covariances/corrélations ne sont pas définies si les intégrales/sommes ne convergent pas
- Ceci est notamment le cas lorsque la distribution de X a des **queues lourdes** : la densité de X décroît trop lentement vers zéro, et X a une probabilité élevée de prendre des valeurs énormes.

Exemple

- Considérons la fonction de densité $f(x) = \alpha x^{-1-\alpha}$ sur $[1, \infty)$ et $f(x) = 0$ pour $x < 1$ (loi **Pareto**). Pour $r \in \mathbb{R}$ on a

$$\mathbb{E}(X^r) = \alpha \int_1^\infty x^{r-1-\alpha} dx = \begin{cases} \frac{\alpha}{\alpha - r} & r < \alpha \\ \infty & r \geq \alpha \end{cases}$$

- En particulier, $\mathbb{E}(X) < \infty$ si et seulement si $\alpha > 1$, et $\mathrm{var}(X) < \infty$ si et seulement si $\alpha > 2$
- Pour α petit la densité tend lentement vers zéro

Espérance d'une variable aléatoire mixte

Théorème de l'espérance totale Pour une partition $A_1, A_2, \ldots$

$$\mathbb{E}(X) = \sum_i \mathbb{E}(X|A_i)\Pr(A_i)$$

Exemple : pluie (diapositive 107) La probabilité qu'il pleuve pendant la journée est 0.2. S'il pleut, la quantité de pluie qui tombe suit une loi exponentielle de parametre $\lambda = 0.05\text{mm}^{-1}$. Trouver l'espérance de la quantité de pluie journalière.

Quantiles

Définition: 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 des variables aléatoires continues, $F(x_p) = p$, donc x_p est tel que $\Pr(X \leq x_p) = p$
- Pour la plupart des variables aléatoires continues, ceci implique que $x_p = F^{-1}(p)$, où F^{-1} est la fonction inverse de F
- “La plupart" : celles ayant une fonction de densité strictement positive (sur $\{x : 0 < F(x) < 1\}$)
- Pour des variables aléatoires discrètes la situation est plus complexe
- Les quantiles empiriques (diapositive 32) sont des estimations (cf les prochains cours) des quantiles à partir des données à disposition.

En particulier, on appelle le 0.5ème quantile la **médiane** de F

Exemple quantiles

Exemple Calculer les quantiles des lois (a) $U(a, b)$, (b) Pareto (diapositive 134)

2.4 Théorèmes fondamentaux de probabilité

Approche expérimentale

- Considérons l'expérience de jeter une pièce de monnaie 10'000 fois et observons le nombre de "face" obtenues
- Soient $X_1, \dots, X_n$ les variables aléatoires indépendantes

$$X_i = \begin{cases} 1, & \text{si le } i\text{ème jet donne "face"}, \\ 0, & \text{si le } i\text{ème jet donne "pile"} \end{cases} \sim B(1, p)$$

- Donc $S_n = X_1 + \cdots + X_n$ représente le nombre de "face" sur n essais et

$$S_n \sim \mathcal{B}(n, p)$$

- La proportion de "Face" sur n jets est $\overline{X}_n := S_n/n$ et

$$\begin{aligned} \mathbb{E}(\overline{X}_n) &= n^{-1}\mathbb{E}(S_n) = n^{-1}\, np = p, \\ \mathrm{var}(\overline{X}_n) &= n^{-2}\mathrm{var}(S_n) = n^{-2}np(1-p) = p(1-p)/n \to 0 \end{aligned}$$

quand $n \to \infty$
- Donc $\overline{X}_n$ se concentre de plus en plus autour de p

Lois des grands nombres

Théorème (loi (faible) des grands nombres) Soient $X_1, X_2, \dots$ des variables aléatoires indépendantes et identiquement distribuées d'espérance $\mu = \mathbb{E}(X_1)$ et variance $\sigma^2 = \mathrm{var}(X_1)$ finies. Alors pour tout $\epsilon > 0$

$$\Pr(|\overline{X}_n - \mu| \geq \epsilon) \to 0, \qquad n \to \infty. \tag{1}$$

***Théorème (loi forte des grands nombres)** Soient $X_1, X_2, \dots$ des variables aléatoires indépendantes et identiquement distribuées d'espérance $\mu = \mathbb{E}(X_1)$ finie. Alors

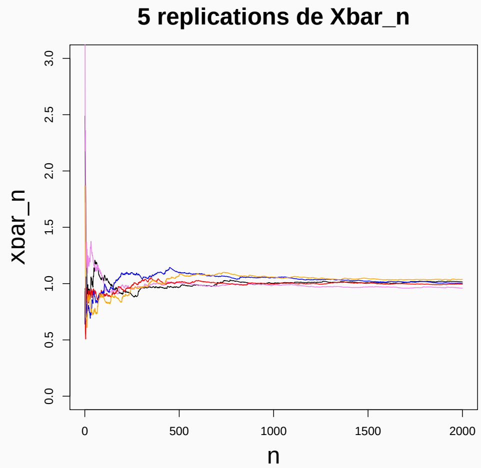
$$\Pr\left(\lim_{n\to\infty} \overline{X}_n = \mu\right) = 1 \tag{2}$$

*Il est donc certain que $\overline{X}_n$ soit proche de μ pour n grand

- *La loi forte est plus forte parce que (2) implique (1) et la variance peut être infinie
- *La loi faible utilise seulement $\mathrm{cov}(X_i, X_j) = 0$ pour $i \neq j$

Pour l'examen il suffit de connaître la loi faible

Illustration de la loi des grands nombres : exp(1)



Vitesse de convergence : Théorème central limite

- $\overline{X}_n \to \mu$ quand $n \to \infty$, mais à quelle vitesse ?
- Comme $\mathbb{E}(\overline{X}_n) = \mu$ et $\mathrm{var}(\overline{X}_n) = \sigma^2/n \in (0, \infty)$, pour tout n

$$Z_n := \frac{\overline{X}_n - \mu}{\sqrt{\sigma^2/n}} = \sqrt{n}\frac{\overline{X}_n - \mu}{\sigma}$$

a espérance 0 et variance 1, suggérant que la vitesse est $\sqrt{n}$

Théorème central limite Soient $X_1, X_2, \dots$ des variables aléatoires indépendantes et identiquement distribuées d'espérance μ et variance $\sigma^2 \in (0, \infty)$. Alors $Z_n := \sqrt{n}(\overline{X}_n - \mu)/\sigma$ satisfait

$$\Pr(Z_n \leq x) \to \Phi(x), \qquad x \in \mathbb{R}$$

La convergence étant uniforme en x, on déduit

$$\Pr(\overline{X}_n \leq x) = \Pr(Z_n \leq \sqrt{n}(x - \mu)/\sigma) \approx \Phi(\sqrt{n}(x - \mu)/\sigma)$$

donc $\overline{X}_n$ suit **approximativement** une loi $\mathcal{N}(\mu, \sigma^2/n)$

Illustration avec des variables exp(1)

On calcule $\sqrt{n}(\overline{X}_n - \mathbb{E}(X_1))/\sqrt{\mathrm{var}(X_1)}$, $R = 5000$ fois

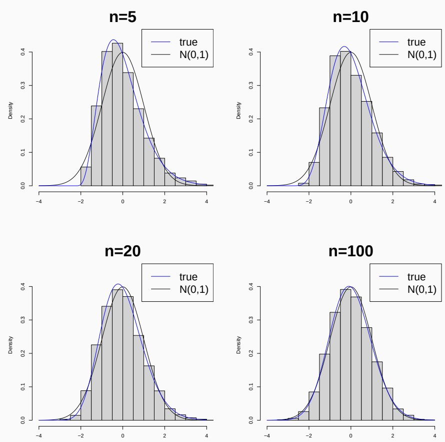


Illustration avec des variables $\exp(1)$

- On s'intéresse à la distribution de
$$\frac{\sqrt{n}(\overline{X}_n - \mathbb{E}(X_1))}{\sqrt{\text{var}(X_1)}}$$
- Fixons $n = 5$ ou 10 ou 20 ou 100 et $R = 5000$
- Générer $z_1^{(1)}, \dots, z_n^{(1)} \overset{iid}{\sim} \exp(1)$, et calculer leur moyenne $\overline{z}^{(1)}$
- Générer $z_1^{(2)}, \dots, z_n^{(2)} \overset{iid}{\sim} \exp(1)$, et calculer leur moyenne $\overline{z}^{(2)}$

$$\vdots$$

- Générer $z_1^{(R)}, \dots, z_n^{(R)} \overset{iid}{\sim} \exp(1)$, et calculer leur moyenne $\overline{z}^{(R)}$
- Les R valeurs
$$\left(\frac{\sqrt{n}(\overline{z}^{(1)} - \mathbb{E}(X_1))}{\sqrt{\text{var}(X_1)}}, \dots, \frac{\sqrt{n}(\overline{z}^{(R)} - \mathbb{E}(X_1))}{\sqrt{\text{var}(X_1)}}\right)$$
sont un échantillon issu de la distribution d'intérêt

Exemple

Exemple Soit $X \sim \mathcal{B}(m, p)$. Donner une approximation de $\Pr(X \leq r)$, pour $r \in \mathbb{R}$.

Solution Exemple 146 :
On a $X = \sum_{i=1}^m Y_i$, où $Y_1, \dots, Y_m \overset{\text{iid}}{\sim} \mathcal{B}(p)$. De plus, $\mathbb{E}(Y_1) = p$ et $\text{Var}(Y_1) = p(1-p)$. Le TCL nous donne donc que $X \overset{\text{app}}{\sim} \mathcal{N}(mp, mp(1-p))$ pour m grand. Ainsi, si Z désigne une variable aléatoire de loi $\mathcal{N}(0,1)$, on a, pour m grand,

$$\Pr(X \leq r) = \Pr\left(\frac{X - mp}{\sqrt{mp(1-p)}} \leq \frac{r - mp}{\sqrt{mp(1-p)}}\right)$$
$$\approx \Pr\left(Z \leq \frac{r - mp}{\sqrt{mp(1-p)}}\right) = \Phi\left(\frac{r - mp}{\sqrt{mp(1-p)}}\right).$$

Utilisation du théorème central limite

- Le théorème central limite est utilisé pour approximer des probabilités impliquant des sommes de variables aléatoires indépendantes
- Sous les conditions précédentes, on a
$$\mathbb{E}\left(\sum_{j=1}^n X_j\right) = n\mu, \quad \text{var}\left(\sum_{j=1}^n X_j\right) = n\sigma^2 \in (0, \infty)$$
- On **standardise** la somme
$$\frac{\sum_{j=1}^n X_j - n\mu}{\sqrt{n\sigma^2}} = \frac{n(\bar{X}_n - \mu)}{\sqrt{n\sigma^2}} = \frac{n^{1/2}(\bar{X}_n - \mu)}{\sigma} = Z_n$$
- Par le théorème central limite Z_n est approximativement $\mathcal{N}(0,1)$ et donc
$$\Pr\left(\sum_{j=1}^n X_j \leq r\right) = \Pr\left\{\frac{\sum_{j=1}^n X_j - n\mu}{\sqrt{n\sigma^2}} \leq \frac{r - n\mu}{(n\sigma^2)^{1/2}}\right\} \approx \Phi\left\{\frac{r - n\mu}{(n\sigma^2)^{1/2}}\right\}.$$

Exemple Un livre de 640 pages a un nombre aléatoire d'erreurs sur chaque page. Si le nombre d'erreurs par page suit une loi de Poisson d'espérance $\lambda = 0.1$, et est indépendant des autres pages, quelle est la probabilité que le livre contienne moins de 50 erreurs ?

Exemple : théorème central limite

Exemple 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 $\lambda = 0.1$, et est indépendant des autres pages, quelle est la probabilité que le livre contienne moins de 50 erreurs ?

Extensions et remarques

- Le théorème centrel limite est **remarquable**, car la distribution des X_i **n'a pas d'importance** : seulement l'espérance et la variance apparaissent. $\sum X_i$ a approximativement la même distribution si $X_i \sim Exp(1)$ ou $X_i \sim$Poiss(1)
- **Méthode delta** Si g est une fonction telle que $g'(\mu)$ existe, alors
$$\sqrt{n}\frac{g(\overline{X}_n) - g(\mu)}{\sigma} = g'(\mu)Z_n + o(Z_n)$$
suit approximativement $\mathcal{N}(0, [g'(\mu)]^2)$ et donc $g(\overline{X}_n) \overset{\text{app}}{\sim} \mathcal{N}\left\{g(\mu), g'(\mu)^2\sigma^2/n\right\}$
- Version **générale** de la méthode delta : si $\sqrt{n}(Y_n - \theta) \overset{\text{app}}{\sim} Y$ pour une constante $\theta \in \mathbb{R}$, alors $\sqrt{n}(g(Y_n) - g(\theta)) \overset{\text{app}}{\sim} g'(\mu)Y$
- **Notation :** $\overset{\text{app}}{\sim}$ indique une distribution **approximative**
- Le théorème centrel limite dépend d'un effect de moyennement, et échoue quand tout dépend d'une fraction minuscule des variables. Il n'est donc pas valable pour les maxima, les minima, l'étendue, ..., pour lequels on a d'autres théorèmes limites

3. Idées fondamentales de la statistique