Réseaux de neurones : rappels de probabilités et statistiques

Fabrice Rossi

http://apiacoa.org/contact.html.

Université Paris-IX Dauphine

Plan des rappels

1. Probabilités
2. Conditionnement et indépendance
3. Variable aléatoire
4. Moments
5. Retour sur le conditionnement
6. Notion d'estimateur
7. Théorèmes limites
8. Vraisemblance
9. Estimation bayesienne

Modélisation stochastique

- Phénomènes intrinsèquement aléatoires : mécanique quantique
- Phénomènes chaotiques :
  - rupture d'équilibre (stylo en équilibre sur sa pointe !)
  - phénomènes très complexes (météo par exemple)
- Connaissance incomplète, exemple :
  - lien entre z et x

$x = 1\ 2\ 2\ 1\ 4\ 5\ 3\ 1\ 2\ 2\ 5\ 4\ 2\ 1\ 5\ 3\ 5\ 2\ 5\ 2\ 5\ 4\ 3\ 5\ 5\ 2\ 3\ 2\ 3\ 2$

$z = 0\ 3\ 1\ 0\ 4\ 5\ 3\ 1\ 1\ 3\ 4\ 5\ 1\ 1\ 6\ 2\ 6\ 2\ 5\ 2\ 6\ 4\ 2\ 5\ 5\ 2\ 4\ 2\ 2\ 2$

  - $P(z = 3|x = 2) = \frac{1}{5}$, $P(z = 2|x = 2) = \frac{1}{2}$ et $P(z = 1|x = 2) = \frac{3}{10}$
  - en fait $z = x + y$ avec et

$y = -1\ 1\ -1\ -1\ 0\ 0\ 0\ 0\ -1\ 1\ -1\ 1\ -1\ 0\ 1\ -1\ 1\ 0\ 0\ 0\ 1\ 0\ -1\ 0\ 0\ 0\ 1\ 0\ -1\ 0$

Probabilités

Deux visions :

1. “objective” :
   - vision “fréquentiste”
   - calcul de fréquence par comptage
   - probabilité : limite des fréquences
   - exemple : probabilité qu’un accident soit une conséquence de la consommation d’alcool
2. “subjective” :
   - vision bayesienne
   - connaissance/croyance *a priori*
   - prise en compte de l’expérience pour modifier la croyance
   - exemple : probabilité qu’une pièce tombe sur pile

Probabilités (2)

Modèle probabiliste, $(\Omega, \mathcal{M}, P)$:

- Ω : l'univers, l'ensemble des possibles :
  - lancé d'**un** dé : $\Omega = \{1, 2, 3, 4, 5, 6\}$
  - lancé d'une pièce **deux** fois de suite : $\Omega = \{(P, P), (P, F), (F, P), (F, F)\}$
- $\mathcal{M}$: les évènements étudiables ($\mathcal{M} \subset \mathcal{P}(\Omega)$) :
  - conditions techniques : une σ-algèbre, i.e. :
    - contient Ω (évènement sûr) et $\emptyset$ (évènement impossible)
    - stable par passage au complémentaire (pour pouvoir dire que A n'a pas eu lieu)
    - stable par union dénombrable (pour pouvoir dire que A ou B ont eu lieu)
  - exemple : $\mathcal{M} = \mathcal{P}(\Omega)$

Probabilités (3)

Modèle probabiliste, $(\Omega, \mathcal{M}, P)$:

- P : une probabilité :
  - une fonction de $\mathcal{M}$ dans $[0, 1]$
  - donne la probabilité d'un évènement
  - $P(\Omega) = 1$
  - si les $(A_i)_{i \in \mathbb{N}}$ sont des évènements deux à deux disjoints

$$P\left(\bigcup_{i=0}^{\infty} A_i\right) = \sum_{i=0}^{\infty} P(A_i)$$

  - exemple pour dé non pipé, $P(\{i\}) = \frac{1}{6}$ pour tout i
- P définit une intégrale, et on a

$$P(B) = \int \mathbf{1}_B dP$$

Conditionnement

Un des concepts les plus importants en probabilité : qu'est-ce qu'apporte une information ?

- $P(A|B)$: la probabilité de l'évènement A sachant que l'évènement B a eu lieu ($P(B) > 0$)
- $$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(AB)}{P(B)}$$
- exemple :
  - $A = \{$le dé donne une valeur $> 4\}$ et $B = \{$le dé donne une valeur $>$ à $2\}$
  - $P(AB) = P(A)$ (car $A \subset B$) et donc $P(AB) = \frac{1}{3}$
  - $P(B) = \frac{2}{3}$
  - $P(A|B) = \frac{1}{2}$

Indépendance

Deux évènements sont indépendants si la connaissance de l'un n'apporte rien sur l'autre :

- A et B indépendants $\Leftrightarrow P(AB) = P(A)P(B)$
- si $P(B) > 0$, A et B indépendants $\Leftrightarrow P(A|B) = P(A)$

Hypothèse très utile en pratique. Par exemple :

- on lance n fois une pièce ($P(P) = P(F) = \frac{1}{2}$)
- univers $\Omega = \{P, F\}^n$
- lancés indépendants, implique $P(\{l_1, \ldots, l_n\}) = \frac{1}{2^n}$
- par indépendance, connaître P pour les évènements réduits à un lancé permet d'obtenir P sur $\mathcal{M}$ toute entière

Une application subtile

Le jeu des portes :

- le joueur est placé devant trois portes
- un cadeau est caché derrière une porte (il n'y a rien derrière les autres)
- le joueur choisit la porte i
- l'animateur lui indique que le cadeau n'est pas derrière la porte j ($j \neq i$!)

Question : le joueur a-t-il intérêt à changer de porte ?

Une application subtile

Le jeu des portes :

- le joueur est placé devant trois portes
- un cadeau est caché derrière une porte (il n'y a rien derrière les autres)
- le joueur choisit la porte i
- l'animateur lui indique que le cadeau n'est pas derrière la porte j ($j \neq i$!)

Question : le joueur a-t-il intérêt à changer de porte ?

Modélisation :

- Ω : position du trésor, choix du joueur, choix de l'animateur
- position du trésor et choix du joueur indépendants (et uniformes)
- choix de l'animateur uniforme (sachant la position du trésor et le choix du joueur)

Une application subtile (2)

Calculs :

- problème totalement symétrique
- $P(T = 1 | J = 1, A = 2)$?

Une application subtile (2)

Calculs :

- problème totalement symétrique
- $P(A = 2|T = 1, J = 1) = \frac{1}{2}$
- $P(T = 1, J = 1) = \frac{1}{3}\frac{1}{3}$
- $P(T = 1, J = 1, A = 2) = \frac{1}{18}$

Une application subtile (2)

Calculs :

- problème totalement symétrique
- $P(A = 2|T = 1, J = 1) = \frac{1}{2}$
- $P(T = 1, J = 1) = \frac{1}{3}\frac{1}{3}$
- $P(T = 1, J = 1, A = 2) = \frac{1}{18}$
- $P(J = 1, A = 2) = P(J = 1, A = 2, T = 1) + P(J = 1, A = 2, T = 3)$
- $P(A = 2|T = 3, J = 1) = 1$ donc $P(J = 1, A = 2, T = 3) = \frac{1}{9}$
- $P(J = 1, A = 2) = \frac{3}{18}$

Une application subtile (2)

Calculs :

- problème totalement symétrique
- $P(A = 2|T = 1, J = 1) = \frac{1}{2}$
- $P(T = 1, J = 1) = \frac{1}{3}\frac{1}{3}$
- $P(T = 1, J = 1, A = 2) = \frac{1}{18}$
- $P(J = 1, A = 2) = P(J = 1, A = 2, T = 1) + P(J = 1, A = 2, T = 3)$
- $P(A = 2|T = 3, J = 1) = 1$ donc $P(J = 1, A = 2, T = 3) = \frac{1}{9}$
- $P(J = 1, A = 2) = \frac{3}{18}$
- donc $P(T = 1|J = 1, A = 2) = \frac{1}{3}$

Une application subtile (2)

Calculs :

- problème totalement symétrique
- $P(A = 2|T = 1, J = 1) = \frac{1}{2}$
- $P(T = 1, J = 1) = \frac{1}{3}\frac{1}{3}$
- $P(T = 1, J = 1, A = 2) = \frac{1}{18}$
- $P(J = 1, A = 2) = P(J = 1, A = 2, T = 1) + P(J = 1, A = 2, T = 3)$
- $P(A = 2|T = 3, J = 1) = 1$ donc $P(J = 1, A = 2, T = 3) = \frac{1}{9}$
- $P(J = 1, A = 2) = \frac{3}{18}$
- donc $P(T = 1|J = 1, A = 2) = \frac{1}{3}$

Moralité : il faut changer de porte !

Variables aléatoires

On se donne D et $\mathcal{A}$ une σ-algèbre de D. Une variable aléatoire sur $(\Omega, \mathcal{M})$ est

- une fonction X de Ω dans D
- mesurable : pour tout $A \in \mathcal{A}$, $X^{-1}(A) \in \mathcal{M}$

Intuitivement, on veut pouvoir calculer la probabilité d'un évènement de la forme $X \in A$. Exemples de variables aléatoires pour Ω défini par le résultat du lancé de deux dés :

- S : la somme des deux dés
- M : la plus grande de deux valeurs obtenues

Si D est au plus dénombrable, on parle de variable aléatoire discrète.

Loi et fonction de répartition

- X définit une probabilité sur D, notée P_X grâce à :

$$P_X(A) = P(X^{-1}(A)) = P(X \in A)$$

P_X est la **loi** de X.

- Si X est à valeurs dans $\mathbb{R}$, on définit la **fonction de répartion** de X, F_X par

$$F_X(y) = P(X \leq y)$$

- une variable aléatoire X admet une **densité** quand il existe une fonction f_X de $\mathbb{R}$ dans $\mathbb{R}$ telle que pour tout $A \in \mathcal{A}$

$$P(X \in A) = \int_A f_X(x)dx$$

Exemples

- S = somme de deux dés :

	2	3	4	5	6	7	8	9	10	11	12
P_S	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$
F_S	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{6}{36}$	$\frac{10}{36}$	$\frac{15}{36}$	$\frac{21}{36}$	$\frac{26}{36}$	$\frac{30}{36}$	$\frac{33}{36}$	$\frac{35}{36}$	1

- M = max de deux dés :

	1	2	3	4	5	6
P_M	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{5}{36}$	$\frac{7}{36}$	$\frac{9}{36}$	$\frac{11}{36}$
F_M	$\frac{1}{36}$	$\frac{4}{36}$	$\frac{9}{36}$	$\frac{16}{36}$	$\frac{25}{36}$	1

- loi exponentielle : $f_{X_\lambda}(x) = \mathbf{1}_{\mathbb{R}^+}(x)\lambda e^{-\lambda x}$,
$F_{X_\lambda}(y) = \mathbf{1}_{\mathbb{R}^+}(y)\left(1 - e^{-\lambda y}\right)$

Espérance

Version mathématique de la notion de moyenne : l'**espérance**.

- Basée sur l'intégrale associée à une probabilité :

$$E(X) = \int X dP$$

- si X est une variable aléatoire positive, $E(X)$ est bien définie (mais on peut avoir $E(X) = \infty$)
- Propriété intéressante

$$E(f(X)) = \int f(X) dP = \int f dP_X$$

- cas discret : $E(X) = \sum_{k=0}^{\infty} \alpha_k P_X(\alpha_k)$

Exemples

- Valeur d'un dé : $E(V) = \frac{7}{2}$
- Somme de deux dés : $E(S) = 7$ (par linéarité de l'espérance !)
- Max de deux dés : $E(M) = \frac{161}{36} \simeq 4.47$
- Variable avec une densité

$$E(X) = \int x f_X(x) dx$$

- v.a. exponentielle : $E(X_\lambda) = \frac{1}{\lambda}$
- v.a. gaussienne (ou normale), notée $\mathcal{N}(\mu, \sigma)$:

$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ et $E(X) = \mu$

Variance

Idée : mesurer la dispersion d'une variable aléatoire autour de sa moyenne.

- Variance : $\sigma^2(X) = E\left((X - E(X))^2\right)$
- Ecart-type : $\sigma(X) = \sqrt{\sigma^2(X)}$ (!)
- $P(|X - E(X)| > \epsilon) \leq \frac{\sigma^2(X)}{\epsilon^2}$ (Bienaymé-Tchebychev)
- Exemples :
  - Variable avec une densité :

$$\sigma^2(X) = \int (x - E(X))^2 f_X(x) dx$$

  - v.a. exponentielle : $\sigma^2(X_\lambda) = \frac{1}{\lambda^2}$
  - v.a. gaussienne : $\sigma^2(X_\lambda) = \sigma^2$. Remarque : $P(|X - E(X)| \leq 1.96\sigma) \simeq 0.95$ (BT donne 0.74)

Corrélation

Idée : mesurer le “lien” entre deux variables aléatoires.

- Covariance : $\Gamma(X,Y) = E\left[(X - E(X))(Y - E(Y))\right]$
- si X et Y sont à valeurs dans $\mathbb{R}^n$, on définit une matrice de covariance :

$$\Gamma(X,Y) = \begin{pmatrix} \Gamma(X_1,Y_1) & \cdots & \Gamma(X_1,Y_n) \\ \vdots & \Gamma(X_i,Y_j) & \vdots \\ \Gamma(X_n,Y_1) & \cdots & \Gamma(X_n,Y_n) \end{pmatrix}$$

- Coefficient de corrélation : $\rho(X,Y) = \frac{\Gamma(X,Y)}{\sigma(X)\sigma(Y)}$
- $\rho(X,Y) = 0$: variables dites non corrélées
- X et Y indépendantes $\Rightarrow$ X et Y non corrélées (le contraire est faux !)
- $\sigma^2(X+Y) = \sigma^2(X) + \sigma^2(Y) + \Gamma(X,Y)$

Conditionnement

Idée : approcher Y par une fonction de X (i.e. “expliquer” Y grâce à X).

- Pour tout couple de v.a. X et Y, il existe une v.a., mesurable au sens de $\sigma(X)$, notée $E(Y|X)$ telle que pour tout B

$$E(Y\mathbf{1}_{X\in B}) = E\left[E(Y|X)\mathbf{1}_{X\in B}\right]$$

ou encore

$$\int_{X\in B} Y dP = \int_{X\in B} E(Y|X) dP$$

- en fait, $E(Y|X) = f(X)$ pour une certaine fonction f
- si Y possède une variance, $f(X)$ est la meilleure approximation de Y par une fonction de X au sens des moindres carrés

Conditionnement (2)

On note $E(Y|X=x) = f(x)$

- point de vue intuitif : $E(Y|X=x)$ associe à x la valeur moyenne des Y qu'on peut obtenir quand $X = x$
- si X est discrète, le point de vue intuitif est exact, i.e.

$$E(Y|X=x) = \frac{E(Y\mathbf{1}_{X=x})}{P(X=x)}$$

- exemple : S, somme de deux dés, X_1, valeur du premier dé.

x	1	2	3	4	5	6
$E(S\|X_1=x)$	$\frac{9}{2}$	$\frac{11}{2}$	$\frac{13}{2}$	$\frac{15}{2}$	$\frac{17}{2}$	$\frac{19}{2}$

S'obtient par indépendance : $E(S|X_1=x) = E(X_1+X_2|X_1=x) = E(x+X_2|X_1=x) = x + E(X_2)$

Conditionnement (3)

- si X et Y ont une densité jointe, i.e., il existe une fonction f_{XY} telle que

$$P\left((X,Y) \in V\right) = \int \int_V f_{XY}(x,y)dxdy$$

- alors $E(Y|X)$ admet pour densité

$$f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)}$$

- avec

$$f_X(x) = \int f_{XY}(x,y)dy$$

Conditionnement (4)

- conditionnement par un évènement

$$E(Y|B) = \frac{1}{P(B)} \int_B Y dP$$

- exemple : M, max des lancés de deux dés, X_1, valeur du premier dé : $E(M|X_1 \geq 3) = \frac{59}{12} \simeq 4.92$
- si $Y = f(X)$, $E(Y|X) = f(X)$
- si $Y = f(X) + \mathcal{E}$ et que $\mathcal{E}$ et X sont indépendants :

$$E(Y|X) = f(X) + E(\mathcal{E})$$

Statistiques

Principe général :

- on observe des variables aléatoires $X_1, \ldots, X_n$ de même loi P
- on cherche obtenir des informations sur la loi P à partir des observations
- par exemple, on fixe une famille de lois $P \in \{P_\theta | \theta \in \Theta\}$, et on cherche à trouver la valeur de θ telle que $P = P_\theta$
- plus généralement, on cherche à estimer $\alpha(\theta)$

Exemples :

- $X_i \sim \mathcal{N}(\mu, 1)$ et on cherche μ
- Modèle linéaire : $Y_i = \alpha x_i + \beta + \mathcal{E}_i$ avec $\mathcal{E}_i \sim \mathcal{N}(0, \sigma)$ et on cherche α et β

Estimateur

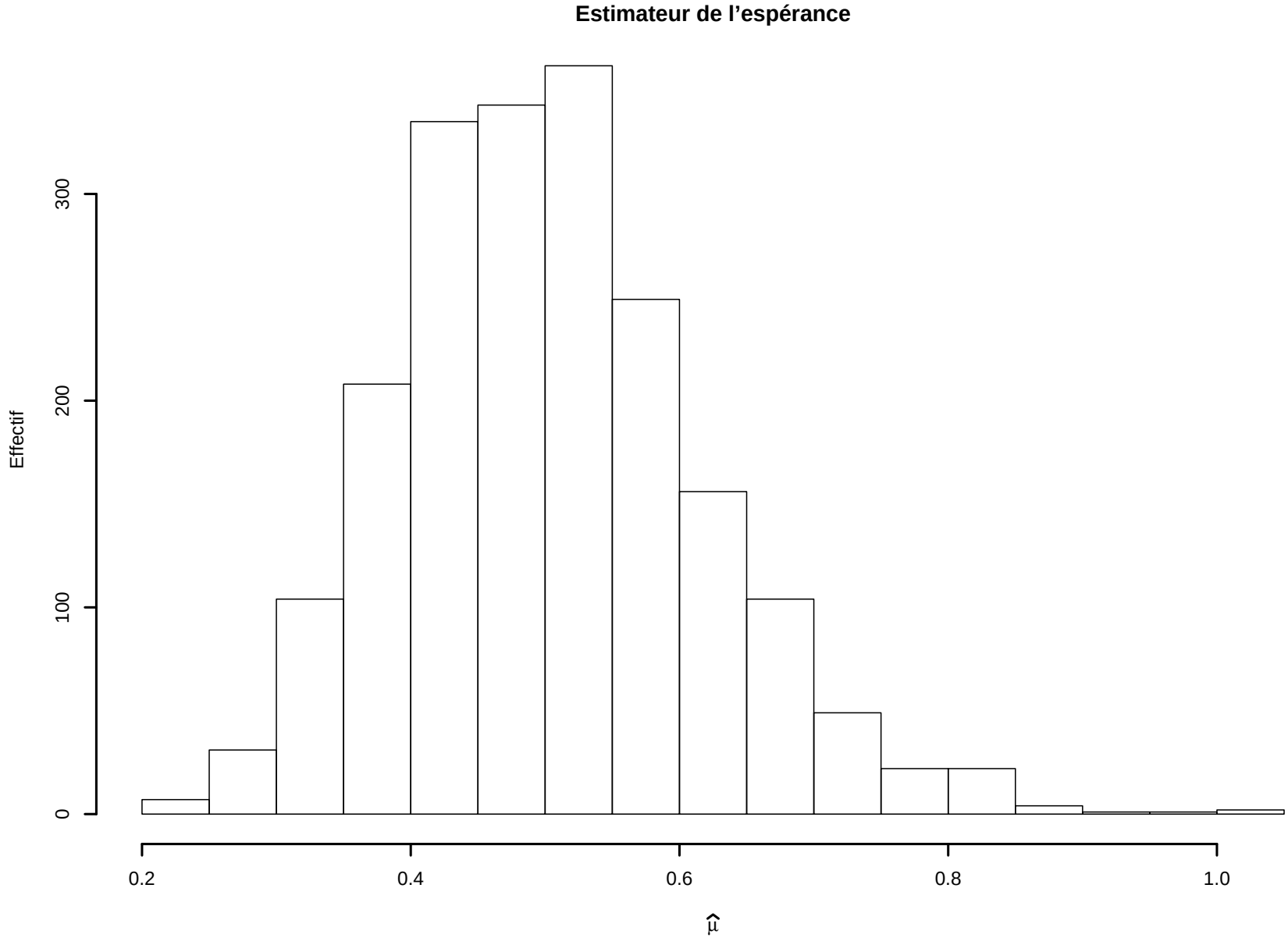
- On suppose les X_i à valeur dans A
- Un estimateur de $\alpha(\theta) \in B$ est une fonction mesurable $\widehat{\alpha}$ de A^n dans l'image de B
- Exemples avec $P_{\mu,\sigma}$ une loi de moyenne μ et d'écart-type σ :
  - Estimateur de μ : la moyenne empirique $\widehat{\mu}(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i$
  - Estimateur de σ^2 : $\widehat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \widehat{\mu})^2$
- **Biais** d'un estimateur : $E(\widehat{\alpha}(X_1, \ldots, X_n)) - \alpha(\theta)$
- Exemples :
  - $E(\widehat{\mu}(X_1, \ldots, X_n)) = \mu$: estimateur sans biais
  - $E\left(\widehat{\sigma^2}(X_1, \ldots, X_n)\right) = \frac{n-1}{n}\sigma^2$: estimateur biaisé (on lui préfère donc $\widehat{\sigma^2} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \widehat{\mu})^2$)

Exemple

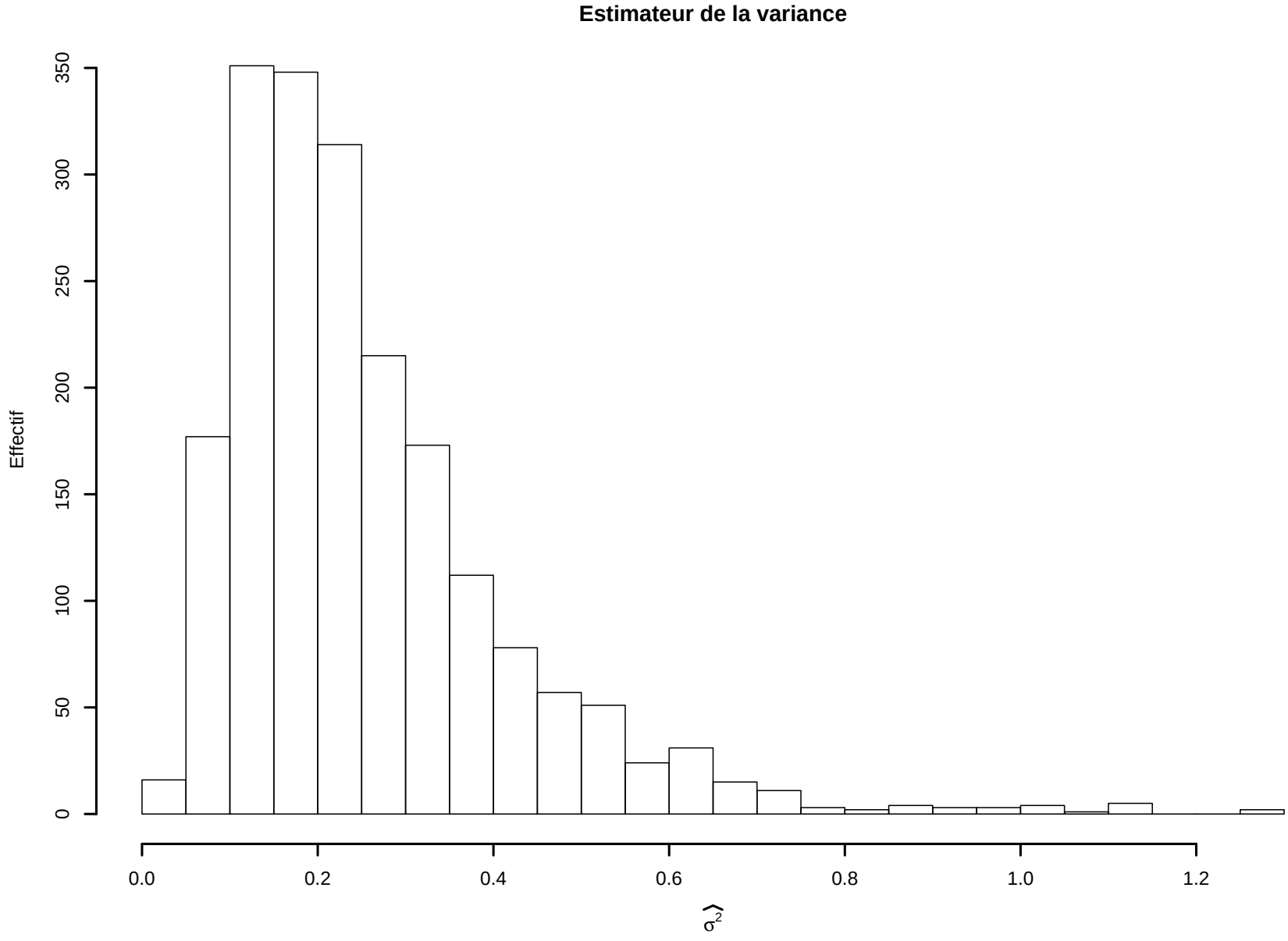
Procédure :

- on prend $X_i \sim \lambda e^{-\lambda x}$, avec $\lambda = 2$, soit :
  - $E(X_i) = \frac{1}{2}$
  - $\sigma^2(X_i) = \frac{1}{4}$
- on considère $n = 20$ réalisations
- on calcule $\widehat{\mu}$ et $\widehat{\sigma^2}$
- on recommence 2000 fois
- pour chaque estimateur, on trace l'histogramme des valeurs obtenues
- on calcule la moyenne des valeurs (approximation de l'espérance d'après la loi des grands nombres) :
  - $E(\widehat{\mu}) \simeq 0.503$
  - $E(\widehat{\sigma^2}) \simeq 0.256$

Exemple (2)


Estimateur de l'espérance
Effectif
0
100
200
300
0.2
0.4
0.6
0.8
1.0
$\hat{\mu}$

Exemple (3)


Estimateur de la variance
Effectif
0
50
100
150
200
250
300
350
0.0
0.2
0.4
0.6
0.8
1.0
1.2
$\widehat{\sigma^2}$

Estimateur (2)

- **Risque** d'un estimateur : $E\left[(\widehat{\alpha}(X_1,\ldots,X_n)-\alpha(\theta))^2\right]$
- **Variance** d'un estimateur : $E\left[(\widehat{\alpha}(X_1,\ldots,X_n)-E(\widehat{\alpha}(X_1,\ldots,X_n)))^2\right]$
- Décomposition biais+variance du risque (on note $\widehat{\alpha}=\widehat{\alpha}(X_1,\ldots,X_n)$) :

$$E\left[(\widehat{\alpha}-\alpha(\theta))^2\right]=(E(\widehat{\alpha})-\alpha(\theta))^2+E\left[(\widehat{\alpha}-E(\widehat{\alpha}))^2\right]$$

- Exemple, risque de la moyenne empirique : $\frac{\sigma^2}{n}$
- Empiriquement (sur l'exemple précédent) :
  - risque de la moyenne : 0.0126 (théorique 0.0125)
  - risque de la variance : 0.0262

Loi forte des grands nombres

- $(X_i)_{i \in \mathbb{N}}$ indépendantes et identiquement distribuées
- $E(X_i) = \mu < \infty$
- Alors

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} X_i = \mu \; p.s.$$

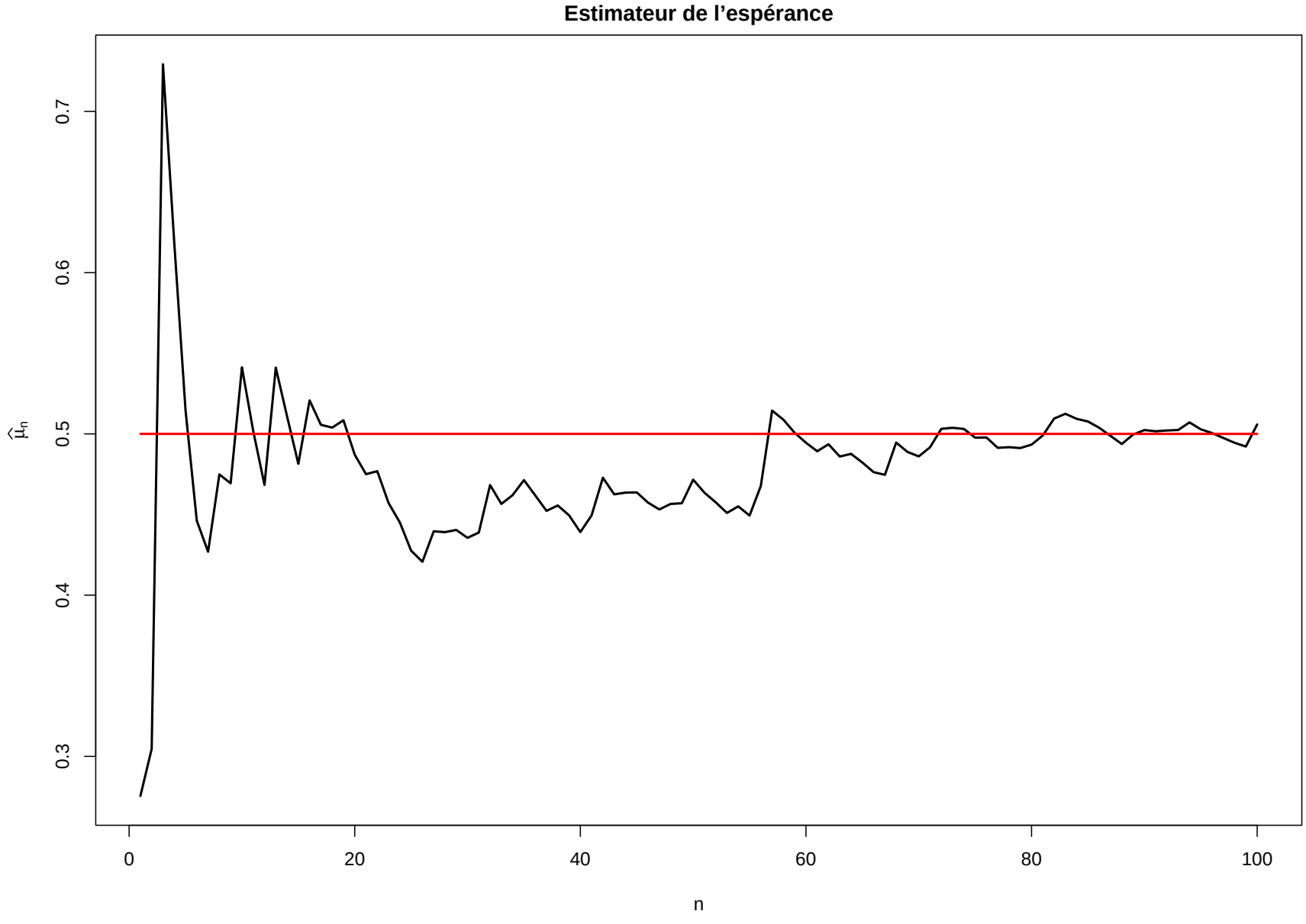
- Version intuitive : la moyenne empirique converge (presque sûrement) vers la moyenne mathématique (i.e., l'espérance)
- Base de l'approche fréquentiste
- Dans le cas de variables indépendantes, l'estimateur de la moyenne est donc **fortement consistant**

Exemple

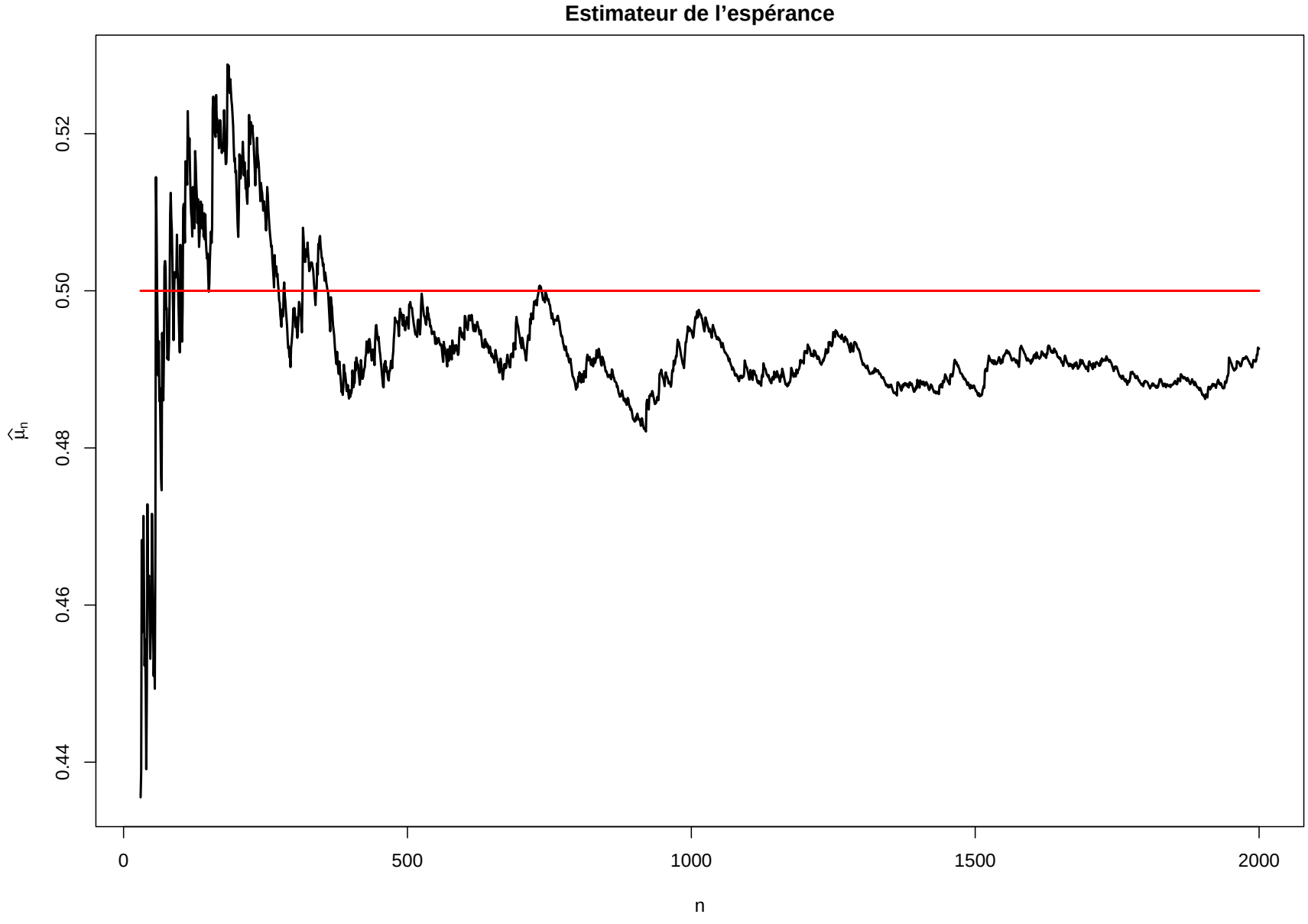
Procédure :

- on prend $X_i \sim \lambda e^{-\lambda x}$, avec $\lambda = 2$, soit :
  - $E(X_i) = \frac{1}{2}$
  - $\sigma^2(X_i) = \frac{1}{4}$
- on considère n réalisations
- on calcule $\widehat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$
- on trace l'évolution de $\widehat{\mu}_n$

Exemple (2)


Estimateur de l'espérance
0.3
0.4
0.5
0.6
0.7
$\hat{\mu}_n$
0
20
40
60
80
100
n

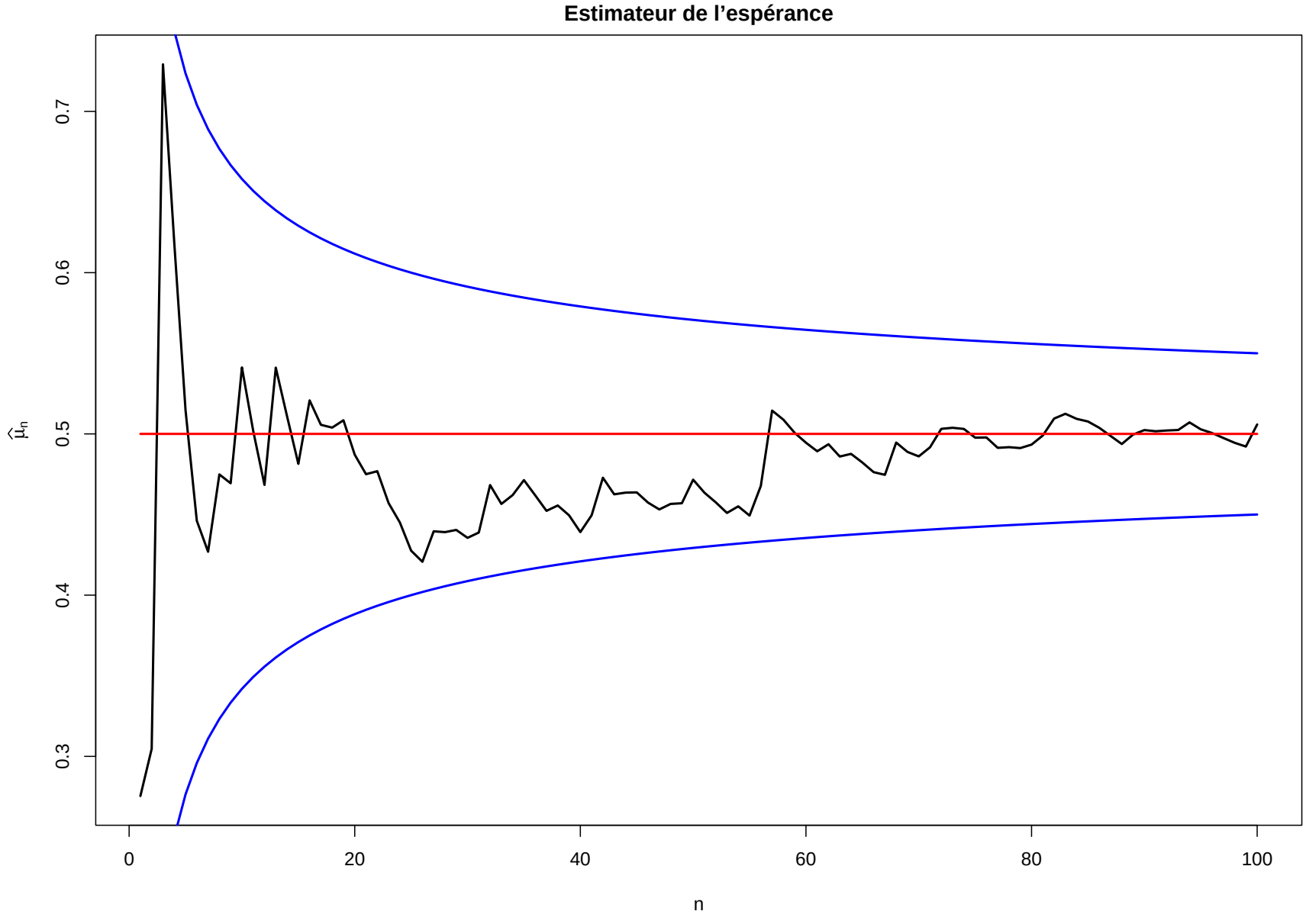
Exemple (3)


Estimateur de l'espérance
0.44
0.46
0.48
0.50
0.52
$\hat{\mu}_n$
0
500
1000
1500
2000
n

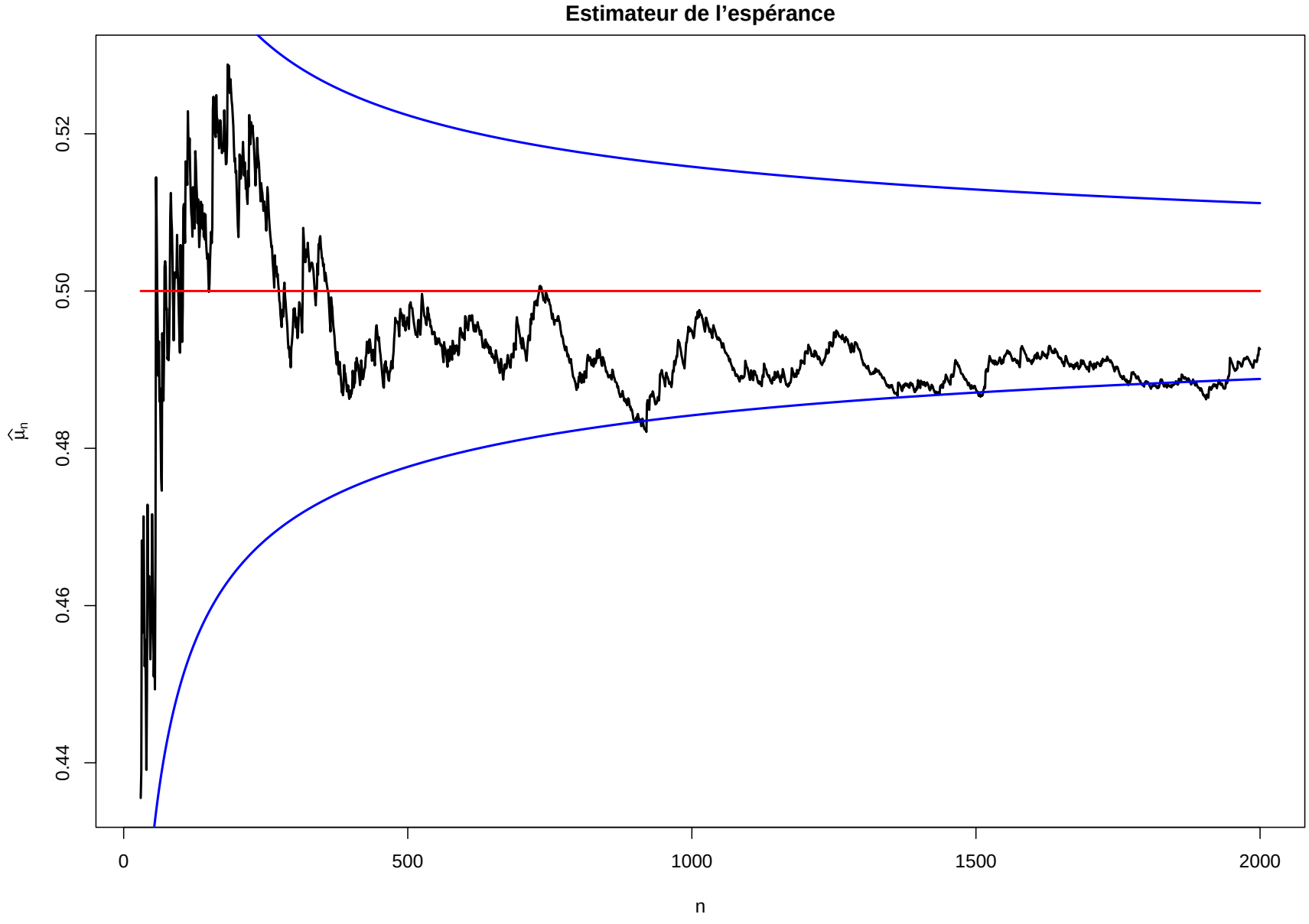
Théorème de la limite centrale

- $(X_i)_{i\in\mathbb{N}}$ indépendantes et identiquement distribuées
- $E(X_i) = \mu < \infty$ et $\sigma^2(X_i) = \sigma^2 < \infty$
- Alors la suite $Y_n = \sqrt{n}(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu)$ converge en loi vers $\mathcal{N}(0, \sigma)$
- Signification mathématique : l'intégrale au sens de la loi P_{Y_n} de toute fonction continue bornée converge vers l'intégrale de cette fonction au sens d'une loi normale $\mathcal{N}(0, \sigma)$
- Signification intuitive : quand n est grand, Y_n se comporte presque comme une variable aléatoire gaussienne
- Donne une idée de la dispersion de la moyenne empirique autour de la moyenne mathématique

Suite de l'exemple (4)



Suite de l'exemple (5)

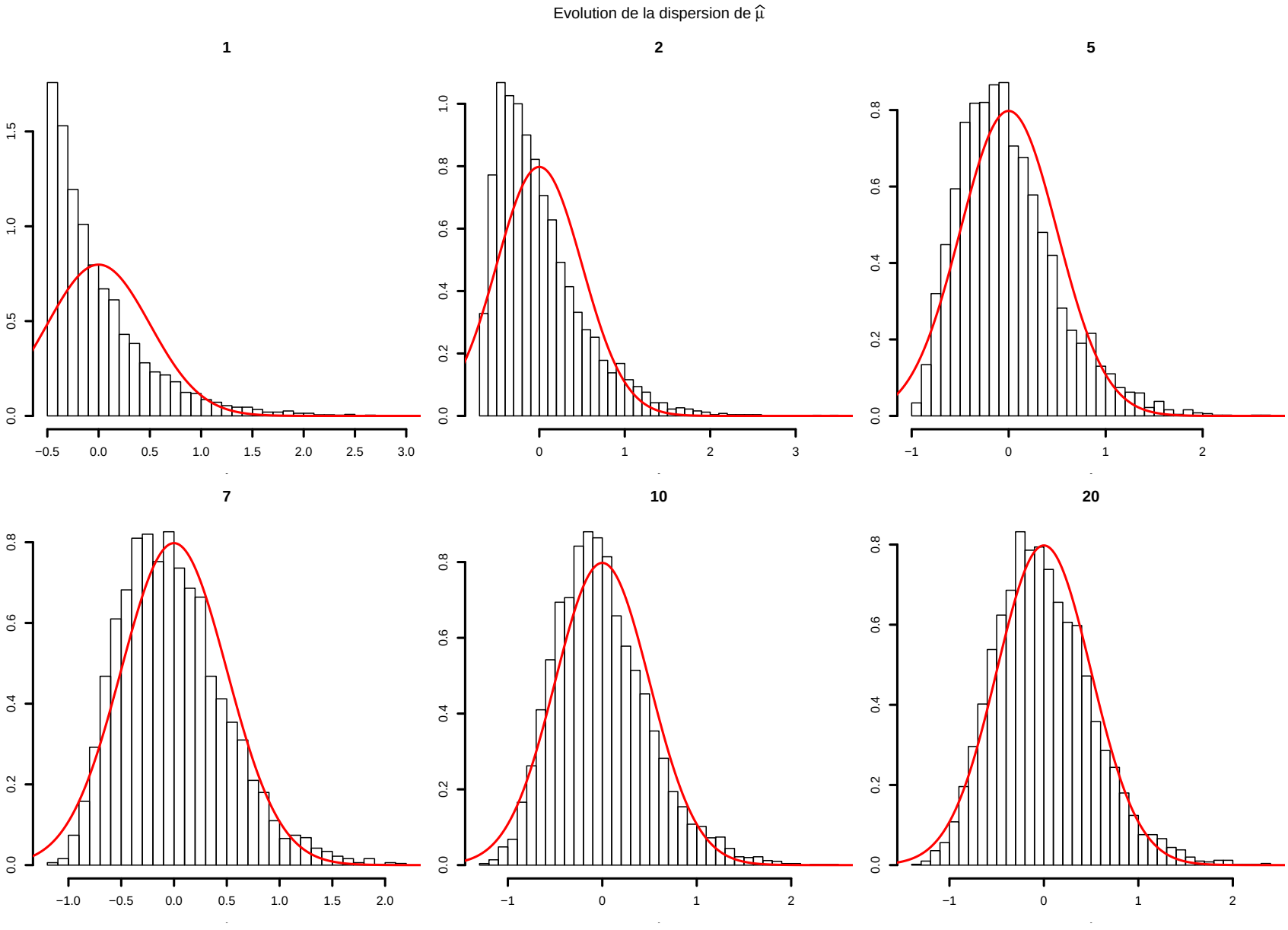


Une autre vérification expérimentale

Procédure :

- on prend $X_i \sim \lambda e^{-\lambda x}$, avec $\lambda = 2$, soit :
  - $E(X_i) = \frac{1}{2}$
  - $\sigma^2(X_i) = \frac{1}{4}$
- pour différentes valeurs de n :
  - on considère n réalisations
  - on calcule $\sqrt{n}(\widehat{\mu}_n - \mu)$
  - on recommence k fois
  - on trace l'histogramme des valeurs obtenues
- on étudie l'évolution de l'histogramme quand n augmente

Résultat



Vraisemblance

- Modèle paramétrique, $P \in \{P_\theta | \theta \in \Theta\}$ dominé : chaque P_θ possède une densité
- Vraisemblance : $\mathcal{L}(\theta, .)$, une densité pour P_θ
- Vraisemblance d'un échantillon i.i.d. : $\mathcal{L}(\theta, x_1, \ldots, x_n) = \prod_{i=1}^{n} \mathcal{L}(\theta, x_i)$
- Estimateur du **maximum de vraisemblance** : estimateur $\widehat{\theta}$ qui maximise $\mathcal{L}(\theta, x_1, \ldots, x_n)$, i.e.

$$\mathcal{L}\left(\widehat{\theta}(x_1, \ldots, x_n), x_1, \ldots, x_n\right) = \sup_{\theta \in \Theta} \mathcal{L}(\theta, x_1, \ldots, x_n)$$

- Idée intuitive : la vraisemblance $\mathcal{L}(\theta, x_1, \ldots, x_n)$ donne une idée de la probabilité d'obtenir les observations si celles-ci sont engendrées par P_θ
- Cas discret : mise en œuvre de l'idée intuitive

Vraisemblance (2)

- En général, on maximise la log-vraisemblance, i.e.

$$\sum_{i=1}^{n} \log \mathcal{L}(\theta, x_i)$$

- Exemple, pile ou face truqué :
  - paramètre p (probabilité d'avoir pile, soit $x_i = 1$)
  - observations i.i.d. (k fois pile)

$$\mathcal{L}(p, x^1, \ldots, x^n) = \prod_{x^i=1} p \prod_{x^i=0} (1-p) = p^k (1-p)^{n-k}$$

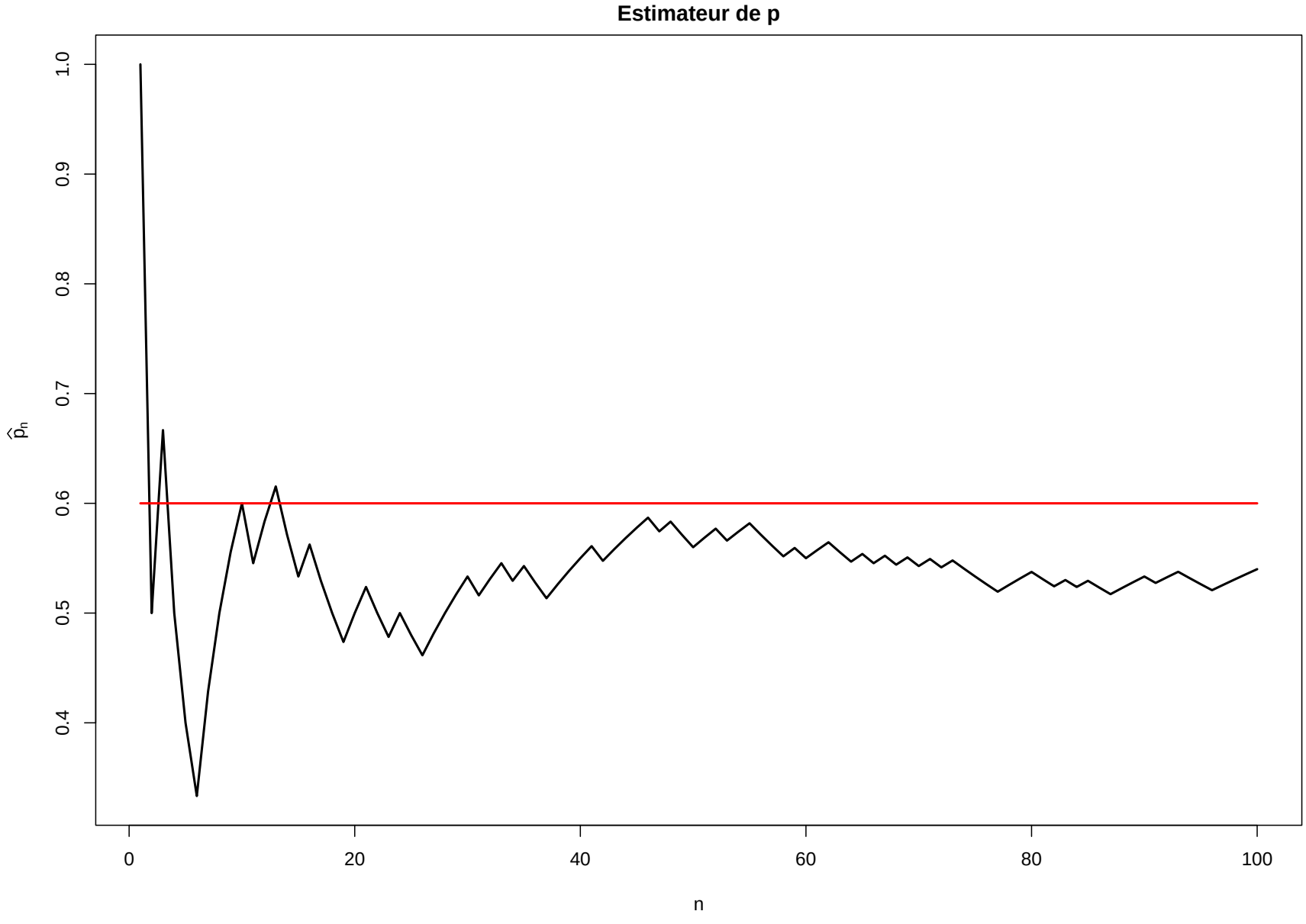
  - estimateur du maximum de vraisemblance : $\widehat{p} = \frac{k}{n}$
- La moyenne empirique est en général l'estimateur du maximum de vraisemblance pour l'espérance

Exemple de mise en œuvre

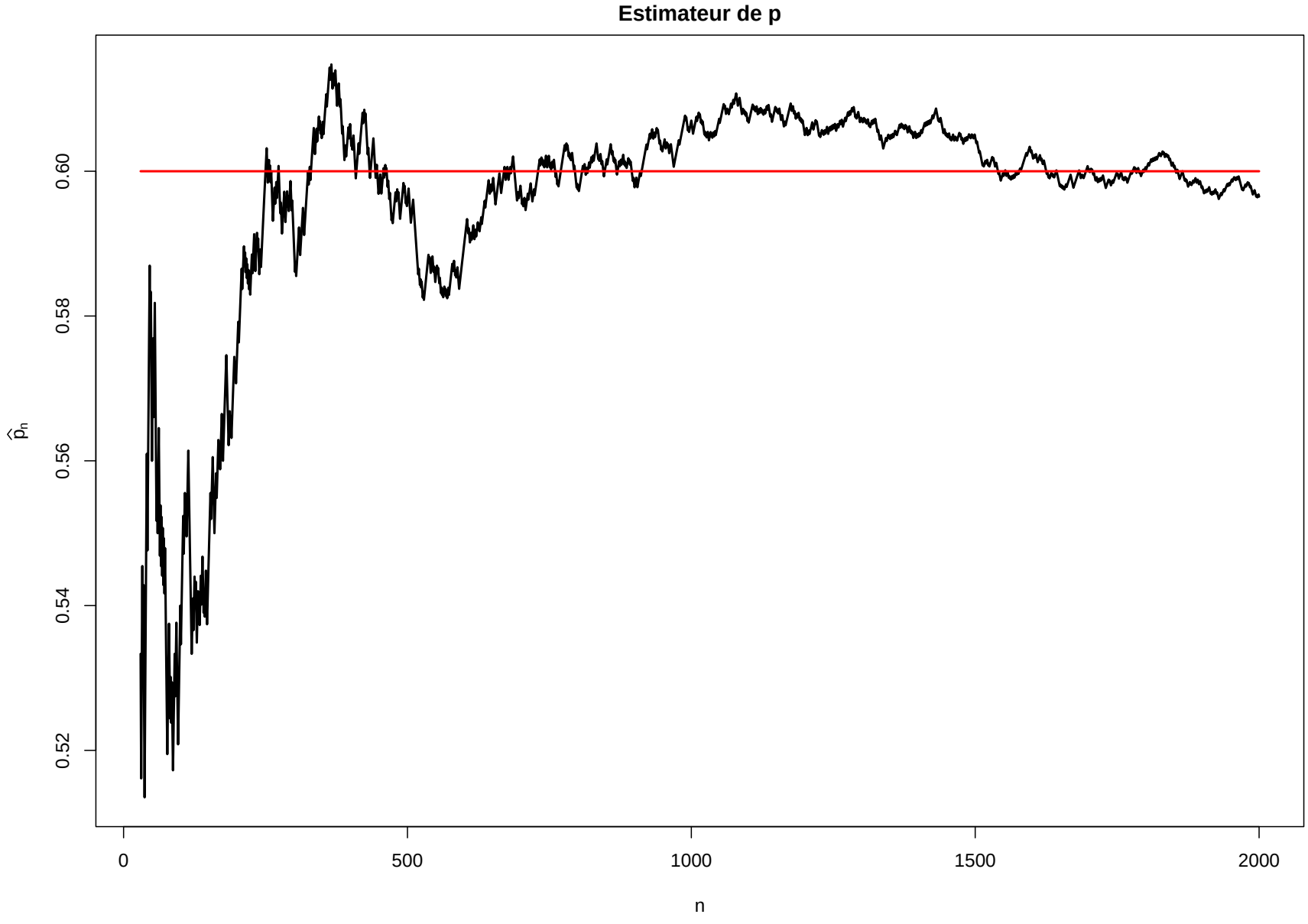
Procédure :

- X_i : pile ou face (probabilité p d'avoir pile)
- k_n : nombre de piles obtenus pour n lancés
- on trace l'évolution de $\widehat{p}_n = \frac{k_n}{n}$
- d'après la loi des grands nombres, $\widehat{p}_n$ converge vers p

Exemple (2)



Exemple (3)


Estimateur de p
$\hat{p}_n$
0.52
0.54
0.56
0.58
0.60
0
500
1000
1500
2000
n

Vraisemblance (3)

Problèmes possibles :

- n'existe pas toujours
- n'est pas toujours sans biais. Exemple :
  - $X_i \sim N(0, \sigma)$
  - $$\mathcal{L}(\sigma, x^1, \ldots, x^n) = \frac{1}{(2\sigma^2\pi)^{\frac{n}{2}}} e^{-\frac{\sum_{i=1}^n x_i^2}{2\sigma^2}}$$
  - on montre que l'estimateur du MV de σ^2 est alors $\frac{1}{n}\sum_{i=1}^n x_i^2$ qui est biaisé !

Néanmoins très pratique, en particulier car un estimateur du MV est en général consistant.

Estimation bayesienne

Idée : mettre une distribution sur Θ, un *a priori*.

- Règle de Bayes

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- Permet d'estimer $P(\theta|x_1, \ldots, x_n)$ grâce à :
  - $P(x_1, \ldots, x_n|\theta)$: modèle paramétrique
  - $P(\theta)$: *a priori* sur Θ
  - $P(x_1, \ldots, x_n)$: par intégration
- Permet de tenir compte de connaissances expertes

Exemple : pile ou face truqué

- probabilité d'obtenir pile : p
- on observe P_n le nombre de piles obtenues dans n lancés
- $P(P_n = k|p) = \binom{n}{k} p^k (1-p)^{n-k}$
- probabilité *a priori* sur p : uniforme sur $[0, 1]$
- on utilise la règle de Bayes avec densité :

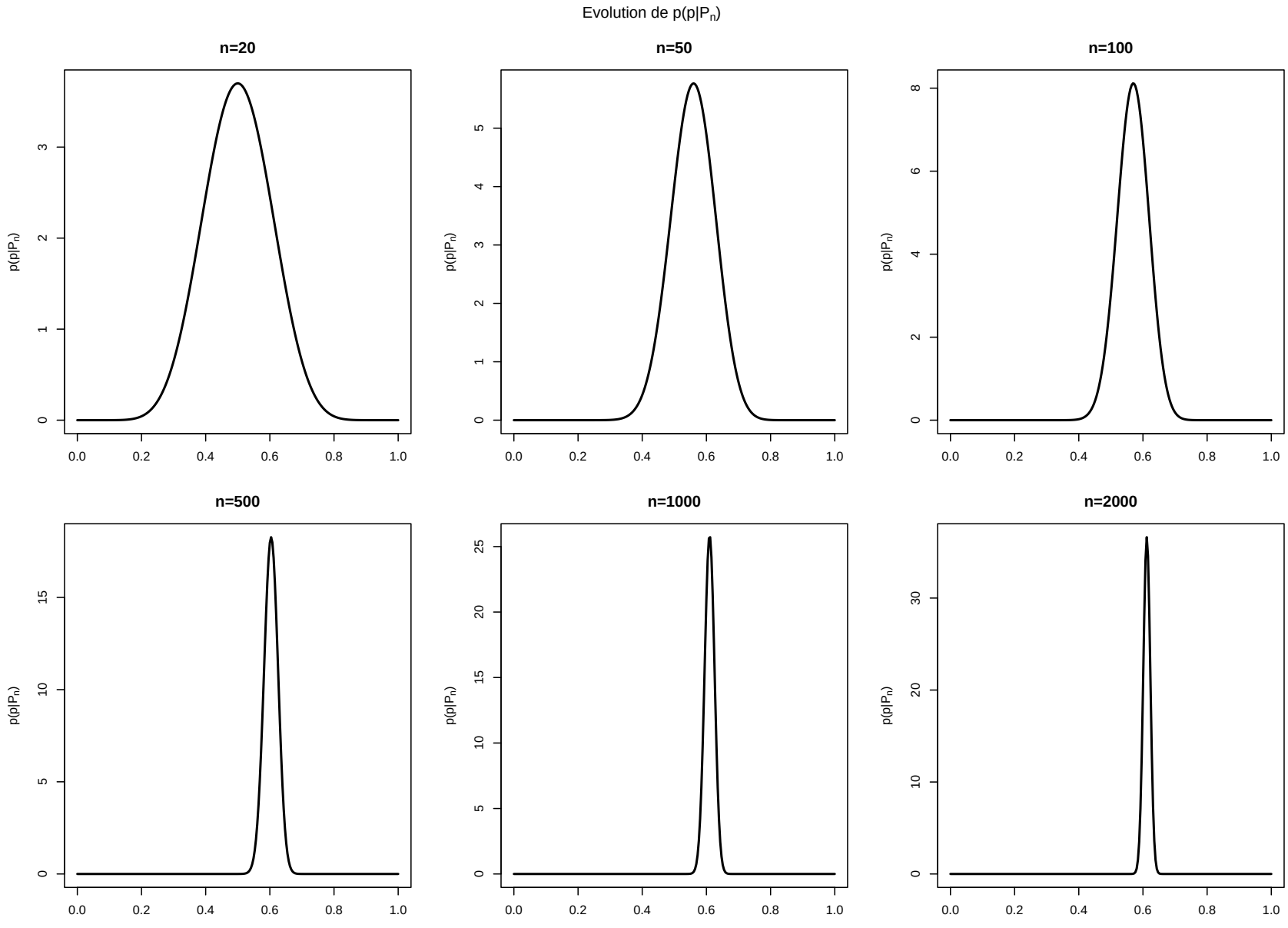
$$p(p|P_n = k) = \frac{P(P_n = k|p)p(p)}{P(P_n = k)}$$

- on obtient $p(p|P_n = k) = \frac{p^k(1-p)^{n-k}}{\int_0^1 p^k(1-p)^{n-k}dp}$
- on reconnaît une loi Beta de paramètres $k+1$ et $n-k+1$, d'espérance $\frac{k+1}{n+2}$ et de variance $\frac{(k+1)(n-k+1)}{(n+2)^2(n+3)}$

Exemple : pile ou face truqué (2)

- loi des grands nombres $\lim_{n\to\infty} \frac{k}{n} = p$
- asymptotiquement $\frac{k+1}{n+2} \sim p$
- asymptotiquement $\frac{(k+1)(n-k+1)}{(n+2)^2(n+3)} \sim \frac{p(1-p)}{n} = \frac{\sigma^2}{n}$
- intuitivement : la distribution *a posteriori* de p se resserre autour de son espérance

Exemple (3)



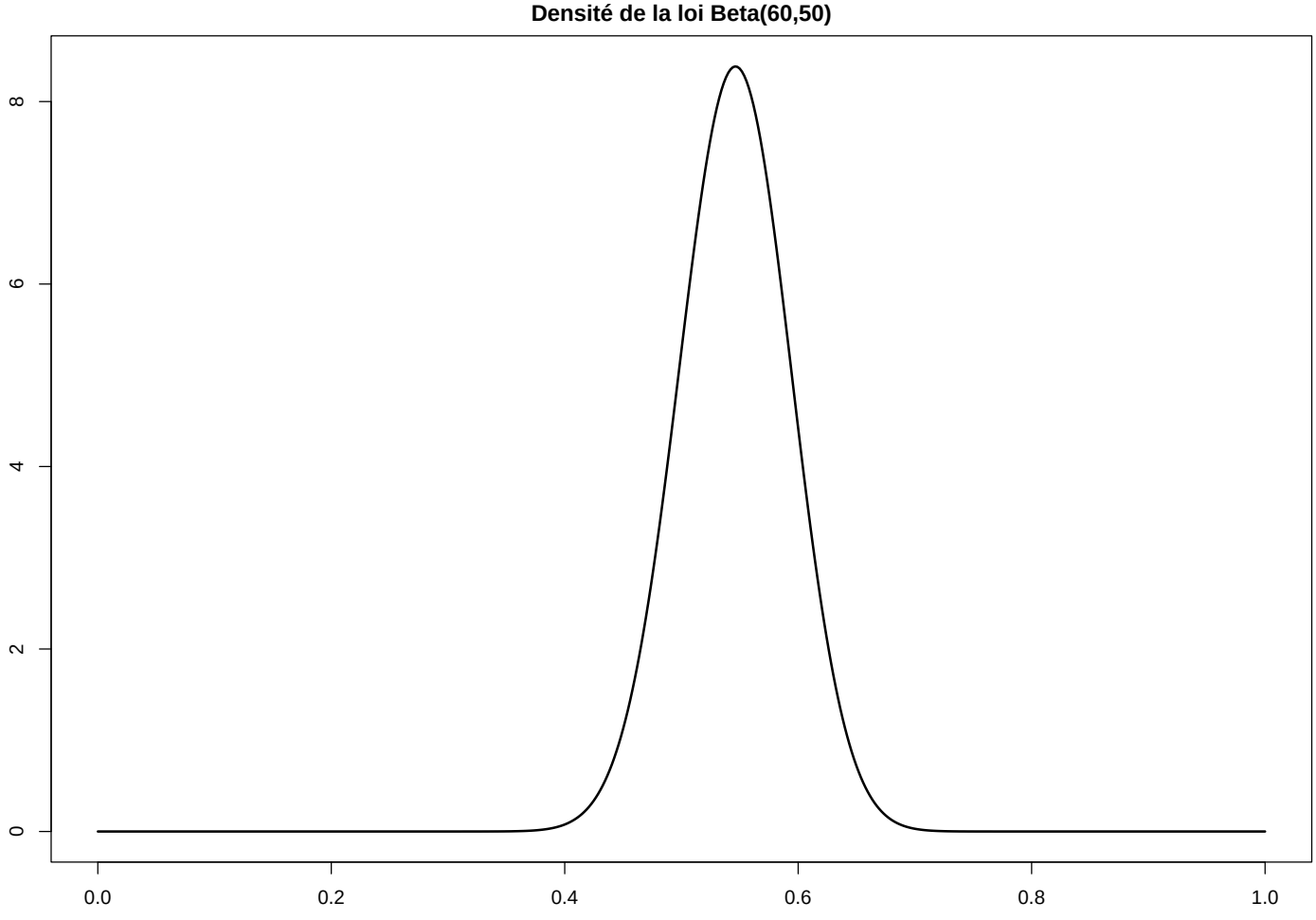
A priori non uniforme

On sait que la pièce est truquée et on a des informations sur la nature du trucage :

- probabilité d'obtenir pile : p distribué selon une loi Beta de paramètres a et b
- on montre alors que $p(p|P_n = k)$ est une Beta de paramètres $k + a$ et $n - k + b$
- espérance de p sachant $P_n = k$: $\frac{k+a}{n+a+b}$
- variance de p sachant $P_n = k$: $\frac{(k+a)(n-k+b)}{(n+a+b)^2(a+b+n+1)}$

Exemple (4)

A priori : loi $Beta(60, 50)$



Exemple (5)